# Advances in Optimal Transport : Low-Rank Structures and Applications in Machine Learning

Meyer Scetbon

▶ **To cite this version:**

Meyer Scetbon. Advances in Optimal Transport : Low-Rank Structures and Applications in Machine Learning. Optimization and Control [math.OC]. Institut Polytechnique de Paris, 2023. English. NNT : 2023IPPAG002 . tel-04100457

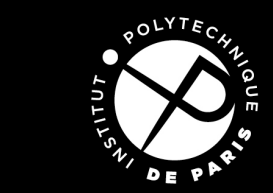
**HAL Id: tel-04100457**

**https://theses.hal.science/tel-04100457v1**

Submitted on 17 May 2023

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2023IPPAG002

Thèse de doctorat

# Advances in Optimal Transport: Low-Rank Structures and Applications in Machine Learning

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à l'École Nationale de la Statistique et de l'Administration Économique

École doctorale n°574 École doctorale de mathématiques Hadamard (EDMH)
Spécialité de doctorat : Mathématiques appliquées

Thèse présentée et soutenue à Palaiseau, le 14 Avril 2023, par

**MEYER SCETBON**

Composition du Jury :

| | |
|---|---|
| François-Xavier Vialard<br>Professeur, Université Gustave Eiffel | Président |
| Philippe Rigollet<br>Professeur, MIT | Rapporteur |
| Facundo Mémoli<br>Professeur, OSU | Rapporteur |
| Anna Korba<br>Professeur Assistant, ENSAE | Examinateur |
| Rémi Flamary<br>Professeur, École Polytechnique | Examinateur |
| Marco Cuturi<br>Professeur, ENSAE | Directeur de thèse |

# Advances in Optimal Transport: Low-Rank Structures and Applications in Machine Learning

Meyer Scetbon

April 2023

**Abstract**

Optimal transport (OT) plays an increasingly important role in machine learning (ML) to compare probability distributions. The OT problem has been used in many applications, and stated with a wide variety of formulations. Among these the Monge ansatz and the Kantorovich linear program stand out. The former involves finding an efficient push-forward map that can morph a measure onto another, while the latter relax the matching of the measures by allowing the splitting of masses. Kantorovich OT is far more amenable to computations and has been the main focus in data sciences. Yet, it poses, in its original form, several challenges when used for applied problems: (i) computing OT between discrete distributions amounts to solving a large and expensive network flow problem which requires a supercubic complexity in the number of points; (ii) estimating OT using sampled measures is doomed by the curse of dimensionality. These issues can be mitigated using an entropic regularization, solved with the Sinkhorn algorithm, which improves on both statistical and computational aspects. While much faster, entropic OT still requires a quadratic complexity with respect to the number of points and therefore remains prohibitive for large-scale problems. Seizing this opportunity, I devoted a significant part of my thesis to work on scalable approaches to OT, which led to my line of work on the introduction of low-rank optimal transport (LOT). I also realized that the fundamental idea proposed by Kantorovich to relax OT could be applied in other settings, and I proposed new approaches using this very same idea to tackle the fair division problem and the adversarial attacks problem through the lens of OT. This thesis is therefore divided in two main parts. In the first part, I present new regularization approaches for the OT problem, as well as its quadratic extension, the Gromov-Wasserstein (GW) problem, by imposing low-rank structures on couplings. This yields a linear complexity both in time and memory with respect to the number of points. In my first attempt towards that goal, I proposed to approximate the iterations of the Sinkhorn algorithm solving entropic OT by forcing a specific low-rank factorization of the kernel involved, resulting in a low non-negative rank factorization of the optimal coupling. Then I propose to generalize this idea and to directly solve the OT problem as well as the GW problem under low non-negative rank constraints on the admissible couplings. We show that these new regularization schemes have better computational and statistical performances compared to the entropic approach and that they can even reach a linear complexity under low-rank assumptions on the ground cost matrices. These new computational schemes pave the way for the use of OT in the large-scale setting. In a second part, I present two settings where the fundamental idea proposed by Kantorovich to relax the OT problem can also be applied, offering

new perspective on longstanding ML problems. More precisely, we propose to relax and lift the fair division problem between multiple agents into the space of distributions by allowing the splitting of resource masses in the partition. By doing so, we show that it is always possible to obtain a fair partition of the resources and we obtain a generalization of the OT problem when multiple costs are involved. We also tackle the problem of adversarial examples using OT. In this problem, the attacker can be represented as a deterministic map that push forward the data distribution towards an adversarial one that aims at maximizing the risk of the classifier. By relaxing the definition of the attacker to be a coupling, we obtain a variational formulation of the adversarial risk which allows us to interpret the adversarial risk minimization problem as a two-player zero-sum game and we study the question of the existence of Nash equilibria in this game.

**Résumé**

Le transport optimal (TO) joue un rôle de plus en plus important en apprentissage automatique (AA) pour comparer des mesures de probabilités. Le problème du TO a été utilisé dans de nombreuses applications et formulé de plusieurs manières. Parmi ces formulations, le problème de Monge et le programme linéaire de Kantorovich se démarquent. La première implique de trouver une transformation efficace pour envoyer une mesure sur une autre, tandis que la seconde relâche la contrainte qu'impose Monge pour faire correspondre des mesures en autorisant la division des masses. Le TO de Kantorovich est beaucoup plus accessible aux calculs et a été la formulation la plus exploitée en sciences des données. Cependant, elle pose, dans sa forme originale, plusieurs défis lorsqu'elle est utilisée pour des problèmes appliqués : (i) calculer le TO entre des distributions discrètes équivaut à résoudre un programme linéaire large et coûteux qui nécessite une complexité super-cubique par rapport aux nombre de points; (ii) estimer le TO en utilisant des mesures échantillonnées est voué à l'échec en raison de la malédiction de la dimensionnalité. Ces problèmes peuvent être atténués en utilisant une régularisation entropique, résolue avec l'algorithme Sinkhorn, qui améliore à la fois les aspects statistiques et computationnels. Bien que beaucoup plus rapide, le TO entropique nécessite toujours une complexité quadratique par rapport au nombre de points et reste donc prohibitif pour les problèmes à grande échelle. Profitant de cette opportunité, j'ai consacré une partie importante de ma thèse à travailler sur des nouvelles approches de calculs pour le TO, ce qui a conduit à ma ligne de travail sur l'introduction du transport optimal de faible rang. J'ai également réalisé que l'idée fondamentale proposée par Kantorovich pour relaxer le TO pouvait être appliquée dans d'autres contextes, et j'ai proposé de nouvelles approches en utilisant cette même idée pour aborder le problème de la division équitable et le problème des attaques adverses à travers le prisme du TO. Cette thèse est donc divisée en deux parties principales. Dans la première partie, je présente de nouvelles approches de régularisation pour le problème du TO, ainsi que son extension quadratique, le problème de Gromov-Wasserstein (GW), en imposant des structures de faible rang sur les couplages. Les algorithmes obtenus possèdent une complexité linéaire à la fois en temps et en mémoire par rapport au nombre de points et permettent donc l'application du transport et ses extensions dans le regime d'un très grands nombre de points. Dans ma première tentative vers cet objectif, je propose d'approcher les itérations de Sinkhorn résolvant le TO entropique en imposant une factorisation de faible rang spécifique du noyau associé, ce qui donne une factorisation de rang non négatif faible du couplage optimal. Ensuite, je propose de généraliser cette idée et de résoudre le problème du TO ainsi que le problème GW en imposant directement

une contrainte de rang non négatif faible sur les couplages admissibles dans le problème d'optimisation du transport. Nous montrons que ces nouveaux schémas de régularisation ont de meilleures performances computationnelles et statistiques que l'approche entropique et qu'ils peuvent même atteindre une complexité linéaire sous des hypothèses de rang faible sur les matrices de coûts associés au problème de transport. Ces nouveaux schémas de calcul ouvrent la voie à l'utilisation du TO à grande échelle. Dans une deuxième partie, je présente deux contextes où l'idée fondamentale proposée par Kantorovich pour résoudre le problème de l'OT peut également être appliquée, offrant ainsi une nouvelle perspective sur des problèmes de ML de longue date. Plus précisément, nous proposons de relaxer le problème de division équitable entre plusieurs agents dans l'espace des distributions en permettant la division des masses de ressources dans leur repartition. Ce faisant, nous montrons qu'il est toujours possible d'obtenir une partition équitable des ressources et nous obtenons une généralisation du problème du TO lorsqu'il y a plusieurs coûts impliqués. Nous abordons également le problème des exemples adverses à l'aide du TO. Dans ce problème, l'attaquant est représenté sous forme d'une fonction déterministe qui projette la distribution des données vers une distribution adverse visant à maximiser le risque du classificateur. En relaxant la définition de l'attaquant pour qu'il soit non plus une fonction mais un couplage, nous obtenons une formulation variationnelle du risque adverse qui nous permet d'interpréter le problème de minimisation du risque adverse comme un jeu à somme nulle à deux joueurs et nous étudions la question de l'existence d'équilibres de Nash dans ce jeu.

# Contents

# Remerciements

Je tiens tout d'abord à te remercier, Marco, pour tout ce que tu m'as appris et apporté durant ces trois années passées a tes cotés. Je te remercie pour ton exemplarité en tant que chercheur qui restera toujours un modèle pour moi, ton exigence que je respecte profondément et qui m'a toujours encouragé à m'améliorer, ton mentorat qui m'a fait grandir tout au long de cette thèse, ta confiance qui m'a permis d'explorer d'autres domaines et qui n'a fait que croitre ma passion pour la recherche, ton soutien qui a été un des piliers de mon épanouissement, et enfin ta générosité à travers tous les conseils que tu m'as prodigué sans limite et qui s'étendent bien au delà de la recherche. Pour tout cela Marco, je te remercie. J'espère que notre collaboration se poursuivra bien après la fin de ma thèse.

I would like to express my sincere gratitude to the members of the jury, whose presence honors me deeply. Thank you very much Facundo for taking the time to review my thesis and thank you for the kind review. Un grand merci à Philippe d'avoir accepté d'être rapporteur de ma thèse et d'avoir rédigé un rapport si détaillé et encourageant. Merci à Rémi et François-Xavier d'avoir eu la gentillesse d'accepter de faire partie de mon jury, c'est un plaisir et un honneur pour moi que de vous avoir dans mon jury. Enfin, merci Anna, avec qui j'ai tant partagé à la fois professionnellement et humainement depuis plus de trois ans, d'avoir toujours été de bons conseils et de me faire l'honneur de conclure ce chapitre de ma vie en étant membre de mon jury.

Je veux aussi remercier Gabriel, qui m'a tant apporté au cours de cette thèse et que je considère comme un second mentor. Merci Gabriel pour ta disponibilité, ton implication dans nos collaborations, et pour toutes les interactions que nous avons eues au cours des trois dernières années. Ces échanges m'ont beaucoup appris. Merci pour ta perspicacité, qui m'a toujours émerveillé et poussé à progresser, ainsi que pour ta gentillesse constante, que ce soit dans le travail comme dans la vie, et enfin merci pour ton soutien que j'apprécie énormément.

Je tiens aussi à remercier Zaid, sans qui cette thèse n'aurait surement jamais eu lieu. Merci Zaid de m'avoir transmis votre passion pour la recherche et qui me rend plus heureux chaque jour. Merci pour votre bienveillance qui m'a toujours extrêmement touché, merci de m'avoir fait confiance en acceptant de me superviser avant le commencement de ma thèse, et surtout merci pour tout ce temps que vous m'avez consacré. Mes séjours passées à vos cotés seront à jamais gravé dans ma mémoire comme la naissance de ma passion pour la recherche.

Je veux aussi remercier Gaël qui m'a donné la chance de découvrir ce monde de la recherche. Merci Gaël de m'avoir initié à la recherche sous ta supervision, pour ton si bel accueil dans ton équipe que je trouve extraordinaire à tous les niveaux, pour ton enthousiasme qui me réjouissait de progresser chaque jour à tes cotés, pour ta générosité dans le partage et l'échange des idées qui me donnait l'envie d'explorer d'avantage, et pour ton implication dans ta supervision qui a rendu cette expérience exceptionnelle.

I am also very grateful to Michael Elad for supervising me for more than four months in his laboratory in Haifa. Thank you Miki for your trust that allowed me to discover a bit of your research field and that I found fascinating, for giving me the chance to work with you, for welcoming me as you did in your group, for your listening and learning that made me discover new research perspectives and for your support even after all these years that I greatly appreciate.

Je tiens aussi à remercier Elvis avec qui j'ai eu la chance de collaborer étroitement pendant plus de cinq mois sous sa supervision. Merci Elvis pour m'avoir offert l'opportunité de découvrir la recherche en dehors du monde académique, pour m'avoir initié à de nouvelles problématiques et guidé tout au long de notre collaboration, pour m'avoir laisser la liberté d'explorer de nouvelles pistes qui n'étaient pas forcément liées au projet initial, pour ta bonne humeur et ta sympathie, et pour avoir passer toutes ces heures côte à côte devant un tableau à se triturer les méninges tout en partageant des moments de complicité qui demeureront des souvenirs mémorables.

Thank you to all my collaborators who made this thesis so rewarding and who gave me the opportunity to discover new areas that I particularly like. Thank you to Peyman for guiding me alongside Miki during my wonderful experience in Israël, thank you to Yaniv for your precious help and expertise, thank you to Nicholas and Soumik for all our discussions and for your involvement which brought me so much, and finally thank you to Rafael, Yann, and Jamal with whom I had the pleasure and the chance to collaborate closely on a fascinating project.

Je n'aurais jamais pris autant de plaisir durant ces trois dernières années sans la rencontre de camarades formidables. Je veux donc remercier tous les doctorants et les chercheurs du CREST qui ont rendu cette aventure mémorable. Je tiens particulièrement à remercier Nicolas pour tous les moments qu'on a partagé et qui j'espère sont loin d'être les derniers, merci à Jaouad pour ta gentillesse et ta bonne humeur, merci à mes camarades du transport, Théo et Nina, qui ont rendu notre bureau si sympathique, et merci également à Flore, Julien, Suzanne, Amir, Arshak, Etienne, Nayel, François-Pierre, Boris, Avo, Geoffrey, Badr, Gabriel, Jules, et Jérémy. Je veux aussi remercier les membres permanents du labo pour avoir rendu mon expérience au CREST si enrichissante. Merci à Arnak, Nicolas, Victor-Emmanuel, Cristina, Matthieu, Guillaume et Sacha. Je tiens aussi à remercier particulièrement Vianney avec qui j'ai tant partager à la fois professionnellement mais surtout humainement, que ce soit à pied ou à vélo, en France ou aux Etats-Unis, devant un match de NBA ou de football, et qui m'a toujours épaulé (parfois même littéralement) et ce depuis mon arrivé à l'ENS de Cachan. Pour tout cela, Vianney, et pour tout le reste, je te remercie.

Je voudrais aussi remercier mes merveilleux camarades que j'ai eu la chance de rencontrer et de côtoyer le long de cette aventure. Merci à l'équipe (anciennement) Parietal de m'avoir si bien accueilli et dont je garderai toujours un souvenir intense. En particulier, je veux remercier Bertrand, Alex, Thomas, David, Pierre, Mathurin, Arthur, et Hicham. I also want to thank my amazing Seattle comrades, namely Vincent, Ema, Lang, and Krishna. Finally, I want to thank my warm comrades from Technion, namely Gregory, Dror, Alona, and Aviad.

Je tiens aussi à remercier mes amis qui me sont si chers et avec qui la vie est chaque jour plus agréable. Trop nombreux pour être cité, je tiens à tous vous remercier pour tout ce que vous m'apportez, pour tous ces moments de joies, pour votre soutien, pour votre présence, et surtout merci pour toute votre affection et votre amour, merci à vous mes frères de cœur. Parmi mes amis, je veux mettre l'accent sur un ami en particulier avec qui j'ai énormément partagé tout au long de cette thèse. Je veux parler de mon ami Laurent. En commençant par la prépa, jusqu'à la thèse, en passant par nos masters ou encore notre virée à Londres pour se familiariser avec le monde de la finance quantitative, nous n'avons cessé de partager des moments forts ces dix dernières années et je te remercie pour tout cela. Je te remercie aussi pour ces magnifiques collaborations et interactions que nous avons eu tout le long de nos thèses, qui nous ont "sauvé" de l'ennuie du COVID, mais qui surtout m'ont fait apprécier d'autant plus la recherche lorsqu'elle menée avec un ami. Merci Laurent.

J'aimerais conclure ce passage pour remercier profondément ma famille. Merci pour votre Amour et votre soutien inconditionnel. Si j'en suis ici aujourd'hui, c'est grâce à vous. Merci de toujours me conforter dans mes choix qui me rendent si heureux aujourd'hui, d'être cette force motrice que me fait avancer et progresser chaque jour, de ne jamais laisser le doute s'installer et d'être toujours là pour moi. Merci à toi Maman de me soutenir avec tant de force et de me faire croire que tout est possible, sans toi je ne serai pas là. Merci à toi Papa de toujours veiller sur moi et pour ta gentillesse infinie que seul un père comme toi sait donner, tu es mon repère et mon modèle dans la vie. Merci à toi Cess d'être la plus merveilleuse des sœurs qu'un frère puisse avoir, tu es rayonnante par ta bonté et ton intelligence et je suis si fier d'être ton frère. Merci à toi Gary de me faire rêvé depuis mon plus jeune age avec tes incroyables aventures qui m'ont toujours fasciné. Merci à toi Philo, ma seconde Maman, pour tout l'amour et la bienveillance que tu me donnes. Enfin merci à toi Adva, הנסיכה שלי , de me rendre plus heureux chaque jour passé à tes cotés. Ces mots ne suffiront jamais à exprimer tout l'amour et la reconnaissance que je vous porte. Vous êtes la source même de mon épanouissement et de ma motivation, je vous aime.

# Introduction

# List of Contributions

During my PhD years, I had the chance to study different areas of machine learning through different collaborations, resulting in publications. For the sake of consistency, I will only discuss in this thesis the contributions concerning the computational aspects of optimal transport and the applications of optimal transport in machine learning. Below is the list of contributions that will be presented in this thesis.

- Low-rank Optimal Transport: Approximation, Statistics and Debiasing, M.S., Marco Cuturi, in *Advances in Neural Information Processing Systems 36*, 2022 [1]
- Linear-Time Gromov Wasserstein Distances using Low Rank Couplings and Costs, M.S., Gabriel Peyré, in *Proceedings of the 37th International Conference on Machine Learning*, 2022 [2]
- Low-Rank Sinkhorn Factorization, Marco Cuturi, Gabriel Peyré, in *Proceedings of the 38th International Conference on Machine Learning*, 2021 [3]
- Mixed Nash Equilibria in the Adversarial Examples Game, Laurent Meunier$^\star$, M.S.$^\star$, Rafael Pinot, Jamal Atif, Yann Chevaleyre in *Proceedings of the 38th International Conference on Machine Learning*, 2021 [4]
- Equitable and Optimal Transport with Multiple Agents, M.S.$^\star$, Laurent Meunier$^\star$, Jamal Atif, Marco Cuturi in *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, 2021 [5]
- Linear Time Sinkhorn Divergences using Positive Features, M.S., Marco Cuturi in *Advances in Neural Information Processing Systems 33*, 2020 [6]

However, I would also like to mention the work I did during my PhD thanks to these collaborations but which will not be developed in this manuscript.

- An Asymptotic Test for Conditional Independence using Analytic Kernel Embeddings, M.S.$^\star$, Laurent Meunier$^\star$, Yaniv Romano, in *Proceedings of the 39th International Conference on Machine Learning*, 2022 [7]

- Triangular Flows for Generative Modeling: Statistical Consistency, Smoothness Classes, and Fast Rates, Nicholas J. Irons, M.S., Soumik Pal, Zaid Harchaoui, in *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, 2022 [8]

- Deep K-SVD Denoising, M.S., Michael Elad, Peyman Milanfar, in *IEEE Transactions on Image Processing*, 2021 [9]

- A Spectral Analysis of Dot-product Kernels, M.S., Zaid Harchaoui, in *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, 2021 [10]

- Harmonic Decompositions of Convolutional Networks, M.S., Zaid Harchaoui, in *Proceedings of the 37th International Conference on Machine Learning*, 2020 [11]

# Outline and Contributions

The ability to compare and manipulate probability distributions is omnipresent in machine learning (ML). For example, supervised learning methods [12] heavily rely on such comparisons to measure the risk. In classification [13], ML practitioners often train a classifier by minimizing the cross-entropy loss [14] between the true conditional distributions of the labels and the learned conditional distributions over the classes. In the regression setting [15], the risk of the prediction is also measured by comparing the true conditional distribution of the data with the learned one under Gaussian assumptions. More generally, maximum likelihood estimation (MLE) [16], which is a standard technique to estimate parameters of a probability distribution that best describes the observed data, aims at minimizing the Kullblack-Leibler (KL) divergence between the true and the modeled distribution. All these fundamental techniques, widely used in ML, rely on the comparison of distributions in order to quantify the uncertainty of the predictions. Being able to compare distributions is also essential in other areas of ML such as in statistical testing [17, 18] or causal discovery [19, 20]. In recent years, this need has received even more attention in the ML community with the development of new generative models capable of solving increasingly complex tasks. Variational Autoencoders [21] are a class of generative models that learn to approximate a target distribution by encoding the data into a lower-dimensional space and involve minimizing a divergence measure between the latent distribution and a prior distribution. Generative Adversarial Networks (GAN) [22] are another popular class of generative models that learn to generate samples by comparing in an adversarial manner the generated probability distribution with the target one. Normalizing flows [23] can be used for both generative modeling as well as density estimation and are learned using MLE. Diffusion models [24] transform a simple base distribution into a more complex target distribution using a sequence of diffusion steps and measure the sequential errors using KL divergences between the generated and the true distributions. Finally, the most illustrative example of this literature is certainly the emergence of transformers [25] which are today one of the most efficient architectures for a wide range of problems in computer vision [26] and natural language processing [27] and aim at treating data as probability distributions. In particular, transformers

deal with signals such as images and sentences by representing them respectively as discrete distributions of patches and words after having encoded the spatial structure of these objects using positional embeddings [28]. Therefore, being able to compare and deal with distributions is becoming an increasingly important challenge in ML and it is in this context that optimal transport (OT) has become a widely used tool.

**OT in data sciences.** When it comes to comparing distributions, the statistics literature provides a rich class of divergence functions to measure the discrepancy between two probability distributions, such as the KL divergence, the total variation (TV) distance, or more generally the family of $\phi$-divergences [29]. Yet, these divergences rely on comparing density functions pointwise, and saturate or diverge when the supports of the probability measures are disjoint limiting their applications only for the comparison of histograms or continuous probability distributions. OT [30] has become an increasingly important alternative in ML thanks to its versatility of applications to compare probability measures. Starting from a cost function (e.g. a distance) on the space on which measures are supported, OT consists in finding a mapping [31] or coupling [32] between both measures that is optimal with respect to that cost. In other words, OT naturally extends the ground cost between two points to a discrepancy function between histograms of points, or probability measures, in the form of an optimization problem. As a result, OT provides a simple and comprehensive framework to compare probability distributions and has inspired many developments in machine learning [33]. A flurry of works have recently connected it to other trending topics, such as normalizing flows or convex neural networks [34, 35, 36], while the scope of its applications has now reached several fields of science such as computed vision [37], signal processing [38, 39, 40], single-cell biology [41], imaging [42, 43], neuroscience [44, 45], graphics [46, 47, 48], or generative modeling [49, 50, 51]. Another major feature of OT is the optimal coupling obtained when solving it which provides an optimal alignment of the probability measures at hand. Such an object, specific to OT and at least as important as the OT cost itself, has found numerous applications in ML to align word embeddings [52, 53, 54], to reconstruct cell trajectories [55, 56], for domain adaptation [57, 58] or even for encoding discrete distributions as in transformers [25] using barycentric projections [59]. In addition, the versatility of the OT framework goes beyond the comparison of probability measures supported on the same space. [60] propose a quadratic version of OT, namely the Gromov-Wasserstein (GW) distance that aims at comparing point clouds or probability measures living in incomparable spaces. While OT seeks an optimal matching

according to a ground cost by minimizing a linear score associated to that cost, GW seeks an assignment that is as close to an isometry as possible, as quantified by a quadratic score. Several problems in ML require comparing datasets that live in heterogeneous spaces. This situation arises typically when realigning two distinct views (or features) from points sampled from similar sources. Recent applications to single-cell genomics [61, 62] provide a case in point: Thousands of cells taken from the same tissue are split in two groups, each processed with a different experimental protocol, resulting in two distinct sets of heterogeneous feature vectors; Despite this heterogeneity, one expects to find a mapping registering points from the first to the second set, since they contain similar overall information. GW has also been used in supervised learning [63], generative modeling [64], domain adaptation [65], structured prediction [66], quantum chemistry [47] and alignment layers [67].

**Challenges of OT in ML.** Solving optimal transport problems at scale poses, however, formidable challenges. The most obvious among them is computational: Instantiating the Kantorovich [32] problem on discrete measures of size $n$ can be solved with a linear program (LP) of complexity $O(n^3 \log n)$ [68, 69, 70]. A second and equally important challenge lies in the statistical performance of using that LP to estimate OT between densities: the LP solution between i.i.d. samples converges exponentially slowly with respect to the ambient dimension to that between densities [71]. [72, 73] obtained refined results and show that estimating OT requires an exponential number of samples w.r.t. the intrinsic dimensionality of the support. It is now increasingly clear that regularizing OT in some way or another is the only way to mitigate these two issues [51, 74, 75]. A popular approach consists in penalizing the OT problem with a strongly convex function of the coupling [76, 77], and some more specific uses of an entropic penalty, to recover so called Sinkhorn divergences [78]. Entropic OT is cheaper to compute than regular OT [79, 80], smooth and programmatically differentiable in their inputs [46, 55], and have a better sample complexity [81, 82]. While entropic OT solvers do lower computational costs from supercubic down to an embarrassingly-parallel quadratic cost, using to compare measures that have more than a few tens of thousands of points remains a challenge. These computational limitations are even more critical in GW, a non-convex quadratic generalization of the OT problem that is NP-hard to solve in general [83]. As OT, GW can be regularized using entropy [84, 85] and [47] propose to apply a mirror descent (MD) scheme to approximate the entropic GW cost that consists in solving a sequence of nested entropic OT problems. Although this heuristic achieves low GW costs, it remains considerably limited in practice due to its cubic complexity w.r.t. the number of

points. Only two broad approaches are known to achieve tractable running times: *(i)* Solve related, yet significantly different, proxies of the GW energy, either by embedding points *as* univariate measures [86, 87], by using a sliced mechanism when restricted to Euclidean settings [88] or by considering tree metrics for supports of each probability measure [89], *(ii)* Reduce the size of the GW problem through quantization of input measures [90] or recursive clustering approaches [91, 62]. However none of these works have tried to accelerate the MD scheme proposed in [47] to approximate GW.

**OT and ML Applications.** Although the OT cost or its solution have been used directly as a loss [49] or in order to align distributions [92] in various ML applications, many links between OT and applied problems remain to be discovered. The origin of the OT theory can be traced back to the 18th century when the French mathematician Gaspard Monge [31] introduced the problem of finding the most efficient way to transport a probability distribution towards another using push forward maps. Although Monge's mathematical formulation of optimal transport was groundbreaking, it was soon discovered that his approach had limitations due to the lack of a provable solution to the problem. It wasn't until 150 years later that significant progress was made in the field of OT theory [93], thanks to the fundamental idea of Kantorovich [32]: He proposed a relaxation of the OT problem by considering probabilistic maps that allowed for the splitting of mass in the matching process. In fact, the limitation encountered by the Monge formulation of OT can be found in several problems which at first sight seem to be unrelated to OT. Fair division [94] has been widely studied by the artificial intelligence [95] and economics [96] communities. It consists in partitioning diverse resources among agents according to some fairness criteria. One of the standard problems in fair division is the fair cake-cutting problem [97, 98]. The cake is an heterogeneous resource, and the agents have heterogeneous preferences over different parts of the cake. Hence, taking into account these preferences, one might share the cake fairly between the agents. This problem has many variants such as the cake-cutting with two cakes [99], or the Multi Type Resource Allocation [100, 101]. However, in all these models it is assumed that there is only one indivisible unit per type of resource available, and therefore partitioning the resources amounts to defining a deterministic map that assigns to each type of resources a unique agent. The constraint on the partition limits considerably the resolution of the fair division problem as it might not admit a solution and yet, no relaxation of the problem has been proposed. Adversarial attacks [102, 103] is also another setting where deterministic maps between distributions appear.

State of the art classifiers are sensitive to imperceptible perturbations of their inputs that make them fail. Last years, research have concentrated on proposing new defense methods [104, 105, 106] and building more and more sophisticated attacks [107, 108, 109, 110]. So far, most defense strategies proved to be vulnerable to these new attacks or are computationally intractable. This asks the following question: can we build classifiers that are robust against any adversarial attack? The answer to this question might be mainly limited due to the restricted definition of the attacker: as defined in adversarial risk, adversarial attacks are defined as argsupermum of the loss over balls centered in the datapoints. Therefore, adversarial attacks can be viewed as deterministic maps that push forward the data distribution toward adversarial distributions. Interestingly, no work has, to our knowledge, tried yet to relax the definition of the adversary to give a principled answer to the above question.

**Contributions of this thesis.** This thesis, started in 2019 under the supervision of Marco Cuturi, makes a few contributions on new computational approaches to tackle large-scale optimal transport problems in machine learning, and studies new methodologies applying optimal transport to solve longstanding challenges in robust optimization and algorithmic fairness. More precisely, the contributions of this thesis are divided into two main parts. In a first part, we present our contributions concerning the algorithmic and theoretical development of new regularization schemes based on low-rank methods to allow the application of optimal transport and its quadratic variant in the large-scale setting. In a second part we present our contributions where optimal transport is used as a tool to understand and study the fair division problem and the problem of adversarial attacks.

In Part II, we present the following four contributions.

- In [6], we propose to speed-up the resolution of entropy regularized OT with the Sinkhorn algorithm by considering a specific low-rank factorization of the kernel matrix $K = \exp(-C/\varepsilon)$ involved in the Sinkhorn iterations. Our low-rank approximation of the kernel $K$ is obtained using parameterized feature maps which associate to any point in the support of the measures a vector in the positive orthant and therefore forces the positiveness of the factorization. We show that our approach can be used to approximate the entropic OT with common cost functions such as the square Euclidean distance. We also illustrate the versatility of our method by extending previously proposed OT-GAN to a new approach that learns adversarially a kernel induced from a positive feature map. This approach is fully differentiable in the feature map and can be used to train a GAN at scale with linear time iterations.

- In [3], we propose a new regularization scheme of the OT problem, called Low-rank Optimal Transport (LOT), which constrains the admissible couplings to have a low nonnegative rank. Instead of factorizing the kernel involved in the Sinkhorn iterations with positive factors, we directly impose a low nonnegative rank constraint on the feasible set of couplings considered in OT problems, with no approximations on the cost or kernel matrices. We introduce a generic approach that can solve the OT problem under low-rank constraints with arbitrary costs. Couplings with low nonnegative rank have a natural low-rank factorization as a product of sub-couplings with a common marginal, which is used to optimize jointly on sub-couplings and the common marginal distribution using a mirror-descent approach. We prove the non-asymptotic stationary convergence of our algorithm, and show that the time complexity of the algorithm is generally quadratic but can become linear when exploiting low rank assumptions on the *cost* (not the kernel which strongly depends on the regularization parameter $\varepsilon$) involved in the OT problem.

- In [1], we aim at improving our knowledge and practical ability to use low-rank factorizations in optimal transport. The paper focuses on the theoretical and practical aspects of low-rank OT (LOT). We generalize the definition of LOT to general probability measures, and derives the rate of convergence of LOT to the true OT for both discrete and general probability measures. We also provides an upper-bound for the statistical error made when estimating LOT using empirical measures and show that it has a parametric rate that is independent of the dimension. We establish links between the bias induced by the low-rank constraints on OT and clustering methods. We introduce a debiased version of LOT that metrizes weak convergence and is suitable for large-scale comparison of measures in machine learning. Finally we propose practical strategies to tune the step-length and initialization of the LOT algorithm, making it a generic and automatized method for the choice of hyperparameters.

- In [2], we focus on the computational aspects of Gromov-Wasserstein and propose a new regularization scheme of the problem based on low-rank constraints. More precisely, we exploit a low-rank factorization of the two input cost matrices to reduce the complexity of recomputing the cost at each iteration of the entropic GW scheme from cubic to quadratic, thereby lowering its total complexity. We show that the low-rank approach for couplings can be used in the GW pipeline to achieve a $O(n^2)$ strategy with no prior assumption on input cost matrices. We also explain why methods that exploit the geometrical properties of the kernels are of little use in a GW setup. We combine both low-rank assumptions on costs and couplings to

achieve GW approximation with linear complexity in time and memory, and demonstrate the effectiveness of our method on simulated and real datasets.

In Part III, we present the following two contributions.

- In our work [5], we introduce EOT (Equitable and Optimal Transport), which is a relaxed version of the fair division problem. In the fair division problem, there are multiple agents who aim to share one or multiple sets of resources, by finding a fair partition of these sets. Here we propose to relax the problem and consider the case where resources are no more sets but rather distributions on these sets where a certain amount of divisible mass is associated to each of the elements. EOT is defined as linear optimization problem under linear constraints that maximizes the minimum of individual utilities. We show that the partition obtained by EOT is equitable, optimal, and proportional, and derive the dual formulation of EOT with strong duality results. We also show that EOT is related to some usual Integral Probability Metrics, and propose an entropic regularized version of the problem with an efficient algorithm similar to the Sinkhorn algorithm to approximate EOT.

- In our work [4], we obtain a game theoretic point of view of the adversarial risk minimization problem using optimal transport. We show that it can be reformulated as a distributionally robust optimization problem over specific Wasserstein balls and we study the existence of Nash equilibria. More precisely, by relaxing the adversary to be a coupling instead of a deterministic map in its original formulation, we obtain a variational formulation of the adversarial risk for deterministic as well as random classifiers, and show that in both settings, the adversarial risk minimization problems can be reformulated as two-players zero-sum games. We show that in the case of mixed strategies, it is always possible to approximate a Nash Equilibrium (and even reach it under some assumptions), meaning that adding randomness in the choice of the classifiers allows to learn a random classifier that is robust to any adversarial perturbations. We also design an algorithm that efficiently learn a finite mixture of classifiers and show empirically improved adversarial robustness over classical deterministic defenses.

We now turn to a more detailed presentation of the chapters constituting this thesis.

# Chapter 1. Optimal Transport: From Monge to Kantorovich

This chapter introduces the key concepts and results on optimal transport on which this thesis builds upon. Because of our focus on ML applications, we state these results for measures supported on $\mathbb{R}^d$. We first present the original Monge formulation of the OT problem and its main limitations, then we present the Kantorovich relaxation and the links between these two formulations, finally we expose some fundamental properties of the Kantorovich OT.

**Monge Optimal Transport.** The original formulation of optimal transport was proposed by Gaspard Monge in 1781, and is known as the Monge problem. Given two measures of equal mass $\mu$ and $\nu$ living in $\mathcal{P}(\mathbb{R}^d)$ and a cost function $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}+$, Monge raised the problem of transporting $\mu$ to $\nu$ optimally w.r.t. $c$. More formally, this problem can be stated as

$$\inf_{T:T\#\mu=\nu} \int_{\mathbb{R}^d} c(x, T(x)) d\mu(x)$$

where $T\#\mu$ is the pushforward measure of $\mu$ by $T$, defined by $T\#\mu(A) := \mu(T^{-1}(A))$ for all $\mu$-measurable sets. When it exists, a transport map satisfying the constraint $T\#\mu = \nu$ assigns to each point $x$ in the support of the initial measure $\mu$ a point $T(x)$ in the support of the target measure $\nu$, and it transports all the mass of $\mu$ located at $x$ to $T(x)$. The Monge problem aims at finding among all these transport maps, one that minimizes the total transportation cost. A sufficient condition for the existence of a transport map is that $\mu$ is atomless, but even when transport maps exist, there may be none that is optimal. More generally, the Monge problem is not always well-posed and even when it is, it can be very hard to solve as both the objective and the constraints are non-convex.

**Kantorovich Optimal Transport.** The Kantorovich formulation of the optimal transport problem relax the Monge problem by seeking instead to minimize the transportation cost over a set of probabilistic maps that specify how much mass is moved from each point in the source distribution to each point in the target distribution. More formally, instead of considering deterministic maps $T$, Kantorovich proposed to consider probabilistic map, i.e. measures over the product space $\mathbb{R}^d \times \mathbb{R}^d$ that have $\mu$ and $\nu$ as marginals:

$$\inf_{\gamma \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} c(x, y) d\gamma(x, y)$$

where $\Pi(\mu,\nu) = \{\gamma \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d) \text{ s.t. } \pi_1\#\gamma = \mu,\ \pi_2\#\gamma = \nu\}$ is the set of transportation plans, and $\pi_1 : (x,y) \to x$, $\pi_2 : (x,y) \to y$ are the canonical projections. The minimizers of this problem are called optimal transport plans between $\mu$ and $\nu$. The Kantorovich formulation is much easier to handle than the Monge problem as it is a linear optimization problem, and a solution to the Kantorovich formulation always exists under weak conditions on the cost function $c$. This new formulatiom of OT is much more flexible as it can handle more general scenarios, including cases where the two distributions have different shapes and sizes, or when one distribution has atoms.

**Links between the Two Formulations.** The relaxation proposed by Kantorovich is in fact a tight extension of the Monge problem. When the initial measure is atomless, [111] shows that the Monge and the Kantorovich formulations coincide, i.e. the two optimal costs are equal. Therefore, the Kantorovich formulation can be seen as the minimal extension of the original Monge problem, which admits a minimizer. We also present some general cases where an optimal Monge map exists and coincide with an optimal coupling solving the Kantorovich optimal transport problem. Specifically, we present the case where the ground cost is of the form $c(x,y) = h(x-y)$ with $h$ a strictly convex function and the initial measure $\mu$ is absolutely continious for which Monge and Kantorivich formulations admit the same unique minimizer. We also present an important special case of the above result, that is when $c$ is the squared Euclidean distance. As shown by Brenier in his seminal paper, the optimal Monge map can be characterized as the gradient of a convex function.

**Some Useful Properties.** Some of the main properties of the Kantorovich optimal transport are also discussed. We introduce the Wasserstein distances that are special cases of the Kantorovich OT when the ground cost is a distance $d(x,y)$ on $\mathbb{R}^d$ to a power $p \geq 1$. These objects define metrics to measure the distance between two probability distributions $\mu$ and $\nu$ with moments of order $p$. The Wasserstein distance satisfies all three metric axioms and it metrizes the weak convergence. We also present the dual formulation of OT defined as an optimization problem that aims to find the supremum cost $\int_{\mathbb{R}^d} f d\mu + \int_{\mathbb{R}^d} g d\nu$ over all possible bounded and continuous functions $f$ and $g$ that satisfy the cost constraint $f \oplus g \leq c$, given $\mu$ and $\nu$ that are probability distributions. More formally the dual OT problem is defined as:

$$\sup \left\{ \int_{\mathbb{R}^d} f d\mu + \int_{\mathbb{R}^d} g d\nu : \text{ s.t. } f, g \in C_b(\mathbb{R}^d) \text{ and } f \oplus g \leq c \right\} .$$

We recall a sufficient condition for the existence of a solution for the dual formulation, which requires that $\mu$ and $\nu$ are compactly and $c$ is continuous. Then under the

same assumption, we present a strong duality result that shows that both the dual and the primal formulation of OT are equal.

# Chapter 2. Optimal Transport: Challenges in Machine Learning

This chapter introduces the practical challenges of applying optimal transport (OT) on data for machine learning applications. The focus is on discrete and finite probability measures, which is the main setting of OT application in machine learning. In particular, the chapter focuses on the discrete formulation of the Monge problem and the challenges in solving it due to its degeneracy. It also covers the discrete formulation of the Kantorovich relaxation, highlighting its limitations in terms of both computational complexity and statistical aspects. Additionally, the chapter introduces the entropy-regularized OT as an approximation of OT that offers improved complexity and faster statistical convergence rates.

**Discrete Optimal Transport.** The discrete optimal transport problem aims at solving OT between probability measures that are discrete (and finite), meaning each measure is a weighted sum of Dirac measures supported on finitely many points. When considering the Monge formulation of OT in the discrete setting which seeks a map minimizing the transportation cost by associating to each point of the initial measure a single point that must push the mass of one measure toward the mass of another, one can encode this map using indices and formulate it as a generalization of the optimal assignment problem. However, both the discrete Monge OT and the assignment problem are limited in that the former is in general degenerate while the latter can only compare uniform histograms of the same size. Additionally, the feasible set for the Monge problem is non-convex, making it difficult to solve in its original formulation. When considering the Kantorovich's approach in the discrete setting which relax the deterministic nature of transportation, allowing mass at a source point to be potentially dispatched across several locations, one can reformulate the OT problem as a simple linear program using the formalism of matrices. This approach defines a valid coupling as a matrix to encode the flexibility of probabilistic transport, which is always symmetric, and the resulting optimal transport problem can be solved using the network simplex algorithm. However solving the problem remains coostly as it requires a supercubic complexity with respect to the number of points and therefore can only be applied for small problems of size smaller than a few thousands of points. The problem of estimating the optimal transport cost between two distributions, $\mu$ and $\nu$, using only samples drawn from these distributions is also presented. A common estimator for the

unknown distance between the true distributions is to compute it between the empirical measures. The rate of convergence of the estimated distance to the true distance is often referred to as the "sample complexity". The sample complexity of the p-Wasserstein distance is presented, which states that the expected value of the absolute difference between the estimated and true distances is of order $n^{-1/d}$. This rate is tight in $\mathbb{R}^d$ if one of the measures has a density with respect to the Lebesgue measure, but it can be refined if the measures are supported on low-dimensional subdomains.

**Entropic Optimal Transport.** Computing exactly the optimal transport cost in the discrete setting requires solving a costly linear program with a supercubic complexity. Moreover, OT suffers from the curse of dimensionality and is therefore likely to be meaningless when used on samples from high-dimensional densities. To alleviate these issues, Cuturi proposes to regularize the OT problem by adding an entropic penalty to the objective. By doing so, one can solve exactly this regularized OT problem using a simple alternate minimization scheme, called the Sinkhorn algorithm, that relies only on matrix/vector products and therefore obtains an improved quadratic complexity in terms of time and memory. More specifically, the optimal coupling solving the entropic OT has the form $P = \mathrm{diag}(u)K\,\mathrm{diag}(v)$ and the scaling vectors $u$ and $v$ are updated at each iteration of the Sinkhorn algorithm using simple rescaling operations. The regularization also helps overcome the curse of dimensionality to have good statistical performances. If enough entropy is added, then entropic OT between empirical measures converges towards the unkown entropic OT cost between the true measures with a parametric rate. However, when the entropic penalty is not large enough, the estimation of entropic OT still suffers from the curse of the dimension.

# Chapter 3. Gromov-Wasserstein: Quadratic Optimal Transport

This chapter presents the Gromov-Wasserstein (GW) problem, which can be seen as the quadratic variant of optimal transport. One of the main motivation of GW is that it allows the comparison of probability measures even if they are supported on incomparable sets. In this chapter, we recall the principal definitions and properties of the GW problem, discuss the computational aspects of the GW problem, including its connections to the Quadratic Assignment Problem (QAP) and its NP-hard nature and present an alternative heuristic based on entropic regularization in order to approximate the solution of the GW problem.

**Introduction to Gromov-Wasserstein.** The Gromov-Wasserstein (GW) problem is an alternative to the Optimal Transport (OT) problem, and is used for situations where probability measures have supports in incomparable spaces. The GW problem involves finding an optimal coupling between two probability measures on two Polish spaces, based on a $\ell_p$ distance between the costs of the two spaces. More formally, Let $c_X : X \times X \to \mathbb{R}$ and $c_Y : Y \times Y \to \mathbb{R}$ be continuous measurable functions, and $\mu \in \mathbb{P}(X)$, $\nu \in \mathbb{P}(Y)$ be probability measures on $X, Y$ two Polish spaces. The Gromov-Wasserstein problem is defined as:

$$\mathrm{GW}_p((\mu, c_X), (\nu, c_Y)) = \inf_{\gamma \in \Pi(\mu,\nu)} \left( \int_{X\times Y} \int_{X\times Y} |c_X(x, x') - c_Y(y, y')|^p \, d\gamma d\gamma \right)^{\frac{1}{p}} .$$

The GW objective is constructed so that if an optimal coupling $\gamma$ maps $x$ to $y$ and $x'$ to $y'$, then the couple $(x, x')$ should be "as similar" in $X$ according to $c_x$ as $(y, y')$ in $Y$ according to $c_Y$. When $c_X$, $c_Y$ are distances, it implies that $x, x'$ are as close in $X$ as $y, y'$ in $Y$. The GW problem always admits a solution given certain regularity assumptions on the costs. In addition, the GW problem defines a distance between (equivalence classes of) metric measure spaces that are triplet including a Polish space, a metric, and a Borel probability measure. GW is invariant with respect to a large class of transformation such as rotations, translations or permutations which is particularly useful when it comes to compare shapes.

**Computational Aspects of Gromov-Wasserstein.** When applied on discrete probability measures, the GW problem can be reformulated as a quadratic non-convex optimization problem over the set of nonnegative matrices satysfing linear constraints. This problem is NP-hard in general and notoriuously hard to approximate. When the discrete measures under considations admits the same support size and are uniforms, then GW can be viewed as a relaxation of the Quardratic Assignment Problem (QAP). Indeed, by restricting the admissible couplings to be those which are supported on a graph of a function, that are in fact the permutations matrices, one recovers exactly the QAP. Due to its computational limitations, Peyré et al. [47] propose to regularize the GW problem by adding an entropic penalty to the objective. By doing so, the authors obtain a simple heuristic of the GW problem which consists in solving iteratively nested entropic OT problems. This computational scheme, while allowing one to compute an efficient approximation of the GW problem remains very costly as it requires in the best case scenario a cubic complexity with respect to the number of points.

# Chapter 4. Linear Time Sinkhorn Divergences using Positive Features

*This chapter is based on [6].*

Because of the statistical and computational hurdles of OT, its applications in ML often rely on some form of regularization to smooth the OT problem, and some more specific uses of an entropic penalty, to recover so called Sinkhorn divergences [112]. These divergences are cheaper to compute than regular OT [79, 80], smooth and programmatically differentiable in their inputs [46, 55], and have a better sample complexity [81] while still defining convex and definite pseudometrics [113]. While Sinkhorn divergences do lower OT costs from supercubic down to an embarassingly parallel quadratic cost, using them to compare measures that have more than a few tens of thousands of points in forward mode remains a challenge. The purpose of this chapter is to introduce an new approximation scheme of the Sinkhorn algorithm which can be computed in linear time with respect to the number of points, opening new perspectives to apply entropic OT at scale.

**Related work.** The definition of Sinkhorn divergences usually starts from that of the ground cost on observations. That cost is often chosen by default to be a $q$-norm between vectors, or a shortest-path distance on a graph when considering geometric domains [114, 115, 116, 44]. Given two measures supported respectively on $n$ and $m$ points, entropic OT instantiates first a $n \times m$ pairwise matrix of costs $C$, to solve a minimization problem of a linear objective penalized by the coupling's entropy. This can be rewritten as a Kullback-Leibler minimization:

$$\min_{\text{couplings } P} \langle C, P \rangle - \varepsilon H(P) = \varepsilon \min_{\text{couplings } P} \mathrm{KL}(P, K)\,, \tag{1}$$

where matrix $K$ is defined as $K := \exp(-C/\varepsilon)$, the elementiwe neg-exponential of a rescaled cost $C$. This problem can then be solved using Sinkhorn's algorithm, which only requires applying repeatedly kernel $K$ to vectors. While faster optimization schemes to compute regularized OT have been been investigated [117, 118, 119], the Sinkhorn algorithm remains, because of its robustness and simplicity of its parallelism, the workhorse of choice to solve entropic OT. Since Sinkhorn's algorithm cost is driven by the cost of applying $K$ to a vector, speeding up that evaluation is the most impactful way to speedup Sinkhorn's algorithm. This is the case when using separable costs on grids (applying $K$ boils down to carrying out a convolution at cost $(n^{1+1/d})$ [120, Remark 4.17]) or when using shortest path metrics on graph in which case applying $K$ can be approximated using a heat-kernel [121]. While it

is tempting to use low-rank matrix approximations of the kernel $K$ using standard techniques, applying them within Sinkhorn iterations requires that the application of the approximated kernel guarantees the positiveness of the output. Indeed, if some values of the kernel $K$ are close to 0 and the approximation of each entry are not sufficiently precise, then the approximate kernel might have an negative entries which is enough to make the Sinkhorn algorithm diverge. In [122], the authors propose to use the Nyström method in order to approximate the kernel $K$ and decrease the complexity of the Sinkhorn algorithm, however, in order to guarantee positivity of the entries of the approximate kernel, their method requires $\varepsilon$ to be sufficiently large and and a tolerance error to be very low.

**Our contributions.** Because regularized OT can be carried out using only the definition of a kernel $K$ with positive entries, we focus instead on kernels $K$ that are guaranteed to have positive entries by design. Indeed, rather than choosing a cost to define a kernel next, we consider instead ground costs of the form $c(x, y) = -\varepsilon \log\langle\varphi(x), \varphi(y)\rangle$ where $\varphi$ is a map from the ground space onto the positive orthant in $\mathbb{R}^r$. This choice ensures that both the Sinkhorn algorithm itself (which can approximate optimal primal and dual variables for the OT problem) and the evaluation of Sinkhorn divergences can be computed exactly with an effort scaling linearly in $r$ and in the number of points, opening new perspectives to apply OT at scale. Starting from the kernel instead of the cost to approximate the entropic OT, our contributions are three fold:

- We introduce a general family of kernels admitting a positive and random feature expansion and prove under some regularity assumptions on the positive feature map that our method is able to reach a $\delta$-approximation of the entropic OT cost in $\mathcal{O}(rn)$ time and memory, where $n$ is the number of samples and $r$ is the number of positive random features considered to approximate the true kernel, as soon as $r$ scales in $\log(n)/\delta^2$.

- We show that kernels built from our positive feature expansions can be used to approximate some usual cost functions including the square Euclidean distance. We provide for each of these usual costs an explicit formulation of the positive feature map associated.

- We illustrate the versatility of our approach by extending previously proposed OT-GAN approaches [50, 81], that focused on learning adversarially cost functions $c_\theta$ and incurred therefore a quadratic cost, to a new approach that learns instead adversarially a kernel $k_\theta$ induced from a positive feature map $\varphi_\theta$. We leverage here the fact that our approach is fully differentiable in the feature map to train a GAN at scale, with linear time iterations.

# Chapter 5. Low-Rank Optimal Transport: an Algorithmic Approach

*This chapter is based on [3].*

It is now increasingly clear that regularizing OT in some way or another is the only way to mitigate the computational as well as the statistical issues [51, 74, 75] of OT. A popular approach consists in penalizing the OT problem with a strongly convex function of the coupling [76, 77]. The most popular regularization scheme for OT remains the entropic approach due to its simplicity and high parallelization capability on GPUs. However its quadratic complexity both in term of time and memory remains a major issue when one wants to apply OT on problems with more than a few tens of thousands of points. A key observation when entropy is added to the coupling is that the more entropy is added, the lower the rank (actually the nonnegative rank). Based on this observation, we explore in this work an alternative, and more direct approach to add regularity in the OT problem: we restrict, instead of adding entropy, the set of feasible couplings to have a small nonnegative rank.

**Related work.** Low-rank factorizations are not new to regularized OT. They have been used to speed-up the resolution of entropy regularized OT with the Sinkhorn algorithm, pending some approximations: Given a data-dependent $n \times m$ cost matrix $C$, the Sinkhorn iterations consist in matrix-vector products of the form $Kv$ or $K^T u$ where $K := \exp(-C/\varepsilon)$ and $u, v$ are $n, m$- vectors. Altschuler et al. [122] and Altschuler and Boix-Adsera [123] have proposed to approximate the kernel $K$ with a product of thin rank $r$ matrices, $\widetilde{K} = AB^T$. Naturally, the ability to approximate $K$ with a low-rank $\widetilde{K}$ degrades as $\varepsilon$ decreases, making this approach valid only for sufficiently large $\varepsilon$. Thanks to this approximation, however, each Sinkhorn iteration is linear in $n$ or $m$ ($\mathcal{O}(n+m)r)$) as long as $r \ll n, m$, and the coupling outputted by the Sinkorn algorithm is of the form $\widetilde{P} = CD^T$ where $C = \mathrm{diag}(u)A$, $D = \mathrm{diag}(v)B$. This approximation results therefore in a *low-rank* solution that is not, however, rigorously optimal for the original problem as defined by $K$ but rather that defined by $\widetilde{K}$. The solution obtained with $\widetilde{K}$ can be arbitrary close to the true solution by increasing the rank $r$ considered at the cost of a higher complexity. Similarly, in Scetbon and Cuturi [6] we consider instead *nonnegative low-rank* approximations for $K$ of the form $\widetilde{K} = QR^T$ where $Q, R > 0$ coordinate-wise. The positivity is key here as it ensures the convergence of the approximate Sinkhorn scheme and so for any choice of $\varepsilon$. By doing so, we ends up with a coupling approximating the optimal solution of the entropic OT and of the form $P = EF^T$ where $E \in \mathbb{R}_+^{n \times r}$ and $F \in \mathbb{R}_+^{m \times r}$. Therefore the coupling

outputted by this scheme admits a nonnegative-rank of at most $r$. However, among all the couplings admitting a nonnegative rank smaller than $r$, the solution obtained by this method is in general not the one that reaches the smallest OT cost, and therefore is not optimal in this respect. To our knowledge, only Forrow et al. [124] have used low rank considerations for couplings, rather than costs or kernels. Their work studies the case where the ground cost is the squared Euclidean distance. They obtain for that cost a proxy for rank-constrained OT problems using 2-Wasserstein barycenters [125]. Their algorithm blends those in [126, 127] and results in an intuitive mass transfer plan that goes through a small number of $r$ points, where $r$ is the coupling's nonnegative rank.

**Our Contributions.** In this work, we propose a new alternative to entropic OT to regularize the OT problem by directly constraining the non-negative rank of admissible couplings. Our approach borrows ideas from [124] but is generic as it applies to all ground costs. Here, we constrain the nonnegative rank of the coupling solution $P$ in the OT problem, rather than relying on a low rank approximation $\widetilde{K}$ for kernel $K = e^{-C/\varepsilon}$. This is a crucial point, because the ability to approximate $K$ with a low rank $\widetilde{K}$ depends implicitly on the choice of $\varepsilon$ which can decrease as $\varepsilon$ goes to 0. By contrast, our approach applies to all ranks, small and large. To tackle this problem, we propose to repamatrize the optimization problem and show that couplings admitting a nonnegative rank smaller than $r$ can be expressed as couplings of the form $P = Q\,\mathrm{diag}(1/g)R^T$ decomposed as the product of two thin sub-couplings $Q \in \mathbb{R}_+^{n\times r}$ and $R \in \mathbb{R}_+^{m\times r}$ with common right marginal $g$, and left-marginal given by those of $P$ on each side. Each of these sub-couplings minimizes a transport cost that involves the original cost matrix $C$ and the other sub-coupling. We handle this problem by optimizing jointly on $Q$, $R$ and $g$ using a mirror-descent approach. We prove the non-asymptotic stationary convergence of this approach. Interestingly, we also show that a low-rank assumption on the cost matrix (not on the kernel) can also be leveraged, providing therefore a "best of both worlds" scenario in which both the *coupling*'s and the *cost*'s (not the kernel) low rank properties can be enforced and exploited. Indeed we show that the time complexity of our algorithm can become linear when exploiting low rank assumptions on the *cost* involved in the OT problem. Finally, a useful parallel can be drawn between our approach and that of the vanilla Sinkhorn algorithm, in the sense that they propose different regularization schemes. Indeed, the (discrete) path of solutions obtained by our algorithm when varying $r$ between 1 and $\min(n, m)$ can be seen as an alternative to the entropic regularization path. Both paths contain at their extremes the original OT solution (maximal rank and minimal entropy) and the product of marginals (minimal rank and maximal entropy).

# Chapter 6. Low-rank Optimal Transport: Theoretical Properties

*This chapter is based on [1].*

While it is always intuitively possible to reduce the size of measures (e.g. using $k$-means) prior to solving an OT between them, a promising line of work proposes to combine both [128, 3, 2]. Conceptually, the low-rank approach solve simultaneously both an optimal clustering/aggregation strategy with the computation of an effective transport. This intuition rests on an explicit factorization of couplings into two sub-couplings. This has several computational benefits, since its computational cost becomes linear in $n$ if the ground cost matrix seeded to the OT problem has itself a low-rank. While these computational improvements, mostly demonstrated empirically, hold several promises, the theoretical properties of these methods are not yet well established. This stands in stark contrast to the Sinkhorn approach, which is comparatively much better understood. In this chapter, we target main theoretical properties and practical aspects of the low-rank approach introduced in [3] in order to cement the impact of low-rank approaches in computational OT.

**Related work.** In an applied setting, we often assume that we only have access to samples drawn from the distributions of interest. An important statistical problem in optimal transport is to approximate the (usually unknown) optimal transport cost between $\mu \in \mathcal{P}(\mathbb{R}^d)$ and $\nu \in \mathcal{P}(\mathbb{R}^d)$ using only samples $(x_i)_{i=1}^n$ from $\mu$ and $(y_j)_{j=1}^m$ from $\nu$. These samples are assumed to be independently identically distributed from their respective distributions. For optimal transport costs, a straightforward estimator of the unknown distance between the true distributions is to compute it directly between the empirical measures $\hat{\mu} := \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ and $\hat{\nu} := \frac{1}{n}\sum_{j=1}^m \delta_{x_j}$, hoping ideally that one can control the rate of convergence of the latter to the former. Note that here both $\hat{\nu}$ and $\hat{\mu}$ are random measures, so $OT(\hat{\mu}, \hat{\nu})$ is a random number. An important question is the speed of convergence of $OT(\hat{\mu}, \hat{\nu})$ toward $OT(\mu, \nu)$ , and this rate is often called the "sample complexity". It is well known that standard OT suffers from the curse of dimensionality [129]: Its sample complexity scales in $\mathcal{O}(n^{-1/d})$ and therefore is exponential in the dimension of the ambient space. Although it was recently proved that this result can be refined to consider the implicit dimension of data [72], the sample complexity of OT appears now to be the major bottleneck for the use of OT in high-dimensional machine learning problems. When entropy is added to the objective of the optimal transport problem, it allows also to improve the statistical rates of OT. It has been shown in [51, 82] that entropic OT enjoys a parametric rates $\mathcal{O}\left(\frac{\varepsilon^{-d/2}}{\sqrt{n}}\right)$ with respect to the number of samples. Therefore, when $\varepsilon$ is sufficiently large,

then the plug-in estimator $OT_\varepsilon(\hat{\mu}, \hat{\nu})$ enjoys a fast rate of convergence towards the true quantity $OT_\varepsilon(\mu, \nu)$, however when $\varepsilon$ goes to 0, entropic OT still suffers from the curse of dimensionality with respect to its hyperparameter $\varepsilon$. Although all theoretical contributions converge towards the fact that in practice entropic transport is much more suitable than true OT when it comes to comparing discrete probability measures, a remaining question concerns which quantity to use to measure the difference between two distributions based on the entropic plans. Indeed, entropic OT is symmetric but it is no longer a distance as it does not satisfy the triangle inequality, nor a divergence as it is not positive, nor even able to separate distributions as in general the entropic OT cost between a measure and itself in not 0. To alleviate these issues, Genevay et al. [78] proposed to subtract debiasing terms from entropic OT, defining the Sinkhorn divergence. Feydy et al. [130] then proved that the Sinkhorn divergence defines a suitable divergence able to interpolate between the Maximum Mean Discrepency (MMD) and OT when varying $\varepsilon$.

**Our Contributions.** The goal of this paper is to advance our knowledge, understanding and practical ability to leverage low-rank factorizations in OT. This paper provides five contributions, targeting theoretical and practical properties of LOT:

- We generalize the definition of Low-rank OT (LOT), introduced in [3] in the discrete case, for general probability measures and study the bias induced by the low-rank constraints. We derive the rate of convergence of the low-rank OT to the true OT for both discrete and general probability measures with respect to the non-nnegative rank parameter.

- We make a first step towards a better understanding of the statistical complexity of LOT by providing an upper-bound of the statistical error, made when estimating LOT using the plug-in estimator. Given samples drawn independently from general probability measures supported on compact subsets of $\mathbb{R}^d$, we show that the empirical version of the LOT cost can be upper-bounded by the LOT cost between the true measures and an additional error term that enjoys a parametric rate $\mathcal{O}(\sqrt{r/n})$ that is independent of the dimension $d$.

- We exhibit links between the bias induced by the low-rank factorization and clustering methods. Because the nonnegative rank constaint induces a bias on the OT problem, the LOT cost between a measure and itself is not necessarily 0. This value and the low nonnegtaive-rank coupling solving this LOT problem reflect geometrical information about the measure: LOT cost tells us how much a measure can be clustered in $r$ clusters according to the ground cost $c$ while the optimal coupling provides the clustering of the points

according to this geometry. Therefore LOT offers a new clustering method and so for any geometry $c$. As a special case, when $c = \|\cdot - \cdot\|_2^2$ we recover the classical k-means clustering method.

- We introduce a debiased version of LOT: as the Sinkhorn divergence [130], we show that debiased LOT is nonnegative, equal to 0 if and only if the two measures are the same, that it metrizes the weak convergence, and that it interpolates between the maximum mean discrepancy [18] and OT when varying the nonnegative rank $r$. While the debiased LOT has all the desirable geometric properties in order to be used as a loss function to compare distributions, it still retains the favorable computational complexity of LOT and therefore is a suitable choice for large-scale application in ML when one aims at learning a distribution as in generative modeling.

- We propose practical strategies to tune the step-length and the initialization of the algorithm presented in [3] allowing to have a generic and automatized method for the choice of these hyperparameters, leaving only one hyperparameter to be chosen by the user, namely the choice of the nonnegative rank $r$, like the choice of $\varepsilon$ in the entropic OT.

# Chapter 7. Low-rank Gromov Wasserstein Distances

*This chapter is based on [2].*

The ability to align points across two related yet incomparable point clouds (e.g. living in different spaces) plays an important role in machine learning. The Gromov-Wasserstein (GW) framework provides an increasingly popular answer to such problems, by seeking a low-distortion, geometry-preserving assignment between these points. As a non-convex, quadratic generalization of optimal transport (OT), GW is NP-hard. Much like OT is a relaxation of the optimal assignment problem, GW is a relaxation of the quadratic assignment problem (QAP). Both GW and QAP are NP-hard [131]. While practitioners often resort to solving a regularized version of GW as a nested sequence of entropy-regularized OT problems, the cubic complexity (in the number $n$ of samples) of that approach is a roadblock. We show in this chapter how our recent variant of the OT problem that restricts the set of admissible couplings to those having a low-rank factorization [3] is remarkably well suited to the resolution of GW.

**Related work.** The GW problem replaces the linear objective in OT by a *non-convex*, *quadratic*, objective $\mathcal{Q}_{A,B}(P) := \text{cst} - 2\langle APB, P\rangle$ parameterized by *two* square cost matrices $A$ and $B$. In practice, linearizing iteratively $\mathcal{Q}_{A,B}$ works

well [84, 85]: recompute a synthetic cost $C_t := AP_{t-1}B$, use Sinkhorn to get $P_t := \text{argmin}_P \langle C_t, P \rangle + \varepsilon \text{reg}(P)$, repeat. This leads to a computational scheme that scales in cubic time and requires a quadratic memory space with respect to the number of samples. Several obstacles stand in the way of speeding up this entropic GW scheme. The re-computation of the cost matrix involved at each outer iteration is an issue, since it requires $O(n^3)$ operations [47, Prop. 1]. We only know of two broad approaches that achieve tractable running times: *(i)* Solve related, yet significantly different, proxies of the GW energy, either by embedding points *as* univariate measures [86, 87], by using a sliced mechanism when restricted to Euclidean settings [88] or by considering tree metrics for supports of each probability measure [89], *(ii)* Reduce the size of the GW problem through quantization of input measures [90] or recursive clustering approaches [91, 62]). Interestingly, no work has, to our knowledge, tried yet to accelerate Sinkhorn iterations withing GW.

**Our contributions.** Our method addresses the problem of approximating the GW using the new regularization scheme proposed in [3] based on low-rank constraints. Our method overcomes limitations that arises from updating the cost matrix $C_t$ at a cubic cost and solving the nested entropic OT problems requiring a quadratic complexity.

- We show that a low-rank factorization (or approximation) of the two input cost matrices that define GW, one for each measure, can be exploited to lower the complexity of recomputing $C_t$ from cubic to quadratic. By doing we do, we are also able to reduce the total complexity of the entropic GW scheme from cubic to quadratic.

- We show next, independently, that using the low-rank approach for *couplings* advocated by [3] to solve OT can be inserted in the GW pipeline and result in a $O(n^2)$ strategy for GW, with no prior assumption on input cost matrices. We also briefly explain why methods that exploit the geometrical properties of $C$ (or its kernel $K = e^{-C}$) to obtain faster iterations are of little use in a GW setup, because of the necessity to re-instantiate a new cost $C_t$ at each outer iteration.

- Finally, we show that both low-rank assumptions (on costs and couplings) can be combined to shave yet another factor and reach GW approximation with linear complexity in time and memory. We provide experiments, on simulated and real datasets, which show that our approach has comparable performance to entropic-regularized GW and its practical ability to reach "good" local minima to GW, for a considerably cheaper computational price,

and with a conceptually different regularization path, yet can scale to millions of points.

# Chapter 8. Equitable and Optimal Transport with Multiple Agent

*This chapter is based on [5].*

Fair division [94] has been widely studied by the artificial intelligence [95] and economics [96] communities. Fair division consists in partitioning diverse resources among agents according to some fairness criteria. One of the standard problems in fair division is the fair cake-cutting problem [97, 98]. The cake is an heterogeneous resource, such as a cake with different toppings, and the agents have heterogeneous preferences over different parts of the cake, i.e., some people prefer the chocolate toppings, some prefer the cherries, others just want a piece as large as possible. Hence, taking into account these preferences, one might share the cake equitably between the agents. A generalization of this problem, for which achieving fairness constraints is more challenging, is when the splitting involves several heterogeneous cakes, and where the agents have linked preferences over the different parts of the cakes. This problem has many variants such as the cake-cutting with two cakes [99], or the Multi Type Resource Allocation [100, 101]. In all these models it is assumed that there is only one indivisible unit per type of resource available in each cake, and once an agent choose it, he or she has to take it all. In this setting, the cake can be seen as a set where each element of the set represents a type of resource, for instance each element of the cake represents a topping. A natural relaxation of these problems is when a divisible quantity of each type of resources is available. Based on the fundamental idea of Kantorovich to relax the OT problem, we introduce in this chapter EOT (Equitable and Optimal Transport), a formulation that solves both the cake-cutting and the cake-cutting with two cakes problems when the resources are divisible.

**Related work.** Fair division of goods has a long standing history in economics and computational choice. A classical problem is the fair cake-cutting that consists in splitting the cake between $N$ individuals according to their heterogeneous preferences. The cake $\mathcal{X}$, viewed as a set, is divided in $\mathcal{X}_1, \dots, \mathcal{X}_N$ disjoint sets among the $N$ individuals. The utility for a single individual $i$ for a slice $S$ is denoted $V_i(S)$. It is often assumed that $V_i(\mathcal{X}) = 1$ and that $V_i$ is additive for disjoint sets. There exists many criteria to assess fairness for a partition $\mathcal{X}_1, \dots, \mathcal{X}_N$ such as proportionality ($V_i(\mathcal{X}_i) \geq 1/N$), envy-freeness ($V_i(\mathcal{X}_i) \geq V_i(\mathcal{X}_j)$) or equitability

$(V_i(\mathcal{X}_i) = V_j(\mathcal{X}_j))$. The cake-cutting problem has applications in many fields such as dividing land estates, advertisement space or broadcast time. An extension of the cake-cutting problem is the cake-cutting with two cakes problem [99] where two heterogeneous cakes are involved. In this problem, preferences of the agents can be coupled over the two cakes. The slice of one cake that an agent prefers might be influenced by the slice of the other cake that he or she might also obtain. The goal is to find a partition of the cakes that satisfies fairness conditions for the agents sharing the cakes. Cloutier et al. [99] studied the envy-freeness partitioning. Both the cake-cutting and the cake-cutting with two cakes problems assume that there is only one indivisible unit of supply per element $x \in \mathcal{X}$ of the cake(s). Therefore sharing the cake(s) consists in obtaining a paritition of the set(s). However in this setting, the problem might not be well posed, and even if it is, solving the problem can be hard in practice. In this chapter, we also establish the links of EOT with some integral probability metrics. IPMs are (semi-)metrics on the space of probability measures. For a set of functions $\mathcal{F}$ and two probability distributions $\mu$ and $\nu$, they are defined as

$$\mathrm{IPM}_{\mathcal{F}}(\mu, \nu) = \sup_{f \in \mathcal{F}} \int f d\mu - \int f d\nu.$$

For instance, when $\mathcal{F}$ is chosen to be the set of bounded functions with uniform norm less or equal than 1, we recover the Total Variation distance [132] (TV). They recently regained interest in the ML community thanks to their application to Generative Adversarial Networks (GANs) [22] where IPMs are natural metrics for the discriminator [133, 134, 135, 136]. They also helped to build consistent two-sample tests [18, 137]. However when a closed form of the IPM is not available, exact computation of IPMs between discrete distributions may not be possible or can be costful. For instance, the Dudley metric can be written as a Linear Program [138] which has at least the same complexity as standard OT.

**Our Contributions.** In this paper we introduce EOT an extension of Optimal Transport which aims at finding an equitable and optimal transportation strategy between multiple agents. We make the following contributions.

- We introduce the EOT which aims at finding an equitable and optimal coupled partition of the resources according the heterogeneous preferences of the agents. Each agent in the problem is represented as an utility (or cost) function, and the resources allocated to each agent is a sub-coupling such that their sum is a valid coupling of the resources satisfying the marginal constraints. Formally, EOT is finding the partition that maximize the smallest utilities among the agents. From a transport point of view, EOT aims at splitting the transportation task from a probabilities measures toward another

between multiples workers (or agents) represented as cost functions in order to obtain a partition of the task that is equitable and optimal. Here EOT is trying to minimize the most expensive total transportation cost among the workers.

- We show that EOT solves a fair division problem where heterogeneous resources have to be shared among multiple agents. More precisely, we show that the EOT always admit a solution and that at optimality, the total utility or cost of the agents are equal and optimal. As a by-product, we also show that the partition obtained is not only equitable and optimal but also proportional which is another important fairness criteria.

- EOT is a linear optimization problem under linear constraints. We derive its dual and prove that strong duality holds. As a by-product, we show that EOT is related to some usual IPMs families and in particular the widely known Dudley metric.To the best of our knowledge, this is the first time a link is given between the Dudley metric and Optimal Transport. As a consequence, we also derive sufficient conditions on the cost functions for EOT to metrize the weak convergence.

- We also tackle the computational aspects of EOT and propose an entropic regularized version of the problem, derive its dual formulation, obtain strong duality. We then provide an efficient algorithm to approximate EOT.

# Chapter 9. Mixed Nash Equilibria in the Adversarial Examples Game

*This chapter is based on [4].*

Adversarial examples [102, 103] are one of the most dizzling problems in machine learning: state of the art classifiers are sensitive to imperceptible perturbations of their inputs that make them fail. This asks the following question: can we build classifiers that are robust against any adversarial attack? Assuming that one can reformulate the adversarial risk minimization problem as a min-max problem, then the above question is equivalent to ask for the existence of a Nash Equilibrium. Showing the existence of such equilibrium would ensure that it is possible to learn a classifier that is robust against any small pertubation of the data, i.e. against any attack that happens after having learned the classifier. In this chapter, we tackles the problem of adversarial examples from a game theoretic point of view using tools from optimal transport and study the open question of the existence of mixed Nash equilibria in the zero-sum game formed by the attacker and the classifier.

**Related work.** A recent line of research argued that randomized classifiers could help countering adversarial attacks [139, 140, 141, 142]. Along this line, [143] demonstrated, using game theory, that randomized classifiers are indeed more robust than deterministic ones against regularized adversaries. However, the findings of these previous works depends on the definition of considered adversary. In particular, they did not investigate scenarios where the adversary also uses randomized strategies, which is essential to account for if we want to give a principled answer to the question on the existence of a classifier that is robust against any adversarial attack. Previous works studying adversarial examples from the scope of game theory investigated the randomized framework (for both the classifier and the adversary) in restricted settings where the adversary is either parametric or has a finite number of strategies [144, 145, 146]. Adversarial examples have been studied under the notions of Stackelberg game in [147], and zero-sum game in [144, 145, 146]. These works considered restricted settings (convex loss, parametric adversaries, etc.) that do not comply with the nature of the problem. Indeed, it has been proven that no convex loss can be a good surrogate for the 0/1 loss in the adversarial setting [148, 149], narrowing the scope of these results. If one can show that for sufficiently separated conditional distributions, an optimal deterministic classifier always exists, necessary and sufficient conditions for the need of randomization are still to be established. Pinot et al. [143] studied partly this question for regularized deterministic adversaries, leaving the general setting of randomized adversaries and mixed equilibria unanswered, which is the very scope of this paper. Bhagoji et al. [150] and Pydi and Jog [151] investigated classifier-agnostic lower bounds on the adversarial risk of any deterministic classifier using OT. These works only evaluate lower bounds on the primal deterministic formulation of the problem, while we study the existence of mixed Nash equilibria. Note that Pydi and Jog [151] started to investigate a way to formalize the adversary using Markov kernels, but did not investigate the impact of randomized strategies on the game. Another line of works [152, 153, 154] studied the problem of adversarial examples through the scope of distributionally robust optimization. In these frameworks, the set of adversarial distributions is defined using an $\ell_p$ Wasserstein ball (the adversary is allowed to have an *average* perturbation of at most $\varepsilon$ in $\ell_p$ norm). This however does not match the usual adversarial attack problem, where the adversary cannot move any point by more than $\varepsilon$.

**Ours Contributions.** In this work, we study the adversarial example game from a game theoretic point of view thanks to optimal transport. More precisely we make the following contributions.

- In the standard formulation of the adversarial risk, the adversary is defined as a function that maps each point of the dataset to an adversary point

restricted to live in a ball centered in the datapoint that maximizes the loss. Using the fundamental idea of Kantorovich, we extend the work of [151] and relax this formulation in order to allow the adversary to be instead a coupling between the data distribution and the adversarial one. This relaxation is in fact tight as we show that both formulation are equal while the latter can be formulated as an optimization problem. Indeed we show that the adversarial risk can be reformulated as a linear maximization problem over distributions restricted to live in a specific Wasserstein ball centered on the data distribution.

- Using our variational formulation of the adversarial risk, we show that the adversarial risk minimization problem can casted as a Distributionally Robust Optimization (DRO) [155]. This formulation naturally leads us to analyze adversarial risk minimization as two-player a zero-sum game.

- We show that in general a Nash equilibrium in such game does not exists and we provide a simple example showing the necessity for using randomized strategies both with the attacker and the classifier. Then we show that in the adversarial example game when both the adversary and the classifier can use randomized strategies, it is always possible to reach a Mixed Nash equilibrium.

- Finally we design efficient algorithms to learn a finite mixture of classifiers. Taking inspiration from robust optimization [152] and subgradient methods [156], we derive a first oracle algorithm to optimize a finite mixture. Then, following the line of work of [76], we introduce an entropic regularization to effectively compute an approximation of the optimal mixture. We validate our findings with experiments on simulated and real datasets, namely CIFAR-10 an CIFAR-100 [157].

# Contributions de cette Thèse

La comparaison et la manipulation de mesures de probabilités sont des tâches omniprésentes en apprentissage automatique (AA). Par exemple, les méthodes d'apprentissage supervisé [12] s'appuient fortement sur ces comparaisons pour mesurer le risque. En classification [13], les praticiens de l'AA apprennent un classifieur en minimisant la fonction perte d'entropie croisée [14] entre les vraies distributions conditionnelles des étiquettes et les distributions conditionnelles apprises sur les classes. Dans le cadre de la régression [15], le risque de la prédiction est également mesuré en comparant la vraie distribution conditionnelle des données avec celle apprise sous des hypothèses gaussiennes. Plus généralement, l'estimation du maximum de vraisemblance (EMV) [16], qui est une technique standard pour estimer les paramètres d'une distribution de probabilité qui décrit le mieux les données observées, vise à minimiser la divergence de Kullblack-Leibler (KL) entre la vraie distribution et la distribution modélisée. Toutes ces techniques fondamentales, largement utilisées en AA, reposent sur la comparaison des distributions afin de quantifier l'incertitude des prédictions. Il est également essentiel de pouvoir comparer les distributions dans d'autres domaines de la science des données, comme par exemple pour la conception de tests statistiques [17, 18] ou la découverte de causes [19, 20]. Ces dernières années, ce besoin a reçu encore plus d'attention de la part de la communauté de l'AA grâce au développement de nouveaux modèles génératifs capables de résoudre des tâches de plus en plus complexes. Les autoencodeurs variationnels [21] sont une classe de modèles génératifs qui apprennent à approximer une distribution cible en encodant les données dans un espace de dimension inférieure et impliquent la minimisation d'une mesure de divergence entre la distribution latente et une distribution a priori. Les réseaux adverses génératifs (GAN) [22] sont une autre classe populaire de modèles génératifs qui apprennent à générer des échantillons en comparant de manière adverse la distribution de probabilité générée avec la distribution cible. Les flux de normalisation [23] peuvent être utilisés pour la modélisation générative ainsi que pour l'estimation de la densité et sont appris à l'aide de l'EMV. Les modèles de diffusion [24] transforment une distribution de base simple en une distribution cible plus complexe à l'aide d'une séquence d'étapes de diffusion et mesurent les erreurs séquentielles à l'aide des

divergences KL entre les distributions générées et les distributions réelles. Enfin, l'exemple le plus illustratif de cette littérature est certainement l'émergence des transformers [25] qui sont aujourd'hui l'une des architectures les plus efficaces pour résoudre un large éventail de problèmes en vision par ordinateur [26] et en traitement du langage naturel [27]. Ces nouvelles architectures visent à traiter les données comme des distributions de probabilité. En particulier, les transformers traitent des signaux tels que des images et des phrases en les représentant respectivement comme des distributions discrètes de "patchs" et de mots après avoir encodé la structure spatiale de ces objets à l'aide de représentations positionnelles [28]. Par conséquent, être capable de comparer et de traiter les distributions devient un défi de plus en plus important en AA et c'est dans ce contexte que le transport optimal (TO) est devenu un outil largement utilisé.

**TO dans la science des données.** Lorsqu'il s'agit de comparer des distributions, la littérature statistique fournit une riche classe de fonctions de divergence pour mesurer l'écart entre deux distributions de probabilité, telles que la divergence de KL, la distance de variation totale (VT), ou plus généralement la famille des $\phi$-divergences [29]. Cependant, ces divergences reposent sur la comparaison ponctuelle des fonctions de densité et saturent ou divergent lorsque les supports des mesures de probabilité sont disjoints, ce qui limite leur application à la comparaison d'histogrammes ou de distributions de probabilité continues de même support. Le TO [30] est devenu une alternative de plus en plus importante en AA grâce à la polyvalence de son utilisation pour comparer les mesures de probabilité. À partir d'une fonction de coût (par exemple, une distance) sur l'espace sur lequel les mesures sont supportées, le TO consiste à trouver une correspondance [31] ou un couplage [32] entre les deux mesures qui est optimal par rapport à ce coût. En d'autres termes, le TO étend naturellement le coût au sol entre deux points à une fonction de divergence entre des histogrammes de points, ou des mesures de probabilité, sous la forme d'un problème d'optimisation. En conséquence, le TO fournit un cadre simple et complet pour comparer les distributions de probabilité et a inspiré de nombreux développements dans l'apprentissage automatique [33]. Une multitude de travaux l'ont récemment relié à d'autres sujets d'actualité, tels que la normalisation des flux ou les réseaux neuronaux convexes [34, 35, 36], tandis que la portée de ses applications a maintenant atteint plusieurs domaines de la science tels que la vision par ordinateur [37], le traitement du signal [38, 39, 40], la biologie unicellulaire [41], l'imagerie [42, 43], les neurosciences [44, 45], ou modélisation générative [49, 50, 51]. Une autre caractéristique majeure du TO est le couplage optimal obtenu lors de sa résolution, qui fournit un alignement optimal des mesures

de probabilité à disposition. Un tel objet, spécifique au TO et au moins aussi important que le coût du TO lui-même, a aussi de nombreuses applications en AA pour aligner des representations de mots [52, 53, 54], pour reconstruire les trajectoires cellulaires [55, 56], pour l'adaptation de domaines [57, 58] ou même pour encoder des distributions discrètes comme dans les transformers [25] à l'aide de projections barycentriques [59]. En outre, la polyvalence du cadre du TO va au-delà de la comparaison des mesures de probabilité supportées sur le même espace. [60] propose une version quadratique du TO, à savoir la distance de Gromov-Wasserstein (GW) qui vise à comparer des nuages de points ou des mesures de probabilité supportées sur des espaces incomparables. Alors que le TO recherche une correspondance optimale en fonction d'un coût au sol en minimisant un score linéaire associé à ce coût, GW recherche une correspondance qui soit aussi proche que possible d'une isométrie, quantifiée par un score quadratique. Plusieurs problèmes en AA nécessitent de comparer des ensembles de données qui vivent dans des espaces hétérogènes. Cette situation se produit généralement lors du réalignement de deux vues distinctes de points échantillonnés à partir de sources similaires. Les applications récentes à la génomique des cellules uniques [61, 62] en sont un bon exemple: Des milliers de cellules prélevées dans le même tissu sont réparties en deux groupes, chacune étant traitée selon un protocole expérimental différent, ce qui donne deux ensembles distincts de vecteurs de caractéristiques hétérogènes. Malgré cette hétérogénéité, on s'attend à trouver une correspondance entre les points du premier et du second ensemble, car ils contiennent des informations globales similaires. GW a également été utilisé dans l'apprentissage supervisé [63], la modélisation générative [64], l'adaptation au domaine [65], la prédiction structurée [66], la chimie quantique [47] et l'alignement de couches [67].

**Défis du TO en AA.** La résolution des problèmes de transport optimal à l'échelle pose cependant des défis redoutables. Le plus évident d'entre eux est d'ordre computationel: l'instanciation du problème Kantorovich [32] sur des mesures discrètes de taille $n$ peut être résolue à l'aide d'un programme linéaire (PL) de complexité $O(n^3 \log n)$ [68, 69, 70]. Un deuxième défi, tout aussi important, réside dans la performance statistique de l'utilisation de ce PL pour estimer le TO entre les densités: la solution du PL entre les échantillons i.i.d. converge exponentiellement lentement par rapport à la dimension ambiante vers celle entre les densités [71]. [72, 73] ont obtenu des résultats fins et montrent que l'estimation du TO nécessite un nombre exponentiel d'échantillons par rapport à la dimensionnalité intrinsèque du support. Il est maintenant de plus en plus clair que la régularisation du TO d'une manière ou d'une autre est le seul moyen d'atténuer ces deux problèmes [51, 74, 75].

Une approche populaire consiste à pénaliser le problème du TO avec une fonction fortement convexe du couplage [76, 77], et certaines utilisations plus spécifiques d'une pénalité entropique, pour récupérer les divergences dites de Sinkhorn [78]. Le TO entropique est moins coûteux à calculer que le TO ordinaire [79, 80], lisse et différentiable dans ses entrées [46, 55], et a une meilleure complexité statistique [81, 82]. Bien que les solveurs entropiques du TO réduisent le coût de calcul de supercubique à un coût quadratique parallèlisable, son utilisation pour comparer des mesures qui sont supportées sur plus de quelques dizaines de milliers de points reste un défi. Ces limitations de calcul sont encore plus critiques dans le cas de GW, une généralisation quadratique non convexe du problème TO, qui est NP-hard à résoudre en général [83]. Comme le TO, GW peut être régularisé en utilisant l'entropie [84, 85] et [47] proposent d'appliquer un schéma de descente en miroir (DM) pour approcher le coût entropique de GW consistant à résoudre une séquence de problèmes de TO entropiques imbriqués. Bien que cette heuristique permette d'obtenir des coûts GW faibles, elle reste considérablement limitée en pratique en raison de sa complexité cubique par rapport au nombre de points. Seules deux approches générales permettent d'obtenir des temps d'exécution raisonnables : *(i)* Résoudre des approximations reliées, mais significativement différentes, de l'énergie de GW, soit en intégrant les points *comme* des mesures univariées [86, 87], soit en utilisant un mécanisme en tranches limité au cas euclidien [88], soit en considérant des métriques arborescentes pour les supports de chaque mesure de probabilité [89], *(ii)* Réduire la taille du problème de GW par la quantification des mesures d'entrée [90] ou par des approches de regroupement récursif [91, 62]. Toutefois, aucun de ces travaux n'a tenté d'accélérer le schéma DM proposé dans [47] pour approximer GW.

**TO et applications AA.** Bien que le coût du TO ou sa solution aient été utilisés directement comme fonction de perte [49] ou pour aligner les distributions [92] dans diverses applications de l'AA, de nombreux liens entre le TO et les problèmes appliqués restent à découvrir. L'origine de la théorie du TO remonte au XVIIIe siècle, lorsque le mathématicien français Gaspard Monge [31] a introduit le problème de la recherche de la manière la plus efficace de transporter une distribution de probabilités vers une autre en utilisant des transformations capable d'envoyer une distribution vers une autre. Bien que la formulation mathématique du transport optimal de Monge ait été révolutionnaire, on a rapidement découvert que son approche avait des limites en raison de l'absence de garentie d'existence d'une solution au problème. Ce n'est que 150 ans plus tard que des progrès significatifs ont été réalisés dans le domaine de la théorie du transport optimal, grâce à l'idée

fondamentale de Kantorovich [32]: Il a proposé une relaxation du problème du TO en considérant des couplages, plutôt que des transformations deterministes, pour faire correspondre les distributions et autorisant ainsi la division des masses dans le processus d'appariement. Il s'avère que la limitation de la formulation de Monge du TO peut se retrouver dans plusieurs problèmes qui, à première vue, ne semblent pas liés au TO. Un premier exemple est le problème de la répartition équitable [94]. Ce problème a été largement étudiée par les communautés de l'intelligence artificielle [95] et de l'économie [96]. Elle consiste à répartir diverses ressources entre des agents en respectant un ou plusieurs critères d'équité. L'un des problèmes classiques de la répartition équitable est le problème de la découpe équitable d'un gâteau [97, 98]. Le gâteau est une ressource hétérogène, et les agents ont des préférences hétérogènes sur les différentes parties du gâteau. Ainsi, en tenant compte de ces préférences, on voudrait partager le gâteau équitablement entre les agents. Ce problème a de nombreuses variantes, telles que le découpage de gâteau avec deux gâteaux ou l'allocation de ressources de type multiple [100, 101]. Cependant, dans tous ces modèles, on suppose qu'il n'y a qu'une seule unité indivisible par type de ressource disponible et que, par conséquent, la partition des ressources revient à définir une transformation déterministe qui attribue à chaque type de ressource un agent unique. La contrainte sur la partition limite considérablement la résolution du problème de la répartition équitable, car il est possible qu'il n'y ait pas de solution, et aucun assouplissement du problème n'a encore été proposé. Un second problème où l'on retrouve des transformations déterministes entre des distributions est celui des attaques adverses [102, 103]. Ce problème cherche à trouver des classifieurs capablent d'avoir des bonnes propriètés de généralisation malgrè des pertubations imperceptibles de la donnée. En effet, les classifieurs actuels sont sensibles à ces perturbations imperceptibles et échouent fatalement en cas d'attaques. Ces dernières années, la recherche s'est concentrée sur la proposition de nouvelles méthodes de défense [104, 105, 106] et sur la construction d'attaques de plus en plus sophistiquées [107, 108, 109, 110]. Jusqu'à présent, la plupart des stratégies de défense se sont révélées vulnérables à ces nouvelles attaques ou sont difficiles à calculer. Cela pose la question suivante: pouvons-nous construire des classifieurs qui sont robustes contre toute attaque adverse ? La réponse à cette question pourrait être principalement limitée en raison de la définition restreinte de l'attaquant: les attaques adverses sont définies comme la solution d'un problème d'optimisation visant à maximiser la fonction de perte sur les boules centrées sur les points de données. Par conséquent, les attaques adverses peuvent être considérées comme des fonctions déterministes qui envoient la distribution des données vers des distributions adverses. Par ailleurs, à notre connaissance, aucun travail n'a encore essayé d'assouplir la définition de l'adversaire pour donner une réponse de principe à la question posée.

**Contributions de cette thèse.** Cette thèse, commencée en 2019 sous la direction de Marco Cuturi, apporte quelques contributions sur de nouvelles approches computationelles pour aborder les problèmes de transport optimal à grande échelle en apprentissage automatique, et étudie de nouvelles méthodologies appliquant le transport optimal pour résoudre des défis de longue date dans l'optimisation robuste et l'équité algorithmique. Plus précisément, les contributions de cette thèse sont divisées en deux parties principales. Dans une première partie, nous présentons nos contributions concernant le développement algorithmique et théorique de nouveaux schémas de régularisation basés sur des méthodes de faible rang pour permettre l'application du transport optimal et de sa variante quadratique dans un cadre à grande échelle. Dans une deuxième partie, nous présentons nos contributions où le transport optimal est utilisé comme outil pour comprendre et étudier le problème de la division équitable et le problème des attaques adverses.

Dans la partie II, nous présentons les quatre contributions suivantes.

- Dans [6], nous proposons d'accélérer la résolution du TO régularisé par l'entropie avec l'algorithme de Sinkhorn en considérant une factorisation spécifique de faible rang de la matrice du noyau $K = \exp(-C/\varepsilon)$ impliquée dans les itérations de Sinkhorn. Notre approximation de faible rang du noyau $K$ est obtenue en utilisant des représentations paramétrées qui associent à tout point du support des mesures un vecteur dans l'orthant positif et forcent donc la positivité de la factorisation. Le couplage obtenu admets donc par construction une factorization de faible rang non négatif. Nous montrons que notre approche peut être utilisée pour approcher le TO entropique avec des fonctions de coût courantes telles que la distance euclidienne au carrée. Nous illustrons également la polyvalence de notre méthode en étendant le TO-GAN précédemment proposé à une nouvelle approche qui apprend de manière adverse un noyau induit à partir de representations positives. Cette approche est entièrement différentiable et peut être utilisée pour apprendre un GAN à l'échelle avec des itérations en temps linéaire.

- Dans [3], nous proposons un nouveau schéma de régularisation du problème de TO, appelé "Low-rank Optimal Transport" (LOT), qui contraint les couplages admissibles à avoir un rang non négatif faible. Au lieu de factoriser le noyau impliqué dans les itérations de Sinkhorn avec des facteurs positifs, nous imposons directement une contrainte de rang non négatif faible sur l'ensemble des couplages admissibles dans le problème de TO, sans approximation sur

les matrices de coût ou de noyau. Nous introduisons une approche générique qui peut résoudre le problème de TO sous des contraintes de faible rang avec des coûts arbitraires. Les couplages de faible rang non négatif ont une factorisation naturelle de faible rang qui s'écrit comme un produit de sous-couplages avec une marge commune et qu'on exploite pour optimiser conjointement les sous-couplages et la distribution marginale commune à l'aide d'une approche de descente en miroir. Nous prouvons la convergence non asymptotique vers un point stationnaire de notre algorithme et montrons que la complexité en temps et en mémoire de l'algorithme est généralement quadratique mais peut devenir linéaire lorsqu'on exploite des hypothèses de faible rang sur le *coût* (pas le noyau qui dépend fortement du paramètre de régularisation $\varepsilon$) impliqué dans le problème du TO.

- Dans [1], nous visons à améliorer nos connaissances et notre capacité pratique à utiliser les factorisations de faible rang dans le transport optimal. L'article se concentre sur les aspects théoriques et pratiques du TO de faible rang (LOT). Nous généralisons la définition de LOT aux mesures de probabilité générales et dérivons la vitesse de convergence de LOT vers le véritable TO pour les mesures de probabilité discrètes et générales. Nous fournissons également une borne supérieure pour l'erreur statistique commise lors de l'estimation de LOT à l'aide de mesures empiriques et montrons qu'elle a un vitesse paramétrique indépendante de la dimension. Nous établissons des liens entre le biais induit par les contraintes de faible rang sur le TO et les méthodes de clustering. Nous introduisons une version débiaisée de LOT qui métrise la convergence faible et convient à la comparaison à grande échelle de mesures dans l'apprentissage automatique. Enfin, nous proposons des stratégies pratiques pour ajuster le pas de gradient dans la descente mirroire et l'initialisation de l'algorithme de LOT, ce qui en fait une méthode générique et automatisée pour le choix des hyperparamètres.

- Dans [2], nous nous concentrons sur les aspects computationel de Gromov-Wasserstein et proposons un nouveau schéma de régularisation du problème basé sur des contraintes de faible rang. Plus précisément, nous exploitons une factorisation de faible rang des deux matrices de coût d'entrée pour réduire la complexité du calcul du coût, à chaque itération du schéma GW entropique, de cubique à quadratique, diminuant ainsi sa complexité totale. Nous montrons que l'approche de faible rang pour les couplages, proposée initialement pour résoudre le problème de TO, peut être utilisée pour GW et nous obtenons une complexité quadratique sans hypothèse sur les matrices de coût d'entrée. Nous expliquons également pourquoi les méthodes qui exploitent les propriétés géométriques des noyaux sont peu utiles dans le

problème de GW. Nous combinons les hypothèses de faible rang sur les coûts et les couplages pour obtenir une approximation de GW avec une complexité linéaire en temps et en mémoire, et nous démontrons l'efficacité de notre méthode sur des ensembles de données simulées et réelles.

Dans la partie III, nous présentons les deux contributions suivantes.

- Dans notre travail [5], nous introduisons EOT (Equitable and Optimal Transport), qui est une version relaxée du problème de la répartition équitable. Dans ce problème, plusieurs agents cherchent à partager un ou plusieurs ensembles de ressources, en trouvant une partition équitable de ces ensembles. Nous proposons ici d'assouplir le problème et de considérer le cas où les ressources ne sont plus des ensembles mais plutôt des distributions sur ces ensembles où une certaine quantité de masse divisible est associée à chacun des éléments de ces ensembles. EOT est défini comme un problème d'optimisation linéaire sous contraintes linéaires qui maximise le minimum des utilités individuelles. Nous montrons que la partition obtenue par EOT est équitable, optimale et proportionnelle, et nous dérivons la formulation duale de EOT avec des résultats de forte dualité. Nous montrons également que EOT est lié à certaines métriques sur l'espace des mesures, comme la métrique de Dudley, et nous proposons une version régularisée du problème avec un algorithme efficace similaire à l'algorithme de Sinkhorn pour calculer EOT.

- Dans notre travail [4], nous obtenons une reformulation du problème de minimisation du risque adverse en utilisant le transport optimal. Nous montrons qu'il peut être reformulé comme un problème d'optimisation robuste distributionnel sur des boules de Wasserstein spécifiques et nous étudions l'existence d'équilibres de Nash. Plus précisément, en assouplissant l'adversaire pour qu'il soit un couplage au lieu d'une transformation déterministe de la distribution des données vers la distribution adverse, nous obtenons une formulation variationnelle du risque adverse pour des classifieurs déterministes et aléatoires, et nous montrons que dans les deux cas, les problèmes de minimisation du risque adverse peuvent être reformulés comme des jeux à somme nulle à deux joueurs. Nous montrons que dans le cas de stratégies mixtes, il est toujours possible d'approcher un équilibre de Nash (et même de l'atteindre sous certaines hypothèses), ce qui signifie que l'ajout d'aléas dans le choix des classifieurs permet d'apprendre une mixture de classsifieurs qui est robuste à toutes perturbations adverses. Nous concevons également un algorithme qui permet d'apprendre efficacement un mélange fini de classifieurs et nous montrons une amélioration empirique de la robustesse contre des attaques adverses usuelles par rapport aux défenses déterministes classiques.

Nous allons maintenant présenter plus en détail les chapitres qui constituent cette thèse.

# Chapitre 1. Transport Optimal: de Monge à Kantorovich

Ce chapitre introduit les concepts et résultats clés sur le transport optimal sur lesquels cette thèse s'appuie. En raison de notre intérêt pour les applications en AA, nous énonçons ces résultats pour des mesures supportées sur $\mathbb{R}^d$. Nous présentons d'abord la formulation originale de Monge du problème du transport optimal et ses principales limitations, puis nous présentons la relaxation de Kantorovich et les liens entre ces deux formulations, et enfin nous exposons quelques propriétés fondamentales du transport optimal de Kantorovich.

**Transport optimal de Monge.** La formulation originale du transport optimal a été proposée par Gaspard Monge en 1781 et est connue sous le nom de problème de Monge. Étant donné deux mesures de masse égale $\mu$ et $\nu$ vivant dans $\mathcal{P}(\mathbb{R}^d)$ et une fonction de coût $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}+$, Monge a proposé le problème du transport optimal de $\mu$ à $\nu$ en fonction de $c$. Plus formellement, ce problème peut être énoncé comme suit

$$\inf_{T:T\#\mu=\nu} \int_{\mathbb{R}^d} c(x, T(x))d\mu(x)$$

où $T\#\mu$ est la mesure pushforward de $\mu$ par $T$, définie par $T\#\mu(A) := \mu(T^{-1}(A))$ pour tous les ensembles mesurables de $\mu$. Lorsqu'elle existe, une transformation de transport satisfaisant la contrainte $T\#\mu = \nu$, assigne à chaque point $x$ du support de la mesure initiale $\mu$ un point $T(x)$ du support de la mesure cible $\nu$, et transporte toute la masse de $\mu$ située à $x$ vers $T(x)$. Le problème de Monge vise à trouver, parmi toutes ces transformations de transport, celle qui minimise le coût total du transport. Une condition suffisante pour l'existence d'une transformation de transport est que $\mu$ soit sans atome, mais même lorsque des tranformations de transport existent, il se peut qu'aucune ne soit optimale. Plus généralement, le problème de Monge n'est pas toujours bien posé et même lorsqu'il l'est, il peut être très difficile à résoudre car l'objectif et les contraintes ne sont pas convexes.

**Transport optimal de Kantorovich.** La formulation de Kantorovich du problème de transport optimal assouplit le problème de Monge en cherchant à minimiser le coût de transport sur un ensemble de transformations probabilistes qui spécifient la quantité de masse déplacée de chaque point de la distribution source à chaque

point de la distribution cible. Plus formellement, au lieu de considérer des transformations déterministes $T$, Kantorovich a proposé de considérer des mesures sur l'espace produit $\mathbb{R}^d \times \mathbb{R}^d$ qui ont $\mu$ et $\nu$ comme marginales :

$$\mathrm{OT}(\mu, \nu) := \inf_{\gamma \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} c(x, y) d\gamma(x, y)$$

où $\Pi(\mu, \nu) = \{\gamma \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d) \text{ s.t. } \pi_1 \# \gamma = \mu,\ \pi_2 \# \gamma = \nu\}$ est l'ensemble des plans de transport, et $\pi_1 : (x, y) \to x$, $\pi_2 : (x, y) \to y$ sont les projections canoniques. Les minimiseurs de ce problème sont appelés plans de transport optimaux entre $\mu$ et $\nu$. La formulation de Kantorovich est beaucoup plus facile à traiter que le problème de Monge car il s'agit d'un problème d'optimisation linéaire, et une solution à la formulation de Kantorovich existe toujours sous des conditions faibles sur la fonction de coût $c$. Cette nouvelle formulation du TO est beaucoup plus flexible car elle peut traiter des scénarios plus généraux, y compris des cas où les deux distributions ont des formes et des tailles différentes, ou lorsqu'une distribution a des atomes.

**Liens entre les deux formulations.** La relaxation proposée par Kantorovich est en fait une extension étroite du problème de Monge. Lorsque la mesure initiale est sans atome, [111] montre que les formulations de Monge et de Kantorovich coïncident, c'est-à-dire que les deux coûts optimaux sont égaux. Par conséquent, la formulation de Kantorovich peut être considérée comme l'extension minimale du problème de Monge original, qui admet un minimiseur. Nous présentons également quelques cas généraux où une transformation de Monge optimale existe et coïncide avec un couplage optimal résolvant le problème de transport optimal de Kantorovich. En particulier, nous présentons le cas où le coût au sol est de la forme $c(x, y) = h(x - y)$ avec $h$ une fonction strictement convexe et la mesure initiale $\mu$ est absolument continue pour laquelle les formulations de Monge et de Kantorivich admettent le même minimiseur unique. Nous présentons également un cas particulier important du résultat ci-dessus, à savoir lorsque $c$ est le carré de la distance euclidienne. Comme l'a montré Brenier dans son article fondateur, la transformtion de Monge optimale peut être caractérisée comme le gradient d'une fonction convexe.

**Certaines propriétés utiles.** Certaines des principales propriétés du transport optimal de Kantorovich sont également discutées. Nous introduisons les distances de Wasserstein qui sont des cas particuliers du transport optimal de Kantorovich lorsque le coût au sol est une distance $d(x, y)$ sur $\mathbb{R}^d$ à une puissance $p \geq 1$. Ces objets définissent des métriques pour mesurer la distance entre deux distributions de probabilité $\mu$ et $\nu$ avec des moments d'ordre $p$. La distance de Wasserstein

satisfait les trois axiomes métriques et elle métrie la convergence faible. Nous présentons également la formulation duale du TO, définie comme un problème d'optimisation visant à trouver le coût maximisant $\int_{\mathbb{R}^d} f d\mu + \int_{\mathbb{R}^d} g d\nu$ sur toutes les fonctions bornées et continues possibles $f$ et $g$ qui satisfont la contrainte de coût $f \oplus g \leq c$, étant donné $\mu$ et $\nu$ qui sont des distributions de probabilités. Plus formellement, le problème dual d'TO est défini comme suit :

$$\sup \left\{ \int_{\mathbb{R}^d} f d\mu + \int_{\mathbb{R}^d} g d\nu : \text{ s.t. } f, g \in C_b(\mathbb{R}^d) \text{ and } f \oplus g \leq c \right\} .$$

Nous rappelons une condition suffisante pour l'existence d'une solution pour la formulation duale, qui requiert que $\mu$ et $\nu$ soient supportées sur des compacts et que $\mu$ soit continue. Ensuite, sous la même hypothèse, nous présentons un résultat de dualité forte qui montre que la formulation duale et la formulation primale du TO sont égales.

# Chapitre 2. Transport Optimal: Les Défis en Apprentissage Automatique

Ce chapitre présente les défis pratiques de l'application du transport optimal sur les données pour les applications en apprentissage automatique. L'accent est mis sur les mesures de probabilité discrètes et finies, qui constituent le cadre principal de l'application du transport optimal dans l'apprentissage automatique. En particulier, le chapitre se concentre sur la formulation discrète du problème de Monge et sur les difficultés à le résoudre en raison de sa dégénérescence. Il couvre également la formulation discrète de la relaxation de Kantorovich, en soulignant ses limites en termes de complexité de calcul et d'aspects statistiques. En outre, le chapitre présente le TO régularisé par l'entropie comme une approximation du TO qui offre une complexité améliorée et des taux de convergence statistique plus rapides.

**Transport optimal discret.** Le problème du transport optimal discret vise à résoudre TO entre des mesures de probabilité qui sont discrètes (et finies), ce qui signifie que chaque mesure est une somme pondérée de mesures de Dirac supportées sur un nombre fini de points. Si l'on considère la formulation de Monge du TO dans le cadre discret qui cherche une transformation minimisant le coût de transport en associant à chaque point de la mesure initiale un point unique de la mesure d'arrivée et qui doit respecter les contraintes de masses, on peut coder cette transformation à l'aide d'indices et la formuler comme une généralisation du problème de l'assignement optimale. Cependant, le TO de Monge discret et le problème d'assignement sont tous deux limités dans la mesure où le premier est

en général dégénéré tandis que le second ne peut comparer que des histogrammes uniformes de même taille. En outre, l'ensemble réalisable du problème de Monge n'est pas convexe, ce qui le rend difficile à résoudre dans sa formulation originale. En considérant l'approche de Kantorovich dans un cadre discret qui relâche la nature déterministe du transport, permettant à une masse associée à un point d'origine d'être potentiellement répartie sur plusieurs sites, on peut reformuler le problème TO comme un simple programme linéaire en utilisant le formalisme des matrices. Cette approche définit un couplage valide comme une matrice pour encoder la flexibilité du transport probabiliste, qui est toujours symétrique, et le problème de transport optimal qui en résulte peut être résolu à l'aide de l'algorithme "network simplex". Cependant, la résolution du problème reste difficile car elle nécessite une complexité supercubique par rapport au nombre de points et ne peut donc être appliquée qu'à de petits problèmes de taille inférieure à quelques milliers de points. Le problème de l'estimation du coût de transport optimal entre deux distributions, $\mu$ et $\nu$, en utilisant uniquement des échantillons tirés de ces distributions est également présenté. Un estimateur courant de la distance de transport inconnue entre les vraies distributions consiste à la calculer entre les mesures empiriques. Le taux de convergence de la distance estimée vers la distance réelle est souvent appelé "complexité de l'échantillon". La complexité de l'échantillon de la distance p-Wasserstein est présentée et l'erreur moyenne entre la distance estimée et la distance réelle est de l'ordre de $n^{-1/d}$. Cette vitesse est optimale dans $\mathbb{R}^d$ si l'une des mesures a une densité par rapport à la mesure de Lebesgue, mais elle peut être raffinée si les mesures sont supportées sur des sous-domaines de faible dimension.

**Transport optimal entropique.** Le calcul exact du coût de transport optimal dans le cadre discret nécessite la résolution d'un programme linéaire coûteux d'une complexité supercubique. En outre, le TO souffre de la malédiction de la dimensionnalité et est donc susceptible d'être dénué de sens lorsqu'il est utilisé sur des échantillons de densités à haute dimension. Pour remédier à ces problèmes, Cuturi propose de régulariser le problème de TO en ajoutant une pénalité entropique à l'objectif. Ce faisant, il est possible de résoudre exactement ce problème du TO régularisé à l'aide d'un schéma de minimisation alternatif simple, appelé algorithme de Sinkhorn, qui repose uniquement sur les produits matriciels/vectoriels et obtient donc une complexité quadratique améliorée en termes de temps et de mémoire. Plus précisément, le couplage optimal résolvant le TO entropique a la forme $P = \mathrm{diag}(u)K\,\mathrm{diag}(v)$ et les vecteurs d'échelle $u$ et $v$ sont mis à jour à chaque itération de l'algorithme de Sinkhorn à l'aide de simples opérations de remise à l'échelle. La régularisation permet également de surmonter la malédiction de la dimensionnalité pour obtenir de bonnes performances statistiques. Si l'on ajoute

suffisamment d'entropie, le TO entropique entre les mesures empiriques converge vers le coût de TO entropique inconnu entre les mesures réelles avec un taux paramétrique. Toutefois, lorsque la pénalité entropique n'est pas assez importante, l'estimation du TO entropique souffre toujours de la malédiction de la dimension.

# Chapitre 3. Gromov-Wasserstein: Transport Optimal Quadratique

Ce chapitre présente le problème de Gromov-Wasserstein (GW), qui peut être considéré comme la variante quadratique du transport optimal. L'une des principales motivations du problème GW est qu'il permet de comparer des mesures de probabilité même si elles sont supportées par des ensembles incomparables. Dans ce chapitre, nous rappelons les principales définitions et propriétés du problème GW, nous discutons des aspects computationels du problème GW, y compris ses liens avec le problème d'assignement quadratique (PAQ) et sa nature NP-hard, et nous présentons une heuristique alternative basée sur la régularisation entropique afin d'approcher la solution du problème GW.

**Introduction à Gromov-Wasserstein.** Le problème de Gromov-Wasserstein (GW) est une alternative au problème de transport optimal (TO) et est utilisé dans les situations où les mesures de probabilité ont des supports dans des espaces incomparables. Le problème GW consiste à trouver un couplage optimal entre deux mesures de probabilité, sur la base d'une distance $\ell_p$ entre les coûts des deux espaces. De manière plus formelle, supposons que $c_X : X \times X \to \mathbb{R}$ et $c_Y : Y \times Y \to \mathbb{R}$ soient des fonctions continues mesurables, et que $\mu \in \mathbb{P}(X)$, $\nu \in \mathbb{P}(Y)$ soient des mesures de probabilités sur $X, Y$ deux espaces polonais. Le problème de Gromov-Wasserstein est défini comme suit:

$$\mathrm{GW}_p((\mu, c_X), (\nu, c_Y)) = \inf_{\gamma \in \Pi(\mu,\nu)} \left( \int_{X\times Y} \int_{X\times Y} |c_X(x, x') - c_Y(y, y')|^p \, d\pi(x, y) d\pi(x', y') \right)^{1/p}.$$

L'objectif de GW est construit de tel sorte que si un couplage optimal $\gamma$ fait correspondre $x$ à $y$ et $x'$ à $y'$, alors le couple $(x, x')$ doit être "aussi similaire" dans $X$ selon $c_x$ que $(y, y')$ dans $Y$ selon $c_Y$. Lorsque $c_X$, $c_Y$ sont des distances, cela implique que $x, x'$ sont aussi proches dans $X$ que $y, y'$ dans $Y$. Le problème GW admet toujours une solution compte tenu de certaines hypothèses de régularité sur les coûts. De plus, le problème GW définit une distance entre (des classes d'équivalence d'espaces de mesure métrique qui sont des triplets comprenant un espace polonais, une métrique et une mesure de probabilité. GW est invariant par rapport à une large classe de transformations telles que les rotations, les translations

ou les permutations, ce qui est particulièrement utile lorsqu'il s'agit de comparer des formes.

**Aspects informatiques de Gromov-Wasserstein.** Lorsqu'il est appliqué à des mesures de probabilité discrètes, le problème GW peut être reformulé comme un problème d'optimisation quadratique non convexe sur l'ensemble des matrices non négatives satisfaisant des contraintes linéaires. Ce problème est NP-hard en général et notoirement difficile à approcher. Lorsque les mesures discrètes considérées admettent la même taille de support et sont uniformes, le GW peut être considéré comme une relaxation du problème d'assignement quadratique (PAQ). En effet, en restreignant les couplages admissibles à ceux qui sont supportés sur le graphe d'une transformation, qui sont en fait les matrices de permutations, on retrouve exactement le PAQ. En raison de ses limitations calculatoires, Peyré et al. [47] proposent de régulariser le problème GW en ajoutant une pénalité entropique à l'objectif. Ce faisant, les auteurs obtiennent une heuristique simple du problème GW qui consiste à résoudre itérativement des problèmes TO entropiques imbriqués. Ce schéma de calcul, bien qu'il permette de calculer une approximation efficace du problème GW, reste très coûteux car il nécessite, dans le meilleur des cas, une complexité cubique par rapport au nombre de points.

# Chapitre 4. Divergences de Sinkhorn en Temps Linéaire

*Ce chapitre est basé sur [6].*

En raison des obstacles statistiques et calculatoires du TO, ses applications en AA reposent souvent sur une forme de régularisation pour lisser le problème du TO, et sur certaines utilisations plus spécifiques d'une pénalité entropique, pour récupérer les divergences dites de Sinkhorn [112]. Ces divergences sont moins coûteuses à calculer que le TO normal [79, 80], lisses et différentiables par programme dans leurs entrées [46, 55], et ont une meilleure complexité d'échantillonnage [81] tout en définissant toujours des pseudométries convexes et définies [113]. Bien que les divergences de Sinkhorn réduisent les coûts du TO de supercubique à un coût quadratique parallèlisable, son utilisation pour comparer des mesures qui ont plus que quelques dizaines de milliers de points en mode direct reste un défi. L'objectif de ce chapitre est d'introduire un nouveau schéma d'approximation de l'algorithme de Sinkhorn qui peut être calculé en temps linéaire par rapport au nombre de points, ce qui ouvre de nouvelles perspectives pour l'application du TO entropique à l'échelle.

**Travaux liés.** La définition des divergences de Sinkhorn commence généralement par celle du coût au sol des observations. Ce coût est souvent choisi par défaut pour être une norme $q$ entre les vecteurs, ou une distance du plus court chemin sur un graphe lorsqu'on considère des domaines géométriques [114, 115, 116, 44]. Étant donné deux mesures supportées respectivement sur $n$ et $m$ points, le TO entropique instancie d'abord une matrice de coûts $C$ par paire $n\ foism$, pour résoudre un problème de minimisation d'un objectif linéaire pénalisé par l'entropie du couplage. Ce problème peut être réécrit comme une minimisation de Kullback-Leibler :

$$\min_{\text{couplings } P} \langle C, P\rangle - \varepsilon H(P) = \varepsilon \min_{\text{couplings } P} \mathrm{KL}(P, K)\,, \tag{2}$$

où la matrice $K$ est définie comme $K := \exp(-C/\varepsilon)$, la fonction nég-exponentielle par coordonnée d'un coût rééchelonné $C$. Ce problème peut ensuite être résolu à l'aide de l'algorithme de Sinkhorn, qui ne nécessite que l'application répétée du noyau $K$ à vecteurs de mise à l'échelle. Bien que des schémas d'optimisation plus rapides pour calculer le TO régularisé aient été étudiés [117, 118, 119], l'algorithme de Sinkhorn reste, en raison de sa robustesse et de la simplicité de son parallélisme, le choix le plus répendu pour résoudre le TO entropique. Comme le coût de l'algorithme de Sinkhorn est déterminé par le coût de l'application de $K$ à un vecteur, l'accélération de cette évaluation est le moyen le plus efficace d'accélérer l'algorithme de Sinkhorn. C'est le cas lorsque l'on utilise des coûts séparables sur des grilles (appliquer $K$ revient à effectuer une convolution au coût de $(n^{1+1/d})$ [120, Remark 4.17]) ou lorsque l'on utilise la métrique du plus court chemin sur un graphe, auquel cas l'application de $K$ peut être approximée à l'aide d'un noyau thermique [121]. Bien qu'il soit tentant d'utiliser des approximations matricielles de faible rang du noyau $K$ à l'aide de techniques standard, leur application dans le cadre des itérations de Sinkhorn nécessite que l'application du noyau approximé garantisse la positivité de la sortie. En effet, si certaines valeurs du noyau $K$ sont proches de 0 et que l'approximation de chaque entrée n'est pas suffisamment précise, alors le noyau approximé peut avoir une entrée négative, ce qui suffit à faire diverger l'algorithme de Sinkhorn. Dans [122], les auteurs proposent d'utiliser la méthode de Nyström afin d'approximer le noyau $K$ et de diminuer la complexité de l'algorithme de Sinkhorn, cependant, afin de garantir la positivité des entrées du noyau approximé, leur méthode nécessite que $\varepsilon$ soit suffisamment grand et d'avoir une erreur de tolérance très faible pour l'approximation du noyau.

**Nos contributions.** Parce que le TO régularisé peut être effectué en utilisant seulement la définition d'un noyau $K$ avec des entrées positives, nous nous concentrons plutôt sur les noyaux $K$ qui sont garantis d'avoir des entrées positives par conception. En effet, plutôt que de choisir un coût pour définir ensuite un noyau, nous considérons plutôt des coûts de base de la forme $c(x, y) = -\varepsilon \log\langle\varphi(x), \varphi(y)\rangle$

où $\varphi$ est une representation de l'espace de base sur l'orthant positif dans $\mathbb{R}^r$. Ce choix garantit que l'algorithme de Sinkhorn lui-même (qui peut approcher les solutions primales et duales optimales du problème de TO) et l'évaluation des divergences de Sinkhorn peuvent être calculés exactement avec un effort s'échelonnant linéairement en le nombre de points et en $r$, ce qui ouvre de nouvelles perspectives pour l'application du TO à l'échelle. En partant du noyau au lieu du coût pour approximer le TO entropique, nos contributions sont de trois ordres:

- Nous introduisons une famille générale de noyaux admettant une expansion positive et aléatoire et prouvons, sous certaines hypothèses de régularité sur les representations positives des points, que notre méthode est capable d'atteindre une approximation $\delta$ du coût entropique TO en $\mathcal{O}(rn)$ temps et mémoire, où $n$ est le nombre d'échantillons et $r$ est le nombre de caractéristiques aléatoires positives considérées pour approcher le vrai noyau, dès que $r$ s'échelonne en $\log(n)/\delta^2$.

- Nous montrons que les noyaux construits à partir de nos expansions positives peuvent être utilisés pour approximer certaines fonctions de coût habituelles, y compris la distance euclidienne carrée. Nous fournissons pour chacun de ces coûts habituels une formulation explicite de la representation positive associée.

- Nous illustrons la polyvalence de notre approche en étendant les approches TO-GAN précédemment proposées [50, 81], qui se concentraient sur l'apprentissage adverse des fonctions de coût $c_\theta$ et encouraient donc un coût quadratique, à une nouvelle approche qui apprend plutôt de façon adverse un noyau $k_\theta$ induit à partir d'une representation positive $\varphi_\theta$. Nous tirons parti du fait que notre approche est entièrement différentiable dans la representation pour entraîner un GAN à l'échelle, avec des itérations en temps linéaire.

# Chapitre 5. Transport Optimal de Rang Faible: Approche Algorithmique

*Ce chapitre est basé sur [3].*

Il est maintenant de plus en plus clair que la régularisation du TO d'une manière ou d'une autre est le seul moyen d'atténuer les problèmes calculatoires et statistiques [51, 74, 75] du TO. Une approche populaire consiste à pénaliser le problème TO avec une fonction fortement convexe du couplage [76, 77]. Le schéma de régularisation le plus populaire pour TO reste l'approche entropique en raison de sa simplicité et de sa grande capacité de parallélisation sur les GPU. Cependant

sa complexité quadratique à la fois en termes de temps et de mémoire reste un problème majeur quand on veut appliquer le TO sur des problèmes avec plus de quelques dizaines de milliers de points. Une observation clé lorsque l'entropie est ajoutée au couplage est que plus l'entropie est ajoutée, plus le rang est bas (en fait le rang non négatif). Sur la base de cette observation, nous explorons dans ce travail une approche alternative et plus directe pour ajouter de la régularité au problème TO: nous restreignons l'ensemble des couplages réalisables pour qu'ils aient un rang non négatif faible.

**Travaux liés.** Les factorisations de bas rang ne sont pas nouvelles pour le TO régularisé. Elles ont été utilisées pour accélérer la résolution du TO régularisé par l'entropie avec l'algorithme de Sinkhorn, sous réserve de certaines approximations: Étant donné une matrice de coût $C$ dépendante des données, les itérations de Sinkhorn consistent en des produits matrice-vecteur de la forme $Kv$ ou $K^T u$ où $K := \exp(-C/\varepsilon)$ et $u, v$ sont des vecteurs $n, m$. Altschuler et al. [122] et Altschuler and Boix-Adsera [123] ont proposé d'approximer le noyau $K$ avec un produit de matrices fines de rang $r$, $\widetilde{K} = AB^T$. Naturellement, la capacité d'approximer $K$ avec une $\widetilde{K}$ de faible rang se dégrade lorsque $\varepsilon$ diminue, ce qui fait que cette approche n'est valable que pour des $\varepsilon$ suffisamment grands. Grâce à cette approximation, cependant, chaque itération de Sinkhorn est linéaire en $n$ ou $m$ ($\mathcal{O}(n+m)r)$) tant que $r \ll n, m$, et le couplage produit par l'algorithme de Sinkorn est de la forme $\widetilde{P} = CD^T$ où $C = \mathrm{diag}(u)A$, $D = \mathrm{diag}(v)B$. Cette approximation aboutit donc à une solution *faible rang* qui n'est cependant pas rigoureusement optimale pour le problème original tel qu'il est défini par $K$ mais plutôt celui défini par $\widetilde{K}$. La solution obtenue avec $\widetilde{K}$ peut être arbitrairement proche de la vraie solution en augmentant le rang $r$ considéré au prix d'une plus grande complexité. De même, dans Scetbon and Cuturi [6], nous considérons plutôt une factorisation *de faible rang non négatif* pour $K$ de la forme $\widetilde{K} = QR^T$ où $Q, R > 0$ coordonnée par coordonnée. La positivité est essentielle ici car elle garantit la convergence du schéma de Sinkhorn approximatif, et ce pour tout choix de $\varepsilon$. Ce faisant, nous obtenons un couplage qui se rapproche de la solution optimale du TO entropique et de la forme $P = EF^T$ où $E \in \mathbb{R}_+^{n\times r}$ et $F \in \mathbb{R}_+^{m\times r}$. Par conséquent, le couplage produit par ce schéma admet un rang non négatif d'au plus $r$. Cependant, parmi tous les couplages admettant un rang non négatif inférieur à $r$, la solution obtenue par cette méthode n'est en général pas celle qui atteint le coût TO le plus faible, et n'est donc pas optimale à cet égard. À notre connaissance, seuls Forrow et al. [124] ont utilisé des considérations de faible rang pour les couplages, plutôt que sur les coûts ou les noyaux. Leur travail étudie le cas où le coût au sol est le carré de la distance euclidienne. Ils obtiennent pour ce coût une approximation des problèmes du TO de faible rangs en utilisant des barycentres de 2-Wasserstein [125]. Leur

algorithme combine ceux de [126, 127] et aboutit à un plan de transfert de masse intuitif qui passe par un petit nombre de $r$ points, où $r$ est le rang non négatif du couplage.

**Nos contributions.** Dans ce travail, nous proposons une nouvelle alternative au TO entropique pour régulariser le problème du TO en contraignant directement le rang non négatif des couplages admissibles. Notre approche emprunte des idées à [124] mais est générique car elle s'applique à tous les coûts de base. Ici, nous contraignons le rang non négatif de la solution de couplage $P$ dans le problème TO, plutôt que de nous appuyer sur une approximation de faible rang $\widetilde{K}$ pour le noyau $K = e^{-C/\varepsilon}$. Il s'agit d'un point crucial, car la capacité d'approximer $K$ avec une approximation de faible rang $\widetilde{K}$ dépend implicitement du choix de $\varepsilon$ qui peut diminuer lorsque $\varepsilon$ tend vers 0. En revanche, notre approche s'applique à tous les rangs, petits et grands. Pour résoudre ce problème, nous exploitons une reformulation des couplages de faible rang non négatif: ces couplages peuvent être exprimés comme des des couplages de la forme $P = Q\,\mathrm{diag}(1/g)R^T$ décomposés comme le produit de deux sous-couplages fins $Q \in \mathbb{R}_+^{n\times r}$ et $R \in \mathbb{R}_+^{m\times r}$ avec une marge droite commune $g$, et une marge gauche donnée par celles de $P$ de chaque côté. Chacun de ces sous-couplages minimise un coût de transport qui implique la matrice de coût originale $C$ et l'autre sous-couplage. Nous traitons ce problème en optimisant conjointement $Q$, $R$ et $g$ à l'aide d'une approche de descente miroir. Nous prouvons la convergence stationnaire non asymptotique de cette approche. Il est intéressant de noter que nous montrons également qu'une hypothèse de faible rang sur la matrice de coût (et non sur le noyau) peut également être exploitée, fournissant ainsi un scénario du "meilleur des deux mondes" dans lequel les propriétés de faible rang du *couplage* et du *coût* (et non du noyau) peuvent être mises en œuvre et exploitées. En effet, nous montrons que la complexité temporelle de notre algorithme peut devenir linéaire lorsqu'on exploite des hypothèses de faible rang sur le *coût* impliqué dans le problème du TO. Enfin, un parallèle utile peut être établi entre notre approche et celle de l'algorithme vanille de Sinkhorn, dans le sens où ils proposent des schémas de régularisation différents. En effet, le chemin (discret) des solutions obtenues par notre algorithme lorsque l'on fait varier $r$ entre 1 et $\min(n, m)$ peut être considéré comme une alternative au chemin de régularisation entropique. Les deux chemins contiennent à leurs extrêmes la solution TO originale (rang maximal et entropie minimale) et le produit des marginaux (rang minimal et entropie maximale).

# Chapitre 6. Transport Optimal de Rang Faible: Propriétés Théoriques

*Ce chapitre est basé sur [1].*

Bien qu'il soit toujours intuitivement possible de réduire la taille des mesures (par exemple en utilisant $k$-means) avant de résoudre un problème de TO entre elles, une ligne de travail prometteuse propose de combiner les deux [128, 3, 2]. D'un point de vue conceptuel, l'approche de faible rang permet de résoudre simultanément une stratégie optimale de regroupement/agrégation et le calcul d'un transport efficace. Cette intuition repose sur une factorisation explicite des couplages en deux sous-couplages. Cette méthode présente plusieurs avantages sur le plan du calcul, puisque son coût devient linéaire en $n$ si la matrice de coût au sol utilisée pour le problème TO est elle-même de faible rang. Bien que ces améliorations en matière de calcul, pour la plupart démontrées de manière empirique, soient prometteuses, les propriétés théoriques de ces méthodes ne sont pas encore bien établies. Cela contraste fortement avec l'approche de Sinkhorn, qui est comparativement beaucoup mieux comprise. Dans ce chapitre, nous ciblons les principales propriétés théoriques et les aspects pratiques de l'approche de bas rang introduite dans [3] afin de cimenter l'impact des approches de bas rang dans l'TO computationnelle.

**Travaux liés.** Dans un contexte appliqué, nous supposons souvent que nous n'avons accès qu'à des échantillons tirés des distributions qui nous intéressent. Un problème statistique important dans le domaine du transport optimal consiste à estimer le coût de transport optimal (généralement inconnu) entre $\mu \in \mathcal{P}(\mathbb{R}^d)$ et $\nu \in \mathcal{P}(\mathbb{R}^d)$ en utilisant uniquement des échantillons $(x_i)_{i=1}^n$ de $\mu$ et $(y_j)_{j=1}^m$ de $\nu$. Ces échantillons sont supposés être distribués de manière indépendante et identique à partir de leurs distributions respectives. Pour les coûts de transport optimaux, un estimateur simple de la distance inconnue entre les vraies distributions consiste à la calculer directement entre les mesures empiriques $\hat{\mu} := \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ et $\hat{\nu} := \frac{1}{n}\sum_{j=1}^m \delta_{x_j}$, en espérant idéalement pouvoir contrôler le taux de convergence du second vers le premier. Notons qu'ici, $\hat{\nu}$ et $\hat{\nu}$ sont des mesures aléatoires, de sorte que $\mathrm{TO}(\hat{\nu}, \hat{\nu})$ est un nombre aléatoire. Une question importante est la vitesse de convergence de $TO(\hat{\mu}, \hat{\nu})$ vers $\mathrm{TO}(\mu, \nu)$ , et ce taux est souvent appelé "complexité de l'échantillon". Il est bien connu que le TO standard souffre de la malédiction de la dimensionnalité [129]: Sa complexité d'échantillon s'échelonne en $\mathcal{O}(n^{-1/d})$ et est donc exponentielle dans la dimension de l'espace ambiant. Bien qu'il ait été récemment prouvé que ce résultat pouvait être affiné pour prendre en compte la dimension implicite des données, la complexité d'échantillonnage du TO semble maintenant être la principal limitation d'utilisation du TO dans les problèmes

d'apprentissage automatique à haute dimension. Lorsque l'entropie est ajoutée à l'objectif du problème de transport optimal, elle permet également d'améliorer les taux statistiques du TO. Il a été montré dans [51, 82] que le TO entropique bénéficie d'un taux paramétrique $\mathcal{O}\left(\frac{\varepsilon^{-d/2}}{\sqrt{n}}\right)$ par rapport au nombre d'échantillons. Par conséquent, lorsque $\varepsilon$ est suffisamment grand, l'estimateur $\mathrm{TO}_\varepsilon(\hat{\mu}, \hat{\nu})$ bénéficie d'un taux de convergence rapide vers la quantité réelle $\mathrm{TO}_\varepsilon(\mu, \nu)$, mais lorsque $\varepsilon$ tend vers 0, le TO entropique souffre toujours de la malédiction de la dimensionnalité par rapport à son hyperparamètre $\varepsilon$. Bien que toutes les contributions théoriques convergent vers le fait qu'en pratique, le transport entropique est beaucoup plus approprié que le véritable TO lorsqu'il s'agit de comparer des mesures de probabilité discrètes, une question subsiste quant à la quantité à utiliser pour mesurer la différence entre deux distributions basées sur les plans entropiques. En effet, le transport entropique est symétrique mais ce n'est plus une distance car il ne satisfait pas l'inégalité triangulaire, ni une divergence car il n'est pas positif, ni même capable de séparer des distributions car en général le coût du transport entropique entre une mesure et elle-même n'est pas de 0. Pour pallier ces problèmes, Genevay et al. [78] a proposé de soustraire les termes de débiaisage du TO entropique, définissant ainsi la divergence de Sinkhorn. Feydy et al. [130] ont ensuite prouvé que la divergence de Sinkhorn définit une divergence appropriée capable d'interpoler entre la discrépance moyenne maximale (DMM) et le TO lorsque l'on fait varier $\varepsilon$.

**Nos contributions.** L'objectif de cet article est de faire progresser notre connaissance, notre compréhension et notre capacité pratique à exploiter les factorisations de faible rang dans le TO. Cette partie présente cinq contributions, ciblant les propriétés théoriques et pratiques de LOT:

- Nous généralisons la définition du TO à faible rang (LOT), introduite dans [3] dans le cas discret, pour des mesures de probabilité générales et nous étudions le biais induit par les contraintes de faible rang. Nous dérivons le taux de convergence du TO de faible rang vers le vrai TO pour les mesures de probabilité discrètes et générales par rapport au paramètre de rang non négatif $r$.

- Nous faisons un premier pas vers une meilleure compréhension de la complexité statistique du LOT en fournissant une borne supérieure de l'erreur statistique commise lors de l'estimation de LOT à l'aide de l'estimateur plug-in. Étant donné des échantillons tirés indépendamment de mesures de probabilité générales supportées sur des sous-ensembles compacts de $\mathbb{R}^d$, nous montrons que la version empirique du coût de LOT peut être borné par le coût de LOT entre les vraies mesures et un terme d'erreur supplémentaire qui bénéficie d'un taux paramétrique $\mathcal{O}(\sqrt{r/n})$ qui est indépendant de la dimension $d$.

- Nous montrons des liens entre le biais induit par la factorisation de bas rang et les méthodes de clustering. Parce que la contriante de rang non négatif induit un biais sur le problème du TO, le coût LOT entre une mesure et elle-même n'est pas nécessairement 0. Cette valeur et le couplage de rang non négatif faible qui permet de résoudre ce problème reflètent des informations géométriques sur la mesure: Le coût de LOT indique dans quelle mesure une probabilité peut être regroupée en $r$ groupes en fonction du coût de base $c$, tandis que le couplage optimal permet de regrouper les points en fonction de cette géométrie. LOT offre donc une nouvelle méthode de clustering, et ce pour toute géométrie $c$. Dans un cas particulier, lorsque $c = \|\cdot - \cdot\|_2^2$ nous retrouvons la méthode classique de clustering k-means.

- Nous introduisons une version débiaisée de LOT: comme la divergence de Sinkhorn [130], nous montrons que LOT débiaisée est non négative, égale à 0 si et seulement si les deux mesures sont les mêmes, qu'elle métrise la convergence faible, et qu'elle interpole entre la divergence moyenne maximale [18] et le TO lorsque l'on fait varier le rang non négatif $r$. LOT débiaisé possède donc toutes les propriétés géométriques souhaitables pour être utilisé comme fonction de perte pour comparer des distributions, et bénéficie des bonnes propriétés calculatoire du transport de faible rang pour être appliqué à grande échelle.

- Nous proposons des stratégies pratiques pour ajuster le pas de gradient et l'initialisation de l'algorithme présenté dans [3] permettant d'avoir une méthode générique et automatisée pour le choix de ces hyperparamètres, ne laissant qu'un seul hyperparamètre à choisir par l'utilisateur, à savoir le choix du rang non négatif $r$, comme le choix de $\varepsilon$ dans le TO entropique.

# Chapter 7. Les Distances de Gromov Wasserstein de Rang Faible

*This chapter is based on [2].*

La capacité à aligner des nuages points d'espaces incomparables (par exemple, vivant dans des espaces différents) joue un rôle important dans l'apprentissage automatique. Le cadre de Gromov-Wasserstein (GW) fournit une réponse de plus en plus populaire à de tels problèmes, en recherchant un assignement à faible distorsion et préservant la géométrie entre ces points. En tant que généralisation non convexe et quadratique du transport optimal, GW est NP-hard. Tout comme TO est une relaxation du problème d'assignement optimale, GW est une relaxation

du problème d'assignement quadratique. Le GW et le PAQ sont tous deux NP-hard [131]. Bien que les praticiens aient souvent recours à la résolution d'une version régularisée de GW sous la forme d'une séquence imbriquée de problèmes de TO régularisés par l'entropie, la complexité cubique (en nombre $n$ d'échantillons) de cette approche constitue un véritable obstacle en pratique. Nous montrons dans ce chapitre comment notre récente variante du problème de TO qui restreint l'ensemble des couplages admissibles à ceux ayant une factorisation de faible rang non négatif est remarquablement bien adaptée à la résolution de GW.

**Travaux Liés.** Le problème GW remplace l'objectif linéaire dans l'TO par un objectif *non convexe*, *quadratique*, $\mathcal{Q}_{A,B}(P) := \text{cst} - 2\langle APB, P\rangle$ paramétré par *deux* matrices de coût carrées $A$ et $B$. En pratique, la linéarisation itérative de $\mathcal{Q}_{A,B}$ fonctionne bien [84, 85]: recalculer un coût synthétique $C_t := AP_{t-1}B$, utiliser Sinkhorn pour obtenir $P_t := \text{argmin}_P \langle C_t, P\rangle + \varepsilon\text{reg}(P)$, répéter. Cela conduit à un schéma de calcul qui s'étend en temps cubique et nécessite un espace quadratique en mémoire par rapport au nombre de points. Plusieurs obstacles s'opposent à l'accélération de ce schéma du GW entropique. Le recalcul de la matrice de coût impliquée à chaque itération extérieure est un problème, car il nécessite $O(n^3)$ opérations [47, Prop. 1]. Nous ne connaissons que deux approches générales qui permettent d'obtenir des temps d'exécution raisonnables: *(i)* Résoudre des approximations liées, mais significativement différentes, de l'énergie GW, soit en intégrant des points *comme* des mesures univariées [86, 87], soit en utilisant un mécanisme en tranches pour des problèmes euclidiens [88], soit en considérant des métriques arborescentes pour les supports de chaque mesure de probabilité [89], *(ii)* Réduire la taille du problème GW par la quantification des mesures d'entrée [90] ou par des approches de clustering récursif [91, 62]. Il est intéressant de noter qu'à notre connaissance, aucun travail n'a encore tenté d'accélérer les itérations de Sinkhorn dans le cadre de GW.

**Nos contributions.** Notre méthode aborde le problème de l'approximation du coût de GW en utilisant le nouveau schéma de régularisation proposé dans [3] basé sur des contraintes de faible rang. Notre méthode surmonte les limitations liées à la mise à jour de la matrice de coût $C_t$ qui demande une complexité cubique et à la résolution des problèmes de TO entropiques imbriqués nécessitant une complexité quadratique.

- Nous montrons qu'une factorisation (ou approximation) de faible rang des deux matrices de coût d'entrée qui définissent GW, une pour chaque mesure, peut être exploitée pour réduire la complexité du calcul de $C_t$ de cubique à quadratique. Ce faisant, nous sommes également en mesure de réduire la complexité totale du schéma de GW entropique de cubique à quadratique.

- Nous montrons ensuite, de manière indépendante, que l'utilisation de l'approche de faible rang pour les *couplages* préconisée par [3] pour résoudre le TO peut être insérée dans le pipeline GW et donner lieu à une stratégie en $O(n^2)$ pour GW, sans hypothèse préalable sur les matrices de coût d'entrée. Nous expliquons aussi brièvement pourquoi les méthodes qui exploitent les propriétés géométriques de $C_t$ (ou de son noyau $K_t = e^{-C_t}$) pour obtenir des itérations plus rapides sont peu utiles dans une configuration GW, en raison de la nécessité de réinstancier un nouveau coût $C_t$ à chaque itération extérieure.

- Enfin, nous montrons que les deux hypothèses de faible rang (sur les coûts et les couplages) peuvent être combinées pour réduire encore un facteur et atteindre une approximation de GW avec une complexité linéaire en temps et en mémoire. Nous présentons des expériences, sur des ensembles de données simulées et réelles, qui montrent que notre approche a des performances comparables à celles de la méthode entropique et sa capacité pratique à atteindre de "bons" minima locaux de la méthode GW, pour un prix de calcul considérablement moins élevé, et avec un chemin de régularisation conceptuellement différent, tout en pouvant s'étendre à des millions de points.

# Chapitre 8. Transport Equitable et Optimal avec des Agents Multiples

*Ce chapitre est basé sur [5].*

La répartition équitable [94] a été largement étudiée par les communautés de l'intelligence artificielle [95] et de l'économie [96]. La division équitable consiste à répartir diverses ressources entre les agents en fonction de certains critères d'équité. L'un des problèmes classiques de la répartition équitable est le problème de la découpe équitable d'un gâteau [97, 98]. Le gâteau est une ressource hétérogène, comme un gâteau avec différentes garnitures, et les agents ont des préférences hétérogènes sur les différentes parties du gâteau, c'est-à-dire que certaines personnes préfèrent les garnitures au chocolat, d'autres préfèrent les cerises, d'autres encore veulent juste une part aussi grande que possible. Par conséquent, en tenant compte de ces préférences, on peut partager le gâteau équitablement entre les agents. Une généralisation de ce problème, pour lequel il est plus difficile d'obtenir des contraintes d'équité, est le cas où le partage concerne plusieurs gâteaux hétérogènes et où les agents ont des préférences liées sur les différentes parties des gâteaux. Ce problème a de nombreuses variantes, comme le découpage de gâteaux avec deux gâteaux [99], ou l'allocation de ressources multi-types [100, 101]. Dans tous ces modèles, on suppose qu'il n'y a qu'une seule unité indivisible par type de

ressource disponible dans chaque gâteau, et qu'une fois qu'un agent l'a choisie, il doit la prendre en entier. Dans ce cadre, le gâteau peut être considéré comme un ensemble dont chaque élément représente un type de ressource, par exemple chaque élément du gâteau représente un nappage. Ces problèmes sont naturellement assouplis lorsqu'une quantité divisible de chaque type de ressources est disponible. Sur la base de l'idée fondamentale de Kantorovich d'assouplir le problème du TO, nous introduisons dans ce chapitre EOT (Equitable and Optimal Transport), une formulation qui résout à la fois les problèmes de découpage de gâteau et de découpage de gâteau avec deux gâteaux lorsque les ressources sont divisibles.

**Travaux liés.** Le partage équitable des biens a une longue histoire en économie et informatique. Un problème classique est celui du partage équitable du gâteau qui consiste à partager le gâteau entre $N$ individus en fonction de leurs préférences hétérogènes. Le gâteau $\mathcal{X}$, considéré comme un ensemble, est divisé en $\mathcal{X}_1, \ldots, \mathcal{X}_N$ ensembles disjoints entre les $N$ individus. L'utilité d'un seul individu $i$ pour une tranche $S \subset \mathcal{X}$ est notée $V_i(S)$. On suppose souvent que $V_i(\mathcal{X}) = 1$ et que $V_i$ est additif pour les ensembles disjoints. Il existe de nombreux critères pour évaluer l'équité d'une partition $\mathcal{X}_1, \ldots, \mathcal{X}_N$ tels que la proportionnalité ($V_i(\mathcal{X}_i) \geq 1/N$), l'absence d'envie ($V_i(\mathcal{X}_i) \geq V_i(\mathcal{X}_j)$) ou l'équitabilité ($V_i(\mathcal{X}_i) = V_j(\mathcal{X}_j)$). Le problème de la découpe du gâteau a des applications dans de nombreux domaines tels que la division des propriétés foncières, l'espace publicitaire ou le temps de diffusion. Une extension du problème du cake-cutting est le problème du cake-cutting avec deux gâteaux [99] où deux gâteaux hétérogènes sont impliqués. Dans ce problème, les préférences des agents peuvent être couplées sur les deux gâteaux. La part d'un gâteau qu'un agent préfère peut être influencée par la part de l'autre gâteau qu'il peut également obtenir. L'objectif est de trouver une partition des gâteaux qui satisfasse les conditions d'équité pour les agents partageant les gâteaux. Cloutier et al. [99] ont étudié la partition sans envie. Les problèmes de découpage de gâteau et de découpage de gâteau avec deux gâteaux supposent qu'il n'y a qu'une seule unité indivisible de ressource par élément $x \in \mathcal{X}$ du (des) gâteau(x). Par conséquent, partager le(s) gâteau(x) consiste à obtenir une partition de l'ensemble. Cependant, dans ce contexte, le problème peut ne pas être bien posé, et même s'il l'est, sa résolution peut s'avérer difficile en pratique. Dans ce chapitre, nous établissons également les liens de EOT avec certaines métriques intégrales de probabilité (MIP). Les MIP sont des (semi-)métriques sur l'espace des mesures de probabilité. Pour un ensemble de fonctions $\mathcal{F}$ et deux distributions de probabilité $\mu$ et $\nu$, elles sont définies comme $\mathrm{MIP}_{\mathcal{F}}(\mu, \nu) = \sup_{f \in \mathcal{F}} \int f d\mu - \int f d\nu$. Par exemple, lorsque $\mathcal{F}$ est choisi comme étant l'ensemble des fonctions bornées dont la norme infini est inférieure ou égale à 1, nous retrouvons la distance de variation totale (VT) [132]. Ces métriques ont récemment regagné l'intérêt de la communauté de l'AA grâce

à leur application aux Generative Adversarial Networks (GANs) [22] où les MIP sont des métriques naturelles pour le discriminateur [133, 134, 135, 136]. Ils ont également contribué à l'élaboration de tests à deux échantillons cohérents [18, 137]. Cependant, lorsqu'une forme explicite de l'MIP n'est pas disponible, le calcul exact des MIP entre les distributions discrètes peut ne pas être possible ou peut très coûteux. Par exemple, la métrique de Dudley peut être écrite sous la forme d'un programme linéaire [138] qui a au moins la même complexité que le TO standard.

**Nos contributions.** Dans cet article, nous présentons EOT une extension du Transport Optimal qui vise à trouver une stratégie de transport équitable et optimale entre plusieurs agents. Nous apportons les contributions suivantes.

- Nous introduisons EOT qui vise à trouver une partition couplée équitable et optimale des ressources en fonction des préférences hétérogènes des agents. Chaque agent du problème est représenté par une fonction d'utilité (ou de coût), et les ressources allouées à chaque agent sont un sous-couplage tel que leur somme est un couplage valide des ressources satisfaisant les contraintes marginales. Formellement, EOT consiste à trouver la partition qui maximise les plus petites utilités parmi les agents. Du point de vue du transport, EOT vise à diviser la tâche de transport d'une mesure de probabilité vers une autre entre plusieurs travailleurs (ou agents) représentés comme des fonctions de coût afin d'obtenir une partition de la tâche qui soit équitable et optimale. Ici, EOT essaie de minimiser le coût total de transport le plus élevé parmi les travailleurs.

- Nous montrons que EOT résout un problème de division équitable où des ressources hétérogènes doivent être partagées entre plusieurs agents. Plus précisément, nous montrons que les EOT admettent toujours une solution et qu'à l'optimalité, l'utilité totale ou le coût des agents sont égaux et optimaux. Comme sous-produit, nous montrons également que la partition obtenue est non seulement équitable et optimale, mais aussi proportionnelle, qui est un autre critère d'équité important.

- EOT est un problème d'optimisation linéaire sous contraintes linéaires. Nous dérivons son dual et prouvons que la dualité est forte. Comme sous-produit, nous montrons que EOT est lié à certaines familles habituelles de MIP et en particulier à la métrique de Dudley. En conséquence, nous dérivons également des conditions suffisantes sur les fonctions de coût pour que EOT métrise la convergence faible.

- Nous abordons également les aspects calculatoires de EOT et proposons une version régularisée entropique du problème, dérivons sa formulation duale,

obtenons une dualité forte. Nous fournissons ensuite un algorithme efficace pour approximer EOT.

# Chapitre 9. Équilibres de Nash mixtes dans le jeu des exemples adverses

*Ce chapitre est basé sur [4].*

Les exemples d'attaques adverses [102, 103] constituent l'un des problèmes des plus complexes de l'apprentissage automatique: les classifieurs les plus performants sont sensibles à des perturbations imperceptibles de leurs entrées qui les font échouer. Cela pose la question suivante: peut-on construire des classificateurs qui soient robustes face à n'importe quelle attaque adverse ? En supposant que l'on puisse reformuler le problème de minimisation du risque adverse comme un problème min-max, la question ci-dessus reviendrait à demander l'existence d'un équilibre de Nash. Démontrer l'existence d'un tel équilibre garantirait qu'il est possible d'apprendre un classifieur robuste à toute petite perturbation des données, c'est-à-dire à toute attaque qui se produit après l'apprentissage du classificateur. Dans ce chapitre, nous abordons le problème des exemples adverses du point de vue de la théorie des jeux en utilisant les outils du transport optimal et nous étudions la question ouverte de l'existence d'équilibres de Nash mixtes dans le jeu à somme nulle formé par l'attaquant et le classifieur.

**Travaux liés.** Une ligne de recherche récente a soutenu que les classifieurs randomisés pourraient aider à contrer les attaques adverses [139, 140, 141, 142]. Dans le même ordre d'idées, [143] a démontré, à l'aide de la théorie des jeux, que les classifieurs randomisés sont en effet plus robustes que les classifieurs déterministes face à des adversaires régularisés. Toutefois, les conclusions de ces travaux antérieurs dépendent de la définition de l'adversaire considéré. En particulier, ils n'ont pas étudié les scénarios dans lesquels l'adversaire utilise également des stratégies aléatoires, ce qui est essentiel si nous voulons donner une réponse de principe à la question de l'existence d'un classifieur robuste contre toute attaque adverse. Les travaux antérieurs qui étudient les exemples adverses dans le cadre de la théorie des jeux ont examiné le cadre aléatoire (à la fois pour le classifieur et l'adversaire) dans des contextes restreints où l'adversaire est soit paramétrique, soit dispose d'un nombre fini de stratégies [144, 145, 146]. Des exemples d'adversaires ont été étudiés sous les notions de jeu de Stackelberg dans [147], et de jeu à somme nulle dans [144, 145, 146]. Ces travaux ont pris en compte des paramètres restreints (perte convexe, adversaires paramétriques, etc.) qui ne correspondent

pas à la nature du problème. En effet, il a été prouvé qu'aucune perte convexe ne peut être un bon substitut pour la perte 0/1 dans le cadre adverse [148, 149], ce qui réduit le champ d'application de ces résultats. Si l'on peut montrer que pour des distributions conditionnelles suffisamment séparées, un classificateur déterministe optimal existe toujours, des conditions nécessaires et suffisantes pour la nécessité de la randomisation restent à établir. Pinot et al. [143] ont étudié en partie cette question pour des adversaires déterministes régularisés, laissant le cadre général des adversaires randomisés et des équilibres mixtes sans réponse, ce qui est la portée même de cet article. Bhagoji et al. [150] et Pydi and Jog [151] ont étudié les limites inférieures du risque adverse de tout classificateur déterministe en utilisant le TO. Ces travaux n'évaluent que les limites inférieures de la formulation déterministe primaire du problème, alors que nous étudions l'existence d'équilibres de Nash mixtes. Notons que Pydi and Jog [151] a commencé à étudier un moyen de formaliser l'adversaire en utilisant des noyaux de Markov, mais n'a pas étudié l'impact des stratégies aléatoires sur le jeu. Une autre ligne de travaux [152, 153, 154] a étudié le problème des exemples adverses dans le cadre de l'optimisation distributionnellement robuste. Dans ces cadres, l'ensemble des distributions adverses est défini à l'aide d'une boule de Wasserstein $\ell_p$ (l'adversaire est autorisé à avoir une perturbation *moyenne* d'au plus $\varepsilon$ dans la norme $\ell_p$). Cependant, cela ne correspond pas au problème habituel de l'attaque adverse, où l'adversaire ne peut déplacer aucun point de plus de $\varepsilon$.

**Nos contributions.** Dans ce chapitre, nous étudions le problème des exemples adverses en le formulant comme un jeu à somme nulle grâce au transport optimal. Plus précisément, nous apportons les contributions suivantes.

- Dans la formulation standard du risque adverse, l'adversaire est défini comme une fonction qui fait correspondre chaque point de l'ensemble de données à un point adverse restreint à vivre dans une boule centrée sur le point de données maximisant la fonction perte. En utilisant l'idée fondamentale de Kantorovich, nous étendons le travail de [151] et assouplissons cette formulation afin de permettre à l'adversaire d'être à la place un couplage entre la distribution des données et celle de l'adversaire. Cette relaxation est en fait étroite puisque nous montrons que les deux formulations sont égales alors que la dernière peut être formulée comme un problème d'optimisation. En effet, nous montrons que le risque adverse peut être reformulé comme un problème linéaire de maximisation sur des distributions restreintes à vivre dans une boule de Wasserstein spécifique centrée sur la distribution des données.

- En utilisant notre formulation variationnelle du risque adverse, nous montrons que le problème de minimisation du risque adverse peut être reformulé comme

un problème d'optimisation distributionnellement robuste (ODR) [155]. Cette formulation nous amène naturellement à analyser la minimisation du risque adverse comme un jeu à deux joueurs et à somme nulle.

- Nous montrons qu'en général, il n'existe pas d'équilibre de Nash dans ce type de jeu et nous fournissons un exemple simple montrant la nécessité d'utiliser des stratégies aléatoires à la fois pour l'attaquant et pour le classifieur. Nous montrons ensuite que dans ce jeu, lorsque l'adversaire et le classifieur sont tous deux autoriser à utiliser des stratégies aléatoires, il est toujours possible d'atteindre un équilibre de Nash.

- Enfin, nous concevons des algorithmes efficaces pour apprendre un mélange fini de classificateurs. En nous inspirant de l'optimisation robuste [152] et des méthodes de sous-gradient [156], nous dérivons un algorithme oracle pour optimiser un mélange fini. Ensuite, en suivant la ligne de travail de [76], nous introduisons une régularisation entropique pour calculer efficacement une approximation du mélange optimal. Nous validons nos résultats par des expériences sur des ensembles de données simulées et réelles, à savoir CIFAR-10 et CIFAR-100 [157].

# Notations

### Operations

| | |
|---|---|
| $\pi_1 : (x, y) \mapsto x$ | The canonical projection on the first variable |
| $\pi_2 : (x, y) \mapsto y$ | The canonical projection on the second variable |
| $\mu \ll \nu$ | $\mu$ is absolutely continuous w.r.t. $\nu$ |
| $\mu \otimes \nu$ | The tensor product of the measures $\mu$ and $\nu$ |
| $A \odot B$ | The Hadamard product of the matrices $A$ and $B$ |
| $f \oplus g$ | The tensor sum of the vectors $f$ and $g$ |
| $f \in \mathcal{O}(g)$ | $f \leq Cg$ for a universal constant $C$ |
| $f \in \Omega(g)$ | $g \leq Qf$ for a universal constant $Q$ |
| $g\#\mu$ | The pushforward measure of $\mu$ by an application $g$ |

### Spaces

| | |
|---|---|
| $\Delta_n$ | The probability simplex of size $n$ |
| $\Delta_n^+$ | The probability simplex of size $n$ of positive histograms |
| $\mathcal{M}(\mathcal{X})$ | The set of Radon measures on $\mathcal{X}$ |
| $\mathcal{M}_+(\mathcal{X})$ | The sets of positive Radon measures on $\mathcal{X}$ |
| $\mathcal{P}(\mathcal{X})$ | The power of probability measures on $\mathcal{X}$ |
| $\mathcal{X}$ | A Polish space |
| $\mathrm{C}^+(\mathcal{X} \times \mathcal{Y})$ | The space of non-negative continuous functions on $\mathcal{X} \times \mathcal{Y}$ |
| $\mathrm{C}_*^-(\mathcal{X} \times \mathcal{Y})$ | The set of negative continuous functions on $\mathcal{X} \times \mathcal{Y}$ |

| | |
|---|---|
| $\mathrm{LSC}(\mathcal{X} \times \mathcal{Y})$ | The space of lower semi-continuous functions on $\mathcal{X} \times \mathcal{Y}$ |
| $\mathrm{LSC}^{+}(\mathcal{X} \times \mathcal{Y})$ | The space of non-negative lower semi-continuous functions on $\mathcal{X} \times \mathcal{Y}$ |
| $\mathrm{LSC}_{*}^{-}(\mathcal{X} \times \mathcal{Y})$ | The set of negative bounded below lower semi-continuous functions on $\mathcal{X} \times \mathcal{Y}$ |
| $\mathbb{R}$ | The set of real numbers |
| $C_b(\mathcal{X})$ | the vector space of bounded continuous functions on $\mathcal{X}$ |

**Other symbols**

| | |
|---|---|
| $\mathbf{1}_n$ | the vector of dimension $n$ with only 1 as entries |
| $\|\cdot\|_2$ | the Euclidean norm |

# Part I

# Background on Optimal Transport

In this part, we introduce the key results and concepts from the OT theory on which this thesis will rely. This presentation puts the accent on the computational aspects of OT, with the end goal of applying OT tools to ML problems.

We start by presenting the original Monge formulation of OT and its Kantorovich relaxation in Chapter 1. More precisely, we start by introducing the original OT problem as introduced by Monge and present its main limitations. Monge defines the OT problem as the minimal cost of transporting a probability measure to another according to a ground cost function $c(x, y)$ measuring the cost of moving one unit of mass from the location $x$ to the location $y$. In his formulation, Monge only allows transporting distributions according to maps $T$ that send each point of the initial measure to a point $T(x)$ in the support of the target measure. Therefore the Monge problem is defined as an optimization problem over those maps that are effectively able to transport exactly the probability measures, however, this constraint on the transportation task makes the problem ill-posed sometimes and even when it is well-posed, it is in general hard to solve. To alleviate these issues, Kantorovich proposes to relax the transportation strategy by replacing these deterministic mappings using instead probabilistic ones, or couplings. By doing so, Kantorovich allows that one point $x$ from the initial measure to be mapped to several points in the target measure by splitting the mass associated to the point $x$. This relaxation admits (almost) always a solution, and can be seen as the minimal convex relaxation of the OT problem admitting a minimizer. Finally, we also present some fundamental properties of this new formulation of OT such as the existence of a dual, or metric properties on the space of distributions.

In Chapter 2, we delve into the practical challenges of applying OT on data for ML applications. We only focus on the application of OT on discrete and finite probability measures which is the main setting of application of OT in ML. More specifically, we first introduce the discrete formulation of the Monge problem and emphasize on the degeneracy as well as the difficulty to solve this formulation in the discrete setting. We then consider the discrete formulation of the Kantorovich relaxation and introduce the analysis on the computational as well as the statistical aspect of OT when dealing with discrete and finite probability measures. While the discrete Monge formulation cannot be solve in general, the discrete Kantorovich one always admits a solution. However solving the Kantorovich OT problem requires to solve a network flow problem with a super-cubic complexity with respect to the number of points, limiting its applications in ML for only small size problems. In addition when the discrete measures at hands are in fact empirical measures associated to two distributions supported on a subset of $\mathbb{R}^d$, an important statistical question is to control the convergence rate of the plug-in estimator using

these empirical measures. Once again, OT is not adapted for high-dimensional applications as its statistical rate suffers from the curse of dimensionality. Then we present entropy-regularized OT initially introduced as an approximation of OT with improved complexity using Sinkhorn's algorithm. Indeed, when some entropy is added to the objective of the OT problem, then this regularized version of OT can be solved in quadratic time with respect to the number of points. In addition, this regularization enjoys fast statistical rate of convergence when enough entropy is added and so even in the high-dimensional setting. However, this new formulation of the transport, although much more advantageous than the unregularized version when working with data, still does not allow the use of OT in large scale problems because of its quadratic complexity.

In Chapter 3, we present the quadratic version of optimal transport, namely the Gromov-Wasserstein problem. We first recall the definition of the problem and its the major properties. GW can be viewed as the quadractic version of the Kantorovich formulation of OT and is deeply connected to the Gromov-Hausdorff distance. Indeed, GW is a relaxed version of the latter where metric spaces have been endowed with a probability measures in order to be compared. Similar to the Kantorovich formulation of OT, GW is defined as a minimization problem and always admit a solution. One of the most important feature of the GW problem is that it defines a distance over a set of very structured objects, namely the set metric measure spaces (quotiented by isomorphisms). These objects are richer that probabilities as they also contains a metric information on the points where are supported the measures. This metric structure added to the measures allows the comparison of probability measures even if they are supported on incomparable sets (e.g. living in different spaces). Additionally, GW is invariant with respect to a large class of transformations such as rotations, translation or permutations and therefore provides an increasingly popular answer to compare shapes. Then we dive in the computational aspects of the GW problem. In the discrete setting, one can establish strong connections between GW and the Quadratic Assignment Problem (QAP). While GW allows the search of the optimal matching between measures on the set of all couplings satisfying the marginal constraints, the QAP restrict its search to deterministic maps only. QAP can be seen as the Monge formulation of GW problem. Although GW is very linked to OT, because of its quadratic formulation, the problem is in general non-convex and NP-hard. Since GW is an NP-hard problem, all its applications rely on heuristics, the most popular being the entropic regularization. Adding an entropic term to the objective allows to derive a simple algorithm that takes advantage of the Sinkhorn scheme by solving a sequence of nested entropy-regularized OT problems. However, that approximation remains costly, requiring $\mathcal{O}(n^3)$ operations when dealing with two datasets of $n$

samples, preventing the use of GW for problems larger than few thousand points.

# Chapter 1

# Optimal Transport: From Monge to Kantorovich

Optimal transport is a branch of mathematics that studies the problem of finding the most efficient way to move one distribution to another. The notion of optimal transport can be traced back to the work of Gaspard Monge in the late 18th century, who studied the problem of transporting soil from one location to another in order to level the ground. In the 20th century, the problem was rediscovered and formalized mathematically, and has since found numerous applications in mathematics, physics, economics, and computer science. In this chapter, we give an introduction to optimal transport and present the main results upon which this thesis is built. Inspired by the reference books of Villani [30, 158], Santambrogio [159], and Peyré and Cuturi [33], we provide a general presentation of the optimal transport problem as originally introduced by Monge [31] and further relaxed by Kantorovich [32].

## 1.1 Monge Optimal Transport

The original formulation of optimal transport was proposed by Gaspard Monge in 1781, and is known as the Monge problem. Given two measures of equal mass $\mu$ and $\nu$ living in $\mathcal{P}(\mathbb{R}^d)$ and a cost function $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$, Monge raised the problem of transporting $\mu$ to $\nu$ optimally w.r.t. $c$. More formally, this problem can be stated as

$$\inf_{T:\ T\#\mu=\nu} \int_{\mathbb{R}^d} c(x, T(x)) d\mu(x) \tag{1.1}$$

where $T\#\mu$ is the pushforward measure of $\mu$ by $T$, defined by $T\#\mu(A) := \mu(T^{-1}(A))$ for all $\mu$-measurable sets. A map $T$ satisfying the constraint $T\#\mu = \nu$ is called a transport map between $\mu$ and $\nu$ and assigns to each point $x$ in the support of

the initial measure $\mu$ a point $T(x)$ in the support of the target measure $\nu$, and transports also all the (infinitesimal) mass of $\mu$ located at $x$ to $T(x)$. When it exists, an optimal transport map solving (1.1) is called a Monge map. However, the Monge problem may not always be well-defined, and even when it is, it can be challenging to solve the optimization problem.

**A transport map might not even exists.** As an example, consider the case where $\mu$ is a Dirac distribution. Then, $T\#\mu$ is necessarily also a Dirac distribution, hence there can be no transport in Monge's sense if $\nu$ is not a Dirac distribution as well. This also highlights the intrinsic asymmetry of (1.1), as conversely, it is always possible to find a Monge map going to a Dirac measure $\delta_y$, by setting $\forall x,\ T(x) = y$. Santambrogio [111] proposes a simple sufficient condition for the existence of a transport map: atomeless measure $\mu$ guarentees that the constraint space of the Monge problem (1.1) is non-empty.

**Theorem 1.1.1** ([111])**.** *If $\mu$, $\nu$ are two probability measures on $\mathbb{R}^d$ and $\mu$ is atomless, then there exists at least a transport map $T$ such that $T\#\mu = \nu$.*

**However, even if a transport map exists, there may be none that is optimal...** In [111], the author propose a simple example where the set of constraint $T\#\mu = \nu$ is not empty, however the infimum cannot be reached by such maps. Indeed, Santambrogio [111] proposes to consider the case where $c(x, y) = \|x - y\|^2$ in $\mathbb{R}^2$, $\mu$ is the uniform probability measure on $\{0\} \times [0, 1]$ and $\nu$ is the mixture of two uniform probability measures, $\nu_1$ and $\nu_2$, with equal weights $1/2$ on respectively $\{-1\} \times [0, 1]$ and $\{1\} \times [0, 1]$, that is $\nu = \frac{1}{2}(\nu_1 + \nu_2)$ (see Figure 1.1). The existence of maps satisfying the constraint of the Monge problem, i.e. $T\#\mu = \nu$ is clear as one can simply split the initial measure $\mu$ in two such that each sub-measure have same total mass, and then sends one towards $\nu_1$ and the other towards $\nu_2$. However none of these maps are able to reach the infimum of the Monge problem because this would imply that a map can send horizontally each point $x$ which is not possible.

In the next section, we present the fundamental relaxation of the Monge problem introduced by Kantorovich [32] in order to overcome the difficulties of solving the Monge problem (1.1).

## 1.2 Kantorovich Optimal Transport

In order to overcome these difficulties, Kantorovich proposed a relaxation of the Monge problem, which is now known as the Kantorovich formulation. In this formulation, instead of seeking a transport map that moves all the mass from each

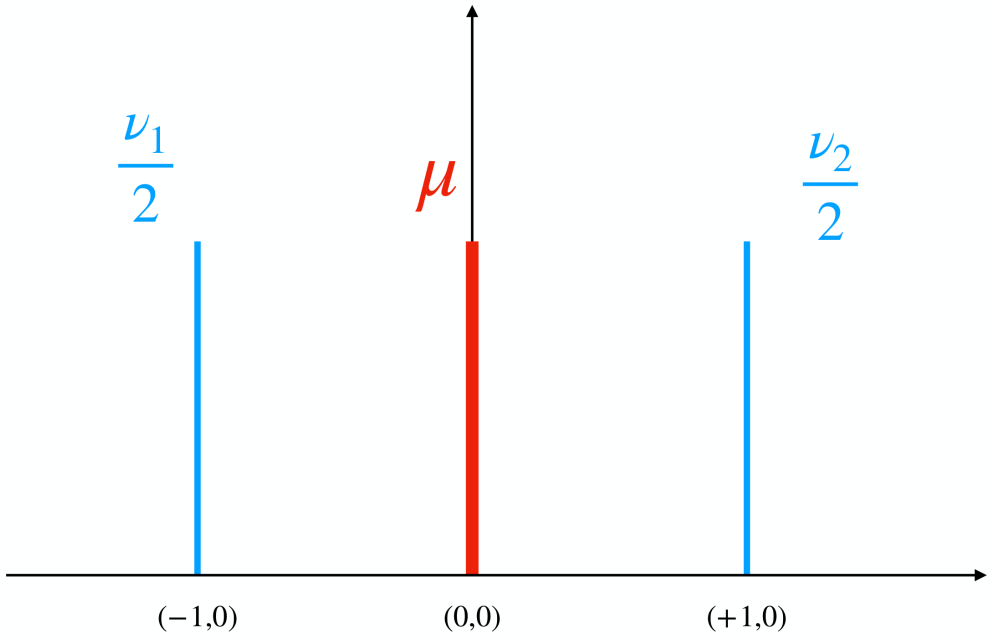


Figure 1.1: Two measures $\mu, \nu \in \mathcal{P}(\mathbb{R}^2)$ where $\nu = \frac{1}{2}(\nu_1 + \nu_2)$ such that there exists an infinite number of maps $T$ satisfying $T\#\mu = \nu$, however none of them realizes the infimum of the Monge problem 1.1.

point of the source distribution to the target distribution, one seeks a transport plan that specify how much mass is moved from each point in the source distribution to each point in the target distribution. The cost function is then minimized over this set of transport plans, which is often easier to compute. More formally, instead of considering deterministic maps $T$, Kantorovich proposed to consider probabilistic map, i.e. measures $\gamma$ over the product space $\mathbb{R}^d \times \mathbb{R}^d$ that have $\mu$ and $\nu$ as marginals:

$$\mathrm{OT}_c(\mu, \nu) := \inf_{\gamma \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} c(x, y) d\gamma(x, y) \tag{1.2}$$

where $\Pi(\mu, \nu) = \{\gamma \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$ s.t. $\pi_1\#\gamma = \mu,\ \pi_2\#\gamma = \nu\}$ is the set of transportation plans, and $\pi_1 : (x, y) \to x$, $\pi_2 : (x, y) \to y$ are the canonical projections. These probability measures over $X \times Y$ are an alternative way to describe the displacement of the particles of $\mu$: we specify for each pair of measurable sets $(A, B)$ how much mass goes from $A$ to $B$. More precisely, the value $\gamma(A \times B)$ denotes the amount of mass moving from $A$ to $B$. It is clear that this description allows for more general movements, since from a single point $x$, masses can a priori move to different destinations $y$. If multiple destinations really occur, then this movement cannot be described through a map $T$. Note that the constraints

on $\gamma(A \times \mathbb{R}^d)$ and $\gamma(\mathbb{R}^d \times B)$ exactly mean that we restrict our attention to the movements that really take the distribution $\mu$ and move it onto the distribution $\nu$. The minimizers for this problem are called optimal transport plans between $\mu$ and $\nu$.

If an optimal transport plan $\gamma$ is of the form $(\mathrm{id}, \times T)\#\mu$ for a measurable map $T\colon X \to Y$ (i.e., when no splitting of the mass occurs), the map $T$ would be called the optimal transport map from $\mu$ to $\nu$. It can be easily checked that $(\mathrm{id}, T)\#\mu$ belongs to $\Pi(\mu, \nu)$ i.f.f $T$ pushes $\mu$ onto $\nu$ and the objective of the Kantorovich problem (1.2) takes the form $\int_{\mathbb{R}^d} c(x, T(x))d\mu(x)$, thus generalizing the Monge problem (1.1). This generalized problem proposed by Kantorovich is much easier to handle than the original one proposed by Monge: it is a linear optimization problem under linear constraints and there always exist transport plans in $\Pi(\mu, \nu)$ (for instance, $\mu \otimes \nu \in \Pi(\mu, \nu)$). Another key advantage of this formulation is that a solution to (1.2) exists under weak conditions on the cost function $c$.

**Theorem 1.2.1** (Santambrogio [111]). *Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ and $c : \mathbb{R}^d \times \mathbb{R}^d \to [0, +\infty]$ be a lower semi-continuous ground cost. Then (1.2) admits a solution.*

Therefore the Kantorovich formulation can handle more general scenarios, including cases where the two distributions have different shapes and sizes, or when one distribution has atoms. Then if one is concerned with the Monge problem, a natural inquiry arises: is the minimum cost obtained in the Kantorovich problem equivalent to the minimum cost in the Monge problem? Furthermore, if this equivalence holds, can the minimum transportation plan $\gamma$ obtained in the Kantorovich problem be realized by a transport map $T$? The next section explores the interconnections between these two approaches to the optimal transport problem.

## 1.3 Links Between the Two Formulations

**Kantorovich and Monge coincide.** From the definitions (1.1), (1.2), we already know that

$$\inf_{T:\ T\#\mu=\nu} \int_{\mathbb{R}^d} c(x, T(x))d\mu(x) \geq \inf_{\gamma \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} c(x, y)d\gamma(x, y)$$

because any transport map $T$ induces a valid coupling $\gamma$. In fact, as soon as the initial measure $\mu$ is atomeless, then both quantities coincides.

**Theorem 1.3.1** (Santambrogio [111]). *Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ compactly supported and $c : \mathbb{R}^d \times \mathbb{R}^d \to [0, +\infty]$ be a lower semi-continuous ground cost. Then if $\mu$ is*

*atomeless then we have:*

$$\inf_{T:\ T\#\mu=\nu} \int_{\mathbb{R}^d} c(x,T(x))d\mu(x) = \min_{\gamma\in\Pi(\mu,\nu)} \int_{\mathbb{R}^d\times\mathbb{R}^d} c(x,y)d\gamma(x,y)$$

Therefore the Kantorovich formulation (1.2) can be seen as the minimal extension of the original problem formulated by Monge (1.1) which admits a minimizer. In light of this consideration, it is now natural to ask when an optimal map solving (1.1) exists in order for the solutions of the problems to coincide.

**On the existence of a Monge map.** For an absolutely continuous measure $\mu$, Theorems 1.3.2 and 1.3.3 below show that under conditions on the cost function and compactness assumptions, solution to (1.1) exists and coincides with the solution of (1.2) in the coupling formalism.

**Theorem 1.3.2** (Santambrogio [111])**.** *Let $\mu,\nu\in\mathcal{P}(\mathbb{R}^d)$ be compactly supported, and such that $\mu$ is a.c. Consider a cost function $c(x,y)=h(x-y)$ where $h$ is a strictly convex function. Then, there exists a unique optimal transport map $T$ and a unique optimal coupling $\gamma$, and $T$ and $\gamma$ are related by $\gamma=(id,T)\#\mu$.*

Hence, under the conditions of Theorem 1.3.2, an optimal Monge map exists and can equivalently be described as an optimal transportation plan supported on its graph. In particular, for absolutely continuous and compactly supported $\mu$ and $\nu$, Theorem 1.3.2 holds when $c(x,y)=\|x-y\|^p$ with $p>1$. The $p=2$ case holds a particular place in the optimal transport theory, as shown by Brenier in his seminal paper [160].

**Theorem 1.3.3** (Brenier [160])**.** *Let $\mu,\nu\in\mathcal{P}(\mathbb{R}^d)$ such that $\mu$ is a.c., and $c(x,y)=\|x-y\|^2$. Then, problem (1.1) admits a unique solution, which is characterized (among all transport maps) as being the gradient of a convex function $\phi$ : $\forall x\in\mathbb{R}^d$, $T(x)=\nabla\phi(x)$.*

Compared to Theorem 1.3.2, the major contribution of Theorem 1.3.3 is the unique characterization of the transport map as the gradient of a convex function. As an example, it implies the following immediate corollary.

**Corollary 1.3.1.** *Let $\mu\in\mathcal{P}(\mathbb{R}^d)$ be a.c., $c(x,y)=\|x-y\|^2$ and $\phi:\ \mathbb{R}^d\to\mathbb{R}$ a convex function. Then, $\nabla\phi$ is the unique optimal Monge map from $\mu$ to $\nabla\phi$.*

Now that we have seen that the Kantorovich and the Monge formulation can be identified in a very general setting, we explore in the next section some important properties of the Kantorovich relaxation.

## 1.4 Some Useful Properties

**Wasserstein distances.** When the ground cost $c$ is actually a distance $d(x, y)$ on $\mathbb{R}^d$ to a power $p \geq 1$ and when $\mu$, $\nu$ have moments of order $p$, the Wasserstein distances can be defined from (1.2).

**Definition 1.4.1** (Wasserstein Distances)**.** *Let $p \geq 1$ and $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$. The $p$-Wasserstein distance is defined as*

$$W_p(\mu, \nu) := \inf_{\gamma \in \Pi(\mu,\nu)} \left( \int_{\mathbb{R}^d \times \mathbb{R}^d} d(x, y)^p d\gamma(x, y)) \right)^{1/p} \tag{1.3}$$

Wasserstein distances satisfy all three metric axioms on $\mathcal{P}_p(\mathbb{R}^d)$ Santambrogio [111, Prop 5.1], and metrize weak convergence plus convergence of moments of order $p$ Santambrogio [111, Thm 5.11]. In machine learning, the metrization of weak convergence is a crucial requirement for measure discrepancies, as we are often interested in minimizing the value of a loss function integrated against probability distributions.

**A dual formulation of OT.** The problem (1.2) is a linear optimization under convex constraints, given by linear equalities or inequalities. Hence, an important tool will be duality theory, which is typically used for convex problems. For now let us just introduce the following optimization problem.

**Definition 1.4.2.** *Given $\mu \in \mathcal{P}(\mathbb{R}^d)$, $\nu \in \mathcal{P}(\mathbb{R}^d)$, and the cost function $c : \mathbb{R}^d \times \mathbb{R}^d \to [0, +\infty]$, we consider the problem:*

$$D_c(\mu, \nu) = \sup \left\{ \int_{\mathbb{R}^d} f d\mu + \int_{\mathbb{R}^d} g d\nu : \ \text{s.t. } f, g \in C_b(\mathbb{R}^d) \text{ and } f \oplus g \leq c \right\} \tag{1.4}$$

First of all, we notice that

$$D(\mu, \nu) \leq \inf_{\gamma \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} c(x, y) d\gamma(x, y) \ ,$$

as it is enough to integrate the condition $f \oplus g \leq c$ according to $\gamma$ to get

$$\int_{\mathbb{R}^d} f d\mu + \int_{\mathbb{R}^d} g d\nu \leq \int_{\mathbb{R}^d \times \mathbb{R}^d} c d\gamma \ .$$

This is valid for every admissible $(f, g)$ and every admissible $\gamma$ and proves the desired inequality. Yet, (1.4) does not admit a straightforward existence result, since the class of admissible functions lacks compactness. In the following proposition, we present a sufficient condition for existence of solutions for (1.4)

**Proposition 1.4.1.** *(Santambrogio [111]) Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ and suppose that $\mu$ and $\nu$ are compactly and $c$ is continuous. Then there exists a solution to problem (1.4).*

We can now present the main theorem showed in [159, Theorem 1.39] stating that the problem defined in (1.4) is the dual of the Kantorovich formulation (1.2) and that strong duality holds under some regularity assumptions on the cost $c$.

**Theorem 1.4.1.** *(Strong Duality) Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ and suppose that $\mu$ and $\nu$ are compactly and $c$ is continuous. Then strong duality holds and we have:*

$$\begin{aligned} &\max\left\{\int_{\mathbb{R}^d} f d\mu + \int_{\mathbb{R}^d} g d\nu : \ \text{s.t. } f, g \in C_b(\mathbb{R}^d) \text{ and } f \oplus g \leq c\right\} \\ &= \min_{\gamma \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} c(x,y) d\gamma(x,y)) \ . \end{aligned} \tag{1.5}$$

*This fundamental idea proposed by Kantorovich to relax the Monge problem will be largely exploited in Part III in order to analyze longstanding ML problems using optimal transport.*

# Chapter 2

# Optimal Transport: Challenges in Machine Learning

In practice, we often work with finite sets of data, which implies that we have access only to discrete empirical measures. Thus, to solve optimal transport problems, we need to consider the discrete setting, where the measures are approximated by finite sets of points. In this case, we have a finite number of mass points, and we need to find an optimal transport plan/map that assigns each point of the source measure to a (or multiple) point(s) of the target measure, minimizing the total cost of transport. This is known as the discrete optimal transport problem, which has become increasingly popular in recent years due to its numerous applications in data analysis, computer vision, machine learning, and other fields. The discrete setting presents some particular challenges, such as computational efficiency and the effect of the discretization of the measures on the statistical properties of OT. This chapter puts the accent on these two challenges, with the end goal of applying OT tools to machine learning (ML) problems.

## 2.1 Discrete Optimal Transport

Discrete optimal transport deals with the case where the measure spaces are discrete. In this case, each measure is a weighted sum of Dirac measures supported on finitely many points. That is, if $\mu$ is a measure on a set $\{x_1, \dots, x_n\} \subset \mathbb{R}^d$, then we can write it as $\mu = \sum_{i=1}^n a_i \delta_{x_i}$, where $a_i > 0$ and $\sum_{i=1}^n a_i = 1$. Similarly, if $\nu$ is a measure on a set $\{y_1, \dots, y_m\}$, then we can write it as $\nu = \sum_{j=1}^m b_j \delta_{y_j}$.

**Discrete Monge optimal transport.** For discrete measures $\mu = \sum_{i=1}^n a_i \delta_{x_i}$ and $\nu = \sum_{j=1}^m b_j \delta_{y_j}$, the Monge problem (1.1) seeks a map that associates to each point $x_i$ a single point $y_j$ and which must push the mass of $\mu$ toward the mass of $\nu$.

Formally, the Monge problem considers map $T : \{x_1, \dots, x_n\} \to \{y_1, \dots, y_m\}$ that must verify that

$$\forall j \in 1, \dots, m, \quad b_j = \sum_{i:T(x_i)=y_j} a_i \ .$$

Then the Monge map should minimize the total transportation cost,

$$\min_T \left\{ \sum_i c(x_i, T(x_i)) \ : \ T\#\mu = \nu \right\} \ . \tag{2.1}$$

If all $x$'s and $y$'s are distinct, a map between discrete points can be represented using indices $\sigma : [|n|] \to [|m|]$, where $j = \sigma(i)$. The mass conservation is expressed as $\forall j \in 1, \dots, m, \quad b_j = \sum_{i \in \sigma^{-1}(j)} a_i$, where $\sigma^{-1}(j)$ denotes the preimage set of $j$. When $n = m$ and all weights are uniform ($a_i = b_j = 1/n$), then the mass conservation constraint implies that $T$ is a bijection, and the Monge problem is equivalent to the optimal assignment problem, which seeks to find the optimal permutation $\sigma$ $T(x_i) = y_{\sigma(i)}$:

$$\min_{\sigma \in \mathrm{Perm}(n)} \frac{1}{n} \sum_i C_{i,\sigma(i)} \tag{2.2}$$

where $\mathrm{Perm}(n)$ is the set of all permutations on $[|n|]$ and $C_{i,\sigma(i)} = c(x_i, y_{\sigma_i})$. Note that when the number of points in the target measure is greater than the number of points in the source measure ($n < m$), there may not even exist a transport map between the two measures, regardless of optimality. This is due to the weight vectors of the measures being incompatible.

**Limitation of the discrete Monge problem.** The classical assignment problem (2.2) and its extension to the Monge problem (2.1) are not always adapted to discrete measures in practical problems. One limitation of the former is that the assignment problem can only compare uniform histograms of the same size, and thus does not account for nonuniform weights. Although Monge's push-forward map formulation can generalize to measures with nonuniform weights and different sizes, it may also become infeasible if mass conservation is not satisfied. Moreover, both the assignment problem and Monge's formulation are combinatorial, and the feasible set of the Monge problem is nonconvex, which makes them challenging to solve in their original form.

**Discrete Kantorovich optimal transport.** Kantorovich introduced a key idea in transportation theory, as described in [32], which relaxes the requirement that a source point $x_i$ must be assigned to a unique location $y_{\sigma_i}$ or $T(x_i)$. Instead,

mass at any point $x_i$ can be split and distributed across multiple locations. This flexibility is achieved by using a coupling matrix $P \in \mathbb{R}_+^{n \times m}$ to describe the amount of mass flowing from $x_i$ to $y_j$, rather than a permutation $\sigma$ or a map $T$. Admissible couplings are easier to characterize than Monge maps:

$$\Pi_{a,b} := \left\{ P \in \mathbb{R}_+^{n \times m} \ : \ P\mathbf{1}_m = a, \ P^\top \mathbf{1}_n = b \right\} .$$

Kantorovich's optimal transport problem now reads

$$\mathrm{OT}_c(\mu, \nu) := \min_{P \in \Pi_{a,b}} \langle C, P \rangle = \sum_{i,j} C_{i,j} P_{i,j} \tag{2.3}$$

where $\forall \ i, j \quad C_{i,j} := c(x_i, y_j)$. This is a linear program, and as is usually the case with such programs, its optimal solutions are not necessarily unique. The problem can be algorithmically solved using the network simplex algorithm, in $\mathcal{O}(n+m)nm \log(n+m)$ time [161]. Hence, although it is tractable, discrete optimal transport can be computationally expensive. However, discrete OT plans are sparse, which is a valuable property in matching-based applications such as domain adaptation [57]. This sparsity comes from the fact that there always exists an optimal plan lying on a vertex of $\Pi_{a,b}$: such a plan has at most $n+m$ nonzero entries. Finally, the discrete Kantorovich problem 2.3 can be naturally paired with its dual formulation, which is also a linear program.

**Proposition 2.1.1.** *The discrete Kantorovich problem 2.3 admits a dual and strong duality holds, that is*

$$\mathrm{OT}_c(\mu, \nu) = \max \left\{ \langle f, a \rangle + \langle g, b \rangle : \ f \in \mathbb{R}^n, g \in \mathbb{R}^m \ \textit{and} \quad \forall i, j \ f_i + g_j \leq c(x_i, y_j) \right\}$$

Note also that Kantorovich's relaxed formulation is always symmetric, in the sense that $\mathrm{OT}_c(\mu, \nu) = \mathrm{OT}_c(\nu, \mu)$ whereas the Monge formulation was intrisically asymmetric.

**From Monge to Kantorovich in the discrete setting.** As in the general setting we can relate the discrete Monge and Kantorovich formulations as in the following proposition. For that purpose, let us denote for any permutation $\sigma \in \mathrm{Perm}(n)$, the permutation matrix associated $P_\sigma$ defined as $\forall (i,j) \in \{1, 2, \ldots, n\}, ; (P_\sigma)_{i,j} = 1$ if $j = \sigma(i)$ and 0 otherwise.

**Proposition 2.1.2.** *If $m = n$ and $a = b = \frac{1}{n}\mathbf{1}_n$, then there exists an optimal solution for Problem (2.3) $P_{\sigma^*}$, which is a permutation matrix associated to an optimal permutation $\sigma^* \in Perm(n)$ for Problem (2.1).*

**Statistical Considerations.** In an applied setting, we often assume that we only have access to samples drawn from the distributions of interest. An important statistical problem in optimal transport is to approximate the (usually unknown) optimal transport cost between $\mu \in \mathcal{P}(\mathbb{R}^d)$ and $\nu \in \mathcal{P}(\mathbb{R}^d)$ using only samples $(x_i)_{i=1}^n$ from $\mu$ and $(y_j)_{j=1}^m$ from $\nu$. These samples are assumed to be independently identically distributed from their respective distributions. For optimal transport costs, a straightforward estimator of the unknown distance between the true distributions is to compute it directly between the empirical measures $\hat{\mu} := \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ and $\hat{\nu} := \frac{1}{n}\sum_{j=1}^m \delta_{x_j}$. An important question for statistical considerations is to control the speed of convergence of $\mathrm{OT}_c(\hat{\mu}, \hat{\nu})$ towards $OT_c(\mu, \nu)$, often called the "sample complexity". In the following theorem, we present the sample complexity of the p-Wasserstein distance $W_p$ as defined in 1.3.

**Theorem 2.1.1.** *(Rates for OT Dudley [129]) Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ supported on a bounded domain, then for any $d > 2$ and $1 \leq p < +\infty$*

$$\mathbb{E}(|W_p(\hat{\mu}, \hat{\nu}) - W_p(\mu, \nu)|) = \mathcal{O}(n^{-1/d})$$

This rate is tight in $\mathbb{R}^d$ if one of the two measures has a density with respect to the Lebesgue measure. This rate can be refined when the measures are supported on low-dimensional subdomains: Weed et al. [72] show that, indeed, the rate depends on the intrinsic dimensionality of the support. Therefore the estimation of OT using the simple plug-in estimator suffers from the curse of dimensionality, meaning that if the samples come from probabilities supported on a high dimensional space, the empirical OT is likely to be meaningless.

**Limitations of discrete OT.** In practice, optimal transport has two major limitations that prevent its use in machine learning. Indeed, it is computationally expensive as it requires a supercubic complexity with respect to the number of points which prevents its application to problems larger than few thousands of points. Second, optimal transport suffers from the curse of dimensionality, which means that its performance degrades exponentially as the number of dimensions increases. As a result, although optimal transport has many useful applications, it is not always feasible to use in practice, and alternative approaches may need to be considered.

## 2.2 Entropic Optimal Transport

In this section, we present the most commonly used numerical schemes for approximating solutions to the Kantorovich formulation of optimal transport introduced by Cuturi [76]. This approach involves adding an entropic penalty to the original

problem, which provides several important advantages, making it a highly useful tool. The regularization allows for the minimization of the regularized problem using a simple alternate minimization scheme, which requires only quadratic time and memory complexity. The scheme translates into iterations that involve simple matrix-vector products, making it highly suitable for execution on GPUs. The regularity brought by the entropy also helps to overcome the curse of dimensionality (at least to some extent), thereby enabling good statistical performance. The discrete entropy of a coupling matrix is defined as

$$H(P) := -\sum_{i,j} P_{i,j}(\log(P_{i,j}) - 1) \ .$$

Note that $H$ is 1-strongly concave, and the entropic regularization uses $-\varepsilon H$, where $\varepsilon$ monitors the strengh of entropy added to the problem, as a regularizing function to obtain approximate solutions to the original transport problem.

$$\mathrm{OT}_{c,\varepsilon}(\mu,\nu) := \min_{P\in\Pi_{a,b}} \langle P, C\rangle - \varepsilon H(P) \tag{2.4}$$

Since the objective is an $\varepsilon$-strongly convex function, Problem (2.4) has a unique optimal solution. A direct interpretation of this regularization can be obtained by simply reformulating the objective. Defining the Kullback–Leibler divergence between couplings as

$$\mathrm{KL}(P,K) := -\sum_{i,j} P_{i,j}\left(\log\left(\frac{P_{i,j}}{K_{i,j}}\right) - 1\right) + K_{i,j}$$

the unique solution $P_\varepsilon$ of (2.4) is the projection onto $\Pi_{a,b}$ of the kernel associated to the cost matrix $C$ defined as $K_{i,j} := \exp(-C_{i,j}/\varepsilon)$. Indeed one has that using the definition above

$$P_\varepsilon = \operatorname*{argmin}_{P\in\Pi_{a,b}} \mathrm{KL}(P,K) \ . \tag{2.5}$$

Therefore entropic OT is seeking for the coupling satisfying the marginal constraints that is the closest to the kernel $K$ w.r.t KL divergence. In addition, as for the discrete OT, the entropic OT enjoys a dual formulation.

**Proposition 2.2.1.** *Let $\mu = \sum_{i=1}^n a_i\delta_{x_i}$ and $\nu = \sum_{j=1}^m b_j\delta_{y_j}$ two discrete probability measures. Then one has*

$$\mathrm{OT}_{c,\varepsilon}(\mu,\nu) = \max_{f\in\mathbb{R}^n,\ g\in\mathbb{R}^m} \langle f, a\rangle + \langle g, b\rangle - \varepsilon\langle \exp(f/\varepsilon), K\exp(g/\varepsilon)\rangle \tag{2.6}$$

*where* $\exp(\cdot)$ *is a coordinate-wise operator.*

**Sinkhorn Algorithm.** In order to solve either (2.4) or (2.6), Cuturi [76] propose to use the Sinkhorn algorithm [162]. Using the first order condition for optimality, one obtains that the optimal coupling has to be of the form $P_\varepsilon = \mathrm{diag}(u)K\,\mathrm{diag}(v)$ where $u \in \mathbb{R}_+^n$ and $v \in \mathbb{R}_+^m$. Then the Sinkhorn algorithm, starting at $u^{(0)} = \mathbf{1}_n$ and $v^{(0)} = \mathbf{1}_m$, simply updates the scaling vectors $u$ and $v$ at each iteration as follows:

$$u^{(\ell+1)} = \frac{a}{Kv^{(\ell)}}, \quad \text{and} \quad v^{(\ell+1)} = \frac{b}{K^\top u^{(\ell+1)}}\,. \tag{2.7}$$

In fact this very simple algorithm has two interpretations depending on whether we apply it on the primal or the dual problem. When applied on (2.4) or (2.5), the Sinkhorn algorithm is equivalent to the iterative Bregman projection (IBP) algorithm [163]. Indeed by splitting the constraints into two linear constraint spaces that are

$$C_a := \{P \in \mathbb{R}^{n\times m} \;:\; P\mathbf{1}_m = a\} \quad \text{and} \quad C_b := \{P \in \mathbb{R}^{n\times m} \;:\; P^\top \mathbf{1}_n = b\}$$

the IBP algorithm starts at $P^{(0)} := K$ and consists in performing the following operations at each iteration:

$$P^{(\ell+1)} = \operatorname*{argmin}_{P\in C_a} \mathrm{KL}(P, P^{(\ell)}), \qquad P^{(\ell+2)} = \operatorname*{argmin}_{P\in C_b} \mathrm{KL}(P, P^{(\ell+1)})$$

which is exactly equivalent to perform the operations presented in (2.7). Another interpretation of the Sinkhorn algorithm is that when it is applied on the dual formulation of the entropic OT (2.6), we obtain a simple coordinate gradient ascent scheme on $f$ and $g$. Indeed, starting from $f^{(0)} = 0_n$ and $g^{(0)} = 0_m$, at each iteration the coordinate gradient ascent algorithm performs:

$$\begin{aligned} f^{(\ell+1)} &= \varepsilon \log(a) - \varepsilon \log(K \exp(g^{(\ell)}/\varepsilon)) \quad \text{and} \\ g^{(\ell+1)} &= \varepsilon \log(b) - \varepsilon \log(K^\top \exp(f^{(\ell+1)}/\varepsilon)) \end{aligned}$$

Then by defining $u^{(\ell)} := \exp(f^{(\ell)}/\varepsilon)$ and $v^{(\ell)} := \exp(g^{(\ell)}/\varepsilon)$, we recovers the updates (2.7).

**Computational Complexity.** When applying the Sinkhorn algorithm, one needs to apply at each iteration a matrix/vector product which requires a quadractic complexity with respect to the number of samples that is $\mathcal{O}(nm)$ algebraic operations. Note also that a simple implementation of the Sinkhorn algorithm would also requires a quadratic memory space as one would need to store in memory the kernel $K$. In addition, Franklin and Lorenz [164] proved the that Sinkhorn algorithm enjoys a linear convergence rate to the global minimizer w.r.t the Hilbert metric.

**On the sample complexity of the entropic OT.** When entropy is added to the objective of the optimal transport problem, it allows also to improve the statistical rates of OT. In order to present the result, we need first to introduce a generalized version of the entropic OT which is defined for any probability measures $\mu, \nu \in \mathbb{P}(\mathbb{R}^d)$ as

$$\mathrm{OT}_{c,\varepsilon}(\mu,\nu) := \min_{\gamma \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} c(x,y) d\gamma + \varepsilon \mathrm{KL}(\gamma, \mu \otimes \nu) \tag{2.8}$$

where we have generalized the definition of $KL$ to any coupling as follows:

$$\mathrm{KL}(\gamma,\zeta) := \begin{cases} \int_{\mathbb{R}^d \times \mathbb{R}^d} \log\left(\frac{d\gamma}{d\zeta}\right) d\gamma + \int_{\mathbb{R}^d \times \mathbb{R}^d} d\zeta - \int_{\mathbb{R}^d \times \mathbb{R}^d} d\gamma, & \text{if } \gamma \ll \zeta \\ +\infty, & \text{otherwise} \end{cases}$$

This is the exact generalization of the discrete case presented earlier in (2.4), as in the discrete setting, one obtains that $\mathrm{KL}(\gamma, \mu \otimes \nu) = \sum_{i,j} \gamma_{i,j}(\log(\gamma_{i,j}) - 1) + \sum_{i,j} a_i b_j + H(a) + H(b) = H(\gamma) + c$ where $c$ is a constant independent of $\gamma$. Therefore the discrete entropic OT introduces in (2.4) and its generalized version (2.8) evaluated on the same discrete measures admits the exact same solution. Let us now consider two distributions $\mu \in \mathcal{P}(\mathbb{R}^d)$ and $\nu \in \mathcal{P}(\mathbb{R}^d)$ and let us denote $\hat{\mu}$ and $\hat{\nu}$ their associated empirical versions of size $n$, then it has been shown in [51, 82] that entropic OT enjoys a parametric rates with respect to the number of samples.

**Theorem 2.2.1.** *Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ supported on a bounded set and let us assume that the cost $c$ is $C^\infty$ and $L$-lipschitz on that domain. Then one has*

$$\mathbb{E}(|\,\mathrm{OT}_{c,\varepsilon}(\hat{\mu},\hat{\nu}) - \mathrm{OT}_{c,\varepsilon}(\mu,\nu)|) = \mathcal{O}\left(\frac{\varepsilon^{-d/2}}{\sqrt{n}}\right)$$

Therefore, when $\varepsilon$ is sufficiently large, then the plug-in estimator $OT_{c,\varepsilon}(\hat{\mu},\hat{\nu})$ enjoys a fast rate of convergence towards the true quantity $OT_{c,\varepsilon}(\mu,\nu)$, however when $\varepsilon$ goes to 0, entropic OT still suffers from the curse of dimensionality as the rate still has an exponential dependency in the dimension w.r.t the entropic strengh $\varepsilon$.

**Debiased Sinkhorn divergence.** When it comes to compare probabilities measures, entropic OT might not be adapted especially when $\varepsilon$ is not sufficiently small: it is no longer a distance as it does not satisfy the triangle inequality, nor a divergence as it is not positive, nor even able to separate distributions as in general $OT_{c,\varepsilon}(\mu,\mu) \neq 0$. To alleviate these issues, Genevay et al. [78] proposed to subtract debiasing terms from $OT_{c,\varepsilon}$, defining the Sinkhorn divergence:

$$S_{c,\varepsilon}(\mu,\nu) := \mathrm{OT}_{c,\varepsilon}(\mu,\nu) - \frac{1}{2}(\mathrm{OT}_{c,\varepsilon}(\mu,\mu) + \mathrm{OT}_{c,\varepsilon}(\nu,\nu))\,. \tag{2.9}$$

Feydy et al. [130] then proved that the Sinkhorn divergence defines a suitable divergence able to interpolate between the Maximum Mean Discrepency (MMD) [18] and OT.

**Proposition 2.2.2.** *Let $c(x,y) = |x-y|^p$, $p > 1$. Then for all compactly supported measures $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$, $S_\varepsilon(\mu,\nu)$ defines a symmetric positive definite divergence which is convex in $\mu$ or $\nu$, and metrizes weak convergence. In addition we have that*

$$S_{c,\varepsilon}(\mu,\nu) \xrightarrow[\varepsilon\to+\infty]{} \frac{1}{2}\,\mathrm{MMD}_{-c}(\mu,\nu)$$
$$S_{c,\varepsilon}(\mu,\nu) \xrightarrow[\varepsilon\to 0^+]{} \mathrm{OT}_c(\mu,\nu)$$

*where*

$$\mathrm{MMD}_k(\mu,\nu) = \int_{\mathbb{R}^d\times\mathbb{R}^d} \left(k(x,x') + k(y,y') - 2k(x,y)\right) d\mu(x)d\mu(x')d\nu(y)d\nu(y')$$

**Entropic OT cannot be applied in the large-scale setting...** Entropic optimal transport is a regularized form of OT that helps to overcome many practical issues associated with the original Kantorovich formulation of OT when working with data and it is backed with a lot of theoretical properties which facilitate its use. In particular, it offers improved computational efficiency and better statistical rates. However, despite these benefits, the quadratic complexity of its numerical algorithm still presents a challenge in the application of OT to large-scale datasets with hundreds of thousands of points. Therefore, there is a growing need to develop scalable OT algorithms that can handle such massive datasets.

*In part II, we will exploit similar ideas and propose new regularization schemes of the optimal transport problem based on low-rank constraints in order to make it applicable in the large-scale setting.*

# Chapter 3

# Gromov-Wasserstein: Quadratic Optimal Transport

The ability to align points across two related yet incomparable point clouds (e.g. living in different spaces) plays an important role in machine learning. This situation arises typically when realigning two distinct views (or features) from points sampled from similar sources. Despite this heterogeneity, one expects to find a mapping registering points from the first to the second set, since they contain similar overall information. That realignment is usually carried out using the Gromov-Wasserstein (GW) machinery proposed by Memoli [165], Mémoli [86]. GW seeks a relaxed assignment matrix that is as close to an isometry as possible, as quantified by a quadratic score. This chapter presents the main definitions and properties of the GW problem and emphisizes on the computational challenges behind this problem.

## 3.1 Introduction to Gromov-Wasserstein

Applying OT can be challenging when dealing with probability measures whose supports lie in incomparable spaces, e.g. when the supports of the measures are not part of a common ground metric space. Defining a meaningful cost function to compare the two measures can be difficult in such cases, and the Wasserstein distance, which offers a natural geometry on the set of distributions supported on the same metric space, may not be applicable. This is particularly true when the dimensions of the two spaces are different, as the distance between a point in one space and a point in the other cannot be defined. Additionally, OT is not invariant to certain transformations, such as rotations or translations, making it less useful when it comes to compare shapes. The Gromov-Wasserstein (GW) framework offers a solution to these issues by using a quadratic optimal transport problem instead

of a linear one, and quantifying the metric distortion when transporting points between spaces. This section introduces the GW problem, its metric properties, and standard numerical solvers. For further details, we recommend referring to [166, 86, 167].

**Gromov-Wasserstein Problem.** Let $(X, d_X)$, $(Y, d_Y)$ be two Polish spaces, $c_X : X \times X \to \mathbb{R}$ and $c_Y : Y \times Y \to \mathbb{R}$ two continuous measurable functions, and let us consider $\mu \in \mathbb{P}(X)$, $\nu \in \mathbb{P}(Y)$ two probability measures on respectively $X, Y$. The Gromov-Wasserstein problem aims at finding:

$$\mathrm{GW}_p((\mu, c_X), (\nu, c_Y)) = \inf_{\gamma \in \Pi(\mu,\nu)} \left( \int_{X\times Y} \int_{X\times Y} |c_X(x, x') - c_Y(y, y')|^p \, d\gamma d\gamma \right)^{\frac{1}{p}}, \tag{3.1}$$

for $p \geq 1$. The choice of costs $c_X$ and $c_Y$ between points in spaces $X$ and $Y$ is a crucial component in the Gromov-Wasserstein problem. One common approach is to use intrinsic metrics $d_X$ and $d_Y$ to define a metric between metric measure spaces, represented as triplets $(X, d_X, \mu)$, which has been studied extensively in prior works like [86, 166]. The objective of the GW problem is to find an optimal coupling $\gamma$ that maps points from $X$ to $Y$ such that pairs of points $(x, x')$ are "similar" in $X$ with respect to $c_X$ as pairs of points $(y, y')$ are in $Y$ with respect to $c_Y$. When $c_X$ and $c_Y$ are distances, this implies that points $x$ and $x'$ are as close in $X$ as $y$ and $y'$ are in $Y$.

**Properties of GW.** As for the Kantorovich formulation of OT, the equation (3.1) always admits a solution under some regularity assumptions on the costs. This result can be seen as a corollary of the one presented in [111] for OT.

**Theorem 3.1.1.** *Let $X$ and $Y$ be two Polish spaces, $c_X$ and $c_Y$ two continuous cost functions and $p \geq 1$. Assuming that*

$$\int_{X\times X} c_X(x, x')^p d\mu(x) \otimes \mu(x') < +\infty \quad \text{and} \quad \int_{Y\times Y} c_Y(y, y')^p d\nu(y) \otimes \nu(y') < +\infty$$

*then the Gromov-Wasserstein problem introduced in (3.1) is finite and admits a minimizer.*

A key feature of the Gromov-Wasserstein problem is its ability to compare probability measures that have supports in distinct and potentially unrelated spaces. This is achieved by evaluating the entire metric measure spaces associated with each probability measure.

**Definition 1** (Metric measure space)**.** *A metric measure space is a triplet* $(X, d_X, \mu)$ *where*

- $(X, d_X)$ *is a Polish space*
- $\mu$ *is a Borel probability measure on* $X$*.*

In order to present the metric properties of GW on the set of metric measure spaces, we need first to introduce a notion of equivalence of two metric measure spaces. This can be done thanks to the notion of isomorphism.

**Definition 2** (Isomorphism)**.** *Let* $(X, d_X)$ *and* $(Y, d_Y)$ *be Polish spaces and* $\mu \in \mathcal{P}(X)$*,* $\nu \in \mathcal{P}(Y)$*. We say that* $(X, d_X, \mu)$ *is isomorphic to* $(Y, d_Y, \nu)$ *if there exists a bijection* $\varphi : \text{supp}(\mu) \to \text{supp}(\nu)$ *such that:*

1. $\varphi$ *is an isometry, i.e.* $d_Y(\varphi(x), \varphi(x')) = d_X(x, x')$ *for all* $x, x' \in \text{supp}(\mu)$*.*
2. $\varphi$ *pushes* $\mu$ *forward to* $\nu$*, i.e.* $\varphi \# \mu = \nu$*.*

The following theorem is fundamental for GW and aims to unify the metric properties of GW given in [86, 166]. It proves that GW defines a metric w.r.t. the isomorphism notion.

**Theorem 3.1.2** (Mémoli [86])**.** *Let* $(X, d_X)$*,* $(Y, d_Y)$ *be Polish spaces and* $p \geq 1$*. Then, the* $\text{GW}_p$ *is symmetric, positive and satisfies the triangle inequality. More precisely, given* $(X, d_X, \mu)$*,* $(Y, d_Y, \nu)$ *and* $(Z, d_Z, m)$ $\text{GW}_p$ *satisfies:*

$$\text{GW}_p((\mu, d_X), (\nu, d_Y)) \leq \text{GW}_p((\mu, d_X), (m, d_Z)) + \text{GW}_p((m, d_Z), (\nu, d_Y)).$$

*Moreover, the* $\text{GW}_p$ *characterizes isomorphisms:* $\text{GW}_p(\mu, \nu) = 0$ *if and only if* $(X, d_X, \mu)$ *and* $(Y, d_Y, \nu)$ *are isomorphic.*

This theorem allows to endow the set of all the metric measure spaces of the form $(X, c_X, \mu)$ with a distance defined by GW, which, however, requires the finiteness of GW. More precisely we define $\mathbb{X}_p$ to be the space of all metric measure spaces with finite $L^p$ cost, i.e., $\mathbb{X}_p := \{(X, d_X, \mu) \mid \int_{X \times X} d_X(x, x')^p d\mu(x) < +\infty\}$ where $(X, d_X)$ is a Polish space and $\mu \in \mathbb{P}(X)$.

**Theorem 3.1.3** (Mémoli [86])**.** $\text{GW}_p$ *is a distance on* $\mathbb{X}_p$ *quotiented by isomorphisms.*

The implications of the above theorem are numerous. First, it endows the space of all metric measure spaces with a topology and geometric structure, induced by Gromov-Wasserstein. Second, it suggests that Gromov-Wasserstein is well-suited for comparing objects with respect to a large class of invariants, such as rotations,

translations, or permutations. This property is particularly important for shape comparison, where the orientation of a shape does not define its nature. Finally, if the Gromov-Wasserstein distance vanishes, it necessarily implies that the objects are isomorphic, which is valuable for detecting such cases. Gromov-Wasserstein is also deeply connected to the Gromov-Hausdorff distance, which measures how far $(X, d_X)$ and $(Y, d_Y)$ are from being isometric and can be used for studying the convergence of metric spaces. However, computing the Gromov-Hausdorff distance results in a highly non-convex optimization problem whose global solution is not tractable. As shown in [86], the introduction of the measures allows to relax the definition of the Gromov-Hausdorff distance and leads to the Gromov-Wasserstein distance.

## 3.2 Computational Aspects of Gromov-Wasserstein

In this section we focus on the computational aspects of the GW problem. In the following $\mu = \sum_{i=1}^{n} a_i \delta_{x_i} \in P(X)$, $\nu = \sum_{j=1}^{m} b_j \delta_{y_j} \in P(Y)$ are discrete probability measures over respectively $(X, d_X)$, $(Y, d_Y)$. We also note $A$, $B$ the matrices of pair-to-pair distances inside each space, i.e. $\forall (i,k) \in [n]^2, A_{i.i'} = d_X(x_i, x_{i'})$ and $\forall (j,j) \in [m]^2, B_{j,j'} = d_Y(y_j, y_{j'})$. The discrete GW problem aims at solving:

$$\mathrm{GW}_p^p((a,A),(b,B)) = \min_{P \in \Pi_{a,b}} \sum_{i,i',j,j'} |A_{i,i'} - B_{j,j'}|^p P_{i,j} P_{i,j} \tag{3.2}$$

The optimization problem 3.2 is a non-convex Quadratic Program (QP), NP-hard in general [83] and can be notoriously hard to approximate. When $p = 2$, equation 3.2 can be recast as

$$\mathrm{GW}_2^2((a,A),(b,B)) = \min_{P \in \Pi_{a,b}} \langle A^{\odot 2} a, a \rangle + \langle B^{\odot 2} b, b \rangle - 2\langle APB, P \rangle \tag{3.3}$$

where $\odot$ is the Hadamard (elementwise) product or power. Even in that case, GW is in general non-convex and NP-hard.

**Relationship with the Quadratic Assignment Problem.** The GW problem is in fact closely related to the so-called Quadratic Assignment Problem (QAP). This problem was first introduced by Koopmans and Beckmann [168] to model a plant location problem and plays many roles in optimization today. Given two matrices $A = (A_{i,j})_{i,j \in [n]}$ and $B = (B_{i,j})_{i,j \in [n]}$, the standard form for the QAP reads:

$$\max_{\sigma \in \mathrm{Perm}(n)} \sum_{i=1}^{n} \sum_{j=1}^{n} A_{\sigma(i),\sigma(j)} B_{i,j}. \tag{3.4}$$

Unfortunately the QAP is NP-Hard in general and only few special cases are known to be computable in polynomial time. The QAP can be seen as the Monge formulation of the GW problem when $p = 2$ and the measures are both uniform on their respective supports and have the number of points $n$. Indeed in that case using equation 3.3, and allowing only couplings that come from a transport map, we obtain that the Monge formulation of the GW (MGW) problem is defined as:

$$\mathrm{MGW}_2^2(\mu, \nu) = \min_{\sigma \in \mathrm{Perm}(n)} \langle A^{\odot 2} a, a \rangle + \langle B^{\odot 2} b, b \rangle - 2 \sum_{i=1}^{n} \sum_{j=1}^{n} A_{\sigma(i),\sigma(j)} B_{i,j}; \tag{3.5}$$

which admits the exact same solution(s) as the QAP defined in (3.4).

**Entropic Regularization.** The original GW problem (3.1) can be regularized using entropy [84, 85], leading to problem:

$$\mathrm{GW}_{p,\varepsilon}^p((a, A), (b, B)) = \min_{P \in \Pi_{a,b}} \sum_{i,i',j,j'} |A_{i,i'} - B_{j,j'}|^p P_{i,j} P_{i,j} - \varepsilon H(P)\,, \tag{3.6}$$

Peyré et al. [47] propose to solve the entropic GW problem using a mirror descent (MD) scheme w.r.t. the KL divergence. Their algorithm boils down to solving a sequence of regularized OT problems. When $p = 2$ the algorithm can be even more simplified: in that case, starting at $P^{(0)} = ab^T$, the proposed algorithm solves at each iteration the following problem:

$$P^{(\ell+1)} = \operatorname*{argmin}_{P \in \Pi_{a,b}} \mathrm{KL}(P, K_\varepsilon^{(\ell)}) \text{ where } K_\varepsilon^{(\ell)} := \exp(4AP^{(\ell)}B/\varepsilon)$$

which is an entropic OT problem and can be solved efficiently using the Sinkhorn algorithm [76]. The algorithm proposed by the authors recovers therefore as a special case the "softassign quadratic assignment" algorithm introduced in prior work [169, 170]. However, the convergence proof provided in [169] is limited to the convergence of functional values, rather than the convergence of the iterates. Additionally, the proof only applies to convex functions being minimized in the optimization problem defined in (3.6), which is not the case for all matrices $(A, B)$. From a computational point of view, this algorithm requires in general (for general $p$) $\mathcal{O}(n^2m^2)$ algebraic operations per iteration and $\mathcal{O}(n^2)$ memory space while in the case of $p = 2$, the time complexity can be reduced to $\mathcal{O}(nm(n + m))$ per iteration.

**Gromov-Wassertein remains too costly.** The original GW problem is a non-convex and NP-hard problem and has to rely on some approximations in order to be used in practice. Entropic GW is one of the most successful attempt towards this goal which allows to obtain an approximation of the GW cost as well as the

optimal coupling solving it in cubic time. However, despite its good performance in practice and its improved complexity, the entropic approach can only be applied for small problems of order of one thousand points. Therefore, there is a need to develop scalable GW algorithms that can handle larger datasets (and hopefully massive ones).

*In part II, we will exploit similar ideas and propose new regularization schemes of the Gromov-Wasserstein problem based on low-rank constraints in order to reach a linear complexity both in time and memory.*

# Part II

# Low-rank Optimal Transport

In this part, we propose new regularization schemes of the OT problem and its quadratic variant, namely the Gromov-Wasserstein problem, by considering low-rank factorization of both the underlying cost and the coupling solving the OT problem itself. These new computational schemes pave the way for the use of OT in the large-scale setting. This part is divided in four contributions.

- In a first contribution, we propose to approximate the iterations of the Sinkhorn algorithm solving the entropic OT problem by using ground costs of the form $c(x, y) = -\log\langle\varphi(x), \varphi(y)\rangle$ where $\varphi$ is a map from the ground space onto the positive orthant $\mathbb{R}^r_+$, with $r \ll n$ where $n$ is the number of samples. By doing so, we obtain a low nonnegative rank approximation of the optimal coupling solving the entropic OT problem and ensures that the cost of Sinkhorn iterations scales linearly w.r.t the number of samples. We show that usual cost functions can be approximated using this form and propose explicit embedding $\varphi$ for each of them. The positivity of the feature embedding $\varphi$ is essential here as it guarantees the convergence of the Sinkhorn algorithm. Additionally, we take advantage of the fact that our approach yields approximation that remain fully differentiable with respect to input distributions, as opposed to previously proposed adaptive low-rank approximations, to train a faster variant of OT-GAN.

- In a second contribution, instead of approximating the optimal coupling of the entropic OT problem using a low nonnegative rank approximation, we introduce a new regularization scheme of the OT problem by constraining directly the nonnegative rank of the couplings. We then propose a generic approach that aims at solving, in full generality, the OT problem under low-nonnegative rank constraints with arbitrary costs. Our algorithm relies on an explicit factorization of low-rank couplings as a product of *sub-coupling* factors linked by a common marginal; similar to an NMF approach, we alternatively updates these factors. Our algorithm enjoys a linear complexity as soon as one has access to a low rank approximation of the cost matrix, which is always the case for distance matrices.

- In a third contribution, we focus on the theoretical properties of the low-rank optimal transport (LOT) approach advocated in our previous contribution. LOT restricts the search for low-cost couplings to those that have a low-nonnegative rank, yielding linear time algorithms in cases of interest. However, these promises can only be fulfilled if the LOT approach is seen as a legitimate contender to entropic regularization when compared on properties of interest, where the scorecard typically includes theoretical properties (statistical complexity and relation to other methods) or practical aspects (debiasing, hyperparameter tuning, initialization). We target each of these areas in

this contribution in order to cement the impact of low-rank approaches in computational OT.

- To conclude this part, we also present a fourth contribution where we show that the low-rank approach can be also particularly beneficial for the Gromov-Wasserstein (GW) problem. The Gromov-Wasserstein (GW) framework provides an increasingly popular answer to the alignment problem of points across two related yet incomparable point clouds (e.g. living in different spaces), by seeking a low-distortion, geometry-preserving assignment between these points. As a non-convex, quadratic generalization of OT, the GW problem is NP-hard. While practitioners often resort to solving GW approximately as a nested sequence of entropy-regularized OT problems, the cubic complexity (($\mathcal{O}(n^3)$ where $n$ in the number of samples) of that approach is a roadblock. We show in this work how our recent variant of the OT problem that restricts the set of admissible couplings to those having a low-rank factorization is remarkably well suited to the resolution of GW: when applied to GW, we show that this approach is not only able to compute a stationary point of the GW problem in time $O(n^2)$, but also uniquely positioned to benefit from the knowledge that the initial cost matrices are low-rank, to yield a linear time $O(n)$ GW approximation. Our approach yields similar results, yet orders of magnitude faster computation than the SoTA entropic GW approaches, on both simulated and real data.

# Chapter 4

# Linear Time Sinkhorn Divergences using Positive Features

Although Sinkhorn divergences are now routinely used in data sciences to compare probability distributions, the computational effort required to compute them remains expensive, growing in general quadratically in the size $n$ of the support of these distributions. Indeed, solving optimal transport (OT) with an entropic regularization requires computing a $n \times n$ kernel matrix (the neg-exponential of a $n \times n$ pairwise ground cost matrix) that is repeatedly applied to a vector. We propose to use instead ground costs of the form $c(x, y) = -\log\langle\varphi(x), \varphi(y)\rangle$ where $\varphi$ is a map from the ground space onto the positive orthant $\mathbb{R}^r_+$, with $r \ll n$. This choice yields, equivalently, a kernel $k(x, y) = \langle\varphi(x), \varphi(y)\rangle$, and ensures that the cost of Sinkhorn iterations scales as $O(nr)$. We show that usual cost functions can be approximated using this form. Additionally, we take advantage of the fact that our approach yields approximation that remain fully differentiable with respect to input distributions, as opposed to previously proposed adaptive low-rank approximations of the kernel matrix, to train a faster variant of OT-GAN [50].

This chapter is based on [6].

## 4.1 Introduction

Optimal transport (OT) theory [171] plays an increasingly important role in machine learning to compare probability distributions, notably point clouds, discrete measures or histograms [120]. As a result, OT is now often used in graphics [46, 47, 48], neuroimaging [44], to align word embeddings [52, 53, 54], reconstruct cell trajectories [55, 41, 56], domain adaptation [57, 58] or estimation of generative models [49, 50, 51]. Yet, in their original form, as proposed by Kantorovich [32], OT distances are not a natural fit for applied problems: they minimize a network flow problem, with a supercubic complexity ($n^3 \log n$) [172] that results in an output that is *not* differentiable with respect to the measures' locations or weights [173, §5]; they suffer from the curse of dimensionality [129, 71] and are therefore likely to be meaningless when used on samples from high-dimensional densities.

Because of these statistical and computational hurdles, all of the works quoted above do rely on some form of regularization to smooth the OT problem, and some more specific uses of an entropic penalty, to recover so called Sinkhorn divergences [112]. These divergences are cheaper to compute than regular OT [79, 80], smooth and programmatically differentiable in their inputs [46, 55], and have a better sample complexity [81] while still defining convex and definite pseudometrics [113]. While Sinkhorn divergences do lower OT costs from supercubic down to an embarassingly parallel quadratic cost, using them to compare measures that have more than a few tens of thousands of points in forward mode (less obviously if backward execution is also needed) remains a challenge.

**Entropic regularization: starting from ground costs.** The definition of Sinkhorn divergences usually starts from that of the ground cost on observations. That cost is often chosen by default to be a $q$-norm between vectors, or a shortest-path distance on a graph when considering geometric domains [114, 115, 116, 44]. Given two measures supported respectively on $n$ and $m$ points, regularized OT instantiates first a $n \times m$ pairwise matrix of costs $C$, to solve a linear program penalized by the coupling's entropy. This can be rewritten as a Kullback-Leibler minimization:

$$\min_{\text{couplings } P} \langle C, P \rangle - \varepsilon H(P) = \varepsilon \min_{\text{couplings } P} \mathrm{KL}(P, K)\,, \tag{4.1}$$

where matrix $K$ appearing in Eq. (4.1) is defined as $K := \exp(-C/\varepsilon)$, the elementiwe neg-exponential of a rescaled cost $C$.As described in more detail in §4.2, this problem can then be solved using Sinkhorn's algorithm, which only requires applying repeatedly kernel $K$ to vectors. While faster optimization schemes to compute regularized OT have been been investigated [117, 118, 119], the Sinkhorn

algorithm remains, because of its robustness and simplicity of its parallelism, the workhorse of choice to solve entropic OT. Since Sinkhorn's algorithm cost is driven by the cost of applying $K$ to a vector, speeding up that evaluation is the most impactful way to speedup Sinkhorn's algorithm. This is the case when using separable costs on grids (applying $K$ boils down to carrying out a convolution at cost $(n^{1+1/d})$ [120, Remark 4.17]) or when using shortest path metrics on graph in which case applying $K$ can be approximated using a heat-kernel [121]. While it is tempting to use low-rank matrix factorization, using them within Sinkhorn iterations requires that the application of the approximated kernel guarantees the positiveness of the output. As shown by [122] this can only be guaranteed, when using the Nyström method, when regularization is high and tolerance very low.

**Starting instead from the Kernel.** Because regularized OT can be carried out using only the definition of a kernel $K$, we focus instead on kernels $K$ that are guaranteed to have positive entries by design. Indeed, rather than choosing a cost to define a kernel next, we consider instead ground costs of the form $c(x, y) = -\varepsilon \log\langle\varphi(x), \varphi(y)\rangle$ where $\varphi$ is a map from the ground space onto the positive orthant in $\mathbb{R}^r$. This choice ensures that both the Sinkhorn algorithm itself (which can approximate optimal primal and dual variables for the OT problem) and the evaluation of Sinkhorn divergences can be computed exactly with an effort scaling linearly in $r$ and in the number of points, opening new perspectives to apply OT at scale.

**Our contributions** are two fold: *(i)* We show that kernels built from positive features can be used to approximate some usual cost functions including the square Euclidean distance using random expansions. *(ii)* We illustrate the versatility of our approach by extending previously proposed OT-GAN approaches [50, 81], that focused on learning adversarially cost functions $c_\theta$ and incurred therefore a quadratic cost, to a new approach that learns instead adversarially a kernel $k_\theta$ induced from a positive feature map $\varphi_\theta$. We leverage here the fact that our approach is fully differentiable in the feature map to train a GAN at scale, with linear time iterations.

## 4.2 Regularized Optimal Transport

**Sinkhorn Divergence.** Let $\mu = \sum_{i=1}^n a_i \delta_{x_i}$ and $\nu = \sum_{j=1}^m b_j \delta_{y_j}$ be two discrete probability measures. The Sinkhorn divergence [174, 175, 50] between $\mu$ and $\nu$ is,

given a constant $\varepsilon > 0$, equal to

$$S_{c,\varepsilon}(\mu,\nu) := \mathrm{OT}_{c,\varepsilon}(\mu,\nu) - \frac{1}{2}\left(\mathrm{OT}_{c,\varepsilon}(\mu,\mu) + \mathrm{OT}_{c,\varepsilon}(\nu,\nu)\right), \text{ where} \tag{4.2}$$

$$\mathrm{OT}_{c,\varepsilon}(\mu,\nu) := \min_{\substack{P\in\mathbb{R}_+^{n\times m} \\ P\mathbf{1}_m = a, P^T\mathbf{1}_n = b}} \langle P, C\rangle - \varepsilon H(P). \tag{4.3}$$

Here $C := [c(x_i, y_j)]_{ij}$ and $H$ is the Shannon entropy, $H(P) := -\sum_{ij} P_{ij}(\log P_{ij} - 1)$. Because computing and differentiating $S_{c,\varepsilon}$ is equivalent to doing so for three evaluations of $\mathrm{OT}_{c,\varepsilon}$ (neglecting the third term in the case where only $\mu$ is a variable) [120, §4], we focus on $\mathrm{OT}_{c,\varepsilon}$ in what follows.

**Primal Formulation.** Problem (4.3) is $\varepsilon$-strongly convex and admits therefore a unique solution $P_\varepsilon$ which, writing first order conditions for problem (4.3), admits the following factorization:

$$\exists u^\star \in \mathbb{R}_+^n, v^\star \in \mathbb{R}_+^m \text{ s.t. } P_\varepsilon = \mathrm{diag}(u^\star) K \mathrm{diag}(v^\star), \text{ where } K := \exp(-C/\varepsilon). \tag{4.4}$$

These *scalings* $u^\star, v^\star$ can be computed using Sinkhorn's algorithm, which consists in initializing $u$ to any arbitrary positive vector in $\mathbb{R}^m$, to apply then fixed point iteration described in Alg. 3.

These two iterations require together $2nm$ operations if $K$ is stored as a matrix and applied directly. The number of Sinkhorn iterations needed to converge to a precision $\delta$ (monitored by the difference between the column-sum of $\mathrm{diag}(u)K\mathrm{diag}(v)$ and $b$) is controlled by the scale of elements in $C$ relative to $\varepsilon$ [164]. That convergence deteriorates with smaller $\varepsilon$, as studied in more detail by [72, 176].

**Algorithm 1** Sinkhorn

**Inputs:** $K, a, b, \delta, u$ **repeat**
  $v \leftarrow b/K^T u,\ u \leftarrow a/Kv$
**until** $\|v \circ K^T u - b\|_1 < \delta$;
**Result:** $u, v$

**Dual Formulation.** The dual of (4.3) plays an important role in our analysis [120, §4.4]:

$$\mathrm{OT}_{c,\varepsilon}(\mu,\nu) = \max_{\alpha\in\mathbb{R}^n, \beta\in\mathbb{R}^m} a^T\alpha + b^T\beta - \varepsilon(e^{\alpha/\varepsilon})^T K e^{\beta/\varepsilon} = \varepsilon\left(a^T \log u_\varepsilon + b^T \log v_\varepsilon - 1\right) \tag{4.5}$$

where we have introduced, next to its definition, its evaluation using optimal scalings $u^\star$ and $v^\star$ described above. This equality comes from that fact that *(i)* one can show that $\alpha^\star := \varepsilon \log u^\star,\ \beta^\star := \varepsilon \log v^\star$, *(ii)* the term $(e^{\alpha/\varepsilon})^T K e^{\beta/\varepsilon} = u^T K v$ is equal to 1, whenever the Sinkhorn loop has been applied even just once, since these

sums describe the sum of a coupling (a probability distribution of size $n \times m$). As a result, given the outputs $u, v$ of Alg. 3 we estimate (4.3) using

$$\widehat{\mathrm{OT}}_{c,\varepsilon}(\mu,\nu) = \varepsilon \left( a^T \log u + b^T \log v - 1 \right). \tag{4.6}$$

Approximating $\mathrm{OT}_{c,\varepsilon}(\mu,\nu)$ can be therefore carried using exclusively calls to the Sinkhorn algorithm, which requires instantiating kernel $K$, in addition to computing inner product between vectors, which can be computed in $\mathcal{O}(n+m)$ algebraic operations; the instantiation of $C$ is never needed, as long as $K$ is given. Using this dual formulation(4.3) we can now focus on kernels that can be evaluated with a linear cost to achieve linear time Sinkhorn divergences.

## 4.3 Linear Sinkhorn with Positive Features

The usual flow in transport dictates to choose a cost first $c(x,y)$ to define a kernel $k(x,y) := \exp(-c(x,y)/\varepsilon)$ next, and adjust the temperature $\varepsilon$ depending on the level of regularization that is adequate for the task. We propose in this work to do exactly the opposite, by choosing instead parameterized feature maps $\varphi_\theta : \mathcal{X} \mapsto (\mathbb{R}_+^*)^r$ which associate to any point in $\mathcal{X}$ a vector in the positive orthant. With such maps, we can therefore build the corresponding positive-definite kernel $k_\theta$ as $k_\theta(x,y) := \varphi_\theta(x)^T \varphi_\theta(y)$ which is a positive function. Therefore as a by-product and by positivity of the feature map, we can define for all $(x,y) \in \mathcal{X} \times \mathcal{X}$ the following cost function

$$c_\theta(x,y) := -\varepsilon \log \varphi_\theta(x)^T \varphi_\theta(y). \tag{4.7}$$

**Remark 1** (Transport on the Positive Sphere.)**.** *Defining a cost as the log of a dot-product as described in (4.7) has already played a role in the recent OT literature. In [177], the author defines a cost $c$ on the sphere $\mathbb{S}^d$, as $c(x,y) = -\log x^T y$, if $x^T y > 0$, and $\infty$ otherwise. The cost is therefore finite whenever two normal vectors share the same halfspace, and infinite otherwise. When restricted to the the positive sphere, the kernel associated to this cost is the linear kernel. See App. 4.7 for an illustration.*

More generally, the above procedure allows us to build cost functions on any cartesian product spaces $\mathcal{X} \times \mathcal{Y}$ by defining $c_{\theta,\gamma}(x,y) := -\varepsilon \log \varphi_\theta(x)^T \psi_\gamma(y)$ where $\psi_\gamma : \mathcal{Y} \mapsto (\mathbb{R}_+^*)^r$ is a parametrized function which associates to any point $\mathcal{Y}$ also a vector in the same positive orthant as the image space of $\varphi_\theta$ but this is out of the scope of this paper.

### 4.3.1 Achieving linear time Sinkhorn iterations with Positive Features.

Choosing a cost function $c_\theta$ as in (4.7) greatly simplifies computations, by design, since one has, writing for the matrices of features for two set of points $x_1, \dots, x_n$ and $y_1, \dots, y_m$

$$\xi := \left[\varphi_\theta(x_1), \dots, \varphi_\theta(x_n)\right] \in (\mathbb{R}_+^*)^{r\times n}, \quad \zeta := \left[\varphi_\theta(y_1), \dots, \varphi_\theta(y_m)\right] \in (\mathbb{R}_+^*)^{r\times m},$$

that the resulting sample kernel matrix $K_\theta$ corresponding to the cost $c_\theta$ is $K_\theta = \left[e^{-c_\theta(x_i,y_j)/\varepsilon}\right]_{i,j} = \xi^T\zeta$. Moreover thanks to the positivity of the entries of the kernel matrix $K_\theta$ there is no duality gap and we obtain that

$$\mathrm{OT}_{c_\theta,\varepsilon}(\mu,\nu) = \max_{\alpha\in\mathbb{R}^n,\beta\in\mathbb{R}^m} a^T\alpha + b^T\beta - \varepsilon(\xi e^{\alpha/\varepsilon})^T\zeta e^{\beta/\varepsilon}. \tag{4.8}$$

Therefore the Sinkhorn iterations in Alg. 3 can be carried out in exactly $r(n+m)$ operations. The main question remains on how to choose the mapping $\varphi_\theta$. In the following, we show that, for some well chosen mappings $\varphi_\theta$, we can approximate the entropic OT for some classical costs in linear time.

### 4.3.2 Approximation properties of Positive Features.

Let $\mathcal{U}$ be a metric space and $\rho$ a probability measure on $\mathcal{U}$. We consider kernels on $\mathcal{X}$ of the form:

$$\text{for } (x,y)\in\mathcal{X}^2,\ k(x,y) = \int_{u\in\mathcal{U}} \varphi(x,u)^T\varphi(y,u)d\rho(u). \tag{4.9}$$

Here $\varphi : \mathcal{X}\times\mathcal{U} \to (\mathbb{R}_+^*)^p$ is such that for all $x\in\mathcal{X}$, $u\in\mathcal{U} \to \|\varphi(x,u)\|_2$ is square integrable (for the measure $d\rho$). Given such kernel and a regularization $\varepsilon$ we define the cost function $c(x,y) := -\varepsilon\log(k(x,y))$. In fact, we will see in the following that for some usual cost functions $\tilde{c}$, e.g. the square Euclidean cost, the Gibbs kernel associated $\tilde{k}(x,y) = \exp(-\varepsilon^{-1}\tilde{c}(x,y))$ admits a decomposition of the form Eq.(4.9). To obtain a finite-dimensional representation, one can approximate the integral with a weighted finite sum. Let $r\geq 1$ and $\theta := (u_1, ..., u_r) \in \mathcal{U}^r$ from which we define the following positive feature map

$$\varphi_\theta(x) := \frac{1}{\sqrt{r}}\left(\varphi(x,u_1), ..., \varphi(x,u_r)\right) \in \mathbb{R}^{p\times r}$$

and a new kernel as $k_\theta(x,y) := \langle\varphi_\theta(x), \varphi_\theta(y)\rangle$. When the $(u_i)_{1\leq i\leq r}$ are sampled independently from $\rho$, $k_\theta$ may approximates the kernel $k$ arbitrary well if the number of random features $r$ is sufficiently large. For that purpose let us now introduce some assumptions on the kernel $k$.

**Assumption 1.** *There exists a constant $\psi > 0$ such that for all $x, y \in \mathcal{X}$:*

$$|\varphi(x,u)^T \varphi(y,u)/k(x,y)| \leq \psi \tag{4.10}$$

**Assumption 2.** *There exists a $\kappa > 0$ such that for ally $x, y \in \mathcal{X}$, $k(x,y) \geq \kappa > 0$ and $\varphi$ is differentiable there exists $V > 0$ such that:*

$$\sup_{x \in \mathcal{X}} \mathbb{E}_\rho \left( \|\nabla_x \varphi(x,u)\|^2 \right) \leq V \tag{4.11}$$

We can now present our main result on our proposed approximation scheme of $\mathrm{OT}_{c,\varepsilon}$ which is obtained in linear time with high probability. See Appendix 4.5.1 for the proof.

**Theorem 4.3.1.** *Let $\delta > 0$ and $r \geq 1$. Then the Sinkhorn Alg. 3 with inputs $K_\theta$, $a$ and $b$ outputs $(u_\theta, v_\theta)$ such that*

$$|\,\mathrm{OT}_{c_\theta,\varepsilon} - \widehat{\mathrm{OT}}_{c_\theta,\varepsilon}| \leq \frac{\delta}{2}$$

*in* $\mathcal{O}\left(\frac{n\varepsilon r}{\delta}\left[Q_\theta - \log \min_{i,j}(a_i, b_j)\right]^2\right)$ *operations where* $Q_\theta = -\log \min_{i,j} k_\theta(x_i, y_j)$. *Moreover if Assumptions 1 and 2 hold then for $\tau > 0$,*

$$r \in \Omega\left(\frac{\psi^2}{\delta^2}\left[\min\left(d\varepsilon^{-1}\|C\|_\infty^2 + d\log\left(\frac{\psi V D}{\tau\delta}\right), \log\left(\frac{n}{\tau}\right)\right)\right]\right) \tag{4.12}$$

*and $u_1, ..., u_r$ drawn independently from $\rho$, with a probability $1-\tau$, $Q_\theta \leq \varepsilon^{-1}\|C\|_\infty^2 + \log(2 + \delta\varepsilon^{-1})$ and it holds*

$$|\,\mathrm{OT}_{c,\varepsilon} - \widehat{\mathrm{OT}}_{c_\theta,\varepsilon}| \leq \delta \tag{4.13}$$

Therefore with a probability $1-\tau$, Sinkhorn Alg. 3 with inputs $K_\theta$, $a$ and $b$ output a $\delta$-approximation of the entropic OT distance in $\tilde{\mathcal{O}}\left(\frac{n}{\varepsilon\delta^3}\|C\|_\infty^4 \psi^2\right)$ algebraic operation where the notation $\tilde{\mathcal{O}}(.)$ omits polylogarithmic factors depending on $R, D, \varepsilon, n$ and $\delta$.

It worth noting that for every $r \geq 1$ and $\theta$, Sinkhorn Alg. 3 using kernel matrix $K_\theta$ will converge towards the solution of the entropic OT problem associated with the cost function $c_\theta$ in linear time thanks to the positivity of the feature maps used. Moreover, to ensure with high probability that the solution obtained approximate an optimal solution for the entropic OT problem associated with the cost function $c$, we need, if the features are chosen randomly, to ensure a minimum number of them. In contrast such result is not possible in [122]. Indeed in their works,

the number of random features $r$ cannot be chosen arbitrarily as they need to ensure the positiveness of the all the coefficients of the approximated kernel matrix obtained by the Nyström algorithm of [178] to run the Sinkhorn iterations and therefore need a very high precision which requires a certain number of random features $r$.

**Remark 2** (Acceleration.)**.** *It is worth noting that our method can also be applied in combination with the accelerated version of the Sinkhorn algorithm proposed in [179]. Indeed for $\tau > 0$, applying our approximation scheme to their algorithm leads with a probability $1-\tau$ to a $\delta/2$-approximation of $\mathrm{OT}_{c,\varepsilon}$ in $\mathcal{O}\left(\frac{nr}{\sqrt{\delta}}[\sqrt{\varepsilon^{-1}}A_\theta]\right)$ algebraic operations where $A_\theta = \inf_{(\alpha,\beta)\in\Theta_\theta} \|(\alpha,\beta)\|_2$, $\Theta_\theta$ is the set of optimal dual solutions of (4.8) and $r$ satisfying Eq.(4.12). See the full statement and the proof in Appendix 4.5.2.*

The number of random features prescribed in Theorem 4.3.1 ensures with high probability that $\widehat{\mathrm{OT}}_{c_\theta,\varepsilon}$ approximates $\mathrm{OT}_{c,\varepsilon}$ well when $u_1,\dots,u_r$ are drawn independently from $\rho$. Indeed, to control the error due to the approximation made through the Sinkhorn iterations, we need to control the error of the approximation of $K$ by $K_\theta$ relatively to $K$. In the next proposition we show with high probability that for all $(x,y)\in\mathcal{X}\times\mathcal{X}$,

$$(1-\delta)k(x,y) \leq k_\theta(x,y) \leq (1+\delta)k(x,y) \tag{4.14}$$

for an arbitrary $\delta > 0$ as soon as the number of random features $r$ is large enough. See Appendix 4.5.3 for the proof.

**Proposition 4.3.1.** *Let $\mathcal{X}\subset\mathbb{R}^d$ compact, $n\geq 1$, $X=\{x_1,...,x_n\}$ and $Y=\{y_1,...,y_n\}$ such that $X,Y\subset\mathcal{X}$, $\delta>0$. If $u_1,...,u_r$ are drawn independently from $\rho$ then under Assumption 1 we have*

$$\mathbb{P}\left(\sup_{(x,y)\in X\times Y}\left|\frac{k_\theta(x,y)}{k(x,y)}-1\right|\geq\delta\right)\leq 2n^2\exp\left(-\frac{r\delta^2}{2\psi^2}\right)$$

*Moreover if in addition Assumption 2 holds then we have*

$$\mathbb{P}\left(\sup_{(x,y)\in\mathcal{X}\times\mathcal{X}}\left|\frac{k_\theta(x,y)}{k(x,y)}-1\right|\geq\delta\right)\leq\frac{(\kappa^{-1}D)^2C_{\psi,V,r}}{\delta^2}\exp\left(-\frac{r\delta^2}{2\psi^2(d+1)}\right)$$

*where $C_{\psi,V,r}=2^9\psi(4+\psi^2/r)V\sup_{x\in\mathcal{X}}k(x,x)$ and $D=\sup_{(x,y)\in\mathcal{X}\times\mathcal{X}}\|(x,y)\|_2$.*

**Remark 3** (Ratio Approximation.). *The uniform bound obtained here to control the ratio gives naturally a control of the form Eq.(4.14). In comparison, in [180], the authors obtain a uniform bound on their difference which leads with high probability to a uniform control of the form*

$$k(x,y) - \tau \leq k_\theta(x,y) \leq k(x,y) + \tau \tag{4.15}$$

*where $\tau$ is a decreasing function with respect to $r$ the number of random features required. To be able to recover Eq.(4.14) from the above control, one may consider the case when $\tau = \inf_{x,y \in X \times Y} k(x,y)\delta$ which can considerably increases the number of of random features $r$ needed to ensure the result with at least the same probability. For example if the kernel is the Gibbs kernel associated to a cost function $c$, then $\inf_{x,y \in X \times Y} k(x,y) = \exp(-\|C\|_\infty/\varepsilon)$. More details are left in Appendix 4.5.3.*

In the following, we provides examples of some usual kernels $k$ that admits a decomposition of the form Eq.(4.9), satisfy Assumptions 1 and 2 and hence for which Theorem 4.3.1 can be applied.

**Arc-cosine Kernels.** Arc-cosine kernels have been considered in several works, starting notably from [181], [182] and [183]. The main idea behind arc-cosine kernels is that they can be written using positive maps for vectors $x, y$ in $\mathbb{R}^d$ and the signs (or higher exponent) of random projections $\mu = \mathcal{N}(0, I_d)$

$$k_s(x,y) = \int_{\mathbb{R}^d} \Theta_s(u^T x)\Theta_s(u^T y) d\mu(u)$$

where $\Theta_s(w) = \sqrt{2}\max(0,w)^s$ is a rectified polynomial function. In fact from these formulations, we build a perturbed version of $k_s$ which admits a decomposition of the form Eq.(4.9) that satisfies the required assumptions. See Appendix 4.5.5 for the full statement and the proof.

**Gaussian kernel.** The Gaussian kernel is in fact an important example as it is both a very widely used kernel on its own and its cost function associated is the square Euclidean metric. A decomposition of the form (4.9) has been obtained in ([184]) for the Gaussian kernel but it does not satisfies the required assumptions. In the following lemma, we built a feature map of the Gaussian kernel that satisfies them. See Appendix 4.5.4 for the proof.

**Lemma 1.** *Let $d \geq 1$, $\varepsilon > 0$ and $k$ be the kernel on $\mathbb{R}^d$ such that for all $x, y \in \mathbb{R}^d$, $k(x,y) = e^{-\|x-y\|_2^2/\varepsilon}$. Let $R > 0$, $q = \frac{R^2}{2\varepsilon d \, \mathrm{OT}_0(R^2/\varepsilon d)}$ where $\mathrm{OT}_0$ is the Lambert*

*function,* $\sigma^2 = q\varepsilon/4$, $\rho = \mathcal{N}(0, \sigma^2 Id)$ *and let us define for all* $x, u \in \mathbb{R}^d$ *the following map*

$$\varphi(x,u) = (2q)^{d/4} \exp\left(-2\varepsilon^{-1}\|x-u\|_2^2\right) \exp\left(\frac{\varepsilon^{-1}\|u\|_2^2}{\frac{1}{2}+\varepsilon^{-1}R^2}\right)$$

*Then for any* $x, y \in \mathbb{R}^d$ *we have* $k(x,y) = \int_{u\in\mathbb{R}^d} \varphi(x,u)\varphi(y,u)d\rho(u)$. *Moreover if* $x, y \in \mathcal{B}(0,R)$ *and* $u \in \mathbb{R}^d$ *we have* $k(x,y) \geq \exp(-4\varepsilon^{-1}R^2) > 0$,

$$|\varphi(x,u)\varphi(y,u)/k(x,y)| \leq 2^{d/2+1}q^{d/2} \quad \textit{and}$$
$$\sup_{x\in\mathcal{B}(0,R)} \mathbb{E}(\|\nabla_x \varphi\|_2^2) \leq 2^{d/2+3}q^{d/2}\left[(R/\varepsilon)^2 + \frac{q}{4\varepsilon}\right].$$

### 4.3.3 Constructive approach to Designing Positive Features: Differentiability

In this section we consider a constructive way of building feature map $\varphi_\theta$ which may be chosen arbitrary, or learned accordingly to an objective defined as a function of the entropic OT distance, e.g. OT-GAN objectives [50, 185]. For that purpose, we want to be able to compute the gradient of $\mathrm{OT}_{c_\theta,\varepsilon}(\mu,\nu)$ with respect to the kernel $K_\theta$, or more specifically with respect to the parameter $\theta$ and the locations of the input measures. In the next proposition we show that this entropic OT distance is differentiable with respect to the kernel matrix. See Appendix 4.6 for the proof.

**Proposition 4.3.2.** *Let* $\epsilon > 0$, $(a,b) \in \Delta_n \times \Delta_m$ *and let us also define for any* $K \in (\mathbb{R}_+^*)^{n\times m}$ *with positive entries the following function:*

$$G(K) := \sup_{(\alpha,\beta)\in\mathbb{R}^n\times\mathbb{R}^m} \langle\alpha, a\rangle + \langle\beta, a\rangle - \varepsilon(e^{\alpha/\varepsilon})^T K e^{\beta/\varepsilon}. \tag{4.16}$$

*Then* $G$ *is differentiable on* $(\mathbb{R}_+^*)^{n\times m}$ *and its gradient is given by*

$$\nabla G(K) = -\varepsilon e^{\alpha^\star/\varepsilon}(e^{\beta^\star/\varepsilon})^T \tag{4.17}$$

*where* $(\alpha^\star, \beta^\star)$ *are optimal solutions of Eq.(4.16).*

Note that when $c$ is the square euclidean metric, the differentiability of the above objective has been obtained in [126]. We can now provide the formula for the gradients of interest. For all $X := \left[x_1, \ldots, x_n\right] \in \mathbb{R}^{d\times n}$, we denote $\mu(X) = \sum_{i=1}^n a_i \delta_{x_i}$ and $\mathrm{OT}_{c_\theta,\varepsilon} = \mathrm{OT}_{c_\theta,\varepsilon}(\mu(X),\nu)$. Assume that $\theta$ is a $M$-dimensional vector for simplicity and that $(x,\theta) \in \mathbb{R}^d \times \mathbb{R}^M \to \varphi_\theta(x) \in (\mathbb{R}_+^*)^r$ is a differentiable map.

Then from proposition 4.3.2 and by applying the chain rule theorem, we obtain that

$$\nabla_\theta \operatorname{OT}_{c_\theta,\varepsilon} = -\varepsilon \left( \left(\frac{\partial \xi}{\partial \theta}\right)^T u_\theta^\star (\zeta v_\theta^\star)^T + \left(\frac{\partial \zeta}{\partial \theta}\right)^T v_\theta^\star (\xi u_\theta^\star)^T \right),$$

$$\nabla_X \operatorname{OT}_{c_\theta,\varepsilon} = -\varepsilon \left(\frac{\partial \xi}{\partial X}\right)^T u_\theta^\star (\zeta v_\theta^\star)^T$$

where $(u_\theta^*, v_\theta^*)$ are optimal solutions of (4.5) associated to the kernel matrix $K_\theta$. Note that $\left(\frac{\partial \xi}{\partial \theta}\right)^T$, $\left(\frac{\partial \zeta}{\partial \theta}\right)^T$ and $\left(\frac{\partial \xi}{\partial X}\right)^T$ can be evaluated using simple differentiation if $\varphi_\theta$ is a simple random feature, or, more generally, using automatic differentiation if $\varphi_\theta$ is the output of a neural network.

**Discussion.** Our proposed method defines a kernel matrix $K_\theta$ and a parametrized entropic OT distance $\operatorname{OT}_{c_\theta,\varepsilon}$ which are differentiable with respect to the input measures and the parameter $\theta$. These proprieties are important and used in many applications, e.g. GANs. However such operations may not be allowed when using a data-dependent method to approximate the kernel matrix such as the Nyström method used in [122]. Indeed there, the approximated kernel $\widetilde{K}$ and the entropic OT distance $\operatorname{OT}_{\varepsilon,\widetilde{c}}$ associated are not well defined on a neighbourhood of the locations of the inputs measures and therefore are not differentiable.

## 4.4 Experiments

**Efficiency vs. Approximation trade-off using positive features.** In Figures 8.8,4.3 we plot the deviation from ground truth, defined as $\mathrm{D} := 100 \times \frac{OT_{c,\varepsilon} - \widehat{OT}_{c,\varepsilon}}{|OT_{c,\varepsilon}|} + 100$, and show the time-accuracy tradeoff for our proposed method **RF**, Nystrom **Nys** [122] and Sinkhorn **Sin** [112], for a range of regularization parameters $\varepsilon$ (each corresponding to a different ground truth $OT_{\varepsilon,c}$) and approximation with $r$ random features in two settings. In particular, we show that our method obtains very high accuracy with order of magnitude faster than **Sin** in a larger regime of regularizations than **Nys**. In Figure 4.5 in Appendix 4.7, we also show the time-accuracy tradeoff in the high dimensional setting.

**Using positive features to learn adversarial kernels in GANs.** Let $P_\mathcal{X}$ a given distribution on $\mathcal{X} \subset \mathbb{R}^D$, $(\mathcal{Z}, \mathcal{A}, \zeta)$ an arbitrary probability space and let $g_\rho : \mathcal{Z} \to \mathcal{X}$ a parametric function where the parameter $\rho$ lives in a topological space $\mathcal{O}$. The function $g_\rho$ allows to generate a distribution on $\mathcal{X}$ by considering the push forward operation through $g_\rho$. Indeed $g_{\rho\sharp}\zeta$ is a distribution on $\mathcal{X}$ and if the function space $\mathcal{F} = \{g_\rho \colon \rho \in \mathcal{O}\}$ is large enough, we may be able to recover $P_\mathcal{X}$ for

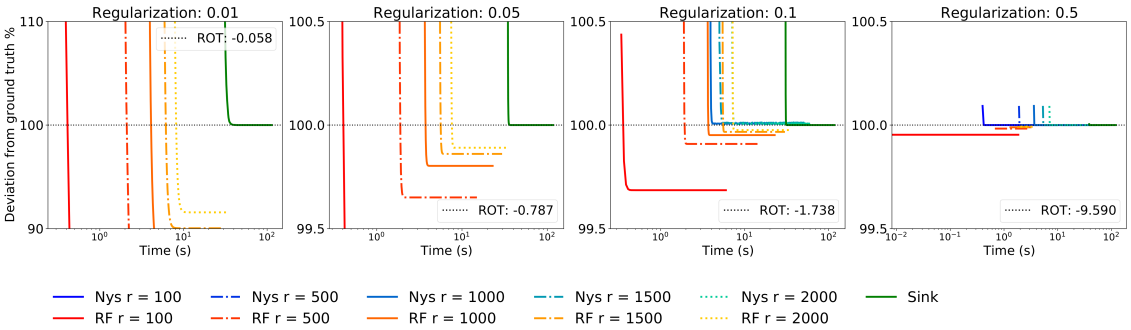



Figure 4.1: In this experiment, we draw 40000 samples from two normal distributions and we plot the deviation from ground truth for different regularizations. These two normal distributions are in $\mathbb{R}^2$. One of them has mean $(1,1)^T$ and identity covariance matrix $I_2$. The other has 0 mean and covariance $0.1 \times I_2$. We compare the results obtained for our proposed method (**RF**) with the one proposed in [122] (**Nys**) and with the Sinkhorn algorithm (**Sin**) proposed in [112]. The cost function considered here is the square Euclidean metric and the feature map used is that presented in Lemma 1. The number of random features (or rank) chosen varies from 100 to 2000. We repeat for each problem 50 times the experiment. Note that curves in the plot start at different points corresponding to the time required for initialization. *Right*: when the regularization is sufficiently large both **Nys** and **RF** methods obtain very high accuracy with order of magnitude faster than **Sin**. *Middle right, middle left*: **Nys** fails to converge while **RF** works for any given random features and provides very high accuracy of the entropic OT cost with order of magnitude faster than **Sin**. *Left*: when the regularization is too small all the methods failed as the Nystrom method cannot be computed, the accuracy of the **RF** method is of order of 10% and Sinkhorn algorithm may be too costly.

a well chosen $\rho$. The goal is to learn $\rho^*$ such that $g_{\rho^*_\sharp}\zeta$ is the closest possible to $P_{\mathcal{X}}$ according to a specific metric on the space of distributions. Here we consider the Sinkhorn distance as introduced in Eq.(4.2). One difficulty when using such metric is to define a well behaved cost to measure the distance between distributions in the ground space. We decide to learn an adversarial cost by embedding the native space $\mathcal{X}$ into a low-dimensional subspace of $\mathbb{R}^d$ thanks to a parametric function $f_\gamma$. Therefore by defining $h_\gamma(x,y) := (f_\gamma(x), f_\gamma(y))$ and given a fixed cost function $c$ on $\mathbb{R}^d$, we can define a parametric cost function on $\mathcal{X}$ as $c \circ h_\gamma(x,y) := c(f_\gamma(x), f_\gamma(y))$. To train a Generative Adversarial Network (GAN), one may therefore optimizes the following objective:

$$\min_\rho \max_\gamma S_{c \circ h_\gamma, \varepsilon}(g_{\rho_\#}\zeta, P_{\mathcal{X}})$$

Indeed, taking the max of the Sinkhorn distance according to $\gamma$ allows to learn a discriminative cost $c \circ h_\gamma$ [185, 50]. However in practice, we do not have access

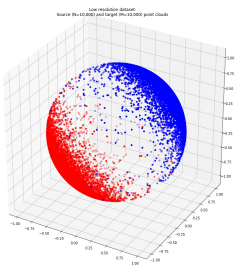


Figure 4.2: Here we show the two distributions considered in the experiment presented in Figure 4.3 to compare the time-accuracy tradeoff between the different methods. All the points are drawn on the unit sphere in $\mathbb{R}^3$, and uniform distributions are considered respectively on the red dots and on the blue dots. There are 10000 samples for each distribution.

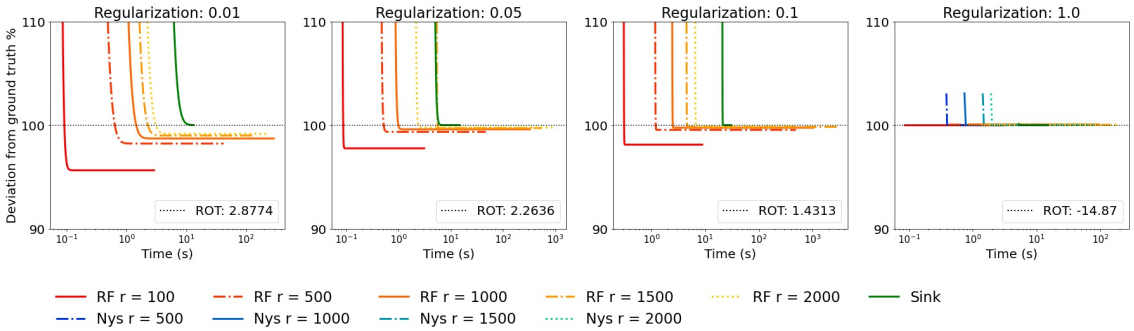



Figure 4.3: In this experiment, we draw 20000 samples from two distributions on the sphere (see Figure 4.2) and we plot the deviation from ground truth for different regularizations. We compare the results obtained for our proposed method (**RF**) with the one proposed in [122] (**Nys**) and with the Sinkhorn algorithm (**Sin**) proposed in [112]. The cost function considered here is the square Euclidean metric and the feature map used is that presented in Lemma 1. The number of random features (or rank) chosen varies from 100 to 2000. We repeat for each problem 10 times the experiment. Note that curves in the plot start at different points corresponding to the time required for initialization. *Right*: when the regularization is sufficiently large both **Nys** and **RF** methods obtain very high accuracy with order of magnitude faster than **Sin**. *Middle right, middle left, left*: **Nys** fails to converge while **RF** works for any given random features and provides very high accuracy of the entropic OT cost with order of magnitude faster than **Sin**.

to the distribution of the data $P_{\mathcal{X}}$, but only to its empirical version $\widehat{P}_X$, where $\widehat{P}_X := \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ and $X := \{x_1, ..., x_n\}$ are the $n$ i.i.d samples drawn from $P_{\mathcal{X}}$. By sampling independently $n$ samples $Z := \{z_1, ..., z_n\}$ from $\zeta$ and denoting $\widehat{\zeta} := \frac{1}{q}\sum_{i=1}^q \delta_{z_i}$ we obtain the following approximation:

$$\min_{\rho} \max_{\gamma} S_{coh_{\gamma,\varepsilon}}(g_{\rho\#}\widehat{\zeta}, \widehat{P}_X)$$

However as soon as $n$ gets too large, the above objective, using the classic Sinkhorn Alg. 3 is very costly to compute as the cost of each iteration of Sinkhorn is quadratic in the number of samples. Therefore one may instead split the data and consider

$B \geq 1$ mini-batches $Z = (Z^b)_{b=1}^B$ and $X = (X^b)_{b=1}^B$ of size $s = \frac{n}{B}$, and obtain instead the following optimisation problem:

$$\min_{\rho} \max_{\gamma} \frac{1}{B} \sum_{b=1}^{B} S_{c \circ h_\gamma, \varepsilon}(g_{\rho\#} \widehat{\zeta}^b, \widehat{P}_X^b)$$

where $\widehat{\zeta}^b := \frac{1}{s} \sum_{i=1}^s \delta_{z_i^b}$ and $\widehat{P}_X^b := \frac{1}{s} \sum_{i=1}^s \delta_{x_i^b}$. However the smaller the batches are, the less precise the approximation of the objective is. To overcome this issue we propose to apply our method and replace the cost function $c$ by an approximation defined as $c_\theta(x, y) = -\epsilon \log \varphi_\theta(x)^T \varphi_\theta(y)$ and consider instead the following optimisation problem:

$$\min_{\rho} \max_{\gamma} \frac{1}{B} \sum_{b=1}^{B} S_{\varepsilon, c_\theta \circ h_\gamma}(g_{\rho\#} \widehat{\zeta}^b, \widehat{P}_X^b).$$

Indeed in that case, the Gibbs kernel associated to the cost function $c_\theta \circ h_\gamma$ is still factorizafable as we have $c_\theta \circ h_\gamma(x, y) = -\epsilon \log \varphi_\theta(f_\gamma(x))^T \varphi_\theta(f_\gamma(y))$. Such procedure allows us to compute the objective in linear time and therefore to largely increase the size of the batches. Note that we keep the batch formulation as we still need it because of memory limitation on GPUs. Moreover, we may either consider a random approximation by drawing $\theta$ randomly for a well chosen distribution or we could learn the random features $\theta$. In the following we decide to learn the features $\theta$ in order to obtain a cost function $c_\theta \circ h_\gamma$ even more discriminative. Finally our objective is:

$$\min_{\rho} \max_{\gamma, \theta} \frac{1}{B} \sum_{b=1}^{B} S_{\varepsilon, c_\theta \circ h_\gamma}(g_{\rho\#} \widehat{\zeta}^b, \widehat{P}_X^b) \tag{4.18}$$

Therefore here we aim to learn an embedding from the input space into the feature space thanks to two operations. The first one consists in taking a sample and embedding it into a latent space thanks to the mapping $f_\gamma$ and the second one is an embedding of this latent space into the feature space thanks to the feature map $\varphi_\theta$. From now on we assume that $g_\rho$ and $f_\gamma$ are neural networks. More precisely we take the exact same functions used in [186, 187] to define $g_\rho$ and $f_\gamma$. Moreover, $\varphi_\theta$ is the feature map associated to the Gaussian kernel defined in Lemma 1 where $\theta$ is initialised with a normal distribution. The number of random features considered has been fixed to be $r = 600$ in the following. The training procedure is the same as [175, 187] and consists in alterning $n_c$ optimisation steps to train the cost function $c_\theta \circ h_\gamma$ and an optimisation step to train the generator $g_\rho$. The code is available at github.com/meyerscetbon/LinearSinkhorn.

| $k_\theta(f_\gamma(x), f_\gamma(z))$ | Image $x$ | Noise $z$ |
|---|---|---|
| Image $x$ | $1802 \times 1e12$ | $2961 \times 1e5$ |
| Noise $z$ | $2961 \times 1e5$ | 48.65 |

Table 4.1: Comparison of the learned kernel $k_\theta$, trained on CIFAR-10 by optimizing the objective (4.18), between images taken from CIFAR-10 and random noises sampled in the native of space of images. The values shown are averages obtained between 5 noise and/or image samples. As we can see the cost learned has well captured the structure of the image space.

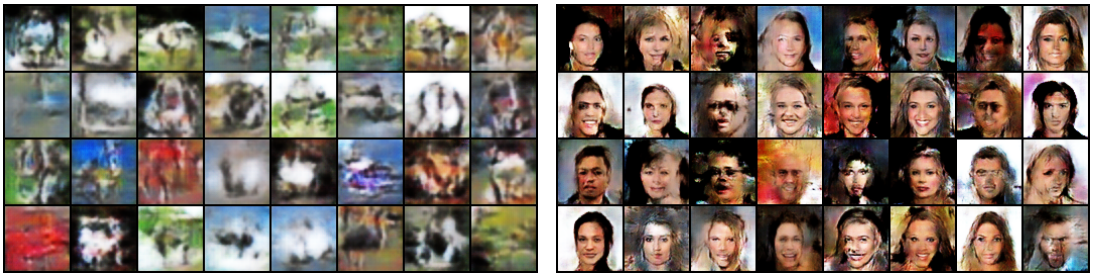


Figure 4.4: Images generated by two learned generative models trained by optimizing the objective (4.18) where we set the number of batches $s = 7000$, the regularization $\varepsilon = 1$, and the number of features $r = 600$. *Left, right:* samples obtained from the proposed generative model trained on respectively CIFAR-10 [188] and celebA [189].

**Optimisation.** Thanks to proposition 4.3.2, the objective is differentiable with respect to $\theta, \gamma$ and $\rho$. We obtain the gradient by computing an approximation of the gradient thanks to the approximate dual variables obtained by the Sinkhorn algorithm. We refers to section 4.3.3 for the expression of the gradient. This strategy leads to two benefits. First it is memory efficient as the computation of the gradient at this stage does not require to keep track of the computations involved in the Sinkhorn algorithm. Second it allows, for a given regularization, to compute with very high accuracy the Sinkhorn distance. Therefore, our method may be applied also for small regularization.

**Results.** We train our GAN models on a Tesla K80 GPU for 84 hours on two different datasets, namely CIFAR-10 dataset [188] and CelebA dataset [189] and learn both the proposed generative model and the adversarial cost function $c_\theta$ derived from the adversarial kernel $k_\theta$. Figure 4.4 illustrates the generated samples and Table 4.1 displays the geometry captured by the learned kernel.

**Discussion.** Our proposed method has mainly two advantages compared to the other Wasserstein GANs (W-GANs) proposed in the literature. First, the computation of the Sinkhorn divergence is linear with respect to the number of samples which allow to largely increase the batch size when training a W-GAN and obtain a better approximation of the true Sinkhorn divergence. Second, our approach is fully differentiable and therefore we can directly compute the gradient of the Sinhkorn divergence with respect the parameters of the network. In [50] the authors do not differentiate through the Wasserstein cost to train their network. In [185] the authors do differentiate through the iterations of the Sinkhorn algorithm but this strategy require to keep track of the computation involved in the Sinkhorn algorithm and can be applied only for large regularizations as the number of iterations cannot be too large.

# Supplementary materials

**Outline.** In Sec. 4.5 we provide the proofs related to the approximation proprieties of our proposed method. In Sec. 4.6 we show the differentiability of the constructive approach. Finally in Sec. 4.7 we add more experiments and illustrations of our proposed method.

## 4.5 Approximation via Random Fourier Features

### 4.5.1 Proof of Theorem 4.3.1

In the following we denote $K = (k(x_i, y_j))_{i,j=1}^n$ $K_\theta = (k_\theta(x_i, y_j))_{i,j=1}^n$ the two gram matrices associated with $k$ and $k_\theta$ respectively. By duality and from these two matrices we can define the two objectives to maximize to obtain $\mathrm{OT}_{c,\varepsilon}$ and $\mathrm{OT}_{c_\theta,\varepsilon}$:

$$\mathrm{OT}_{c,\varepsilon} = \max_{\alpha,\beta} f(\alpha,\beta) := \langle \alpha, a\rangle + \langle \beta, b\rangle - \varepsilon\langle e^{\alpha/\varepsilon}, K e^{\beta/\varepsilon}\rangle$$
$$\mathrm{OT}_{c_\theta,\varepsilon} = \max_{\alpha,\beta} f_\theta(\alpha,\beta) := \langle \alpha, a\rangle + \langle \beta, b\rangle - \varepsilon\langle e^{\alpha/\varepsilon}, K_\theta e^{\beta/\varepsilon}\rangle$$

Moreover as $k$ and $\varphi$ are assumed to be positive, there exists unique (up to a scalar translation) $(\alpha^*, \beta^*)$ and $(\alpha_\theta^*, \beta_\theta^*)$ respectively solutions of $\max_{\alpha,\beta} f(\alpha,\beta)$ and $\max_{\alpha,\beta} f_\theta(\alpha,\beta)$.

**Proof.** *Let us first show the following proposition:*

**Proposition 1.** *Let $\delta > 0$ and $r \geq 1$. Assume that for all $(x, y) \in X \times Y$,*

$$\left|\frac{k(x,y) - k_\theta(x,y)}{k(x,y)}\right| \leq \frac{\delta\varepsilon^{-1}}{2 + \delta\varepsilon^{-1}} \tag{4.19}$$

*then Sinkhorn Alg. 3 with inputs $a, b, K_\theta$ outputs $(\alpha_\theta, \beta_\theta)$ in*

$$\mathcal{O}\left(\frac{nr}{\delta\varepsilon^{-1}}\left[\log\left(\frac{1}{\iota}\right) + \log\left(2 + \delta\varepsilon^{-1}\right) + \varepsilon^{-1}R^2\right]^2\right)$$

*where*

$$\iota = \min_{i,j}(a_i, b_j) \quad \text{and} \quad R = \max_{(x,y)\in X\times Y} c(x,y). \tag{4.20}$$

*such that:*

$$|\,\mathrm{OT}_{c,\varepsilon} - f_\theta(\alpha_\theta, \beta_\theta)| \leq \delta$$

**Proof.** *We remark that:*

$$\begin{aligned}|f(\alpha^*,\beta^*)-f_\theta(\alpha_\theta,\beta_\theta)| &\leq |f(\alpha^*,\beta^*)-f(\alpha^*_\theta,\beta^*_\theta)|\\ &+|f(\alpha^*_\theta,\beta^*_\theta)-f_\theta(\alpha^*_\theta,\beta^*_\theta)|\\ &+|f_\theta(\alpha^*_\theta,\beta^*_\theta)-f_\theta(\alpha_\theta,\beta_\theta)|\end{aligned}$$

*Moreover we have that:*

$$\begin{aligned}|f(\alpha^*,\beta^*)-f(\alpha^*_\theta,\beta^*_\theta)| &= f(\alpha^*,\beta^*)-f(\alpha^*_\theta,\beta^*_\theta)\\ &= f(\alpha^*,\beta^*)-f_\theta(\alpha^*_\theta,\beta^*_\theta)+f_\theta(\alpha^*_\theta,\beta^*_\theta)-f(\alpha^*_\theta,\beta^*_\theta)\\ &\leq |f(\alpha^*,\beta^*)-f_\theta(\alpha^*,\beta^*)|+|f_\theta(\alpha^*_\theta,\beta^*_\theta)-f(\alpha^*_\theta,\beta^*_\theta)|\end{aligned}$$

*Therefore we obtain that:*

$$\begin{aligned}|f(\alpha^*,\beta^*)-f_\theta(\alpha_\theta,\beta_\theta)| &\leq 2|f(\alpha^*_\theta,\beta^*_\theta)-f_\theta(\alpha^*_\theta,\beta^*_\theta)|+|f(\alpha^*,\beta^*)-f_\theta(\alpha^*,\beta^*)|\\ &+|f_\theta(\alpha^*_\theta,\beta^*_\theta)-f_\theta(\alpha_\theta,\beta_\theta)|\end{aligned}$$

*Let us now introduce the following lemma:*

**Lemma 2.** *Let $1>\tau>0$ and let us assume that for all $(x,y)\in X\times Y$,*

$$\left|\frac{k(x,y)-k_\theta(x,y)}{k(x,y)}\right|\leq\tau$$

*then for any $\alpha,\beta\in\mathbb{R}^n$ it holds*

$$|f(\alpha,\beta)-f_\theta(\alpha,\beta)|\leq\varepsilon\tau[\langle e^{\varepsilon^{-1}\alpha},Ke^{\varepsilon^{-1}\beta}\rangle] \tag{4.21}$$

*and*

$$|f(\alpha,\beta)-f_\theta(\alpha,\beta)|\leq\varepsilon\frac{\tau}{1-\tau}[\langle e^{\varepsilon^{-1}\alpha},K_\theta e^{\varepsilon^{-1}\beta}\rangle] \tag{4.22}$$

**Proof.** *Let $\alpha,\beta\in\mathbb{R}^n$. We remarks that:*

$$f(\alpha,\beta)-f_\theta(\alpha,\beta)=\varepsilon[\langle e^{\varepsilon^{-1}\alpha},(K_\theta-K)e^{\varepsilon^{-1}\beta}\rangle]$$

*Therefore we obtain that:*

$$|f(\alpha,\beta)-f_\theta(\alpha,\beta)|\leq\varepsilon\sum_{i,j=1}^{n}e^{\varepsilon^{-1}\alpha_i}e^{\varepsilon^{-1}\beta_j}|[K_\theta]_{i,j}-K_{i,j}|$$

*And the first inequality follows from the fact that $|[K_\theta]_{i,j}-K_{i,j}|\leq\tau|K_{i,j}|$ for all $i,j\in\{1,...,n\}$ and that $k$ is positive. Moreover from the same inequality we obtain that:*

$$|[K_\theta]_{i,j}-K_{i,j}|\leq\frac{\tau}{1-\tau}[K_\theta]_{i,j}$$

*Therefore the second inequality follows.*

*Therefore thanks to lemma 2, we obtain that:*

$$|f(\alpha_\theta^*, \beta_\theta^*) - f_\theta(\alpha_\theta^*, \beta_\theta^*)| \leq \varepsilon \frac{\tau}{1-\tau} [\langle e^{\varepsilon^{-1}\alpha_\theta^*}, K_\theta e^{\varepsilon^{-1}\beta_\theta^*} \rangle]$$

*But as $(\alpha_\theta^*, \beta_\theta^*)$ is the optimum of $f_\theta$, the first order conditions give us that $\langle e^{\varepsilon^{-1}\alpha_\theta^*}, K_\theta e^{\varepsilon^{-1}\beta_\theta^*} \rangle = 1$ and finally we have:*

$$|f(\alpha_\theta^*, \beta_\theta^*) - f_\theta(\alpha_\theta^*, \beta_\theta^*)| \leq \varepsilon \frac{\tau}{1-\tau}$$

*Thanks to lemma 2, we also deduce that:*

$$|f(\alpha^*, \beta^*) - f_\theta(\alpha^*, \beta^*)| \leq \varepsilon\tau$$

*Let us now introduce the following theorem:*

**Theorem 4.5.1.** *([118]) Given $K_\theta \in \mathbb{R}^{n \times n}$ with positive entries and $a, b \in \Delta_n$ the Sinkhorn Alg. 3 computes $(\alpha_\theta, \beta_\theta)$ such that*

$$|f_\theta(\alpha_\theta^*, \beta_\theta^*) - f_\theta(\alpha_\theta, \beta_\theta)| \leq \frac{\delta}{2}$$

*in $\mathcal{O}\left(\delta^{-1}\varepsilon \log\left(\frac{1}{\iota \min_{i,j}[K_\theta]_{i,j}}\right)^2\right)$ iterations where $\iota = \min_{i,j}(a_i, b_j)$ and each of which requires $\mathcal{O}(1)$ matrix-vector products with $K_\theta$ and $\mathcal{O}(n)$ additional processing time.*

*Moreover from Eq. (4.19) we have that*

$$[K_\theta]_{i,j} \geq (1-\tau) K_{i,j}$$

*where $\tau = \frac{\delta\varepsilon^{-1}}{2+\delta\varepsilon^{-1}}$, therefore $\log\left(\frac{1}{\min_{i,j}[K_\theta]_{i,j}}\right) \leq \log\left(\frac{1}{(1-\tau)\min_{i,j} K_{i,j}}\right) \leq \log\left(\frac{1}{1-\tau}\right) + \varepsilon^{-1}R^2$ where $R = \max_{(x,y)\in X\times Y} c(x,y)$ and we obtain that*

$$|f(\alpha^*, \beta^*) - f_\theta(\alpha_\theta, \beta_\theta)| \leq 2\varepsilon \frac{\tau}{1-\tau} + \varepsilon\tau + \frac{\delta}{2}$$

*By replacing $\tau$ by its value, we obtain the desired result.*

*We are now ready to prove the theorem. Let $r \geq 1$. From theorem 4.5.1, we obtain directly that:*

$$|f(\alpha^*, \beta^*) - f_\theta(\alpha_\theta, \beta_\theta)| \leq \frac{\delta}{2}$$

*in* $\mathcal{O}\left(\frac{nr}{\delta}\left[\log\left(\frac{1}{\iota}\right)+Q_\theta\right]^2\right)$ *algebric operations. Moreover let* $\tau>0$ *and*

$$r\in\Omega\left(\frac{\psi^2}{\delta^2}\left[\min\left(d\varepsilon^{-1}R^2+d\log\left(\frac{\psi VD}{\tau\delta}\right),\log\left(\frac{n}{\tau}\right)\right)\right]\right)$$

*and* $u_1,...,u_r$ *drawn independently from* $\rho$*. Then from Proposition 4.3.1 we obtain that with a probability of at least* $1-\delta$ *it holds for all* $(x,y)\in X\times Y$*,*

$$\left|\frac{k(x,y)-k_\theta(x,y)}{k(x,y)}\right|\leq\frac{\delta\varepsilon^{-1}}{2+\delta\varepsilon^{-1}}$$

*and the result follows from Proposition 1.*

### 4.5.2 Accelerated Version

[179] show that one can accelarated the Sinkhorn algorithm (see Alg. 2) and obtain a $\delta$-approximation of the ROT distance. For that purpose, [179] introduce a reformulation of the dual problem (4.8) and obtain

$$\mathrm{OT}_{c_\theta,\varepsilon}=\sup_{\eta_1,\eta_2}F_\theta(\eta_1,\eta_2):=\varepsilon\left[\langle\eta_1,a\rangle+\langle\eta_2,b\rangle-\log\left(\langle e^{\eta_2},K_\theta e^{\eta_2}\rangle\right)\right]$$

which can be shown to be an $L$-smooth function ([190]) where $L\leq 2\varepsilon^{-1}$. Let us now present our result using the accelerated Sinkhorn algorithm.

**Theorem 4.5.2.** *Let* $\delta>0$ *and* $r\geq 1$*. Then the Accelerated Sinkhorn Alg. 2 with inputs* $K_\theta$*,* $a$ *and* $b$ *outputs* $(\alpha_\theta,\beta_\theta)$ *such that*

$$|\,\mathrm{OT}_{c_\theta,\varepsilon}-F_\theta(\alpha_\theta,\beta_\theta)|\leq\frac{\delta}{2}$$

*in* $\mathcal{O}\left(\frac{nr}{\sqrt{\delta}}[\sqrt{\varepsilon^{-1}}A_\theta]\right)$ *algebraic operations where* $A_\theta=\inf_{(\alpha,\beta)\in\Theta_\theta}\|(\alpha,\beta)\|_2$ *and* $\Theta_\theta$ *is the set of optimal dual solutions of (4.8). Moreover let* $\tau>0$*,*

$$r\in\Omega\left(\frac{\psi^2}{\delta^2}\left[\min\left(d\varepsilon^{-1}\|C\|_\infty^2+d\log\left(\frac{\psi VD}{\delta\delta}\right),\log\left(\frac{n}{\delta}\right)\right)\right]\right)$$

*and* $u_1,...,u_r$ *drawn independently from* $\rho$*, then with a probability* $1-\tau$ *it holds*

$$|\,\mathrm{OT}_{c,\varepsilon}-F_\theta(\alpha_\theta,\beta_\theta)|\leq\delta$$

**Proof.** *Let us first introduce the theorem presented in [179]:*

**Algorithm 2** Accelerated Sinkhorn Algorithm.

**Input:** Initial estimate of the Lipschitz constant $L_0$, $a$, $b$, and $K$
**Init:** $A_0 = \alpha_0 = 0$, $\eta^0 = \zeta^0 = \lambda^0 = 0$.
**for** $k \geq 0$ **do**
  $L_{k+1} = L_k/2$
  **while** *True* **do**
    Set $L_{k+1} = L_k/2$
    Set $a_{k+1} = \frac{1}{2L_{k+1}} + \sqrt{\frac{1}{4L_{k+1}^2} + a_k^2 \frac{L_k}{L_{k+1}}}$
    Set $\tau_k = \frac{1}{a_{k+1}L_{k+1}}$
    Set $\lambda^k = \tau_k \zeta^k + (1 - \tau_k)\zeta^k$
    Choose $i_k = \underset{i \in \{1,2\}}{\operatorname{argmax}} \|\nabla_i \phi(\lambda^k)\|_2$
    **if** $i_k = 1$ **then**
      $\eta_1^{k+1} = \lambda_1^k + \log(a) - \log(e^{\lambda_1^k} \circ K e^{\lambda_2^k})$
      $\eta_2^{k+1} = \lambda_2^{k+1}$
      **else**
        $\eta_1^{k+1} = \lambda_1^{k+1}$
        $\eta_2^{k+1} = \lambda_2^k + \log(b) - \log(e^{\lambda_2^k} \circ K^T e^{\lambda_1^k})$
      **end**
    **end**
    Set $\zeta^{k+1} = \zeta^k - a_{k+1} \nabla F_\theta(\lambda^k)$
    **if** $\phi(\eta^k + 1) \leq \phi(\lambda^k) - \frac{\|\nabla F_\theta(\lambda^k)\|^2}{2L_{k+1}}$ **then**
      Set $z = \operatorname{Diag}(e^{\lambda_1^k}) \circ K \circ \operatorname{Diag}(e^{\lambda_2^k})$
      Set $c = \langle e^{\lambda_1^k}, K e^{\lambda_2^k} \rangle$
      Set $\hat{x}^{k+1} = \frac{a_{k+1} c^{-1} z + L_k a_k^2 \hat{x^k}}{L_{k+1} a_{k+1}^2}$
      **Break**
    **end**
    Set $L_{k+1} = 2L_{k+1}$
  **end**
**end**
**Result**: Transport Plan $\hat{x}^{k+1}$ and dual points $\eta^{k+1} = (\eta_1^{k+1}, \eta_2^{k+1})^T$

**Theorem 4.5.3.** *Given $K_\theta \in \mathbb{R}^{n \times n}$ with positive entries and $a, b \in \Delta_n$ the Accelerated Sinkhorn Alg. (2) computes $(\alpha_\theta, \beta_\theta)$ such that*

$$|\operatorname{OT}_{c_\theta, \varepsilon} - F_\theta(\alpha_\theta, \beta_\theta)| \leq \delta$$

*in $\mathcal{O}\left(\sqrt{\frac{\eta}{\delta}} A_\theta\right)$ iterations where $A_\theta = \inf_{(\alpha_\theta^*, \beta_\theta^*) \in \Theta^*} \|(\alpha_\theta^*, \beta_\theta^*)\|_2$ and $\Theta^*$ is the set of*

*optimal dual solutions. Moreover each of which requires $\mathcal{O}(1)$ matrix-vector products with $K_\theta$ and $\mathcal{O}(n)$.*

*From the above result and applying an analogue proof of Theorem 4.5.1, we obtain the desired result.*

### 4.5.3 Proof of Proposition 4.3.1

**Proof.** *The proof is given for $p = 1$ but it hold also for any $p \geq 1$ after making some simple modifications. To obtain the first inequality we remarks that*

$$\mathbb{P}\left(\sup_{(x,y)\in\mathcal{X}\times\mathcal{X}}\left|\frac{k_\theta(x,y)}{k(x,y)}-1\right|\geq\delta\right)\leq\sum_{(x,y)\in X\times Y}\mathbb{P}\left(\left|\frac{k_\theta(x,y)}{k(x,y)}-1\right|\geq\delta\right)$$

*Moreover as $\mathbb{E}_\rho\left(\frac{\varphi(x,u)\varphi(y,u)}{k(x,y)}\right)=1$, the result follows by applying Hoeffding's inequality.*

*To show the second inequality, we follow the same strategy adopted in [180]. Let us denote $f(x,y)=\frac{k_\theta(x,y)}{k(x,y)}-1$ and $\mathcal{M}:=\mathcal{X}\times\mathcal{X}$. First we remarks that $|f(x,y)|\leq K+1$ and $\mathbb{E}_\rho(f)=0$. As $\mathcal{M}$ is a compact, we can find an $\mu$-net that covers $\mathcal{M}$ with $\mathcal{N}(\mathcal{M},\mu)=\left(\frac{4R}{\mu}\right)^{2d}$ where $R=\sup_{(x,y)}\|(x,y)\|_2$ balls of radius $\delta$. Let us denote $z_1,...,z_{\mathcal{N}(\mathcal{M},\mu)}\in\mathcal{M}$ the centers of these balls, and let $L_f$ denote the Lipschitz constant of $f$. As $f$ is differentiable We have therefore $L_f=\sup_{z\in\mathcal{M}}\|\nabla f(z)\|_2$. Moreover we have:*

$$\begin{aligned}\nabla f(z)&=\frac{\nabla k_\theta(z)}{k(z)}-\frac{k_\theta(z)}{k(z)}\nabla k(z)\\&=\frac{1}{k(z)}\left[(\nabla k_\theta(z)-\nabla k(z))+\nabla k(z)\left(1-\frac{k_\theta(z)}{k(z)}\right)\right]\end{aligned}$$

*Therefore we have*

$$\mathbb{E}(\|\nabla f(z)\|^2)\leq\frac{2}{k(z)^2}\left[\mathbb{E}(\|\nabla k_\theta(z)-\nabla k(z)\|^2)+\|\nabla k(z)\|^2\mathbb{E}\left(1-\frac{k_\theta(z)}{k(z)}\right)^2\right]$$

*But for any $z\in\mathcal{M}$ we have from Eq. (4.15) :*

$$\mathbb{E}\left(1-\frac{k_\theta(z)}{k(z)}\right)^2=\int_{t\geq0}\mathbb{P}\left(\left(1-\frac{k_\theta(z)}{k(z)}\right)^2\geq t\right)\tag{4.23}$$

$$\leq\frac{K^2}{r}\tag{4.24}$$

*Moreover, we have:*

$$\nabla k_\theta(z) = \frac{1}{r}\sum_{i=1}^{r} \nabla_x \varphi(x, u_i)\varphi(y, u_i) + \varphi(x, u_i)\nabla_y \varphi(y, u_i)$$

*Therefore we have:*

$$\begin{aligned}\|\nabla k_\theta(z)\|^2 = &\frac{1}{r^2}\sum_{i,j=1}^{r} \langle \nabla_x \varphi(x, u_i), \nabla_x \varphi(x, u_j)\rangle \varphi(y, u_i)\varphi(y, u_j) \\ &+ \frac{1}{r^2}\sum_{i,j=1}^{r} \nabla_y \varphi(y, u_i), \nabla_y \varphi(y, u_j)\rangle \varphi(x, u_i)\varphi(x, u_j) \\ &+ \frac{2}{r^2}\sum_{i,j=1}^{r} \nabla_x \varphi(x, u_i), \nabla_y \varphi(x, u_j)\rangle \varphi(y, u_i)\varphi(x, u_j)\end{aligned}$$

*Moreover as:*

$$\begin{aligned}|\varphi(y, u_i)\varphi(x, u_j)| &\leq \frac{\varphi(y, u_i)^2 + \varphi(x, u_j)^2}{2} \\ &\leq K \sup_{x\in\mathcal{X}} k(x, x)\end{aligned}$$

*And:*

$$\begin{aligned}|\langle \nabla_x \varphi(x, u_i), \nabla_y \varphi(y, u_j)\rangle| &\leq \|\nabla_x \varphi(x, u_i)\|\|\nabla_y \varphi(y, u_j)\| \\ &\leq \frac{\|\nabla_x \varphi(x, u_i)\|^2 + \|\nabla_y \varphi(y, u_j)\|^2}{2}\end{aligned}$$

*And by denoting:*

$$V := \sup_{x\in\mathcal{X}} \mathbb{E}_\rho \left(\|\nabla_x \varphi(x, u)\|^2\right)$$

*Therefore we have:*

$$\mathbb{E}\left(|\langle \nabla_x \varphi(x, u_i), \nabla_y \varphi(y, u_j)\rangle|\right) \leq V \tag{4.25}$$

*We can now derive the following upper bound:*

$$\mathbb{E}(\|\nabla k_\theta(z) - \nabla k(z)\|^2) = \mathbb{E}(\|\nabla k_\theta(z)\|^2) - \|\nabla k(z)\|^2 \leq 4VK \sup_{x\in\mathcal{X}} k(x, x)$$

*Moreover by convexity of the $\ell_2$ square norm, we also obtain that:*

$$\|\nabla k(z)\|^2 \leq VK \sup_{x\in\mathcal{X}} k(x, x)$$

*Therefore we have*

$$\mathbb{E}(\|\nabla f(z)\|^2) \leq 2\kappa^{-2}VK \sup_{x\in\mathcal{X}} k(x,x) \left[4 + \frac{K^2}{r}\right]$$

*Then by applying Markov inequality we obtain that:*

$$\mathbb{P}\left(L_f \geq \frac{\delta}{2\mu}\right) \leq 2\kappa^{-2}VK \sup_{x\in\mathcal{X}} k(x,x) \left[4 + \frac{K^2}{r}\right] \left(\frac{2\mu}{\delta}\right)^2 \tag{4.26}$$

*Moreover, the union bound followed by Hoeffding's inequality applied to the anchors in the $\mu$-net gives*

$$\mathbb{P}\left(\cup_{i=1}^{\mathcal{N}(\mathcal{M},\mu)} |f(z_i)| \geq \delta\right) \leq 2\mathcal{N}(\mathcal{M},\mu) \exp\left(-\frac{r\delta^2}{2K^2}\right) \tag{4.27}$$

*Then by combining Eq. (4.26) and Eq.(4.27) we obtain that:*

$$\begin{aligned}\mathbb{P}\left(\sup_{z\in\mathcal{M}} |f(z)| \geq \delta\right) \leq & 2\left(\frac{4R}{\mu}\right)^{2d} \exp\left(-\frac{r\delta^2}{2K^2}\right) \\ & + 2\kappa^{-2}VK \sup_{x\in\mathcal{X}} k(x,x) \left[4 + \frac{K^2}{r}\right] \left(\frac{2\mu}{\delta}\right)^2\end{aligned}$$

*Therefore by denoting*

$$\begin{aligned}A_1 &:= 2\,(4R)^{2d} \exp\left(-\frac{r\delta^2}{2K^2}\right) \\ A_2 &:= 2\kappa^{-2}VK \sup_{x\in\mathcal{X}} k(x,x) \left[4 + \frac{K^2}{r}\right] \left(\frac{2}{\delta}\right)^2\end{aligned}$$

*and by choosing $\mu = \frac{A_1}{A_2}^{\frac{1}{2d+2}}$, we obtain that:*

$$\begin{aligned}\mathbb{P}\left(\sup_{z\in\mathcal{M}} |f(z)| \geq \delta\right) \leq & 2^9 \left[\frac{\kappa^{-2}KV \sup_{x\in\mathcal{X}} k(x,x) \left[4 + \frac{K^2}{r}\right] R^2}{\delta^2}\right] \\ & \times \exp\left(-\frac{r\delta^2}{2K^2(d+1)}\right)\end{aligned}$$

**Ratio Approximation.** Let us assume here that $p = 1$ for simplicity. The uniform bound obtained on the ratio gives naturally a control of the form Eq.(4.14) with a prescribed number of random features $r$. This result allows to control the error when using the kernel matrix $K_\theta$ instead of the true kernel matrix $K$ in

the Sinkhorn iterations. In the proposition above, we obtain such a result with a probability of at least $1-2n^2\exp\left(-\frac{r\delta^2}{2\psi^2}\right)$ where $r$ is the number of random features and $\psi$ is defined as

$$\psi := \sup_{u\in\mathcal{U}} \sup_{(x,y)\in X\times Y} \left|\frac{\varphi(x,u)\varphi(y,u)}{k(x,y)}\right|.$$

In comparison, in [180], the authors obtain a uniform bound on their difference and by denoting

$$\phi = \sup_{u\in\mathcal{U}} \sup_{(x,y)\in X\times Y} |\varphi(x,u)\varphi(y,u)|,$$

one obtains that with a probability of at least $1-2n^2\exp\left(-\frac{r\tau^2}{2\phi^2}\right)$ for all $(x,y)\in X\times Y$

$$k(x,y)-\tau \leq k_\theta(x,y) \leq k(x,y)+\tau \tag{4.28}$$

To be able to recover Eq.(4.14) from the above control, we need to take $\tau = \inf_{x,y\in X\times Y} k(x,y)\delta$ and by denoting $\phi' = \frac{\phi}{\inf_{x,y\in X\times Y} k(x,y)}$ we obtain that with a probability of at least $1-2n^2\exp\left(-\frac{r\delta^2}{2\phi'^2}\right)$ for all $(x,y)\in X\times Y$

$$(1-\delta)k(x,y) \leq k_\theta(x,y) \leq (1+\delta)k(x,y)$$

Therefore the number of random features needed to guarantee Eq.(4.14) from a control between the difference of the two kernels with at least a probability $1-\delta$ has to be larger than $\left(\frac{\phi'}{\psi}\right)^2$ times the number of random features needed from the control of Proposition 4.3.1 to guarantee Eq.(4.14) with at least the same probability $1-\delta$. But we always have that

$$\psi = \sup_{u\in\mathcal{U}} \sup_{(x,y)\in X\times Y} \left|\frac{\varphi(x,u)\varphi(y,u)}{k(x,y)}\right| \leq \frac{\sup_{u\in\mathcal{U}} \sup_{(x,y)\in X\times Y} |\varphi(x,u)\varphi(y,u)|}{\inf_{x,y\in X\times Y} k(x,y)} = \phi'$$

and in some cases the ratio $\left(\frac{\phi'}{\psi}\right)^2$ can be huge. Indeed, as we will see in the following, for the Gaussian kernel,

$$k(x,y) = \exp(-\varepsilon^{-1}\|x-y\|_2^2)$$

there exists $\varphi$ and $\mathcal{U}$ such that for all $x, y$ and $u\in\mathcal{U}$:

$$\varphi(x,u)\varphi(y,u) = k(x,y)h(u,x,y)$$

where for all $(x_0, y_0) \in X \times Y$,

$$\sup_{u \in \mathcal{U}} |h(u, x_0, y_0)| = \sup_{u \in \mathcal{U}} \sup_{(x,y) \in X \times Y} |h(u, x, y)|.$$

Therefore by denoting $M = \sup_{(x,y) \in X \times Y} \|x - y\|_2$ and $m = \inf_{(x,y) \in X \times Y} \|x - y\|_2$ , we obtain that

$$\left(\frac{\phi'}{\psi}\right)^2 = \left(\frac{\sup_{x,y \in X \times Y} k(x, y)}{\inf_{x,y \in X \times Y} k(x, y)}\right)^2 = \exp\left(2\varepsilon^{-1}[M^2 - m^2]\right)$$

### 4.5.4 Proof of Lemma 1

**Proof.** *Let $\varepsilon > 0$ and $x, y \in \mathbb{R}^d$. We have that:*

$$\exp\left(-2\varepsilon^{-1}\|x - u\|_2^2\right) \exp\left(-2\varepsilon^{-1}\|y - u\|_2^2\right) = \exp\left(-\varepsilon^{-1}\|x - y\|_2^2\right) \times \exp\left(-4\varepsilon^{-1}\left\|u - \left(\frac{x + y}{2}\right)\right\|_2^2\right)$$

*And as the LHS is integrable we have:*

$$\int_{u \in \mathbb{R}^d} \exp\left(-2\varepsilon^{-1}\|x - u\|_2^2\right) \exp\left(-2\varepsilon^{-1}\|y - u\|_2^2\right) du = \int_{u \in \mathbb{R}^d} e^{-\varepsilon^{-1}\|x-y\|_2^2} \exp\left(-4\varepsilon^{-1}\left\|u - \left(\frac{x + y}{2}\right)\right\|_2^2\right) du$$

*Therefore we obtain that:*

$$e^{-\varepsilon^{-1}\|x-y\|_2^2} = \left(\frac{4}{\pi\varepsilon}\right)^{d/2} \int_{u \in \mathbb{R}^d} \exp\left(-2\varepsilon^{-1}\|x - u\|_2^2\right) \exp\left(-2\varepsilon^{-1}\|y - u\|_2^2\right) du$$

*Now we want to transform the above expression as the one stated in 4.9. To do so, let $q > 0$ and let us denote $f_q$ the probability density function associated with the multivariate Gaussian distribution $\rho_q \sim \mathcal{N}\left(0, \frac{q}{4\varepsilon^{-1}} Id\right)$. We can rewrite the RHS of*

*the above equation as the following:*

$$\begin{aligned}
&\left(\frac{4}{\pi\varepsilon}\right)^{d/2}\int_{u\in\mathbb{R}^d}\exp\left(-2\varepsilon^{-1}\|x-u\|_2^2\right)\exp\left(-2\varepsilon^{-1}\|x-u\|_2^2\right)du\\
&=\left(\frac{4}{\pi\varepsilon}\right)^{d/2}\int_{u\in\mathbb{R}^d}\exp\left(-2\varepsilon^{-1}\|x-u\|_2^2\right)\exp\left(-2\varepsilon^{-1}\|x-u\|_2^2\right)\frac{f_q(u)}{f_q(u)}d(u)\\
&=\left(\frac{4}{\pi\varepsilon}\right)^{d/2}\int_{u\in\mathbb{R}^d}\exp\left(-2\varepsilon^{-1}\|x-u\|_2^2\right)\exp\left(-2\varepsilon^{-1}\|x-u\|_2^2\right)\\
&\times\left[\left(2\pi\frac{q}{4\varepsilon^{-1}}\right)^{d/2}e^{\frac{2\varepsilon^{-1}\|u\|_2^2}{q}}\right]d\rho_q(u)\\
&=(2q)^{d/2}\int_{u\in\mathbb{R}^d}\exp\left(-2\varepsilon^{-1}\|x-u\|_2^2\right)\exp\left(-2\varepsilon^{-1}\|x-u\|_2^2\right)e^{\frac{2\varepsilon^{-1}\|u\|_2^2}{q}}d\rho_q(u)
\end{aligned}$$

*Therefore for each $q>0$, we obtain a feature map of $k$ in $L^2(d\rho_q)$ which is defined as:*

$$\varphi(x,u)=(2q)^{d/4}\exp\left(-2\varepsilon^{-1}\|x-u\|_2^2\right)e^{\frac{\varepsilon^{-1}\|u\|_2^2}{q}}.$$

*Moreover we have also:*

$$\begin{aligned}
\varphi(x,u)\varphi(y,u)&=(2q)^{d/2}\exp\left(-2\varepsilon^{-1}\|x-u\|_2^2\right)\exp\left(-2\varepsilon^{-1}\|y-u\|_2^2\right)e^{\frac{2\varepsilon^{-1}\|u\|_2^2}{q}}\\
&=(2q)^{d/2}\exp\left(-\varepsilon^{-1}\|x-y\|_2^2\right)\exp\left(-4\varepsilon^{-1}\left\|u-\left(\frac{x+y}{2}\right)\right\|_2^2\right)e^{\frac{2\varepsilon^{-1}\|u\|_2^2}{q}}
\end{aligned}$$

*Therefore we have:*

$$\begin{aligned}
\frac{\varphi(x,u)\varphi(y,u)}{k(x,y)}&=(2q)^{d/2}\exp\left(-4\varepsilon^{-1}\left\|u-\left(\frac{x+y}{2}\right)\right\|_2^2\right)e^{\frac{2\varepsilon^{-1}\|u\|_2^2}{q}}\\
&=(2q)^{d/2}\exp\left(-4\varepsilon^{-1}\left(1-\frac{1}{2q}\right)\left\|u-\left(1-\frac{1}{2q}\right)\left(\frac{x+y}{2}\right)\right\|_2^2\right)\\
&\quad\exp\left(\frac{4\varepsilon^{-1}}{2q-1}\left\|\left(\frac{x+y}{2}\right)\right\|_2^2\right)
\end{aligned}$$

*Finally by choosing*

$$q=\frac{\varepsilon^{-1}R^2}{2dW\left(\frac{\varepsilon^{-1}R^2}{d}\right)}$$

*where $W$ is the positive real branch of the Lambert function, we obtain that for any $x,y\in\mathcal{B}(0,R)$:*

$$0\leq\frac{\varphi(x,u)\varphi(y,u)}{k(x,y)}\leq 2\times(2q)^{d/2}$$

*Moreover we have:*

$$\varphi(x,u) = (2q)^{d/4} \exp\left(-2\varepsilon^{-1}\|x-u\|_2^2\right) e^{\frac{\varepsilon^{-1}\|u\|_2^2}{q}}$$

*Therefore $\varphi$ is differentiable with respect to $x$ and we have:*

$$\begin{aligned}\|\nabla_x \varphi\|_2^2 &= 4\varepsilon^{-2}\|x-u\|_2^2 \varphi(x,u)^2 \\ &\leq 4\varepsilon^{-2}\psi \sup_{x\in\mathcal{X}} k(x,x)\|x-u\|_2^2\end{aligned}$$

*where $\psi = 2\times(2q)^{d/2}$. But by definition of the kernel we have $\sup_{x\in\mathcal{B}(0,R)} k(x,x) = 1$ and finally we have that for all $x \in \mathcal{B}(0,R)$:*

$$\mathbb{E}(\|\nabla_x \varphi\|_2^2) \leq 4\varepsilon^{-2}\psi\left[R^2 + \frac{q}{4\varepsilon^{-1}}\right]$$

### 4.5.5 Another example: Arc-cosine kernel

**Lemma 3.** *Let $d \geq 1$, $s \geq 0$, $\kappa > 0$ and $k_{s,\kappa}$ be the perturbed arc-cosine kernel on $\mathbb{R}^d$ defined as for all $x, y \in \mathbb{R}^d$, $k_{s,\kappa}(x,y) = k_s(x,y) + \kappa$. Let also $\sigma > 1$, $\rho = \mathcal{N}(0, \sigma^2 Id)$ and let us define for all $x, u \in \mathbb{R}^d$ the following map:*

$$\varphi(x,u) = \left(\sigma^{d/2}\sqrt{2}\max(0,u^T x)^s \exp\left(-\frac{\|u\|^2}{4}\left[1-\frac{1}{\sigma^2}\right]\right), \sqrt{\kappa}\right)^T$$

*Then for any $x, y \in \mathbb{R}^d$ we have:*

$$k_{s,\kappa}(x,y) = \int_{u\in\mathbb{R}^d} \varphi(x,u)^T \varphi(y,u) d\rho(u)$$

*Moreover we have for all $x, y \in \mathbb{R}^d$ $k_{s,\kappa}(x,y) \geq \kappa > 0$ and for any compact $\mathcal{X} \subset \mathbb{R}^d$ we have:*

$$\sup_{u\in\mathbb{R}^d} \sup_{(x,y)\in\mathcal{X}\times\mathcal{X}} \left|\frac{\varphi(x,u)\varphi(y,u)}{k(x,y)}\right| < +\infty \quad \text{and} \quad \sup_{x\in\mathcal{X}} \mathbb{E}(\|\nabla_x\varphi\|_2^2) < +\infty$$

**Proof.** *Let $s \geq 0$. From [182], we have that:*

$$k_s(x,y) = \int_{\mathbb{R}^d} \Theta_s(u^T x)\Theta_s(u^T y) \frac{e^{-\frac{\|u\|_2^2}{2}}}{(2\pi)^{d/2}} du$$

*where $\Theta_s(w) = \max(0,w)^s$. Let $\sigma > 1$ and $f_\sigma$ the probability density function associated with the distribution $\mathcal{N}(0,\sigma^2 Id)$. Therefore we have that*

$$\begin{aligned}k_s(x,y) &= \int_{\mathbb{R}^d} \Theta_s(u^T x)\Theta_s(u^T y) \frac{e^{-\frac{\|u\|_2^2}{2}}}{(2\pi)^{d/2}} \frac{f_\sigma(u)}{f_\sigma(u)} du \\ &= \sigma^d \int_{\mathbb{R}^d} \Theta_s(u^T x)\Theta_s(u^T y) \exp\left(-\frac{\|u\|^2}{2}\left[1-\frac{1}{\sigma^2}\right]\right) d\rho(u)\end{aligned}$$

*where* $\rho = \mathcal{N}(0, \sigma^2 Id)$. *And by defining for all* $x, u \in \mathbb{R}^d$ *the following map:*

$$\varphi(x,u) = \left(\sigma^{d/2}\sqrt{2}\max(0,u^T x)^s \exp\left(-\frac{\|u\|^2}{4}\left[1-\frac{1}{\sigma^2}\right]\right), \sqrt{\kappa}\right)^T$$

*we obtain that any* $x, y \in \mathbb{R}^d$*:*

$$\begin{aligned}\int_{u\in\mathbb{R}^d} \varphi(x,u)^T\varphi(y,u)d\rho(u) &= \kappa + \sigma^d \int_{\mathbb{R}^d} \Theta_s(u^T x)\Theta_s(u^T y)\exp\left(-\frac{\|u\|^2}{2}\left[1-\frac{1}{\sigma^2}\right]\right)d\rho \\ &= \kappa + k_s(x,y) \\ &= k_{s,\kappa}(x,y)\end{aligned}$$

*Moreover from the definition of the feature map* $\varphi$*, it is clear that* $k_{s,\kappa} \geq \kappa > 0$*,*

$$\sup_{u\in\mathbb{R}^d}\sup_{(x,y)\in\mathcal{X}\times\mathcal{X}}\left|\frac{\varphi(x,u)\varphi(y,u)}{k(x,y)}\right| < +\infty \quad \textit{and} \quad \sup_{x\in\mathcal{X}}\mathbb{E}(\|\nabla_x\varphi\|_2^2) < +\infty.$$

## 4.6 Constructive Method: Differentiability

### 4.6.1 Proof of Proposition 4.3.2

**Proof.** *Let us first introduce the following Lemma:*

**Lemma 4.** *Let* $(\alpha^*, \beta^*)$ *solution of (4.5), then we have*

$$\begin{aligned}\max_i \alpha_i^* - \min_i \alpha_i^* &\leq \varepsilon R(K) \\ \max_j \beta_j^* - \min_j \beta_j^* &\leq \varepsilon R(K)\end{aligned}$$

*where* $R(K) = -\log\left(\iota\frac{\min_{i,j} K_{i,j}}{\max_{i,j} K_{i,j}}\right)$ *with* $\iota := \min_{i,j}(a_i, b_j)$.

**Proof 4.6.1.** *Indeed at optimality, the primal-dual relationship between optimal variables gives us that for all* $i = 1, ..., n$*:*

$$e^{\alpha_i^*/\varepsilon}\langle K_{i,:}, e^{\beta^*/\varepsilon}\rangle = a_i \leq 1$$

*Moreover we have that*

$$\min_{i,j} K_{i,j}\langle \mathbf{1}, e^{\beta^*/\varepsilon}\rangle \leq \langle K_{i,:}, e^{\beta^*/\varepsilon}\rangle \leq \max_{i,j} K_{i,j}\langle \mathbf{1}, e^{\beta^*/\varepsilon}\rangle$$

*Therefore we obtain that*

$$\max_i \alpha_i^* \leq \varepsilon \log\left(\frac{1}{\min_{i,j} K_{i,j}\langle \mathbf{1}, e^{\beta^*/\varepsilon}\rangle}\right)$$

*and*

$$\min_i \alpha_i^* \geq \varepsilon \log \left( \frac{\iota}{\langle \mathbf{1}, e^{\beta^*/\varepsilon} \rangle \max_{i,j} K_{i,j}} \right)$$

*Therefore we obtain that*

$$\max_i \alpha_i^* - \min_i \alpha_i^* \geq -\varepsilon \log \left( \iota \frac{\min_{i,j} K_{i,j}}{\max_{i,j} K_{i,j}} \right)$$

*An analogue proof for $\beta^*$ leads to similar result.*

*Let us now define for any $K \in (\mathbb{R}_+^*)^{n \times m}$ with positive entries the following objective function:*

$$F(K, \alpha, \beta) := \langle \alpha, a \rangle + \langle \beta, a \rangle - \varepsilon (e^{\alpha/\varepsilon})^T K e^{\beta/\varepsilon}.$$

*Let us first show that*

$$G(K) := \sup_{(\alpha,\beta) \in \mathbb{R}^n \times \mathbb{R}^m} F(K, \alpha, \beta) \tag{4.29}$$

*is differentiable on $(\mathbb{R}_+^*)^{n \times m}$. For that purpose let us introduce for any $\gamma_1, \gamma_2 > 0$, the following objective function:*

$$G_{\gamma_1,\gamma_2}(K) := \sup_{\substack{(\alpha,\beta) \in B_\infty^n(0,\gamma_1) \times B_\infty^m(0,\gamma_2) \\ \alpha^T e_1 = 0}} F(K, \alpha, \beta)$$

*where $B_\infty^n(0, \gamma)$ denote the ball of radius $\gamma$ according to the infinite norm and $e_1 = (1, 0, ....0)^T \in \mathbb{R}^n$. In the following we denote by*

$$S_{\gamma_1,\gamma_2} := \left\{ (\alpha, \beta) \in B_\infty^n(0, \gamma_1) \times B_\infty^m(0, \gamma_2) \quad : \quad \alpha^T e_1 = 0 \right\}.$$

*Let us now introduce the following Lemma:*

**Lemma 5.** *Let $\varepsilon > 0$, $(a, b) \in \Delta_n \times \Delta_m$, $K \in (\mathbb{R}_+^*)^{n \times m}$ with positive entries. Then*

$$\max_{\alpha \in \mathbb{R}^n, \beta \in \mathbb{R}^m} a^T \alpha + b^T \beta - \varepsilon (e^{\alpha/\varepsilon})^T K e^{\beta/\varepsilon}$$

*admits a unique solution $(\alpha^*, \beta^*)$ such that $\alpha^T e_1 = 0$, $\|\alpha^*\|_\infty \leq \varepsilon R_1(K)$, and , $\|\beta^*\|_\infty \leq \varepsilon [R_1(K) + R_2(K)]$ where $R_1(K) = -\log \left( \iota \frac{\min_{i,j} K_{i,j}}{\max_{i,j} K_{i,j}} \right)$, $R_2(K) = \log \left( n \frac{\max_{i,j} K_{i,j}}{\iota} \right)$ and $\iota := \min_{i,j}(a_i, b_j)$.*

**Proof 4.6.2.** *In fact the existence and uncity up to a scalar transformation is a well known result. See for example [112]. Therefore there is a unique solution* $(\alpha^0, \beta^0)$ *such that* $(\alpha^0)^T e_1 = 0$. *Moreover thanks to Lemma 4, we have that for any* $(\alpha^*, \beta^*)$ *optimal solution that*

$$\max_i \alpha_i^* - \min_i \alpha_i^* \leq \varepsilon R(K) \tag{4.30}$$

$$\max_j \beta_j^* - \min_j \beta_j^* \leq \varepsilon R(K) \tag{4.31}$$

*Therefore we have* $\|\alpha^0\|_\infty \leq \max_i \alpha_i^0 - \min_i \alpha_i^0 \leq \varepsilon R(K)$. *Moreover, the first order optimality conditions for the dual variables* $(\alpha, \beta)$ *implies that for all* $j = 1, .., m$

$$\beta_j^0 = -\varepsilon \log \left( \sum_{i=1}^n \frac{K_{i,j}}{b_j} \exp\left(\frac{\alpha_i^0}{\varepsilon}\right) \right)$$

*Therefore we have that:*

$$\|\beta^0\|_\infty \leq \|\alpha^0\|_\infty + \varepsilon \log \left( n \frac{\max_{i,j} K_{i,j}}{\iota} \right)$$

*and the result follows.*

*Let* $K_0 \in (\mathbb{R}_+^*)^{n \times m}$, *and let us denote* $M_0 = \max_{i,j} K_0[i,j]$, $m_0 = \min_{i,j} K_0[i,j]$ *and*

$$A_\omega := \left\{ K \in (\mathbb{R}_+^*)^{n \times m} \quad \text{such that} \quad \|K - K_0\|_\infty < \omega \right\}$$

*By considering* $\omega_0 = \frac{m_0}{2}$, *we obtain that for any* $K \in A_{\omega_0}$,

$$R_1(K) \leq \log \left( \frac{1}{\iota} \frac{2M_0 + m_0}{m_0} \right)$$

$$R_2(K) \leq \log \left( n \frac{2M_0 + m_0}{2\iota} \right)$$

*Therefore by denoting*

$$\gamma_1^0 = \varepsilon \log \left( \frac{1}{\iota} \frac{2M_0 + m_0}{m_0} \right)$$

$$\gamma_2^0 = \varepsilon \left[ \log \left( \frac{1}{\iota} \frac{2M_0 + m_0}{m_0} \right) + \log \left( n \frac{2M_0 + m_0}{2\iota} \right) \right]$$

*Therefore, from Lemma 5, we have that for all* $K \in A_{\omega_0}$ *there exists a unique optimal solution* $(\alpha, \beta) \in B_\infty^n(0, \gamma_1^0) \times B_\infty^m(0, \gamma_2^0)$ *satisfying* $\alpha^T e_1 = 0$. *Therefore we have first that for all* $K \in A_{\omega_0}$

$$G_{\gamma_1^0, \gamma_2^0}(K) = G(K) \tag{4.32}$$

*and moreover for all $K \in A_{\omega_0}$, the following set*

$$Z_K := \left\{ (\alpha, \beta) \in S_{\gamma_1^0, \gamma_2^0} \quad \textit{such that} \quad F(K, \alpha, \beta) = \sup_{(\alpha,\beta) \in S_{\gamma_1^0, \gamma_2^0}} F(K, \alpha, \beta) \right\}$$

*is a singleton. Let us now consider the restriction of $F$ on $A_{\omega_0} \times S_{\gamma_1^0, \gamma_2^0}$ denoted $F_0$. It is clear from their definition that $A_{\omega_0}$ is an open convex set, and $S_{\gamma_1^0, \gamma_2^0}$ is compact. Moreover $F_0$ is clearly continuous, and for any $(\alpha, \beta) \in S_{\gamma_1^0, \gamma_2^0}$, $F_0(\cdot, \alpha, \beta)$ is convex. Moreover for any $K \in A_{\omega_0}$ the set $Z_K$ is a singleton, therefore from Danskin theorem [191], we deduce that $G_{\gamma_1^0, \gamma_2^0}$ is convex and differentiable on $A_{\omega_0}$ and we have for all $K \in A_{\omega_0}$*

$$\nabla G_{\gamma_1^0, \gamma_2^0}(K) = -\varepsilon e^{\alpha^*/\varepsilon} (e^{\beta^*/\varepsilon})^T \tag{4.33}$$

*where $(\alpha^*, \beta^*) \in Z_K$. Note that any solutions of Eq.(4.29) can be used to evaluated $\nabla G_{\gamma_1^0, \gamma_2^0}(K)$. Moreover thanks to Eq.(4.32), we deduce also that $G$ is also differentiable on $A_{\omega_0}$. Finally the reasoning hold for any $K_0 \in (\mathbb{R}_+^*)^{n \times m}$, therefore $G$ is differentiable and we have:*

$$\nabla G(K) = -\varepsilon e^{\alpha^*/\varepsilon} (e^{\beta^*/\varepsilon})^T \tag{4.34}$$

## 4.7 Illustrations and Experiments

In Figure 4.5, we show the time-accuracy tradeoff in the high dimensional setting. Here the samples are taken from the higgs dataset[1] [192] where the sample lives in $\mathbb{R}^{28}$. This dataset contains two class of signals: a signal process which produces Higgs bosons and a background process which does not. We take randomly 5000 samples from each of these two distributions.

[1] https://archive.ics.uci.edu/ml/datasets/HIGGS

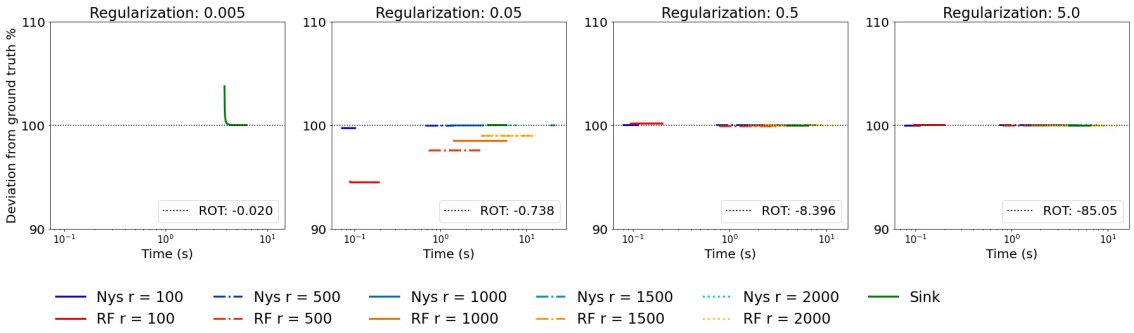



Figure 4.5: In this experiment, we take randomly 10000 samples from the two distributions of the higgs dataset and we plot the deviation from ground truth for different regularizations. We compare the results obtained for our proposed method (**RF**) with the one proposed in [122] (**Nys**) and with the Sinkhorn algorithm (**Sin**) proposed in [112]. The cost function considered here is the square Euclidean metric and the feature map used is that presented in Lemma 1. The number of random features (or rank) chosen varies from 100 to 2000. We repeat for each problem 10 times the experiment. Note that curves in the plot start at different points corresponding to the time required for initialization. *Right, middle right*: when the regularization is sufficiently large both **Nys** and **RF** methods obtain very high accuracy with order of magnitude faster than **Sin**. *Middle left*: both methods manage to obtain high accuracy of the ROT with order of magnitude faster than **Sin**. Note that **Nys** performs better in this setting than our proposed method. *Left*: both methods fail to obtain a good approximation of the ROT.

In Figure 4.6, we consider a discretization of the positive sphere using $50^2 = 2,500$ points and generate three simple histograms of blurred pixels located in the three corners of the simplex.

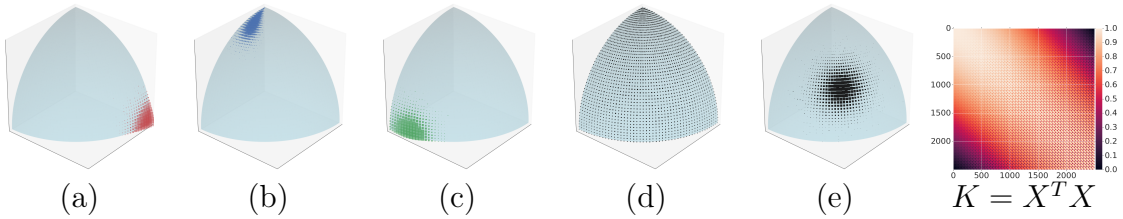



Figure 4.6: Using a discretization of the positive sphere with $50^2 = 2,500$ points we generate three simple histograms (a,b,c) located in the three corners of the simplex. (d) Wasserstein barycenter with a cost $c(x, y) = -\log(x^T y)$ using the method by [127]. (e) Soft-max with temperature 1000 of that barycenter (strongly increasing the relative influence of peaks) reveals that mass is concentrated in areas that would make sense from the more usual $c(x, y) = \arccos x^T y$ distance on the sphere. The kernel corresponding to that cost, here the simple outer product of a matrix $X$ of dimsension $3 \times 2500$.

# Chapter 5

# Low-Rank Optimal Transport: an Algorithmic Approach

Several recent applications of optimal transport (OT) theory to machine learning have relied on regularization, notably entropy and the Sinkhorn algorithm. Because matrix-vector products are pervasive in the Sinkhorn algorithm, several works have proposed to *approximate* kernel matrices appearing in its iterations using low-rank factors. Another route lies instead in imposing low-nonnegative rank constraints on the feasible set of couplings considered in OT problems, with no approximations on cost nor kernel matrices. This route was first explored by Forrow et al. [124], who proposed an algorithm tailored for the squared Euclidean ground cost, using a proxy objective that can be solved through the machinery of regularized 2-Wasserstein barycenters. Building on this, we introduce in this work a generic approach that aims at solving, in full generality, the OT problem under low-nonnegative rank constraints with arbitrary costs. Our algorithm relies on an explicit factorization of low-rank couplings as a product of *sub-coupling* factors linked by a common marginal; similar to an NMF approach, we alternatively updates these factors. We prove the non-asymptotic stationary convergence of this algorithm and illustrate its efficiency on benchmark experiments.

This chapter is based on [3].

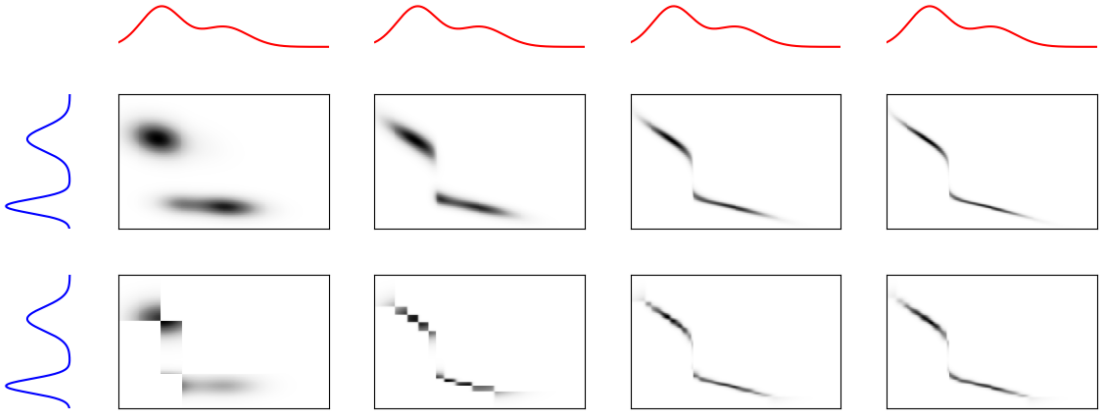


Figure 5.1: Two Gaussian mixture densities evaluated on $n = 200$ and $m = 220$ sized grids in 1D, displayed as blue/red curves. Between them, $n \times m$ optimal coupling matrices obtained by our proposed low-rank OT method for varying rank constraint values $r$ (in increasing order, bottom row) and the Sinkhorn algorithm, for various $\varepsilon$ (in decreasing order, top row). The ground cost is the 1.5-norm.

## 5.1 Introduction

By providing a simple and comprehensive framework to compare probability distributions, optimal transport (OT) theory has inspired many developments in machine learning [33]. A flurry of works have recently connected it to other trending topics, such as normalizing flows or convex neural networks [34, 35, 36], while the scope of its applications has now reached several fields of science such as single-cell biology [41, 56], imaging [42, 43] or neuroscience [44, 45].

**Challenges when computing OT.** Solving optimal transport problems at scale poses, however, formidable challenges. The most obvious among them is computational: Instantiating the Kantorovich [32] problem on discrete measures of size $n$ can be solved with a linear program (LP) of complexity $O(n^3 \log n)$. A second and equally important challenge lies in the statistical performance of using that LP to estimate OT between densities: the LP solution between i.i.d samples converges very slowly to that between densities [71]. It is now increasingly clear that regularizing OT in some way or another is the only way to mitigate these two issues [51, 74, 75]. A popular approach consists in penalizing the OT problem with a strongly convex function of the coupling [76, 77]. We explore in this work an alternative, and more direct approach to add regularity: we restrict, instead, the set of feasible couplings to have a small nonnegative rank.

**Low-Rank Kernel Factorization.** Low-rank factorizations are not new to regularized OT. They have been used to speed-up the resolution of entropy regularized OT with the Sinkhorn algorithm, pending some approximations: Given a data-dependent $n \times m$ cost matrix $C$, the Sinkhorn iterations consist in matrix-vector products of the form $Kv$ or $K^T u$ where $K := \exp(-C/\varepsilon)$ and $u, v$ are $n, m$- vectors. Altschuler et al. [122] and Altschuler and Boix-Adsera [123] have proposed to approximate the kernel $K$ with a product of thin rank $r$ matrices, $\widetilde{K} = AB^T$. Naturally, the ability to approximate $K$ with a low-rank $\widetilde{K}$ degrades as $\varepsilon$ decreases, making this approach valid only for sufficiently large $\varepsilon$. Thanks to this approximation, however, each Sinkhorn iteration is linear in $n$ or $m$, and the coupling outputted by the Sinkorn algorithm is of the form $\widetilde{P} = CD^T$ where $C = \mathrm{diag}(u)A$, $D = \mathrm{diag}(v)B$. This approximation results therefore in a *low-rank* solution that is not, however, rigorously optimal for the original problem as defined by $K$ but rather that defined by $\widetilde{K}$. However the solution obtained with $\widetilde{K}$ can be arbitrary close to the true solution by increasing the rank considered. Similarly, Scetbon and Cuturi [6] consider instead *nonnegative low-rank* approximations for $K$ of the form $\widetilde{K} = QR^T$ where $Q, R > 0$.

**Low-Nonnegative Rank Couplings.** To our knowledge, only Forrow et al. [124] have used low rank considerations for couplings, rather than costs or kernels. Their work studies the case where the ground cost is the squared Euclidean distance. They obtain for that cost a proxy for rank-constrained OT problems using 2-Wasserstein barycenters [125]. Their algorithm blends those in [126, 127] and results in an intuitive mass transfer plan that goes through a small number of $r$ points, where $r$ is the coupling's nonnegative rank.

**Our Contributions.** In this work, we tackle directly the low-rank problem formulated by [124] but make no assumption on the cost matrix; we address instead the low-rank OT problem in its full generality. We consider couplings $P = Q\,\mathrm{diag}(1/g)R^T$ decomposed as the product of two sub-couplings $Q, R$, with common right marginal $g$, and left-marginal given by those of $P$ on each side. Each of these sub-couplings minimizes a transport cost that involves the original cost matrix $C$ and the other sub-coupling. We handle this problem by optimizing jointly on $Q$, $R$ and $g$ using a mirror-descent approach. We prove the non-asymptotic stationary convergence of this approach. In addition, we show that the time complexity of our algorithm can become linear when exploiting low rank assumptions on the *cost* (not the kernel) involved in the OT problem.

**Differences with previous work.** Our approach borrows ideas from [124] but is generic as it applies to all ground costs. Our approach constrains the non-negative

rank of the coupling solution $P$ by construction, rather than relying on a low rank approximation $\widetilde{K}$ for kernel $K = e^{-C/\varepsilon}$. This is a crucial point, because the ability to approximate $K$ with a low rank $\widetilde{K}$ significantly degrades as $\varepsilon$ decreases. By contrast, our approach applies to all ranks, small and large. Interestingly, we also show that a low-rank assumption on the cost matrix (not on the kernel) can also be leveraged, providing therefore a "best of both worlds" scenario in which both the *coupling*'s and the *cost*'s (not the kernel) low rank properties can be enforced and exploited. Finally, a useful parallel can be drawn between our approach and that of the vanilla Sinkhorn algorithm, in the sense that they propose different regularization schemes. Indeed, the (discrete) path of solutions obtained by our algorithm when varying $r$ between 1 and $\min(n, m)$ can be seen as an alternative to the entropic regularization path. Both paths contain at their extremes the original OT solution (maximal rank and minimal entropy) and the product of marginals (minimal rank and maximal entropy), as illustrated in Fig. 5.1.

## 5.2 Discrete Optimal Transport

**OT as a linear program.** Let $a$ and $b$ be two histograms in $\Delta_n, \Delta_m$, the probability simplices of respective size $n, m$. Assuming $a > 0$ and $b > 0$, set $X := \{x_1, \dots, x_n\}$ and $Y := \{y_1, \dots, y_m\}$ two families of points taken each within arbitrary sets, and define discrete distributions $\mu := \sum_{i=1}^n a_i \delta_{x_i}$ and $\nu := \sum_{j=1}^m b_j \delta_{y_j}$. The set of couplings with marginals $a, b$ is:

$$\Pi_{a,b} := \{P \in \mathbb{R}_+^{n \times m} \text{ s.t. } P\mathbf{1}_m = a, P^T\mathbf{1}_n = b\}\,.$$

Given a cost function $c$ defined on pairs of points in $X, Y$ and writing $C := [c(x_i, y_j)]_{i,j}$ its associated matrix, the optimal transport (OT) problem can be written as follows:

$$\mathrm{OT}(\mu, \nu) := \min_{P \in \Pi_{a,b}} \langle C, P \rangle\,. \tag{5.1}$$

**Entropic regularization.** Several works have shown recently [51, 74] that when $X$ and $Y$ are sampled from a continuous space, it is preferable to regularize (5.1) using, for instance, an entropic regularizer [76] to achieve both better computational and statistical efficiency,

$$\mathrm{OT}_\varepsilon(\mu, \nu) := \min_{P \in \Pi_{a,b}} \langle C, P \rangle - \varepsilon H(P)\,. \tag{5.2}$$

where $\varepsilon \geq 0$ and $H$ is the Shannon entropy defined as $H(P) := -\sum_{ij} P_{ij}(\log P_{ij} - 1)$. If $\varepsilon$ goes to 0, one recovers the classical OT problem and for any $\varepsilon > 0$, Eq. (5.2)

becomes $\varepsilon$-strongly convex on $\Pi_{a,b}$ and admits a unique solution $P_\varepsilon$, of the form

$$\exists u^\star \in \mathbb{R}_+^n, v^\star \in \mathbb{R}_+^m \text{ s.t. } P_\varepsilon = \mathrm{diag}(u^\star) K \mathrm{diag}(v^\star) \tag{5.3}$$

where $K := \exp(-C/\varepsilon)$. Cuturi [76] shows that the scaling vectors $u^\star$ and $v^\star$ can be obtained efficiently thanks to the Sinkhorn algorithm (see Alg. 3, where $\odot$ and $/$ denote entry-wise operation). Each iteration can be performed in $\mathcal{O}(nm)$ algebraic operations as it involves only matrix-vector products. The number of Sinkhorn iterations needed to converge to a precision $\delta$ (monitored by the difference between the column-sum of $\mathrm{diag}(u)K\mathrm{diag}(v)$ and $b$) is controlled by the scale of elements in $C$ relative to $\varepsilon$ [164]. That convergence deteriorates with smaller $\varepsilon$, as studied in more detail by [117, 176].

**Algorithm 3** Sinkhorn$(K, a, b, \delta)$

**Inputs:** $K, a, b, \delta, u$

**repeat**

$\quad v \leftarrow b/K^T u,\ u \leftarrow a/Kv$

**until** $\|u \odot Kv - a\|_1 + \|v \odot K^T u - b\|_1 < \delta$;

**Result:** $u, v$

**Mirror descent and $\varepsilon$ schedule.** A possible interpretation of the entropic regularization in the OT problem is that it can be seen as the $k_\varepsilon$-th update of a Mirror Descent (MD) algorithm applied to the objective (5.1) where $k_\varepsilon \geq 1$ depends on $\varepsilon$ and the gradient steps used in the MD. Several works have proposed such links between a gradual decrease in $\varepsilon$ to obtain a better approximation of the unregularized OT problem [193, 119, 194]. More precisely, the MD algorithm associated to the Kullback–Leibler divergence (KL) applied to the objective (5.1) makes for all $k \geq 0$ the following update:

$$Q^{k+1} := \operatorname*{argmin}_{Q \in \Pi_{a,b}} \langle C, Q \rangle + \frac{1}{\gamma_k} \mathrm{KL}(Q, Q_k) \tag{5.4}$$

where $(\gamma_k)_{k \geq 0}$ is a sequence of positive real numbers, $Q_0 \in \Pi_{a,b}$ is an initial point and KL is the Kullback–Leibler divergence defined asw. If $Q_0 := ab^T$, then one obtains that for all $k \geq 0$, updating the coupling according to Eq. (5.4) is the same as solving

$$Q^{k+1} := \operatorname*{argmin}_{Q \in \Pi_{a,b}} \langle C, Q \rangle - \varepsilon_k H(Q)$$

where $\varepsilon_k := (\sum_{j=0}^k \gamma_j)^{-1}$. Therefore the MD algorithm applied to (5.1) produces the sequence $(P_{\varepsilon_k})_{k \geq 0}$ of optimal couplings according to the objective (5.2). We show next that this viewpoint can be applied when one adds also some structures to the couplings considered in the OT problem (5.1), leading to a new regularized approach.

## 5.3 Nonnegative Factorization of the Optimal Coupling

Here we aim at regularizing the OT problem by decomposing the couplings involved into a product of two low-rank couplings. We introduce the associated non-convex problem and develop a mirror-descent algorithm which operates by solving a succession of convex programs.

### 5.3.1 Low-Rank and Factored Couplings

We introduce low rank couplings and explain how they can be parameterized as factored couplings.

**Definition 5.3.1.** *Given $M \in \mathbb{R}_+^{n\times m}$, the nonnegative rank of $M$ is the smallest number of nonnegative rank-one matrices into which the matrix can be decomposed additively:*

$$\mathrm{rk}_+(M) := \min\left\{ q | M = \sum_{i=1}^{q} R_i, \forall i, \mathrm{rk}(R_i) = 1, R_i \geq 0 \right\}.$$

Let $r \geq 1$, and let us denote

$$\Pi_{a,b}(r) := \{P \in \Pi_{a,b}, \mathrm{rk}_+(P) \leq r\}.$$

From Definition 6.2.1, one can write $\Pi_{a,b}(r)$ as

$$\left\{ \sum_{i=1}^{r} g_i q_i r_i^T \text{ s.t. } \forall\ i\ q_i \in \Delta_n,\ r_i \in \Delta_m,\ g \in \Delta_r,\ \sum_{i=1}^{r} g_i q_i = a\ ,\ \sum_{i=1}^{r} g_i r_i = b \right\}$$

from which we deduce directly that $\Pi_{a,b}(r)$ is compact. Moreover for $g \in \Delta_r^* := \{h \in \Delta_r \colon \forall i\ h_i > 0\}$, we write

$$\Pi_{a,g,b} := \left\{ P \in \mathbb{R}_+^{n\times m}, P = Q\,\mathrm{diag}(1/g)R^T,\ Q \in \Pi_{a,g},\ \text{and } R \in \Pi_{b,g} \right\}.$$

Note that $\Pi_{a,g,b}$ is compact and a subset of $\Pi_{a,b}(r)$ since for all $P \in \Pi_{a,g,b}$, $P \in \Pi_{a,b}$ and one has $\mathrm{rk}(P) \leq \mathrm{rk}_+(P) \leq r$. Moreover, for any $P \in \Pi_{a,b}$ such that $\mathrm{rk}_+(P) \leq r$, there exists $g \in \Delta_r^*$, $Q \in \Pi_{a,g}$ and $R \in \Pi_{b,g}$ such that $P = Q\,\mathrm{diag}(1/g)R^T$ [195]. Therefore

$$\bigcup_{g\in\Delta_r^*} \Pi_{a,g,b} = \Pi_{a,b}(r). \tag{5.5}$$

We exploit next this identity to build an efficient algorithm in order to solve the optimal transport problem under low nonnegative rank constraints.

### 5.3.2 The Low-rank OT Problem (LOT)

The problem of interest in this work is:

$$\mathrm{LOT}_r(\mu,\nu) := \min_{P\in\Pi_{a,b}(r)} \langle C, P\rangle. \tag{5.6}$$

Here the minimum is always attained as $\Pi_{a,b}(r)$ is nonempty and compact and the objective is continuous. Thanks to (5.5), problem (5.6) is equivalent to

$$\min_{(Q,R,g)\in\mathcal{C}(a,b,r)} \langle C, Q\,\mathrm{diag}(1/g)R^T\rangle \tag{5.7}$$

where $\mathcal{C}(a,b,r) := \mathcal{C}_1(a,b,r)\cap\mathcal{C}_2(r)$, with

$$\mathcal{C}_1(a,b,r) := \left\{(Q,R,g)\in\mathbb{R}_+^{n\times r}\times\mathbb{R}_+^{m\times r}\times(\mathbb{R}_+^*)^r \text{ s.t. } Q\mathbf{1}_r = a, R\mathbf{1}_r = b\right\}$$
$$\mathcal{C}_2(r) := \left\{(Q,R,g)\in\mathbb{R}_+^{n\times r}\times\mathbb{R}_+^{m\times r}\times\mathbb{R}_+^{r} \text{ s.t. } Q^T\mathbf{1}_n = R^T\mathbf{1}_m = g\right\}.$$

In the following, we also consider regularized version of the problem (6.4) by adding an entropic term to the objective which leads for all $\varepsilon\geq 0$ to the following problem

$$\mathrm{LOT}_{r,\varepsilon}(\mu,\nu) := \inf_{(Q,R,g)\in\mathcal{C}(a,b,r)} \langle C, Q\,\mathrm{diag}(1/g)R^T\rangle - \varepsilon H((Q,R,g)). \tag{5.8}$$

Here the entropy of $(Q,R,g)$ is to be understood as that of the values of the three respective entropies evaluated for each term. We will see that adding an entropic term to the objective allows to stabilize the MD scheme employed to solve (5.6). For all $\varepsilon\geq 0$, the objective function defined in (5.8) is lower semi-continuous, and admits therefore a minimum in $\overline{\mathcal{C}_1(a,b,r)}\cap\mathcal{C}_2(r)$ where $\overline{\mathcal{C}_1(a,b,r)}$ is the closure of $\mathcal{C}_1(a,b,r)$. However, the existence of a solution for problem (5.8) requires more care, as shown in the following proposition.

**Proposition 5.3.1.** *If $\varepsilon = 0$ then the infimum of (5.8) is always attained. If $\varepsilon > 0$, then if $r = 1$, the infimum of (5.8) is attained and for $r\geq 2$, problem (5.8) admits a minimum if $\mathrm{LOT}_{r,\varepsilon}(\mu,\nu) < \mathrm{LOT}_{r-1,\varepsilon}(\mu,\nu)$.*

**Stabilized Formulation using Lower Bounds** In order to ensure stability of the mirror descent method, and enable its theoretical analysis, we introduce a lower bound $\alpha$ on the weight vector $g$. Let us assume in the following that we consider $(r,\varepsilon)$ satisfying the conditions of Proposition 5.3.1. In particular if $\varepsilon = 0$, $r$ can be arbitrarily chosen and we recover the problem defined in (5.6). Under this assumption, there exists $(Q_\varepsilon^*, R_\varepsilon^*, g_\varepsilon^*)\in\mathcal{C}_1(a,b,r)\cap\mathcal{C}_2(r)$ solution of Eq. (5.8) from

which follows the existence of $\frac{1}{r} \geq \alpha^* > 0$, such that $g_\varepsilon^* \geq \alpha^*$ coordinate-wise. Let us now define for any $\frac{1}{r} \geq \alpha > 0$, the following set

$$\mathcal{C}_1(a, b, r, \alpha) := \Big\{(Q, R, g) \in \mathbb{R}_+^{n\times r} \times \mathbb{R}_+^{m\times r} \times \mathbb{R}_+^r \text{ s.t. } Q\mathbf{1}_r = a, R\mathbf{1}_r = b,\ g \geq \alpha\Big\}.$$

Then if $\alpha$ is sufficiently small (i.e. $\alpha \leq \alpha^*$) we have that the problem (5.8) is equivalent to

$$\mathrm{LOT}_{r,\varepsilon,\alpha}(\mu, \nu) = \min_{(Q,R,g)\in\mathcal{C}(a,b,r,\alpha)} \langle C, Q\,\mathrm{diag}(1/g)R^T\rangle - \varepsilon H((Q, R, g)), \tag{5.9}$$

where $\mathcal{C}(a, b, r, \alpha) := \mathcal{C}_1(a, b, r, \alpha) \cap \mathcal{C}_2(r)$. Note that for any $\frac{1}{r} \geq \alpha > 0$, the set of constraints is not empty, compact and the minimum always exists.

### 5.3.3 Mirror Descent Optimization Scheme

**Mirror descent outer loop.** We propose to use a Mirror Descent scheme with a KL divergence to solve Eq. (5.9). It leads, for all $k \geq 0$, to the following updates which necessitate the solution of a convex problem at each step

$$(Q_{k+1}, R_{k+1}, g_{k+1}) := \operatorname*{argmin}_{\boldsymbol{\zeta}\in\mathcal{C}(a,b,r,\alpha)} \mathrm{KL}(\boldsymbol{\zeta}, \boldsymbol{\xi}_k) \tag{5.10}$$

where $(Q_0, R_0, g_0) \in \mathcal{C}(a, b, r, \alpha)$ is an initial point such that $Q_0 > 0$ and $R_0 > 0$, $\boldsymbol{\xi}_k := (\xi_k^{(1)}, \xi_k^{(2)}, \xi_k^{(3)})$, $\xi_k^{(1)} := \exp(-\gamma_k C R_k\,\mathrm{diag}(1/g_k) - (\gamma_k\varepsilon - 1)\log(Q_k))$, $\xi_k^{(2)} := \exp(-\gamma_k C^T Q_k\,\mathrm{diag}(1/g_k) - (\gamma_k\varepsilon - 1)\log(R_k))$, $\xi_k^{(3)} := \exp(\gamma_k\omega_k/g_k^2 - (\gamma_k\varepsilon - 1)\log(g_k))$ with $[\omega_k]_i := [Q_k^T C R_k]_{i,i}$ for all $i \in \{1, \dots, r\}$ and $(\gamma_k)_{k\geq 0}$ is a sequence of positive step sizes. Note that for all $k \geq 0$, $(Q_k, R_k, g_k)$ live in $(\mathbb{R}_+^*)^{n\times r} \times (\mathbb{R}_+^*)^{m\times r} \times (\mathbb{R}_+^*)^r$, and therefore $\boldsymbol{\xi}_k$ is well defined and lives also in $(\mathbb{R}_+^*)^{n\times r} \times (\mathbb{R}_+^*)^{m\times r} \times (\mathbb{R}_+^*)^r$.

**Dykstra's inner loop.** In order to solve Eq. (6.7), we use the Dykstra's Algorithm [196]. Given a closed convex set $\mathcal{C} \subset \mathbb{R}_+^{n\times r} \times \mathbb{R}_+^{m\times r} \times \mathbb{R}_+^r$, we denote for all $\boldsymbol{\xi} \in (\mathbb{R}_+^*)^{n\times r} \times (\mathbb{R}_+^*)^{m\times r} \times (\mathbb{R}_+^*)^r$ the projection according to the Kullback-Leibler divergence as

$$\mathcal{P}_{\mathcal{C}}^{\mathrm{KL}}(\boldsymbol{\xi}) := \operatorname*{argmin}_{\boldsymbol{\zeta}\in\mathcal{C}} \mathrm{KL}(\boldsymbol{\zeta}, \boldsymbol{\xi}).$$

Starting from $\boldsymbol{\zeta}_0 := \boldsymbol{\xi}$ and $\boldsymbol{q}_0 = \boldsymbol{q}_{-1} = (\mathbf{1}, \mathbf{1}, \mathbf{1}) \in \mathbb{R}_+^{n\times r} \times \mathbb{R}_+^{m\times r} \times \mathbb{R}_+^r$, the Dykstra's Algorithm consists in computing for all $j \geq 0$,

$$\begin{aligned}
\boldsymbol{\zeta}_{2j+1} &= \mathcal{P}^{\mathrm{KL}}_{\mathcal{C}_1(a,b,r,\alpha)}(\boldsymbol{\zeta}_{2j} \odot \boldsymbol{q}_{2j-1}) \\
\boldsymbol{q}_{2j+1} &= \boldsymbol{q}_{2j-1} \odot \frac{\boldsymbol{\zeta}_{2j}}{\boldsymbol{\zeta}_{2j+1}} \\
\boldsymbol{\zeta}_{2j+2} &= \mathcal{P}^{\mathrm{KL}}_{\mathcal{C}_2(r)}(\boldsymbol{\zeta}_{2j+1} \odot \boldsymbol{q}_{2j}) \\
\boldsymbol{q}_{2j+2} &= \boldsymbol{q}_{2j} \odot \frac{\boldsymbol{\zeta}_{2j+1}}{\boldsymbol{\zeta}_{2j+2}}.
\end{aligned}$$

As $\mathcal{C}_1(a, b, r, \alpha)$ and $\mathcal{C}_2(r)$ are closed convex subspaces and $\boldsymbol{\xi} \in (\mathbb{R}_+^*)^{n\times r} \times (\mathbb{R}_+^*)^{m\times r} \times (\mathbb{R}_+^*)^r$, one can show that $(\boldsymbol{\zeta}_j)_{j\geq 0}$ converges towards the unique solution of Eq. (6.7), [197]. The following propositions detail how to compute the relevant projections involved in the Dykstra's Algorithm.

**Proposition 5.3.2.** *For $\tilde{\boldsymbol{\xi}} := (\tilde{Q}, \tilde{R}, \tilde{g}) \in (\mathbb{R}_+^*)^{n\times r} \times (\mathbb{R}_+^*)^{n\times r} \times (\mathbb{R}_+^*)^r$, one has, denoting $\hat{g} := \max(\tilde{g}, \alpha)$*

$$\mathcal{P}^{KL}_{\mathcal{C}_1(a,b,r,\alpha)}(\tilde{\boldsymbol{\xi}}) = \left(\operatorname{diag}\left(\frac{a}{\tilde{Q}\mathbf{1}_r}\right)\tilde{Q}, \operatorname{diag}\left(\frac{b}{\tilde{R}\mathbf{1}_r}\right)\tilde{R}, \hat{g}\right).$$

Let us now show the solution of the projection on $\mathcal{C}_2(r)$.

**Proposition 5.3.3.** *For $\tilde{\boldsymbol{\xi}} := (\tilde{Q}, \tilde{R}, \tilde{g}) \in (\mathbb{R}_+^*)^{n\times r} \times (\mathbb{R}_+^*)^{n\times r} \times (\mathbb{R}_+^*)^r$, the projection $(Q, R, g) = \mathcal{P}^{KL}_{\mathcal{C}_2(r)}(\tilde{\boldsymbol{\xi}})$ satisfies*

$$Q = \tilde{Q}\operatorname{diag}(g/\tilde{Q}^T\mathbf{1}_n), \quad R = \tilde{R}\operatorname{diag}(g/\tilde{R}^T\mathbf{1}_m), \quad g = (\tilde{g} \odot \tilde{Q}^T\mathbf{1}_n \odot \tilde{R}^T\mathbf{1}_m)^{1/3}.$$

**Efficient computation of the updates.** The projection obtained in Proposition 5.3.2, 5.3.3 lead to simple updates of the couplings. Indeed, starting with $\boldsymbol{\zeta}_0 := \boldsymbol{\xi} = (\xi^{(1)}, \xi^{(2)}, \xi^{(3)})$ the Dysktra's Algorithm applied to our problem (6.7) needs only to compute scaling vectors as presented in Alg. 4. More precisely, the Dykstra's Algorithm produces the iterates $(\boldsymbol{\zeta}_j)_{j\geq 0}$ which satisfy for all $j \geq 0$ $\boldsymbol{\zeta}_j = (Q_j, R_j, g_j)$ where

$$Q_j = \operatorname{diag}(u_j^1)\xi^{(1)}\operatorname{diag}(v_j^1), \quad R_j = \operatorname{diag}(u_j^2)\xi^{(2)}\operatorname{diag}(v_j^2)$$

for the sequences $(u_j^i, v_j^i)_{j\geq 0}$ initialized as, $u_0^i := \mathbf{1}_n$, $v_0^i := \mathbf{1}_m$ for all $i \in \{1,2\}$, $q_{0,1}^{(3)} = q_{0,2}^{(3)} = q_0^{(1)} = q_0^{(2)} = \mathbf{1}_r$ and computed with the iterations

$$
\begin{aligned}
u_{n+1}^{k,i} &= \frac{p_i}{\xi_k^i v_n^{k,i}} \\
\tilde{g}_{n+1} &= \max(\alpha, g_n \odot q_{n,1}^{(3)}),\ q_{n+1,1}^{(3)} = (g_n \odot q_{n,1}^{(3)})/\tilde{g}_{n+1} \\
g_{n+1} &= (\tilde{g}_{n+1} \odot q_{n,2}^{(3)})^{1/3} \prod_{i=1}^{2} (v_n^{k,i} \odot q_n^{(i)} \odot (\xi_k^i)^T u_n^{k,i})^{1/3} \\
v_{n+1}^{k,i} &= \frac{g_{n+1}}{(\xi_k^i)^T u_n^{k,i}} \\
q_{n+1}^{(i)} &= (v_n^{k,i} \odot q_n^{(i)})/v_{n+1}^{k,i},\ q_{n+1,2}^{(3)} = (\tilde{g}_{n+1} \odot q_{n,2}^{(3)})/g_{n+1}
\end{aligned}
$$

where we have denoted $p_1 := a$ and $p_2 := b$ to simplify the notations.

**Algorithm 4** LR-Dykstra$((\xi^{(i)})_{1\leq i\leq 3}, p_1, p_2, \alpha, \delta)$

**Inputs:** $\xi^{(1)}, \xi^{(2)}, \tilde{g} := \xi^{(3)}, p_1, p_2, \alpha, \delta, q_1^{(3)} = q_2^{(3)} = \mathbf{1}_r, \forall i \in \{1,2\},\ \tilde{v}^{(i)} = \mathbf{1}_r, q^{(i)} = \mathbf{1}_r$

**repeat**

$u^{(i)} \leftarrow p_i/\xi^{(i)}\tilde{v}^{(i)}\ \forall i \in \{1,2\}$,

$g \leftarrow \max(\alpha, \tilde{g} \odot q_1^{(3)}),\ q_1^{(3)} \leftarrow (\tilde{g} \odot q_1^{(3)})/g,\ \tilde{g} \leftarrow g$,

$g \leftarrow (\tilde{g} \odot q_2^{(3)})^{1/3} \prod_{i=1}^{2} (v^{(i)} \odot q^{(i)} \odot (\xi^{(i)})^T u^{(i)})^{1/3}$,

$v^{(i)} \leftarrow g/(\xi^{(i)})^T u^{(i)}\ \forall i \in \{1,2\}$,

$q^{(i)} \leftarrow (\tilde{v}^{(i)} \odot q^{(i)})/v^{(i)}\ \forall i \in \{1,2\},\ q_2^{(3)} \leftarrow (\tilde{g} \odot q_2^{(3)})/g$,

$\tilde{v}^{(i)} \leftarrow v^{(i)}\ \forall i \in \{1,2\},\ \tilde{g} \leftarrow g$

**until** $\sum_{i=1}^{2} \|u^{(i)} \odot \xi^{(i)} v^{(i)} - p_i\|_1 < \delta$;

$Q \leftarrow \mathrm{diag}(u^{(1)})\xi^{(1)}\,\mathrm{diag}(v^{(1)})$

$R \leftarrow \mathrm{diag}(u^{(2)})\xi^{(2)}\,\mathrm{diag}(v^{(2)})$

**Result:** $Q, R, g$

Let us now introduce the proposed MD algorithm applied to (5.9). By denoting $\mathcal{D}(\cdot)$ the operator extracting the diagonal of a square matrix we obtain Alg. 5.

**Algorithm 5** LOT$(C, a, b, r, \alpha, \delta)$

**Inputs:** $C, a, b, (\gamma_k)_{k\geq 0}, Q, R, g, \alpha, \delta$

**for** $k = 1, \dots$ **do**

$\xi^{(1)} \leftarrow \exp(-\gamma_k CR\,\mathrm{diag}(1/g) - (\gamma_k\varepsilon - 1)\log(Q))$,

$\xi^{(2)} \leftarrow \exp(-\gamma_k C^T Q\,\mathrm{diag}(1/g) - (\gamma_k\varepsilon - 1)\log(R))$,

$\omega \leftarrow \mathcal{D}(Q^T CR)$,

$\xi^{(3)} \leftarrow \exp(\gamma_k\omega/g^2 - (\gamma_k\varepsilon - 1)\log(g))$,

$Q, R, g \leftarrow$ LR-Dykstra$((\xi^{(i)})_{1\leq i\leq 3}, a, b, \alpha, \delta)$ (Alg. 4)

**end**

**Result:** $\langle C, Q\,\mathrm{diag}(1/g)R^T\rangle$

**Computational Cost.** Note that $(\xi^{(i)})_{1\leq i\leq 3}$ considered in Alg. 5 live in $\mathbb{R}_+^{n\times r} \times \mathbb{R}_+^{m\times r} \times \mathbb{R}_+^r$ and therefore given those matrices, each iteration of Alg. 4 requires $\mathcal{O}((n+m)r)$ algebraic operations, since it involves only matrix/vector multiplications of the form $\xi^{(i)}v_i$ and $(\xi^{(i)})^T u_i$. However without any assumption on the cost matrix $C$, computing $(\xi^{(i)})_{1\leq i\leq 3}$ requires $\mathcal{O}(nmr)$ algebraic operations since $CR$ and $C^TQ$ must be evaluated. We show in §5.3.5 how to reduce the quadratic cost of computing $(\xi^{(i)})_{1\leq i\leq 3}$ to a linear cost with respect to the number of samples if one assumes that the considered *cost* matrix can be factored, either exactly (ensured with a squared Euclidean distance cost) or approximately if that cost is a distance. Writing $N$ the number of iterations of the MD scheme and $T$ the number of iterations considered in Algorithm 4 at each step of the MD, we end up with a total computational cost of $\mathcal{O}(NT(n+m)r + Nnmr)$.

**Remark 4.** *Note that our algorithm can be applied in the specific case where $\varepsilon = 0$ in order to solve Eq. (5.6). Moreover, our algorithm can be applied for an arbitrary choice of the cost function. For example in Figure 5.4, we run our algorithm on graphs where the distance considered in the shortest-path distance.*

### 5.3.4 Convergence of the Mirror Descent

Even if the objective (5.9) is not convex in $(Q, R, g)$, we obtain the non-asymptotic stationary convergence of the MD algorithm in this setting. For that purpose we introduce a stronger convergence criterion than the one presented in [198] to obtain non-asymptotic stationary convergence of the MD scheme. Indeed let $F_\varepsilon$ be the objective function of the problem (5.9) defined on $\mathcal{C}(a, b, r, \alpha)$ and let us denotes for any $\gamma > 0$ and $\boldsymbol{\xi} \in \mathcal{C}(a, b, r, \alpha)$

$$\mathcal{G}_{\varepsilon,\alpha}(\boldsymbol{\xi}, \gamma) := \underset{\boldsymbol{\zeta}\in\mathcal{C}(a,b,r,\alpha)}{\mathrm{argmin}} \{\langle \nabla F_\varepsilon(\boldsymbol{\xi}), \boldsymbol{\zeta}\rangle + \frac{1}{\gamma} KL(\boldsymbol{\zeta}, \boldsymbol{\xi})\}.$$

Then the criteron used in [198] to show the stationary convergence of the MD scheme is defined as the square norm of the following vector:

$$P_{\mathcal{C}(a,b,r,\alpha)}(\boldsymbol{\xi},\gamma) := \frac{1}{\gamma}(\boldsymbol{\xi} - \mathcal{G}_{\varepsilon,\alpha}(\boldsymbol{\xi},\gamma)).$$

This vector can be seen as a generalized projected gradient of $F_\varepsilon$ at $\boldsymbol{\xi}$. Indeed if $X = \mathbb{R}^d$ and by replacing the *Bregman Divergence* $\mathrm{KL}(u,x)$ by $\frac{1}{2}\|u-x\|_2^2$, we would have $P_X(x,\gamma) = \nabla F_\varepsilon(x)$. Here we consider instead the following criterion to establish convergence:

$$\Delta_{\varepsilon,\alpha}(\boldsymbol{\xi},\gamma) := \frac{1}{\gamma^2}(\mathrm{KL}(\boldsymbol{\xi}, \mathcal{G}_{\varepsilon,\alpha}(\boldsymbol{\xi},\gamma)) + \mathrm{KL}(\mathcal{G}_{\varepsilon,\alpha}(\boldsymbol{\xi},\gamma), \boldsymbol{\xi})).$$

Such criterion is in fact stronger than the one used in [198] as we have

$$\begin{aligned}\Delta_{\varepsilon,\alpha}(\boldsymbol{\xi},\gamma) &= \frac{1}{\gamma^2}(\langle \nabla h(\mathcal{G}_{\varepsilon,\alpha}(\boldsymbol{\xi},\gamma)) - \nabla h(\boldsymbol{\xi}), \mathcal{G}_{\varepsilon,\alpha}(\boldsymbol{\xi},\gamma) - \boldsymbol{\xi}\rangle \\ &\geq \frac{1}{2\gamma^2}\|\mathcal{G}_{\varepsilon,\alpha}(\boldsymbol{\xi},\gamma) - \boldsymbol{\xi}\|_1^2 \\ &= \frac{1}{2}\|P_{\mathcal{C}(a,b,r,\alpha)}(\boldsymbol{\xi},\gamma)\|_1^2\end{aligned}$$

where $h$ denotes the minus entropy function and the last inequality comes from the strong convexity of $h$ on $\mathcal{C}(a,b,r,\alpha)$. For any $\frac{1}{r} \geq \alpha > 0$, we show in the following proposition the non-asymptotic stationary convergence of the MD scheme applied to the problem (5.9). To prove this result, we show that for any $\varepsilon \geq 0$, the objective is smooth relatively to the negative entropy function [199] and we extend the proof of [198] to this case.

**Proposition 5.3.4.** *Let $\varepsilon \geq 0$, $\frac{1}{r} \geq \alpha > 0$ and $N \geq 1$. By denoting*

$$L_{\varepsilon,\alpha} := \sqrt{3\left(2\frac{\|C\|_2^2}{\alpha^4} + \left(\varepsilon + \frac{2\|C\|_2}{\alpha^3}\right)^2\right)}$$

*and by considering a constant stepsize in the MD scheme (6.7) such that for all $k = 1, \ldots, N$ $\gamma_k = \frac{1}{2L_{\varepsilon,\alpha}}$, we obtain that*

$$\min_{1\leq k\leq N} \Delta_{\varepsilon,\alpha}((Q_k, R_k, g_k), \gamma_k) \leq \frac{4L_{\varepsilon,\alpha}D_0}{N}.$$

*where $D_0 := F_\varepsilon(Q_0, R_0, g_0) - \mathrm{LOT}_{r,\varepsilon,\alpha}$ is the distance of the initial value to the optimal one.*

Thanks to Proposition 7.4.2, for $\alpha$ sufficiently small (i.e. $\alpha \leq \alpha^*$), we have $\mathrm{LOT}_{r,\varepsilon,\alpha} = \mathrm{LOT}_{r,\varepsilon}$ and therefore we obtain a stationary point of (5.8). In particular, if $\varepsilon = 0$, the proposed algorithm converges towards a stationary point of (5.6).

**Remark 5.** *We also propose an algorithm to directly solve (5.8). The main difference is that the updates of the MD can be solved using the Iterative Bregman Projections (IBP) Algorithm. See Appendix 5.10 for more details.*

**Remark 6.** *For all $\varepsilon \geq 0$, the MD scheme implies that each iteration $k$ of our proposed algorithm outputs $(Q_k, R_k, g_k) \in \mathcal{C}_1(a, b, r, \alpha) \cap \mathcal{C}_2(r)$, and therefore the matrix obtained a each iteration $P_k^{LOT} = Q_k \operatorname{diag}(1/g_k) R_k^T$ is a coupling which sastifies the marginal constraints while in the Sinkhorn algorithm, the matrix defined at each iteration by $P_k^{Sin} = \operatorname{diag}(u_k) K \operatorname{diag}(v_k)$ becomes a coupling which satisfies the marginal constraints only at convergence.*

In the following section, we aim at accelerating our method in order to obtain a linear time algorithm to solve (5.8).

### 5.3.5 Linear time approximation of the Low-Rank Optimal Transport

Here we aim at obtaining the optimal solution of Eq. (5.8) in linear time with respect to the number of samples. For that purpose let us introduce our main assumption on the cost matrix $C$.

**Assumption 1.** *Assume that $C$ admits a low-rank factorization, that is there exists $A \in \mathbb{R}^{n \times d}$ and $B \in \mathbb{R}^{m \times d}$ such that $C = AB^T$.*

From the Assumption 1 one can in fact accelerate the computation in the iterations of the proposed Alg. (5) and obtain a linear time algorithm with respect to the number of samples. Indeed recall that given $\boldsymbol{\xi} = (\xi^{(i)})_{1 \leq i \leq 3}$, each iteration of the Dykstra's Alg. (4) can be performed in linear time. Moreover, thanks to Assumption 1, the computation of $\boldsymbol{\xi}$, which requires to compute both $CR$ and $C^TQ$ can be performed in $\mathcal{O}((n+m)dr)$ algebraic operations and thus Alg. (5) requires only a linear number of algebraic operations with respect to the number of samples at each iteration.

Let us now justify why the Assumption 1 of a low-rank factorization for the cost matrix is well suited in the problem of computing the Optimal Transport.

**Squared Euclidean Metric.** In the specific case where $C$ is a Square Euclidean distance matrix, it admits a low-rank decomposition. Indeed let $X := [x_1, \ldots, x_n] \in \mathbb{R}^{d \times n}$, let $Y := [y_1, \ldots, y_m] \in \mathbb{R}^{d \times m}$ and let $D := (\|x_i - y_j\|_2^2)_{i,j}$. Then by denoting

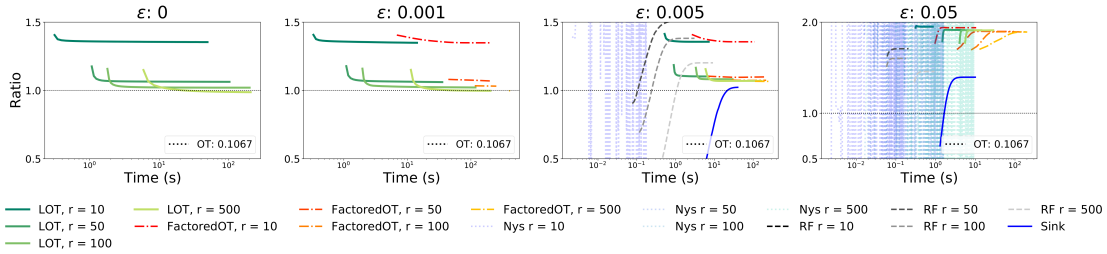



Figure 5.2: In this experiment, we consider two Gaussian distributions evaluated on $n = m = 5000$ in 2D. The first one has a mean of $(1, 1)^T$ and identity covariance matrix $I_2$ while the other has 0 mean and covariance $0.1 \times I_2$. The ground cost is the squared Euclidean distance. Note that for this cost, an exact low-rank factorization of the cost is available, and therefore all low-rank methods, including ours, have a linear time complexity. *Left:* we show that when $\varepsilon = 0$ our method is able to quickly obtain the exact OT by forcing the nonnegative rank of the coupling to be relatively small compared to the number of samples. Note that in this setting, all the other methods cannot be applied. *Middle left, middle right:* In these plots, we show that our method can obtain high accuracy for either estimate the true OT or its regularized version with order of magnitude faster than the other low-rank methods for any rank $r$. Moreover, our methods outperforms **Sin** in these regimes of small regularizations. Note that **Sin** does not converge for $\varepsilon = 0.001$ as we do not consider its stabilized version using log-sum-exp function but rather its classical version which is less costly to compute. *Right:* Here we change the scale of the $y$-axis of the plot. We see that the regime of the entropic regularizations for the Sinkhorn algorithm and our method differs. Indeed, the Sinkhorn algorithm has a larger range of $\varepsilon$ such that it provides an efficient approximation of the OT, whereas **LOT** is regularizing *twice*, namely with respect to both rank *and* entropy.

$p = [\|x_1\|_2^2, \dots, \|x_n\|_2^2]^T \in \mathbb{R}^n$ and $q = [\|y_1\|_2^2, \dots, \|y_m\|_2^2]^T \in \mathbb{R}^m$ we can rewrite $D$ as the following:

$$D = p\mathbf{1}_m^T + \mathbf{1}_n q^T - 2X^T Y.$$

Therefore by denoting $A = [p, \mathbf{1}_n, -2X^T] \in \mathbb{R}^{n\times(d+2)}$ and $B = [\mathbf{1}_m, q, Y^T] \in \mathbb{R}^{n\times(d+2)}$ we obtain that

$$D = AB^T.$$

**General Case: Distance Matrix.** In the following we denote a distance matrix $D \in \mathbb{R}^{n\times m}$, any matrix such that there exists a metric space $(\mathcal{X}, d)$, $\{x_i\}_{i=1}^n \in \mathcal{X}^n$ and $\{y_j\}_{j=1}^m \in \mathcal{X}^m$ which satisfy for all $i, j$, $D_{i,j} = d(x_i, y_j)$. In fact it is always possible to obtain a low-rank approximation of a distance matrix in linear

time. In [200, 201], the authors proposed an algorithm such that for any distance matrix $D \in \mathbb{R}^{n\times m}$ and $\gamma > 0$ it outputs matrices $M \in \mathbb{R}^{n\times d}$, $N \in \mathbb{R}^{m\times d}$ in $\mathcal{O}((m+n)\text{poly}(\frac{d}{\gamma}))$ algebraic operations such that with probability at least 0.99 we have

$$\|D - MN^T\|_F^2 \leq \|D - D_d\|_F^2 + \gamma\|D\|_F^2$$

where $D_d$ denotes the best rank-$d$ approximation to $D$. Therefore one can always obtain a low-rank factorization of a distance matrix in linear time with respect to the number of samples. See Appendix 5.8 for more details.

## 5.4 Numerical Results

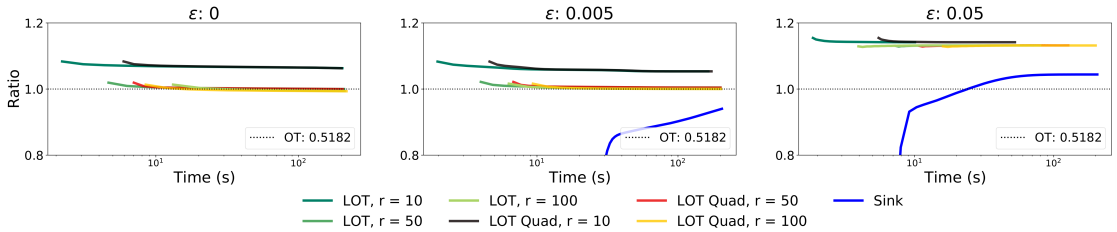



Figure 5.3: Here we consider two Gaussian mixture densities sampled with $n = m = 10000$ points in 2D (See Appendix 5.11 for more details). The ground cost is the Euclidean distance. As this cost is a distance, we can apply our linear version of the algorithm and we denote **LOT Quad** to refer to its quadratic counterpart. We see that **LOT** and **LOT Quad** provide similar results while **LOT** is faster. All kernel-based methods (**Nys**, **RF**) fail to converge in this setting. As in Fig. 5.2, we see that our method is able to approximate faster than **Sin** the true OT thanks to the low-rank constraint.

### 5.4.1 Comparison with other regularization schemes

We consider three synthetic problems in which we study the time-accuracy trade-off as well as the couplings obtained, by comparing our method with other low-rank methods, as well as Sinkhorn's algorithm. More precisely, we compare our proposed method, **LOT**, with the factored Optimal Transport [124], **FactoredOT**, the Nystrom-based method [122], **Nys**, the random features-based method [6], **RF** and the Sinkhorn algorithm [76], **Sin**. For **LOT**, we set the lower bound on $g$ to $\alpha = 10^{-5}$.

**Time-accuracy Tradeoff** We consider two problems where the ground cost involved in the OT problem is either the *squared Euclidean* distance or the *Euclidean* distance. In the first one, we consider measures supported on $n = 5000$ points in $\mathbb{R}^2$, while the second we consider $n = 10000$ samples in $\mathbb{R}^2$. The method proposed by [124] can only be used with the squared Euclidean distance (2-Wasserstein) while ours works for any cost. For all the low-ranks methods, we vary the ranks between 10 and 500. For all the randomized methods, we consider the mean over 10 runs to estimate the OT. In Fig. 5.2, 5.3 we plot the ratio w.r.t. the (non-regularized) optimal transport cost defined as $\mathrm{R} := \langle C, \widetilde{P}\rangle / \langle C, P^*\rangle$ where $\widetilde{P}$ is the coupling obtained by the method considered and $P^*$ is the ground truth (we ensure this optimal cost is large enough to avoid spurious divisions by 0). We present the time-accuracy tradeoffs of the methods for different regularizations $\varepsilon$ and ranks $r$. We show that our method provides consistently a better approximation of the OT while being much faster than the other low-rank methods for various targeted rank values $r$. We also show that our method is able to approximate arbitrarily well the OT and so faster than the Sinkhorn algorithm thanks to the low-rank constraints. We compare the methods in the same setting but we increase the dimensionality of the problems considered and we observe similar results. See Appendix 5.11 for more details.

**Remark 7.** *Adding an entropic regularization in our objective allows to stabilize the MD scheme and therefore obtain faster convergence. Indeed if $\varepsilon > 0$, then the number of iterations required to solve each iteration of the MD scheme (6.7) by Algorithm (4) is monitored by $\varepsilon$ given a certain precision $\delta$ while in the case where $\varepsilon = 0$, the number of iterations required for Algorithm 4 to reach the precision $\delta$ increases as the number of iterations in the MD scheme increases.*

**Comparison of the Couplings** Seeking to take a deeper look at the phenomenon highlighted in Fig. 5.1, we study differences in the regularization paths of **LOT** and **Sin**. We consider distributions supported on graphs of $n = 1000$ nodes, endowed with the shortest path distance [202]. We consider **LOT** with *no* entropic regularization (i.e. $\varepsilon = 0$ in Eq. (5.9)) against **Sin** for various pairs of regularizers. Results are displayed in Fig. 5.4, where the discrete path of regularizations parameterized by the rank $r$ of **LOT** is compared with that obtained by **Sin** when varying $\varepsilon$. The gaps in couplings (in $\ell_1$) between the two methods are displayed. Both methods are able to approximate arbitrarily well the OT but offer two different paths to interpolate from the independent coupling $ab^T$ of rank 1 to the optimal one. More precisely, we see that the range of $\varepsilon$ for which the entropic OT provides an efficient approximation of the true coupling is very localized, while the rank $r$ needed for **LOT** to obtain such approximation is wider. Moreover, we see that the decay of the ratio of **LOT** with respect to $r$ is faster than the decay of **Sin** w.r.t.

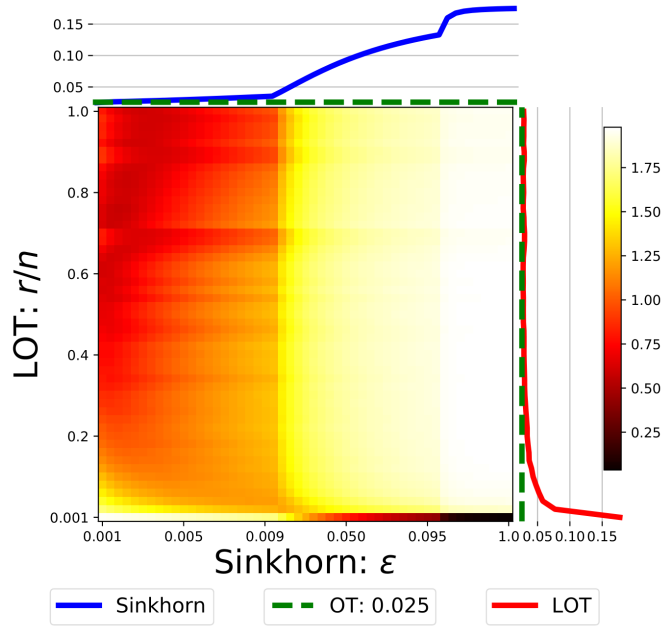



Figure 5.4: We illustrate in this plot the gaps between the couplings reached by **Sin** and **LOT** for varying regularization strengths. Measures were sampled on a complete graph obtained by sampling $2n = 2000$ points from a 2-D standard normal distribution, the edge weights set to their squared Euclidean distances. The supports are obtained by randomly splitting the nodes of the graphs into two subsets of same size. We vary the entropic regularization $\varepsilon$ and the nonnegative rank $r$. We consider $\varepsilon$ in log-scale ranging from 0.001 to 1 and $r$ ranging from 1 to 1000, represented as a fraction of $n$. The blue (resp. red) curve stands for **Sin** (resp. **LOT**). We plot the $\ell_1$ distance between their respective couplings.

$\varepsilon$.

**Remark 8.** *A comparative advantage of using the low-rank parameterization of OT over the Sinkhorn approach lies in the simple bounds that $r$ admits, between 1 and $n$, and the fact that $r$ encodes directly, through an integer, a direct property of the resulting coupling. In that sense, the same value $r$ can be used across experiments that compare measures of various sizes and supports. By contrast, selecting a suitable regularization strength $\varepsilon$ in the Sinkhorn algorithm is usually challenging, as the parameter is continuous and its magnitude depends directly on the cost matrix values, making a common choice across experiments difficult.*

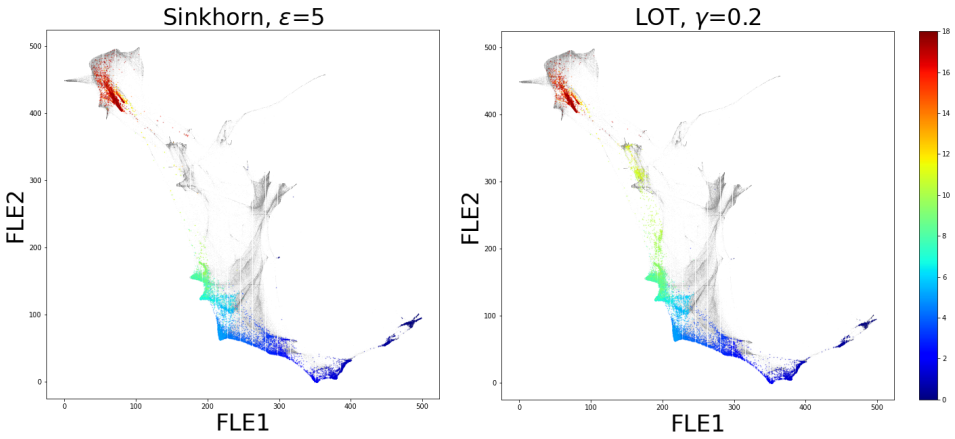



Figure 5.5: Here we compare the paths recovered by both the Sinkhorn algorithm with $\varepsilon = 5$, and our method with $\gamma = 1/\varepsilon$ and $r500$. Each sub-optimal transport problem between two temporal snapchots contains $n \simeq 5000$ cells.

**Real World Application** In Figure 5.5 we consider the single-cell trajectory inference problem [41] where the goal is to infer the ancestors of some specific cells (iPSCs) from temporal snapshots sampled several times a day for a period of 18 days. We apply the exact same pre-processing suggested by [41], and we obtain that our proposed method is able to recover a similar path as the one obtained by the Sinkhorn algorithm.

### 5.4.2 On the non-convexity of LOT

As our problem (5.6) is non-convex, we investigate the effect of the initialization as well as the choice of the gradient step $\gamma$ in the proposed MD scheme. In addition, we consider a specific situation where the optimal coupling solution of (5.1) admits a nonnegative low-rank to see if our method is able to recover the global minimum in such situation. In the following experiments we set $\varepsilon = 0$ and the lower bound on $g$ to $\alpha = 10^{-5}$.

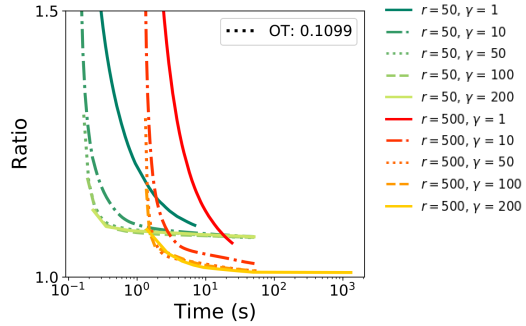



Figure 5.6: In this experiment, we consider the same situation as in Figure 5.2 with $n = m = 1000$ varying $\gamma$ for $r = 50$ or 500.

**Effect of $\gamma$.** In Figure 5.6, we plot the ratio on the same experiment presented in Figure 5.2 when varing $\gamma$. We show that our algorithm is robust to the choice of $\gamma$ as it manages to converge for a large range of $\gamma$. Moreover if the rank is large enough, our method is able to find the optimal solution of the true OT problem (5.1).

**Effect of the Initialization.** In Figure 5.7 we plot the ratios to LP solution of LOT costs, with 50 random initializations (Gaussian entries for $Q, R$, rescaled to have left/right marginals $a$ and $b$). We show that our method is robust to the choice of the initialization. We also design an OT problem where the ground truth OT solution of Eq. (5.1) has low nonnegative rank. Indeed, by fixing $z_1, \dots, z_r \in \mathbb{R}^d$ anchors and by defining the cost $c(x, y) = \min_{k \in 1,..,r} \|x - z_k\| + \|z_k - y\|$, we show that the true optimal coupling has a low nonnegative rank $r$. Our algorithm recovers consistently the OT coupling for multiple random initializations. See Appendix 5.12 for more details.

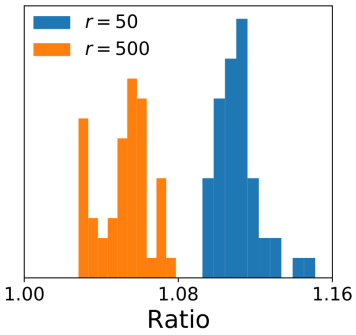



Figure 5.7: Same setting as in Figure 5.6.

**Conclusion** We proposed a new approach to regularize the OT problem by restricting solutions to have a small nonnegative rank. Our algorithm leverages both low-rank constraints and entropic smoothing. Our method can leverage the factorization of the ground cost (and *not* that of the kernel usually associated to Sinkhorn) to propose a linear time complexity alternative to solve OT problems.

## Supplementary material

In Sec. 5.5, we introduce some important notions linked to the mirror-descent scheme. We also prove in this section a general result which states the non-asymptotic stationary convergence of the mirror-descent according to a specific criterion introcuded in this work. In Sec. 5.6, we detail the computation of the Dykstra's algorithm 4 for which we have obtained a simple expression of the updates of the couplings. In Sec. 7.7, we provides all the proofs of the Propositions introduced in this work in the main text. In Sec 5.8, we detail the algorithm presented in [201]. In Sec. 5.9, 5.10, we give two variants of our algorithm when either the marginal $g$ is fixed or when no lower bound is provided on the coordinates of $g$. In Sec. 5.11, we provides more experiment to illustrate our method.

## 5.5 Mirror Descent Algorithm

Let $\mathcal{X}$ a closed convex subset in a Euclidean space $\mathbb{R}^q$, $f : \mathcal{X} \to \mathbb{R}$ continuously differentiable and let us consider the following problem

$$\min_{x \in \mathcal{X}} f(x). \tag{5.11}$$

Given a convex function $h : \mathcal{X} \to \mathbb{R}$ continuously differentiable, one can define the *Bregman Divergence* associated to $h$ as

$$D_h(x, z) := h(x) - h(z) - \langle \nabla h(z), x - z \rangle.$$

To solve Eq. (5.11), one can employ the mirror-descent (MD) algorithm. Given an initial point $x_0 \in \mathcal{X}$ and a sequence of positive step-size $(\gamma_k)_{k \geq 0}$, the mirror-descent scheme associated to the *prox-function* $D_h$ computes

$$x_{k+1} = \operatorname*{argmin}_{x \in \mathcal{X}} \langle \nabla f(x_k), x \rangle + \frac{1}{\gamma_k} D_h(x, x_k).$$

In the following, we need to introduce two notions of relative strong convexity and relative smoothness in order to prove non-asymptotic stationary convergence of the MD scheme.

**Definition** (Relative smoothness.)**.** *Let $L > 0$ and $f$ continuously differentiable on $\mathcal{X}$. $f$ is said to be $L$-smooth relatively to $h$ if*

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + L D_h(y, x)$$

**Definition** (Relative strong convexity)**.** *Let $\alpha > 0$ and $f$ continuously differentiable on $\mathcal{X}$. $f$ is said to be $\alpha$-strongly convex relatively to $h$ if*

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \alpha D_h(y, x) \ \forall \ x, y \in \mathcal{X}$$

Note that $h$ is always 1-strongly convex relatively to $h$. Let us now prove a general result to show non-asymptotic stationary convergence of the MD scheme. For that purpose, we introduce for all $k \geq 0$ the following criterion to establish convergence:

$$\Delta_k := \frac{1}{\gamma_k^2}(D_h(x_k, x_{k+1}) + D_h(x_{k+1}, x_k)).$$

**Proposition 5.5.1.** *Let $N \geq 1$, $f$ continuously differentiable on $\mathcal{X}$ which is $L$-smooth relatively to $h$. By considering for all $k = 1, \ldots, N$, $\gamma_k = 1/2L$, and by denoting $D_0 = f(x_0) - \min_{x \in \mathcal{X}} f(x)$, we have*

$$\min_{0 \leq k \leq N-1} \Delta_k \leq \frac{4LD_0}{N}.$$

*Proof.* Let $k \geq 0$, then by $L$-smoothness of $f$, we have

$$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + LD_h(x_{k+1}, x_k),$$

and by optimality of $x_{k+1}$, we have for all $x \in \mathcal{X}$,

$$\langle \nabla f(x_k) + \frac{1}{\gamma_k}[\nabla h(x_{k+1}) - \nabla h(x_k)], x - x_{k+1} \rangle \geq 0,$$

which implies, by taking $x = x_k$, that

$$\begin{aligned}\langle \nabla f(x_k), x_k - x_{k+1} \rangle &\geq \frac{1}{\gamma_k}[-\langle \nabla h(x_{k+1}), x_k - x_{k+1} \rangle - \langle \nabla h(x_k), x_{k+1} - x_k \rangle] \\ &\geq \frac{1}{\gamma_k}[D_h(x_k, x_{k+1}) + D_h(x_{k+1}, x_k)].\end{aligned}$$

Then we have

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{\gamma_k}[D_h(x_k, x_{k+1}) + D_h(x_{k+1}, x_k)] + LD_h(x_{k+1}, x_k) + LD_h(x_k, x_{k+1})$$

where the last term is added by positivity of $D_h(\cdot, \cdot)$ (as $h$ is supposed to be convex on $\mathcal{X}$). Finally we obtain that

$$\left( \sum_{k=0}^{N-1} \gamma_k(1 - \gamma_k L)\Delta_k \right) \leq f(x_0) - f(x_N) \leq D_0,$$

and as soon as $\gamma_k < \frac{1}{L}$, we have

$$\min_{0 \leq k \leq N-1} \Delta_k \leq \frac{D_0}{\left( \sum_{k=0}^{N-1} \gamma_k(1 - \gamma_k L) \right)}.$$

Then by taking $\gamma_k = \frac{1}{2L}$, the result follows. □

In this paper, we consider $h$ to be the negative entropy function defined on $\Delta_q^*$ as

$$h(x) = \sum_{i=1}^{q} x_i \log(x_i). \tag{5.12}$$

Therefore the *prox-function* associated is just the Kullback–Leibler divergence (KL) defined as,

$$\mathrm{KL}(x, z) = \sum_{i=1}^{q} x_i \log(x_i / z_i).$$

Moreover if $\mathcal{X} \subset \prod_{i=1}^{p} \Delta_{q_i}^*$ for $p \geq 1$, we consider instead

$$h((x^{(1)}, \ldots, x^{(p)})) := \sum_{i=1}^{p} \sum_{j=1}^{q_i} x_j^{(i)} \log(x_j^{(i)})$$

where the associated *prox-function* is

$$D_h((x^{(1)}, \ldots, x^{(p)}), (z^{(1)}, \ldots, z^{(p)})) = \sum_{i=1}^{p} \mathrm{KL}(x^{(i)}, z^{(i)}).$$

## 5.6 The Dykstra's Algorithm

In order to solve Eq. (6.7), we use the Dykstra's Algorithm [196]. Given a closed convex set $\mathcal{C} \subset \mathbb{R}_+^{n\times r} \times \mathbb{R}_+^{m\times r} \times \mathbb{R}_+^r$, we denote for all $\boldsymbol{\xi} \in (\mathbb{R}_+^*)^{n\times r} \times (\mathbb{R}_+^*)^{m\times r} \times (\mathbb{R}_+^*)^r$ the projection according to the Kullback-Leibler divergence as

$$\mathcal{P}_{\mathcal{C}}^{\mathrm{KL}}(\boldsymbol{\xi}) := \operatorname*{argmin}_{\boldsymbol{\zeta}\in\mathcal{C}} \mathrm{KL}(\boldsymbol{\zeta}, \boldsymbol{\xi}).$$

Starting from $\boldsymbol{\zeta}_0 := \boldsymbol{\xi}$ and $\boldsymbol{q}_0 = \boldsymbol{q}_{-1} = (\mathbf{1}, \mathbf{1}, \mathbf{1}) \in \mathbb{R}_+^{n\times r} \times \mathbb{R}_+^{m\times r} \times \mathbb{R}_+^r$, the Dykstra's Algorithm 4 applied to our problem consists in computing for all $j \geq 0$,

$$\begin{aligned}
\boldsymbol{\zeta}_{2j+1} &= \mathcal{P}_{\mathcal{C}_1(a,b,r,\alpha)}^{\mathrm{KL}}(\boldsymbol{\zeta}_{2j} \odot \boldsymbol{q}_{2j-1}) \\
\boldsymbol{q}_{2j+1} &= \boldsymbol{q}_{2j-1} \odot \frac{\boldsymbol{\zeta}_{2j}}{\boldsymbol{\zeta}_{2j+1}} \\
\boldsymbol{\zeta}_{2j+2} &= \mathcal{P}_{\mathcal{C}_2(r)}^{\mathrm{KL}}(\boldsymbol{\zeta}_{2j+1} \odot \boldsymbol{q}_{2j}) \\
\boldsymbol{q}_{2j+2} &= \boldsymbol{q}_{2j} \odot \frac{\boldsymbol{\zeta}_{2j+1}}{\boldsymbol{\zeta}_{2j+2}}.
\end{aligned}$$

In fact these operations can be simplified to simple matrix/vector multiplications. More precisely, the Dykstra's Algorithm produces the iterates $(\boldsymbol{\zeta}_j)_{j\geq 0}$ which satisfy for all $j \geq 0$ $\boldsymbol{\zeta}_j = (Q_j, R_j, g_j)$ where

$$\begin{aligned} Q_j &= \operatorname{diag}(u_j^1)\xi^{(1)} \operatorname{diag}(v_j^1) \\ R_j &= \operatorname{diag}(u_j^2)\xi^{(2)} \operatorname{diag}(v_j^2) \end{aligned}$$

for the sequences $(u_j^i, v_j^i)_{j\geq 0}$ initialized as, $u_0^i := \mathbf{1}_n$, $v_0^i := \mathbf{1}_m$ for all $i \in \{1, 2\}$, $q_{0,1}^{(3)} = q_{0,2}^{(3)} = q_0^{(1)} = q_0^{(2)} = \mathbf{1}_r$ and computed with the iterations

$$\begin{aligned} u_{n+1}^{k,i} &= \frac{p_i}{\xi_k^i v_n^{k,i}} \\ \tilde{g}_{n+1} &= \max(\alpha, g_n \odot q_{n,1}^{(3)}),\ q_{n+1,1}^{(3)} = (g_n \odot q_{n,1}^{(3)})/\tilde{g}_{n+1} \\ g_{n+1} &= (\tilde{g}_{n+1} \odot q_{n,2}^{(3)})^{1/3} \prod_{i=1}^{2} (v_n^{k,i} \odot q_n^{(i)} \odot (\xi_k^i)^T u_n^{k,i})^{1/3} \\ v_{n+1}^{k,i} &= \frac{g_{n+1}}{(\xi_k^i)^T u_n^{k,i}} \\ q_{n+1}^{(i)} &= (v_n^{k,i} \odot q_n^{(i)})/v_{n+1}^{k,i},\ q_{n+1,2}^{(3)} = (\tilde{g}_{n+1} \odot q_{n,2}^{(3)})/g_{n+1} \end{aligned}$$

## 5.7 Proofs

### 5.7.1 Proof of Proposition 5.3.1

*Proof.* The case when $\varepsilon = 0$ is clear. Assume now that $\varepsilon > 0$. When $r = 1$, note that $\mathcal{C}_1(a, b, r) \cap \mathcal{C}_2(r)$ is closed as $g = 1$ and bounded, therefore and by continuity of the objective the mininum exists. Let $r \geq 2$. First remarks that we always have $\mathrm{LOT}_{r,\varepsilon}(\mu, \nu) \leq \mathrm{LOT}_{r-1,\varepsilon}(\mu, \nu)$. Let us assume that (5.8) does not admits a minimum. Because the objective $F_\varepsilon$ is a lower semi-continuous function on $\overline{\mathcal{C}_1(a, b, r)} \cap \mathcal{C}_2(r)$, and by compacity of $\overline{\mathcal{C}_1(a, b, r)} \cap \mathcal{C}_2(r)$, the objective function admits a minimum $(Q, R, g) \in \overline{\mathcal{C}_1(a, b, r)} \cap \mathcal{C}_2(r)$ and we have $\mathrm{LOT}_{r,\varepsilon}(\mu, \nu) = F_\varepsilon(Q, R, g)$. But as the minimum is not attained on $\mathcal{C}_1(a, b, r) \cap \mathcal{C}_2(r)$, it means that there exists at least one coordinate $i \in \{1, \ldots, r\}$ such that $g_i = 0$. Then because the constraints, $Q$ and $R$ both admit a column which is the null vector. By deleting these coordinates in $Q, R, g$, we obtain that $\mathrm{LOT}_{r,\varepsilon}(\mu, \nu) = \mathrm{LOT}_{r-1,\varepsilon}(\mu, \nu)$. □

### 5.7.2 Proof of Proposition 5.3.2

*Proof.* The first oder conditions of the projection gives that there exists $(\lambda_1, \lambda_2, \lambda_3) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^r_+$ such that

$$\begin{aligned} \log(Q/\tilde{Q}) + \lambda_1 \mathbf{1}^T &= 0 \\ \log(R/\tilde{R}) + \lambda_2 \mathbf{1}^T &= 0 \\ \log(g/\tilde{g}) + \lambda_3 &= 0 \end{aligned}$$

Moreover the conditions $Q\mathbf{1} = a$, $R\mathbf{1} = b$ and $g \geq \alpha$ imply that

$$\begin{aligned} Q &= \mathrm{Diag}(a/\tilde{Q}\mathbf{1})\tilde{Q} \\ R &= \mathrm{Diag}(b/\tilde{R}\mathbf{1})\tilde{R} \\ g &= \max(\alpha, \tilde{g}). \end{aligned}$$

□

### 5.7.3 Proof of Proposition 5.3.3

*Proof.* The first order conditions of the projection states that there exists $(\lambda_1, \lambda_2) \in \mathbb{R}^r \times \mathbb{R}^r$ such that

$$\begin{aligned} \log(Q/\tilde{Q}) + \mathbf{1}_n \lambda_1^T &= 0 \\ \log(R/\tilde{R}) + \mathbf{1}_m \lambda_2^T &= 0 \\ \log(g/\tilde{g}) - (\lambda_1 + \lambda_2) &= 0 \end{aligned}$$

Moreover the conditions $Q^T\mathbf{1}_n = R^T\mathbf{1}_m = g$ imply that

$$\begin{aligned} Q &= \tilde{Q}\mathrm{Diag}(g/\tilde{Q}^T\mathbf{1}_n) \\ R &= \tilde{R}\mathrm{Diag}(g/\tilde{R}^T\mathbf{1}_m) \\ g^3 &= \tilde{g} \odot \tilde{Q}^T\mathbf{1}_n \odot \tilde{R}^T\mathbf{1}_m \end{aligned}$$

from which the result follows. □

### 5.7.4 Proof of Proposition 7.4.2

*Proof.* To show the result, we just need to show that

$$F_\varepsilon : (Q, R, g) \in \mathcal{C}(a, b, r, \alpha) \to \langle C, Q\,\mathrm{diag}(1/g)R^T \rangle - \varepsilon H(Q, R, g)$$

is smooth relatively to

$$H(Q,R,g) := \sum_{i,j} Q_{i,j}\log(Q_{i,j}) + \sum_{i,j} R_{i,j}\log(R_{i,j}) + \sum_j g_j \log(g_j),$$

then by applying Proposition 7.7.1, the result will follow. Let us now show that $F_\varepsilon$ is $L_{\varepsilon,\alpha}$-smooth. To do so, it is enough to show that [203, 204]

$$\|\nabla F_\varepsilon(Q_1,R_1,g_1) - \nabla F_\varepsilon(Q_2,R_2,g_2)\|_2 \leq L_{\varepsilon,\alpha}\|H(Q_1,R_1,g_1) - H(Q_2,R_2,g_2)\|_2.$$

We first have that

$$\begin{aligned}
\nabla_Q F_\varepsilon &= CR\,\mathrm{diag}(1/g) + \varepsilon(\log Q + \mathbf{1}\\
\nabla_R F_\varepsilon &= C^T Q\,\mathrm{diag}(1/g) + \varepsilon(\log R + \mathbf{1}\\
\nabla_g F_\varepsilon &= -\mathcal{D}(Q^T RC)/g^2 + \varepsilon(\log g + 1)
\end{aligned}$$

Now we have,

$$\begin{aligned}
\|\nabla F_\varepsilon(Q_1) - \nabla F_\varepsilon(Q_2)\|_2^2 &\leq \|CR_1\,\mathrm{diag}(1/g_1) - CR_2\,\mathrm{diag}(1/g_2)\|_2^2 + \varepsilon^2\|\log Q_1 - \log Q_2\|_2^2\\
&\quad + 2\varepsilon\|\log Q_1 - \log Q_2\|_2\|CR_1\,\mathrm{diag}(1/g_1) - CR_2\,\mathrm{diag}(1/g_2)\|_2\\
&\leq \|C\|_2^2\|(R_1-R_2)\,\mathrm{diag}(1/g_1) + (\mathrm{diag}(1/g_1) - \mathrm{diag}(1/g_2))R_2\|_2^2\\
&\quad + \varepsilon^2\|\log Q_1 - \log Q_2\|_2^2\\
&\quad + 2\varepsilon\|\log Q_1 - \log Q_2\|_2\|CR_1\,\mathrm{diag}(1/g_1) - CR_2\,\mathrm{diag}(1/g_2)\|_2\\
&\leq \|C\|_2^2\left[\frac{\|R_1-R_2\|_2^2}{\alpha^2} + \|1/g_1 - 1/g_2\|_2^2 + \frac{\|R_1-R_2\|\|1/g_1 - 1/g_2\|_2}{\alpha}\right]\\
&\quad + \varepsilon^2\|\log Q_1 - \log Q_2\|_2^2\\
&\quad + 2\varepsilon\|\log Q_1 - \log Q_2\|_2\|CR_1\,\mathrm{diag}(1/g_1) - CR_2\,\mathrm{diag}(1/g_2)\|_2.
\end{aligned}$$

As $Q \to H(Q)$ is 1-strongly convex w.r.t to the $\ell_2$-norm on $\Delta_{n\times r}$, we have

$$\begin{aligned}
\|Q_1 - Q_2\|_2^2 &\leq \langle \log Q_1 - \log Q_2, Q_1 - Q_2\rangle\\
&\leq \|\log Q_1 - \log Q_2\|_2\|Q_1 - Q_2\|_2
\end{aligned}$$

from which follows that

$$\|Q_1 - Q_2\|_2 \leq \log Q_1 - \log Q_2\|_2.$$

Moreover we have

$$\|1/g_1 - 1/g_2\|_2 \leq \frac{\|g_1 - g_2\|_2}{\alpha^2} \leq \|\frac{\|\log g_1 - \log g_2\|_2}{\alpha^2}$$

Therefore we obtain that

$$\begin{aligned}&\|\nabla F_\varepsilon(Q_1) - \nabla F_\varepsilon(Q_2)\|_2^2 \\ &\leq \left(\frac{\|C\|_2}{\alpha}\|\log R_1 - \log R_2\|_2 + \frac{\|C\|_2}{\alpha^2}\|\log g_1 - \log g_2\|_2 + \varepsilon\|\log Q_1 - \log Q_2\|_2\right)^2.\end{aligned}$$

An analogue proof leads to

$$\begin{aligned}&\|\nabla F_\varepsilon(R_1) - \nabla F_\varepsilon(R_2)\|_2^2 \\ &\leq \left(\frac{\|C\|_2}{\alpha}\|\log Q_1 - \log Q_2\|_2 + \frac{\|C\|_2}{\alpha^2}\|\log g_1 - \log g_2\|_2 + \varepsilon\|\log R_1 - \log R_2\|_2\right)^2.\end{aligned}$$

Let us now consider smoothness of $F_\varepsilon$ w.r.t $g$,

$$\begin{aligned}\|\nabla F_\varepsilon(g_1) - \nabla F_\varepsilon(g_2)\|_2^2 \leq &\left\|\frac{\mathcal{D}(Q_1^T C R_1)}{g_1^2} - \frac{\mathcal{D}(Q_2^T C R_2)}{g_2^2}\right\|_2^2 + \varepsilon^2\|\log g_1 - \log g_2\|_2^2 \\ &+ 2\varepsilon\left\|\frac{\mathcal{D}(Q_1^T C R_1)}{g_1^2} - \frac{\mathcal{D}(Q_2^T C R_2)}{g_2^2}\right\|_2 \|\log g_1 - \log g_2\|_2.\end{aligned}$$

but we have that

$$\begin{aligned}\left\|\frac{\mathcal{D}(Q_1^T C R_1)}{g_1^2} - \frac{\mathcal{D}(Q_2^T C R_2)}{g_2^2}\right\|_2^2 \leq &\|(1/g_1^2 - 1/g_2^2)\operatorname{diag}(Q_1^T C R_1)\|_2^2 + \\ &\|\mathcal{D}(Q_1^T C R_1) - \mathcal{D}(Q_2^T C R_2)/g_2^2\|_2^2 \\ &+ 2\|(1/g_1^2 - 1/g_2^2)\operatorname{diag}(Q_1^T C R_1)\|_2 \\ &\|\mathcal{D}(Q_1^T C R_1) - \mathcal{D}(Q_2^T C R_2)/g_2^2\|_2 \\ &\leq \left(\frac{2\|C\|_2}{\alpha^3}\|\log g_1 - \log g_2\|_2 + W\right)^2,\end{aligned}$$

where

$$W = \frac{\|C\|_2}{\alpha^2}\left[\|Q_1 - Q_2\|_2^2 + \|R_1 - R_2\|_2\right]$$

Therefore we obtain that

$$\begin{aligned}&\|\nabla F_\varepsilon(g_1) - \nabla F_\varepsilon(g_2)\|_2^2 \\ &\leq \left(\left(\varepsilon + \frac{2\|C\|_2}{\alpha^3}\right)\|\log g_1 - \log g_2\|_2 + \frac{\|C\|}{\alpha^2}\|Q_1 - Q_2\|_2 + + \frac{\|C\|}{\alpha^2}\|R_1 - R_2\|_2\right)^2\end{aligned}$$

Finally we obtain that

$$\begin{aligned}
&\|\nabla F_\varepsilon(Q_1,R_1,g_1)-\nabla F_\varepsilon(Q_2,R_2,g_2)\|_2^2\\
&\leq 3\left(\frac{\|C\|_2^2}{\alpha^2}+\frac{\|C\|_2^2}{\alpha^4}+\varepsilon^2\right)\left[\|\log Q_1-\log Q_2\|_2^2+\|\log R_1-\log R_2\|_2^2\right]\\
&+3\left(\frac{2\|C\|_2^2}{\alpha^4}+\left(\varepsilon+\frac{2\|C\|_2}{\alpha^3}\right)^2\right)\|\log g_1-\log g_2\|_2^2
\end{aligned}$$

Thus we obtain that

$$\|\nabla F_\varepsilon(Q_1,R_1,g_1)-\nabla F_\varepsilon(Q_2,R_2,g_2)\|_2\leq L_{\varepsilon,\alpha}\|\nabla H(Q_1,R_1,g_1)-\nabla H(Q_2,R_2,g_2)\|_2$$

and the result follows. □

## 5.8 Low-Rank Factorization of Distance Matrix

In this section we present the algorithm used to perform a low-rank approximation of a distance matrix [200, 201]. Given a metric space $(\mathcal{X},d)$, $X=\{x_i\}_{i=1}^n\in\mathcal{X}^n$ and $Y=\{y_j\}_{j=1}^m\in\mathcal{X}^m$ we aim at obtaining a low-rank approximation of the distance matrix $D=(d(x_i,y_j))_{i,j}$ with a precision $\gamma>0$. Let us now present the algorithm considered where we have denoted $t=\lfloor r/\gamma\rfloor$.

**Algorithm 6** LR-Distance$(X,Y,r,\gamma)$

**Inputs:** $X,Y,r,\gamma$

Choose $i^*\in\{1,\dots,n\}$, and $j^*\{1,\dots,m\}$ uniformly at random.

For $i=1,\dots,n$, $p_i\leftarrow d(x_i,y_j^*)^2+d(x_i^*,y_j^*)^2+\frac{1}{m}\sum_{j=1}^m d(x_i^*,y_j)^2$.

Independently choose $i^{(1)},\dots,i^{(t)}$ according $(p_1,\dots,p_n)$.

$X^{(t)}\leftarrow[x_{i^{(1)}},\dots,x_{i^{(t)}}]$, $P^{(t)}\leftarrow[\sqrt{tp_{i^{(1)}}},\dots,\sqrt{tp_{i^{(t)}}}]$, $S\leftarrow d(X^{(t)},Y)/P^{(t)}$

Denote $S=[S^{(1)},\dots,S^{(m)}]$,

For $j=1,\dots,m$, $q_j\leftarrow\|S^{(j)}\|_2^2/\|S\|_F^2$

Independently choose $j^{(1)},\dots,j^{(t)}$ according $(q_1,\dots,q_m)$.

$S^{(t)}\leftarrow[S^{j^{(1)}},\dots,S^{j^{(t)}}]$, $Q^{(t)}\leftarrow[\sqrt{tq_{j^{(1)}}},\dots,\sqrt{tq_{j^{(t)}}}]$, $W\leftarrow S^{(t)}/Q^{(t)}$

$U_1,D_1,V_1\leftarrow\mathrm{SVD}(W)$ (decreasing order of singular values).

$N\leftarrow[U_1(1),\dots,U_1^{(r)}]$, $N\leftarrow S^TN/\|W^TN\|_F$

Choose $j^{(1)},\dots,j^{(t)}$ uniformly at random in $\{1,\dots,m\}$.

$Y^{(t)}\leftarrow[y_{j^{(1)}},\dots,y_{j^{(t)}}]$, $D^{(t)}\leftarrow d(X,Y^{(t)})/\sqrt{t}$.

$U_2,D_2,V_2=\mathrm{SVD}(N^TN)$, $U_2\leftarrow U_2/D_2$, $N^{(t)}\leftarrow[(N^T)^{(j^{(1)})},\dots,(N^T)^{(j^{(t)})}]$, $B\leftarrow U_2^TN^{(t)}/\sqrt{t}$, $A\leftarrow(BB^T)^{-1}$.

$Z\leftarrow AB(D^{(t)})^T$, $M\leftarrow Z^TU_2^T$

**Result:** $M,N$

## 5.9 Positive low-rank factorization with fixed marginal

Let $g \in \Delta_r^*$, and let us for now consider the following problem

$$\mathrm{LOT}_{r,g}(\mu,\nu) := \min_{P \in \Pi_{a,g,b}} \langle C, P \rangle. \tag{5.13}$$

By definition of $\Pi_{a,g,b}$, this problem can be formulated as follows:

$$\mathrm{LOT}_{r,g}(\mu,\nu) = \min_{\substack{Q \in \Pi_{a,g} \\ R \in \Pi_{b,g}}} \langle C, Q\mathrm{Diag}(1/g)R^T \rangle. \tag{5.14}$$

As in the classical OT problem, one can extend the above objective and consider for any $\varepsilon \geq 0$ an entropic version of the problem defined as

$$\mathrm{LOT}_{r,g,\varepsilon}(\mu,\nu) := \min_{\substack{Q \in \Pi_{a,g} \\ R \in \Pi_{b,g}}} \langle C, Q\mathrm{Diag}(1/g)R^T \rangle - \varepsilon H((Q,R)) \tag{5.15}$$

Note that for any $\varepsilon \geq 0$, the minimum always exists as the objective is continuous and $\Pi_{a,g,b}$ is compact. Moreover we clearly have that $\mathrm{LOT}_{r,g,0}(\mu,\nu) = \mathrm{LOT}_{r,g}(\mu,\nu)$. Applying a MD method to the objective (5.14) leads for all $k \geq 0$ to the following updates

$$\begin{aligned} Q_{k+1} &:= \operatorname*{argmin}_{Q \in \Pi_{a,g}} \langle C_k^{(1)}, Q \rangle - \frac{1}{\gamma_k} H(Q) \\ R_{k+1} &:= \operatorname*{argmin}_{R \in \Pi_{a,g}} \langle C_k^{(2)} R \rangle - \frac{1}{\gamma_k} H(R) \end{aligned}$$

where, $(Q_0, R_0) \in \Pi_{a,g} \times \Pi_{b,g}$ is an initial point, $C_k^{(1)} := CR_k\mathrm{Diag}(1/g) + (\varepsilon - \frac{1}{\gamma_k}) \log(Q_k)$, $C_k^{(2)} := C^T Q_k \mathrm{Diag}(1/g) + (\varepsilon - \frac{1}{\gamma_k}) \log(R_k)$ and $\gamma_k$ is a sequence of positive real numbers. Therefore a MD method bowls down to solve at each iteration two regularized OT problems which can be done efficiently using the Sinkhorn algorithm (3).

**Convergence of the Mirror Descent.** Even if the objective (5.14) is not convex in $(Q,R)$, one can obtain the non-asymptotic stationary convergence of the MD algorithm in this setting.

Let $f_\varepsilon$ be the objective function of the problem (5.15) defined on $X := \Pi_{a,g} \times \Pi_{b,g}$ and let us denotes for any $\gamma > 0$ and $x \in X$

$$\mathcal{G}_\varepsilon(x,\gamma) := \operatorname*{argmin}_{u \in X} \{ \langle \nabla f_\varepsilon(x), u \rangle + \frac{1}{\gamma} KL(u,x) \}.$$

Let us now define the following criterion to establish convergence:

$$\Delta_\varepsilon(x,\gamma) := \frac{1}{\gamma^2}(KL(x,\mathcal{G}_\varepsilon(x,\gamma)) + KL(\mathcal{G}_\varepsilon(x,\gamma),x)).$$

To show the non-asymptotic stationary convergence, we show that for any $\varepsilon \geq 0$, the objective is smooth relative to the entropy function [199] and we extend the proof of [198] to this case.

**Proposition.** *Let $\varepsilon \geq 0$ and $N \geq 1$. By denoting $L_\varepsilon := \sqrt{2(\|C\|_2^2\|Diag(1/g)\|_2^2 + \varepsilon^2)}$ and by considering a constant stepsize in the MD scheme such that for all $k = 1,\dots,N$ $\gamma_k = \frac{1}{L_\varepsilon}$, we obtain that*

$$\min_{1\leq k\leq N} \Delta_\varepsilon((Q_k,R_k),\gamma_k) \leq \frac{2L_\varepsilon D_0}{N}.$$

*where $D_0 := f_\varepsilon(Q_0,R_0) - LOT_{r,g,\varepsilon}$ is the distance of the initial value to the optimal one.*

*Proof.* A similar proof of the one given for Proposition 7.4.2 gives that $f_\varepsilon$ is $L_\varepsilon$-smooth relatively to $H$. □

Let us now introduce our first algorithm (7) to compute a positive low-rank factorization of the optimal coupling. Here we consider the case where $g := \mathbf{1}_r/r$. Before introducing our algorithm it is worth noting that a trivial initialization may lead to a trivial fixed point in the MD updates. Indeed if one initialize $Q := ag^T$ and $R := bg^T$, then $CR\text{Diag}(1/g) = Ca\mathbf{1}^T$ and $C^TQ\text{Diag}(1/g) = C^Tb\mathbf{1}^T$ and therefore $(Q,R)$ is a fixed point of the MD. To avoid this, we initialize our algorithm in the following way: let $\lambda := \min_{i,j,k}(a_i,b_j,g_k)/2$, $a_1 \in \Delta_n^*\backslash\{a\}$, $a_2 := (a-\lambda a_1)/(1-\lambda)$, $b_1 \in \Delta_n^*\backslash\{b\}$, $b_2 := (b-\lambda b_1)/(1-\lambda)$, $g_1 \in \Delta_r^*\backslash\{g\}$ and $g_2 := (g-\lambda g_1)/(1-\lambda)$. We can now define our initialization as $Q := \lambda a_1 g_1^T + (1-\lambda)a_2 g_2^T$, $R := \lambda b_1 g_1^T + (1-\lambda)b_2 g_2^T$.

**Algorithm 7** LOT-F$(C, a, b, \delta)$

**Inputs:** $C, a, b, \delta, Q, R, g, \gamma, \delta_{\mathrm{S}}$

**repeat**

$Q_{\text{old}} \leftarrow Q,\ R_{\text{old}} \leftarrow R$

$C^{(1)} \leftarrow CR\mathrm{Diag}(1/g) - \frac{1}{\gamma}\log(Q)$,

$C^{(2)} \leftarrow C^T Q\mathrm{Diag}(1/g) - \frac{1}{\gamma}\log(R)$,

$K^{(1)} \leftarrow \exp(-\gamma C^{(1)})$,

$K^{(2)} \leftarrow \exp(-\gamma C^{(2)})$,

$u, v \leftarrow \mathrm{Sinkhorn}(K^{(1)}, a, g, \delta_{\mathrm{S}})$ (Algorithm (3)),

$Q \leftarrow \mathrm{Diag}(u)K^{(1)}\mathrm{Diag}(v)$,

$u, v \leftarrow \mathrm{Sinkhorn}(K^{(2)}, a, g, \delta_{\mathrm{S}})$ (Algorithm (3)),

$R \leftarrow \mathrm{Diag}(u)K^{(2)}\mathrm{Diag}(v)$

**until** $\Delta((Q, R), \gamma) < \delta$;

**Result:** $Q, R$

**Computational Cost.** Note that the kernels $(K^{(i)})_{1\leq i\leq 2}$ considered in algorithm (7) live in $\mathbb{R}_+^{n\times r} \times \mathbb{R}_+^{m\times r}$ and therefore each iteration of both Sinkhorn algorithms can be computed either in $\mathcal{O}(nr)$ or in $\mathcal{O}(mr)$ algebraic operations as it involves only matrix/vector multiplications of the form $K^{(i)}v$ and $(K^{(i)})^T u$. However without any assumption on the cost matrix $C$, computing $(K^{(i)})_{1\leq i\leq 2}$ costs $\mathcal{O}(nmr)$ algebraic operations as it requires to compute both $CR$ and $C^T Q$ at each iteration. Thanks to assumption 1, such multiplications can be performed in $\mathcal{O}((n+m)dr)$ algebraic operations and thus algorithm (7) requires only a linear number of algebraic operations with respect to the number of samples at each iteration.

In the following, we will see that if we do not fix the marginal, the problem can also be solved efficiently as each iteration of the MD algorithm can be seen as a wasserstein barycenter problem.

## 5.10 A Positive low-rank factorization with free marginal

Applying a MD method to the objective (5.8) leads, for all $k \geq 0$, to the following updates

$$(Q_{k+1}, R_{k+1}, g_{k+1}) := \underset{\boldsymbol{\zeta}\in\mathcal{C}_1(a,b,r)\cap\mathcal{C}_2(r)}{\operatorname{argmin}} \mathrm{KL}(\boldsymbol{\zeta}, \boldsymbol{\xi}_k) \tag{5.16}$$

where $(Q_0, R_0, g_0) \in \mathcal{C}_1(a, b, r)\cap\mathcal{C}_2(r)$ is an initial point, $\boldsymbol{\xi}_k := (\xi_k^{(1)}, \xi_k^{(2)}, \xi_k^{(3)})$, $\xi_k^{(1)} := \exp(-\gamma_k CR_k\mathrm{Diag}(1/g_k)_k - (\gamma_k\varepsilon - 1)\log(Q_k))$, $\xi_k^{(2)} := \exp(-\gamma_k C^T Q_k\mathrm{Diag}(1/g_k) -$

$(\gamma_k\varepsilon - 1)\log(R_k))$, $\xi_k^{(3)} := \exp(\gamma_k\omega_k/g_k^2 - (\gamma_k\varepsilon-1)\log(g_k))$ with $[\omega_k]_i := [Q_k^T C R_k]_{i,i}$ for all $i \in \{1,\dots,r\}$ and $(\gamma_k)_{k\geq 0}$ is a sequence of positive real numbers.

Eq. (5.16) is well defined. Indeed as the kernels $(\xi_k^{(i)})$ are matrices with positive coefficients, the infimum is attained in $\mathcal{C}_1(a,b,r) \cap \mathcal{C}_2(r)$ and the problem admits a unique solution. Moreover solving Eq. (5.16) bowls down to solve

$$(Q_{k+1}, R_{k+1}, g_{k+1}) := \operatorname*{argmin}_{\zeta \in \overline{\mathcal{C}_1(a,b,r)} \cap \mathcal{C}_2(r)} \mathrm{KL}(\zeta, \boldsymbol{\xi}_k) \tag{5.17}$$

In order to solve Eq. (5.17), we consider the Iterative Bregman Projections (IBP) algorithm. Starting from $\boldsymbol{\zeta}_0^{(k)} := \boldsymbol{\xi}_k$, the IBP algorithm consists in computing for all $j \geq 0$,

$$\begin{aligned}\boldsymbol{\zeta}_{2j+1}^{(k)} &= \mathcal{P}^{\mathrm{KL}}_{\overline{\mathcal{C}_1(a,b,r)}}(\boldsymbol{\zeta}_{2j}^{(k)}) \\ \boldsymbol{\zeta}_{2j+2}^{(k)} &= \mathcal{P}^{\mathrm{KL}}_{\mathcal{C}_2(r)}(\boldsymbol{\zeta}_{2j+1}^{(k)}).\end{aligned}$$

As $\overline{\mathcal{C}_1(a,b,r)}$ and $\mathcal{C}_2(r)$ are affine subspaces (note that nonnegativity constraints are already in the definition of the objective) one can show that $\boldsymbol{\zeta}_j^{(k)}$ converges towards the unique solution of Eq. (5.17), [163]. Remarks that the projection on $\overline{\mathcal{C}_1(a,b,r)}$ can be computed very easily as one has for any $\tilde{\boldsymbol{\xi}} := (\tilde{Q}, \tilde{R}, \tilde{g}) \in \mathbb{R}_+^{n\times r} \times \mathbb{R}_+^{n\times r} \times \mathbb{R}_+^r$,

$$\mathcal{P}^{\mathrm{KL}}_{\overline{\mathcal{C}_1(a,b,r)}}(\tilde{\boldsymbol{\xi}}) = \left(\mathrm{Diag}\left(\frac{a}{\tilde{Q}\mathbf{1}_r}\right)\tilde{Q}, \mathrm{Diag}\left(\frac{b}{\tilde{R}\mathbf{1}_r}\right)\tilde{R}, \tilde{g}\right)$$

and the solution of the projection on $\mathcal{C}_2(r)$ is already given in Proposition 5.3.3.

**Efficient computation of the updates.** For all $k \geq 0$, starting with $\boldsymbol{\zeta}_0^{(k)} := \boldsymbol{\xi}_k$ the IBP algorithm leads to a simple algorithm (8) which computes only scaling vectors. More precisely, the IBP algorithm produces the iterates $(\boldsymbol{\zeta}_n^{(k)})_{n\geq 0}$ which satisfy for all $n \geq 0$ $\boldsymbol{\zeta}_n^{(k)} = (Q_n^{(k)}, R_n^{(k)}, g_n^{(k)})$ where

$$\begin{aligned}Q_n^{(k)} &= \mathrm{Diag}(u_n^{k,1})\xi_k^1\mathrm{Diag}(v_n^{k,1}) \\ R_n^{(k)} &= \mathrm{Diag}(u_n^{k,2})\xi_k^2\mathrm{Diag}(v_n^{k,2})\end{aligned}$$

for the sequences $(u_n^{k,i}, v_n^{k,i})$ initialized as $v_0^{k,i} := \mathbf{1}$ for all $i \in \{1,2\}$ and computed with the iterations

$$\begin{aligned}u_n^{k,i} &= \frac{p_i}{\xi_k^i v_n^{k,i}} \\ g_{n+1}^{(k)} &= (g_n^{(k)})^{1/3}\prod_{i=1}^{2}(v_n^{k,i} \odot (\xi_k^i)^T u_n^{k,i})^{1/3} \\ v_{n+1}^{k,i} &= \frac{g_{n+1}^{(k)}}{(\xi_k^i)^T u_n^{k,i}}\end{aligned}$$

where we have denoted $p_1 := a$ and $p_2 := b$ to simplify the notations.

**Algorithm 8** LR-IBP$((\xi^{(i)})_{1\leq i\leq 3}, p_1, p_2, \delta)$

**Inputs:** $\xi^{(1)}, \xi^{(2)}, g := \xi^{(3)}, p_1, p_2, \delta, v^{(i)}$

**repeat**

$u^{(i)} \leftarrow p_i / \xi^{(i)} v^{(i)} \ \forall i \in \{1, 2\}$,

$g \leftarrow (g)^{1/3} \prod_{i=1}^{2} (v^{(i)} \odot (\xi^{(i)})^T u^{(i)})^{1/3}$,

$v^{(i)} \leftarrow g / (\xi^{(i)})^T u^{(i)} \ \forall i \in \{1, 2\}$

**until** $\sum_{i=1}^{2} \|u^{(i)} \odot \xi^{(i)} v^{(i)} - p_i\|_1 < \delta$;

$Q \leftarrow \mathrm{Diag}(u^{(1)}) \xi_k^{(1)} \mathrm{Diag}(v^{(1)})$

$R \leftarrow \mathrm{Diag}(u^{(2)}) \xi_k^{(2)} \mathrm{Diag}(v^{(2)})$

**Result:** $Q, R, g$

Let us now introduce the proposed MD algorithm applied to (6.4). By denoting $\mathcal{D}(\cdot)$ the operator extracting the diagonal of a square matrix we obtain the following algorithm (9) to solve Eq. (5.6). We initialize our algorithm with the exact same procedure as in algorithm (7).

**Algorithm 9** LOT$(C, a, b, r, \delta)$

**Inputs:** $C, a, b, (\gamma_k)_{k\geq 0}, Q, R, g, \delta$

**for** $k = 1, \ldots$ **do**

$\xi^{(1)} \leftarrow \exp(-\gamma_k C R \mathrm{Diag}(1/g) - (\gamma_k \varepsilon - 1) \log(Q))$,

$\xi^{(2)} \leftarrow \exp(-\gamma_k C^T Q \mathrm{Diag}(1/g) - (\gamma_k \varepsilon - 1) \log(R))$,

$\omega \leftarrow \mathcal{D}(Q^T C R)$, $\xi^{(3)} \leftarrow \exp(\gamma_k \omega / g^2 - (\gamma_k \varepsilon - 1) \log(g))$,

$Q, R, g \leftarrow$ LR-IBP$((\xi^{(i)})_{1\leq i\leq 3}, a, b, \delta)$ (Algorithm (8))

**end**

**Result:** $\langle C, Q \mathrm{Diag}(1/g) R^T \rangle$

**Computational Cost.** Note that $(\xi^{(i)})_{1\leq i\leq 3}$ considered in algorithm (9) lives in $\mathbb{R}_+^{n\times r} \times \mathbb{R}_+^{m\times r} \times \times \mathbb{R}_+^r$ and therefore each iteration of algorithm (8) can be computed in $\mathcal{O}((n+m)r)$ algebraic operations as it involves only matrix/vector multiplications of the form $\xi^{(i)} v_i$ and $(\xi^{(i)})^T u_i$. However without any assumption on the cost matrix $C$, computing $(\xi^{(i)})_{1\leq i\leq 3}$ costs $\mathcal{O}(nmr)$ algebraic operations as it requires to compute both $CR$ and $C^T Q$ at each iteration. Thanks to assumption 1, such multiplications can be performed in $\mathcal{O}((n+m)dr)$ algebraic operations and thus algorithm (9) requires only a linear number of algebraic operations with respect to the number of samples at each iterations.

Figure 5.8: Plot of the Gaussian mixtures considered in Fig. 5.3.

## 5.11 Addiational Experiments

In Fig. 5.3, we compare two Gaussian mixture densities sampled with $n = m = 10000$ points in 2D. The two densities considered are

$$f_X(x) = \frac{1}{3}\frac{\exp\left((x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right)}{\sqrt{2\pi|\Sigma|}} + \frac{1}{3}\frac{\exp\left((x-\mu_2)^T\Sigma^{-1}(x-\mu_2)\right)}{\sqrt{2\pi|\Sigma|}} + \frac{1}{3}\frac{\exp\left((x-\mu_3)^T\Sigma^{-1}(x-\mu_3)\right)}{\sqrt{2\pi|\Sigma|}}$$

$$f_Y(x) = \frac{1}{2}\frac{\exp\left((x-\nu_1)^T\Sigma^{-1}(x-\nu_1)\right)}{\sqrt{2\pi|\Sigma|}} + \frac{1}{2}\frac{\exp\left((x-\nu_2)^T\Sigma^{-1}(x-\nu_2)\right)}{\sqrt{2\pi|\Sigma|}}$$

where

$$\mu_1 = [0,0], \quad \mu_2 = [0,1], \quad \mu_3 = [1,1],$$
$$\nu_1 = [0.5,0.5], \quad \nu_2 = [-0.5,0.5], \quad \Sigma = 0.05 \times \mathrm{Id}_2.$$

We show in Fig. 5.8 a plot of the two distributions considered. In Fig. 5.9, we consider the exact same setting as the one presented in Fig. 5.3 but we increase the dimension of the problem. More precisely we consider two Gaussian mixture densities samples with $n = m = 10000$ points in 10D where

$$\mu_1 = [0,\dots,0],\ \mu_2 = [0,1,0,\dots,0],\ \mu_3 = [1,1,0,\dots,0],$$
$$\nu_1 = [0.5,0.5,0,\dots,0],\ \nu_2 = [-0.5,0.5,0,\dots,0],$$
$$\Sigma = 0.05 \times \mathrm{Id}_{10}.$$

Similarly as in Fig. 5.3, we observe that **LOT** and **LOT Quad** provide similar results while **LOT** is faster. All kernel-based methods fail to converge in this

setting. Moreover we see that for small regularizations $\varepsilon$, our method is able to approximate faster than **Sin** the true OT thanks to the low-rank constraint. Note also that we observe again a difference between the two entropic regularizations of the **Sin** objective and **LOT** objective. Indeed the range of $\varepsilon$ where **Sin** provides an efficient approximation of the true OT is larger than the one of **LOT**. Indeed recall that for **LOT**, we regularize *twice* as we constraint the nonnegative rank of the couplings and we add an entropic term to regularize the objective.

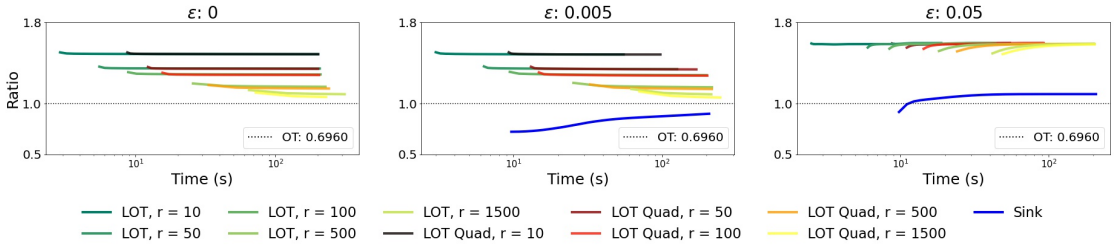



Figure 5.9: Comparison of the time-accuracy tradeoff for different methods for estimating the OT or its regularized version between two mixture of gaussians in 10D.

In Fig. 5.2, we compare the time-accuracy tradeoff for different methods on a synthetic problem where we aim at estimating either the OT or its regularized version between two gaussians in 2D. Here we consider the exact same setting but we increase the dimension of the problem: $d = 10$. As in Fig. 5.2, our proposed method obtains an efficient approximation of the OT or its regularized version for all rank $r$ faster than other low-rank methods in the regime of small $\varepsilon$. We also see that for all low-rank methods, a rank of $r = 500$ is not enough in this setting to obtain the exact OT, but as the rank increases, the approximation gets better.

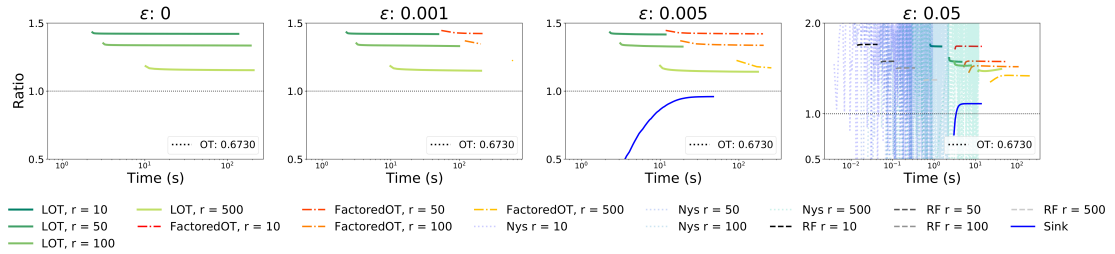



Figure 5.10: In this experiment, we consider two Gaussian distributions evaluated on $n = m = 5000$ in 10D. The first one has a mean of $(1, \dots, 1)^T \in \mathbb{R}^{10}$ and identity covariance matrix $I_{10}$ while the other has 0 mean and covariance $0.1 \times I_{10}$. The ground cost is the squared Euclidean distance.

## 5.12 Tight solution

Let $X = (x_1, \dots, x_n)$, $Y = (y_1, \dots, y_m)$, $a, b \in \Sigma_n, \Sigma_m$ be probability weights, and $Z = (z_1, \dots, z_k)$ be points in a set endowed with a cost $c$. We consider the network problem from sources $X$ to target $Y$ passing through $Z$. This is equivalent to solving the regular $n \times m$ OT problem with cost matrix $C_{ij} = \min_k c(x_i, z_k) + c(z_k, y_j)$. We write $k_{ij} = \operatorname{argmin}_k c(x_i, z_k) + c(z_k, y_j)$, $D = [c(x_i, z_k)]_{ik}$ and $D' = [c(z_k, y_j)]_{kj}$.

**Lemma 6.** *Let $P^\star$ be an optimal solution for the problem $\min_{P \in U(a,b)} \langle P, C \rangle$. Write*

$$g_k^* = \sum_{i,j} P_{ij} 1_{k=k_{ij}}, U_{ik}^* = \sum_j P_{ij} 1_{k=k_{ij}}, V_{kj}^* = \sum_i P_{ij} 1_{k=k_{ij}}$$

*Then matrices $U^* \in U(a, g^*), V^* \in U(g^*, b)$ and are respectively optimal for the OT problems with costs $D$ and $D'$ respectively. Additionally, $\langle P^\star, C \rangle = \langle U^*, D \rangle + \langle V^*, D' \rangle$.*

*Proof.* It is easy to check that $U^* \in U(a, g^*), V^* \in U(g^*, b)$ and that we have:

$$\langle P^\star, C \rangle = \langle U^*, D \rangle + \langle V^*, D' \rangle$$

Moreover let $U \in U(a, g^*), V \in U(g^*, b)$, then we have

$$\begin{aligned}
\langle P^\star, C \rangle \leq \langle C, UD(1/g^*)V \rangle &= \sum_k \frac{1}{g_k^*} \sum_{ij} C_{ij} U_{ik} V_{kj} \\
&= \sum_k \frac{1}{g_k^*} \sum_{i,j} \min_{k'} (D_{ik'} + D_{k'j}) U_{ik} V_{kj} \\
&\leq \sum_k \frac{1}{g_k^*} \sum_{i,j} (D_{ik} + D_{kj}) U_{ik} V_{kj} \\
&\leq \sum_k \frac{1}{g_k^*} \sum_{i,j} D_{ik} U_{ik} V_{kj} + \sum_{i,j} D_{kj} U_{ik} V_{kj} \\
&\leq \sum_k \sum_i D_{ik} U_{ik} + \sum_j D_{kj} V_{kj} \\
&= \langle U, D \rangle + \langle V, D' \rangle
\end{aligned}$$

Therefore for any $U \in U(a, g), V \in U(g, b)$ we have

$$\langle U^*, D \rangle + \langle V^*, D' \rangle \leq \langle U, D \rangle + \langle V, D' \rangle$$

from which follows the optimality of $U^*$ and $V^*$. □

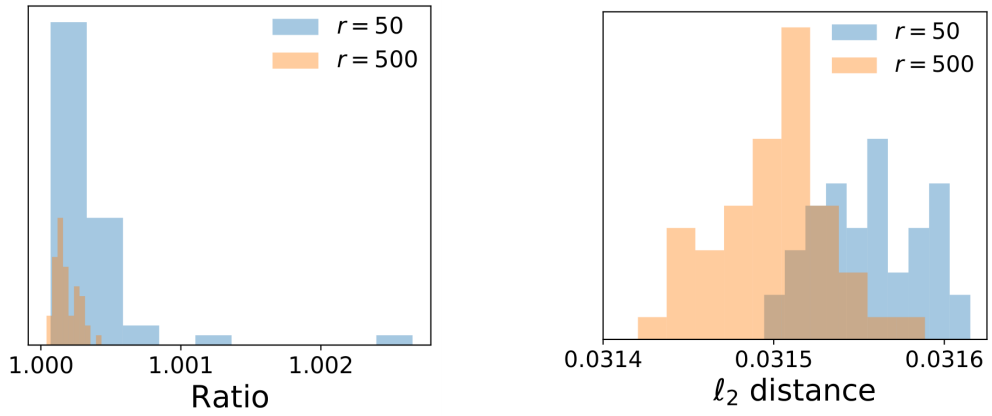



Figure 5.11: Here we consider the same setting as in Figure 5.9 where the cost functions is defined as $c(x, y) = \min_{k\in 1,..,r} \|x - z_k\| + \|z_k - y\|$ and $z_1, \ldots, z_r \in \mathbb{R}^{10}$ are fixed anchors.

**Proposition 5.12.1.** *$U^* D(1/g^*)V^*$ is optimal for the OT problem between $X$ and $Y$ with costs $C$.*

*Proof.* Obviously $U^* D(1/g^*)V^*$ has the right marginals. Moreover from the computation obtained in the proof of Lemma 6, we have

$$\langle C, U^* D(1/g^*)V^* \rangle \leq \langle U^*, D \rangle + \langle V^*, D' \rangle = \langle P^\star, C \rangle$$

from which follows the optimality of $U^* D(1/g^*)V^*$. □

In the following experiment we aim at showing that our method is able to recover the exact true solution of Eq. (5.1) when the optimal coupling admits a low nonnegative rank. Moreover we show that our algorithm is robust to the choice of the initialization. Indeed in Figure 5.11, we plot both the histograms of the ratios to the LP solution of LOT costs and the $\ell_2$ distance between the true optimal coupling and the coupling obtained by our algorithm for multiple random initializations. We show that our method is able to recover consistenly the true optimal coupling.

# Chapter 6

# Low-rank Optimal Transport: Theoretical Properties

The matching principles behind optimal transport (OT) play an increasingly important role in machine learning, a trend which can be observed when OT is used to disambiguate datasets in applications (e.g. single-cell genomics) or used to improve more complex methods (e.g. balanced attention in transformers or self-supervised learning). To scale to more challenging problems, there is a growing consensus that OT requires solvers that can operate on millions, not thousands, of points. The low-rank optimal transport (LOT) approach advocated in [3] holds several promises in that regard, and was shown to complement more established entropic regularization approaches, being able to insert itself in more complex pipelines, such as quadratic OT. LOT restricts the search for low-cost couplings to those that have a low-nonnegative rank, yielding linear time algorithms in cases of interest. However, these promises can only be fulfilled if the LOT approach is seen as a legitimate contender to entropic regularization when compared on properties of interest, where the scorecard typically includes theoretical properties (statistical complexity and relation to other methods) or practical aspects (debiasing, hyperparameter tuning, initialization). We target each of these areas in this paper in order to cement the impact of low-rank approaches in computational OT.

This chapter is based on [1].

## 6.1 Introduction

Optimal transport (OT) is used across data-science to put in correspondence different sets of observations. These observations may come directly from datasets, or, in more advanced applications, depict intermediate layered representations of data. OT theory provides a single grammar to describe and solve increasingly complex matching problems (linear, quadratic, regularized, unbalanced, etc...), making it gain a stake in various areas of science such as as single-cell biology [41, 56, 205], imaging [42, 43, 206] or neuroscience [44, 45].

**Regularized approaches to OT.** Solving OT problems at scale poses, however, formidable challenges. The most obvious among them is computational: the [32] problem on discrete measures of size $n$ is a linear program that requires $O(n^3 \log n)$ operations to be solved. A second and equally important challenge lies in the estimation of OT in high-dimensional settings, since it suffers from the curse-of-dimensionality [71]. The advent of regularized approaches, such as entropic regularization [76], has pushed these boundaries thanks for faster algorithms [74, 75] and improved statistical aspects [51]. Despite these clear strengths, regularized OT solvers remain, however, costly as they typically scale quadratically in the number of observations.

**Scaling up OT using low-rank couplings.** While it is always intuitively possible to reduce the size of measures (e.g. using $k$-means) prior to solving an OT between them, a promising line of work proposes to combine both [128, 3, 2]. Conceptually, these low-rank approaches solve simultaneously both an optimal clustering/aggregation strategy with the computation of an effective transport. This intuition rests on an explicit factorization of couplings into two sub-couplings. This has several computational benefits, since its computational cost becomes linear in $n$ if the ground cost matrix seeded to the OT problem has itself a low-rank. While these computational improvements, mostly demonstrated empirically, hold several promises, the theoretical properties of these methods are not yet well established. This stands in stark contrast to the Sinkhorn approach, which is comparatively much better understood.

**Our Contributions.** The goal of this paper is to advance our knowledge, understanding and practical ability to leverage low-rank factorizations in OT. This paper provides five contributions, targeting theoretical and practical properties of LOT: *(i)* We derive the rate of convergence of the low-rank OT to the true OT with respect to the non-nnegative rank parameter. *(ii)* We make a first step towards a better understanding of the statistical complexity of LOT by providing an upper-bound of

the statistical error, made when estimating LOT using the plug-in estimator; that upper-bound has a parametric rate $\mathcal{O}(\sqrt{1/n})$ that is independent of the dimension. *(iii)* We introduce a debiased version of LOT: as the Sinkhorn divergence [130], we show that debiased LOT is nonnegative, metrizes the weak convergence, and that it interpolates between the maximum mean discrepancy [18] and OT. *(iv)* We exhibit links between the bias induced by the low-rank factorization and clustering methods. *(v)* We propose practical strategies to tune the step-length and the initialization of the algorithm in [3].

## 6.2 Background on Low-rank Optimal Transport

Let $\mu \in \mathcal{M}_1^+(\mathcal{X})$, $\nu \in \mathcal{M}_1^+(\mathcal{Y})$ and $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ a nonnegative and continuous function. The Kantorovitch formulation of optimal transport between $\mu$ and $\nu$ is defined by

$$\mathrm{OT}_c(\mu, \nu) := \min_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y) d\pi(x, y)\,, \tag{6.1}$$

where the feasible set is the set of distributions over the product space $\mathcal{X} \times \mathcal{Y}$ with marginals $\mu$ and $\nu$:

$$\Pi(\mu, \nu) := \left\{\pi \in \mathcal{M}_1^+(\mathcal{X} \times \mathcal{Y}) \;\text{ s.t. }\; P_{1\#\pi} = \mu,\, P_{2\#\pi} = \nu\right\}\,,$$

with $P_{1\#\pi}$ (resp. $P_{2\#\pi}$), the pushforward probability measure of $\pi$ using the projection maps $P_1(x, y) = x$ (resp. $P_2(x, y) = y$). When there exists an optimal coupling solution of (6.1) supported on a graph of a function, we call such function a Monge map. In the discrete setting, one can reformulate the optimal transport problem as a linear program over the space of nonnegative matrices satisfying the marginal constraints. More precisely, let $a$ and $b$ be respectively elements of $\Delta_n^*$ and $\Delta_m^*$ and let also $X := \{x_1, \ldots, x_n\}$ and $Y := \{y_1, \ldots, y_m\}$ be respectively two subsets of $\mathcal{X}$ and $\mathcal{Y}$. By denoting $\mu_{a,X} := \sum_{i=1}^n a_i \delta_{x_i}$ and $\nu_{b,Y} := \sum_{j=1}^m b_j \delta_{y_j}$ the two discrete distributions associated and writing $C := [c(x_i, y_j)]_{i,j}$, the discrete optimal transport problem can be formulated as

$$\mathrm{OT}_c(\mu_{a,X}, \nu_{b,Y}) = \min_{P \in \Pi_{a,b}} \langle C, P\rangle \quad \text{where} \quad \Pi_{a,b} := \{P \in \mathbb{R}_+^{n\times m} \text{ s.t. } P\mathbf{1}_m = a, P^T\mathbf{1}_n = b\}\,. \tag{6.2}$$

Scetbon et al. [3] propose to constrain the discrete optimal transport problem to couplings that have a low-nonnegative rank:

**Definition 6.2.1.** *Given $M \in \mathbb{R}_+^{n\times m}$, the nonnegative rank of $M$ is defined by:* $\mathrm{rk}_+(M) := \min\{q | M = \sum_{i=1}^q R_i, \forall i, \mathrm{rk}(R_i) = 1, R_i \geq 0\}$.

Note that for any $M \in \mathbb{R}_+^{n\times m}$, we always have that $\mathrm{rk}_+(M) \leq \min(n,m)$. For $r \geq 1$, we consider the set of couplings satisfying marginal constaints with nonnegative-rank of at most $r$ as $\Pi_{a,b}(r) := \{P \in \Pi_{a,b},\ \mathrm{rk}_+(P) \leq r\}$. The discrete Low-rank Optimal Transport (LOT) problem is defined by:

$$\mathrm{LOT}_{c,r}(\mu_{a,X}, \nu_{b,Y}) := \min_{P\in\Pi_{a,b}(r)} \langle C, P\rangle \,. \tag{6.3}$$

To solve this problem, [3] show that Problem (6.3) is equivalent to

$$\min_{(Q,R,g)\in\mathcal{C}_1(a,b,r)\cap\mathcal{C}_2(r)} \langle C, Q\,\mathrm{diag}(1/g)R^T\rangle \,, \tag{6.4}$$

where $\mathcal{C}_1(a,b,r) := \left\{(Q,R,g) \in \mathbb{R}_+^{n\times r}\times\mathbb{R}_+^{m\times r}\times(\mathbb{R}_+^*)^r \text{ s.t. } Q\mathbf{1}_r = a, R\mathbf{1}_r = b\right\}$ and $\mathcal{C}_2(r) := \left\{(Q,R,g) \in \mathbb{R}_+^{n\times r}\times\mathbb{R}_+^{m\times r}\times\mathbb{R}_+^{r} \text{ s.t. } Q^T\mathbf{1}_n = R^T\mathbf{1}_m = g\right\}$. They propose to solve it using a mirror descent scheme and prove the non-asymptotic stationary convergence of their algorithm. While [3] only focus on the discrete setting, we consider here its extension for arbitrary probability measures. Following [128], we define the set $\Pi_r(\mu,\nu)$ of rank-$r$ couplings satisfying marginal constraints by:

$$\{\pi \in \Pi(\mu,\nu) : \exists (\mu_i)_{i=1}^r \in \mathcal{M}_1^+(\mathcal{X})^r,\ (\nu_i)_{i=1}^r \in \mathcal{M}_1^+(\mathcal{Y})^r,\ \lambda \in \Delta_r^* \text{ s.t. } \pi = \sum_{i=1}^r \lambda_i \mu_i \otimes \nu_i\} \,.$$

This more general definition of LOT between $\mu \in \mathcal{M}_1^+(\mathcal{X})$ and $\nu \in \mathcal{M}_1^+(\mathcal{Y})$ reads:

$$\mathrm{LOT}_{c,r}(\mu,\nu) := \inf_{\pi\in\Pi_r(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)d\pi(x,y) \,. \tag{6.5}$$

Note that this definition of $\mathrm{LOT}_{c,r}$ is consistent as it coincides with the one defined in (6.3) on discrete probability measures. Observe also that $\Pi_r(\mu,\nu)$ is compact for the weak topology and therefore the infimum in (6.5) is attained. See Appendix 6.8 for more details.

## 6.3 Approximation Error of LOT to original OT as a function of rank

Our goal in this section is to obtain a control of the error induced by the low-rank constraint when trying to approximate the true OT cost. We provide first a control of the approximation error in the discrete setting. The proof is given in Appendix 6.9.1.

**Proposition 6.3.1.** *Let $n, m \geq 2$, $X := \{x_1, \ldots, x_n\} \subset \mathcal{X}$, $Y := \{y_1, \ldots, y_m\} \subset \mathcal{Y}$ and $a \in \Delta_n^*$ and $b \in \Delta_m^*$. Then for $2 \leq r \leq \min(n, m)$, we have that*

$$|\mathrm{LOT}_{r,c}(\mu_{a,X}, \nu_{b,Y}) - \mathrm{OT}_c(\mu_{a,X}, \nu_{b,Y})| \leq \|C\|_\infty \; ln(\min(n, m)/(r-1))$$

**Remark 9.** *Note that this result improves the control obtained in [207], where they obtain that $|\mathrm{LOT}_{r,c}(\mu_{a,X}, \nu_{b,Y}) - \mathrm{OT}_c(\mu_{a,X}, \nu_{b,Y})| \lesssim \|C\|_\infty \sqrt{nm}(\min(n, m) - r)$ as we have for any $z, z' \geq 1$, $|\ln(z) - \ln(z')| \leq |z - z'|$.*

It is in fact possible to obtain another control of the approximation error by partitioning the space where the measures are supported. Let us now introduce the notion of entropy numbers.

**Definition 6.3.1.** *Let $(\mathcal{Z}, d)$ a metric space, $\mathcal{W} \subset \mathcal{Z}$ and $k \geq 1$ an integer. Then by denoting $B_{\mathcal{Z}}(z, \varepsilon) := \{y \in \mathcal{Z} : \; d(z, y) \leq \varepsilon\}$, we define the $k$-th (dyadic) entropy number of $\mathcal{W}$ as*

$$\mathcal{N}_k(\mathcal{W}, d) := \inf\{\varepsilon \; s.t. \; \exists \; z_1, \ldots, z_{2^k} \in \mathcal{Z} : \quad \mathcal{W} \subset \cup_{i=1}^{2^k} B_{\mathcal{Z}}(z_i, \varepsilon)\} \, .$$

For example, any compact set $\mathcal{W}$ of $\mathbb{R}^d$ admits finite entropy numbers, and by denoting $R := \sup_{w \in \mathcal{W}} \|w\|_2$, we have $\mathcal{N}_k(\mathcal{W}, \|\cdot\|_2) \leq 4R/2^{k/d}$. We obtain next a control of the approximation error of $\mathrm{LOT}_{c,r}$ to the true OT cost using entropy numbers (see proof in Appendix 6.9.2).

**Proposition 6.3.2.** *Let $\mu \in \mathcal{M}_1^+(\mathcal{X})$, $\nu \in \mathcal{M}_1^+(\mathcal{Y})$ and assume that $c$ is $L$-Lipschitz w.r.t. $x$ and $y$. Then for any $r \geq 1$, we have*

$$|\mathrm{LOT}_{c,r}(\mu, \nu) - \mathrm{OT}_c(\mu, \nu)| \leq 2L \max(\mathcal{N}_{\lfloor \log_2(\lfloor \sqrt{r} \rfloor) \rfloor}(\mathcal{X}, d_{\mathcal{X}}), \mathcal{N}_{\lfloor \log_2(\lfloor \sqrt{r} \rfloor) \rfloor}(\mathcal{Y}, d_{\mathcal{Y}}))$$

This results in the following bound for the $p$-Wasserstein distance for any $p \geq 1$ on $\mathbb{R}^d$.

**Corollary 6.3.1.** *Let $d \geq 1$, $p \geq 1$, $\mathcal{X}$ a compact subspace of $\mathbb{R}^d$ and $\mu, \nu \in \mathcal{M}_1^+(\mathcal{X})$. By denoting $R := \sup_{x \in \mathcal{X}} \|x\|_2$, we obtain that for any $r \geq 1$,*

$$|\mathrm{LOT}_{\|\cdot\|_2^p, r}(\mu, \nu) - \mathrm{OT}_{\|\cdot\|_2^p}(\mu, \nu)| \leq 4dp \frac{(8R^2)^p}{r^{p/2d}} \, .$$

As per the Proof of Proposition 6.3.2 we can provide a tighter control, assuming a Monge map exists.

**Corollary 6.3.2.** *Under the same assumptions of Proposition 6.3.2 and by assuming in addition that there exists a Monge map solving $\mathrm{OT}_c(\mu, \nu)$, we obtain that for any $r \geq 1$,*

$$|\mathrm{LOT}_{c,r}(\mu, \nu) - \mathrm{OT}_c(\mu, \nu)| \leq L\mathcal{N}_{\lfloor \log_2(r) \rfloor}(\mathcal{Y}, d_{\mathcal{Y}}) \, .$$

When $\mathcal{X} = \mathcal{Y}$ are a subspaces of $\mathbb{R}^d$, a sufficient condition for a Monge map to exists is that either $\mu$ or $\nu$ is absolutely continuous with respect to the Lebesgue measure and that $c$ is of the form $h(x-y)$ where $h : \mathcal{X} \to \mathbb{R}_+$ is a strictly convex function [111, Theorem 1.17]. Therefore if $\mu$ is absolutely continuous with respect to the Lebesgue measure, we obtain for any $r \geq 1$ and $p > 1$

$$|\mathrm{LOT}_{\|\cdot\|_2^p,r}(\mu,\nu) - \mathrm{OT}_{\|\cdot\|_2^p}(\mu,\nu)| \leq 2dp\frac{(8R^2)^p}{r^{p/d}} \ .$$

## 6.4 Sample Complexity of LOT

We now focus on the statistical performance of the plug-in estimator for LOT. In the following we assume that $\mathcal{X} = \mathcal{Y}$ for simplicity. Given $\mu, \nu \in \mathcal{M}_1^+(\mathcal{X})$, we denote the empirical measures associated $\hat{\mu}_n := \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ and $\hat{\nu}_n := \frac{1}{n}\sum_{i=1}^n \delta_{Y_i}$, where $(X_i, Y_i)_{i=1}^n$ are sampled independently from $\mu \otimes \nu$. We consider the plug-in estimator defined as $\mathrm{LOT}_{c,r}(\hat{\mu}_n, \hat{\nu}_n)$, and we aim at quantifying the rate at which it converges towards the true low-rank optimal transport cost $\mathrm{LOT}_{c,r}(\mu,\nu)$. Before doing so, in the next Proposition we show that this estimator is consistent on compact spaces. The proof is given in Appendix 6.9.3.

**Proposition 6.4.1.** *Let $r \geq 1$ and $\mu, \nu \in \mathcal{M}_1^+(\mathcal{X})$, then* $\mathrm{LOT}_{c,r}(\hat{\mu}_n, \hat{\nu}_n) \xrightarrow[n\to+\infty]{} \mathrm{LOT}_{c,r}(\mu,\nu)$ *a.s.*

Next we aim at obtaining the convergence rates of our plug-in estimator. In the following Proposition, we obtain a non-asymptotic upper-bound of the statistical error. See Appendix 6.9.4 for the proof.

**Proposition 6.4.2.** *Let $r \geq 1$ and $\mu, \nu \in \mathcal{M}_1^+(\mathcal{X})$. Then, there exists a constant $K_r$ such that for any $\delta > 0$ and $n \geq 1$, we have, with a probability of at least $1 - 2\delta$, that*

$$\begin{aligned}\mathrm{LOT}_{c,r}(\hat{\mu}_n, \hat{\nu}_n) \leq \mathrm{LOT}_{c,r}(\mu,\nu) &+ 11\|c\|_\infty\sqrt{\frac{r}{n}} \\ &+ K_r\|c\|_\infty\left[\sqrt{\frac{\log(40/\delta)}{n}} + \frac{\sqrt{r}\log(40/\delta)}{n}\right] .\end{aligned}$$

This result is, to the best of our knowledge, the first attempt at providing a statistical control of low-rank optimal transport. We provide an upper-bound of the plug-in estimator which converges towards $\mathrm{LOT}_{c,r}$ at a parametric rate and which is independent of the dimension on general compact metric spaces. While we fall short of providing a lower bound that could match that upper bound, and therefore provide a complete statistical complexity result, we believe this result might provide

a first explanation on why, in practice, $\mathrm{LOT}_{c,r}$ displays better statistical properties than unregularized OT and its curse of dimensionality [129]. In addition, that upper bound compares favorably to known results on entropic optimal transport. The rate of entropy regularized OT does not depend on the ambient dimension with respect to $n$, but carries an exponential dependence in dimension with respect to the regularization parameter $\varepsilon$ [82]. By contrast, the term associated with the nonnegative rank $r$ in our bound has no direct dependence on dimension.

Our next aim is to obtain an explicit rate with respect to $r$ and $n$. In Proposition 6.4.2, we cannot control explicitly $K_r$ in the general setting. Indeed, in our proof, we obtain that $K_r := 14/\min_i \lambda_i^*$ where $(\lambda_i^*)_{i=1}^r \in \Delta_r^*$ are the weights involved in the decomposition of one optimal solution of the true $\mathrm{LOT}_{c,r}(\mu,\nu)$. Therefore the control of $K_r$ requires additional assumptions on the optimal solutions of $\mathrm{LOT}_{c,r}(\mu,\nu)$. In the following Proposition, we obtain an explicit upper-bound of the plug-in estimator with respect to $r$ and $n$ in the asymptotic regime.

**Proposition 6.4.3.** *Let $r \geq 1$, $\delta > 0$ and $\mu, \nu \in \mathcal{M}_1^+(\mathcal{X})$. Then there exists a constant $N_{r,\delta}$ such that if $n \geq N_{r,\delta}$ then with a probability of at least $1-2\delta$, we have*

$$\mathrm{LOT}_{c,r}(\hat{\mu}_n, \hat{\nu}_n) \leq \mathrm{LOT}_{c,r}(\mu,\nu) + 11\|c\|_\infty \sqrt{\frac{r}{n}} + 77\|c\|_\infty \sqrt{\frac{\log(40/\delta)}{n}} \,.$$

Note that one cannot recover the result obtained in Proposition 6.4.3 from the one obtained in Proposition 6.4.2 as we have that $K_r \geq 14r \xrightarrow[r\to+\infty]{} +\infty$. In order to prove the above result, we use an extension of the McDiarmid's inequality when differences are bounded with high probability [208]. See proof in Appendix 6.9.5 for more details.

## 6.5 Debiased Formulation of LOT

We introduce here the debiased formulation of $\mathrm{LOT}_{c,r}$ and show that it is able to distinguish two distributions, metrize the convergence in law and can be used as a new objective in order to learn distributions. We focus next on the debiasing terms involving measures with themselves $\mathrm{LOT}_{c,r}(\mu,\mu)$ in this new divergence, and show that they can be interpreted as defining a new clustering method generalizing $k$-means for any geometry.

### 6.5.1 On the Proprieties of the Debiased Low-rank Optimal Transport

When it comes to learn (or generate) a distribution in ML applications given samples, it is crucial to consider a divergence that is able to distinguish between

two distributions and metrize the convergence in law. In general, $\mathrm{LOT}_{c,r}(\mu,\mu) \neq 0$ and the minimum of $\mathrm{LOT}_{c,r}(\nu,\mu)$ with respect to $\nu$ will not necessarily recover $\mu$. In order to alleviate this issue we propose a debiased version of $\mathrm{LOT}_{c,r}$ defined for any $\mu,\nu \in \mathcal{M}_1^+(\mathcal{X})$ as

$$\mathrm{DLOT}_{c,r}(\mu,\nu) := \mathrm{LOT}_{c,r}(\mu,\nu) - \frac{1}{2}[\mathrm{LOT}_{c,r}(\mu,\mu) + \mathrm{LOT}_{c,r}(\nu,\nu)] \; .$$

Note that $\mathrm{DLOT}_{c,r}(\nu,\nu) = 0$. In the next Proposition, we show that, as the Sinkhorn divergence [209, 130], $\mathrm{DLOT}_{c,r}$ interpolates between the Maximum Mean Discrepancy (MMD) and OT. See proof in Appendix 6.9.6.

**Proposition 6.5.1.** *Let $\mu,\nu \in \mathcal{M}_1^+(\mathcal{X})$. Let us assume that $c$ is symmetric, then we have*

$$\mathrm{DLOT}_{1,c}(\mu,\nu) = \frac{1}{2}\int_{\mathcal{X}^2} -c(x,y)d[\mu-\nu]\otimes d[\mu-\nu](x,y) \; .$$

*If in addition we assume the $c$ is Lipschitz w.r.t to $x$ and $y$, then we have*

$$\mathrm{DLOT}_{c,r}(\mu,\nu) \xrightarrow[r\to+\infty]{} \mathrm{OT}_c(\mu,\nu) \; .$$

Next, we aim at showing some useful properties of the debiased low-rank OT for machine learning applications. For that purpose, let us first recall some definitions.

**Definition 6.5.1.** *We say that the cost $c : \mathcal{X}\times\mathcal{X} \to \mathbb{R}_+$ is a semimetric on $\mathcal{X}$ if for all $x,x' \in \mathcal{X}$, $c(x,x') = c(x',x)$ and $c(x,x') = 0$ if and only if $x = x'$. In addition we say that $c$ has a negative type if $\forall n \geq 2$, $x_1,\dots,x_n \in \mathcal{X}$ and $\alpha_1,\dots,\alpha_n \in \mathbb{R}$ such that $\sum_{i=1}^n \alpha_i = 0$, $\sum_{i,j=1}^n \alpha_i\alpha_j c(x_i,x_j) \leq 0$. We say also that $c$ has a strong negative type if for all $\mu,\nu \in \mathcal{M}_1^+(\mathcal{X})$, $\mu \neq \nu \implies \int_{\mathcal{X}^2} c(x,y)d[\mu-\nu]\otimes[\mu-\nu] < 0$.*

Note that if $c$ has a strong negative type, then $c$ has a negative type too. For example, all Euclidean spaces and even separable Hilbert spaces endowed with the metric induced by their inner products have strong negative type. Also, on $\mathbb{R}^d$, the squared Euclidean distance has a negative type [210]. We can now provide stronger geometric guarantees for $\mathrm{DLOT}_{c,r}$. In the next Proposition, we show that for a large class of cost functions, $\mathrm{DLOT}_{c,r}$ is nonnegative, able to distinguish two distributions, and metrizes the convergence in law. The proof is given in Appendix 6.9.8.

**Proposition 6.5.2.** *Let $r \geq 1$, and let us assume that $c$ is a semimetric of negative type. Then for all $\mu,\nu \in \mathcal{M}_1^+(\mathcal{X})$, we have that*

$$\mathrm{DLOT}_{c,r}(\mu,\nu) \geq 0 \; .$$

*In addition, if $c$ has strong negative type then we have also that*

$$\begin{aligned} \mathrm{DLOT}_{c,r}(\mu,\nu) = 0 &\iff \mu = \nu \quad \textit{and} \\ \mu_n \to \mu &\iff \mathrm{DLOT}_{c,r}(\mu_n,\mu) \to 0\ . \end{aligned}$$

*where the convergence of the sequence of probability measures considered is the convergence in law.*

Observe that when $c$ has strong negative type, $\nu \to \mathrm{DLOT}_{c,r}(\nu,\mu) \geq 0$ and it admits a unique global minimizer at $\nu = \mu$. Therefore, $\mathrm{DLOT}_{c,r}$ has desirable properties to be used as a loss.It is also worth noting that, in order to obtain the metrization of the convergence in law, we show the following Proposition. See proof in Appendix 6.9.7.

**Proposition 6.5.3.** *Let $r \geq 1$ and $(\mu_n)_{n\geq 0}$ and $(\nu_n)_{n\geq 0}$ two sequences of probability measures such that $\mu_n \to \mu$ and $\nu_n \to \nu$ with respect to the convergence in law. Then we have that*

$$\mathrm{LOT}_{c,r}(\mu_n,\nu_n) \to \mathrm{LOT}_{c,r}(\mu,\nu)\ .$$

### 6.5.2 Low-Rank Transport Bias and Clustering

We turn next to the debiasing terms appearing in DLOT and exhibit links between LOT and clustering methods. Indeed, in the discrete setting, the low-rank bias of a probability measure $\mu$ defined as $\mathrm{LOT}_{c,k}(\mu,\mu)$ can be seen as a generalized version of the $k$-means method for any geometry. In the next Proposition we obtain a new formulation of $\mathrm{LOT}_{c,k}(\mu,\mu)$ viewed as a general clustering method on arbitrary metric space. See proof in Appendix 6.9.9.

**Proposition 6.5.4.** *Let $n \geq k \geq 1$, $X := \{x_1,\dots,x_n\} \subset \mathcal{X}$ and $a \in \Delta_n^*$. If $c$ is a semimetric of negative type, then by denoting $C = (c(x_i,x_j))_{i,j}$, we have that*

$$\mathrm{LOT}_{c,k}(\mu_{a,X},\mu_{a,X}) = \min_{Q} \langle C, Q\,diag(1/Q^T\mathbf{1}_n)Q^T\rangle \ : Q \in \mathbb{R}_+^{n\times k},\ Q\mathbf{1}_k = a\ . \quad (6.6)$$

Let us now explain in more details the link between (6.6) and $k$-means. When $\mathcal{X}$ is a subspace of $\mathbb{R}^d$, $c$ is the squared Euclidean distance and $a = \mathbf{1}_n$, we recover exactly the $k$-means algorithm.

**Corollary 6.5.1.** *Let $n \geq k \geq 1$ and $X := \{x_1,\dots,x_n\} \subset \mathbb{R}^d$. We have that*

$$\frac{\mathrm{LOT}_{\|\cdot\|_2^2,k}(\mu_{\mathbf{1}_n,X},\mu_{a,X})}{2} = \min_{Q,z_1,\dots,z_k} \sum_{i=1}^{n}\sum_{q=1}^{k} Q_{i,q}\|x_i - z_q\|_2^2 \ : Q \in \{0,1\}^{n\times k},\ Q\mathbf{1}_k = \mathbf{1}_n\ .$$

In the general setting, solving $\mathrm{LOT}_{c,k}(\mu_{a,X},\mu_{a,X})$ for a given geometry $c$, and a prescribed histrogram $a$ offers a new clustering method where the assignment of the points to the clusters is determined by the matrix $Q^*$ solution of (6.6).

## 6.6 Computing LOT: Adaptive Stepsizes and Better Initializations

We target in this section practical issues that arises when using [3, Algo.3] to solve (6.4). Scetbon et al. [3] propose to apply a mirror descent scheme with respect to the Kullback-Leibler divergence which boils down to solve at each iteration $k \geq 0$ the following convex problem using the Dykstra's Algorithm [196]:

$$(Q_{k+1}, R_{k+1}, g_{k+1}) := \underset{\boldsymbol{\zeta} \in \mathcal{C}_1(a,b,r) \cap \mathcal{C}_2(r)}{\operatorname{argmin}} \operatorname{KL}(\boldsymbol{\zeta}, \boldsymbol{\xi}_k)\,. \tag{6.7}$$

where $(Q_0, R_0, g_0) \in \mathcal{C}_1(a,b,r) \cap \mathcal{C}_2(r)$, $\boldsymbol{\xi}_k := (\xi_k^{(1)}, \xi_k^{(2)}, \xi_k^{(3)})$, $\xi_k^{(1)} := Q_k \odot \exp(-\gamma_k C R_k \operatorname{diag}(1/g_k))$, $\xi_k^{(2)} := R_k \odot \exp(-\gamma_k C^T Q_k \operatorname{diag}(1/g_k))$, $\xi_k^{(3)} := g_k \odot \exp(\gamma_k \omega_k / g_k^2)$ with $[\omega_k]_i := [Q_k^T C R_k]_{i,i}$ for all $i \in \{1, \dots, r\}$, $\operatorname{KL}(\mathbf{w}, \mathbf{r}) := \sum_i w_i \log(w_i / r_i)$ and $(\gamma_k)_{k \geq 0}$ is a sequence of positive step sizes. In the general setting, each iteration of their algorithm requires $\mathcal{O}(nmr)$ operations and when the ground cost matrix $C$ admits a low-rank factorization of the form $C = AB^T$ where $A \in \mathbb{R}^{n \times q}$ and $B \in \mathbb{R}^{m \times q}$ with $q \ll \min(n, m)$, then the total complexity per iteration becomes linear $\mathcal{O}((n+m)rq)$. Note that for the squared Euclidean cost on $\mathbb{R}^d$, we have that $q = d + 2$. In the following we investigate two practical aspects of the algorithm: the choice of the step sizes and the initialization.

**Adaptive choice of $\gamma_k$.** [3] show experimentally that the choice of $(\gamma_k)_{k \geq 0}$ does not impact the solution obtained upon convergence, but rather the speed at which it is attained. Indeed the larger $\gamma_k$ is, the faster the algorithm will converge. As a result, their algorithm simply relies on a fixed $\gamma$ schedule. However, the range of admissible $\gamma$ depends on the problem considered and it may vary from one problem to another. Indeed, the algorithm might fail to converge as one needs to ensure at each iteration $k$ of the mirror descent scheme that the kernels $\boldsymbol{\xi}_k$ do not admit 0 entries in order to solve (6.7) using the Dykstra's Algorithm. Such a situation can occur when the terms involved in the exponentials become too large which may depend on the problem considered. Therefore, it may be of particular interest for practitioners to have a generic range of admissible values for $\gamma$ independently of the considered problem, in order to alleviate parameter tuning issues. We propose to consider instead an adaptive choice of $(\gamma_k)_{k \geq 0}$ along iterations. [211, 212] have proposed adaptive mirror descent schemes where, at each iteration, the step-size is normalized by the squared dual-norm of the gradient. Applying such a strategy in our case amounts to consider at each iteration

$$\gamma_k = \frac{\gamma}{\| \left(CR \operatorname{diag}(1/g), C^T Q \operatorname{diag}(1/g), -\mathcal{D}(Q^T R C)/g^2\right) \|_\infty^2}, \tag{6.8}$$

where the initial $\gamma > 0$ is fixed. By doing so, we are able to guarantee a lower-bound of the exponential terms involved in the expression of the kernels $\boldsymbol{\xi}_k$ at each iteration and prevent them from having 0 entries. We recommend to set such as global $\gamma \in [1, 10]$, and observe that this range works whatever the problem considered.

**On the choice of the initialization.** As $\mathrm{LOT}_{c,r}$ (6.4) is a non-convex optimization problem, the question of choosing an efficient initialization arises in practice. [3] show experimentally that the convergence of the algorithm does not depend on the initalization chosen if no stopping criterion is used. Indeed, their experimental findings support that only well behaved local minimas are attractive. However, in practice one needs to use a stopping criterion in order to terminate the algorithm. We do observe in many instances that using trivial initializers may result in spurious local minima, which trigger the stopping criterion early on and prevent the algorithm to reach a good solution. Based on various experimentations, we propose to consider a novel initialization of the algorithm. Our initialization aims at being close to a well-behaved local minimum by clustering the input measures. When the measures are supported on Euclidean space, we propose to find $r$ centroids $(z_i)_{i=1}^r$ of one of the two input discrete probability measures using $k$-means and to solve the following convex barycenter problem:

$$\begin{aligned} &\min_{Q,R} \langle C_{X,Z}, Q\rangle + \langle C_{Y,Z}, R\rangle - \varepsilon H(Q) - \varepsilon H(R) \;\; \text{s.t.} \\ &Q\mathbf{1}_n = a, \quad R\mathbf{1}_n = b, \quad Q^T\mathbf{1}_r = R^T\mathbf{1}_r\,, \end{aligned} \tag{6.9}$$

where $C_{X,Z} = (c(x_i, z_j))_{i,j}$, $C_{Y,Z} = (c(y_i, z_j))_{i,j}$, and $H(P) = -\sum_{i,j} P_{i,j}(\log(P_{i,j}-1)$. In practice we fix $\varepsilon = 1/10$ and we then initialize $\mathrm{LOT}_{c,r}$ using $(Q, R)$ solution of (6.9) and $g := Q^T\mathbf{1}_r (= R^T\mathbf{1}_r)$. Note that $(Q, R, g)$ is an admissible initialization and finding the centroids as well as solving (6.9) requires $\mathcal{O}((n+m)r)$ algebraic operations. Therefore such initialization does not change the total complexity of the algorithm. In the general (non-Euclidean) case, we propose to initialize the algorithm by applying our generalized $k$-means approach defined in (6.6) on each input measure where we fix the common marginal to be $g = \mathbf{1}_r/r$. More precisely, by denoting $C_{X,X} = (c(x_i, x_j))_{i,j}$ and $C_{Y,Y} = (c(y_i, y_j))_{i,j}$, we initialize the algorithm by solving:

$$\begin{aligned} &Q \in \operatorname*{argmin}_Q \langle C_{X,X}, Q\mathrm{diag}(1/Q^T\mathbf{1}_n)Q^T\rangle \;\; \text{s.t.} \;\; Q \in \mathbb{R}_+^{n\times k} \;\;, \;\; Q\mathbf{1}_k = a, \;\; Q^T\mathbf{1}_n = \mathbf{1}_r/r\,. \\ &R \in \operatorname*{argmin}_R \langle C_{Y,Y}, R\mathrm{diag}(1/R^T\mathbf{1}_m)R^T\rangle \;\; \text{s.t.} \;\; R \in \mathbb{R}_+^{m\times k} \;\;, \;\; R\mathbf{1}_k = b, \;\; R^T\mathbf{1}_n = \mathbf{1}_r/r\,. \end{aligned} \tag{6.10}$$

Note that again the $(Q, R, g)$ obtained is an admissible initialization and the complexity of solving (6.10) is of the same order as solving (6.4), thus the total complexity of the algorithm remains the same.

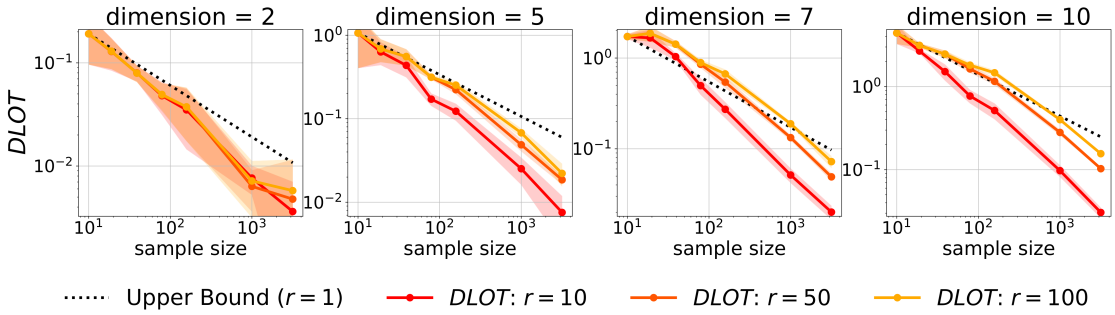



Figure 6.1: In this experiment, we consider a mixture of 10 anisotropic Gaussians supported on $\mathbb{R}^d$ and we plot the value of $\mathrm{DLOT}_{c,r}$ between two independent empirical measures associated to this mixture when varying the number of samples $n$ and the dimension $d$ for multiple ranks $r$. The ground cost considered is the squared Euclidean distance. Note that $\mathrm{LOT}_r(\mu,\mu) \neq 0$ and therefore we use $\mathrm{DLOT}_{c,r}(\mu,\mu)$ instead to evaluate the rates. Each point has been obtained by repeating 10 times the experiment. We compare the empirical rates obtained with the theoretical one derived in Proposition 6.4.2 for $r = 1$. We observe that our theoretical results match the empirical ones and, as expected, the rates do not depend on $d$.

## 6.7 Experiments

In this section, we illustrate experimentally our theoretical findings and show how our initialization provide practical improvements. For that purpose we consider 3 synthetic problems and one real world dataset to: *(i)* provide illustrations on the statistical rates of $\mathrm{LOT}_{c,r}$, *(ii)* exhibit the gradient flow of the debiased formulation $\mathrm{DLOT}_{c,r}$, *(iii)* use the clustering method induced by $\mathrm{LOT}_{c,r}$, and *(iv)* show the effect of the initialization. All experiments were run on a MacBook Pro 2019 laptop.

**Statistical rates.** We aim at showing the statistical rates of the plug-in estimator of $\mathrm{LOT}_{c,r}$. As $\mathrm{LOT}_{c,r}(\mu,\mu) \neq 0$ and as we do not have access to this value given samples from $\mu$, we consider instead the debiased version of the low-rank optimal transport, $\mathrm{DLOT}_{c,r}$. In figure 6.1, we show that the empiricial rates match the theoretical bound obtained in Proposition 6.4.2. In particular, we show that that these rates does not depend on the dimension of the ground space. Note also that we recover our theoretical dependence with respect to the rank $r$: the higher the rank, the slower the convergence.

**Gradient Flows using DLOT.** We illustrate here a practical use of DLOT for ML application. In figure 6.6, we consider $Y_1, \ldots, Y_n$ independent samples from a moon shape distribution in 2D, and by denoting $\hat{\nu}_n$ the empirical measure associated, we show the iterations obtained by a gradient descent scheme on the

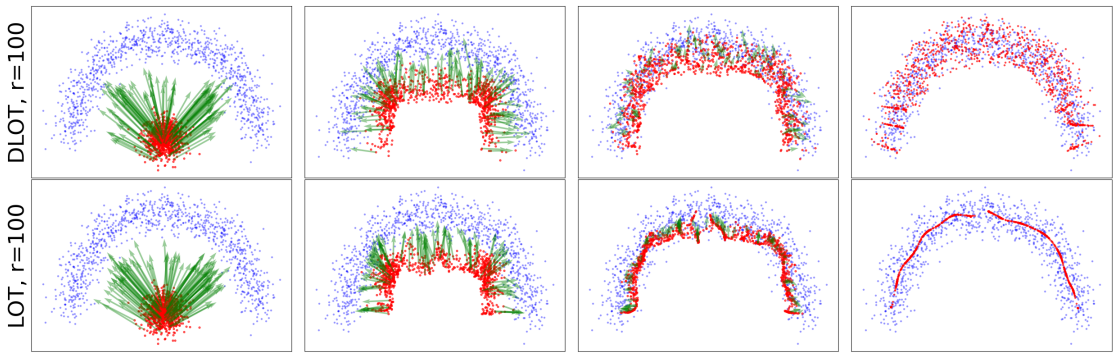



Figure 6.2: We compare the gradient flows $(\mu_t)_{t\geq 0}$ (in red) starting from a Gaussian distribution, $\mu_0$, to a moon shape distribution (in blue), $\nu$, in 2D when minimizing either $L(\mu) := \mathrm{DLOT}_{c,r}(\mu,\nu)$ or $L(\mu) := \mathrm{LOT}_{c,r}(\mu,\nu)$. The ground cost is the squared Euclidean distance and we fix $r = 100$. We consider 1000 samples from each distribution and and we plot the evolution of the probability measure obtained along the iterations of a gradient descent scheme. We also display in green the vector field in the descent direction. We show that the debiased version allows to recover the target distribution while $\mathrm{LOT}_{c,r}$ is learning a biased version with a low-rank structure.

following optimization problem:

$$\min_{\mathbf{X}\in\mathbb{R}^{n\times 2}} \mathrm{DLOT}_{c,r}(\mu_{\mathbf{1}_n/n,\mathbf{X}}, \hat{\nu}_n) \ .$$

We initialize the algorithm using $n = 1000$ samples drawn from a Gaussian distribution. We show that the gradient flow of our debiased version is able to recover the target distribution. We also compare it with the gradient flow of the biased version (LOT) and show that it fails to reproduce the target distribution as it is learning a biased one with a low-rank structure.

**Application to Clustering.** In this experiment we show some applications of the clustering method induced by $\mathrm{LOT}_{c,r}$. In figure 6.3, we consider 6 datasets with different structure and we aim at recovering the clusters using (6.6) for some well chosen costs. We compare the clusters obtained when considering either the squared Euclidean cost (which amounts at applying the $k$-means) and the shortest-path distance on the data viewed as a graph. We show that our method is able to recover the clusters on these settings for well chosen costs and therefore the proposed algorithm in [3] can be seen as a new alternative in order to clusterize data.

**Effect of the Initialization.** Our goal here is to show the effect of the initialization. In figure 6.4, we display the evolution of the cost as well as the value of

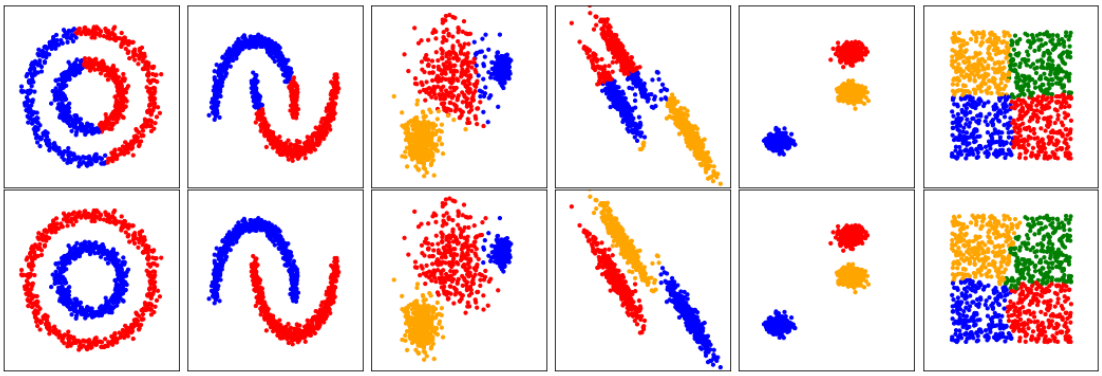


Figure 6.3: In this experiment, we draw 1000 samples from multiple distributions from the python package scikit-learn [213] and we apply the method proposed in (6.6) for two different costs: in the top row we consider the squared Euclidean distance while in the bottom row, we consider the shortest path distance on the graph associated with the ground cost $c(x, y) = 1 - k(x, y)$ where $k$ is a Gaussian kernel. In the two first problem (starting from the left), we fix $r = 2$, in the next three problem we fix $r = 3$ and in the last one we fix $r = 4$. We observe that the flexibility of our method allows to recover the clustering for a well chosen ground cost.

the stopping criterion along the iterations of the MD scheme solving (6.4) when considering different initialization. The $x$-axis corresponds to the total number of algebraic operations. This number is computed at each iteration of the outer loop of the algorithm proposed in [3] and is obtained by computing the complexity of all the operations involved in their algorithm to reach it. We consider this notion of time instead of CPU/GPU time as we do not want to be architecture/machine dependent. Recall also that the stopping criterion $\Delta_k$ introduced in [3] is defined for all $k \geq 1$ by

$$\frac{1}{\gamma_k^2}(\mathrm{KL}((Q_k, R_k, g_k), (Q_{k-1}, R_{k-1}, g_{k-1})) + \mathrm{KL}((Q_{k-1}, R_{k-1}, g_{k-1}), (Q_k, R_k, g_k))),$$

where $((Q_k, R_k, g_k))_{k\geq 0}$ is the sequence solution of (6.7). First, we show that whatever the initialization chosen, the algorithm manages to converge to an efficient solution if no stopping criterion is used. However, the choice of the initialization may impact the termination of the algorithm as some initialization might be too close to some spurious local minima. Indeed, the initial points obtained using a "rank 2" or random initialization can be close to spurious and non-attractive local minima, which may trigger the stopping criterion too early and prevent the algorithm from continuing to run in order to converge towards an attractive and well behaved local minimum. We show also that the initialization we propose

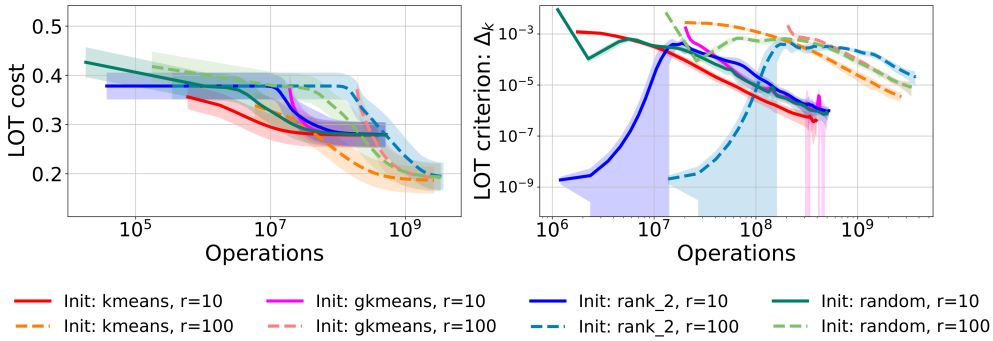



Figure 6.4: In this experiment, we consider the Newsgroup20 dataset [213] constituted of texts and we embed them into distributions in 50D using the same pre-processing steps as in [214]. We compare different initialization when applying the algorithm of [3] to compare random texts viewed as distributions for multiple choices of rank $r$. The ground cost considered in the squared Euclidean distance. We repeat the experiments 50 times by sampling randomly multiple problems of similar size ($\simeq 250$ samples). We normalize the cost matrix by its maximum value in order to have comparable LOT cost. We consider 4 different initialization: the one using $k$-means algorithm (6.9), the one using the generalized $k$-means (6.10), the rank-2 initialization [3] and a random initialization where $Q, R$ and $g$ are drawn from Gaussians. We compare both the cost value and the criterion value ($\Delta_k$) along the iterations of the MD scheme. Note that the curves obtained do not start at the same point in time as we start plotting the curves after obtaining the initial point which in some case requires more algebraic operations (e.g. kmeans methods). First we observe that whatever the initialization considered, the algorithm converges toward the same value. In addition, we observe that both $k$-means and general $k$-means are able to initialize well the algorithm by avoiding bad local minima at initialization while the two other initialization are close to spurious local minima at initialization.

in (6.9) and (6.10) are sufficiently far away from bad local minima and allow the algorithm to converge directly toward the desired solution. The right figure of Fig.4 shows two main observations: (i) that the initial point obtained using a "rank 2" or random initialization can be close to spurious and non-attractive local minima, which may trigger the stopping criterion too early and prevent the algorithm from continuing to run in order to converge towards an attractive and well behaved local minimum. (ii) When initialiazing the algorithm using kmeans methods, we show that our stopping criterion is a decreasing function of time meaning that the algorithm converges directly towards the desired solution.

**Conclusion.** We assembled in this work theoretical and practical arguments to support low-rank factorizations for OT. We have presented two controls: one concerning the approximation error to the true optimal transport and another concerning the statistical rates of the plug-in estimator. The latter is showed to be independent of the dimension, which is of particular interest when studying OT in ML settings. We have motivated further the use of LOT as a loss by introducing its debiased version and showed that it possesses desirable properties: positivity and metrization of the convergence in law. We have also presented the links between the bias induced by such regularization and clustering methods, and studied empirically the effects of hyperparameters involved in the practical estimation of LOT. The strong theoretical foundations provided in this paper motivate further studies of the empirical behaviour of LOT estimator, notably on finding suitable local minima and on improvements on the convergence of the MD scheme using other adaptive choices for step sizes.

# Supplementary materials

## 6.8 On the Definition of $\text{LOT}_{c,r}$

Let $(\mathcal{X}, d_{\mathcal{X}})$ and $(\mathcal{Y}, d_{\mathcal{Y}})$ two nonempty compact Polish spaces, $\mu \in \mathcal{M}_1^+(\mathcal{X})$, $\nu \in \mathcal{M}_1^+(\mathcal{Y})$ two probability measures on these spaces and $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ a nonnegative and continuous function. We define the generalized low-rank optimal transport between $\mu$ and $\nu$ as

$$\text{LOT}_{c,r}(\mu, \nu) := \inf_{\pi \in \Pi_r(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y) d\pi(x, y) \ .$$

where $\Pi_r(\mu, \nu)$ is defined as

$$\{\pi \in \Pi(\mu, \nu) : \exists (\mu_i)_{i=1}^r \in \mathcal{M}_1^+(\mathcal{X})^r, \ (\nu_i)_{i=1}^r \in \mathcal{M}_1^+(\mathcal{Y})^r, \ \lambda \in \Delta_r^* \text{ s.t. } \pi = \sum_{i=1}^r \lambda_i \mu_i \otimes \nu_i\} \ .$$

As $\mathcal{X}$ and $\mathcal{Y}$ are compact, $\Pi_r(\mu, \nu)$ is tight, then Prokhorov's theorem applies and the closure of $\Pi_r(\mu, \nu)$ is sequentially compact. Let us now show that $\Pi_r(\mu, \nu)$ is closed. Indeed, Let $(\pi_n)_{n\geq 0}$ a sequence of $\Pi_r(\mu, \nu)$ converging towards $\pi_*$. Then by definition there exists for all $k \in [|1, r|]$, $(\mu_n^{(k)})_{n\geq 0}$, $(\nu_n^{(k)})_{n\geq 0}$ and $(\lambda_n^{(k)})_{n\geq 0}$ such that for all $n \geq 0$

$$\pi_n = \sum_{i=1}^r \lambda_n^{(k)} \mu_n^{(k)} \otimes \nu_n^{(k)} \ .$$

However, $(\mu_n^{(k)})_{n\geq 0}$ and $(\nu_n^{(k)})_{n\geq 0}$ are also tight, and Prokhorov's theorem applies, therefore we can extract a common subsequence such that for all $k$,

$$\mu_n^{(k)} \to \mu_*^{(k)} \ \text{ and } \nu_n^{(k)} \to \nu_*^{(k)}$$

In addition as $(\lambda_n)_{n\geq 0}$ live in the simplex $\Delta_r$, we can also extract a sub-sequence, such that $\lambda_n \to \lambda_* \in \Delta_r$. Finally by unicity of the limit we obtain that

$$\pi_* = \sum_{k=1}^r \lambda_*^{(k)} \mu_*^{(k)} \otimes \nu_*^{(k)} \ .$$

Finally, by denoting $I := \{k : \lambda_*^{(k)} > 0\}$, and by considering $i^* \in I$, we obtain that

$$\pi_* = \sum_{i\in I\setminus\{i^*\}}^r \lambda_*^{(i)} \mu_*^{(i)} \otimes \nu_*^{(i)} + \sum_{j=1}^{r-|I|+1} \frac{\lambda_*^{(i^*)}}{r - |I| + 1} \mu_*^{(i^*)} \otimes \nu_*^{(i^*)} \ .$$

from which follows that $\pi_* \in \Pi_r(\mu, \nu)$.

## 6.9 Proofs

### 6.9.1 Proof of Proposition 6.3.1

**Proposition.** *Let $n, m \geq 2$, $X := \{x_1, \dots, x_n\} \subset \mathcal{X}$, $Y := \{y_1, \dots, y_m\} \subset \mathcal{Y}$ and $a \in \Delta_n^*$ and $b \in \Delta_m^*$. Then for $2 \leq r \leq \min(n, m)$, we have that*

$$|\mathrm{LOT}_{c,r}(\mu_{a,X}, \nu_{b,Y}) - \mathrm{OT}_c(\mu_{a,X}, \nu_{b,Y})| \leq \|C\|_\infty \; ln(\min(n, m)/(r-1))$$

*Proof.* Let $P \in \mathrm{argmin}_{P \in \Pi_{a,b}} \langle C, P \rangle$. As $P$ is a nonnegative matrix, its nonnegative rank cannot exceed $\min(n, m)$. Assume for simplicity, that $n = m$, then there exists $(R_i)_{i=1}^n$ nonnegative matrices of rank 1 such that

$$P = \sum_{i=1}^{n} R_i \; .$$

As for all $i \in [|1, n|]$, $R_i$ is a rank 1 matrix, there exist $\tilde{q}_i, \tilde{r}_i \in \mathbb{R}_+^n$ such that $R_i = \tilde{q}_i \tilde{r}_i^T$. Then by denoting $q_i = \tilde{q}_i / |\tilde{q}_i|$, $r_i = \tilde{r}_i / |\tilde{r}_i|$ and $\lambda_i = |\tilde{q}_i||\tilde{r}_i|$ where for any $h \in \mathbb{R}^n$ $|h| := \sum_{i=1}^n h_i$, we obtain that

$$P = \sum_{i=1}^{n} \lambda_i q_i r_i^T \; .$$

Without loss of generality, we can consider the case where $\lambda_1 \geq \cdots \geq \lambda_n$. Let us now denote $\lambda := (\lambda_1, \dots, \lambda_n)$, and by using the fact the $P$ is a coupling we obtain that $\lambda \in \Delta_n$. Also, by definition of $\lambda$, we have that for all $k \in [|1, n|]$, $\lambda_k \leq 1/k$. Let us now define

$$\tilde{P} := \sum_{i=1}^{r-1} \lambda_i q_i r_i^T + \left( \sum_{i=r}^{n} \lambda_i \right) \alpha_r \beta_r^T$$

where

$$\alpha_r := \frac{\sum_{i=r}^n \lambda_i q_i}{\sum_{i=r}^n \lambda_i}$$
$$\beta_r := \frac{\sum_{i=r}^n \lambda_i r_i}{\sum_{i=r}^n \lambda_i}$$

Remark that $\tilde{P} \in \Pi_{a,b}(r)$, therefore we obtain that

$$
\begin{aligned}
&|\mathrm{LOT}_{c,r}(\mu_{a,X}, \nu_{b,Y}) - \mathrm{OT}_c(\mu_{a,X}, \nu_{b,Y})| = \mathrm{LOT}_{c,r}(\mu_{a,X}, \nu_{b,Y}) - \mathrm{OT}_x(\mu_{a,X}, \nu_{b,Y}) \\
&\leq \langle C, \tilde{P} \rangle - \langle C, P \rangle \\
&\leq \langle C, \left( \sum_{i=r}^{n} \lambda_i \right) \alpha_r \beta_r^T \rangle - \langle C, \sum_{i=r}^{n} \lambda_i q_i r_i^T \rangle \\
&\leq \langle C, \left( \sum_{i=r}^{n} \lambda_i \right) \alpha_r \beta_r^T \rangle \\
&\leq \|C\|_\infty \sum_{i=r}^{n} \lambda_i \leq \|C\|_\infty \sum_{i=r}^{n} \frac{1}{i} \leq \|C\|_\infty \ln(n/(r-1))
\end{aligned}
$$

□

### 6.9.2 Proof of Proposition 6.3.2

**Proposition 6.9.1.** *Let $\mu \in \mathcal{M}_1^+(\mathcal{X})$, $\nu \in \mathcal{M}_1^+(\mathcal{Y})$ and let us assume that $c$ is $L$-Lipschitz w.r.t. $x$ and $y$ . Then for any $r \geq 1$, we have*

$$
|LOT_{c,r}(\mu, \nu) - OT_c(\mu, \nu)| \leq 2L \max(\mathcal{N}_{\lfloor \log_2(\lfloor \sqrt{r} \rfloor) \rfloor}(\mathcal{X}, d_{\mathcal{X}}), \mathcal{N}_{\lfloor \log_2(\lfloor \sqrt{r} \rfloor) \rfloor}(\mathcal{Y}, d_{\mathcal{Y}}))
$$

*Proof.* As $\mathcal{X}$ and $\mathcal{Y}$ are compact, $\mathcal{N}_{\lfloor \log_2(\lfloor \sqrt{r} \rfloor) \rfloor}(\mathcal{X}, d), \mathcal{N}_{\lfloor \log_2(\lfloor \sqrt{r} \rfloor) \rfloor}(\mathcal{Y}, d) < +\infty$ and then by denoting $\varepsilon_{\mathcal{X}} := \mathcal{N}_{\lfloor \log_2(\lfloor \sqrt{r} \rfloor) \rfloor}(\mathcal{X}, d_{\mathcal{X}})$, there exists $x_1, \ldots, x_{\lfloor \sqrt{r} \rfloor} \in \mathcal{X}$, such that $\mathcal{X} \subset \bigcup_{i=1}^{r} \mathcal{B}_{\mathcal{X}}(x_i, \varepsilon)$ from which we can extract a partition $(S_{i,\mathcal{X}})_{i=1}^{\lfloor \sqrt{r} \rfloor}$ of $\mathcal{X}$ such that for all $i \in [|1, \lfloor \sqrt{r} \rfloor|]$, and $x, y \in S_{i,\mathcal{X}}$, $d_{\mathcal{X}}(x, y) \leq \varepsilon_{\mathcal{X}}$. Similarly we can build a partition $(S_{i,\mathcal{Y}})_{i=1}^{\lfloor \sqrt{r} \rfloor}$ of $\mathcal{Y}$. Let us now define for all $k \in [|1, \lfloor \sqrt{r} \rfloor|]$,

$$
\mu_k := \frac{\mu|_{S_{k,\mathcal{X}}}}{\mu(S_{k,\mathcal{X}})} \quad \text{and} \quad \nu_k := \frac{\nu|_{S_{k,\mathcal{Y}}}}{\nu(S_{k,\mathcal{Y}})}
$$

with the convention that $\frac{0}{0} = 0$, we can define

$$
\pi_r := \sum_{i,j=1}^{\lfloor \sqrt{r} \rfloor} \pi^*(S_{i,\mathcal{X}} \times S_{j,\mathcal{Y}}) \nu_j \otimes \mu_i \, .
$$

First remarks that $\pi_r \in \Pi_r(\mu,\nu)$. Indeed we have for any measurable set $B$

$$\begin{aligned}\pi_r(\mathcal{X}\times B) &= \sum_{j=1}^{\lfloor\sqrt{r}\rfloor^2}\nu_j(B)\sum_{i=1}^{r}\pi^*(S_{i,\mathcal{X}}\times S_{j,\mathcal{Y}})\\ &= \sum_{j=1}^{\lfloor\sqrt{r}\rfloor}\nu_j(B)\nu(S_{j,\mathcal{Y}})\\ &= \sum_{j=1}^{\lfloor\sqrt{r}\rfloor}\nu|_{S_{j,\mathcal{X}}}(B)\\ &= \nu(B)\ ,\end{aligned}$$

similarly $\pi_r(A\times\mathcal{Y}) = \mu(A)$ and we have that $\lfloor\sqrt{r}\rfloor^2 \leq r$. Therefore we obtain that

$$\begin{aligned}|\mathrm{LOT}_{c,r}(\mu,\nu) - \mathrm{OT}_c(\mu,\nu)| &= \mathrm{LOT}_{c,r}(\mu,\nu) - \mathrm{OT}_c(\mu,\nu)\\ &\leq \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)d\pi_r(x,y) - \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)d\pi^*(x,y)\\ &\leq \sum_{i,j=1}^{\lfloor\sqrt{r}\rfloor}\int_{S_{i,\mathcal{X}}\times S_{j,\mathcal{Y}}} c(x,y)d[\pi_r(x,y)-\pi^*(x,y)]\\ &\leq \sum_{i,j=1}^{\lfloor\sqrt{r}\rfloor}\pi^*(S_{i,\mathcal{X}}\times S_{j,\mathcal{Y}})\\ &\times\Big[\sup_{(x,y)\in S_{i,\mathcal{X}}\times S_{j,\mathcal{Y}}} c(x,y) - \inf_{(x,y)\in S_{i,\mathcal{X}}\times S_{j,\mathcal{Y}}} c(x,y)\Big]\\ &\leq L[\varepsilon_{\mathcal{X}}+\varepsilon_{\mathcal{Y}}]\end{aligned}$$

from which the result follows. □

**Corollary.** *Under the same assumptions of Proposition 6.3.2 and by assuming in addition that there exists a Monge map solving $OT_c(\mu,\nu)$, we obtain that for any $r \geq 1$,*

$$|\mathrm{LOT}_{c,r}(\mu,\nu) - \mathrm{OT}_c(\mu,\nu)| \leq L\mathcal{N}_{\lfloor\log_2(r)\rfloor}(\mathcal{Y},d_{\mathcal{Y}})$$

*Proof.* Let us denote $T$ a Monge map solution of $\mathrm{OT}_c(\mu,\nu)$ and as in the proof above, let us consider a partition of $(S_{i,\mathcal{Y}})_{i=1}^r$ of $\mathcal{Y}$ such that for all $i \in [|1,r|]$, and $x,y \in S_{i,\mathcal{Y}}$, $d_{\mathcal{Y}}(x,y) \leq \varepsilon_{\mathcal{Y}}$ with $\varepsilon_{\mathcal{Y}} := \mathcal{N}_{\lfloor\log_2(r)\rfloor}(\mathcal{Y},d_{\mathcal{Y}})$. Let us now define for all $k \in [|1,\lfloor\sqrt{r}\rfloor|]$,

$$\mu_k := \frac{\mu|_{T^{-1}(S_{k,\mathcal{Y}})}}{\mu(T^{-1}(S_{k,\mathcal{Y}}))} \quad \text{and} \quad \nu_k := \frac{\nu|_{S_{k,\mathcal{Y}}}}{\nu(S_{k,\mathcal{Y}})}$$

with the convention that $\frac{0}{0} = 0$, we can define

$$\pi_r := \sum_{k=1}^{r} \pi^*(T^{-1}(S_{k,\mathcal{Y}}) \times S_{k,\mathcal{Y}})\nu_k \otimes \mu_k \ .$$

Again we have that $\pi_r \in \Pi_r(\mu,\nu)$, and we obtain that

$$\begin{aligned}
&|\mathrm{LOT}_{c,r}(\mu,\nu) - \mathrm{OT}_c(\mu,\nu)| = \mathrm{LOT}_{c,r}(\mu,\nu) - \mathrm{OT}_c(\mu,\nu) \\
&\leq \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)d\pi_r(x,y) - \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)d\pi^*(x,y) \\
&\leq \sum_{k=1}^{r} \pi^*(T^{-1}(S_{k,\mathcal{Y}}) \times S_{k,\mathcal{Y}}) \int_{T^{-1}(S_{k,\mathcal{Y}})\times S_{k,\mathcal{Y}}} c(x,y)d\mu_k(y) \otimes \nu_k(y) \\
&- \sum_{k=1}^{r} \int_{T^{-1}(S_{k,\mathcal{Y}})} c(x,T(x))d\mu(x) \\
&\leq \sum_{k=1}^{r} \pi^*(T^{-1}(S_{k,\mathcal{Y}}) \times S_{k,\mathcal{Y}}) \int_{T^{-1}(S_{k,\mathcal{Y}})\times S_{k,\mathcal{Y}}} c(x,y)d\mu_k(y) \otimes \nu_k(y) \\
&- \sum_{k=1}^{r} \pi^*(T^{-1}(S_{k,\mathcal{Y}}) \times S_{k,\mathcal{Y}}) \int_{T^{-1}(S_{k,\mathcal{Y}})\times S_{k,\mathcal{Y}}} c(x,T(x))d\mu_k(x) \otimes \nu_k(y) \\
&\leq \sum_{k=1}^{r} \pi^*(T^{-1}(S_{k,\mathcal{Y}}) \times S_{k,\mathcal{Y}}) \int_{T^{-1}(S_{k,\mathcal{Y}})\times S_{k,\mathcal{Y}}} [c(x,y) - c(x,T(x))]d\mu_k \otimes \nu_k \\
&\leq L\varepsilon_{\mathcal{Y}}
\end{aligned}$$

from which the result follows. Note that to obtain the above inequalities, we use the fact that $\pi^*$ is supported on the graph of $T$, and therefore we have have for all $k \in [|1,r|]$,

$$\pi^*(T^{-1}(S_{k,\mathcal{Y}}) \times S_{k,\mathcal{Y}}) = \mu(T^{-1}(S_{k,\mathcal{Y}})) = \nu(S_{k,\mathcal{Y}}).$$

□

### 6.9.3 Proof of Proposition 6.4.1

**Proposition.** *Let $r \geq 1$ and $\mu,\nu \in \mathcal{M}_1^+(\mathcal{X})$, then*

$$\mathrm{LOT}_{c,r}(\hat{\mu}_n,\hat{\nu}_n) \xrightarrow[n\to+\infty]{} \mathrm{LOT}_{c,r}(\mu,\nu) \quad a.s.$$

*Proof.* Let $\pi^*$ solution of $\mathrm{LOT}_{c,r}(\mu,\nu)$. Then there exists $\lambda^* \in \Delta_r^*$, $(\mu_i^*)_{i=1}^r, (\nu_i^*)_{i=1}^r \in \mathcal{M}_1^+(\mathcal{X})^r$ such that

$$\pi^* = \sum_{i=1}^{r} \lambda_i^* \mu_i^* \otimes \nu_i^*.$$

Note that by definition, we have that

$$\mu = \sum_{i=1}^{r} \lambda_i^* \mu_i^* \quad \text{and} \quad \nu = \sum_{i=1}^{r} \lambda_i^* \nu_i^* \ .$$

Let us now define $\pi_\mu$ and $\pi_\mu$ both elements of $\mathcal{M}_1^+(\mathcal{X} \times [|1, r|])$ as follows:

$$\pi_\mu(A \times \{k\}) := \lambda_k \mu_k(A) \quad \text{and} \quad \pi_\nu(A \times \{k\}) := \lambda_k \nu_k(A)$$

for any measurable set $A$ and $k \in [|1, r|]$. Observe that the right marginals of $\pi_\mu$ and $\pi_\nu$ is the same and we will denote it $\rho$. We can now define for all $x, y \in \mathcal{X}$ the family of kernels $(k_\mu(\cdot, x))_{x \in \mathcal{X}} \in \mathcal{M}_1^+([|1, r|])^{\mathcal{X}}$ and $(k_\nu(\cdot, y))_{y \in \mathcal{X}} \in \mathcal{M}_1^+([|1, r|])^{\mathcal{X}}$ corresponding to the disintegration with respect to the projection of respectively $\mu$ and $\nu$. Let us now consider $n$ independent samples $(Z_i^\mu)_{i=1}^n$ and $(Z_i^\nu)_{i=1}^n$ such that for all $i \in [|1, n|]$, $Z_i^\mu \sim k_\mu(\cdot, X_i)$ and $Z_i^\nu \sim k_\nu(\cdot, Y_i)$ and let us define for all $k \in [|1, r|]$

$$\tilde{\mu}_k := \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{Z_i^\mu = k} \delta_{X_i} \quad \text{and} \quad \tilde{\nu}_k := \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{Z_i^\nu = k} \delta_{Y_i} \ .$$

Let us now define

$$\begin{aligned}
\tilde{\pi} := & \sum_{k=1}^{r-1} \frac{\min(|\tilde{\mu}_k|, |\tilde{\nu}_k|)}{|\tilde{\mu}_k||\tilde{\nu}_k|} \tilde{\mu}_k \otimes \tilde{\nu}_k \\
& + \frac{1}{1 - \sum_{k=1}^{r-1} \min(|\tilde{\mu}_k|, |\tilde{\nu}_k|)} \left[ \hat{\mu} - \sum_{k=1}^{r-1} \frac{\min(|\tilde{\mu}_k|, |\tilde{\nu}_k|)}{|\tilde{\mu}_k|} \tilde{\mu}_k \right] \otimes \left[ \hat{\nu} - \sum_{k=1}^{r-1} \frac{\min(|\tilde{\mu}_k|, |\tilde{\nu}_k|)}{|\tilde{\nu}_k|} \tilde{\nu}_k \right]
\end{aligned}$$

with the convention that $\frac{0}{0} = 0$. Now it is easy to check that $\tilde{\pi} \in \Pi_r(\hat{\mu}, \hat{\nu})$, indeed we have that

$$\begin{aligned}
\tilde{\pi}(A \times \mathcal{X}) = & \sum_{k=1}^{r-1} \frac{\min(|\tilde{\mu}_k|, |\tilde{\nu}_k|)}{|\tilde{\mu}_k|} \tilde{\mu}_k(A) \\
& + \frac{1}{1 - \sum_{k=1}^{r-1} \min(|\tilde{\mu}_k|, |\tilde{\nu}_k|)} \left[ \hat{\mu}(A) - \sum_{k=1}^{r-1} \frac{\min(|\tilde{\mu}_k|, |\tilde{\nu}_k|)}{|\tilde{\mu}_k|} \tilde{\mu}_k(A) \right] \\
& \times \left[ 1 - \sum_{k=1}^{r-1} \min(|\tilde{\mu}_k|, |\tilde{\nu}_k|) \right] \\
= & \ \hat{\mu}(A)
\end{aligned}$$

in addition by construction we have that

$$\left|\hat{\mu} - \sum_{k=1}^{r-1} \frac{\min(|\tilde{\mu}_k|, |\tilde{\nu}_k|)}{|\tilde{\mu}_k|}\tilde{\mu}_k\right| = \left|\hat{\nu} - \sum_{k=1}^{r-1} \frac{\min(|\tilde{\mu}_k|, |\tilde{\nu}_k|)}{|\tilde{\nu}_k|}\tilde{\nu}_k\right| = 1 - \sum_{k=1}^{r-1} \min(|\tilde{\mu}_k|, |\tilde{\nu}_k|)$$

and both $\hat{\mu} - \sum_{k=1}^{r-1} \frac{\min(|\tilde{\mu}_k|,|\tilde{\nu}_k|)}{|\tilde{\mu}_k|}\tilde{\mu}_k$ and $\hat{\nu} - \sum_{k=1}^{r-1} \frac{\min(|\tilde{\mu}_k|,|\tilde{\nu}_k|)}{|\tilde{\nu}_k|}\tilde{\nu}_k$ are positive measures. Therefore we obtain that

$$\mathrm{LOT}_{c,r}(\hat{\mu}, \hat{\nu}) \leq \int_{\mathcal{X}^2} c(x, y) d\tilde{\pi}(x, y)$$

Now we aim at showing at $\int_{\mathcal{X}^2} c(x, y) d\tilde{\pi}(x, y) \to \mathrm{LOT}_{c,r}(\mu, \nu) \quad a.s.$. Indeed first observe that from the law of large numbers we have that for all $k \in [|1, r|]$, $|\tilde{\mu}_k| \to \lambda_k^*$ and similarly $|\tilde{\nu}_k| \to \lambda_k^*$. In addition, for all $k, q$ we have that almost surely, $\tilde{\mu}_k \otimes \tilde{\nu}_q$ converges weakly towards $\lambda_k^* \lambda_q^* \mu_k \otimes \nu_q$. Indeed one can consider the following algebra $\mathcal{F} := \{(x, y) \in \mathcal{X}^2 \to f(x)g(y) \quad f, g \in \mathcal{C}(\mathcal{X})\}$, and then by Stone-Weierstrass, one obtains by density the desired result. Now remark that

$$\begin{aligned}
\int_{\mathcal{X}^2} c(x, y) d\tilde{\pi}(x, y) = &\sum_{k=1}^{r-1} \frac{\min(|\tilde{\mu}_k|, |\tilde{\nu}_k|)}{|\tilde{\mu}_k||\tilde{\nu}_k|} \int_{\mathcal{X}^2} c(x, y) d\tilde{\mu}_k \otimes \tilde{\nu}_k \\
&+ \frac{1}{\tilde{\lambda}_r} \int_{\mathcal{Z}^2} c(x, y) d\tilde{\mu}_r \otimes \tilde{\nu}_r \\
&+ \frac{1}{\tilde{\lambda}_r} \sum_{k=1}^{r-1} \left(1 - \frac{\min(|\tilde{\mu}_k|, |\tilde{\nu}_k|)}{|\tilde{\nu}_k|}\right) \int_{\mathcal{X}^2} c(x, y) d\tilde{\mu}_r \otimes \tilde{\nu}_k \\
&+ \frac{1}{\tilde{\lambda}_r} \sum_{k=1}^{r-1} \left(1 - \frac{\min(|\tilde{\mu}_k|, |\tilde{\nu}_k|)}{|\tilde{\mu}_k|}\right) \int_{\mathcal{X}^2} c(x, y) d\tilde{\mu}_k \otimes \tilde{\nu}_r \\
&+ \frac{1}{\tilde{\lambda}_r} \sum_{k,q=1}^{r-1} \int_{\mathcal{X}^2} \left(1 - \frac{\min(|\tilde{\mu}_k|, |\tilde{\nu}_k|)}{|\tilde{\mu}_k|}\right) \\
&\times \left(1 - \frac{\min(|\tilde{\mu}_q|, |\tilde{\nu}_q|)}{|\tilde{\nu}_q|}\right) c(x, y) d\tilde{\mu}_k(x) d\tilde{\nu}_q(y)
\end{aligned}$$

from which follows directly that $\int_{\mathcal{X}^2} c(x, y) d\tilde{\pi}(x, y) \to \mathrm{LOT}_{c,r}(\mu, \nu) \quad a.s.$ Let us now denote for all $n \geq 1$, $\pi_n$ a solution of $\mathrm{LOT}_{c,r}(\hat{\mu}, \hat{\nu})$. Let $\omega \in \Omega$ an element of the probability space where live the random variables $(X_i)_{i\geq 0}$ and $(Y_i)_{i\geq 0}$ such that $\int_{\mathcal{X}^2} c(x, y) d\tilde{\pi}^{(\omega)}(x, y) \to \mathrm{LOT}_{c,r}(\mu, \nu)$. As $\mathcal{X}$ is compact Thanks to Prokhorov's Theorem, we can extract a sequence such that $(\pi_n^{(\omega)})_{n\geq 0}$ converge weakly towards $\pi^{(\omega)} \in \Pi_r(\mu, \nu)$. In addition we have that for all $n \geq 1$

$$\int_{\mathcal{X}^2} c(x, y) d\pi_n^{(\omega)}(x, y) \leq \int_{\mathcal{X}^2} c(x, y) d\tilde{\pi}^{(\omega)}(x, y)$$

And by considering the limit we obtain that

$$\int c(x,y)d\pi^{(\omega)}(x,y) \leq \mathrm{LOT}_{c,r}(\mu,\nu)$$

However $\pi^{(\omega)} \in \Pi_r(\mu,\nu)$ and by optimality we obtain that

$$\int c(x,y)d\pi^{(\omega)}(x,y) = \mathrm{LOT}_{c,r}(\mu,\nu)$$

This holds for an arbitrary subsequence of $(\pi_n^{(\omega)})_{n\geq 0}$, from which follows that $\int c(x,y)d\pi_n^{(\omega)}(x,y) \to \mathrm{LOT}_{c,r}(\mu,\nu)$. Finally this holds almost surely and the result follows. □

### 6.9.4 Proof of Proposition 6.4.2

**Proposition.** *Let $r \geq 1$ and $\mu,\nu \in \mathcal{M}_1^+(\mathcal{X})$. Then, there exists a constant $K_r$ such that for any $\delta > 0$ and $n \geq 1$, we have, with a probability of at least $1-2\delta$, that*

$$\mathrm{LOT}_{c,r}(\hat{\mu}_n,\hat{\nu}_n) - \mathrm{LOT}_{c,r}(\mu,\nu) \leq 11\|c\|_\infty\sqrt{\frac{r}{n}} + K_r\|c\|_\infty\left[\sqrt{\frac{\log(40/\delta)}{n}} + \frac{\sqrt{r}\log(40/\delta)}{n}\right]$$

*Proof.* We reintroduce the same notation as in the proof of Proposition 6.4.1. Let $\pi^*$ solution of $\mathrm{LOT}_{c,r}(\mu,\nu)$. Then there exists $\lambda^* \in \Delta_r^*$, $(\mu_i^*)_{i=1}^r, (\nu_i^*)_{i=1}^r \in \mathcal{M}_1^+(\mathcal{Z})^r$ such that

$$\pi^* = \sum_{i=1}^r \lambda_i^*\mu_i^* \otimes \nu_i^*.$$

As before let us also consider $\pi_\mu$ and $\pi_\mu$ defined as $\pi_\mu(A\times\{k\}) := \lambda_k\mu_k(A)$ and $\pi_\nu(A\times\{k\}) := \lambda_k\nu_k(A)$ for any measurable set $A$ and $k \in [|1,r|]$ and denote $\rho$ their common right marginal. We also consider $n$ independent samples $(Z_i^\mu)_{i=1}^n$ and $(Z_i^\nu)_{i=1}^n$ such that for all $i \in [|1,n|]$, $Z_i^\mu \sim k_\mu(\cdot,X_i)$ and $Z_i^\nu \sim k_\nu(\cdot,Y_i)$ and we denote for all $k \in [|1,r|]$

$$\tilde{\mu}_k := \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{Z_i^\mu=k}\delta_{X_i} \quad \text{and} \quad \tilde{\nu}_k := \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{Z_i^\nu=k}\delta_{Y_i}$$

Let us now define

$$\hat{\pi} := \sum_{i=1}^r \frac{1}{\lambda_k^*}\tilde{\mu}_k \otimes \tilde{\nu}_k\ .$$

Our goal is to control the following quantity:

$$\left|\mathrm{LOT}_{c,r}(\mu,\nu) - \int_{\mathcal{Z}^2} c(x,y)d\hat{\pi}(x,y)\right|,$$

First observe that

$$\begin{aligned}\mathbb{E}\left[\int_{\mathcal{Z}^2} c(x,y)d\hat{\pi}(x,y)\right] &= \sum_{i=1}^{r}\frac{1}{\lambda_k^*}\mathbb{E}\left[\int_{\mathcal{Z}^2} c(x,y)d\tilde{\mu}_k(x)d\tilde{\nu}_k(y)\right]\\ &= \sum_{i=1}^{r}\frac{1}{\lambda_k^* n^2}\times\sum_{i,j}\mathbb{E}\left[c(X_i,Y_j)\mathbf{1}_{Z_i^\mu=k}\mathbf{1}_{Z_j^\nu=k}\right]\end{aligned}$$

Moreover, we have that

$$\begin{aligned}\mathbb{E}\left[c(X_i,Y_j)\mathbf{1}_{Z_i^\mu=k}\mathbf{1}_{Z_j^\nu=k}\right] &= \int_{(\mathcal{Z}\times[|1,r|])^2} c(x,y)\mathbf{1}_{z=k}\mathbf{1}_{z'=k}d\pi_\mu(x,z)d\pi_\nu(y,z')\\ &= \int_{(\mathcal{Z}\times[|1,r|])^2} c(x,y)\mathbf{1}_{z=k}\mathbf{1}_{z'=k}d\mu_z(x)d\nu_{z'}(y)d\rho(z)d\rho(z')\\ &= \lambda_k^2\int_{\mathcal{Z}^2} c(x,y)d\mu_k(x)d\nu_k(y)\end{aligned}$$

from which follows that

$$\mathbb{E}\left[\int_{\mathcal{Z}^2} c(x,y)d\hat{\pi}(x,y)\right] = \sum_{i=1}^{r}\lambda_k^*\int_{\mathcal{Z}^2} c(x,y)d\mu_k(x)d\nu_k(y) = \mathrm{LOT}_{c,r}(\mu,\nu)$$

Now let us define for all $(x_i,z_i)_{i=1}^n,(y_i,z_i')\in(\mathcal{Z}\times[|1,r|])^n$,

$$g((x_1,z_1),\dots,(x_n,z_n),(y_1,z_1'),\dots,(y_n,z_n')) := \sum_{q=1}^{r}\frac{1}{\lambda_q^* n^2}\sum_{i,j} c(x_i,y_j)\mathbf{1}_{z_i=q}\mathbf{1}_{z_j'=q}\ ,$$

since $\mathcal{Z}$ is compact and $c$ is continuous, we have that

$$\begin{aligned}&|g(\dots,(x_k,z_k),\dots) - g(\dots,(\tilde{x}_k,\tilde{z}_k),\dots)| =\\ &\left|\sum_{q=1}^{r}\frac{1}{\lambda_q^* n^2}\sum_j[c(x_k,y_j)\mathbf{1}_{z_k=q} - c(\tilde{x}_k,y_j)\mathbf{1}_{\tilde{z}_k=q}]\mathbf{1}_{z_j'=q}\right|\\ &= \left|\frac{1}{\lambda_{z_k}^* n^2}\sum_{j=1}^{n} c(x_k,y_j)\mathbf{1}_{z_j'=z_k} - \frac{1}{\lambda_{\tilde{z}_k}^* n^2}\sum_{j=1}^{n} c(\tilde{x}_k,y_j)\mathbf{1}_{z_j'=\tilde{z}_k}\right|\\ &\leq \frac{\|c\|_\infty}{n^2}\left[\frac{\sum_{j=1}^n \mathbf{1}_{z_j'=z_k}}{\lambda_{z_k}^*} + \frac{\sum_{j=1}^n \mathbf{1}_{z_j'=\tilde{z}_k}}{\lambda_{\tilde{z}_k}^*}\right]\\ &\leq \frac{2\|c\|_\infty}{\min_{1\leq q\leq r}\lambda_q^*}\frac{1}{n}\end{aligned}$$

Then by applying the McDiarmid's inequality we obtain that for $\delta > 0$, with a probability at least of $1 - \delta$, we have

$$\left| \mathrm{LOT}_{c,r}(\mu, \nu) - \int_{\mathcal{Z}^2} c(x, y) d\hat{\pi}(x, y) \right| \leq \frac{2\|c\|_\infty}{\min\limits_{1 \leq q \leq r} \lambda_q^*} \sqrt{\frac{\log(2/\delta)}{n}}$$

Now we aim at building a coupling $\tilde{\pi} \in \Pi_r(\hat{\mu}, \hat{\nu})$ from $\hat{\pi}$. Let us consider the same as the one introduce in the proof of Proposition 6.9.3, that is

$$\begin{aligned}
\tilde{\pi} &:= \sum_{k=1}^{r-1} \frac{\min(|\tilde{\mu}_k|, |\tilde{\nu}_k|)}{|\tilde{\mu}_k||\tilde{\nu}_k|} \tilde{\mu}_k \otimes \tilde{\nu}_k \\
&+ \frac{1}{1 - \sum_{k=1}^{r-1} \min(|\tilde{\mu}_k|, |\tilde{\nu}_k|)} \left[ \hat{\mu} - \sum_{k=1}^{r-1} \frac{\min(|\tilde{\mu}_k|, |\tilde{\nu}_k|)}{|\tilde{\mu}_k|} \tilde{\mu}_k \right] \otimes \left[ \hat{\nu} - \sum_{k=1}^{r-1} \frac{\min(|\tilde{\mu}_k|, |\tilde{\nu}_k|)}{|\tilde{\nu}_k|} \tilde{\nu}_k \right]
\end{aligned}$$

with the convention that $\frac{0}{0} = 0$. Let us now expand the above expression, and by denoting $\tilde{\lambda}_r = 1 - \sum_{k=1}^{r-1} \min(|\tilde{\mu}_k|, |\tilde{\nu}_k|)$ we obtain that

$$\begin{aligned}
\tilde{\pi} &= \sum_{k=1}^{r-1} \frac{\min(|\tilde{\mu}_k|, |\tilde{\nu}_k|)}{|\tilde{\mu}_k||\tilde{\nu}_k|} \tilde{\mu}_k \otimes \tilde{\nu}_k \\
&+ \frac{1}{\tilde{\lambda}_r} \tilde{\mu}_r \otimes \tilde{\nu}_r \\
&+ \frac{1}{\tilde{\lambda}_r} \tilde{\mu}_r \otimes \left[ \sum_{k=1}^{r-1} \left( 1 - \frac{\min(|\tilde{\mu}_k|, |\tilde{\nu}_k|)}{|\tilde{\nu}_k|} \right) \tilde{\nu}_k \right] \\
&+ \frac{1}{\tilde{\lambda}_r} \left[ \sum_{k=1}^{r-1} \left( 1 - \frac{\min(|\tilde{\mu}_k|, |\tilde{\nu}_k|)}{|\tilde{\mu}_k|} \right) \tilde{\mu}_k \right] \otimes \tilde{\nu}_r \\
&+ \frac{1}{\tilde{\lambda}_r} \left[ \sum_{k=1}^{r-1} \left( 1 - \frac{\min(|\tilde{\mu}_k|, |\tilde{\nu}_k|)}{|\tilde{\mu}_k|} \right) \tilde{\mu}_k \right] \otimes \left[ \sum_{k=1}^{r-1} \left( 1 - \frac{\min(|\tilde{\mu}_k|, |\tilde{\nu}_k|)}{|\tilde{\nu}_k|} \right) \tilde{\nu}_k \right]
\end{aligned}$$

Now we aim at controlling the following quantity $\left| \int_{\mathcal{Z}^2} c(x, y) d\hat{\pi}(x, y) - \int_{\mathcal{Z}^2} c(x, y) d\tilde{\pi}(x, y) \right|$

and we observe that

$$\int_{\mathcal{Z}^2} c(x,y) d[\hat{\pi}(x,y) - \tilde{\pi}(x,y)] \tag{6.11}$$

$$= \sum_{k=1}^{r-1} \int_{\mathcal{Z}^2} c(x,y) \left[ \frac{1}{\lambda_k^*} - \frac{\min(|\tilde{\mu}_k|, |\tilde{\nu}_k|)}{|\tilde{\mu}_k||\tilde{\nu}_k|} \right] d\tilde{\mu}_k(x)\tilde{\nu}_k(y) \tag{6.12}$$

$$+ \int_{\mathcal{Z}^2} c(x,y) \left[ \frac{1}{\lambda_r^*} - \frac{1}{\tilde{\lambda}_r} \right] d\tilde{\mu}_r(x)\tilde{\nu}_r(y) \tag{6.13}$$

$$+ \frac{1}{\tilde{\lambda}_r} \sum_{k=1}^{r-1} \int_{\mathcal{Z}^2} \left( 1 - \frac{\min(|\tilde{\mu}_k|, |\tilde{\nu}_k|)}{|\tilde{\nu}_k|} \right) c(x,y) d\tilde{\mu}_r(x) d\tilde{\nu}_k(y) \tag{6.14}$$

$$+ \frac{1}{\tilde{\lambda}_r} \sum_{k=1}^{r-1} \int_{\mathcal{Z}^2} \left( 1 - \frac{\min(|\tilde{\mu}_k|, |\tilde{\nu}_k|)}{|\tilde{\mu}_k|} \right) c(x,y) d\tilde{\mu}_k(x) d\tilde{\nu}_r(y) \tag{6.15}$$

$$+ \frac{1}{\tilde{\lambda}_r} \sum_{k,q=1}^{r-1} \int_{\mathcal{Z}^2} \left( 1 - \frac{\min(|\tilde{\mu}_k|, |\tilde{\nu}_k|)}{|\tilde{\mu}_k|} \right) \left( 1 - \frac{\min(|\tilde{\mu}_q|, |\tilde{\nu}_q|)}{|\tilde{\nu}_q|} \right) c(x,y) d\tilde{\mu}_k d\tilde{\nu}_q \tag{6.16}$$

Let us now control each term of the RHS of the above equality. Let us first consider the term in Eq. 6.12, remark that we have

$$\begin{aligned}
&\left| \int_{\mathcal{Z}^2} c(x,y) \left[ \frac{1}{\lambda_k^*} - \frac{\min(|\tilde{\mu}_k|, |\tilde{\nu}_k|)}{|\tilde{\mu}_k||\tilde{\nu}_k|} \right] d\tilde{\mu}_k(x)\tilde{\nu}_k(y) \right| \\
&\leq \left| \left[ \frac{1}{\lambda_k^*} - \frac{\min(|\tilde{\mu}_k|, |\tilde{\nu}_k|)}{|\tilde{\mu}_k||\tilde{\nu}_k|} \right] \right| \|c\|_\infty |\tilde{\mu}_k||\tilde{\nu}_k| \\
&\leq \left| \left[ \frac{|\tilde{\mu}_k||\tilde{\nu}_k|}{\lambda_k^*} - \min(|\tilde{\mu}_k|, |\tilde{\nu}_k|) \right] \right| \|c\|_\infty \\
&\leq \min(|\tilde{\mu}_k|, |\tilde{\nu}_k|) \left| \frac{\max(|\tilde{\mu}_k|, |\tilde{\nu}_k|)}{\lambda_k^*} - 1 \right| \|c\|_\infty \\
&\leq \frac{\min(|\tilde{\mu}_k|, |\tilde{\nu}_k|)}{\lambda_k^*} |\max(|\tilde{\mu}_k|, |\tilde{\nu}_k|) - \lambda_k^*| \|c\|_\infty \\
&\leq \frac{\min(|\tilde{\mu}_k|, |\tilde{\nu}_k|)}{\lambda_k^*} \max(\|\tilde{\lambda}_\mu - \lambda^*\|_\infty, \|\tilde{\lambda}_\nu - \lambda^*\|_\infty) \|c\|_\infty \\
&\leq \|c\|_\infty \max\left( \left\| \frac{\tilde{\lambda}_\mu}{\lambda^*} \right\|_\infty, \left\| \frac{\tilde{\lambda}_\nu}{\lambda^*} \right\|_\infty \right) \max(\|\tilde{\lambda}_\mu - \lambda^*\|_\infty, \|\tilde{\lambda}_\nu - \lambda^*\|_\infty)
\end{aligned}$$

where we have denoted $\tilde{\lambda}_\mu := (|\tilde{\mu}_k|)_{k=1}^r$ and $\tilde{\lambda}_\nu := (|\tilde{\nu}_k|)_{k=1}^r$. Now observe that

$$\begin{aligned}\mathbb{P}\left(\max(\|\tilde{\lambda}_\mu - \lambda^*\|_\infty, \|\tilde{\lambda}_\nu - \lambda^*\|_\infty) \geq t\right) &\leq 2\mathbb{P}\left(\|\tilde{\lambda}_\mu - \lambda^*\|_\infty \geq t\right) \\ &\leq \mathbb{P}\left(d_K(\lambda^*, \tilde{\lambda}_\mu) \geq \frac{t}{2}\right) \\ &\leq 4\exp(-nt^2/2)\end{aligned}$$

where $d_K$ is the Kolmogorov distance. In addition we have

$$\max\left(\left\|\frac{\tilde{\lambda}_\mu}{\lambda^*}\right\|_\infty, \left\|\frac{\tilde{\lambda}_\nu}{\lambda^*}\right\|_\infty\right) \leq 1 + \frac{1}{\min\limits_{1\leq i\leq r}\lambda_i^*}\max\left(\|\tilde{\lambda}_\mu - \lambda^*\|_\infty, \|\tilde{\lambda}_\nu - \lambda^*\|_\infty\right)$$

Combining the two above controls, we obtain that for all $\delta > 0$, with a probability of at least $1-\delta$,

$$\left|\int_{\mathcal{Z}^2} c(x,y)\left[\frac{1}{\lambda_k^*} - \frac{\min(|\tilde{\mu}_k|, |\tilde{\nu}_k|)}{|\tilde{\mu}_k||\tilde{\nu}_k|}\right] d\tilde{\mu}_k(x)\tilde{\nu}_k(y)\right| \leq \|c\|_\infty\sqrt{\frac{2\ln 8/\delta}{n}} + \frac{\|c\|_\infty}{n}\frac{2\ln 8/\delta}{\min\limits_{1\leq i\leq r}\lambda_i^*}$$

Let us now consider the term in Eq. 6.13, we have that

$$\begin{aligned}&\left|\int_{\mathcal{Z}^2} c(x,y)\left[\frac{1}{\lambda_r^*} - \frac{1}{\tilde{\lambda}_r}\right] d\tilde{\mu}_r(x)\tilde{\nu}_r(y)\right| \leq \frac{|\tilde{\mu}_r||\tilde{\nu}_r|}{\lambda_r^*\tilde{\lambda}_r}\left|1 - \sum_{i=1}^r \min(|\tilde{\mu}_k|, |\tilde{\nu}_k|) - \lambda_r\right| \|c\|_\infty \\ &\leq \max\left(\left\|\frac{\tilde{\lambda}_\mu}{\lambda^*}\right\|_\infty, \left\|\frac{\tilde{\lambda}_\nu}{\lambda^*}\right\|_\infty\right)\sum_{k=1}^{r-1}|\lambda_k^* - \min(|\tilde{\mu}_k|, |\tilde{\nu}_k|)|\,\|c\|_\infty \\ &\leq \max\left(\left\|\frac{\tilde{\lambda}_\mu}{\lambda^*}\right\|_\infty, \left\|\frac{\tilde{\lambda}_\nu}{\lambda^*}\right\|_\infty\right)\|c\|_\infty(\|\lambda^* - \tilde{\lambda}_\mu\|_1 + \|\lambda^* - \tilde{\lambda}_\nu\|_1) \\ &\leq 2\|c\|_\infty\max\left(\left\|\frac{\tilde{\lambda}_\mu}{\lambda^*}\right\|_\infty, \left\|\frac{\tilde{\lambda}_\nu}{\lambda^*}\right\|_\infty\right)\max(\|\lambda^* - \tilde{\lambda}_\mu\|_1, \|\lambda^* - \tilde{\lambda}_\nu\|_1)\end{aligned}$$

However we have that

$$\mathbb{P}\left(\max(\|\lambda^* - \tilde{\lambda}_\mu\|_1, \|\lambda^* - \tilde{\lambda}_\nu\|_1) \geq t\right) \leq 2\mathbb{P}\left(\|\lambda^* - \tilde{\lambda}_\mu\|_1 \geq t\right)$$

In addition we have that $\mathbb{E}(\|\lambda^* - \tilde{\lambda}_\mu\|_1) \leq \sqrt{\frac{r}{n}}$ and by applying the McDiarmid's Inequality, we obtain that for all $\delta > 0$, with a probability of $1-\delta$

$$\|\lambda^* - \tilde{\lambda}_\mu\|_1 \leq \sqrt{\frac{r}{n}} + \sqrt{\frac{2\ln(2/\delta)}{n}}$$

Therefore we obtain that with a probability of at least $1-\delta$,

$$\left|\int_{\mathcal{Z}^2} c(x,y)\left[\frac{1}{\lambda_r^*}-\frac{1}{\tilde{\lambda}_r}\right]d\tilde{\mu}_r(x)\tilde{\nu}_r(y)\right| \leq$$
$$2\|c\|_\infty\left[\sqrt{\frac{r}{n}}+\sqrt{\frac{2\ln(8/\delta)}{n}}+\frac{2\ln(8/\delta)+\sqrt{2r\ln(8/\delta)}}{n\times\min\limits_{1\leq i\leq r}\lambda_i^*}\right]$$

For the term in Eq. 6.14 and 6.15, we obtain that

$$\begin{aligned}
&\left|\frac{1}{\tilde{\lambda}_r}\sum_{k=1}^{r-1}\int_{\mathcal{Z}^2}\left(1-\frac{\min(|\tilde{\mu}_k|,|\tilde{\nu}_k|)}{|\tilde{\nu}_k|}\right)c(x,y)d\tilde{\mu}_r(x)d\tilde{\nu}_k(y)\right| \\
&\leq \frac{|\tilde{\mu}_r|}{\tilde{\lambda}_r}\sum_{k=1}^{r-1}\left(|\tilde{\nu}_k|-\min(|\tilde{\mu}_k|,|\tilde{\nu}_k|)\right)\|c\|_\infty \\
&\leq \frac{|\tilde{\mu}_r|}{\tilde{\lambda}_r}[\tilde{\lambda}_r-|\tilde{\nu}_r|]\|c\|_\infty \\
&\leq [|\tilde{\lambda}_r-\lambda_r^*|+|\lambda_r^*-\tilde{\nu}_r|]\|c\|_\infty \\
&\leq 3\|c\|_\infty\max(\|\lambda^*-\tilde{\lambda}_\mu\|_1,\|\lambda^*-\tilde{\lambda}_\nu\|_1)
\end{aligned}$$

Therefore we obtain that with a probability of at least $1-\delta$,

$$\begin{aligned}
&\left|\frac{1}{\tilde{\lambda}_r}\sum_{k=1}^{r-1}\int_{\mathcal{Z}^2}\left(1-\frac{\min(|\tilde{\mu}_k|,|\tilde{\nu}_k|)}{|\tilde{\nu}_k|}\right)c(x,y)d\tilde{\mu}_r(x)d\tilde{\nu}_k(y)\right| \\
&\leq 3\|c\|_\infty\left[\sqrt{\frac{r}{n}}+\sqrt{\frac{2\ln(2/\delta)}{n}}\right]
\end{aligned}$$

Finally the last term in Eq. 6.16 can be controlled as the following:

$$\begin{aligned}
&\left|\frac{1}{\tilde{\lambda}_r}\sum_{k,q=1}^{r-1}\int_{\mathcal{Z}^2}\left(1-\frac{\min(|\tilde{\mu}_k|,|\tilde{\nu}_k|)}{|\tilde{\mu}_k|}\right)\left(1-\frac{\min(|\tilde{\mu}_q|,|\tilde{\nu}_q|)}{|\tilde{\nu}_q|}\right)c(x,y)d\tilde{\mu}_k(x)d\tilde{\nu}_q(y)\right| \\
&\leq \frac{\|c\|_\infty}{\tilde{\lambda}_r}\sum_{k,q=1}^{r-1}\left(1-\frac{\min(|\tilde{\mu}_k|,|\tilde{\nu}_k|)}{|\tilde{\mu}_k|}\right)\left(1-\frac{\min(|\tilde{\mu}_q|,|\tilde{\nu}_q|)}{|\tilde{\nu}_q|}\right)|\tilde{\mu}_k||\tilde{\nu}_q| \\
&\leq \frac{\|c\|_\infty}{\tilde{\lambda}_r}\sum_{k=1}^{r-1}\left(|\tilde{\mu}_k|-\min(|\tilde{\mu}_k|,|\tilde{\nu}_k|)\right)\sum_{k=1}^{r-1}\left(|\tilde{\nu}_k|-\min(|\tilde{\mu}_k|,|\tilde{\nu}_k|)\right) \\
&\leq 3\|c\|_\infty\max(\|\lambda^*-\tilde{\lambda}_\mu\|_1,\|\lambda^*-\tilde{\lambda}_\nu\|_1)
\end{aligned}$$

and we obtain that with a probability of at least $1-\delta$,

$$\left|\frac{1}{\tilde{\lambda}_r}\sum_{k,q=1}^{r-1}\int_{\mathcal{Z}^2}\left(1-\frac{\min(|\tilde{\mu}_k|,|\tilde{\nu}_k|)}{|\tilde{\mu}_k|}\right)\left(1-\frac{\min(|\tilde{\mu}_q|,|\tilde{\nu}_q|)}{|\tilde{\nu}_q|}\right)c(x,y)d\tilde{\mu}_k(x)d\tilde{\nu}_q(y)\right|$$
$$\leq 3\|c\|_\infty\left[\sqrt{\frac{r}{n}}+\sqrt{\frac{2\ln(2/\delta)}{n}}\right]$$

Then by applying a union bound we obtain that with a probability of at least $1-\delta$

$$\left|\int_{\mathcal{Z}^2}c(x,y)d[\hat{\pi}(x,y)-\tilde{\pi}(x,y)]\right|\leq$$
$$\|c\|_\infty\left[11\sqrt{\frac{r}{n}}+12\sqrt{\frac{2\ln 40/\delta}{n}}+\frac{6\ln(40/\delta)+2\sqrt{2r\ln(40/\delta)}}{n\times\min_{1\leq i\leq r}\lambda_i^*}\right]$$

Now observe that

$$\begin{aligned}\mathrm{LOT}_{c,r}(\hat{\mu},\hat{\nu})-\mathrm{LOT}_{c,r}(\mu,\nu)&\leq\int_{\mathcal{Z}^2}c(x,y)d\tilde{\pi}(x,y)-\int_{\mathcal{Z}^2}c(x,y)d\pi^*(x,y)\\&\leq\int_{\mathcal{Z}^2}c(x,y)d[\tilde{\pi}-\hat{\pi}](x,y)+\int_{\mathcal{Z}^2}c(x,y)d[\hat{\pi}-\pi^*](x,y)\end{aligned}$$

and by combining the two control we obtain that with a probability of at least $1-2\delta$,

$$\begin{aligned}&\mathrm{LOT}_{c,r}(\hat{\mu},\hat{\nu})-\mathrm{LOT}_{c,r}(\mu,\nu)\\&\leq\|c\|_\infty\left[11\sqrt{\frac{r}{n}}+12\sqrt{\frac{2\ln 40/\delta}{n}}+\frac{1}{\alpha}\left(2\sqrt{\frac{\log(2/\delta)}{n}}+\frac{6\ln(40/\delta)+2\sqrt{2r\ln(40/\delta)}}{n}\right)\right]\\&\leq 11\|c\|_\infty\sqrt{\frac{r}{n}}+\frac{14\|c\|_\infty}{\alpha}\sqrt{\frac{\log(40/\delta)}{n}}+\frac{2\|c\|_\infty\max(6,\sqrt{2r})\log(40/\delta)}{n\alpha}\end{aligned}$$

where $\alpha:=\min_{1\leq i\leq r}\lambda_i^*$ and the result follows. □

### 6.9.5 Proof Proposition 6.4.3

**Proposition.** *Let $r\geq 1$, $\delta>0$ and $\mu,\nu\in\mathcal{M}_1^+(\mathcal{X})$. Then there exists a constant $N_{r,\delta}$ such that if $n\geq N_{r,\delta}$ then with a probability of at least $1-2\delta$, we have*

$$\mathrm{LOT}_{c,r}(\hat{\mu}_n,\hat{\nu}_n)-\mathrm{LOT}_{c,r}(\mu,\nu)\leq 11\|c\|_\infty\sqrt{\frac{r}{n}}+77\|c\|_\infty\sqrt{\frac{\log(40/\delta)}{n}}\,.$$

*Proof.* We consider the same notations as in the proof of Proposition 6.4.2. In particular let us define for all $(x_i, z_i)_{i=1}^n, (y_i, z_i') \in (\mathcal{Z} \times [|1, r|])^n$,

$$g((x_1, z_1), \dots, (x_n, z_n), (y_1, z_1'), \dots, (y_n, z_n')) := \sum_{q=1}^{r} \frac{1}{\lambda_q^* n^2} \sum_{i,j} c(x_i, y_j) \mathbf{1}_{z_i = q} \mathbf{1}_{z_j' = q} \ ,$$

Recall that we have

$$|g(\dots, (x_k, z_k), \dots) - g(\dots, (\tilde{x}_k, \tilde{z}_k), \dots)| \leq \frac{\|c\|_\infty}{n^2} \left[ \frac{\sum_{j=1}^n \mathbf{1}_{z_j' = z_k}}{\lambda_{z_k}^*} + \frac{\sum_{j=1}^n \mathbf{1}_{z_j' = \tilde{z}_k}}{\lambda_{\tilde{z}_k}^*} \right]$$

$$\leq \frac{2\|c\|_\infty}{n} \max \left( \left\| \frac{\tilde{\lambda}_\mu}{\lambda^*} \right\|_\infty, \left\| \frac{\tilde{\lambda}_\nu}{\lambda^*} \right\|_\infty \right)$$

$$\leq \frac{2\|c\|_\infty}{n} + \frac{2\|c\|_\infty}{n \times \min_{1 \leq i \leq r} \lambda_i^*} \max \left( \|\tilde{\lambda}_\mu - \lambda^*\|_\infty, \|\tilde{\lambda}_\nu - \lambda^*\|_\infty \right)$$

In fact if we have a control in probability of the bounded difference we can use an extension of the McDiarmid's Inequality. For that purpose let us first introduce the following definition.

**Definition 6.9.1.** *Let $(X_i)_{i=1}^m$, $m$ independent random variables and $g$ a measurable function. We say that $g$ is weakly difference-bounded with respect to $(X_i)_{i=1}^m$ by $(b, \beta, \delta)$ if*

$$\mathbb{P}\left(|g(X_1, \dots, X_m) - g(X_1', \dots, X_m')| \leq \beta\right) \geq 1 - \delta$$

*with $X_i' = X_i$ except for one coordinate $k$ where $X_k'$ is an independent copy of $X_k$. Furthermore for any $(x_i)_{i=1}^m$ and $(x_i')_{i=1}^m$ where for all coordinate except on $x_j = x_j'$*

$$|g(x_1, \dots, x_m) - g(x_1', \dots, x_m')| \leq b \ .$$

Let us now introduce an extension of McDiarmid's Inequality [208].

**Theorem 6.9.1.** *Let $(X_i)_{i=1}^m$, $m$ independent random variables and $g$ a measurable function which is weakly difference-bounded with respect to $(X_i)_{i=1}^m$ by $(b, \beta/m, \exp(-Km))$, then if $0 < \tau \leq T(b, \beta, K)$ and $m \geq M(b, \beta, K, \tau)$, then*

$$\mathbb{P}(|g(X_1, \dots, X_m) - \mathbb{E}(g(X_1, \dots, X_m))| \geq \tau) \leq 4 \exp\left( \frac{-\tau^2 m}{8\beta^2} \right)$$

*where*

$$T(b, \beta, K) := \min \left( \frac{14c}{2}, 4\beta\sqrt{K}, \frac{\beta^2 K}{b} \right)$$

$$M(b, \beta, K, \tau) := \max \left( \frac{b}{\beta}, \beta\sqrt{40}, 3\left( \frac{24}{K} + 3 \right) \log\left( \frac{24}{K} + 3 \right), \frac{1}{\tau} \right)$$

Given the above Theroem we can obtain an asymptotic control of the deviation of $g$ from its mean. Let $\delta' > 0$ and let us denote

$$\begin{aligned} m &:= 2n \\ b &:= \frac{2\|c\|_\infty}{n \times \min\limits_{1\leq i\leq r} \lambda_i^*} \\ K &:= \frac{\log(1/\delta')}{2n} \\ \beta &:= 4\|c\|_\infty \left[1 + \frac{1}{\min\limits_{1\leq i\leq r} \lambda_i^*}\sqrt{\frac{2\log(4/\delta')}{n}}\right] \end{aligned}$$

Observe now that with a probability of at least $1 - \exp(-Km)$

$$|g(\dots,(x_k,z_k),\dots) - g(\dots,(\tilde{x}_k,\tilde{z}_k),\dots)| \leq \frac{2\|c\|_\infty}{n}\left[1 + \frac{1}{\min\limits_{1\leq i\leq r} \lambda_i^*}\sqrt{\frac{2\log(4/\delta')}{n}}\right]$$

Let us now fix $\delta > 0$ and let us choose $\delta'$ such that $\delta' := 4/n$ and $\tau := \beta\sqrt{\frac{4\log(4/\delta)}{n}}$, then we obtain that for $n$ sufficiently large (such that $n \geq M(b,\beta,K,\tau)/2$ and $\tau \leq T(b,\beta,K)$), we have that with a probability of at least $1-\delta$

$$\begin{aligned} \left|\mathrm{LOT}_{c,r}(\mu,\nu) - \int_{\mathcal{Z}^2} c(x,y)d\hat{\pi}(x,y)\right| &\leq 4\|c\|_\infty\left[1 + \frac{1}{\min\limits_{1\leq i\leq r} \lambda_i^*}\sqrt{\frac{2\log(n)}{n}}\right]\sqrt{\frac{4\log(4/\delta)}{n}} \\ &\leq 4\|c\|_\infty\sqrt{\frac{4\log(4/\delta)}{n}} + \frac{16\sqrt{5}\|c\|_\infty\sqrt{\log(n)\log(4/\delta)}}{n \times \min\limits_{1\leq i\leq r} \lambda_i^*} \end{aligned}$$

Recall also from the proof of Proposition 6.4.2, that we have with a probability of at least $1-\delta$

$$\begin{aligned} &\left|\int_{\mathcal{Z}^2} c(x,y)d[\hat{\pi}(x,y) - \tilde{\pi}(x,y)]\right| \\ &\leq \|c\|_\infty\left[11\sqrt{\frac{r}{n}} + 12\sqrt{\frac{2\ln 40/\delta}{n}} + \frac{6\ln(40/\delta) + 2\sqrt{2r\ln(40/\delta)}}{n \times \min\limits_{1\leq i\leq r} \lambda_i^*}\right] \end{aligned}$$

Finally by imposing in addition that

$$\sqrt{\frac{n}{\log(n)}} \geq \frac{1}{\min\limits_{1\leq i\leq r} \lambda_i^*} \quad , \quad \sqrt{n} \geq \frac{\sqrt{\log(40/\delta)}}{\min\limits_{1\leq i\leq r} \lambda_i^*} \quad \text{and} \quad \sqrt{n} \geq \frac{\sqrt{r}}{\min\limits_{1\leq i\leq r}}$$

we obtain that for $n$ is large enough (such that (such that $n \geq M(b, \beta, K, \tau)/2$ and $\tau \leq T(b, \beta, K)$) and satysfing the above inequalities, we have with a probability of at least $1 - 2\delta$ that

$$\mathrm{LOT}_{c,r}(\hat{\mu}, \hat{\nu}) - \mathrm{LOT}_{c,r}(\mu, \nu) \leq 11\|c\|_\infty \sqrt{\frac{r}{n}} + 77\|c\|_\infty \sqrt{\frac{\log(40/\delta)}{n}}$$

□

### 6.9.6 Proof Proposition 6.5.1

**Proposition.** *Let $\mu, \nu \in \mathcal{M}_1^+(\mathcal{X})$. Let us assume that $c$ is symmetric, then we have*

$$\mathrm{DLOT}_{1,c}(\mu, \nu) = \frac{1}{2} \int_{\mathcal{X}^2} -c(x, y) d[\mu - \nu] \otimes d[\mu - \nu](x, y) \ .$$

*If in addition we assume the $c$ is Lipschitz w.r.t to $x$ and $y$, then we have*

$$\mathrm{DLOT}_{c,r}(\mu, \nu) \xrightarrow[r \to +\infty]{} \mathrm{OT}_c(\mu, \nu) \ .$$

*Proof.* When $r = 1$, it is clear that for any $\mu, \nu \in \mathcal{M}_1^+(\mathcal{X})$, $\Pi_r(\mu, \nu) = \{\mu \otimes \nu\}$ and thanks to the symmetry of $c$, we have directly that

$$\mathrm{DLOT}_{1,c}(\mu, \nu) = \frac{1}{2} \int_{\mathcal{X}^2} -c(x, y) d[\mu - \nu] \otimes d[\mu - \nu](x, y) = \frac{1}{2} \mathrm{MMD}_{-c}(\mu, \nu) \ .$$

The limit is a direct consequence of Proposition 6.3.2. □

### 6.9.7 Proof of Proposition 6.5.3

**Proposition.** *Let $r \geq 1$ and $(\mu_n)_{n \geq 0}$ and $(\nu_n)_{n \geq 0}$ two sequences of probability measures such that $\mu_n \to \mu$ and $\nu_n \to \nu$ with respect to the convergence in law. Then we have that*

$$LOT_{c,r}(\mu_n, \nu_n) \to LOT_{c,r}(\mu, \nu) \ .$$

*Proof.* Let us denote $\pi$ an optimal solution of $\mathrm{LOT}_{c,r}(\mu, \nu)$ and let us denote $(\mu^{(i)})_{i=1}^r$, $(\nu^{(i)})_{i=1}^r$ and $(\lambda^{(i)})_{i=1}^r$ the decomposition associated. In the following Lemma, we aim at building specific decompositions of the sequences $(\mu_n)_{n \geq 0}$ and $(\nu_n)_{n \geq 0}$.

**Lemma 7.** *Let $r \geq 1$, $\mu \in \mathcal{M}_1^+(\mathcal{X})$ and $(\mu^{(i)})_{i=1}^r \in \mathcal{M}_1^+(\mathcal{X})$ and $(\lambda^{(i)})_{i=1}^r \in \Delta_r^*$ such that $\mu = \sum_{i=1}^r \lambda_i \mu^{(i)}$. Then for any sequence of probability measures $(\mu_n)_{\geq 0}$*

*such that $\mu_n \to \mu$, there exist for all $i \in [|1, r|]$ a sequence of nonnegative measures $(\mu_n^{(i)})_{n\geq 0}$ such that*

$$\mu_n^{(i)} \to \lambda_i \mu^{(i)} \quad \textit{for all } i \in [|1, r|] \textit{ and}$$
$$\sum_{i=1}^{r} \mu_n^{(i)} = \mu_n \quad \textit{for all } n \geq 0$$

*Proof.* For $r = 1$ the result is clear. Let us now show the result for $r = 2$. Let us denote $(\tilde{\mu}_n^{(1)})$ a sequence converging weakly towards $\lambda_1 \mu^{(1)}$. Then by denoting $\mu_n^{(1)} := \mu_n - (\mu_n - \tilde{\mu}_n^{(1)})_+$ where $(\cdot)_+$ correspond to the non-negative part of the measure, we have that

$$\begin{aligned}
&\mu_n^{(1)} \geq 0, \quad \mu_n^{(1)} \to \lambda_1 \mu^{(1)}, \\
&\mu_n^{(2)} := \mu_n - \mu_n^{(1)} \geq 0, \quad \mu_n^{(2)} \to \lambda_2 \mu^{(2)} \quad \text{and} \\
&\mu_n = \mu_n^{(1)} + \mu_n^{(2)} \quad \text{for all} \quad n \geq 0
\end{aligned}$$

which is the result. Let $r \geq 2$ and let us assume that the result holds for all $1 \leq k \leq r$. Let us now consider a decomposition of $\mu$ such that $\mu = \sum_{i=1}^{r+1} \lambda_i \mu^{(i)}$. By denoting $\tilde{\mu}^{(1)} := \frac{\sum_{i=1}^{r} \lambda_i \mu^{(i)}}{\sum_{i=1}^{r} \lambda_i}$, we obtain that

$$\mu = \left( \sum_{i=1}^{r} \lambda_i \right) \tilde{\mu}^{(1)} + \lambda_{r+1} \mu^{(r+1)} \, .$$

Then by recursion we have that there exists sequences of nonnegative measures $(\tilde{\mu}_n^{(1)})$ and $(\mu_n^{(r+1)})$ such that

$$\tilde{\mu}_n^{(1)} \to \left( \sum_{i=1}^{r} \lambda_i \right) \tilde{\mu}^{(1)}, \quad \mu_n^{(r+1)} \to \lambda_{r+1} \mu^{(r+1)} \quad \text{and} \quad \mu_n = \tilde{\mu}_n^{(1)} + \mu_n^{(r+1)} \quad \text{for all} \quad n \geq 0$$

Now observe that $\frac{\tilde{\mu}_n^{(1)}}{|\tilde{\mu}_n^{(1)}|} \to \tilde{\mu}^{(1)} = \sum_{i=1}^{r} \frac{\lambda_i}{\sum_{i=1}^{r} \lambda_i} \mu^{(i)}$. Therefore applying the recursion on this problem allows us to obtain a decomposition of $\tilde{\mu}_n^{(1)}$ of the form

$$\begin{aligned}
&\frac{\tilde{\mu}_n^{(1)}}{|\tilde{\mu}_n^{(1)}|} = \sum_{i=1}^{r} \mu_n^{(i)} \quad \text{where} \\
&\mu_n^{(i)} \geq 0 \quad \text{and} \quad \mu_n^{(i)} \to \frac{\lambda_i}{\sum_{i=1}^{r} \lambda_i} \mu^{(i)} \, .
\end{aligned}$$

Therefore we obtain that

$$\begin{aligned}\mu_n &= \sum_{i=1}^{r} |\tilde{\mu}_n^{(1)}|\mu_n^{(i)} + \mu_n^{(r+1)} \quad \text{where} \\ \mu_n^{(i)} &\geq 0, \quad |\tilde{\mu}_n^{(1)}|\mu_n^{(i)} \to \lambda_i \mu^{(i)} \quad \text{for all} \quad i \in [|1, r|] \quad \text{and} \\ \mu_n^{(r+1)} &\geq 0, \quad \mu_n^{(r+1)} \to \lambda_{r+1}\mu^{(r+1)}\end{aligned}$$

from which follows the result. □

Let us now consider such decompositions of $(\mu_n)_{n\geq 0}$ and $(\nu_n)_{n\geq 0}$ such that each factor converges toward the target decomposition of $\mu$. Now let us build the following coupling:

$$\begin{aligned}\tilde{\pi}_n := &\sum_{k=1}^{r-1} \frac{\min(|\mu_n^{(k)}|, |\nu_n^{(k)}|)}{|\mu_n^{(k)}||\nu_n^{(k)}|}\mu_n^{(k)} \otimes \mu_n^{(k)} \\ &+ \frac{1}{1 - \sum_{k=1}^{r-1} \min(|\mu_n^{(k)}|, |\nu_n^{(k)}|)} \\ &\left[|\mu_n| - \sum_{k=1}^{r-1} \frac{\min(|\mu_n^{(k)}|, |\nu_n^{(k)}|)}{|\mu_n^{(k)}|}\mu_n^{(k)}\right] \otimes \left[\nu_n - \sum_{k=1}^{r-1} \frac{\min(|\mu_n^{(k)}|, |\nu_n^{(k)}|)}{|\nu_n^{(k)}|}\nu_n^{(k)}\right]\end{aligned}$$

with the convention that $\frac{0}{0} = 0$. Now it is easy to check that $\tilde{\pi}_n \in \Pi_r(\mu_n, \nu_n)$, and we have that

$$\mathrm{LOT}_{c,r}(\mu_n, \nu_n) \leq \int_{\mathcal{X}^2} d(x, y) d\tilde{\pi}_n(x, y) \to \mathrm{LOT}_{c,r}(\mu, \nu)$$

and by Prokhorov's theorem and the optimality of the limit of $(\tilde{\pi}_n)_{n\geq 0}$ (up to an extraction) we obtain that $\mathrm{LOT}_{c,r}(\mu_n, \nu_n) \to \mathrm{LOT}_{c,r}(\mu, \nu)$. □

### 6.9.8 Proof Proposition 6.5.2

**Proposition.** *Let $r \geq 1$, and let us assume that $c$ is a semimetric of negative type. Then for all $\mu, \nu \in \mathcal{M}_1^+(\mathcal{X})$, we have that*

$$DLOT_{c,r}(\mu, \nu) \geq 0 \, .$$

*In addition, if $c$ has strong negative type then we have also that*

$$\begin{aligned}\mathrm{DLOT}_{c,r}(\mu, \nu) = 0 \iff \mu = \nu \quad &\textit{and} \\ \mu_n \to \mu \iff \mathrm{DLOT}_{c,r}(\mu_n, \mu) \to 0 \, .&\end{aligned}$$

*where the convergence of the sequence of probability measures considered is the convergence in law.*

*Proof.* Let $\pi^*$ solution of $\mathrm{LOT}_{c,r}(\mu,\nu)$. Then there exists $\lambda^* \in \Delta_r^*$, $(\mu_i^*)_{i=1}^r, (\nu_i^*)_{i=1}^r \in \mathcal{M}_1^+(\mathcal{X})^r$ such that

$$\pi^* = \sum_{i=1}^r \lambda_i^* \mu_i^* \otimes \nu_i^*.$$

Note that by definition, we have that

$$\mu = \sum_{i=1}^r \lambda_i^* \mu_i^* \quad \text{and} \quad \nu = \sum_{i=1}^r \lambda_i^* \nu_i^*,$$

By definition we have also that

$$\mathrm{LOT}_{c,r}(\mu,\mu) \leq \sum_{k=1}^r \lambda_k^* \int_{\mathcal{X}^2} c(x,y) d\mu_k^* \otimes \mu_k^*$$

similarly for $\mathrm{LOT}_{c,r}(\nu,\nu)$ we have

$$\mathrm{LOT}_{c,r}(\nu,\nu) \leq \sum_{k=1}^r \lambda_k^* \int_{\mathcal{X}^2} c(x,y) d\nu_k^* \otimes \nu_k^*$$

Therefore we have

$$\begin{aligned}
\mathrm{DLOT}_{c,r}(\mu,\nu) &\geq \sum_{k=1}^r \lambda_k^* \left( \int_{\mathcal{X}^2} c(x,y) d\mu_k^* \otimes \nu_k^* - \frac{1}{2} \left[ \int_{\mathcal{X}^2} c(x,y) d\mu_k^* \otimes \mu_k^* + \int_{\mathcal{X}^2} c(x,y) d\nu_k^* \otimes \nu_k^* \right] \right) \\
&\geq \sum_{k=1}^r \lambda_k^* \int_{\mathcal{X}^2} -c(x,y) d[\mu_k^* - \nu_k^*] \otimes [\mu_k^* - \nu_k^*] \\
&\geq \sum_{k=1}^r \frac{\lambda_k^*}{2} D_c(\mu_k^*, \nu_k^*)
\end{aligned}$$

where for any any probability measures $\alpha, \beta$ on $\mathcal{X}$ we define

$$D_c(\alpha,\beta) := 2 \int_{\mathcal{X}^2} c(x,y) d\alpha \otimes \beta - \int_{\mathcal{X}^2} c(x,y) d\alpha \otimes \alpha - \int_{\mathcal{X}^2} c(x,y) d\beta \otimes \beta$$

However, as $c$ is assumed to have a negative type, we have that

$$D_c(\mu_k^*, \nu_k^*) \geq 0 \quad \forall k$$

In addition if we assume that $c$ has a strong negative type, then we obtain directly that

$$\mathrm{DLOT}_{c,r}(\mu,\nu) = 0 \implies \mu_k^* = \nu_k^* \quad \forall k \, .$$

Let us now show that $\mathrm{DLOT}_{c,r}$ metrize the convergence in law. The direct implication is a direct consequence of the Proposition 6.5.3. Conversely, if $\mathrm{DLOT}_{c,r}(\mu_n, \mu) \to 0$, then by compacity of $\mathcal{X}$ and thanks to the Prokhorov's theorem we can extract a subsequence of $\mu_n \to \mu^*$, and thanks to Proposition 6.5.3, we also obtain that $\mathrm{DLOT}_{c,r}(\mu_n, \mu) \to \mathrm{DLOT}_{c,r}(\mu^*, \mu)$. Finally we deduce that $\mathrm{DLOT}_{c,r}(\mu^*, \mu) = 0$ and $\mu^* = \mu$.

□

### 6.9.9 Proof Proposition 6.5.4

**Proposition.** *Let $n \geq k \geq 1$, $X := \{x_1, \dots, x_n\} \subset \mathcal{X}$ and $a \in \Delta_n^*$. If $c$ is a semimetric of negative type, then by denoting $C = (c(x_i, x_j))_{i,j}$, we have that*

$$\mathrm{LOT}_{c,k}(\mu_{a,X}, \mu_{a,X}) = \min_{Q} \langle C, Q\, diag(1/Q^T \mathbf{1}_n) Q^T \rangle \quad s.t. \quad Q \in \mathbb{R}_+^{n \times k} \quad , \quad Q\mathbf{1}_k = a \; . \tag{6.17}$$

*Proof.* First remarks that one can reformulate the $\mathrm{DLOT}_{c,k}$ problem as

$$\mathrm{DLOT}_{c,k}(\mu, \mu) := \min_{g \in \Delta_k^*} \min_{(X,Y) \in K_{a,g}^2} \sum_{i=1}^{k} \frac{X_i^T C Y_i}{g_i}$$

where

$$\begin{aligned} K_{a,g} &:= \{X \in \mathbb{R}^{nk} \text{ s.t. } AX = [a, g]^T, \; X \geq 0\} \\ A &:= \begin{pmatrix} \mathbf{1}_n^T \otimes \mathbb{I}_k \\ \mathbb{I}_n^T \otimes \mathbf{1}_k \end{pmatrix} \quad \text{and} \\ X_i &:= [x_{(i-1)\times n+1}, \dots, x_{i \times n}]^T, \quad Y_i := [y_{(i-1)\times n+1}, \dots, y_{i \times n}]^T \quad \text{for all} \quad i \in [|1, k|] \end{aligned}$$

Indeed the above optimization problem is just a reformulation of $\mathrm{DLOT}_{c,k}(\mu, \mu)$ where we have vectorized the couplings in a column-wise order. Let us now show the following lemma from which the result will follow.

**Lemma 8.** *Under the same assumption of Proposition 6.5.4 we have that for all $g \in \Delta_k^*$*

$$\min_{(X,Y) \in K_{a,g}^2} \sum_{i=1}^{k} \frac{X_i^T C Y_i}{g_i} = \min_{X \in K_{a,g}} \sum_{i=1}^{k} \frac{X_i^T C X_i}{g_i}$$

*Proof.* Let $(X^*, Y^*)$ solution of the LHS optimization problem. Then we have that

$$\sum_{i=1}^{k} \frac{(X_i^*)^T C X_i^*}{g_i} \geq \sum_{i=1}^{k} \frac{(X_i^*)^T C Y_i^*}{g_i}$$
$$\sum_{i=1}^{k} \frac{(Y_i^*)^T C Y_i^*}{g_i} \geq \sum_{i=1}^{k} \frac{(X_i^*)^T C Y_i^*}{g_i}$$

Therefore we obtain that

$$0 \leq \sum_{i=1}^{k} \frac{(X_i^*)^T C X_i^*}{g_i} - \sum_{i=1}^{k} \frac{(X_i^*)^T C Y_i^*}{g_i} = \sum_{i=1}^{k} \frac{(X_i^*)^T C (X_i^* - Y_i^*)}{g_i}$$
$$0 \leq \sum_{i=1}^{k} \frac{(Y_i^*)^T C Y_i^*}{g_i} - \sum_{i=1}^{k} \frac{(X_i^*)^T C Y_i^*}{g_i} = \sum_{i=1}^{k} \frac{(Y_i^* - X_i^*)^T C Y_i^*}{g_i}$$

Then by symmetry of $C$, we obtain by adding the two terms that

$$\sum_{i=1}^{k} \frac{(X_i^* - Y_i^*)^T C (X_i^* - Y_i^*)}{g_i} \geq 0$$

However, thanks to the linear constraints, we have that for all $i \in [|1, k|]$,

$$\sum_{q=0}^{n-1} x^*_{(i-1)\times n+1+q} = \sum_{q=0}^{n-1} y^*_{(i-1)\times n+1+q} = g_i$$

Therefore $(X_i^* - Y_i^*)^T \mathbf{1}_n = 0$ and thanks to the negativity of the cost function $c$ we obtain that

$$(X_i^* - Y_i^*)^T C (X_i^* - Y_i^*) \leq 0$$

Therefore we have that

$$(X_i - Y_i)^T C (X_i - Y_i) = 0$$

from which follows that

$$\sum_{i=1}^{k} \frac{(X_i^*)^T C X_i^*}{g_i} = \sum_{i=1}^{k} \frac{(X_i^*)^T C Y_i^*}{g_i} = \sum_{i=1}^{k} \frac{(Y_i^*)^T C Y_i^*}{g_i}$$

and the result follows. □

As the above result holds for any $g \in \Delta_k^*$, we obtain that

$$\mathrm{DLOT}_{c,k}(\mu,\mu) = \min_{g\in\Delta_k^*} \min_{X\in K_{a,g}} \sum_{i=1}^{k} \frac{(X_i^*)^T C X_i^*}{g_i}$$

Then by formulating back this problem in term of matrices, we obtain that

$$\mathrm{DLOT}_{c,k}(\mu,\mu) = \min_{g\in\Delta_k^*} \min_{Q\in\Pi_{a,g}} \langle C, Q\mathrm{diag}(1/g)Q^T\rangle$$

from which the result follows. □

## 6.10 Additional Experiments

### 6.10.1 Comparison of the $\gamma$ schedules

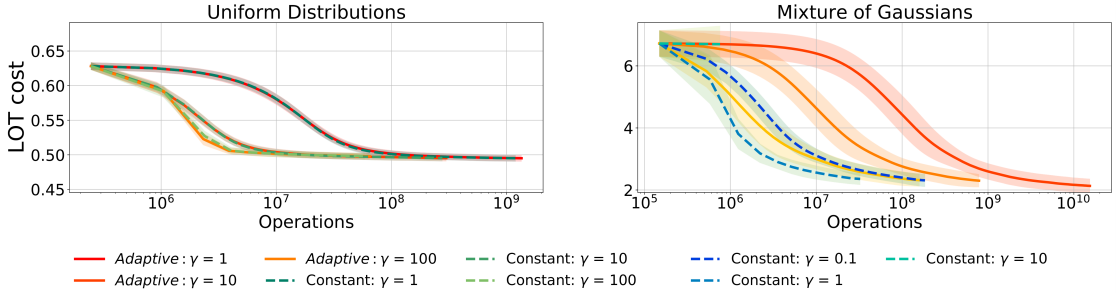



Figure 6.5: In this experiment, we compare two strategies for the choice of the step-size in the MD scheme proposed by [3] on two different problems. More precisely, we compare the constant $\gamma$ schedule with the proposed adaptive one and compare them when the distributions are sampled from either uniform distributions (*left*) or mixtures of anisotropic Gaussians (*right*). We show that the range of admissible $\gamma$ when considering a constant schedule varies from one problem to another. Indeed, in the right plot, we observe that the algorithm converges only when $\gamma \leq 1$, while in the left plot, the algorithm manages to converge for $\gamma \leq 100$. We also observe that our adaptive strategy allows to have a consistent choice of admissible values for $\gamma$ whatever the problem considered. It is worth noticing that whatever the $\gamma$ chosen, the algorithm converges towards the same value, however the larger $\gamma$ is chosen in its admissible range, the faster the algorithm converges.

### 6.10.2 Gradient Flows between two Moons

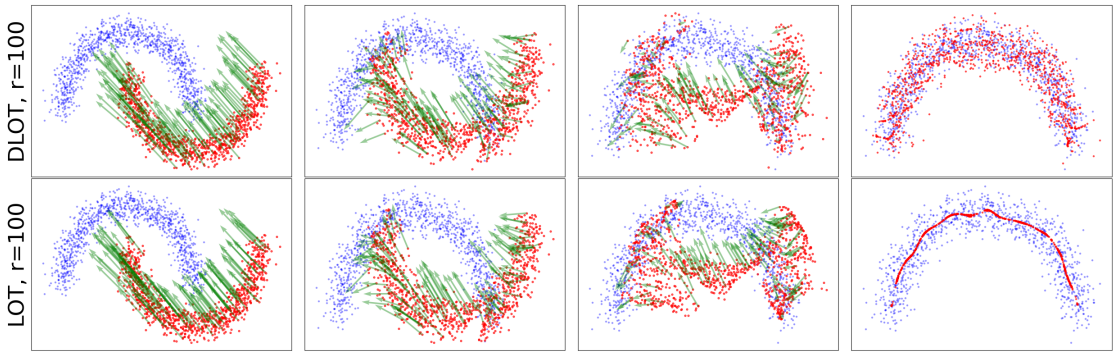



Figure 6.6: We compare the gradient flows $(\mu_t)_{t\geq 0}$ (in red) starting from a moon shape distribution, $\mu_0$, to another moon shape distribution (in blue), $\nu$, in 2D when minimizing either $L(\mu) := \mathrm{DLOT}_{c,r}(\mu, \nu)$ or $L(\mu) := \mathrm{LOT}_{c,r}(\mu, \nu)$. The ground cost is the squared Euclidean distance and we fix $r = 100$. We consider 1000 samples from each distribution and and we plot the evolution of the probability measure obtained along the iterations of a gradient descent scheme. We also display in green the vector field in the descent direction. We show that the debiased version allows to recover the target distribution while $\mathrm{LOT}_{c,r}$ is learning a biased version with a low-rank structure.

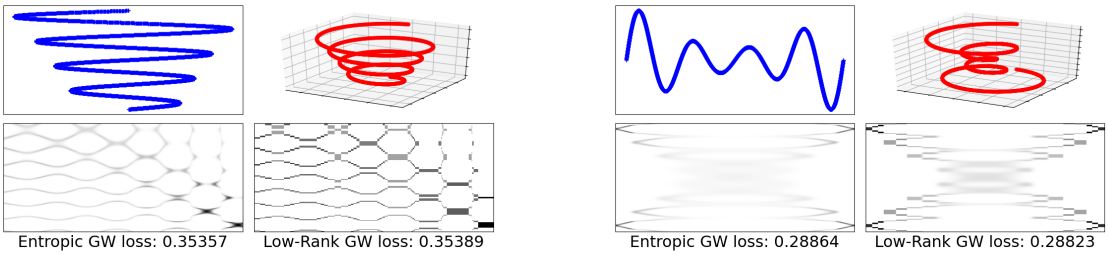
# Chapter 7

# Low-rank Gromov Wasserstein Distances

The ability to align points across two related yet incomparable point clouds (e.g. living in different spaces) plays an important role in machine learning. The Gromov-Wasserstein (GW) framework provides an increasingly popular answer to such problems, by seeking a low-distortion, geometry-preserving assignment between these points. As a non-convex, quadratic generalization of optimal transport (OT), GW is NP-hard. While practitioners often resort to solving GW approximately as a nested sequence of entropy-regularized OT problems, the cubic complexity (in the number $n$ of samples) of that approach is a roadblock. We show in this work how a recent variant of the OT problem that restricts the set of admissible couplings to those having a low-rank factorization is remarkably well suited to the resolution of GW: when applied to GW, we show that this approach is not only able to compute a stationary point of the GW problem in time $O(n^2)$, but also uniquely positioned to benefit from the knowledge that the initial cost matrices are low-rank, to yield a linear time $O(n)$ GW approximation. Our approach yields similar results, yet orders of magnitude faster computation than the SoTA entropic GW approaches, on both simulated and real data.

This chapter is based on [2].

## 7.1 Introduction




Example 1 Example 2

Figure 7.1: Two examples where we compare the low-rank regularization with the entropic one for the GW problem. For each example we describe the structure of the associated figure in the following. *Top row:* Two curves in 2D and 3D, with $n = 5000$ points. *Bottom row:* coupling and GW loss obtained with the SoTA $O(n^3)$ entropic approach [47] (left) and with our linear $O(n)$ method (right) when using the squared Euclidean distances as the ground costs. See Appendix 7.11.1 for more details.

**Increasing interest for Gromov-Wasserstein...** Several problems in machine learning require comparing datasets that live in heterogeneous spaces. This situation arises typically when realigning two distinct views (or features) from points sampled from similar sources. Recent applications to single-cell genomics [61, 62] provide a case in point: Thousands of cells taken from the same tissue are split in two groups, each processed with a different experimental protocol, resulting in two distinct sets of heterogeneous feature vectors; Despite this heterogeneity, one expects to find a mapping registering points from the first to the second set, since they contain similar overall information. That realignment is usually carried out using the Gromov-Wasserstein (GW) machinery proposed by [86] . GW seeks a relaxed assignment matrix that is as close to an isometry as possible, as quantified by a quadratic score. GW has practical appeal: It has been used in supervised learning [63], generative modeling [64], domain adaptation [65], structured prediction [66], quantum chemistry [47] and alignment layers [67].

**...despite being hard to solve.** Since GW is an NP-hard problem, all applications above rely on heuristics, the most popular being the sequential resolution of nested entropy-regularized OT problems. That approximation remains costly, requiring $\mathcal{O}(n^3)$ operations when dealing with two datasets of $n$ samples. Our goal

is to reduce that complexity, by exploiting and/or enforcing low-rank properties of matrices arising *both* in data and variables of the GW problem.

**OT: from cubic to linear complexity.** Compared to GW, aligning two populations embedded in the *same* space is far simpler, and corresponds to the usual optimal transport (OT) problem [33].Given a $n \times m$ cost matrix $C$ and two marginals, the OT problem minimizes $\mathcal{L}_C(P) := \langle C, P \rangle$ w.r.t. a coupling matrix $P$ satisfying these marginal constraints. For computational and statistical reasons, most practitioners rely on regularized approaches $\mathcal{L}_C^\varepsilon(P) := \langle C, P \rangle + \varepsilon \mathrm{reg}(P)$. When reg is the neg-entropy, Sinkhorn algorithm [215] can be efficiently employed [76, 117, 216]. The Sinkhorn iteration has $O(nm)$ complexity, but this can be sped-up using either a low-rank factorizations (or approximations) of the *kernel* matrix $K := e^{-C/\varepsilon}$ [121, 122, 217, 6], or, alternatively and as proposed by [3, 128], by imposing a low-rank *constraint* on the coupling $P$. A goal in this paper is to show that this latter route is remarkably well suited to the GW problem.

**GW: from NP-hard to cubic approximations.** The GW problem replaces the linear objective in OT by a *non-convex*, *quadratic*, objective $\mathcal{Q}_{A,B}(P) := \mathrm{cst} - 2\langle APB, P \rangle$ parameterized by *two* square cost matrices $A$ and $B$. Much like OT is a relaxation of the optimal assignment problem, GW is a relaxation of the quadratic assignment problem (QAP). Both GW and QAP are NP-hard [131]. In practice, linearizing iteratively $\mathcal{Q}_{A,B}$ works well [84, 85]: recompute a synthetic cost $C_t := AP_{t-1}B$, use Sinkhorn to get $P_t := \mathrm{argmin}_P \langle C_t, P \rangle + \varepsilon \mathrm{reg}(P)$, repeat. This is akin to a mirror-descent scheme [47], interpreted as a bi-linear relaxation in certain cases [218].

**Challenges to speed up GW.** Several obstacles stand in the way of speeding up GW. The re-computation of $C_t = AP_{t-1}B$ at each outer iteration is an issue, since it requires $O(n^3)$ operations [47, Prop. 1]. We only know of two broad approaches that achieve tractable running times: *(i)* Solve related, yet significantly different, proxies of the GW energy, either by embedding points *as* univariate measures [86, 87], by using a sliced mechanism when restricted to Euclidean settings [88] or by considering tree metrics for supports of each probability measure [89], *(ii)* Reduce the size of the GW problem through quantization of input measures [90]. or recursive clustering approaches [91, 62]). Interestingly, no work has, to our knowledge, tried yet to accelerate Sinkhorn iterations withing GW.

**Our contributions: a quadratic to linear GW approximation.** Our method addresses the problem by taking the GW as it is, overcoming limitations that may arise from a changing cost matrix $C_t$. We show first that a low-rank factorization

(or approximation) of the two input cost matrices that define GW, one for each measure, can be exploited to lower the complexity of recomputing $C_t$ from cubic to quadratic. We show next, independently, that using the low-rank approach for *couplings* advocated by [3] to solve OT can be inserted in the GW pipeline and result in a $O(n^2)$ strategy for GW, with no prior assumption on input cost matrices. We also briefly explain why methods that exploit the geometrical properties of $C$ (or its kernel $K = e^{-C}$) to obtain faster iterations are of little use in a GW setup, because of the necessity to re-instantiate a new cost $C_t$ at each outer iteration. Finally, we show that both low-rank assumptions (on costs and couplings) can be combined to shave yet another factor and reach GW approximation with linear complexity in time and memory. We provide experiments, on simulated and real datasets, which show that our approach has comparable performance to entropic-regularized GW and its practical ability to reach "good" local minima to GW, for a considerably cheaper computational price, and with a conceptually different regularization path (see examples in Fig. 7.1), yet can scale to millions of points.

## 7.2 Background on Gromov-Wasserstein

**Comparing metric measure spaces.** Let $(\mathcal{X}, d_{\mathcal{X}})$ and $(\mathcal{Y}, d_{\mathcal{Y}})$ be two metric spaces, and $\mu := \sum_{i=1}^{n} a_i \delta_{x_i}$ and $\nu := \sum_{i=j}^{m} b_j \delta_{y_j}$ two discrete probability measures, where $n, m \geq 1$; $a$, $b$ are probability vectors in the simplicies $\Delta_n, \Delta_m$ of size $n$ and $m$; and $(x_1, \dots, x_n)$, $(y_1, \dots, y_m)$ are families in $\mathcal{X}$ and $\mathcal{Y}$. Given $q \geq 1$, the following square pairwise *cost* matrices encode the geometry *within* $\mu$ and $\nu$,

$$A := [d^q_{\mathcal{X}}(x_i, x_{i'})]_{1 \leq i,i' \leq n}, \; B := [d^q_{\mathcal{Y}}(x_j, x_{j'})]_{1 \leq i,i' \leq m}$$

The GW discrepancy between these two discrete metric measure spaces $(\mu, d_{\mathcal{X}})$ and $(\nu, d_{\mathcal{Y}})$ is the solution of the following non-convex quadratic problem, written for simplicity as a function of $(a, A)$ and $(b, B)$:

$$\begin{aligned} \mathrm{GW}((a, A), (b, B)) &= \min_{P \in \Pi_{a,b}} \mathcal{Q}_{A,B}(P), \\ \text{where } \Pi_{a,b} &:= \{P \in \mathbb{R}_+^{n \times m} | P\mathbf{1}_m = a, P^T \mathbf{1}_n = b\}, \end{aligned} \tag{7.1}$$

and the energy $\mathcal{Q}_{A,B}$ is a quadratic function of $P$ designed to measure the distortion of the assignment:

$$\mathcal{Q}_{A,B}(P) := \sum_{i,j,i',j'} (A_{i,i'} - B_{j,j'})^2 P_{i,j} P_{i',j'} \; . \tag{7.2}$$

[86] proves that $\mathrm{GW}^{\frac{1}{2}}$ defines a distance on the space of metric measure spaces quotiented by measure-preserving isometries. (7.2) can be evaluated in $\mathcal{O}(n^2 m +$

$nm^2$) operations, rather than using $n^2m^2$ terms:

$$\mathcal{Q}_{A,B}(P) = \langle A^{\odot 2}a, a\rangle + \langle B^{\odot 2}b, b\rangle - 2\langle APB, P\rangle \,, \tag{7.3}$$

where $\odot$ is the Hadamard (elementwise) product or power.

**Entropic Gromov-Wasserstein.** The original GW problem (7.1) can be regularized using entropy [84, 85], leading to problem:

$$\mathrm{GW}_\varepsilon((a,A),(b,B)) = \min_{P\in\Pi_{a,b}} \mathcal{Q}_{A,B}(P) - \varepsilon H(P)\,, \tag{7.4}$$

where $H(P) := -\sum_{i,j} P_{i,j}(\log(P_{i,j}) - 1)$ is $P$'s entropy. [47] propose to solve that problem using mirror descent (MD), w.r.t. the KL divergence. Their algorithm boils down to solving a sequence of regularized OT problems, as in Algo. 10: Each KL projection in Line 5 is solved efficiently with the Sinkhorn algorithm [76].

**Algorithm 10** Entropic-GW

**Input :** $a \in \Delta_n,\ A \in \mathbb{R}^{n\times n},\ b \in \Delta_m,\ B \in \mathbb{R}^{m\times m},\ \varepsilon > 0$

1 $P = ab^T$ `nm`
2 **for** $t = 0, \dots$ **do**
3 $\quad C \leftarrow -4APB$ `nm(n+m)`
4 $\quad K_\varepsilon \leftarrow \exp(-C/\varepsilon)$ `nm`
5 $\quad P \leftarrow \underset{P\in\Pi(a,b)}{\operatorname{argmin}} \mathrm{KL}(P, K_\varepsilon)$ $\mathcal{O}$`(nm)`
6 $\mathrm{GW} = \mathcal{Q}_{A,B}(P)$ `nm(n+m)`

**Result:** GW

**Computational complexity.** Given a cost matrix $C$, the KL projection of $K_\varepsilon$ onto the polytope $\Pi(a,b)$, where $\mathrm{KL}(P,Q) = \langle P, \log(P/Q) - 1\rangle$, is carried out in Line 5 of the inner loop of Algo. 10 using the Sinkhorn algorithm, through matrix-vector products. This quadratic complexity (in **red**) is dominated by the cost of updating matrix $C$ at each iteration in Line 3, which requires $\mathcal{O}(n^2m + nm^2)$ algebraic operations (cubic, in **violet**). As noted above, evaluating the objective $\mathcal{Q}_{A,B}(P)$ in Line 6 is also cubic.

**Step-by-step guide to reaching linearity.** We show next in §7.3 that these iterations can be sped up when the distance matrices are low-rank (or have low-rank approximations), in which case the cubic updates in $C$ and evaluation of $\mathcal{Q}_{A,B}$ in Lines 3, 6 become quadratic. Independently, we show in §7.4 that, with *no assumption* on these cost matrices, replacing the Sinkhorn call in Line 5 with a low-rank approach [3] can lower the cost of Lines 3, 6 to quadratic (while also

making Line 5 linear). Remarkably, we show in §7.5 that these two approaches can be combined in Lines 3, 6, to yield, to the best of our knowledge, the first linear time/memory algorithm able to match the performance of the Entropic-GW approach.

## 7.3 Low-rank (Approximated) Costs

**Exact factorization for distance matrices.** consider

**Assumption 2.** *$A$ and $B$ admit a low-rank factorization: there exists $A_1, A_2 \in \mathbb{R}^{n\times d}$ and $B_1, B_2 \in \mathbb{R}^{m\times d'}$ s.t. $A = A_1 A_2^T$ and $B = B_1 B_2^T$, where $d \ll n, d' \ll m$.*

A case in point is when both $A$ and $B$ are *squared* Euclidean distance matrices, with a sample size that is much larger than ambient dimension. This case is highly relevant in practice, since it covers most applications of OT to ML. Indeed, the $d \ll n$ assumption usually holds, since cases where $d \gg n$ fall in the "curse of dimensionality" regime where OT is less useful [219, 72]. Writing $X = [x_1, \dots, x_n] \in \mathbb{R}^{d\times n}$, if $A = [\|x_i - x_j\|_2^2]_{i,j}$, then one has, writing $z = (X^{\odot 2})^T \mathbf{1}_d \in \mathbb{R}^n$ that $A = z\mathbf{1}_n^T + \mathbf{1}_n z^T - 2X^T X$. Therefore by denoting $A_1 = [z, \mathbf{1}_n, -\sqrt{2}X^T] \in \mathbb{R}^{n\times(d+2)}$ and $A_2 = [\mathbf{1}_n, z, \sqrt{2}X^T] \in \mathbb{R}^{n\times(d+2)}$ we obtain the factorization above. Under Assumption 2, the complexity of Algo. 10 is reduced to $O(n^2)$: Line 3 reduces to:

$$C = -4A_1 A_2^T P B_1 B_2^T \,,$$

in $nm(d+d')+dd'(n+m)$ algebraic operations, while Line 6, using the reformulation of $\mathcal{Q}_{A,B}(P)$ in (7.3), becomes quadratic as well. Indeed, writing $G_1 := A_1^T P B_2$ and $G_2 := A_2^T P B_1$, both in $\mathbb{R}^{d\times d'}$, one has $\langle APB, P\rangle = \mathbf{1}_d^T (G_1 \odot G_2)\mathbf{1}_{d'}$. Computing $G_1, G_2$ given $P$ requires only $2(nmd + mdd')$, and computing their dot product adds $dd'$ algebraic operations. The overall complexity to compute $\mathcal{Q}_{A,B}(P)$ is $\mathcal{O}(nmd + mdd')$.

**General distance matrices.** When the original cost matrices $A, B$ are not low-rank but describe distances, we build upon recent works that output their low-rank approximation in linear time [200, 201]. These algorithms produce, for any distance matrix $A \in \mathbb{R}^{n\times m}$ and $\tau > 0$, matrices $A_1 \in \mathbb{R}^{n\times d}$, $A_2 \in \mathbb{R}^{m\times d}$ in $\mathcal{O}((m+n)\text{poly}(\frac{d}{\tau}))$ operations such that, with probability at least 0.99,

$$\|A - A_1 A_2^T\|_F^2 \leq \|A - A_d\|_F^2 + \tau\|A\|_F^2 \,,$$

where $A_d$ denotes the best rank-$d$ approximation to $A$ in the Frobenius sense. The rank $d$ should be selected to trade off approximation of $A$ and speed-ups for the method, e.g. such that $d/\tau \ll m + n$. We fall back on this approach to obtain a low-rank factorization of a distance matrix in linear time whenever needed, aware that this incurs an additional approximation (see Appendix 7.9).

**Algorithm 11** Quadratic Entropic-GW

**Input :** $A_1, A_2 \in \mathbb{R}^{n\times d}, B_1, B_2, \in \mathbb{R}^{m\times d'} a, b, \varepsilon$

$P = ab^T$ `nm`

**for** $t = 0, \dots$ **do**

  $G_2 \leftarrow A_2^T P B_1$ `nmd + mdd'`

  $C \leftarrow -4A_1 G_2 B_2^T$ `nmd' + ndd'`

  $K_\varepsilon \leftarrow \exp(-C/\varepsilon)$ `nm`

  $P \leftarrow \underset{P\in\Pi(a,b)}{\operatorname{argmin}} \mathrm{KL}(P, K_\varepsilon)$ $\mathcal{O}$(`nm`)

$c_1 \leftarrow a^T(A_1 A_2^T)^{\odot 2} a + b^T(B_1 B_2^T)^{\odot 2} b$ `n²d'+m²d'`

$G_2 \leftarrow A_2^T P B_1$ `nmd + mdd'`

$G_1 \leftarrow A_1^T P B_2$ `nmd + mdd'`

$c_2 \leftarrow -2\mathbf{1}_d^T (G_1 \odot G_2)\mathbf{1}_{d'}$ `dd'`

$\mathcal{Q}_{A,B}(P) \leftarrow c_1 + c_2$

**Return:** $\mathcal{Q}_{A,B}(P)$

## 7.4 Low-rank Constraints for Couplings

In this section, we shift our attention to a different opportunity for speed-ups, *without* Assumption 2: we consider the GW problem on couplings that are *low-rank*, in the sense that they are factorized using two low-rank couplings linked by a common marginal $g$ in $\Delta_r^*$, the *interior* of $\Delta_r$ (all entries positive). Writing the set of couplings with a nonnegative rank smaller than $r$ [3, §3.1]:

$$\Pi_{a,b}(r) := \left\{ P \in \mathbb{R}_+^{n\times m}, \exists g \in \Delta_r^* \text{ s.t. } P = Q\,\mathrm{diag}(1/g)R^T,\ Q \in \Pi_{a,g},\ \text{and } R \in \Pi_{b,g} \right\},$$

we can define the low-rank GW problem, written GW-LR$^{(r)}((a,A),(b,B))$ as the solution of

$$\min_{(Q,R,g)\in\mathcal{C}(a,b,r)} \mathcal{Q}_{A,B}(Q\,\mathrm{diag}(1/g)R^T)\,, \tag{7.5}$$

where $\mathcal{C}(a,b,r) := \mathcal{C}_1(a,b,r) \cap \mathcal{C}_2(r)$, with

$$\mathcal{C}_1(a,b,r) := \left\{ (Q,R,g) \in \mathbb{R}_+^{n\times r} \times \mathbb{R}_+^{m\times r} \times (\mathbb{R}_+^*)^r \text{ s.t. } Q\mathbf{1}_r = a, R\mathbf{1}_r = b \right\},$$
$$\mathcal{C}_2(r) := \left\{ (Q,R,g) \in \mathbb{R}_+^{n\times r} \times \mathbb{R}_+^{m\times r} \times \mathbb{R}_+^{r} \text{ s.t. } Q^T\mathbf{1}_n = R^T\mathbf{1}_m = g \right\}.$$

**Mirror Descent Scheme.** We propose to use a MD scheme with respect to the generalized KL divergence to solve (7.5). If one chooses $(Q_0, R_0, g_0) \in \mathcal{C}(a,b,r)$ an initial point such that $Q_0 > 0$ and $R_0 > 0$, this results in,

$$(Q_{k+1}, R_{k+1}, g_{k+1}) := \underset{\boldsymbol{\zeta}\in\mathcal{C}(a,b,r)}{\mathrm{argmin}} \mathrm{KL}(\boldsymbol{\zeta}, \boldsymbol{K}_k)\,, \tag{7.6}$$

where the three matrices $\boldsymbol{K}_k := (K_k^{(1)}, K_k^{(2)}, K_k^{(3)})$ are

$$\begin{aligned} K_k^{(1)} &:= \exp(4\gamma A P_k B R_k \operatorname{diag}(1/g_k) + \log(Q_k)) \\ K_k^{(2)} &:= \exp(4\gamma B P_k^T A Q_k \operatorname{diag}(1/g_k) + \log(R_k)) \\ K_k^{(3)} &:= \exp(-4\gamma \omega_k / g_k^2 + \log(g_k)) \end{aligned}$$

with $[\omega_k]_i := [Q_k^T A P_k B R_k]_{i,i}$ for all $i \in \{1, \dots, r\}$ and $\gamma > 0$ is a step size. Solving (7.6) can be done efficiently thanks to Dykstra's Algorithm as proposed in [3]. See Algo. 12 and Appendix 7.10.

**Avoiding vanishing components.** As in $k$-means optimization, the algorithm above might run into cases in which entries of the histogram $g$ vanish to 0. Following [3] we can avoid this by setting a lower bound $\alpha$ on the weight vector $g$, such that $g \geq \alpha$ coordinate-wise. Practically, we introduce truncated feasible sets $\mathcal{C}(a, b, r, \alpha) := \mathcal{C}_1(a, b, r, \alpha) \cap \mathcal{C}_2(r)$ where $\mathcal{C}_1(a, b, r, \alpha) := \mathcal{C}_1(a, b, r) \cap \{(Q, R, g) \mid g \geq \alpha\}$.

**Initialization.** To initialize our algorithm, we adapt the *first lower bound* of [86] to the low-rank setting and prove the following Proposition (see appendix 7.7 for proof).

**Proposition 7.4.1.** *Let us denote* $\tilde{x} = A^{\odot 2} a \in \mathbb{R}^n$, $\tilde{y} = B^{\odot 2} b \in \mathbb{R}^m$ *and* $\tilde{C} = (|\sqrt{\tilde{x}_i} - \sqrt{\tilde{y}_j}|^2)_{i,j} \in \mathbb{R}^{n \times m}$. *Then for all* $r \geq 1$ *we have,*

$$\text{GW-LR}_\alpha^{(r)}((a, A), (b, B)) \geq \text{LOT}_\alpha^{(r)}(\tilde{C}, a, b), \text{where}$$

$$\text{LOT}_\alpha^{(r)}(\tilde{C}, a, b) := \min_{(Q,R,g) \in \mathcal{C}(a,b,r,\alpha)} \langle \tilde{C}, Q \operatorname{diag}(1/g) R^T \rangle \ .$$

$\text{LOT}_\alpha^{(r)}(\tilde{C}, a, b)$ can be solved with [3]. The cost $\tilde{C}$ is the squared Euclidean distance between two families $\{\tilde{x}_1, \dots, \tilde{x}_n\}$ and $\{\tilde{y}_1, \dots, \tilde{y}_m\}$ in 1-D, which admits a trivial rank 2 factorization. We can therefore apply the linear-time version of their algorithm to compute the lower bound. Algo. 12 summarizes this, where $\mathcal{D}(\cdot)$ denotes the operator extracting the diagonal of a square matrix. In practice we observe that such initialization outperforms trivial or random initializations (see Section 7.6).

**Computational Cost.** Our initialization requires $\tilde{x}$ and $\tilde{y}$, obtained in $\mathcal{O}(n^2 + m^2)$ operations. Running [3, Algo.3] with a squared Euclidean distances between two families in 1-D has cost $\mathcal{O}((n + m)r)$. Solving the barycenter problem as defined in (7.6) can be done efficiently thanks to Dykstra's Algorithm. Indeed, each iteration

of [3, Algo. 2], assuming $(K_k^{(1)}, K_k^{(2)}, K_k^{(3)})$ is given, requires only $\mathcal{O}((n+m)r)$ algebraic operations. However, computing kernel matrices $(K_k^{(1)}, K_k^{(2)}, K_k^{(3)})$ at each iteration of Algorithm 12 requires a quadratic complexity with respect to the number of samples. Overall the proposed algorithm, while faster than the cubic implementation proposed in [47], still needs $\mathcal{O}((n^2+m^2)r)$ operations per iteration.

**Dykstra Iterations.** In our complexity analysis, we do not take into account the number of iterations required to terminate Dykstra's Algorithm. We show experimentally (see Fig. 7.2) that, as usually observed for Sinkhorn [76, Fig. 5], this number does not depend on problem size $n, m$, but rather on the geometric characteristics of $A, B$ and $\gamma$.

**Convergence of MD.** Although objective (7.5) is not convex in $(Q, R, g)$, we obtain the non-asymptotic stationary convergence of our proposed method. In [3], the authors study the convergence of the MD scheme when applied to the low-rank formulation of OT. In the GW setting, such strategy makes even more sense as the GW problem is a NP-hard non-convex problem and obtaining global guarantees is out of reach in a general framework. Therefore we follow the strategy proposed in [3] and consider the following convergence criterion,

$$\Delta_\alpha(\boldsymbol{\xi}, \gamma) := \frac{1}{\gamma^2}(\mathrm{KL}(\boldsymbol{\xi}, \mathcal{G}_\alpha(\boldsymbol{\xi}, \gamma)) + \mathrm{KL}(\mathcal{G}_\alpha(\boldsymbol{\xi}, \gamma), \boldsymbol{\xi}))$$

where $\mathcal{G}_\alpha(\boldsymbol{\xi}, \gamma) := \mathrm{argmin}_{\boldsymbol{\zeta} \in \mathcal{C}(a,b,r,\alpha)}\{\langle \nabla \mathcal{Q}_{A,B}(\boldsymbol{\xi}), \boldsymbol{\zeta}\rangle + \frac{1}{\gamma}\mathrm{KL}(\boldsymbol{\zeta}, \boldsymbol{\xi})\}$. This convergence criterion is in fact stronger than the one using the (generalized) projected gradient presented in [198] to obtain non-asymptotic stationary convergence of the MD scheme. Indeed the criterion used there is defined as the square norm of the following vector:

$$P_{\mathcal{C}(a,b,r,\alpha)}(\boldsymbol{\xi}, \gamma) := \frac{1}{\gamma}(\boldsymbol{\xi} - \mathcal{G}_\alpha(\boldsymbol{\xi}, \gamma)) \ ,$$

which can be seen as a generalized projected gradient of $\mathcal{Q}_{A,B}$ at $\boldsymbol{\xi}$. By denoting $X := \mathbb{R}^d$ and by replacing the *Bregman Divergence* $\mathrm{KL}(\boldsymbol{\zeta}, \boldsymbol{\xi})$ by $\frac{1}{2}\|\boldsymbol{\zeta} - \boldsymbol{\xi}\|_2^2$ in the MD scheme, we would have $P_X(\xi, \gamma) = \nabla \mathcal{Q}_{A,B}(\xi)$. Now observe that we have

$$\begin{aligned}
\Delta_\alpha(\boldsymbol{\xi}, \gamma) &= \frac{1}{\gamma^2}(\langle \nabla h(\mathcal{G}_\alpha(\boldsymbol{\xi}, \gamma)) - \nabla h(\boldsymbol{\xi}), \mathcal{G}_\alpha(\boldsymbol{\xi}, \gamma) - \boldsymbol{\xi}\rangle \\
&\geq \frac{1}{2\gamma^2}\|\mathcal{G}_\alpha(\boldsymbol{\xi}, \gamma) - \boldsymbol{\xi}\|_1^2 \\
&= \frac{1}{2}\|P_{\mathcal{C}(a,b,r,\alpha)}(\boldsymbol{\xi}, \gamma)\|_1^2
\end{aligned}$$

where $h$ denotes the minus entropy function and the last inequality comes from the strong convexity of $h$ on $\mathcal{C}(a, b, r, \alpha)$. Therefore $\Delta_\alpha(\boldsymbol{\xi}, \gamma)$ dominates $\|P_{\mathcal{C}(a,b,r,\alpha)}(\boldsymbol{\xi}, \gamma)\|_1$ and characterizes a stronger convergence.

For any $1/r \geq \alpha > 0$, Proposition 7.4.2 shows the non-asymptotic stationary convergence of the MD scheme for Problem (7.5). See Appendix 7.7 for the proof.

**Proposition 7.4.2.** *Let $\frac{1}{r} \geq \alpha > 0, N \geq 1$ and $L_\alpha := 27(\|A\|_2\|B\|_2/\alpha^4)$. Consider a constant stepsize $\gamma = \frac{1}{2L_\alpha}$ in the MD scheme (7.6). Writing $D_0 := \mathcal{Q}_{A,B}(Q_0 \operatorname{diag}(1/g_0)R_0^T) - \text{GW-LR}_\alpha^{(r)}((a, A), (b, B))$ the gap between initial value and optimum, one has*

$$\min_{1\leq k\leq N} \Delta_\alpha((Q_k, R_k, g_k), \gamma) \leq \frac{4L_\alpha D_0}{N}.$$

Since for $\alpha$ small enough, $\text{GW-LR}_\alpha^{(r)}((a, A), (b, B)) = \text{GW-LR}^{(r)}((a, A), (b, B))$, Proposition 7.4.2 shows that our algorithm reaches a stationary point of (7.5). This Proposition claims that within at most $N$ iterations the minimum of the $(\Delta_\alpha((Q_t, R_t, g_t), \gamma))_{1\leq t\leq N}$ is of order $\mathcal{O}(1/N)$. Note that this is a standard way to obtain the stationary convergence (see e.g. [198]. In practice, this is sufficient to define a stopping criteria, as one could simply compute at each iteration the criterion and keep only in memory the smallest value at each iteration.

## 7.5 Double Low-rank GW

Almost all operations in Algorithm 12 only require linear memory storage and time, except for the computations of $\tilde{x} = A^{\odot 2}a$ and $\tilde{y} = B^{\odot 2}b$ in Line 1, and the four updates involving $C_1$ and $C_2$ in Lines 6,7,14,15 which all require a quadratic number of algebraic operations. When adding Assumption 2 from §7.3 to the rank constrained approach from §7.4, we show that the strengths of both approaches can work hand in hand, both in easier initial evaluations of $\tilde{x}, \tilde{y}$, but, most importantly, at each new recomputation of a *factorized* linearization of the quadratic objective:

**Linear-time Norms in Line 1** Because $A$ admits a low-rank factorization, one can obtain a low-rank factorization for $A^{\odot 2}$ pending the condition $d^2 \ll n$. Indeed, remark that for $u, v \in \mathbb{R}^d$, $\langle u, v\rangle^2 = \langle uu^T, vv^T\rangle$. Therefore, if one describes $A_1 := [u_1; \ldots; u_n]$ and $A_2 := [v_1; \ldots; v_n]$ row-wise, and one uses the flattened out-product operator $\psi(u) := \text{vec}(uu^T) \in \mathbb{R}^{d^2}$ where $\text{vec}(\cdot)$ flattens a matrix,

$$A^{\odot 2} = \tilde{A}_1 \tilde{A}_2^T \text{ where } \tilde{A}_1 = [\psi(u_1; \ldots; \psi(u_n)],$$
$$\tilde{A}_2 = [\psi(v_1); \ldots; \psi(v_n)] .$$

**Algorithm 12** Low-Rank GW

**Input :** $A, B, a, b, r, \alpha, \gamma$

$\tilde{x} \leftarrow A^{\odot 2}a$, $\tilde{y} \leftarrow B^{\odot 2}b$ $\quad \mathtt{m}^2 + \mathtt{n}^2$

$z_1 \leftarrow \tilde{x}^{\odot 2}$, $z_2 \leftarrow \tilde{y}^{\odot 2}$ $\quad \mathtt{m} + \mathtt{n}$

$\tilde{C}_1 \leftarrow [z_1, \mathbf{1}_n, -\sqrt{2}\tilde{x}]$, $\tilde{C}_2 \leftarrow [\mathbf{1}_m, z_2, \sqrt{2}\tilde{y}]^T$ $\quad \mathtt{n} + \mathtt{m}$

$(Q, R, g) \leftarrow \mathrm{LOT}_{\alpha}^{(r)}(\tilde{C}_1\tilde{C}_2, a, b)$ $\quad \mathcal{O}((\mathtt{n} + \mathtt{m})\mathtt{r})$

**for** $t = 1, \ldots$ **do**

- $C_1 \leftarrow -AQ\,\mathrm{diag}(1/g)$ $\quad \mathcal{O}(\mathtt{n}^2\mathtt{r})$
- $C_2 \leftarrow R^T B$ $\quad \mathcal{O}(\mathtt{m}^2\mathtt{r})$
- $K^{(1)} \leftarrow Q \odot e^{4\gamma C_1 C_2 R\,\mathrm{diag}(1/g)}$ $\mathcal{O}((\mathtt{m} + \mathtt{n})\mathtt{r}^2)$
- $K^{(2)} \leftarrow R \odot e^{4\gamma C_2^T C_1^T Q\,\mathrm{diag}(1/g)}$ $\mathcal{O}((\mathtt{m} + \mathtt{n})\mathtt{r}^2)$
- $\omega \leftarrow \mathcal{D}(Q^T C_1 C_2 R)$ $\mathcal{O}(\mathtt{n}\mathtt{r}^2)$
- $K^{(3)} \leftarrow g \odot e^{-4\gamma\omega/g^2}$ $\mathcal{O}(\mathtt{r})$
- $Q, R, g \leftarrow \underset{\boldsymbol{\zeta} \in \mathcal{C}(a,b,r,\alpha)}{\mathrm{argmin}}\ \mathrm{KL}(\boldsymbol{\zeta}, \mathbf{K})$ $\mathcal{O}((\mathtt{m} + \mathtt{n})\mathtt{r})$

$c_1 \leftarrow \langle \tilde{x}, a \rangle + \langle \tilde{y}, b \rangle$ $\quad \mathtt{n} + \mathtt{m}$

$C_1 \leftarrow -AQ\,\mathrm{diag}(1/g)$ $\quad \mathcal{O}(\mathtt{n}^2\mathtt{r})$

$C_2 \leftarrow R^T B$ $\quad \mathcal{O}(\mathtt{m}^2\mathtt{r})$

$G \leftarrow C_2 R$, $G \leftarrow C_1 G$ $\quad \mathcal{O}((\mathtt{m} + \mathtt{n})\mathtt{r}^2)$

$c_2 \leftarrow -2\langle Q, G\,\mathrm{diag}(1/g) \rangle$ $\quad \mathcal{O}(\mathtt{n}\mathtt{r})$

$\mathcal{Q} \leftarrow c_1 + c_2$

**Return:** $\mathcal{Q}$

Line 1 in Algo. 12 can be replaced by $\tilde{x} \leftarrow \tilde{A}_1 \tilde{A}_2{}^T a$ and $\tilde{y} \leftarrow \tilde{B}_1 \tilde{B}_2{}^T b$. Pending the condition $d^2 \ll n, d'^2 \ll m$, this results in $nd^2 + m(d')^2$ operations. Note that Algo. 11 (line 7) can also benefit from this factorization, however as its complexity is already quadratic, the linearization of this operation has no effect on the global computational cost.

**Linearization of Lines 6,7,14,15.** The critical step in Algo. 10 that requires updating $C$ at each outer iteration is cubic. As described earlier in Algo. 12 and Algo. 11, a low-rank constraint on the coupling or a low-rank assumption on costs $A$ and $B$ reduce this cost to quadratic. Remarkably, both can be combined to yield linear time by replacing in Algo. 12, Lines 6, 7, 14, 15 by

$$C_1 \leftarrow -A_1 A_2^T Q \operatorname{diag}(1/g) \quad \text{and} \quad C_2 \leftarrow R^T B_2 B_1^T \ .$$

Note that this speed-up would not be achieved using other approaches that output a low rank approximation of the transport plan [217, 122, 6]. The crucial obstacle to using these methods here is that the cost matrix $C$ in GW changes throughout iterations, and is synthetic–the output of a matrix product $APB$ involving the very last transport $P$. This stands in stark contrast with the requirements in [217, 122, 6] that the *kernel* matrix corresponding to $K_\varepsilon = e^{-C/\varepsilon}$ admits favorable properties, such as being p.s.d or admitting an explicit (random or not) finite dimensional feature approximation.

**Linear time GW.** We have shown that (red) quadratic operations appearing in Algo. (12) can be replaced by linear alternatives. The iterations that have not been modified had an overall complexity of $\mathcal{O}(mr(r+d') + nr(r+d))$. The initialization and linearization steps can now be performed in linear time and complexity, respectively in $\mathcal{O}(n(r+d^2) + m((d')^2 + r))$ and $\mathcal{O}((nr(r+d) + mr(r+d'))$.

## 7.6 Experiments

Our goal in this section is to provide practical guidance on how to use our method (to set stepsize $\gamma$, lower bound $\alpha$ on entries of $g$ and rank $r$) and compare its practical performance with other baselines, both in terms of running times and relevance, on 5 simulated datasets and 2 real world applications. We consider our quadratic approach **LR** (Algo. 12) and its linear time counterpart **Lin LR** (§7.5). We compare them with **Ent**, the cubic implementation of [47], and its improved quadratic version **Quad Ent** introduced in this paper (Algo. 11). We also use **MREC** as implemented in [62]. Because all these approaches admit different hyperparameters, we evaluate them by stressing GW loss as a function

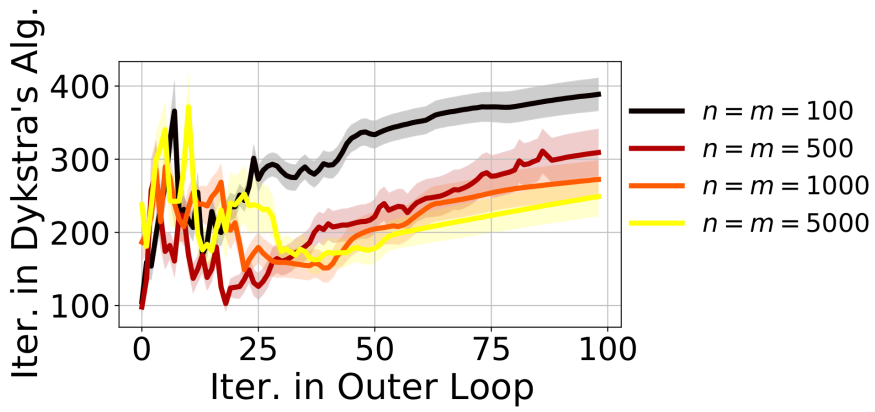



Figure 7.2: We consider samples of a mixture of 10 anisotropic Gaussians in resp. 10 and 15-D endowed with the squared Eucl. metric. The number of iterations of Dykstra's algorithm required to reach a precision of $\delta = 1e-3$ along the iterations of the Algo. 12 is not impacted significantly by varying $n$, the sample size.

of computational effort, as well as performance in downstream metrics. Because the couplings obtained by **MREC** do *not* satisfy marginal constraints, computing its GW loss is irrelevant, but its matching can be used in the single cell genomics experiments we consider. Experiments were run on a MacBook Pro 2019 laptop, and data from `github.com/rsinghlab/SCOT`. The code is available at https://github.com/meyerscetbon/LinearGromov.

**Initialization.** For a fair comparison with the entropic approach, we adapt the *first lower bound* of [86, Def. 6.1] to the entropic case to initialize it. In all experiments displaying time-accuracy tradeoffs, we report computation budget as number of operations. Accuracy is measured by evaluating the ground-truth energy $\mathcal{Q}_{A,B}$ (even in scenarios when the method uses a low rank approximation for $A, B$ at optimization time). We repeat all experiments 10 times on random resampling of the measures in all synthetic problems, to obtain error bars.

**On the iterations of Dykstra's Algorithm.** In this experiment, we show that the number of iterations involved in the Dykstra's Algorithm does not depends on $n$ the number of samples when applying Algo. 12. In Fig. 7.2, we consider samples of mixtures of (10 and 15) anisotropic Gaussians in resp. 10 and 15-D and report the number of iterations of the Dykstra's Algorithm required to reach a precision $\delta = 1e-3$ along the iterations of Algo. 12. We observe that the number of iterations in Dykstra does not depend on $n$ the number of samples considered. Note that for all the sample sizes considered, we need far fewer iterations (usually

$\leq 25$) for the outer loop to converge: the plots show a larger $x$-axis than what is observed in practice.

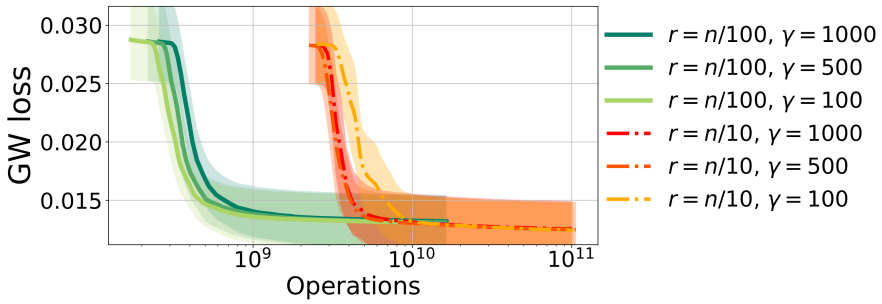
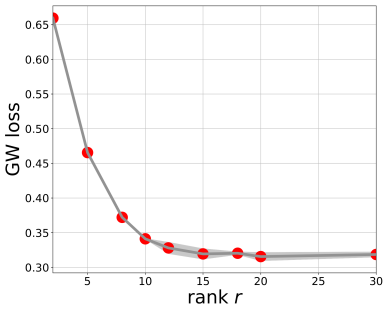



Figure 7.3: We consider two $n = m = 1000$ samples of mixtures of (2 and 3) Gaussians in resp. 5 and 10-D, endowed with the squared Euclidean metric, compared with **Lin LR**. The time/loss tradeoff illustrated in these plots show that our method is only mildly impacted by step size $\gamma$ for both ranks $r = n/100$ and $n/10$.

**Sensitity to $\gamma$ and $\alpha$.** We study how optimization parameters $\gamma$ and $\alpha$ impact results. We consider $n = m = 1000$ samples drawn from two mixtures of (2 and 3) anisotropic Gaussians in respectively 5-D and 10-D (details in Appendix 7.11.2). Fig. 7.3, reports the time vs. GW loss tradeoff of our method when varying $\gamma$, both for $r = n/100$ or $n/10$ illustrating its robustness to that choice. Fig. 7.11 in Appendix 7.11.2 shows similar conclusions with respect to $\alpha$. Recall that $\alpha$ was only used to lower bound the weights of barycenter $g$, to ensure no collapse. In all other experiments, we always set $\gamma = 100$ and $\alpha = 10^{-10}$ for our methods, and only focus on rank $r$.

**Effect of the rank.** We study the impact of rank $r$ on our method. We consider samples from two Gaussian mixtures, with respectively 10 and 20 centers in 10-D and 15-D and $n = m = 5000$. We compute the GW cost obtained by **Lin LR** in the squared Euclidean setting as a function of $r$ the rank. We observe that the loss decreases as the rank increases until the rank $r$ reaches 20 (the largest number of clusters in our mixtures). Therefore, our method is able to capture the clustered structure of data (See Appendix 7.11.3). In practice $r$ should be selected such that it corresponds to the number of clusters in the data.



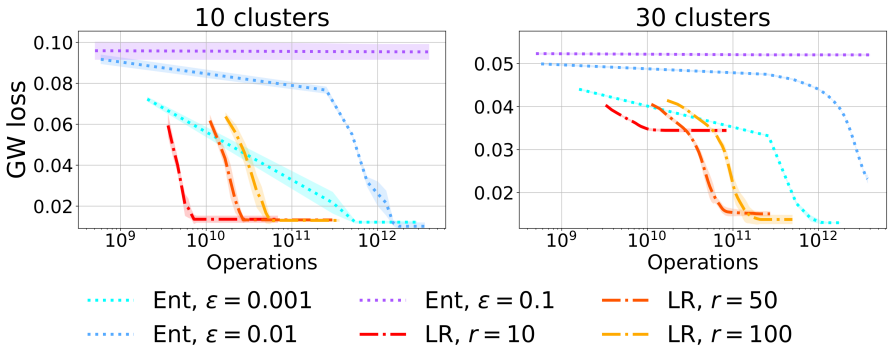



Figure 7.4: We sample $n = m = 5000$ points from two anisotropic Gaussian blobs, respectively in 10 and 15-D, with either 10 or 30 clusters, endowed with the Euclidean distance. We compare our quadratic method **LR** with the cubic Entropic GW **Ent**, which requires instantiating matrices $A$ and $B$. We vary both $r$ (our method) and $\varepsilon$ (entropic). Our method obtains similar GW loss, while being orders of magnitude faster. Note the gap in performance between $r = 10$ and $r = 50$ when the input measures have 30 clusters: the GW loss decreases as the rank $r$ increases until it reaches the number of clusters in the data.

**Synthetic low-rank problem.** We consider two anisotropic Gaussian blobs with the same number of blobs in respectively 10-D and 15-D. We constrain the distance between the centroids of the clusters to be larger than the dimension (see Appendix 7.11.4 for illustrations). In Figures 7.4 and 7.6, when the underlying cost is the (*not* squared) Euclidean distance, our methods manage to consistently obtain similar GW loss that those obtained by entropic methods, using very low rank $r = n/100$, while being orders of magnitude faster. Fig. 7.7 explores the more favorable case where the underlying cost is the *squared* Euclidean distance, reaching similar conclusions.

**Large scale experiment.** In this experiment, we show that our method is able to compute an approximation of the GW cost in the large sample setting. In Fig. 7.5, we samples $n = m = 1e5$ samples from the unit square in 2-D and we compare the time/loss tradeoff when varying the rank $r$. We show that our method is the only one able to approach the GW cost in such regimes.

**Experiments on Single Cell Genomics Data.** We reproduce the single-cell alignment experiments introduced in [61]. The datasets consist in single-cell multi-omics data generated by co-assays, provided with a ground truth one–to-one correspondence, which can be used to benchmark GW strategies. The SNAREseq dataset [220], with $n = m = 1047$ points in $\mathbb{R}^{19}$, describes a real-world experiment;

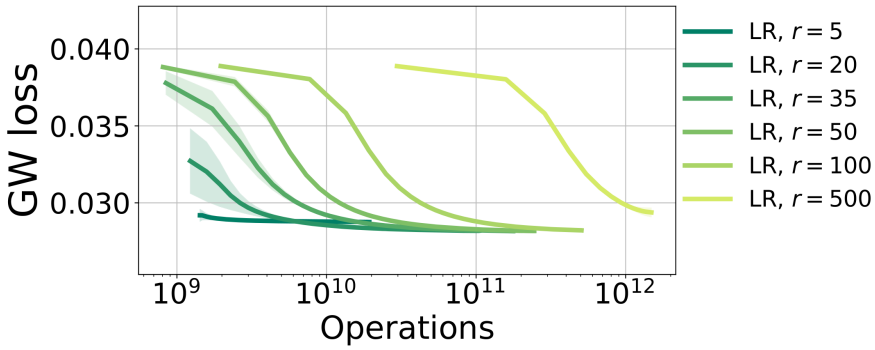



Figure 7.5: We sample $n = 1e5$ points from the unit square in 2-D. The underlying cost considered is the squared Euclidean cost. In this regime, only **Lin GW-LR** can be computed. We plot the time/loss tradeoff when varying $r$.

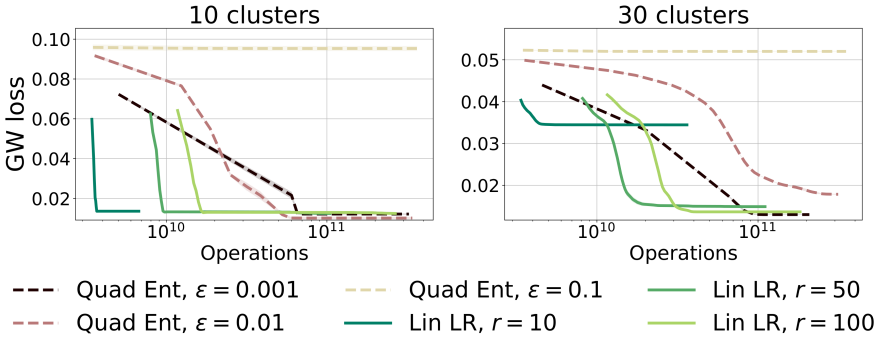



Figure 7.6: Same setting as Fig. 7.4, using a low-rank approximation of the Euclidean distance (see §7.3) to introduce our linear method **Lin LR** and compare it with **Quad Ent**. The rank of their factorizations is set to $d = d' = 100$. We vary $\varepsilon$ and rank $r$ to reach similar conclusions to those outlined in Fig. 7.4. Note also that both **Lin LR** and **Quad Ent** reach similar GW loss as those obtained by their full-rank counterparts, while being orders of magnitude faster.

the Splatter dataset [221] with $n = m = 5000$ points in $\mathbb{R}^{500}$ is synthetic. We use the pre-processing from [61] to prepare intra-domain distance matrices $A$ and $B$ using a k-NN graph based on correlations, to compute shortest path distances. Note that in that case, one cannot obtain directly in linear time a low-rank factorization of $A$ and $B$ using [200, 201], since the shortest path distances need to be computed first. Therefore, we only use our quadratic approach **LR** and the cubic implementation of the entropic method **Ent**, along with **MREC**. In Fig. 7.8 we compare both the time/GW loss tradeoffs and the alignment performances through the "fraction of samples closer than the true match" (FOSCTTM) error introduced in [222]. Note that we cannot compare the time-accuracy tradeoff of **MREC** with our method

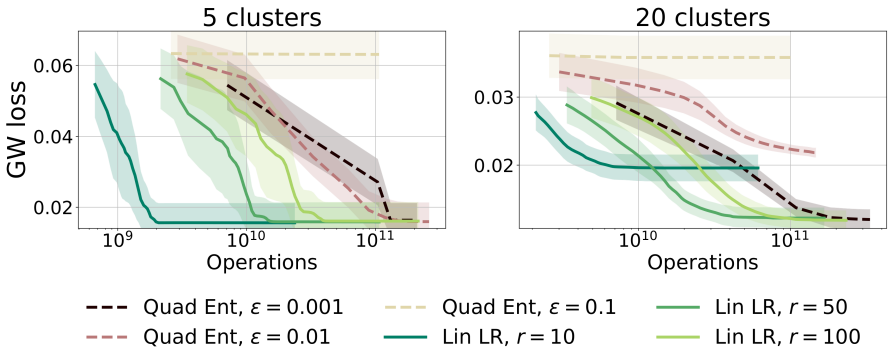



Figure 7.7: Setting as in Fig. 7.4, with $n = m = 10000$ samples from anisotropic Gaussian blobs of 5 or 20 clusters, endowed with the squared Eucl. distance. We compare **Lin LR** and **Quad Ent** using exact factorizations of $A$ and $B$.

as the coupling obtained does not satisfy the marginal constraints. **LR** reaches similar loss, while being orders of magnitude faster than **Ent**, even for a very small rank $r = n/100$.

**Experiment on BRAIN.** We reproduce the experiment proposed in [62]. We consider the dataset introduced in [223] of single cells sampled from the human brain with eight different cell labels. The dataset contains two groups with different representations: one contains $n = 34079$ cells represented by their genes expressions, while the second contains $m = 27906$ cells represented by their DNA region accessibilities. We reuse the preprocessing in [62], by applying the method proposed in [224] and available in Scanpy [225] to the first group and a TF-IDF representation to the second one. A PCA is then performed on each group to reduce dimensions to 50, endowed with the squared Euclidean distance. These datasets are too large to be handled with entropic approaches, and show the potential of our linear approach **Lin LR** to handle larger scale problems. To compare **Lin LR** with **MREC**, we measure the accuracy of their matchings, as proposed in [62], by computing the fraction of points in the first group whose associated points under the matching given by the method share the same label in the second group. In Figure 7.9, we plot the accuracy against the rank (or the number of clusters in MREC) for both **Lin LR** and **MREC**. We also consider multiple versions of **MREC** by varying its entropic regularization parameter, $\varepsilon$, involved in the inner matching of the recursive method. Our method obtains consistently better accuracy than that obtained by **MREC**.

**Discussion.** While the factorization introduced in [3] held the promise to speed up classic OT, we show in this work that it delivers an even larger impact when

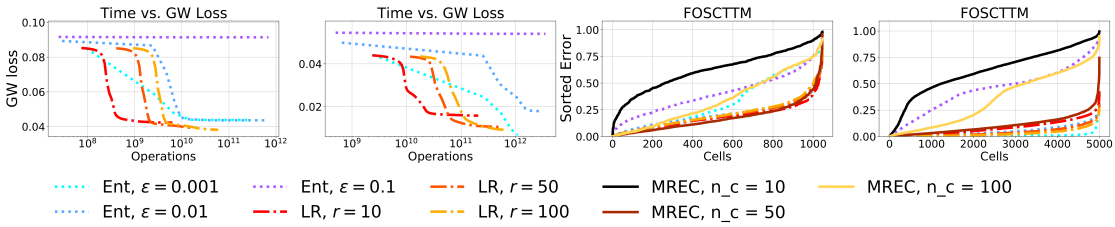



Figure 7.8: We consider both the SNAREseq dataset (*left, middle-right*) which consists in two point clouds of $n = m = 1047$ samples in respectively 10-D and 19-D and the Splatter dataset (*middle-left, right*) composed of two point clouds of $n = m = 5000$ samples in respectively 50-D and 500-D. The cost considered is the shortest-path distance of a $k - NN$ graph. We compare both the time-accuracy tradeoffs of our method with the Entropic-GW (*left, middle-left*) and the FOSCTTMs ranked in the increasing order of **LR**, **Ent** and **MREC** when varying their hyperparameters (*middle-right, right*). Because the coupling returned by **MREC** does not satisfy marginal constraints, we do not include it in left plots. Our method reaches similar accuracy while being order of magnitude faster than **Ent** even for a small rank $r = n/100$. We notice that the alignments obtained by our method are robust to the choice of $r$, with similar performance for all methods.

applied to GW: indeed, the combination of low-rank Sinkhorn factorization with-low rank cost matrices is the only one, to our knowledge, that achieves linear time/memory complexity for the Gromov-Wasserstein problem. The GW problem is NP-hard, its optimal solution out of reach and approximate solutions can only be reached using an inductive bias. Here we propose to compute *efficiently* a coupling whose GW cost is low. By adding low-rank constraints, our goal is no longer to approach the optimal coupling, but rather to promote low-rank solutions among many that have a low GW cost. Our low-rank constraint obtains similar performance as the entropic regularization, the current default approach, while being much faster to compute. We show in experiments that low-rank couplings can reach low GW costs, and that they are directly useful in real-world tasks. Our approach has, however, a few limitations compared to the entropic one: setting $\gamma$, while not problematic in most of our experiments, could require a bit of tuning in order to obtain faster runs in challenging situations. Our assumptions to reach linearity, as discussed in §7.4 and 7.5 mostly rests on two important assumptions: the rank of distance matrices (the intrisic dimensionality of data points) must be such that $d, d'$ are dominated by $n, m$ and that a small enough rank $r$ be able to capture the configuration of the input measures. Pending these constraints, which are valid in most relevant experimental setups we know of, we have demonstrated that our approach is versatile, remains faithful to the original GW formulation, and scales to sizes that are out of reach for the SoTA entropic solver.

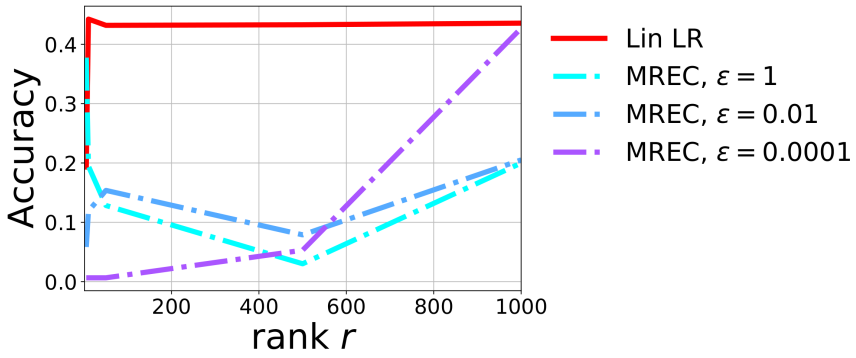



Figure 7.9: Using the BRAIN dataset (two point clouds of $n = 34079$ and $m = 27906$ samples in 50-D, endowed with squared Euclidean distance) we compare the GW loss against the rank (or the number of clusters) for both **Lin LR** and **MREC** for multiple choices of $\varepsilon$ in **MREC**. We show that our method is robust to the choice of the rank and obtains consistently better accuracy than **MREC**.

# Supplementary material

## 7.7 Proofs

### 7.7.1 Proof of Proposition 7.4.1

*Proof.* Let $(Q, R, g) \in \mathcal{C}(a, b, r, \alpha)$, $P := Q\,\mathrm{diag}(1/g)R^T$. Remarks that for all $i, j$,

$$\begin{aligned}\sqrt{\sum_{i',j'} |A_{i,i'} - B_{j,j'}|^2 P_{i',j'}} &\geq \left| \sqrt{\sum_{i',j'} |A_{i,i'}|^2 P_{i',j'}} - \sqrt{\sum_{i',j'} |B_{j,j'}|^2 P_{i',j'}} \right| \\ &\geq |\sqrt{\tilde{x}_i} - \sqrt{\tilde{y}_j}|\end{aligned}$$

Therefore we have

$$\begin{aligned}\sqrt{\sum_{i,i',j,j'} |A_{i,i'} - B_{j,j'}|^2 P_{i',j'} P_{i,j}} &= \sqrt{\sum_{i,j} \sum_{i',j'} |A_{i,i'} - B_{j,j'}|^2 P_{i',j'} P_{i,j}} \\ &\geq \sqrt{\sum_{i,j} |\sqrt{\tilde{x}_i} - \sqrt{\tilde{y}_j}|^2 P_{i,j}}\end{aligned}$$

Finally we obtain that

$$\sum_{i,i',j,j'} |A_{i,i'} - B_{j,j'}|^2 P_{i',j'} P_{i,j} - \varepsilon H(Q, R, g) \geq \sum_{i,j} |\sqrt{\tilde{x}_i} - \sqrt{\tilde{y}_j}|^2 P_{i,j} - \varepsilon H(Q, R, g)$$

and by taking the infimum over all $(Q, R, g) \in \mathcal{C}(a, b, r, \alpha)$, the results follows. □

### 7.7.2 Proof of Proposition 7.4.2

To show the result, we first need to recall some notions linked to the relative smoothness. Let $\mathcal{X}$ a closed convex subset in a Euclidean space $\mathbb{R}^q$. Given a convex function $H : \mathcal{X} \to \mathbb{R}$ continuously differentiable, one can define the *Bregman divergence* associated to $H$ as

$$D_H(x, z) := H(x) - H(z) - \langle \nabla H(z), x - z \rangle.$$

Let us now introduce the definition of the relative smoothness with respect the $H$.

**Definition 7.7.1** (Relative smoothness.)**.** *Let $L > 0$ and $f$ continuously differentiable on $\mathcal{X}$. $f$ is said to be $L$-smooth relatively to $H$ if*

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + L D_H(y, x)$$

In [3], the authors show the following general result on the non-asymptotic stationary convergence of the mirror-descent scheme defined by the following recursion:

$$x_{k+1} = \operatorname*{argmin}_{x\in\mathcal{X}} \langle \nabla f(x_k), x\rangle + \frac{1}{\gamma_k} D_h(x, x_k)$$

where $(\gamma_k)$ a sequence of positive step-size.

**Proposition 7.7.1** ([3]). *Let $N \geq 1$, $f$ continuously differentiable on $\mathcal{X}$ which is $L$-smooth relatively to $H$. By considering for all $k = 1, \dots, N$, $\gamma_k = 1/2L$, and by denoting $D_0 = f(x_0) - \min_{x\in\mathcal{X}} f(x)$, we have*

$$\min_{0\leq k\leq N-1} \Delta_k \leq \frac{4LD_0}{N}.$$

*where for all $k = 1, \dots, N$*

$$\Delta_k := \frac{1}{\gamma_k^2}(D_H(x_k, x_{k+1}) + D_H(x_{k+1}, x_k)).$$

Let us now show that our objective function is relatively smooth with respect the the KL divergence [203, 204]. The result of Propostion 7.4.2 will then follow from Proposition 7.7.1. Here $\mathcal{X} = \mathcal{C}(a, b, r, \alpha)$, $H$ is the negative entropy defined as

$$H(Q, R, g) := \sum_{i,j} Q_{i,j}(\log(Q_{i,j}) - 1) + \sum_{i,j} R_{i,j}(\log(R_{i,j}) - 1) + \sum_j g_j(\log(g_j) - 1),$$

and let us define for all $(Q, R, g) \in \mathcal{C}(a, b, r, \alpha)$

$$F_\varepsilon(Q, R, g) := -2\langle AQ\,\mathrm{diag}(1/g)R^T B, Q\,\mathrm{diag}(1/g)R^T\rangle + \varepsilon H(Q, R, g)\ .$$

Let us now show the following proposition.

**Proposition 7.7.2.** *Let $\varepsilon \geq 0$, $\frac{1}{r} \geq \alpha > 0$ and let us denote $L_{\varepsilon,\alpha} := 27(\|A\|_2\|B\|_2/\alpha^4 + \varepsilon)$. Then for all $(Q_1, R_1, g_1), (Q_2, R_2, g_2) \in \mathcal{C}(a, b, r, \alpha)$, we have*

$$\|\nabla F_\varepsilon(Q_1, R_1, g_1) - \nabla F_\varepsilon(Q_2, R_2, g_2)\|_2 \leq L_{\varepsilon,\alpha}\|H(Q_1, R_1, g_1) - H(Q_2, R_2, g_2)\|_2$$

*Proof.* Let $(Q, R, g) \in \mathcal{C}(a, b, r, \alpha)$ and let us denote $P = Q\,\mathrm{diag}(1/g)R^T$. We first have that

$$\nabla F_\varepsilon(Q, R, g) = (\nabla_Q F_\varepsilon(Q, R, g), \nabla_R F_\varepsilon(Q, R, g), \nabla_g F_\varepsilon(Q, R, g))$$

where

$$\begin{aligned}\nabla_Q F_\varepsilon(Q,R,g) &:= -4APBR\,\mathrm{diag}(1/g) + \varepsilon \log Q\\ \nabla_R F_\varepsilon(Q,R,g) &:= -4BP^TAQ\,\mathrm{diag}(1/g) + \varepsilon \log R\\ \nabla_g F_\varepsilon(Q,R,g) &:= -4\mathcal{D}(Q^TAPBR)/g^2 + \varepsilon \log g\end{aligned}$$

First remarks that

$$\begin{aligned}\|\nabla_Q F_\varepsilon(Q_1,R_1,g_1) - \nabla_Q F_\varepsilon(Q_2,R_2,g_2)\|_2 \le\ & 4\|AP_1BR_1\,\mathrm{diag}(1/g_1) - AP_2BR_2\,\mathrm{diag}(1/g_2)\|_2\\ & + \varepsilon\|\log Q_1 - \log Q_2\|_2\ .\end{aligned}$$

Moreover we have

$$\begin{aligned}AP_1BR_1\,\mathrm{diag}(1/g_1) - AP_2BR_2\,\mathrm{diag}(1/g_2) = A((P_1-P_2)BR_1\,\mathrm{diag}(1/g_1)\\ + P_2B(R_1\,\mathrm{diag}(1/g_1) - R_2\,\mathrm{diag}(1/g_2))\end{aligned}$$

where

$$P_1 - P_2 = (Q_1-Q_2)\,\mathrm{diag}(1/g_1)R_1^T + Q_2(\mathrm{diag}(1/g_1)R_1^T - \mathrm{diag}(1/g_2)R_2^T)$$

and

$$R_1\,\mathrm{diag}(1/g_1) - R_2\,\mathrm{diag}(1/g_2) = (R_1-R_2)\,\mathrm{diag}(1/g_1) + R_2(\mathrm{diag}(1/g_1) - \mathrm{diag}(1/g_2))\ .$$

Moreover we have

$$\begin{aligned}\|AP_1BR_1\,\mathrm{diag}(1/g_1) - AP_2BR_2\,\mathrm{diag}(1/g_2)\| \le & \|A\|\|B\|\|P_1-P_2\|/\alpha\\ & + \|A\|\|B\|\|R_1\,\mathrm{diag}(1/g_1) - R_2\,\mathrm{diag}(1/g_2)\|\end{aligned}$$

then remark that

$$\|P_1-P_2\| \le \|Q_1-Q_2\|/\alpha + \|R_1\,\mathrm{diag}(1/g_1) - R_2\,\mathrm{diag}(1/g_2)\|$$

and

$$\|R_1\,\mathrm{diag}(1/g_1) - R_2\,\mathrm{diag}(1/g_2)\| \le \|R_1-R_2\|/\alpha + \|1/g_1 - 1/g_2\|$$

from which follows that

$$\begin{aligned}&\|AP_1BR_1\,\mathrm{diag}(1/g_1) - AP_2BR_2\,\mathrm{diag}(1/g_2)\|\\ &\le \frac{\|A\|\|B\|}{\alpha}\left(\frac{\|Q_1-Q_2\|}{\alpha} + \frac{\|R_1-R_2\|}{\alpha} + \|1/g_1 - 1/g_2\|\right)\\ &+ \|A\|\|B\|\left(\frac{\|R_1-R_2\|}{\alpha} + \|1/g_1 - 1/g_2\|\right)\ .\end{aligned}$$

As $Q \to H(Q)$ is 1-strongly convex w.r.t to the $\ell_2$-norm on $\Delta_{n\times r}$, we have

$$\begin{aligned}\|Q_1 - Q_2\|_2^2 &\leq \langle \log Q_1 - \log Q_2, Q_1 - Q_2\rangle \\ &\leq \|\log Q_1 - \log Q_2\|_2 \|Q_1 - Q_2\|_2\end{aligned}$$

from which follows that

$$\|Q_1 - Q_2\|_2 \leq \|\log Q_1 - \log Q_2\|_2.$$

Moreover we have

$$\|1/g_1 - 1/g_2\|_2 \leq \frac{\|g_1 - g_2\|_2}{\alpha^2} \leq \frac{\|\log g_1 - \log g_2\|_2}{\alpha^2}$$

Then we obtain that

$$\begin{aligned}\|\nabla_Q F_\varepsilon(Q_1, R_1, g_1) - \nabla_Q F_\varepsilon(Q_2, R_2, g_2)\|_2 \leq &\left(\frac{4\|A\|\|B\|}{\alpha^2} + \varepsilon\right)\|\log Q_1 - \log Q_2\|_2 \\ &+ (1 + 1/\alpha)\frac{4\|A\|\|B\|}{\alpha}\|\log R_1 - \log R_2\|_2 \\ &(1 + 1/\alpha)\frac{4\|A\|\|B\|}{\alpha^2}\|\log g_1 - \log g_2\|_2\end{aligned}$$

Similarly we obtain that Then we obtain that

$$\begin{aligned}\|\nabla_R F_\varepsilon(Q_1, R_1, g_1) - \nabla_R F_\varepsilon(Q_2, R_2, g_2)\|_2 \leq &\left(\frac{4\|A\|\|B\|}{\alpha^2} + \varepsilon\right)\|\log R_1 - \log R_2\|_2 \\ &+ (1 + 1/\alpha)\frac{4\|A\|\|B\|}{\alpha}\|\log Q_1 - \log Q_2\|_2 \\ &(1 + 1/\alpha)\frac{4\|A\|\|B\|}{\alpha^2}\|\log g_1 - \log g_2\|_2\end{aligned}$$

Moreover we have

$$\begin{aligned}\|\nabla_g F_\varepsilon(Q_1, R_1, g_1) - \nabla_g F_\varepsilon(Q_2, R_2, g_2)\|_2 \leq &4\|\mathcal{D}(Q_1^T A P_1 B R_1)/g_1^2 - \mathcal{D}(Q_2^T A P_2 B R_2)/g_2^2\| \\ &+ \varepsilon\|\log g_1 - \log g_2\|\end{aligned}$$

and

$$\begin{aligned}\mathcal{D}(Q_1^T A P_1 B R_1)/g_1^2 - \mathcal{D}(Q_2^T A P_2 B R_2)/g_2^2 = &(1/g_1^2 - 1/g_2^2)\mathcal{D}(Q_1^T A P_1 B R_1) \\ &+ \frac{1}{g_2^2}(\mathcal{D}(Q_1^T A P_1 B R_1) - \mathcal{D}(Q_2^T A P_2 B R_2));.\end{aligned}$$

Note also that

$$\|(1/g_1^2 - 1/g_2^2)\mathcal{D}(Q_1^T A P_1 B R_1)\| \leq \frac{2\|A\|\|B\|}{\alpha^3}\|\log g_1 - \log g_2\|$$

and

$$Q_1^T AP_1BR_1 - Q_2^T AP_2BR_2 = (Q_1^T - Q_2^T)AP_1BR_1 + Q_2^T A(P_1BR_1 - P_2BR_2)$$
$$= (Q_1^T - Q_2^T)AP_1BR_1 + Q_2^T A((P_1 - P_2)BR_1 + P_2B(R_1 - R_2))$$

from which follows that

$$\|\frac{1}{g_2^2}(\mathcal{D}(Q_1^T AP_1BR_1) - \mathcal{D}(Q_2^T AP_2BR_2))\| \leq \frac{\|A\|\|B\|}{\alpha^2}\left(\|\log Q_1 - \log Q_2\| + \|\log R_1 - \log R_2\|\right)$$
$$+ \frac{\|A\|\|B\|}{\alpha^2}\|P_1 - P_2\|$$

and we obtain that

$$\|\nabla_g F_\varepsilon(Q_1, R_1, g_1) - \nabla_g F_\varepsilon(Q_2, R_2, g_2)\|_2 \leq \left(\frac{4\|A\|\|B\|}{\alpha^2} + \frac{1}{\alpha}\right)\|\log Q_1 - \log Q_2\|$$
$$+ \left(\frac{4\|A\|\|B\|}{\alpha^2} + \frac{1}{\alpha}\right)\|\log R_1 - \log R_2\|$$
$$+ \left(\frac{4\|A\|\|B\|}{\alpha^4} + \frac{8\|A\|\|B\|}{\alpha^3} + \varepsilon\right)\|\log g_1 - \log g_2\|$$

Finally we have

$$\|\nabla F_\varepsilon(Q_1, R_1, g_1) - \nabla F_\varepsilon(Q_2, R_2, g_2)\|_2^2$$
$$\leq 3\left[\left(\frac{4\|A\|\|B\|}{\alpha^2} + \varepsilon\right)^2 + (1 + 1/\alpha)^2\frac{16\|A\|^2\|B\|^2}{\alpha^2} + \left(\frac{4\|A\|\|B\|}{\alpha^2} + \frac{1}{\alpha}\right)^2\right]$$
$$\times \left(\|\log Q_1 - \log Q_2\|^2 + \|\log R_1 - \log R_2\|^2\right)$$
$$+ 3\left[2(1 + 1/\alpha)^2\frac{16\|A\|\|^2B\|^2}{\alpha^4} + \left(\frac{4\|A\|\|B\|}{\alpha^4} + \frac{8\|A\|\|B\|}{\alpha^3} + \varepsilon\right)^2\right]\|\log g_1 - \log g_2\|^2$$

from which we obtain that

$$\|\nabla F_\varepsilon(Q_1, R_1, g_1) - \nabla F_\varepsilon(Q_2, R_2, g_2)\|_2^2 \leq L_{\varepsilon,\alpha}^2\left(\|\log Q_1 - \log Q_2\|^2 + \|\log R_1 - \log R_2\|^2\right)$$
$$+ L_{\varepsilon,\alpha}^2\|\log g_1 - \log g_2\|^2$$

and the result follows. □

## 7.8 Double Regularization Scheme

Another way to stabilize the method is by considering a double regularization scheme as proposed in [3] where in addition of constraining the nonnegative rank

of the coupling, we regularize the objective by adding an entropic term in $(Q, R, g)$, which is to be understood as that of the values of the three respective entropies evaluated for each term.

$$\text{GW-LR}^{(r)}_{\varepsilon,\alpha}((a,A),(b,B)) := \min_{(Q,R,g)\in\mathcal{C}(a,b,r,\alpha)} \mathcal{E}_{A,B}(Q\,\text{diag}(1/g)R^T) - \varepsilon H((Q,R,g))\,. \tag{7.7}$$

**Mirror Descent Scheme.** We propose to use a MD scheme with respect to the KL divergence to approximate GW-LR$^{(r)}_{\varepsilon,\alpha}$ defined in (7.7). More precisely, for any $\varepsilon \geq 0$, the MD scheme leads for all $k \geq 0$ to the following updates which require solving a convex barycenter problem per step:

$$(Q_{k+1}, R_{k+1}, g_{k+1}) := \operatorname*{argmin}_{\boldsymbol{\zeta}\in\mathcal{C}(a,b,r,\alpha)} \text{KL}(\boldsymbol{\zeta}, \boldsymbol{K}_k) \tag{7.8}$$

where $(Q_0, R_0, g_0) \in \mathcal{C}(a,b,r)$ is an initial point such that $Q_0 > 0$ and $R_0 > 0$, $P_k := Q_k\,\text{diag}(1/g_k)R_k^T$, $\boldsymbol{K}_k := (K_k^{(1)}, K_k^{(2)}, K_k^{(3)})$, $K_k^{(1)} := \exp(4\gamma A P_k B R_k\,\text{diag}(1/g_k) - (\gamma\varepsilon - 1)\log(Q_k))$,
$K_k^{(2)} := \exp(4\gamma B P_k^T D Q_k\,\text{diag}(1/g_k) - (\gamma\varepsilon - 1)\log(R_k))$, $K_k^{(3)} := \exp(-4\gamma\omega_k/g_k^2 - (\gamma\varepsilon - 1)\log(g_k))$ with $[\omega_k]_i := [Q_k^T A P_k B R_k]_{i,i}$ for all $i \in \{1,\dots,r\}$ and $\gamma$ is a positive step size. Solving (7.6) can be done efficiently thanks to the Dykstra's Algorithm as showed in [3]. See Appendix 7.10 for more details.

**Convergence of the mirror descent.** Even if the objective (7.7) is not convex in $(Q, R, g)$, we obtain the non-asymptotic stationary convergence of the MD algorithm in this setting. For that purpose we consider the same convergence criterion as the one proposed in [3] to obtain non-asymptotic stationary convergence of the MD scheme defined as

$$\Delta_{\varepsilon,\alpha}(\boldsymbol{\xi},\gamma) := \frac{1}{\gamma^2}(\text{KL}(\boldsymbol{\xi}, \mathcal{G}_{\varepsilon,\alpha}(\boldsymbol{\xi},\gamma)) + \text{KL}(\mathcal{G}_{\varepsilon,\alpha}(\boldsymbol{\xi},\gamma), \boldsymbol{\xi}))$$

where $\mathcal{G}_{\varepsilon,\alpha}(\boldsymbol{\xi},\gamma) := \operatorname{argmin}_{\boldsymbol{\zeta}\in\mathcal{C}(a,b,r,\alpha)}\{\langle\nabla\mathcal{E}_{A,B}(\boldsymbol{\xi}), \boldsymbol{\zeta}\rangle + \frac{1}{\gamma}\text{KL}(\boldsymbol{\zeta}, \boldsymbol{\xi})\}$. For any $1/r \geq \alpha > 0$, we show in the following proposition the non-asymptotic stationary convergence of the MD scheme applied to the problem (7.7). See Appendix 7.7 for the proof.

**Proposition 7.8.1.** *Let $\varepsilon \geq 0$, $\frac{1}{r} \geq \alpha > 0$ and $N \geq 1$. By denoting $L_{\varepsilon,\alpha} := 27(\|A\|_2\|B\|_2/\alpha^4 + \varepsilon)$ and by considering a constant stepsize in the MD scheme (7.6) $\gamma = \frac{1}{2L_{\varepsilon,\alpha}}$, we obtain that*

$$\min_{1\leq k\leq N} \Delta_{\varepsilon,\alpha}((Q_k, R_k, g_k), \gamma) \leq \frac{4L_{\varepsilon,\alpha}D_0}{N}.$$

*where $D_0 := \mathcal{E}_{A,B}(Q_0\,\text{diag}(1/g_0 R_0^T) - \text{GW-LR}^{(r)}((a,A),(b,B))$ is the distance of the initial value to the optimal one.*

## 7.9 Low-rank Approximation of Distance Matrices

Here we recall the algorithm used to perform a low-rank approximation of a distance matrix [200, 201]. We use the implementation of [3].

**Algorithm 13** LR-Distance$(X, Y, r, \gamma)$ [200, 201]

**Inputs:** $X, Y, r, \gamma$

Choose $i^* \in \{1, \dots, n\}$, and $j^*\{1, \dots, m\}$ uniformly at random.

For $i = 1, \dots, n$, $p_i \leftarrow d(x_i, y_j^*)^2 + d(x_i^*, y_j^*)^2 + \frac{1}{m}\sum_{j=1}^m d(x_i^*, y_j)^2$.

Independently choose $i^{(1)}, \dots, i^{(t)}$ according $(p_1, \dots, p_n)$.

$X^{(t)} \leftarrow [x_{i^{(1)}}, \dots, x_{i^{(t)}}]$, $P^{(t)} \leftarrow [\sqrt{tp_{i^{(1)}}}, \dots, \sqrt{tp_{i^{(t)}}}]$, $S \leftarrow d(X^{(t)}, Y)/P^{(t)}$

Denote $S = [S^{(1)}, \dots, S^{(m)}]$,

For $j = 1, \dots, m$, $q_j \leftarrow \|S^{(j)}\|_2^2/\|S\|_F^2$

Independently choose $j^{(1)}, \dots, j^{(t)}$ according $(q_1, \dots, q_m)$.

$S^{(t)} \leftarrow [S^{j^{(1)}}, \dots, S^{j^{(t)}}]$, $Q^{(t)} \leftarrow [\sqrt{tq_{j^{(1)}}}, \dots, \sqrt{tq_{j^{(t)}}}]$, $W \leftarrow S^{(t)}/Q^{(t)}$

$U_1, D_1, V_1 \leftarrow \mathrm{SVD}(W)$ (decreasing order of singular values).

$N \leftarrow [U_1(1), \dots, U_1^{(r)}]$, $N \leftarrow S^T N/\|W^T N\|_F$

Choose $j^{(1)}, \dots, j^{(t)}$ uniformly at random in $\{1, \dots, m\}$.

$Y^{(t)} \leftarrow [y_{j^{(1)}}, \dots, y_{j^{(t)}}]$, $D^{(t)} \leftarrow d(X, Y^{(t)})/\sqrt{t}$.

$U_2, D_2, V_2 = \mathrm{SVD}(N^T N)$, $U_2 \leftarrow U_2/D_2$, $N^{(t)} \leftarrow [(N^T)^{(j^{(1)})}, \dots, (N^T)^{(j^{(t)})}]$, $B \leftarrow U_2^T N^{(t)}/\sqrt{t}$, $A \leftarrow (BB^T)^{-1}$.

$Z \leftarrow AB(D^{(t)})^T$, $M \leftarrow Z^T U_2^T$

**Result:** $M, N$

## 7.10 Nonnegative Low-rank Factorization of the Couplings

In this section, we recall the algorithm presented in [3] to solve problem (7.6) where we denote $p_1 := a$ and $p_2 := b$.

**Algorithm 14** LR-Dykstra($(K^{(i)})_{1\leq i\leq 3}, p_1, p_2, \alpha, \delta$) [3]

**Inputs:** $K^{(1)}, K^{(2)}, \tilde{g} := K^{(3)}, p_1, p_2, \alpha, \delta, q_1^{(3)} = q_2^{(3)} = \mathbf{1}_r, \forall i \in \{1,2\}$, $\tilde{v}^{(i)} = \mathbf{1}_r, q^{(i)} = \mathbf{1}_r$

**repeat**

$u^{(i)} \leftarrow p_i / K^{(i)}\tilde{v}^{(i)}\ \forall i \in \{1,2\}$,

$g \leftarrow \max(\alpha, \tilde{g} \odot q_1^{(3)}),\ q_1^{(3)} \leftarrow (\tilde{g} \odot q_1^{(3)})/g,\ \tilde{g} \leftarrow g$,

$g \leftarrow (\tilde{g} \odot q_2^{(3)})^{1/3} \prod_{i=1}^{2} (v^{(i)} \odot q^{(i)} \odot (K^{(i)})^T u^{(i)})^{1/3}$,

$v^{(i)} \leftarrow g/(K^{(i)})^T u^{(i)}\ \forall i \in \{1,2\}$,

$q^{(i)} \leftarrow (\tilde{v}^{(i)} \odot q^{(i)})/v^{(i)}\ \forall i \in \{1,2\},\ q_2^{(3)} \leftarrow (\tilde{g} \odot q_2^{(3)})/g$,

$\tilde{v}^{(i)} \leftarrow v^{(i)}\ \forall i \in \{1,2\},\ \tilde{g} \leftarrow g$

**until** $\sum_{i=1}^{2} \|u^{(i)} \odot K^{(i)} v^{(i)} - p_i\|_1 < \delta$;

$Q \leftarrow \mathrm{diag}(u^{(1)}) K^{(1)} \mathrm{diag}(v^{(1)})$

$R \leftarrow \mathrm{diag}(u^{(2)}) K^{(2)} \mathrm{diag}(v^{(2)})$

**Result:** $Q, R, g$

## 7.11 Additional Experiements

### 7.11.1 Illustration

In Fig. 7.10, we show the time-accuracy tradeoffs of the two methods presented in Figure 7.1 on the same example. We see that our method, **Lin GW-LR**, manages to obtain similar accuracy as the one obtained by **Quad Entropic-GW** even when the rank $r = n/1000$ while being much faster with order of magnitude.

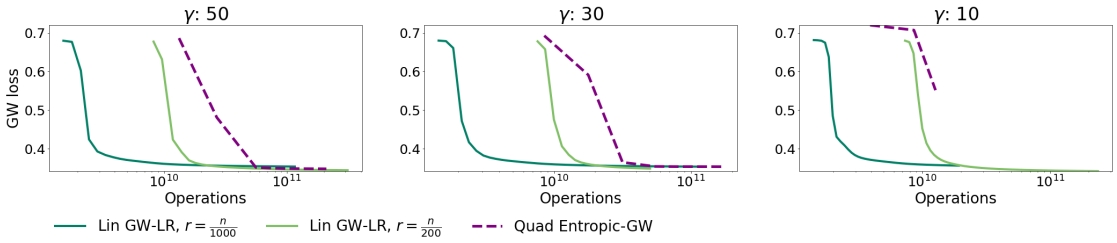



Figure 7.10: Here $n = m = 10000$, and the ground cost considered is the squared Euclidean distance. Note that for in that case we have an exact low-rank factorization of the cost. Therefore we compare only **Quad Entropic-GW** and **Lin GW-LR**. We plot the time-accuracy tradeoff when varying $\gamma$ for multiple ranks $r$. $\varepsilon = 1/\gamma$ for **Quad Entropic-GW** and $\varepsilon = 0$ for **Lin GW-LR**.

### 7.11.2 Effect of $\gamma$ and $\alpha$

In Fig. 5.6 and 7.11, we consider two Gaussian mixture densities in respectively 5-D and 10-D where we generate randomly the mean and covariance matrice of each Gaussian density using the wishart distribution.

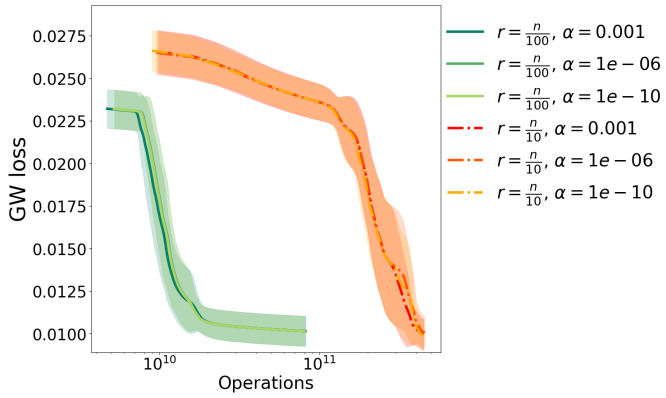



Figure 7.11: We consider $n = m = 5000$ samples of mixtures of (2 and 3) Gaussians in resp. 5 and 10-D, endowed with the squared Euclidean metric, compared with **Lin LR**. The time/loss tradeoff illustrated in these plots show that our method is not impacted by step size $\alpha$ for both ranks $r = n/100$ and $n/10$.

### 7.11.3 Effect of the Rank

In this experiment we compare two isotropic Gaussian blobs with respectively 10 and 20 centers in 10-D and 15-D and $n = m = 5000$ samples. In Fig. 7.12, we show the two first coordinates of the dataset considered.

### 7.11.4 Low-rank Problem

In Fig. 7.4, 7.6 and 7.7, we consider two distributions in respectively 10-D and 15-D where the support is a concatenation of clusters of points. In Fig. 7.13, we show an illustration of the distributions considered in smaller dimensions.

### 7.11.5 Ground Truth Experiment

In this experiment we aim at comparing the different methods when the optimal coupling solving the GW problem has a full rank. For that purpose we consider a certain shape in 2-D which corresponds to the support of the source distribution and we apply two isometric transformations to it, which are a rotation and a

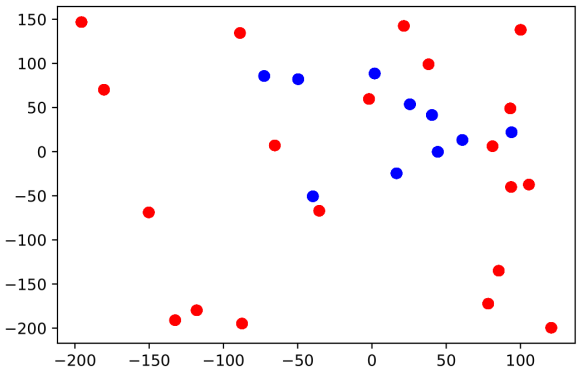



Figure 7.12: We consider two isotropic Gaussian blobs with respectively 10 and 20 centers in 10-D and 15-D and $n = m = 5000$ samples and we plot their 2 first coordinates.

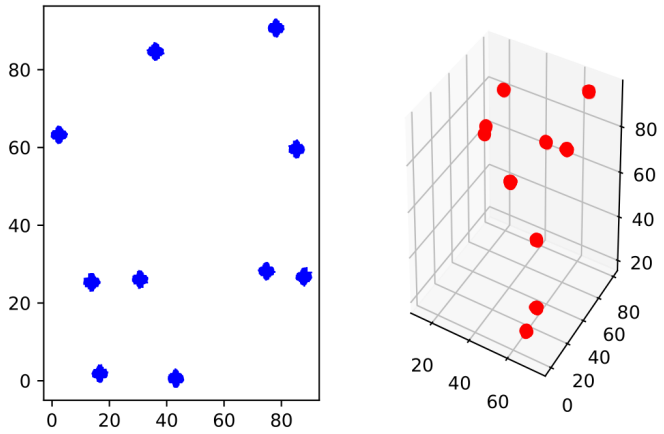



Figure 7.13: The source distribution and the target distribution live respectively in $\mathbb{R}^2$ and $\mathbb{R}^3$. Both distributions have the same number of samples $n = m = 10000$, the same number of clusters which is set to be 10 here, the same number of points in each cluster, and we force the distance between the centroids of the cluster to be larger than $\beta = 10$ in each distribution.

translation to obtain the support the target distribution. See Figure 7.14 (*left*) for an illustration of the dataset. Here we set $a$ and $b$ to be uniform distributions and the underlying cost is the squared Euclidean distance. Therefore the optimal coupling solution of the GW problem is the identity matrix and the GW loss must be 0. In Figure 7.15, we compare the time-accuracy tradeoffs, and we show that even in that case, our methods obtain a better time-accuracy tradeoffs for all $\gamma$.

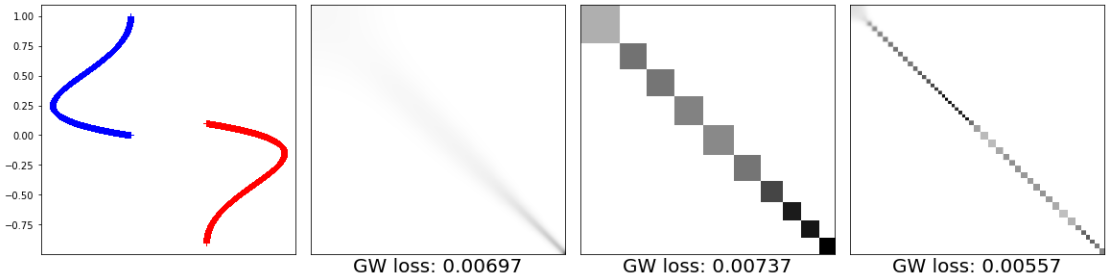



Figure 7.14: We compare the couplings obtained when the ground truth is the identity matrix in the same setting as in Figure 7.10. Here the comparison is done when $\gamma = 250$. *Left:* illustration of the dataset considered. *Middle left:* we show the coupling as well as the GW loss obtained by **Quad Entropic-GW**. *Middle right, right:* we show the couplings and the GW losses obtained by **Lin GW-LR** when the rank is respectively $r = 10$ and $r = 100$.

See also Figure 7.14 for a comparison of the couplings obtained by the different methods.

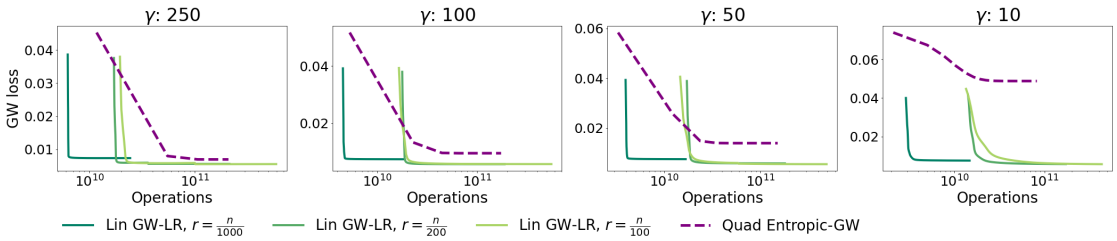



Figure 7.15: The ground truth here is the identity matrix and the true GW loss to achieve is 0. We set the number of samples to be $n = m = 10000$. As we consider the squared Euclidean distance, only **Quad Entropic-GW** and **Lin GW-LR** are compared. We plot the time-accuracy tradeoff when varying $\gamma$ for multiple choices of rank $r$. $\varepsilon = 1/\gamma$ for **Quad Entropic-GW** and $\varepsilon = 0$ for **Lin GW-LR**.

# Part III

# Applications of OT in Machine Learning

In this part, we show that OT can also offer new perspective on longstanding ML problems once lifted into the set of distributions. We adopt this point of view on two applied problems in fairness and robustness respectively and propose new approaches to tackle them using OT. This part consists of two contributions.

- In a fifth contribution, we propose to relax and lift the fair cake-cutting problem into the space of distributions and introduce an extension of the OT problem when multiple costs are involved. Considering each cost (or utility) as an agent, we aim to partition equitably resources between agents according to their heterogeneous preferences. To do so, we aim to maximize the utility of the least advantaged agent. This is a fair division problem. A transportation point of view of this problem is when the goal is to share equally between agents the work of transporting one distribution to another. Here we minimize the transportation cost of the agent who works the most. Like optimal transport, the problem can be cast as a linear optimization problem. When there is only one agent, we recover OT. When two agents are considered, we are able to recover Integral Probability Metrics (IPMs) defined by $\alpha$-Hölder functions, which include the widely-known Dudley metric. To the best of our knowledge, this is the first time a link is given between the Dudley metric and Optimal Transport.

- Finally in a sixth contribution, we tackles the problem of adversarial examples using OT. By lifting the attacker into the space of distributions, we obtain a variational formulation of the adversarial risk for deterministic as well as random classifiers where the adversary is restricted to live in a specific wasserstein ball. This new formulation of the adversarial risk allows us to interpret the adversarial risk minimization problem as a two-player zero-sum game between the attacker and the classifier. We then study the open question of the existence of mixed Nash equilibria in this zero-sum game. While previous works usually allow only one player to use randomized strategies, we show the necessity of considering randomization for both the classifier and the attacker. We demonstrate that this game has no duality gap, meaning that it always admits approximate Nash equilibria. We also provide the first optimization algorithms to learn a mixture of a finite number of classifiers that approximately realizes the value of this game, *i.e.* procedures to build an optimally robust randomize classifier.

# Chapter 8

# Equitable and Optimal Transport with Multiple Agents

We introduce an extension of the Optimal Transport problem when multiple costs are involved. Considering each cost as an agent, we aim to share equally between agents the work of transporting one distribution to another. To do so, we minimize the transportation cost of the agent who works the most. Another point of view is when the goal is to partition equitably goods between agents according to their heterogeneous preferences. Here we aim to maximize the utility of the least advantaged agent. This is a fair division problem. Like Optimal Transport, the problem can be cast as a linear optimization problem. When there is only one agent, we recover the Optimal Transport problem. When two agents are considered, we are able to recover Integral Probability Metrics defined by $\alpha$-Hölder functions, which include the widely-known Dudley metric. To the best of our knowledge, this is the first time a link is given between the Dudley metric and Optimal Transport. We provide an entropic regularization of that problem which leads to an alternative algorithm faster than the standard linear program.

This chapter is based on [5].

## 8.1 Introduction

Optimal Transport (OT) has gained interest last years in machine learning with diverse applications in neuroimaging [44], generative models [134, 50], supervised learning [92], word embeddings [226], reconstruction cell trajectories [56, 41] or adversarial examples [227]. The key to use OT in these applications lies in the gain of computation efficiency thanks to regularizations that smoothes the OT problem. More specifically, when one uses an entropic penalty, one recovers the so called Sinkhorn distances [76]. In this paper, we introduce a new family of variational problems extending the optimal transport problem when multiple costs are involved with various applications in fair division of goods/work and operations research problems.

Fair division [94] has been widely studied by the artificial intelligence [95] and economics [96] communities. Fair division consists in partitioning diverse resources among agents according to some fairness criteria. One of the standard problems in fair division is the fair cake-cutting problem [97, 98]. The cake is an heterogeneous resource, such as a cake with different toppings, and the agents have heterogeneous preferences over different parts of the cake, i.e., some people prefer the chocolate toppings, some prefer the cherries, others just want a piece as large as possible. Hence, taking into account these preferences, one might share the cake equitably between the agents. A generalization of this problem, for which achieving fairness constraints is more challenging, is when the splitting involves several heterogeneous cakes, and where the agents have linked preferences over the different parts of the cakes. This problem has many variants such as the cake-cutting with two cakes [99], or the Multi Type Resource Allocation [100, 101]. In all these models it is assumed that there is only one indivisible unit per type of resource available in each cake, and once an agent choose it, he or she has to take it all. In this setting, the cake can be seen as a set where each element of the set represents a type of resource, for instance each element of the cake represents a topping. A natural relaxation of these problems is when a divisible quantity of each type of resources is available. We introduce EOT (**E**quitable and **O**ptimal **T**ransport), a formulation that solves both the cake-cutting and the cake-cutting with two cakes problems in this setting.

Our problem expresses as an optimal transportation problem. Hence, we prove duality results and provide fast computation based on Sinkhorn algorithm. As interesting properties, some Integral Probability Metrics (IPMs) [228] as Dudley metric [219], or standard Wasserstein metric [30] are particular cases of the EOT problem.

**Contributions.** In this paper we introduce EOT an extension of Optimal Transport which aims at finding an equitable and optimal transportation strategy between multiple agents. We make the following contributions:

- In Section 8.3, we introduce the problem and show that it solves a fair division problem where heterogeneous resources have to be shared among multiple agents. We derive its dual and prove strong duality results. As a by-product, we show that EOT is related to some usual IPMs families and in particular the widely known Dudley metric.

- In Section 8.4, we propose an entropic regularized version of the problem, derive its dual formulation, obtain strong duality. We then provide an efficient algorithm to compute EOT. Finally we propose other applications of EOT for Operations Research problems.

## 8.2 Related Work

**Optimal Transport.** Optimal transport aims to move a distribution towards another at lowest cost. More formally, if $c$ is a cost function on the ground space $\mathcal{X} \times \mathcal{Y}$, then the relaxed Kantorovich formulation of OT is defined for $\mu$ and $\nu$ two distributions as

$$\mathrm{OT}_c(\mu, \nu) := \inf_{\gamma} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y) d\gamma(x, y)$$

where the infimum is taken over all distributions $\gamma$ with marginals $\mu$ and $\nu$. Kantorovich theorem states the following strong duality result under mild assumptions [30]

$$\mathrm{OT}_c(\mu, \nu) = \sup_{f,g} \int_{\mathcal{X}} f(x) d\mu(x) + \int_{\mathcal{Y}} g(y) d\nu(y)$$

where the supremum is taken over continuous bounded functions satisfying for all $x, y$, $f(x) + g(y) \leq c(x, y)$. The question of considering an optimal transport problem when multiple costs are involved has already been raised in recent works. For instance, [229] proposed a robust Wasserstein distance where the distributions are projected on a $k$-dimensional subspace that maximizes their transport cost. In that sense, they aim to choose the most expensive cost among Mahalanobis square distances with kernels of rank $k$. In articles [230, 231], the authors aim to learn a cost given observed matchings by inversing the optimal transport problem [232]. In [233] the authors study "feature-robust" optimal transport, which can be also seen as a robust cost selection for optimal transport. In articles [78, 6], the authors learn an adversarial cost to train a generative adversarial network. Here, we do not aim to consider a worst case scenario among the available costs but rather consider

that the costs work together in order to split equitably the transportation problem among them at lowest cost.

**Entropic Relaxation of OT.** Computing exactly the optimal transport cost requires solving a linear program with a supercubic complexity ($n^3 \log n$) [172] that results in an output that is *not* differentiable with respect to the measures' locations or weights [173]. Moreover, OT suffers from the curse of dimensionality [129, 71] and is therefore likely to be meaningless when used on samples from high-dimensional densities. Following the line of work introduced by [76], we propose an approximated computation of our problem by regularizing it with an entropic term. Such regularization in OT accelerates the computation, makes the problem differentiable with regards to the distributions [130] and reduces the curse of dimensionality [51]. Taking the dual of the approximation, we obtain a smooth and convex optimization problem under a simplicial constraint.

**Fair Division.** Fair division of goods has a long standing history in economics and computational choice. A classical problem is the fair cake-cutting that consists in splitting the cake between $N$ individuals according to their heterogeneous preferences. The cake $\mathcal{X}$, viewed as a set, is divided in $\mathcal{X}_1, \ldots, \mathcal{X}_N$ disjoint sets among the $N$ individuals. The utility for a single individual $i$ for a slice $S$ is denoted $V_i(S)$. It is often assumed that $V_i(\mathcal{X}) = 1$ and that $V_i$ is additive for disjoint sets. There exists many criteria to assess fairness for a partition $\mathcal{X}_1, \ldots, \mathcal{X}_N$ such as proportionality ($V_i(\mathcal{X}_i) \geq 1/N$), envy-freeness ($V_i(\mathcal{X}_i) \geq V_i(\mathcal{X}_j)$) or equitability ($V_i(\mathcal{X}_i) = V_j(\mathcal{X}_j)$). The cake-cutting problem has applications in many fields such as dividing land estates, advertisement space or broadcast time. An extension of the cake-cutting problem is the cake-cutting with two cakes problem [99] where two heterogeneous cakes are involved. In this problem, preferences of the agents can be coupled over the two cakes. The slice of one cake that an agent prefers might be influenced by the slice of the other cake that he or she might also obtain. The goal is to find a partition of the cakes that satisfies fairness conditions for the agents sharing the cakes. Cloutier et al. [99] studied the envy-freeness partitioning. Both the cake-cutting and the cake-cutting with two cakes problems assume that there is only one indivisible unit of supply per element $x \in \mathcal{X}$ of the cake(s). Therefore sharing the cake(s) consists in obtaining a paritition of the set(s). In this paper, we show that EOT is a relaxation of the cutting cake and the cake-cutting with two cakes problems, when there is a divisible amount of each element of the cake(s). In that case, cakes are no more sets but distributions that we aim to divide between the agents according to their coupled preferences.

**Integral Probability Metrics.** In our work, we make links with some integral probability metrics. IPMs are (semi-)metrics on the space of probability measures. For a set of functions $\mathcal{F}$ and two probability distributions $\mu$ and $\nu$, they are defined as

$$\mathrm{IPM}_{\mathcal{F}}(\mu,\nu) = \sup_{f\in\mathcal{F}} \int f d\mu - \int f d\nu.$$

For instance, when $\mathcal{F}$ is chosen to be the set of bounded functions with uniform norm less or equal than 1, we recover the Total Variation distance [132] (TV). They recently regained interest in the Machine Learning community thanks to their application to Generative Adversarial Networks (GANs) [22] where IPMs are natural metrics for the discriminator [133, 134, 135, 136]. They also helped to build consistent two-sample tests [18, 137]. However when a closed form of the IPM is not available, exact computation of IPMs between discrete distributions may not be possible or can be costful. For instance, the Dudley metric can be written as a Linear Program [138] which has at least the same complexity as standard OT. Here, we show that the Dudley metric is in fact a particular case of our problem and obtain a faster approximation thanks to the entropic regularization.

## 8.3 Equitable and Optimal Transport

**Notations.** Let $\mathcal{Z}$ be a Polish space, we denote $\mathcal{M}(\mathcal{Z})$ the set of Radon measures on $\mathcal{Z}$. We call $\mathcal{M}_+(\mathcal{Z})$ the sets of positive Radon measures, and $\mathcal{P}(\mathcal{Z})$ the set of probability measures. We denote $\mathcal{C}_b(\mathcal{Z})$ the vector space of bounded continuous functions on $\mathcal{Z}$. Let $\mathcal{X}$ and $\mathcal{Y}$ be two Polish spaces. We denote for $\mu \in \mathcal{M}(\mathcal{X})$ and $\nu \in \mathcal{M}(\mathcal{Y})$, $\mu\otimes\nu$ the tensor product of the measures $\mu$ and $\nu$, and $\mu \ll \nu$ means that $\nu$ dominates $\mu$. We denote $\pi_1 : (x,y) \in \mathcal{X}\times\mathcal{Y} \mapsto x$ and $\pi_2 : (x,y) \in \mathcal{X}\times\mathcal{Y} \mapsto y$ respectively the projections on $\mathcal{X}$ and $\mathcal{Y}$, which are continuous applications. For an application $g$ and a measure $\mu$, we denote $g_{\#}\mu$ the pushforward measure of $\mu$ by $g$. For $\mathcal{X}$ and $\mathcal{Y}$ two Polish spaces, we denote $\mathrm{LSC}(\mathcal{X}\times\mathcal{Y})$ the space of lower semi-continuous functions on $\mathcal{X}\times\mathcal{Y}$, $\mathrm{LSC}^+(\mathcal{X}\times\mathcal{Y})$ the space of non-negative lower semi-continuous functions on $\mathcal{X}\times\mathcal{Y}$ and $\mathrm{LSC}^-_*(\mathcal{X}\times\mathcal{Y})$ the set of negative bounded below lower semi-continuous functions on $\mathcal{X}\times\mathcal{Y}$ . We also denote $\mathrm{C}^+(\mathcal{X}\times\mathcal{Y})$ the space of non-negative continuous functions on $\mathcal{X}\times\mathcal{Y}$ and $\mathrm{C}^-_*(\mathcal{X}\times\mathcal{Y})$ the set of negative continuous functions on $\mathcal{X}\times\mathcal{Y}$. Let $N \geq 1$ be an integer and denote $\Delta_N^+ := \{\lambda \in \mathbb{R}_+^N \text{ s.t. } \sum_{i=1}^N \lambda_i = 1\}$, the probability simplex of $\mathbb{R}^N$. For two positive measures of same mass $\mu \in \mathcal{M}_+(\mathcal{X})$ and $\nu \in \mathcal{M}_+(\mathcal{Y})$, we define the set of couplings with marginals $\mu$ and $\nu$:

$$\Pi(\mu,\nu) := \{\gamma \text{ s.t. } \pi_1\#\gamma = \mu \ , \ \pi_2\#\gamma = \nu\} \ .$$

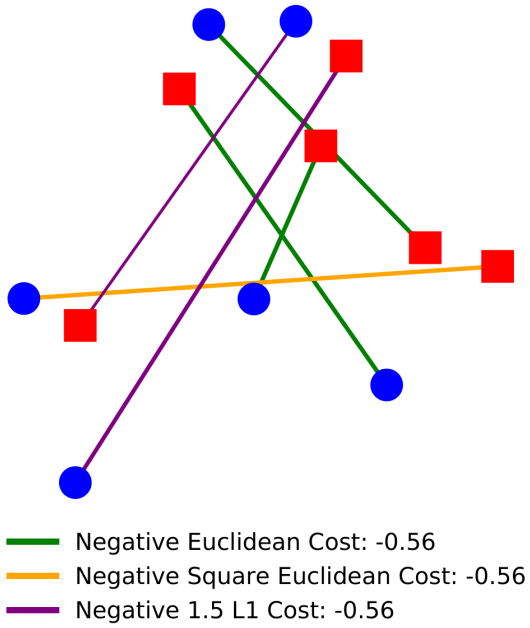



Figure 8.1: Equitable and optimal division of the resources between $N = 3$ different negative costs (i.e. utilities) given by EOT. Utilities have been normalized. Blue dots and red squares represent the different elements of resources available in each cake. We consider the case where there is exactly one unit of supply per element in the cakes, which means that we consider uniform distributions. Note that the partition between the agents is equitable (i.e. utilities are equal) and proportional (i.e. utilities are larger than $1/N$).

We introduce the subset of $(\mathcal{M}_+(\mathcal{X}) \times \mathcal{M}_+(\mathcal{Y}))^N$ representing marginal decomposition:

$$\Upsilon^N(\mu,\nu) := \Big\{ (\mu_i,\nu_i)_{i=1}^N \text{ s.t. } \sum_i \mu_i = \mu, \sum_i \nu_i = \nu \text{ and } \forall i,\ \mu_i(\mathcal{X}) = \nu_i(\mathcal{Y}) \Big\}.$$

We also define the following subset of $\mathcal{M}_+(\mathcal{X} \times \mathcal{Y})^N$ corresponding to the coupling decomposition:

$$\Gamma^N(\mu,\nu) := \Big\{ (\gamma_i)_{i=1}^N \text{ s.t. } \pi_1 \# \sum \gamma_i = \mu \ ,\ \pi_2 \# \sum \gamma_i = \nu \Big\}.$$

### 8.3.1 Primal Formulation

Consider a fair division problem where several agents aim to share two sets of resources, $\mathcal{X}$ and $\mathcal{Y}$, and assume that there is a divisible amount of each resource $x \in \mathcal{X}$ (resp. $y \in \mathcal{Y}$) that is available. Formally, we consider the case where

resources are no more sets but rather distributions on these sets. Denote $\mu$ and $\nu$ the distribution of resources on respectively $\mathcal{X}$ and $\mathcal{Y}$. For example, one might think about a situation where agents want to share fruit juices and ice creams and there is a certain volume of each type of fruit juices and a certain mass of each type of ice creams available. Moreover each agent defines his or her paired preferences for each couple $(x, y) \in \mathcal{X} \times \mathcal{Y}$. Formally, each person $i$ is associated to an upper semi-continuous mapping $u_i : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$ corresponding to his or her preference for any given pair $(x, y)$. For example, one may prefer to eat chocolate ice cream with apple juice, but may prefer pineapple juice when it comes with vanilla ice cream. The total utility for an individual $i$ and a pairing $\gamma_i \in \mathcal{M}_+(\mathcal{X} \times \mathcal{Y})$ is then given by $V_i(\gamma_i) := \int u_i d\gamma_i$. To partition fairly among individuals, we maximize the minimum of individual utilities.

From a transport point of view, let assume that there are $N$ workers available to transport a distribution $\mu$ to another one $\nu$. The cost of a worker $i$ to transport a unit mass from location $x$ to the location $y$ is $c_i(x, y)$. To partition the work among the $N$ workers fairly, we minimize the maximum of individual costs.

These problems are in fact the same where the utility $u_i$, defined in the fair division problem, might be interpreted as the opposite of the cost $c_i$ defined in the transportation problem, i.e. for all $i$, $c_i = -u_i$. The two above problem motivate the introduction of EOT defined as follows.

**Definition 3** (Equitable and Optimal Transport)**.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be Polish spaces. Let $\mathbf{c} := (c_i)_{1 \leq i \leq N}$ be a family of bounded below lower semi-continuous cost functions on $\mathcal{X} \times \mathcal{Y}$, and $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$. We define the equitable and optimal transport primal problem:*

$$\mathrm{EOT}_{\mathbf{c}}(\mu, \nu) := \inf_{(\gamma_i)_{i=1}^N \in \Gamma^N(\mu,\nu)} \max_i \int c_i d\gamma_i \ . \tag{8.1}$$

We prove along with Theorem 8.3.1 that the problem is well defined and the infimum is attained. Lower-semi continuity is a standard assumption in OT. In fact, it is the weakest condition to prove Kantorovich duality [30, Chap. 1]. Note that the problem defined here is a linear optimization problem and when $N = 1$ we recover standard optimal transport. Figure 8.1 illustrates the equitable and optimal transport problem we consider. Figure 8.5 in Appendix 8.9 shows an illustration with respect to the transport viewpoint in the exact same setting, i.e. $c_i = -u_i$. As expected, the couplings obtained in the two situations are not the same. We now show that in fact, EOT optimum satisfies equality constraints in case of constant sign costs, i.e. total utility/cost of each individual are equal in the optimal partition. See Appendix 8.6.2 for the proof.

**Proposition 8.3.1** (EOT solves the problem under equality constraints)**.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be Polish spaces. Let $\mathbf{c} := (c_i)_{1\leq i\leq N} \in \mathrm{LSC}^+(\mathcal{X}\times\mathcal{Y})^N \cup \mathrm{LSC}^-_*(\mathcal{X}\times\mathcal{Y})^N$, $\mu\in\mathcal{P}(\mathcal{X})$ and $\nu\in\mathcal{P}(\mathcal{Y})$. Then the following are equivalent:*

- $(\gamma_i^*)_{i=1}^N \in \Gamma^N(\mu,\nu)$ *is solution of Eq. (8.1),*
- $(\gamma_i^*)_{i=1}^N \in \underset{(\gamma_i)_{i=1}^N\in\Gamma^N(\mu,\nu)}{\operatorname{argmin}} \left\{t \text{ s.t. } \forall i \ \int c_i d\gamma_i = t\right\}$.

*Moreover, we have that*

$$\mathrm{EOT}_{\mathbf{c}}(\mu,\nu) = \min_{(\gamma_i)_{i=1}^N\in\Gamma^N(\mu,\nu)} \left\{ t \text{ s.t. } \forall i \ \int c_i d\gamma_i = t \right\} .$$

This property highly relies on the sign of the costs. For instance if two costs are considered, one always positive and the other always negative, then the constraints cannot be satisfied. When the cost functions are non-negatives, EOT refers to a transportation problem while when the costs are all negatives, costs become utilities and EOT refers to a fair division problem. The two points of view are concordant, but proofs and interpretations rely on the sign of the costs.

### 8.3.2 An Equitable and Proportional Division

When the cost functions considered $c_i$ are all negatives, EOT become a fair division problem where the utility functions are defined as $u_i := -c_i$. Indeed according to Proposition 8.3.1, EOT solves

$$\max_{(\gamma_i)_{i=1}^N\in\Gamma^N(\mu,\nu)} \left\{ t \text{ s.t. } \forall i, \ \int u_i d\gamma_i = t \right\} .$$

Recall that in our model, the total utility of the agent $i$ is given by $V_i(\gamma_i) := \int u_i d\gamma_i$. Therefore EOT aims to maximize the total utility of each agent $i$ while ensuring that they are all equal. Let us now analyze which fairness conditions the partition induced by EOT verifies. Assume that the utilities are normalized, i.e., $\forall i$, there exists $\gamma_i \in \mathcal{P}(\mathcal{X}\times\mathcal{Y})$ such that $V_i(\gamma_i) = 1$. For example one might consider the cases where $\forall i$, $\gamma_i = \mu\otimes\nu$ or $\gamma_i \in \operatorname{argmin}_{\gamma\in\Pi_{\mu,\nu}} \int c_i d\gamma$. Then any solution $(\gamma_i^*)_{i=1}^N \in \Gamma^N_{\mu,\nu}$ of EOT satisfies:

- **Proportionality**: for all $i$, $V_i(\gamma_i^*) \geq 1/N$,
- **Equitablity**: for all $i,j$, $V_i(\gamma_i^*) = V_j(\gamma_j^*)$.

Proportionality is a standard fair division criterion for which a resource is divided among $N$ agents, giving each agent at least $1/N$ of the heterogeneous resource by his/her own subjective valuation. Therefore here, this situation corresponds to

the case where the normalized utility of each agent is at least $1/N$. Moreover, an equitable division is a division of an heterogeneous resource, in which each partner is equally happy with his/her share. Here this corresponds to the case where the utility of each agent are all equal. The problem solved by EOT is a fair division problem where heterogeneous resources have to be shared among multiple agents according to their preferences. This problem is a relaxation of the two cake-cutting problem when there are a divisible amount of each item of the cakes. In that case, cakes are distributions and EOT makes a proportional and equitable partition of them. Details are left in Appendix 8.6.2.

**Fair Cake-cutting.** Consider the case where the cake is an heterogeneous resource and there is a certain divisible quantity of each type of resource available. For example chocolate and vanilla are two types of resource present in the cake for which a certain mass is available. In that case, each type of resource in the cake is pondered by the actual quantity present in the cake. Up to a normalization, the cake is no more the set $\mathcal{X}$ but rather a distribution on this set. Note that for the two points of view to coincide, it suffices to assume that there is exactly the same amount of mass for each type of resources available in the cake. In that case, the cake can be represented by the uniform distribution over the set $\mathcal{X}$, or equivalently the set $\mathcal{X}$ itself. When cakes are distributions, the fair cutting cake problem can be interpreted as a particular case of EOT when the utilities of the agents do not depend on the variable $y \in \mathcal{Y}$. In short, we consider that utilities are functions of the form $u_i(x, y) = v_i(x)$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$. The normalization of utilities can be cast as follows: $\forall i$, $V_i(\mu) = \int v_i(x) d\mu(x) = 1$. Then Proposition 8.3.1 shows that the partition of the cake made by EOT is proportional and equitable. Note that for EOT to coincide with the classical cake-cutting problem, one needs to consider that the uniform masses of the cake associated to each type of resource cannot be splitted. This can be interpreted as a Monge formulation [30] of EOT which is out of the scope of this paper.

### 8.3.3 Optimality of EOT

We next investigate the coupling obtained by solving EOT. In the next proposition, we show that under the same assumptions of Proposition 8.3.1, EOT solutions are optimal transportation plans. See Appendix 8.6.3 for the proof.

**Proposition 8.3.2** (EOT realizes optimal plans)**.** *Under the same conditions of Proposition 8.3.1, for any $(\gamma_i^*)_{i=1}^N \in \Gamma^N(\mu, \nu)$ solution of Eq. (8.1), we have for all $i \in \{1, \dots, N\}$*

$$\gamma_i^* \in \underset{\gamma \in \Pi(\mu_i^*, \nu_i^*)}{\operatorname{argmin}} \int c_i d\gamma \quad \textit{where} \qquad \mu_i^* := \pi_1 \# \gamma_i^* \ , \ \nu_i^* := \Pi_2 \# \gamma_i^* \ , \tag{8.2}$$

*and*

$$\mathrm{EOT}_{\mathbf{c}}(\mu,\nu) = \min_{(\mu_i,\nu_i)_{i=1}^N \in \Upsilon^N(\mu,\nu)} t \quad \text{s.t.} \quad \forall i \; OT_{c_i}(\mu_i,\nu_i) = t \; . \tag{8.3}$$

Given the optimal matchings $(\gamma_i^*)_{i=1}^N \in \Gamma^N(\mu,\nu)$, one can easily obtain the partition of the agents of each marginals. Indeed for all $i$, $\mu_i^* := \pi_1 \# \gamma_i^*$ and $\nu_i^* := \pi_2 \# \gamma_i^*$ represent respectively the portion of the agent $i$ from distributions $\mu$ and $\nu$.

**Remark 10** (Utilitarian and Optimal Transport)**.** *To contrast with* EOT*, an alternative problem is to maximize the sum of the total utilities of agents, or equivalently minimize the sum of the total costs of agents. This problem can be cast as follows:*

$$\inf_{(\gamma_i)_{i=1}^N \in \Gamma^N(\mu,\nu)} \sum_i \int c_i d\gamma_i \tag{8.4}$$

*Here one aims to maximize the total utility of all the agents, while in* EOT *we aim to maximize the total utility per agent under egalitarian constraint. The solution of (8.4) is not fair among agents and one can show that this problem is actually equal to* $OT_{\min_i(c_i)}(\mu,\nu)$*. Details can be found in Appendix 8.8.1.*

### 8.3.4 Dual Formulation

Let us now introduce the dual formulation of the problem and show that strong duality holds under some mild assumptions. See Appendix 8.6.4 for the proof.

**Theorem 8.3.1** (Strong Duality)**.** *Let* $\mathcal{X}$ *and* $\mathcal{Y}$ *be Polish spaces. Let* $\mathbf{c} := (c_i)_{i=1}^N$ *be bounded below lower semi-continuous costs. Then* strong duality holds*, i.e. for* $(\mu,\nu) \in \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y})$*:*

$$\mathrm{EOT}_{\mathbf{c}}(\mu,\nu) = \sup_{\substack{\lambda \in \Delta_N^+ \\ (f,g) \in \mathcal{F}_{\mathbf{c}}^\lambda}} \int f d\mu + \int g d\nu \tag{8.5}$$

*where* $\mathcal{F}_{\mathbf{c}}^\lambda := \{(f,g) \in \mathcal{C}_b(\mathcal{X}) \times \mathcal{C}_b(\mathcal{Y}) \text{ s.t. } \forall i \in \{1,...,N\}, \; f \oplus g \leq \lambda_i c_i\}$.

This theorem holds under the same hypothesis and follows the same reasoning as the one in [30, Theorem 1.3]. While the primal formulation of the problem is easy to understand, we want to analyse situations where the dual variables also play a role. For that purpose we show in the next proposition a simple characterisation of the primal-dual optimality in case of constant sign cost functions. See Appendix 8.6.5 for the proof.

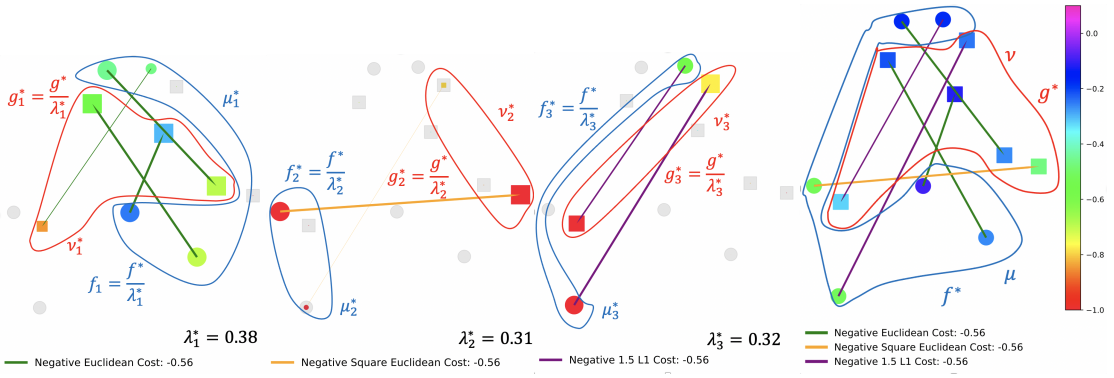



Figure 8.2: *Left, middle left, middle right*: the size of dots and squares is proportional to the weight of their representing atom in the distributions $\mu_k^*$ and $\nu_k^*$ respectively. The utilities $f_k^*$ and $g_k^*$ for each point in respectively $\mu_k^*$ and $\nu_k^*$ are represented by the color of dots and squares according to the color scale on the right hand side. The gray dots and squares correspond to the points that are ignored by agent $k$ in the sense that there is no mass or almost no mass in distributions $\mu_k^*$ or $\nu_k^*$. *Right*: the size of dots and squares are uniform since they correspond to the weights of uniform distributions $\mu$ and $\nu$ respectively. The values of $f^*$ and $g^*$ are given also by the color at each point. Note that each agent gets exactly the same total utility, corresponding exactly to EOT. This value can be computed using dual formulation (8.5) and for each figure it equals the sum of the values (encoded with colors) multiplied by the weight of each point (encoded with sizes).

**Proposition 8.3.3.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be compact Polish spaces. Let $\mathbf{c} := (c_i)_{1\leq i\leq N} \in \mathrm{C}^+(\mathcal{X}\times\mathcal{Y})^N \cup \mathrm{C}_*^-(\mathcal{X}\times\mathcal{Y})^N$, $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$. Let also $(\gamma_k)_{k=1}^N \in \Gamma^N(\mu,\nu)$ and $(\lambda, f, g) \in \Delta_n^+ \times \mathcal{C}_b(\mathcal{X}) \times \mathcal{C}_b(\mathcal{Y})$. Then Eq. (8.5) admits a solution and the following are equivalent:*

- *$(\gamma_k)_{k=1}^N$ is a solution of Eq. (8.1) and $(\lambda, f, g)$ is a solution of Eq. (8.5).*
- 1. $\forall i \in \{1,...,N\},\ f \oplus g \leq \lambda_i c_i$
  2. $\forall i,j \in \{1,...,N\}\ \int c_i d\gamma_i = \int c_j d\gamma_j$
  3. $f \oplus g = \lambda_i c_i \quad \gamma_i$-$a.e.$

**Remark 11.** *It is worth noting that when we assume that $\mathbf{c} := (c_i)_{1\leq i\leq N} \in \mathrm{C}_*^+(\mathcal{X}\times\mathcal{Y})^N \cup \mathrm{C}_*^-(\mathcal{X}\times\mathcal{Y})^N$, then we can refine the second point of the equivalence presented in Proposition 8.3.3 by adding the following condition: $\forall i \in \{1,...,N\}\ \lambda_i \neq 0$.*

Given two distributions of resources represented by the measures $\mu$ and $\nu$, and $N$ utility functions denoted $(u_i)_{i=1}^N$, we want to find an *equitable* and *stable* partition among the agents in case of *transferable utilities*. Let $k$ be an agent. We say that

his or her utility is transferable when once $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ get matched, he or she has to decide how to split his or her associated utility $u_k(x, y)$ . She or he divides $u_k(x, y)$ into a quantity $f_k(x)$ which can be seen as the utility of having $x$ and $g_k(y)$ for having $y$. Therefore in that problem we ask for $(\gamma_k, f_k, g_k)_{k=1}^N$ such that

$$u_k(x, y) = f_k(x) + g_k(y) \quad \gamma_k\text{-a.e.} \tag{8.6}$$

Moreover, for the partition to be *stable* [234], we want to ensure that, for every agent $k$, none of the resources $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ that have not been matched together for this agent would increase their utilities, $f_k(x)$ and $g_k(y)$, if there were matched together in the current matching instead. Formally we ask that for $k \in \{1, \dots, N\}$ and all $(x, y) \in \mathcal{X} \times \mathcal{Y}$,

$$f_k(x) + g_k(y) \geq u_k(x, y) \ . \tag{8.7}$$

Indeed if there exist $k$, $x$ and $y$ such that $u_k(x, y) > f_k(x) + g_k(y)$, then $x$ and $y$ will not be matched together in the share of the agent $k$ and he can improve his utility for both $x$ and $y$ by matching $x$ with $y$.

Finally we aim to share equitably the resources among the agents which boils down to ask

$$\forall i, j \in \{1, ..., N\} \quad \int u_i d\gamma_i = \int u_j d\gamma_j \tag{8.8}$$

Thanks to Proposition 8.3.3, finding $(\gamma_k, f_k, g_k)_{k=1}^N$ satisfying (8.6), (8.7) and (8.8) can be done by solving Eq. (8.1) and Eq. (8.5). Indeed let $(\gamma_k)_{k=1}^N$ an optimal solution of Eq. (8.1) and $(\lambda, f, g)$ an optimal solution of Eq. (8.5). Then by denoting for all $k = 1, \dots, N$, $f_k = \frac{f}{\lambda_k}$ and $g_k = \frac{g}{\lambda_k}$, we obtain that $(\gamma_k, f_k, g_k)_{k=1}^N$ solves the *equitable* and *stable* partition problem in case of *transferable utilities*. Note that again, we end up with equality constraints for the optimal dual variables. Indeed, for all $i, j \in \{1, \dots, N\}$, at optimality we have $\int f_i + g_i d\gamma_i = \int f_j + g_j d\gamma_j$ . Figure 8.2 illustrates this formulation of the problem with dual potentials. Figure 8.7 in Appendix 8.9 shows the dual solutions with respect to the transport viewpoint in the exact same setting, i.e. $c_i = -u_i$. Once again, the obtained solutions differ.

### 8.3.5 Link with other Probability Metrics

In this section, we provide some topological properties on the object defined by the EOT problem. In particular, we make links with other known probability metrics, such as Dudley and Wasserstein metrics and give a tight upper bound.

When $N = 1$, recall from the definition (8.1) that the problem considered is

exactly the standard OT problem. Moreover any EOT problem with $k \leq N$ costs can always be rewritten as a EOT problem with $N$ costs. See Appendix 8.8.2 for the proof. From this property, it is interesting to note that, for any $N \geq 1$, EOT generalizes standard Optimal Transport.

**Optimal Transport.** Given a cost function $c$, if we consider the problem EOT with $N$ costs such that, for all $i$, $c_i = N \times c$ then, the problem $\mathrm{EOT}_{\mathbf{c}}$ is exactly $\mathrm{OT}_c$. See Appendix 8.8.2 for the proof.

Now we have seen that all standard OT problems are sub-cases of the EOT problem, one may ask whether EOT can recover other families of metrics different from standard OT. Indeed we show that the EOT problem recovers an important family of IPMs with supremum taken over the space of $\alpha$-Hölder functions with $\alpha \in (0, 1]$. See Appendix 8.6.6 for the proof.

**Proposition 8.3.4.** *Let $\mathcal{X}$ be a Polish space. Let $d$ be a metric on $\mathcal{X}^2$ and $\alpha \in (0, 1]$. Denote $c_1 = 2 \times \mathbf{1}_{x \neq y}$, $c_2 = d^\alpha$ and $\mathbf{c} := (c_1, (N-1) \times c_2, ..., (N-1) \times c_2) \in \mathrm{LSC}(\mathcal{X} \times \mathcal{X})^N$ then for any $(\mu, \nu) \in \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X})$*

$$\mathrm{EOT}_{\mathbf{c}}(\mu, \nu) = \sup_{f \in B_{d^\alpha}(\mathcal{X})} \int_{\mathcal{X}} f d\mu - \int_{\mathcal{X}} f d\nu \tag{8.9}$$

*where $B_{d^\alpha}(\mathcal{X}) := \left\{ f \in C^b(\mathcal{X}) \colon \|f\|_\infty + \|f\|_\alpha \leq 1 \right\}$ and $\|f\|_\alpha := \sup_{x \neq y} \frac{|f(x) - f(y)|}{d^\alpha(x,y)}$.*

**Dudley Metric.** When $\alpha = 1$, then for $(\mu, \nu) \in \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X})$, we have

$$\mathrm{EOT}_{\mathbf{c}}(\mu, \nu) = \mathrm{EOT}_{(c_1, d)}(\mu, \nu) = \beta_d(\mu, \nu)$$

where $\beta_d$ is the *Dudley Metric* [219]. In other words, the Dudley metric can be interpreted as an equitable and optimal transport between the measures with the trivial cost and a metric $d$. We acknowledge that [235] made a link between Unbalanced Optimal Transport and the "flat metric", an IPM close to the Dudley metric, defined on the space $\{f \colon \|f\|_\infty \leq 1,\ \|f\|_1 \leq 1\}$.

**Weak Convergence.** When $d$ is an unbounded metric on $\mathcal{X}$, it is well known that $\mathrm{OT}_{d^p}$ with $p \in (0, +\infty)$ metrizes a convergence a bit stronger than weak convergence [30, Chap. 7]. A sufficient condition for Wasserstein distances to metrize weak convergence on the space of distributions is that the metric $d$ is bounded. In contrast, metrics defined by Eq. (8.9) do not require such assumptions and $\mathrm{EOT}_{(\mathbf{1}_{x \neq y}, d^\alpha)}$ metrizes the weak convergence of probability measures [30, Chap. 1-7].

For an arbitrary choice of costs $(c_i)_{1\leq i\leq N}$, we obtain a tight upper control of EOT and show how it is related to the OT problem associated to each cost involved. See Appendix 8.6.7 for the proof.

**Proposition 8.3.5.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be Polish spaces. Let $\mathbf{c} := (c_i)_{1\leq i\leq N}$ be a family of nonnegative lower semi-continuous costs. For any $(\mu,\nu) \in \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y})$*

$$\mathrm{EOT}_{\mathbf{c}}(\mu,\nu) \leq \left(\sum_{i=1}^{N} \frac{1}{OT_{c_i}(\mu,\nu)}\right)^{-1} \tag{8.10}$$

Proposition 8.3.5 means that the minimal cost to transport all goods under the constraint that all workers contribute equally is lower than the case where agents share equitably and optimally the transport with distributions $\mu_i$ and $\nu_i$ respectively proportional to $\mu$ and $\nu$, which equals the harmonic sum written in Equation (8.10).

**Example.** *Applying the above result in the case of the Dudley metric recovers the following inequality [138, Proposition 5.1]*

$$\beta_d(\mu,\nu) \leq \frac{\mathrm{TV}(\mu,\nu)\, OT_d(\mu,\nu)}{\mathrm{TV}(\mu,\nu) + OT_d(\mu,\nu)}.$$

## 8.4 Entropic Relaxation

In their original form, as proposed by Kantorovich [32], Optimal Transport distances are not a natural fit for applied problems: they minimize a network flow problem, with a supercubic complexity $(n^3 \log n)$ [172]. Following the work of [76], we propose an entropic relaxation of EOT, obtain its dual formulation and derive an efficient algorithm to compute an approximation of EOT.

### 8.4.1 Primal-Dual Formulation

Let us first extend the notion of Kullback-Leibler divergence for positive Radon measures. Let $\mathcal{Z}$ be a Polish space, for $\mu,\nu \in \mathcal{M}_+(\mathcal{Z})$, we define the generalized Kullback-Leibler divergence as $\mathrm{KL}(\mu,\nu) = \int \log \frac{d\mu}{d\nu} d\mu + \int d\nu - \int d\mu$ if $\mu \ll \nu$, and $+\infty$ otherwise. We introduce the following regularized version of EOT.

**Definition 4** (Entropic relaxed primal problem)**.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be two Polish spaces, $\mathbf{c} := (c_i)_{1\leq i\leq N}$ a family of bounded below lower semi-continuous costs lower*

*semi-continuous costs on* $\mathcal{X} \times \mathcal{Y}$ *and* $\boldsymbol{\varepsilon} := (\varepsilon_i)_{1 \leq i \leq N}$ *be non negative real numbers. For* $(\mu, \nu) \in \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y})$*, we define the* EOT regularized primal problem*:*

$$\mathrm{EOT}_{\mathbf{c}}^{\boldsymbol{\varepsilon}}(\mu, \nu) := \inf_{\gamma \in \Gamma^N(\mu,\nu)} \max_i \int c_i d\gamma_i + \sum_{j=1}^{N} \varepsilon_j \mathrm{KL}(\gamma_j, \mu \otimes \nu)$$

Note that here we sum the generalized Kullback-Leibler divergences since our objective is function of $N$ measures in $\mathcal{M}_+(\mathcal{X} \times \mathcal{Y})$. This problem can be compared with the one from standard regularized OT. In the case where $N = 1$, we recover the standard regularized OT. For $N \geq 1$, the underlying problem is $\sum_{i=1}^{N} \varepsilon_i-$strongly convex. Moreover, we prove the essential property that as $\boldsymbol{\varepsilon} \to 0$, the regularized problem converges to the standard problem. See Appendix 8.8.3 for the full statement and the proof. As a consequence, entropic regularization is a consistent approximation of the original problem we introduced in Section 8.3.1. Next theorem shows that strong duality holds for lower semi-continuous costs and compact spaces. This is the basis of the algorithm we will propose in Section 8.4.2. See Appendix 8.6.8 for the proof.

**Theorem 8.4.1** (Duality for the regularized problem)**.** *Let* $\mathcal{X}$ *and* $\mathcal{Y}$ *be two compact Polish spaces,* $\mathbf{c} := (c_i)_{1 \leq i \leq N}$ *a family of bounded below lower semi-continuous costs on* $\mathcal{X} \times \mathcal{Y}$ *and* $\boldsymbol{\varepsilon} := (\varepsilon_i)_{1 \leq i \leq N}$ *be non negative numbers. For* $(\mu, \nu) \in \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y})$*, strong duality holds:*

$$\begin{aligned} \mathrm{EOT}_{\mathbf{c}}^{\boldsymbol{\varepsilon}}(\mu, \nu) = \sup_{\lambda \in \Delta_N^+} \sup_{\substack{f \in \mathcal{C}_b(\mathcal{X}) \\ g \in \mathcal{C}_b(\mathcal{Y})}} & \int f d\mu + \int g d\nu \\ & - \sum_{i=1}^{N} \varepsilon_i \left( \int e^{\frac{f(x)+g(y)-\lambda_i c_i(x,y)}{\varepsilon_i}} d\mu(x) d\nu(y) - 1 \right) \end{aligned} \tag{8.11}$$

*and the infimum of the primal problem is attained.*

As in standard regularized optimal transport there is a link between primal and dual variables at optimum. Let $\gamma^*$ solving the reguralized primal problem and $(f^*, g^*, \lambda^*)$ solving the dual one:

$$\forall i, \ \gamma_i^* = \exp\left( \frac{f^* + g^* - \lambda_i^* c_i}{\varepsilon_i} \right) \cdot \mu \otimes \nu$$

.

### 8.4.2 Proposed Algorithms

We can now present algorithms obtained from entropic relaxation to approximately compute the solution of EOT. Let $\mu = \sum_{i=1}^{n} a_i \delta_{x_i}$ and $\nu = \sum_{j=1}^{m} b_j \delta_{y_j}$ be discrete

**Algorithm 15** Projected Alternating Maximization

**Input:** $\mathbf{C} = (C_i)_{1 \leq i \leq N}$, $a$, $b$, $\varepsilon$, $L_\lambda$

**Init:** $f^0 \leftarrow \mathbf{1}_n$; $g^0 \leftarrow \mathbf{1}_m$; $\lambda^0 \leftarrow (1/N, ..., 1/N) \in \mathbb{R}^N$

**for** $k = 1, 2, ...$ **do**

$K^k \leftarrow \sum_{i=1}^N K_i^{\lambda_i^{k-1}}$,

$c_k \leftarrow \langle f^{k-1}, K^k g^{k-1} \rangle$, $f^k \leftarrow \frac{c_k a}{K^k g^{k-1}}$,

$d_k \leftarrow \langle f^k, K^k g^{k-1} \rangle$, $g^k \leftarrow \frac{d_k b}{(K^k)^T f^k}$,

$\lambda^k \leftarrow \mathrm{Proj}_{\Delta_N^+}\left(\lambda^{k-1} + \frac{1}{L_\lambda} \nabla_\lambda F_{\mathbf{C}}^{\varepsilon}(\lambda^{k-1}, f^k, g^k)\right)$.

**end**

**Result**: $\lambda, f, g$

probability measures where $a \in \Delta_n^+$, $b \in \Delta_m^+$, $\{x_1, ..., x_n\} \subset \mathcal{X}$ and $\{y_1, ..., y_m\} \subset \mathcal{Y}$. Moreover for all $i \in \{1, \ldots, N\}$ and $\lambda > 0$, define $\mathbf{C} := (C_i)_{1 \leq i \leq N} \in (\mathbb{R}^{n \times m})^N$ with $C_i := (c_i(x_k, y_l))_{k,l}$ the $N$ cost matrices and $K_i^\lambda := \exp(-\lambda C_i/\varepsilon)$. Assume that $\varepsilon_1 = \cdots = \varepsilon_N = \varepsilon$. Compared to the standard regularized OT, the main difference here is that the problem contains an additional variable $\lambda \in \Delta_N^+$. When $N = 1$, one can use Sinkhorn algorithm. However when $N \geq 2$, we do not have a closed form for updating $\lambda$ when the other variables of the problem are fixed. In order to enjoy from the strong convexity of the primal formulation, we consider instead the dual associated with the equivalent primal problem given when the additional trivial constraint $\mathbf{1}_n^T \left(\sum_i P_i\right) \mathbf{1}_m = 1$ is considered. In that the dual obtained is

$$\mathrm{EOT}_{\mathbf{c}}^{\boldsymbol{\epsilon}}(\mu, \nu) = \sup_{\substack{\lambda \in \Delta_N^+ \\ f \in \mathbb{R}^n,\ g \in \mathbb{R}^m}} \langle f, a \rangle + \langle g, b \rangle - \varepsilon \left[ \log\left( \sum_i \langle e^{f/\varepsilon}, K_i^{\lambda_i} e^{g/\varepsilon} \rangle \right) + 1 \right]$$

We show that the new objective obtained above is smooth w.r.t $(\lambda, f, g)$. See Appendix 8.8.4 for the proof. One can apply the accelerated projected gradient ascent [236, 237] which enjoys an optimal convergence rate for first order methods of $\mathcal{O}(k^{-2})$ for $k$ iterations.

It is also possible to adapt Sinkhorn algorithm to our problem. See Algorithm 15. We denoted by $\mathrm{Proj}_{\Delta_N^+}$ the orthogonal projection on $\Delta_N^+$ [238], whose complexity is in $\mathcal{O}(N \log N)$. The smoothness constant in $\lambda$ in the algorithm is $L_\lambda = \max_i \|C_i\|_\infty^2/\varepsilon$. In practice Alg. 15 gives better results than the accelerated gradient descent. Note that the proposed algorithm differs from the Sinkhorn algorithm in many points and therefore the convergence rates cannot be applied here. Analyzing the rates of a *projected* alternating maximization method is, to the best of our knowledge, an unsolved problem. Further work will be devoted to study the convergence of this algorithm. We illustrate Algorithm 15 by showing the convergence

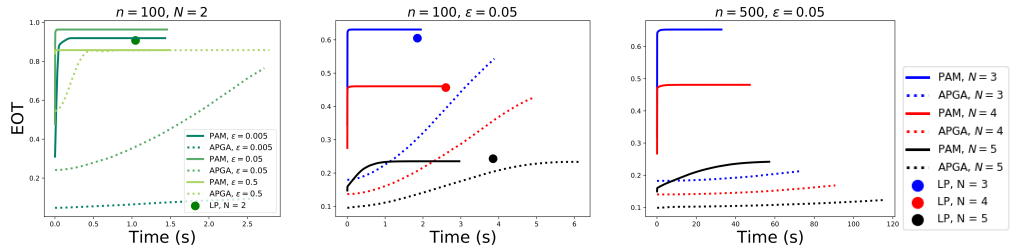



Figure 8.3: Comparison of the time-accuracy tradeoffs between the different proposed algorithms. *Left:* we consider the case where the number of days is $N = 2$, the size of support for both measures is $n = m = 100$ and we vary $\varepsilon$ from 0.005 to 0.5. *Middle:* we fix $n = m = 100$ and the regularization $\varepsilon = 0.05$ and we vary the number of days $N$ from 3 to 5. *Right:* the setting considered is the same as in the figure in the middle, however we increase the sample size such that $n = m = 500$. Note that in that case, **LP** is too costly to be computed.

of the regularized version of EOT towards the ground truth when $\varepsilon \to 0$ in the case of the Dudley Metric. See Figure 8.8 in Appendix 8.9.

## 8.5 Other applications of EOT

**Minimal Transportation Time.** Assume there are $N$ internet service providers who propose different debits to transport data across locations, and one needs to transfer data from multiple servers to others, the fastest as possible. We assume that $c_i(x, y) \geq 0$ corresponds to the transportation time needed by provider $i$ to transport one unit of data from a server $x$ to a server $y$. For instance, the unit of data can be one Megabit. Then $\int c_i d\gamma_i$ corresponds the time taken by provider $i$ to transport $\mu_i = \Pi_{1\sharp}\gamma_i$ to $\nu_i = \Pi_{2\sharp}\gamma_i$. Assuming the transportation can be made in parallel and given a partition of the transportation task $(\gamma_i)_{i=1}^N$, $\max_i \int c_i d\gamma_i$ corresponds to the total time of transport the data $\mu = \Pi_{1\sharp}\sum \gamma_i$ to the locations $\nu = \Pi_{2\sharp}\sum \gamma_i$ according to this partition. Then EOT, which minimizes $\max_i \int c_i d\gamma_i$, is finding the fastest way to transport the data from $\mu$ to $\nu$ by splitting the task among the $N$ internet service providers. Note that at optimality, all the internet service providers finish their transportation task at the same time (see Proposition 8.3.1).

**Sequential Optimal Transport.** Consider the situation where an agent aims to transport goods from some stocks to some stores in the next $N$ days. The cost to transport one unit of good from a stock located at $x$ to a store located at $y$ may vary across the days. For example the cost of transportation may depend on

the price of gas, or the daily weather conditions. Assuming that he or she has a good knowledge of the daily costs of the $N$ coming days, he or she may want a transportation strategy such that his or her daily cost is as low as possible. By denoting $c_i$ the cost of transportation the $i$-th day, and given a strategy $(\gamma_i)_i^N$, the maximum daily cost is then $\max_i \int c_i d\gamma_i$, and EOT therefore finds the cheapest strategy to spread the transport task in the next $N$ days such that the maximum daily cost is minimized. Note that at optimality he or she has to spend the exact same amount everyday.

In Figure 8.3 we aim to simulate the Sequential OT problem and compare the time-accuracy trade-offs of the proposed algorithms. Let us consider a situation where one wants to transport merchandises from $\mu = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ to $\nu = \frac{1}{m}\sum_{j=1}^m \delta_{y_j}$ in $N$ days. Here we model the locations $\{x_i\}$ and $\{y_j\}$ by drawing them independently from two Gaussian distributions in $\mathbb{R}^2$: $\forall i,\ x_i \sim \mathcal{N}\left(\binom{3}{3}, \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)\right)$ and $\forall j,\ y_j \sim \mathcal{N}\left(\binom{4}{4}, \left(\begin{smallmatrix} 1 & -.2 \\ -.2 & 1 \end{smallmatrix}\right)\right)$. We assume that everyday there is wind modeled by a vector $w \sim \mathcal{U}(B(0,1))$ where $B(0,1)$ is the unit ball in $\mathbb{R}^2$ that is perfectly known in advance. We define the cost of transportation on day $i$ as $c_i(x,y) = \|y - x\| - 0.7\langle w_i, y - x\rangle$ to model the effect of the wind on the transportation cost. In the following figures we plot the estimates of EOT obtained from the proposed algorithms in function of the runtime for various sample sizes $n$, number of days $N$ and regularizations $\varepsilon$. **PAM** denotes Alg. 15, **APGA** denotes Alg. 16 (See Appendix C.4), **LP** denotes the linear program which solves exactly the primal formulation of the EOT problem. Note that when **LP** is computable (i.e. $n \leq 100$), it is therefore the ground truth. We show that in all the settings, **PAM** performs better than **APGA** and provides very high accuracy with order of magnitude faster than LP.

# Supplementary material

## 8.6 Proofs

### 8.6.1 Notations

Let $\mathcal{Z}$ be a Polish space, we denote $\mathcal{M}(\mathcal{Z})$ the set of Radon measures on $\mathcal{Z}$ endowed with total variation norm: $\|\mu\|_{\mathrm{TV}} = \mu_+(\mathcal{Z}) + \mu_-(\mathcal{Z})$ with $(\mu_+, \mu_-)$ is the Dunford decomposition of the signed measure $\mu$. We call $\mathcal{M}_+(\mathcal{Z})$ the sets of positive Radon measures, and $\mathcal{P}(\mathcal{Z})$ the set of probability measures. We denote $\mathcal{C}_b(\mathcal{Z})$ the vector space of bounded continuous functions on $\mathcal{Z}$ endowed with $\|\cdot\|_\infty$ norm. We recall the *Riesz-Markov theorem*: if $\mathcal{Z}$ is compact, $\mathcal{M}(\mathcal{Z})$ is the topological dual of $\mathcal{C}_b(\mathcal{Z})$. Let $\mathcal{X}$ and $\mathcal{Y}$ be two Polish spaces. It is immediate that $\mathcal{X} \times \mathcal{Y}$ *is a Polish space*. We denote for $\mu \in \mathcal{M}(\mathcal{X})$ and $\nu \in \mathcal{M}(\mathcal{Y})$, $\mu \otimes \nu$ the tensor product of the measures $\mu$ and $\nu$, and $\mu \ll \nu$ means that $\nu$ dominates $\mu$. We denote $\pi_1 : (x, y) \in \mathcal{X} \times \mathcal{Y} \mapsto x$ and $\pi_2 : (x, y) \in \mathcal{X} \times \mathcal{Y} \mapsto y$ respectively the projections on $\mathcal{X}$ and $\mathcal{Y}$, which are continuous applications. For an application $g$ and a measure $\mu$, we denote $g_\sharp \mu$ the pushforward measure of $\mu$ by $g$. For $f : \mathcal{X} \to \mathbb{R}$ and $g : \mathcal{Y} \to \mathbb{R}$, we denote $f \oplus g : (x, y) \in \mathcal{X} \times \mathcal{Y} \mapsto f(x) + g(y)$ the tensor sum of $f$ and $g$. For $\mathcal{X}$ and $\mathcal{Y}$ two Polish spaces, we denote $\mathrm{LSC}(\mathcal{X} \times \mathcal{Y})$ the space of lower semi-continuous functions on $\mathcal{X} \times \mathcal{Y}$, $\mathrm{LSC}^+(\mathcal{X} \times \mathcal{Y})$ the space of non-negative lower semi-continuous functions on $\mathcal{X} \times \mathcal{Y}$ and $\mathrm{LSC}^-_*(\mathcal{X} \times \mathcal{Y})$ the set of negative bounded below lower semi-continuous functions on $\mathcal{X} \times \mathcal{Y}$ . Let $N \geq 1$ be an integer and denote $\Delta_N^+ := \{\lambda \in \mathbb{R}_+^N \text{ s.t. } \sum_{i=1}^N \lambda_i = 1\}$, the probability simplex of $\mathbb{R}^N$. For two positive measures of same mass $\mu \in \mathcal{M}_+(\mathcal{X})$ and $\nu \in \mathcal{M}_+(\mathcal{Y})$, we define the set of couplings with marginals $\mu$ and $\nu$:

$$\Pi(\mu, \nu) := \{\gamma \text{ s.t. } \pi_{1\sharp}\gamma = \mu \ , \ \pi_{2\sharp}\gamma = \nu\} \ .$$

For $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$, we introduce the subset of $(\mathcal{M}_+(\mathcal{X}) \times \mathcal{M}_+(\mathcal{Y}))^N$ representing marginal decomposition:

$$\Upsilon^N(\mu, \nu) := \left\{(\mu_i, \nu_i)_{i=1}^N \text{ s.t. } \textstyle\sum_i \mu_i = \mu, \sum_i \nu_i = \nu \text{ and } \forall i, \ \mu_i(\mathcal{X}) = \nu_i(\mathcal{Y})\right\}.$$

We also define the following subset of $\mathcal{M}_+(\mathcal{X} \times \mathcal{Y})^N$ corresponding to the coupling decomposition:

$$\Gamma^N(\mu, \nu) := \left\{(\gamma_i)_{i=1}^N \text{ s.t. } \pi_{1\sharp} \sum_i \gamma_i = \mu \ , \ \pi_{2\sharp} \sum_i \gamma_i = \nu \right\}.$$

### 8.6.2 Proof of Proposition 8.3.1

*Proof.* First, it is clear that $\mathrm{EOT}_{\mathbf{c}}(\mu, \nu) \geq \inf_{\gamma \in \Gamma^N(\mu,\nu)}\{t \text{ s.t. } \forall i, \ t = \int c_i d\gamma_i\}$. Let us now show that in fact it is an equality. Thanks to Theorem 8.3.1, the infimum

is attained for $\inf_{\gamma\in\Gamma_{\mu,\nu}} \max_i \int c_i d\gamma_i$. Indeed recall that $\Gamma^N(\mu,\nu)$ is compact and that the objective is lower semi-continuous. Let $\gamma^*$ be such a minimizer. Let $I$ be the set of indices $i$ such that $\int c_i d\gamma_i^* = \mathrm{EOT}_{\mathbf{c}}(\mu,\nu)$. Assume that there exists $j$ such that, $\mathrm{EOT}_{\mathbf{c}}(\mu,\nu) > \int c_j d\gamma_j^*$.

In case of costs of $\mathrm{LSC}^+(\mathcal{X}\times\mathcal{Y})$, for all $i\in I$, there exists $(x_i,y_i)\in \mathrm{Supp}(\gamma_i^*)$ such that $c_i(x_i,y_i)>0$. Let us denote $A_{(x_i,y_i)}$ measurable sets such that $(x_i,y_i)\in A_{(x_i,y_i)}$ and let us denote $\tilde{\gamma}$ defined as for all $k\notin I\cup\{j\}$, $\tilde{\gamma}_k=\gamma_k^*$, for $i\in I$, $\tilde{\gamma}_i = \gamma_i^* - \epsilon \mathbf{1}_{A_{(x_i,y_i)}}\gamma_i^*$ and $\tilde{\gamma}_j = \gamma_j^* + \sum_{i\in I}\epsilon\mathbf{1}_{A_{(x_i,y_i)}}\gamma_i^*$ for $\epsilon$ sufficiently small so that $\tilde{\gamma}\in\Gamma^N(\mu,\nu)$. Now, $\max_k \int c_k d\gamma_k^* > \max_k \int c_k d\tilde{\gamma}_k$, which contradicts that $\gamma^*$ is a minimizer. Then for $i,j$, $\int c_i d\gamma_i^* = \int c_j d\gamma_j^*$. And then: $\mathrm{EOT}_{\mathbf{c}}(\mu,\nu) = \inf_{\gamma\in\Gamma^N(\mu,\nu)}\max_i \int c_i d\gamma_i$.

In case of costs in $\mathrm{LSC}_*^-(\mathcal{X}\times\mathcal{Y})$, there exists $(x_0,y_0)\in\mathrm{Supp}(\gamma_j^*)$ such that $c_j(x_0,y_0)<0$. Let us denote $A_{(x_0,y_0)}$ a measurable set such that $(x_0,y_0)\in A_{(x_0,y_0)}$ and let us denote $\tilde{\gamma}$ defined as for all $k\notin I\cup\{j\}$, $\tilde{\gamma}_k=\gamma_k^*$ and for all $i\in I$, $\tilde{\gamma}_i = \gamma_i^* + \frac{\epsilon}{|I|}\mathbf{1}_{A_{(x_0,y_0)}}\gamma_j^*$ and $\tilde{\gamma}_j = \gamma_j^* - \epsilon\mathbf{1}_{A_{(x_0,y_0)}}\gamma_j^*$ for $\epsilon$ sufficiently small so that $\tilde{\gamma}\in\Gamma^N(\mu,\nu)$. Now, $\max_k\int c_k d\gamma_k^* > \max_k \int c_k d\tilde{\gamma}_i$, which contradicts that $\gamma^*$ is a minimizer. Then for $i,j$, $\int c_i d\gamma_i^* = \int c_j d\gamma_j^*$. And then: $\mathrm{EOT}_{\mathbf{c}}(\mu,\nu) = \inf_{\gamma\in\Gamma^N(\mu,\nu)}\max_i\int c_i d\gamma_i$.

It is clear that equitability is verified thanks to the previous proof. For proportionality, assume the normalization: $\forall i$, there exists $\gamma_i\in\mathcal{P}(\mathcal{X}\times\mathcal{Y})$ such that $V_i(\gamma_i)=1$. Then for each $i$, $V_i(\gamma_i/N)=1/N$ and $(\gamma_i)_i\in\Gamma^N(\mu,\nu)$. Then at optimum: $\forall i$, $V_i(\gamma_i^*)\geq 1/N$ and proportionality is verified.

□

### 8.6.3 Proof of Proposition 8.3.2

*Proof.* We prove along with Theorem 8.3.1 that the infimum defining $\mathrm{EOT}_{\mathbf{c}}(\mu,\nu)$ is attained. Let $\gamma^*$ be this infimum. Then at optimum we have shown that for all $i,j$, $\int c_i d\gamma_i^* = \int c_j d\gamma_j^* = t$. Let denote for all $i$, $\mu_i = \pi_{1\sharp}\gamma_i^*$ and $\nu_i = \pi_{2\sharp}\gamma_i^*$.

Let assume there exists $i$ such that $\int c_i d\gamma_i^* > \mathrm{OT}_{c_i}(\mu_i,\nu_i)$. Let $\gamma_i'$ realising the infimum of $\mathrm{OT}_{c_i}(\mu_i,\nu_i)$. Let $\epsilon>0$ be sufficiently small, then let define $\tilde{\gamma}$ as follows: for all $j\neq i$, $\tilde{\gamma}_j = (1-\epsilon)\gamma_j^*$. and $\tilde{\gamma}_i = \gamma_i' + \epsilon\sum_{j\neq i}\gamma_j^*$. Then for all $j\neq i$, $\int c_j d\tilde{\gamma}_j = (1-\epsilon)t$ and $\int c_i d\tilde{\gamma}_i = \mathrm{OT}_{c_i}(\mu_i,\nu_i) + \epsilon\sum_{j\neq i}\int c_i d\gamma_j^*$. It is clear that $\tilde{\gamma}\in\Gamma^N(\mu,\nu)$. For $\epsilon>0$ sufficiently small, $\max_i\int c_i d\tilde{\gamma}_i = (1-\epsilon)t < t$, which contradicts the optimality of $\gamma^*$.

A possible reformulation for EOT is:

$$\mathrm{EOT}_{\mathbf{c}}(\mu,\nu) = \min_{\substack{(\mu_i,\nu_i)_{i=1}^N\in\Upsilon^N(\mu,\nu)\\ \forall i,\ \gamma_i\in\pi(\mu,\nu)}} \left\{ t \text{ s.t. } \int c_i d\gamma_i = t \right\}$$

We previously show that at optimum the couplings are optimal transport plans, then:

$$\mathrm{EOT}_{\mathbf{c}}(\mu,\nu) = \min_{(\mu_i,\nu_i)_{i=1}^N \in \Upsilon^N(\mu,\nu)} \{t \text{ s.t. } \forall i,\ \mathrm{OT}_{c_i}(\mu_i,\nu_i) = t\}$$

which concludes the proof. □

### 8.6.4 Proof of Theorem 8.3.1

To prove this theorem, one need to prove the three following technical lemmas. The first one shows the weak compacity of $\Gamma^N(\mu,\nu)$.

**Lemma 9.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be Polish spaces, and $\mu$ and $\nu$ two probability measures respectively on $\mathcal{X}$ and $\mathcal{Y}$. Then $\Gamma^N(\mu,\nu)$ is sequentially compact for the weak topology induced by $\|\gamma\| = \max_{i=1,..,N} \|\gamma_i\|_{\mathrm{TV}}$.*

*Proof.* Let $(\gamma^n)_{n\geq 0}$ a sequence in $\Gamma^N(\mu,\nu)$, and let us denote for all $n \geq 0$, $\gamma^n = (\gamma_i^n)_{i=1}^N$. We first remark that for all $i \in \{1,...,N\}$ and $n \geq 0$, $\|\gamma_i^n\|_{\mathrm{TV}} \leq 1$ therefore for all $i \in \{1,...,N\}$, $(\gamma_i^n)_{n\geq 0}$ is uniformly bounded. Moreover as $\{\mu\}$ and $\{\nu\}$ are tight, for any $\delta > 0$, there exist $K \subset \mathcal{X}$ and $L \subset \mathcal{Y}$ compact sets such that

$$\mu(K^c) \leq \frac{\delta}{2} \quad \text{and} \quad \nu(L^c) \leq \frac{\delta}{2}. \tag{8.12}$$

Therefore, we obtain that for any for all $i \in \{1,...,N\}$,

$$\begin{aligned}
\gamma_i^n(K^c \times L^c) &\leq \sum_{k=1}^N \gamma_k^n(K^c \times L^c) && (8.13)\\
&\leq \sum_{k=1}^N \gamma_k^n(K^c \times \mathcal{Y}) + \gamma_k^n(\mathcal{X} \times L^c) && (8.14)\\
&\leq \mu(K^c) + \nu(L^c) = \delta. && (8.15)
\end{aligned}$$

Therefore, for all $i \in \{1,...,N\}$, $(\gamma_i^n)_{n\geq 0}$ is tight and uniformly bounded and Prokhorov's theorem [239, Theorem A.3.15] guarantees for all $i \in \{1,...,N\}$, $(\gamma_i^n)_{n\geq 0}$ admits a weakly convergent subsequence. By extracting a common convergent subsequence, we obtain that $(\gamma^n)_{n\geq 0}$ admits a weakly convergent subsequence. By continuity of the projection, the limit also lives in $\Gamma^N(\mu,\nu)$ and the result follows. □

Next lemma generalizes Rockafellar-Fenchel duality to our case.

**Lemma 10.** *Let $V$ be a normed vector space and $V^*$ its topological dual. Let $V_1, ..., V_N$ be convex functions and lower semi-continuous on $V$ and $E$ a convex function on $V$. Let $V_1^*, ...V_N^*, E^*$ be the Fenchel-Legendre transforms of $V_1, ...V_N, E$. Assume there exists $z_0 \in V$ such that for all $i$, $V_i(z_0) < \infty$, $E(z_0) < \infty$, and for all $i$, $V_i$ is continuous at $z_0$. Then:*

$$\inf_{u \in V} \sum_i V_i(u) + E(u) = \sup_{\substack{\gamma_1 ..., \gamma_N, \gamma \in V^* \\ \sum_i \gamma_i = \gamma}} - \sum_i V_i^*(-\gamma_i) - E^*(\gamma)$$

*Proof.* This Lemma is an immediate application of Rockafellar-Fenchel duality theorem [240, Theorem 1.12] and of Fenchel-Moreau theorem [240, Theorem 1.11]. Indeed, $V = \sum_{i=1}^{N} V_i(u)$ is a convex function, lower semi-continuous and its Legendre-Fenchel transform is given by:

$$V^*(\gamma^*) = \inf_{\sum_{i=1}^{N} \gamma_i^* = \gamma^*} \sum_{i=1}^{N} V_i^*(\gamma_i^*). \tag{8.16}$$

□

Last lemma is an application of Sion's Theorem to this problem.

**Lemma 11.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be Polish spaces. Let $\mathbf{c} = (c_i)_{1 \leq i \leq N}$ be a family of bounded below lower semi-continuous costs on $\mathcal{X} \times \mathcal{Y}$, then for $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$, we have*

$$\mathrm{EOT}_{\mathbf{c}}(\mu, \nu) = \sup_{\lambda \in \Delta_N^+} \inf_{\gamma \in \Gamma_{\mu,\nu}^N} \sum_{i=1}^{N} \lambda_i \int_{\mathcal{X} \times \mathcal{Y}} c_i(x, y) d\gamma_i(x, y) \tag{8.17}$$

*and the infimum is attained.*

*Proof.* Taking for granted that a minmax principle can be invoked, we have

$$\begin{aligned} \sup_{\lambda \in \Delta_N^+} \inf_{\gamma \in \Gamma_{\mu,\nu}^N} \sum_{i=1}^{N} \lambda_i \int_{\mathcal{X} \times \mathcal{Y}} c_i(x, y) d\gamma_i(x, y) &= \inf_{\gamma \in \Gamma_{\mu,\nu}^N} \sup_{\lambda \in \Delta_N^+} \sum_{i=1}^{N} \lambda_i \int_{\mathcal{X} \times \mathcal{Y}} c_i(x, y) d\gamma_i(x, y) \\ &= \mathrm{EOT}_{\mathbf{c}}(\mu, \nu) \end{aligned}$$

But thanks to Lemma 9, we have that $\Gamma_{\mu,\nu}^N$ is compact for the weak topology. And $\Delta_N^+$ is convex. Moreover the objective function $f : (\lambda, \gamma) \in \Delta_N^+ \times \Gamma^N(\mu, \nu) \mapsto \sum_{i=1}^{N} \lambda_i \int_{\mathcal{X} \times \mathcal{Y}} c_i^n d\gamma_i$ is bilinear, hence convex and concave in its variables, and continuous with respect to $\lambda$. Moreover, let $(c_i^n)_n$ be non-decreasing sequences of

bounded cost functions such that $c_i = \sup_n c_i^n$. By monotone convergence, we get $f(\lambda, \gamma) = \sup_n \sum_i \lambda_i \int c_i^n d\gamma_i$, $f(\lambda, .)$. So $f$ the supremum of continuous functions, then $f$ is lower semi-continuous with respect to $\gamma$, therefore Sion's minimax theorem [241] holds.

□

We are now able to prove Theorem 8.3.1.

*Proof.* Let $\mathcal{X}$ and $\mathcal{Y}$ be two Polish spaces. For all $i \in \{1, .., N\}$, we define $c_i : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ a bounded below lower-semi cost function. The proof follows the exact same steps as those in the proof of [30, Theorem 1.3]. First we suppose that $\mathcal{X}$ and $\mathcal{Y}$ are compact and that for all $i$, $c_i$ is continuous, then we show that it can be extended to $X$ and $Y$ non compact and finally to $c_i$ only lower semi continuous.

First, let assume $\mathcal{X}$ and $\mathcal{Y}$ are compact and that for all $i$, $c_i$ is continuous. Let fix $\lambda \in \Delta_N^+$. We recall the topological dual of the space of bounded continuous functions $\mathcal{C}_b(\mathcal{X} \times \mathcal{Y})$ endowed with $\|.\|_\infty$ norm, is the space of Radon measures $\mathcal{M}(\mathcal{X} \times \mathcal{Y})$ endowed with total variation norm. We define, for $u \in \mathcal{C}_b(\mathcal{X} \times \mathcal{Y})$:

$$V_i^\lambda(u) = \begin{cases} 0 & \text{if} \quad u \geq -\lambda_i c_i \\ +\infty & \text{else} \end{cases}$$

and:

$$E(u) = \begin{cases} \int f d\mu + \int g d\nu & \text{if} \quad \exists (f, g) \in \mathcal{C}_b(\mathcal{X}) \times \mathcal{C}_b(\mathcal{Y}),\ u = f + g \\ +\infty & \text{else} \end{cases}$$

One can show that for all $i$, $V_i^\lambda$ is convex and lower semi-continuous (as the sublevel sets are closed) and $E^\lambda$ is convex. More over for all $i$, these functions continuous in $u_0 \equiv 1$ the hypothesis of Lemma 10 are satisfied.

Let now compute the Fenchel-Legendre transform of these function. Let $\gamma \in \mathcal{M}(\mathcal{X} \times \mathcal{Y})$ :

$$\begin{aligned} V_i^{\lambda *}(-\gamma) &= \sup_{u \in \mathcal{C}_b(\mathcal{X} \times \mathcal{Y})} \left\{ -\int u d\gamma; \quad u \geq -\lambda_i c_i \right\} \\ &= \begin{cases} \int \lambda_i c_i d\gamma & \text{if} \quad \gamma \in \mathcal{M}_+(\mathcal{X} \times \mathcal{Y}) \\ +\infty & \text{otherwise} \end{cases} \end{aligned}$$

On the other hand:

$$E^{\lambda *}(\gamma) = \begin{cases} 0 & \text{if} \quad \forall (f, g) \in \mathcal{C}_b(\mathcal{X}) \times \mathcal{C}_b(\mathcal{Y}),\ \int f d\mu + \int g d\nu = \int (f + g) d\gamma \\ +\infty & \text{else} \end{cases}$$

This dual function is finite and equals 0 if and only if that the marginals of the dual variable $\gamma$ are $\mu$ and $\nu$.

Applying Lemma 10, we get:

$$\inf_{u\in\mathcal{C}_b(\mathcal{X}\times\mathcal{Y})}\sum_i V_i^\lambda(u)+E(u)=\sup_{\substack{\gamma_1,\dots,\gamma_N,\gamma\in\mathcal{M}(\mathcal{X}\times\mathcal{Y})\\ \sum\gamma_i=\gamma}}\sum -V_i^{\lambda*}(\gamma_i)-E^{\lambda*}(-\gamma)$$

Hence, we have shown that, when $\mathcal{X}$ and $\mathcal{Y}$ are compact sets, and the costs $(c_i)_i$ are continuous:

$$\sup_{(f,g)\in\mathcal{F}_{\mathbf{c}}^\lambda}\int f d\mu+\int g d\nu=\inf_{\gamma\in\Gamma^N(\mu,\nu)}\sum_i\lambda_i\int c_i d\gamma_i$$

Let now prove the result holds when the spaces $\mathcal{X}$ and $\mathcal{Y}$ are not compact. We still suppose that for all $i$, $c_i$ is uniformly continuous and bounded. We denote $\|\mathbf{c}\|_\infty:=\sup_i\sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}}|c_i(x,y)|$. Let define $I^\lambda(\gamma):=\sum_i\lambda_i\int_{\mathcal{X}\times\mathcal{Y}}c_i d\gamma_i$

Let $\gamma^*\in\Gamma^N(\mu,\nu)$ such that $I^\lambda(\gamma^*)=\min_{\gamma\in\Gamma^N(\mu,\nu)}I^\lambda(\gamma)$. The existence of the minimum comes from the lower-semi continuity of $I^\lambda$ and the compacity of $\Gamma^N(\mu,\nu)$ for weak topology.

Let fix $\delta\in(0,1)$. $\mathcal{X}$ and $\mathcal{Y}$ are Polish spaces then $\exists\mathcal{X}_0\subset\mathcal{X},\mathcal{Y}_0\subset\mathcal{Y}$ compacts such that $\mu(\mathcal{X}_0^c)\le\delta$ and $\mu(\mathcal{Y}_0^c)\le\delta$. It follows that $\forall i$, $\gamma_i^*((\mathcal{X}_0\times\mathcal{Y}_0)^c)\le 2\delta$. Let define $\gamma^{*0}$ such that for all $i$, $\gamma_i^{*0}=\frac{\mathbf{1}_{\mathcal{X}_0\times\mathcal{Y}_0}}{\sum_i\gamma_i^*(\mathcal{X}_0\times\mathcal{Y}_0)}\gamma_i^*$. We define $\mu_0=\pi_1\#\sum_i\gamma_i^{*0}$ and $\nu_0=\pi_2\#\sum_i\gamma_i^{*0}$. We then naturally define

$$\Gamma_0^N(\mu_0,\nu_0):=\left\{(\gamma_i)_{1\le i\le N}\in\mathcal{M}_+(\mathcal{X}_0\times\mathcal{Y}_0)^N\text{ s.t. }\pi_1\#\sum_i\gamma_i=\mu_0\text{ and }\pi_2\#\sum_i\gamma_i=\nu_0\right\}$$

and

$$I_0^\lambda(\gamma_0):=\sum_i\lambda_i\int_{\mathcal{X}_0\times\mathcal{Y}_0}c_i d\gamma_{0,i}\quad\text{for}\gamma_0\in\Gamma_0^N(\mu_0,\nu_0)$$

.

Let $\tilde{\gamma}_0$ verifying $I_0^\lambda(\tilde{\gamma}_0)=\min_{\gamma_0\in\Gamma_0^N(\mu_0,\nu_0)}I_0^\lambda(\gamma_0)$. Let $\tilde{\gamma}=(\sum_i\gamma_i^*(\mathcal{X}_0\times\mathcal{Y}_0))\,\tilde{\gamma}_0+\mathbf{1}_{(\mathcal{X}_0\times\mathcal{Y}_0)^c}\gamma^*\in\Gamma^N(\mu,\nu)$. Then we get

$$I^\lambda(\tilde{\gamma})\le\min_{\gamma_0\in\Gamma^N_{0,\mu_0,\nu_0}}I_0^\lambda(\gamma_0)+2\sum|\lambda_i|\|\mathbf{c}\|_\infty\delta$$

We have already proved that:

$$\sup_{(f,g)\in\mathcal{F}_{0,\mathbf{c}}^\lambda}J_0^\lambda(f,g)=\inf_{\gamma_0\in\Gamma^N_{0,\mu_0,\nu_0}}I_0^\lambda(\gamma_0)$$

with $J_0^\lambda(f,g) = \int f d\mu_0 + \int g d\nu_0$ and $\mathcal{F}_{0,\mathbf{c}}^\lambda$ is the set of $(f,g) \in \mathcal{C}_b(\mathcal{X}_0) \times \mathcal{C}_b(\mathcal{Y}_0)$ satisfying, for every $i$, $f \oplus g \leq \min_i \lambda_i c_i$. Let $(\tilde{f}_0, \tilde{g}_0) \in \mathcal{F}_{\mathbf{0},\mathbf{c}}^\lambda$ such that :

$$J_0^\lambda(\tilde{f}_0, \tilde{g}_0) \geq \sup_{(f,g)\in\mathcal{F}_{0,\mathbf{c}}^\lambda} J_0^\lambda(f,g) - \delta$$

Since $J_0^\lambda(0,0) = 0$, we get $\sup J_0^\lambda \geq 0$ and then, $J_0^\lambda(\tilde{f}_0, \tilde{g}_0) \geq \delta \geq -1$. For every $\gamma_0 \in \Gamma_{0,\mu_0,\nu_0}^N$:

$$J_0^\lambda(\tilde{f}_0, \tilde{g}_0) = \int (\tilde{f}_0(x) + \tilde{g}_0(y)) d\gamma_0(x,y)$$

then we have the existence of $(x_0, y_0) \in \mathcal{X}_0 \times \mathcal{Y}_0$ such that : $\tilde{f}_0(x_0) + \tilde{g}_0(y_0) \geq -1$. If we replace $(\tilde{f}_0, \tilde{g}_0)$ by $(\tilde{f}_0 - s, \tilde{g}_0 + s)$ for an accurate $s$, we get that: $\tilde{f}_0(x_0) \geq \frac{1}{2}$ and $\tilde{g}_0(y_0) \geq \frac{1}{2}$, and then $\forall (x,y) \in \mathcal{X}_0 \times \mathcal{Y}_0$:

$$\begin{aligned}
\tilde{f}_0(x) &\leq c'(x, y_0) - \tilde{g}_0(y_0) \leq c'(x, y_0) + \frac{1}{2} \\
\tilde{g}_0(y) &\leq c'(x_0, y) - \tilde{f}_0(x_0) \leq c'(x_0, y) + \frac{1}{2}
\end{aligned}$$

where $c' := \min_i \lambda_i c_i$. Let define $\bar{f}_0(x) = \inf_{y\in\mathcal{Y}_0} c'(x,y) - \tilde{g}_0(y)$ for $x \in \mathcal{X}$. Then $\tilde{f}_0 \leq \bar{f}_0$ on $\mathcal{X}_0$. We then get $J_0^\lambda(\bar{f}_0, \tilde{g}_0) \geq J_0^\lambda(\tilde{f}_0, \tilde{g}_0)$ and $\bar{f}_0 \leq c'(., y_0) + \frac{1}{2}$ on $\mathcal{X}$. Let define $\bar{g}_0(y) = \inf_{x\in\mathcal{X}} c'(x,y) - \bar{f}_0(y)$. By construction $(f_0, g_0) \in \mathcal{F}_{\mathbf{c}}^\lambda$ since the costs are uniformly continuous and bounded and $J_0^\lambda(\bar{f}_0, \bar{g}_0) \geq J_0^\lambda(\bar{f}_0, \tilde{g}_0) \geq J_0^\lambda(\tilde{f}_0, \tilde{g}_0)$. We also have $\bar{g}_0 \geq c'(x_0, .) + \frac{1}{2}$ on $\mathcal{Y}$. Then we have in particular: $\bar{g}_0 \geq -\|\mathbf{c}\|_\infty - \frac{1}{2}$ on $\mathcal{X}$ and $\bar{f}_0 \geq -\|\mathbf{c}\|_\infty - \frac{1}{2}$ on $\mathcal{Y}$. Finally:

$$
\begin{aligned}
J^\lambda(\bar f_0,\bar g_0) &:= \int_{\mathcal{X}_0} \bar f d\mu_0 + \int_{\mathcal{Y}_0} \bar g_0 d\nu \\
&= \sum_i \gamma_i^*(\mathcal{X}_0\times\mathcal{Y}_0)\int_{\mathcal{X}_0\times\mathcal{Y}_0} (\bar f_0(x)+\bar g_0(y))d\left(\sum_i \gamma_i^{*0}(x,y)\right) \\
&+ \int_{(\mathcal{X}_0\times\mathcal{Y}_0)^c} \bar f_0(x)+\bar g_0(y) d\left(\sum_i \gamma_i^*(x,y)\right) \\
&\geq (1-2\delta)\left(\int_{\mathcal{X}_0}\bar f_0 d\mu_0 + \int_{\mathcal{Y}_0}\bar g_0 d\nu_0\right) - (2\|\mathbf{c}\|_\infty+1)\sum_i \gamma^*((\mathcal{X}_0\times\mathcal{Y}_0)^c) \\
&\geq (1-2\delta)J_0^\lambda(\bar f_0,\bar g_0) - 2\sum|\lambda_i|(2\|\mathbf{c}\|_\infty+1)\delta \\
&\geq (1-2\delta)J_0^\lambda(\tilde f_0,\tilde g_0) - 2\sum|\lambda_i|(2\|\mathbf{c}\|_\infty+1)\delta \\
&\geq (1-2\delta)(\inf I_0^\lambda - \delta) - 2\sum|\lambda_i|(2\|\mathbf{c}\|_\infty+1)\delta \\
&\geq (1-2\delta)(\inf I^\lambda - (2\sum|\lambda_i|\|\mathbf{c}\|_\infty+1)\delta) - 2\sum|\lambda_i|(2\|\mathbf{c}\|_\infty+1)\delta
\end{aligned}
$$

This being true for arbitrary small $\delta$, we get $\sup J^\lambda \geq \inf I^\lambda$. The other sens is always true then:

$$
\sup_{(f,g)\in\mathcal{F}_{\mathbf{c}}^\lambda} \int f d\mu + \int g d\nu = \inf_{\gamma\in\Gamma^N(\mu,\nu)} \sum_i \lambda_i \int c_i d\gamma_i
$$

for $c_i$ uniformly continuous and $\mathcal{X}$ and $\mathcal{Y}$ non necessarily compact.

Let now prove that the result holds for lower semi-continuous costs. Let $\mathbf{c} := (c_i)_i$ be a collection of lower semi-continuous costs. Let $(c_i^n)_n$ be non-decreasing sequences of bounded below cost functions such that $c_i = \sup_n c_i^n$. Let fix $\lambda\in\Delta_N^+$. From last step, we have shown that for all $n$:

$$
\inf_{\gamma\in\Gamma^N(\mu,\nu)} I_n^\lambda(\gamma) = \sup_{(f,g)\in\mathcal{F}_{\mathbf{c}^n}^\lambda} \int f d\mu + \int g d\nu \tag{8.18}
$$

where $I_n^\lambda(\gamma) = \sum_i \lambda_i \int c_i^n d\gamma_i$. First it is clear that:

$$
\sup_{(f,g)\in\mathcal{F}_{\mathbf{c}}^\lambda} \int f d\mu + \int g d\nu \leq \sup_{(f,g)\in\mathcal{F}_{\mathbf{c}^n}^\lambda} \int f d\mu + \int g d\nu \tag{8.19}
$$

Let show that:

$$
\inf_{\gamma\in\Gamma^N(\mu,\nu)} I^\lambda(\gamma) = \sup_n \inf_{\gamma\in\Gamma^N(\mu,\nu)} I_n^\lambda(\gamma) = \lim_n \inf_{\gamma\in\Gamma^N(\mu,\nu)} I_n^\lambda(\gamma)
$$

where $I^\lambda(\gamma) = \sum_i \lambda_i \int c_i d\gamma_i$.

Let $(\gamma^{n,k})_k$ a minimizing sequence of $\Gamma^N(\mu,\nu)$ for the problem $\inf_{\gamma\in\Gamma^N(\mu,\nu)} \sum_i \lambda_i \int c_i^n d\gamma_i$. By Lemma 9, up to an extraction, there exists $\gamma^n \in \Gamma^N(\mu,\nu)$ such that $(\gamma^{n,k})_k$ converges weakly to $\gamma^n$. Then:

$$\inf_{\gamma\in\Gamma^N(\mu,\nu)} I_n^\lambda(\gamma) = I_n^\lambda(\gamma^n)$$

Up to an extraction, there also exists $\gamma^* \in \Gamma^N(\mu,\nu)$ such that $\gamma^n$ converges weakly to $\gamma^*$. For $n \geq m$, $I_n^\lambda(\gamma^n) \geq I_m^\lambda(\gamma^n) \geq I_m^\lambda(\gamma^m)$, so by continuity of $I_m^\lambda$:

$$\lim_n I_n^\lambda(\gamma^n) \geq \limsup_n I_m^\lambda(\gamma^n) \geq I_m^\lambda(\gamma^*)$$

By monotone convergence, $I_m^\lambda(\gamma^*) \to I^\lambda(\gamma^*)$ and $\lim_n I_n^\lambda(\gamma_n) \geq I^\lambda(\gamma^*) \geq \inf_{\gamma\in\Gamma^N(\mu,\nu)} I^\lambda(\gamma)$.

Along with Eqs. 8.18 and 8.19, we get that:

$$\inf_{\gamma\in\Gamma^N(\mu,\nu)} I^\lambda(\gamma) \leq \sup_{(f,g)\in\mathcal{F}_{\mathbf{c}}^\lambda} \int f d\mu + \int g d\nu$$

The other sens being always true, we have then shown that, in the general case we still have:

$$\inf_{\gamma\in\Gamma^N(\mu,\nu)} I^\lambda(\gamma) = \sup_{(f,g)\in\mathcal{F}_{\mathbf{c}}^\lambda} \int f d\mu + \int g d\nu$$

To conclude, we apply Lemma 11, and we get:

$$\begin{aligned}\sup_{\lambda\in\Delta_N^+} \sup_{(f,g)\in\mathcal{F}_{\mathbf{c}}^\lambda} \int f d\mu + \int g d\nu &= \sup_{\lambda\in\Delta_N^+} \inf_{\gamma\in\Gamma^N(\mu,\nu)} I^\lambda(\gamma) \\ &= \mathrm{EOT}_{\mathbf{c}}(\mu,\nu)\end{aligned}$$

□

### 8.6.5 Proof of Proposition 8.3.3

*Proof.* Let recall that, from standard optimal transport results:

$$\mathrm{EOT}_{\mathbf{c}}(\mu,\nu) = \sup_{u\in\Phi_{\mathbf{c}}} \int u d\mu d\nu$$

with $\Phi_{\mathbf{c}} := \left\{u \in \mathcal{C}_b(\mathcal{X}\times\mathcal{Y}) \text{ s.t. } \exists\lambda\in\Delta_N^+,\ \exists\phi\in\mathcal{C}_b(\mathcal{X}),\ u = \phi^{cc}\oplus\phi^c \text{ with } c = \min_i \lambda_i c_i\right\}$ where $\phi^c$ is the $c$-transform of $\phi$, i.e. for $y\in\mathcal{Y}$, $\phi^c(y) = \inf_{x\in\mathcal{X}} c(x,y) - \phi(x)$.

Let denote $\omega_1, \ldots, \omega_N$ the continuity modulii of $c_1, ..., c_N$. The existence of continuity modulii is ensured by the uniform continuity of $c_1, \ldots, c_N$ on the compact sets $\mathcal{X} \times \mathcal{Y}$ (Heine's theorem). Then a modulus of continuity for $\min_i \lambda_i c_i$ is $\sum_i \lambda_i \omega_i$. As $\phi^c$ and $\phi^{cc}$ share the same modulus of continuity than $c = \min_i \lambda_i c_i$, for $u$ is $\Phi_{\mathbf{c}}$, a common modulus of continuity is $2 \times \sum_i \omega_i$. More over, it is clear that for all $x, y$, $\{u(x, y) \text{ s.t. } u \in \Phi_c\}$ is compact. Then, applying Ascoli's theorem, we get, that $\Phi_{\mathbf{c}}$ is compact for $\|.\|_\infty$ norm. By continuity of $u \to \int u d\mu d\nu$, the supremum is attained, and we get the existence of the optimum $u^*$. The existence of optima $(\lambda^*, f^*, g^*)$ immediately follows.

Let first assume that $(\gamma_k)_{k=1}^N$ is a solution of Eq. (8.1) and $(\lambda, f, g)$ is a solution of Eq. (8.5). Then it is clear that for all $i, j$, $f \oplus g \leq \lambda_i c_i$, $(\gamma_k)_{k=1}^N \in \Gamma^N(\mu, \nu)$ and $\int c_j d\gamma_j = \int c_i d\gamma_i$ (by Proposition 8.3.1). Let $k \in \{1, \ldots, N\}$. Moreover, by Theorem 8.3.1:

$$
\begin{aligned}
0 &= \int f d\mu + \int g d\nu - \int c_i d\gamma_i \\
&= \sum \int (f(x) + g(y)) d\gamma_i(x, y) - \sum_i \lambda_i \int c_i(x, y) d\gamma_i(x, y) \\
&= \sum \int (f(x) + g(y) - \lambda_i c_i(x, y)) d\gamma_i(x, y)
\end{aligned}
$$

Since $f \oplus g \leq \lambda_i c_i$ and $\gamma_i$ are positive measures then $f \oplus g = \lambda_i c_i$, $\gamma_i$-almost everywhere.

Reciprocally, let assume that there exist $(\gamma_k)_{k=1}^N \in \Gamma^N(\mu, \nu)$ and $(\lambda, f, g) \in \Delta_n^+ \times \mathcal{C}_b(\mathcal{X}) \times \mathcal{C}_b(\mathcal{Y})$ such that $\forall i \in \{1, ..., N\}$, $f \oplus g \leq \lambda_i c_i$, $\forall i, j \in \{1, ..., N\}$ $\int c_i d\gamma_i = \int c_j d\gamma_j$ and $f \oplus g = \lambda_i c_i$ $\gamma_i$-a.e.. Then, for any $k$:

$$
\begin{aligned}
\int c_k d\gamma_k &= \sum_i \lambda_i \int c_i d\gamma_i \\
&= \sum_i \int (f(x) + g(y)) d\gamma_i(x, y) \\
&= \int f(x) d\mu(x) + \int g(y) d\nu(y) \\
&\leq \mathrm{EOT}_{\mathbf{c}}(\mu, \nu) \text{ by Theorem 8.3.1}
\end{aligned}
$$

then $\gamma_k$ is solution of the primal problem. We also have for any $k$:

$$\begin{aligned}\int f d\mu + \int g d\nu &= \sum_i \int (f(x) + g(y)) d\gamma_i(x,y) \\ &= \sum_i \int \lambda_i c_i d\gamma_i \\ &= \int c_k d\gamma_k \\ &\geq \mathrm{EOT}_{\mathbf{c}}(\mu,\nu)\end{aligned}$$

□

then, thanks to Theorem 8.3.1, $(\lambda, f, g)$ is solution of the dual problem.

Let now proof the result stated in Remark 11. Let assume the costs are strictly positive or strictly negative. If there exist $i$ such that $\lambda_i = 0$, thanks to the condition $f \oplus g \leq \lambda_i c_i$, we get $f \oplus g \leq 0$ and then $f \oplus g = 0$ which contradicts the conditions $f \oplus g = \lambda_k c_k$ for all $k$.

### 8.6.6 Proof of Proposition 8.3.4

Before proving the result let us first introduce the following lemma.

**Lemma 12.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be Polish spaces. Let $\mathbf{c} := (c_i)_{1\leq i\leq N}$ a family of bounded below continuous costs. For $(x,y) \in \mathcal{X}\times\mathcal{Y}$ and $\lambda \in \Delta_N^+$, we define*

$$c_\lambda(x,y) := \min_{i=1,\dots,N} (\lambda_i c_i(x,y))$$

*then for any $(\mu,\nu) \in \mathcal{P}(\mathcal{X})\times\mathcal{P}(\mathcal{Y})$*

$$\mathrm{EOT}_{\mathbf{c}}(\mu,\nu) = \sup_{\lambda\in\Delta_N^+} OT_{c_\lambda}(\mu,\nu) \tag{8.20}$$

*Proof.* Let $(\mu,\nu) \in \mathcal{P}(\mathcal{X})\times\mathcal{P}(\mathcal{Y})$ and $\mathbf{c} := (c_i)_{1\leq i\leq N}$ cost functions on $\mathcal{X}\times\mathcal{Y}$. Let $\lambda \in \Delta_N^+$, then by Proposition 8.3.1:

$$\mathrm{EOT}_{\mathbf{c}}(\mu,\nu) = \sup_{\lambda\in\Delta_N^+} \sup_{(f,g)\in\mathcal{F}_{\mathbf{c}}^\lambda} \int_{\mathcal{X}} f(x)d\mu(x) + \int_{\mathcal{Y}} g(y)d\nu(y)$$

Therefore by denoting $c_\lambda := \min_i(\lambda_i c_i)$ which is a continuous. The dual form of the classical Optimal Transport problem gives that:

$$\sup_{(f,g)\in\mathcal{F}_{\mathbf{c}}^\lambda} \int_{\mathcal{X}} f(x)d\mu(x) + \int_{\mathcal{Y}} g(y)d\nu(y) = \mathrm{OT}_{c_\lambda}(\mu,\nu)$$

and the result follows. □

Let us now prove the result of Proposition 8.3.4.

*Proof.* Let $\mu$ and $\nu$ be two probability measures. Let $\alpha \in (0,1]$. Note that if $d$ is a metric then $d^\alpha$ too. Therefore in the following we consider $d$ a general metric on $\mathcal{X} \times \mathcal{X}$. Let $c_1 : (x,y) \to 2 \times \mathbf{1}_{x \neq y}$ and $c_2 = d^\alpha$. For all $\lambda \in [0,1)$:

$$c_\lambda(x,y) := \min(\lambda c_1(x,y), (1-\lambda)c_2(x,y)) = \min(2\lambda, (1-\lambda)d(x,y))$$

defines a distance on $\mathcal{X} \times \mathcal{X}$. Then according to [30, Theorem 1.14]:

$$\mathrm{OT}_{c_\lambda}(\mu,\nu) = \sup_{f \text{ s.t. } f \text{ } 1-c_\lambda \text{ Lipschitz}} \int f d\mu - \int f d\nu$$

Then thanks to Lemma 12 we have

$$\mathrm{EOT}_{(c_1,c_2)}(\mu,\nu) = \sup_{\lambda \in [0,1], f \text{ s.t. } f \text{ } 1-c_\lambda \text{ Lipschitz}} \int f d\mu - \int f d\nu$$

Let now prove that in this case: $\mathrm{EOT}_{(c_1,c_2)}(\mu,\nu) = \beta_d(\mu,\nu)$. Let $\lambda \in [0,1)$ and $f$ a $c_\lambda$ Lipschitz function. $f$ is lower bounded: let $m = \inf f$ and $(u_n)_n$ a sequence satisfying $f(u_n) \to m$. Then for all $x, y$, $f(x) - f(y) \leq 2\lambda$ and $f(x) - f(y) \leq (1-\lambda)d(x,y)$. Let define $g = f - m - \lambda$. For $x$ fixed and for all $n$, $f(x) - f(u_n) \leq 2\lambda$, so taking the limit in $n$ we get $f(x) - m \leq 2\lambda$. So we get that for all $x, y$, $g(x) \in [-\lambda, +\lambda]$ and $g(x) - g(y) \in [-(1-\lambda)d(x,zy), (1-\lambda)d(x,y)]$. Then $, g_{,\infty} \leq \lambda$ and $, g_{,d} \leq 1-\lambda$. By construction, we also have $\int f d\mu - \int f d\nu = \int g d\mu - \int g d\nu$.Then $, g_{,\infty} +, g_{,d} \leq 1$. So we get that $\mathrm{EOT}_{(c_1,c_2)}(\mu,\nu) \leq \beta_d(\mu,\nu)$. Reciprocally, let $g$ be a function satisfying $, g_{,\infty} +, g_{,d} \leq 1$. Let define $f = g+, g_{,\infty}$ and $\lambda =, g_{,\infty}$. Then, for all $x, y$, $f(x) \in [0, 2\lambda]$ and so $f(x) - f(y) \leq 2\lambda$. It is immediate that $f(x) - f(y) \in [-(1-\lambda)d(x,y), (1-\lambda)d(x,y)]$. Then we get $f(x) - f(y) \leq \min(\lambda, (1-\lambda)d(x,y))$. And by construction, we still have $\int f d\mu - \int f d\nu = \int g d\mu - \int g d\nu$. So $\mathrm{EOT}_{(c_1,c_2)}(\mu,\nu) \geq \beta_d(\mu,\nu)$.

Finally we get $\mathrm{EOT}_{(c_1,c_2)}(\mu,\nu) = \beta_d(\mu,\nu)$ when $c_1 : (x,y) \to 2 \times \mathbf{1}_{x \neq y}$ and $c_2 = d$ a distance on $\mathcal{X} \times \mathcal{X}$. □

### 8.6.7 Proof of Proposition 8.3.5

**Lemma 13.** *Let $x_1, \ldots, x_N \geq 0$, then:*

$$\sup_{\lambda \in \Delta_N^+} \min_i \lambda_i x_i = \frac{1}{\sum_i \frac{1}{x_i}}$$

*Proof.* First if there exists $i$ such that $x_i = 0$, we immediately have $\sup_{\lambda \in \Delta_N^+} \min_i \lambda_i x_i = 0$.

$g : \lambda \mapsto \min_i \lambda_i x_i$ is a continuous function on the compact set $\lambda \in \Delta_N^+$. Let denote $\lambda^*$ the maximum of $g$.
Let show that for all $i, j$, $\lambda_i^* x_i = \lambda_j^* x_j$. Let denote $i_0, \dots, i_k$ the indices such that $\lambda_{i_l}^* x_{i_l} = \min_i \lambda_i^* x_i$. Let assume there exists $j_0$ such that: $\lambda_{j_0}^* x_{j_0} > \min_i \lambda_i^* x_i$, and that all other indices $i$ have a larger $\lambda_i^* x_i \geq \lambda_{j_0}^* x_{j_0}$. Then for $\epsilon > 0$ sufficiently small, let $\tilde{\lambda}$ defined as: $\tilde{\lambda}_{j_0} = \lambda_{j_0}^* - \epsilon$, $\tilde{\lambda}_{i_l} = \lambda_{i_l}^* + \epsilon/k$ for all $l \in \{1, \dots, k\}$ and $\tilde{\lambda}_i = \lambda_i^*$ for all other indices. Then $\tilde{\lambda} \in \Delta_N^+$ and $g(\lambda^*) < g(\tilde{\lambda})$, which contradicts that $\lambda^*$ is the maximum.
Then at the optimum for all $i, j$, $\lambda_i^* x_i = \lambda_j^* x_j$. So $\lambda_i^* x_i = C$ for a certain constant $C$. Moreover $\sum_i \lambda_i^* = 1$. Then $1/C = \sum_i 1/x_i$. Finally, for all $i$,

$$\lambda_i^* = \frac{1/x_i}{\sum_i 1/x_i}$$

and then:

$$\sup_{\lambda \in \Delta_N^+} \min_i \lambda_i x_i = \frac{1}{\sum_i \frac{1}{x_i}}.$$

□

*Proof.* Let $\mu$ and $\nu$ be two probability measures respectively on $\mathcal{X}$ and $\mathcal{Y}$. Let $\mathbf{c} := (c_i)_i$ be a family of cost functions. Let define for $\lambda \in \Delta_N^+$, $c_\lambda(x, y) := \min_i(\lambda_i c_i(x, y))$. We have, by linearity $\mathrm{OT}_{c_\lambda}(\mu, \nu) \leq \min_i(\lambda_i \mathrm{OT}_{c_i}(\mu, \nu))$. So we deduce by Lemma 12:

$$\begin{aligned}
\mathrm{EOT}_{\mathbf{c}}(\mu, \nu) &= \sup_{\lambda \in \Delta_N^+} \mathrm{OT}_{c_\lambda}(\mu, \nu) \\
&\leq \sup_{\lambda \in \Delta_N^+} \min_i \lambda_i \mathrm{OT}_{c_i}(\mu, \nu) \\
&= \frac{1}{\sum_i \frac{1}{\mathrm{OT}_{c_i}(\mu, \nu)}} \text{ by Lemma 13}
\end{aligned}$$

which concludes the proof. □

### 8.6.8 Proof of Theorem 8.4.1

*Proof.* To show the strong duality of the regularized problem, we use the same sketch of proof as for the strong duality of the original problem. Let first assume

that, for all $i$, $c_i$ is continuous on the compact set $\mathcal{X} \times \mathcal{Y}$. Let fix $\lambda \in \Delta_N^+$. We define, for all $u \in \mathcal{C}_b(\mathcal{X} \times \mathcal{Y})$:

$$V_i^\lambda(u) = \varepsilon_i \left( \int_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \exp \frac{-u(x,y) - \lambda_i c_i(x,y)}{\varepsilon_i} d\mu(x) d\nu(y) - 1 \right)$$

and:

$$E(u) = \begin{cases} \int f d\mu + \int g d\nu & \text{if} \quad \exists (f,g) \in \mathcal{C}_b(\mathcal{X}) \times \mathcal{C}_b(\mathcal{Y}),\ u = f + g \\ +\infty & \text{else} \end{cases}$$

Let compute the Fenchel-Legendre transform of these functions. Let $\gamma \in \mathcal{M}(\mathcal{X} \times \mathcal{Y})$:

$$V_i^{\lambda *}(-\gamma) = \sup_{u\in\mathcal{C}_b(\mathcal{X}\times\mathcal{Y})} - \int u d\gamma - \varepsilon_i \left( \int_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \exp \frac{-u(x,y) - \lambda_i c_i(x,y)}{\varepsilon_i} d\mu(x) d\nu(y) - 1 \right)$$

However, by density of $\mathcal{C}_b(\mathcal{X} \times \mathcal{Y})$ in $L^1_{d\mu\otimes\nu}(\mathcal{X} \times \mathcal{Y})$, the set of integrable functions for $\mu \otimes \nu$ measure, we deduce that

$$V_i^{\lambda *}(-\gamma) = \sup_{u\in L^1_{d\mu\otimes\nu}(\mathcal{X}\times\mathcal{Y})} - \int u d\gamma - \varepsilon_i \left( \int_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \exp \frac{-u(x,y) - \lambda_i c_i(x,y)}{\varepsilon_i} d\mu(x) d\nu(y) - 1 \right)$$

This supremum equals $+\infty$ if $\gamma$ is not positive and not absolutely continuous with regard to $\mu \otimes \nu$. Let us now denote

$$F_{\gamma,\lambda}(u) := - \int u d\gamma - \varepsilon_i \left( \int_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \exp \frac{-u(x,y) - \lambda_i c_i(x,y)}{\varepsilon_i} d\mu(x) d\nu(y) - 1 \right).$$

$F_{\gamma,\lambda_*}$ is Fréchet differentiable and its maximum is attained for $u^* = \varepsilon_i \log\left(\frac{d\gamma}{d\mu\otimes\nu}\right) + \lambda_i c_i$. Therefore we obtain that

$$\begin{aligned} V_i^{\lambda *}(-\gamma) &= \varepsilon_i \left( \int \log\left(\frac{d\gamma}{d\mu\otimes\nu}\right) d\gamma + 1 - \gamma(\mathcal{X}\times\mathcal{Y}) \right) + \lambda_i \int c_i d\gamma \\ &= \lambda_i \int c_i d\gamma + \varepsilon_i \mathrm{KL}(\gamma_i, \mu \times \nu) \end{aligned}$$

Thanks to the compactness of $\mathcal{X} \times \mathcal{Y}$, all the $V_i^\lambda$ for $i \in \{1, ..., N\}$ are continuous on $\mathcal{C}_b(\mathcal{X} \times \mathcal{Y})$. Therefore by applying Lemma 10, we obtain that:

$$\inf_{u\in\mathcal{C}_b(\mathcal{X}\times\mathcal{Y})} \sum_i V_i^\lambda(u) + E(u) = \sup_{\substack{\gamma_1 ..., \gamma_N, \gamma \in \mathcal{M}(\mathcal{X}\times\mathcal{Y}) \\ \sum_i \gamma_i = \gamma}} - \sum_i V_i^{\lambda *}(\gamma_i) - E^*(-\gamma)$$

$$\sup_{f\in\mathcal{C}_b(\mathcal{X}),\ g\in\mathcal{C}_b(\mathcal{Y})} \int f d\mu + \int g d\nu$$
$$- \sum_{i=1}^{N} \varepsilon_i \left( \int_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \exp \frac{f(x)+g(y)-\lambda_i c_i(x,y)}{\varepsilon_i} d\mu(x) d\nu(y) - 1 \right)$$
$$= \inf_{\gamma\in\Gamma^N(\mu,\nu)} \sum_{i=1}^{N} \lambda_i \int c_i d\gamma_i + \varepsilon_i \mathrm{KL}(\gamma_i, \mu\otimes\nu)$$

Therefore by considering the supremum over the $\lambda \in \Delta_N$, we obtain that

$$\sup_{\lambda\in\Delta_N^+} \sup_{f\in\mathcal{C}_b(\mathcal{X}),\ g\in\mathcal{C}_b(\mathcal{Y})} \int f d\mu + \int g d\nu$$
$$- \sum_{i=1}^{N} \varepsilon_i \left( \int_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \exp \frac{f(x)+g(y)-\lambda_i c_i(x,y)}{\varepsilon_i} d\mu(x) d\nu(y) - 1 \right)$$
$$= \sup_{\lambda\in\Delta_N^+} \inf_{\gamma\in\Gamma^N(\mu,\nu)} \sum_{i=1}^{N} \lambda_i \int c_i d\gamma_i + \varepsilon_i \mathrm{KL}(\gamma_i, \mu\otimes\nu)$$

Let $f : (\lambda, \gamma) \in \Delta_N^+ \times \Gamma^N(\mu,\nu) \mapsto \sum_{i=1}^N \lambda_i \int c_i d\gamma_i + \varepsilon_i \mathrm{KL}(\gamma_i, \mu\otimes\nu)$. $f$ is clearly concave and continuous in $\lambda$. Moreover $\gamma \mapsto \mathrm{KL}(\gamma_i, \mu\otimes\nu)$ is convex and lower semi-continuous for weak topology [239, Lemma 1.4.3]. Hence $f$ is convex and lower-semi continuous in $\gamma$. $\Delta_N^+$ is convex, and $\Gamma^N(\mu,\nu)$ is compact for weak topology (see Lemma 9). So by Sion's theorem, we get the expected result:

$$\min_{\gamma\in\Gamma^N(\mu,\nu)} \sup_{\lambda\in\Delta_N^+} \sum_i \lambda_i \int c_i d\gamma_i + \sum_i \varepsilon_i \mathrm{KL}(\gamma_i, \mu\otimes\nu)$$
$$= \sup_{\lambda\in\Delta_N^+} \sup_{(f,g)\in\mathcal{C}_b(\mathcal{X})\times\mathcal{C}_b(\mathcal{Y})} \int_{\mathcal{X}} f(x) d\mu(x) + \int_{\mathcal{Y}} g(y) d\nu(y)$$
$$- \sum_{i=1}^{N} \varepsilon_i \left( \int_{\mathcal{X}\times\mathcal{Y}} e^{\frac{f(x)+g(y)-\lambda_i c_i(x,y)}{\varepsilon_i}} d\mu(x) d\nu(y) - 1 \right)$$

Moreove by fixing $\gamma \in \Gamma^N(\mu,\nu)$, we have

$$\sup_{\lambda\in\Delta_N^+} \sum_i \lambda_i \int c_i d\gamma_i + \sum_i \varepsilon_i \mathrm{KL}(\gamma_i, \mu\otimes\nu)$$
$$= \max_i \int c_i d\gamma_i + \sum_j \varepsilon_j \mathrm{KL}(\gamma_j, \mu\otimes\nu)$$

which concludes the proof in case of continuous costs. A similar proof as the one of the Theorem 8.4.1 allows to extend the results for lower semi-continuous cost functions. □

## 8.7 Discrete cases

### 8.7.1 Exact discrete case

Let $a \in \Delta_N^+$ and $b \in \Delta_m^+$ and $\mathbf{C} := (C_i)_{1\leq i\leq N} \in (\mathbb{R}^{n\times m})^N$ be $N$ cost matrices. Let also $X := \{x_1, ..., x_n\}$ and $Y := \{y_1, ..., y_m\}$ two subset of $\mathcal{X}$ and $\mathcal{Y}$ respectively. Moreover we define the two following discrete measure $\mu = \sum_{i=1}^n a_i \delta_{x_i}$ and $\nu = \sum_{i=1}^n b_i \delta_{y_i}$ and for all $i$, $C_i = (c_i(x_k, y_l))_{1\leq k\leq n, 1\leq l\leq m}$ where $(c_i)_{i=1}^N$ a family of cost functions. The discretized multiple cost optimal transport primal problem can be written as follows:

$$\mathrm{EOT}_{\mathbf{c}}(\mu, \nu) = \inf_{P\in\Gamma_{a,b}^N} \max_i \langle P_i, C_i\rangle$$

where $\Gamma_{a,b}^N := \left\{(P_i)_{1\leq i\leq N} \in \left(\mathbb{R}_+^{n\times m}\right)^N \text{ s.t. } (\sum_i P_i)\mathbf{1}_m = a \text{ and } (\sum_i P_i^T)\mathbf{1}_n = b\right\}$. As in the continuous case, strong duality holds and we can rewrite the dual in the discrete case also.

**Proposition 8.7.1** (Duality for the discrete problem)**.** *Let $a \in \Delta_N^+$ and $b \in \Delta_m^+$ and $\mathbf{C} := (C_i)_{1\leq i\leq N} \in (\mathbb{R}^{n\times m})^N$ be $N$ cost matrices. Strong duality holds for the discrete problem and*

$$\mathrm{EOT}_{\mathbf{c}}(\mu, \nu) = \sup_{\lambda\in\Delta_N^+} \sup_{(f,g)\in\mathcal{F}_{\mathbf{C}}^\lambda} \langle f, a\rangle + \langle g, b\rangle.$$

*where $\mathcal{F}_{\mathbf{C}}^\lambda := \{(f, g) \in \mathbb{R}_+^n \times \mathbb{R}_+^m \text{ s.t. } \forall i \in \{1, ..., N\},\ f\mathbf{1}_m^T + \mathbf{1}_n g^T \leq \lambda_i C_i\}$.*

### 8.7.2 Entropic regularized discrete case

We now extend the regularization in the discrete case. Let $a \in \Delta_n^+$ and $b \in \Delta_m^+$ and $\mathbf{C} := (C_i)_{1\leq i\leq N} \in (\mathbb{R}^{n\times m})^N$ be $N$ cost matrices and $\boldsymbol{\varepsilon} = (\varepsilon_i)_{1\leq i\leq N}$ be nonnegative real numbers. The discretized regularized primal problem is:

$$\mathrm{EOT}_{\mathbf{c}}^{\boldsymbol{\varepsilon}}(\mu, \nu) = \inf_{P\in\Gamma_{a,b}^N} \max_i \langle P_i, C_i\rangle - \sum_{i=1}^N \varepsilon_i \mathrm{H}(P_i)$$

where $\mathrm{H}(P) = \sum_{i,j} P_{i,j}(\log P_{i,j} - 1)$ for $P = (P_{i,j})_{i,j} \in \mathbb{R}_+^{n\times m}$ is the discrete entropy. In the discrete case, strong duality holds thanks to Lagrangian duality and Slater sufficient conditions:

**Proposition 8.7.2** (Duality for the discrete regularized problem)**.** *Let $a \in \Delta_n^+$ and $b \in \Delta_m^+$ and $\mathbf{C} := (C_i)_{1\leq i\leq N} \in (\mathbb{R}^{n\times m})^N$ be $N$ cost matrices and $\boldsymbol{\varepsilon} := (\varepsilon_i)_{1\leq i\leq N}$*

*be non negative reals. Strong duality holds and by denoting* $K_i^{\lambda_i} = \exp(-\lambda_i C_i/\varepsilon_i)$, *we have*

$$\mathrm{EOT}_{\mathbf{c}}^{\boldsymbol{\varepsilon}}(\mu,\nu) = \sup_{\lambda\in\Delta_N^+} \sup_{f\in\mathbb{R}^n,\ g\in\mathbb{R}^m} \langle f,a\rangle + \langle g,b\rangle - \sum_{i=1}^N \varepsilon_i \langle e^{\mathbf{f}/\varepsilon_i}, K_i^{\lambda_i} e^{\mathbf{g}/\varepsilon_i}\rangle.$$

The objective function for the dual problem is strictly concave in $(\lambda, f, g)$ but is neither smooth or strongly convex.

*Proof.* The proofs in the discrete case are simpler and only involves Lagrangian duality [242, Chapter 5]. Let do the proof in the regularized case, the one for the standard problem follows exactly the same path.

Let $a \in \Delta_N^+$ and $b \in \Delta_m^+$ and $\mathbf{C} := (C_i)_{1\le i\le N} \in (\mathbb{R}^{n\times m})^N$ be $N$ cost matrices.

$$\begin{aligned}
\mathrm{EOT}_{\mathbf{c}}^{\boldsymbol{\varepsilon}}(\mu,\nu) &= \inf_{P\in\Gamma_{a,b}^N} \max_{1\le i\le N} \langle P_i, C_i\rangle - \sum_{i=1}^N \varepsilon_i \mathrm{H}(P_i) \\
&= \inf_{\substack{(t,P)\in\mathbb{R}\times\left(\mathbb{R}_+^{n\times m}\right)^N \\ (\sum_i P_i)\mathbf{1}_m = a \\ (\sum_i P_i^T)\mathbf{1}_n = b \\ \forall j,\ \langle P_j, C_j\rangle \le t}} t - \sum_{i=1}^N \varepsilon_i \mathrm{H}(P_i) \\
&= \inf_{(t,P)\in\mathbb{R}\times\left(\mathbb{R}_+^{n\times m}\right)^N} \sup_{f\in\mathbb{R}^n,\ g\in\mathbb{R}^m,\ \lambda\in\mathbb{R}_+^N} t + \sum_{j=1}^N \lambda_j(\langle P_j, C_j\rangle - t) - \sum_{i=1}^N \varepsilon_i \mathrm{H}(P_i) \\
&\quad + f^T\left(a - \sum_i P_i \mathbf{1}_m\right) + g^T\left(b - \sum_i P_i^T \mathbf{1}_n\right)
\end{aligned}$$

The constraints are qualified for this convex problem, hence by Slater's sufficient condition [242, Section 5.2.3], strong duality holds and:

$$\begin{aligned}
\mathrm{EOT}_{\mathbf{c}}^{\boldsymbol{\varepsilon}}(\mu,\nu) &= \sup_{f\in\mathbb{R}^n,\ g\in\mathbb{R}^m,\ \lambda\in\mathbb{R}_+^N} \inf_{(t,P)\in\mathbb{R}\times\left(\mathbb{R}_+^{n\times m}\right)^N} t + \sum_{j=1}^N \lambda_j(\langle P_j, C_j\rangle - t) - \sum_{j=1}^N \varepsilon_j \mathrm{H}(P_j) \\
&\quad + f^T\left(a - \sum_{j=1}^N P_i \mathbf{1}_m\right) + g^T\left(b - \sum_{j=1}^N P_i^T \mathbf{1}_n\right) \\
&= \sup_{\substack{f\in\mathbb{R}^n \\ g\in\mathbb{R}^m \\ \lambda\in\Delta_N^+}} \langle f,a\rangle + \langle g,b\rangle + \sum_{j=1}^N \inf_{P_j\in\mathbb{R}_+^{n\times m}} \left(\langle P_j, \lambda_j C_j - f\mathbf{1}_n^T - \mathbf{1}_m g^T\rangle - \varepsilon_j \mathrm{H}(P_j)\right)
\end{aligned}$$

But for every $i = 1, .., N$ the solution of

$$\inf_{P_j \in \mathbb{R}_+^{n \times m}} \left( \langle P_j, \lambda_j C_j - f \mathbf{1}_n^T - \mathbf{1}_m g^T \rangle - \varepsilon_j \mathrm{H}(P_j) \right)$$

is

$$P_j = \exp\left( \frac{f \mathbf{1}_n^T + \mathbf{1}_m g^T - \lambda_j C_j}{\varepsilon_i} \right)$$

Finally we obtain that

$$\mathrm{EOT}_{\mathbf{c}}^{\boldsymbol{\epsilon}}(\mu, \nu) = \sup_{f \in \mathbb{R}^n,\ g \in \mathbb{R}^m,\ \lambda \in \Delta_N^+} \langle f, a \rangle + \langle g, b \rangle - \sum_{k=1}^{N} \varepsilon_k \sum_{i,j} \exp\left( \frac{f_i + g_j - \lambda_k C_k^{i,j}}{\varepsilon_k} \right)$$

□

## 8.8 Other results

### 8.8.1 Utilitarian and Optimal Transport

**Proposition 8.8.1.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be Polish spaces. Let $\mathbf{c} := (c_i)_{1\leq i\leq N}$ be a family of bounded below continuous cost functions on $\mathcal{X}\times\mathcal{Y}$, and $\mu\in\mathcal{P}(\mathcal{X})$ and $\nu\in\mathcal{P}(\mathcal{Y})$. Then we have:*

$$\inf_{(\gamma_i)_{i=1}^N\in\Gamma^N(\mu,\nu)}\sum_i\int c_i d\gamma_i = OT_{\min_i(c_i)}(\mu,\nu) \tag{8.21}$$

*Proof.* The proof is a by-product of the proof of Theorem 8.3.1. The continuity of the costs is necessary since $\min_i(c_i)$ is not necessarily lower semi-continuous when the costs are supposed lower semi-continuous. □

**Remark 12.** *We thank an anonymous reviewer for noticing that the utilitarian problem can be written also as an Optimal Transport on the space $\mathcal{Z} = (\mathcal{X}\times\{1,\dots,N\})\times(\mathcal{Y}\times\{1,\dots,N\})$:*

$$\min_{\gamma\in\tilde{\Gamma}(\mu,\nu)}\int_{x,i,y,j} c((x,i),(y,j))d\gamma(x,i,y,j)$$

*where the constraint space is $\tilde{\Gamma}(\mu,\nu) := \left\{\gamma\in\mathcal{M}_1^+(\mathcal{Z})\ s.t.\ \pi_{\mathcal{X}}\gamma=\mu,\ \pi_{\mathcal{Y}}\gamma=\nu\right\}$.*

### 8.8.2 MOT generalizes OT

**Proposition 8.8.2.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be Polish spaces. Let $N\geq 0$, $\mathbf{c} = (c_i)_{1\leq i\leq N}$ be a family of nonnegative lower semi-continuous costs and let us denote for all $k\in\{1,\dots,N\}$, $\mathbf{c}_k = (c_i)_{1\leq i\leq k}$. Then for all $k\in\{1,\dots,N\}$, there exists a family of costs $\mathbf{d}_k\in LSC(\mathcal{X}\times\mathcal{Y})^N$ such that*

$$\mathrm{EOT}_{\mathbf{d}_k}(\mu,\nu) = \mathrm{EOT}_{\mathbf{c}_k}(\mu,\nu) \tag{8.22}$$

*Proof.* For all $k\in\{1,...,N\}$, we define $\mathbf{d}_k := (c_1,...,(N-k+1)\times c_k,...,(N-k+1)\times c_k)$. Therefore, thanks to Lemma 12 we have

$$\mathrm{EOT}_{\mathbf{d}_k}(\mu,\nu) = \sup_{\lambda\in\Delta_N^+}\mathrm{OT}_{c_\lambda}(\mu,\nu) \tag{8.23}$$

$$= \sup_{(\lambda,\gamma)\in\Delta_n^k}\inf_{\gamma\in\Gamma_{\mu,\nu}}\int_{\mathcal{X}\times\mathcal{Y}}\min(\lambda_1c_1,..,\lambda_{k-1}c_{k-1},\lambda_k c_k)d\gamma \tag{8.24}$$

where $\Delta_n^k := \{(\lambda, \gamma) \in \Delta_N^+ \times \mathbb{R}_+ : \quad \gamma = (N-k+1) \times \min(\lambda_k, ..., \lambda_N)\}$. First remarks that

$$\gamma = 1 - \sum_{i=1}^{k-1} \lambda_i \iff (N-k+1) \times \min(\lambda_k, ..., \lambda_N) = \sum_{i=k}^{N} \lambda_i \tag{8.25}$$

$$\iff \lambda_k = ... = \lambda_N \tag{8.26}$$

But in that case $(\lambda_1, ..., \lambda_{k-1}, \gamma) \in \Delta_k$ and therefore we obtain that

$$\mathrm{EOT}_{\mathbf{d}_k}(\mu, \nu) \geq \sup_{\lambda \in \Delta_k} \inf_{\gamma \in \Gamma_{\mu,\nu}} \int_{\mathcal{X}\times\mathcal{Y}} \min(\lambda_1 c_1, .., \lambda_{k-1} c_{k-1}, \gamma c_k) d\gamma = \mathrm{EOT}_{\mathbf{c}_k}(\mu, \nu)$$

Finally by definition we have $\gamma \leq \sum_{i=k}^{N} \lambda_i = 1 - \sum_{i=1}^{k-1} \lambda_i$ and therefore

$$\int_{\mathcal{X}\times\mathcal{Y}} \min(\lambda_1 c_1, .., \lambda_{k-1} c_{k-1}, \gamma c_k) d\gamma \leq \int_{\mathcal{X}\times\mathcal{Y}} \min\left(\lambda_1 c_1, .., \lambda_{k-1} c_{k-1}, \left(1 - \sum_{i=1}^{k-1} \lambda_i\right) c_k\right)$$

Then we obtain that

$$\mathrm{EOT}_{\mathbf{d}_k}(\mu, \nu) \leq \mathrm{EOT}_{\mathbf{c}_k}(\mu, \nu)$$

and the result follows. □

**Proposition 8.8.3.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be Polish spaces and $\mathbf{c} := (c_i)_{1 \leq i \leq N}$ a family of nonnegative lower semi-continuous costs on $\mathcal{X} \times \mathcal{Y}$. We suppose that, for all $i$, $c_i = N \times c_1$. Then for any $(\mu, \nu) \in \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y})$*

$$\mathrm{EOT}_{\mathbf{c}}(\mu, \nu) = \mathrm{EOT}_{c_1}(\mu, \nu) = OT_{c_1}(\mu, \nu). \tag{8.27}$$

*Proof.* Let $c := (c_i)_{1 \leq i \leq N}$ such that for all $i$, $c_i = c_1$. for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and $\lambda \in \Delta_N^+$, we have:

$$c_\lambda(x, y) := \min_i(\lambda_i c_i(x, y)) = \min_i(\lambda_i) c_1(x, y)$$

Therefore we obtain from Lemma 12 that

$$\mathrm{EOT}_c(\mu, \nu) = \sup_{\lambda \in \Delta_N^+} \mathrm{OT}_{c_\lambda}(\mu, \nu) \tag{8.28}$$

But we also have that:

$$\begin{aligned}
\mathrm{OT}_{c_\lambda}(\mu, \nu) &= \inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} \min_i(\lambda_i c_i(x, y)) d\gamma(x, y) \\
&= \min_i(\lambda_i) \inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c_1(x, y) d\gamma(x, y) \\
&= \min_i(\lambda_i) \mathrm{OT}_{c_1}(\mu, \nu)
\end{aligned}$$

Finally by taking the supremum over $\lambda \in \Delta_N^+$ we conclude the proof. □

### 8.8.3 Regularized EOT tends to EOT

**Proposition 8.8.4.** *For* $(\mu,\nu) \in \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y})$ *we have* $\lim_{\boldsymbol{\varepsilon}\to 0} \mathrm{EOT}_{\mathbf{c}}^{\boldsymbol{\varepsilon}}(\mu,\nu) = \mathrm{EOT}_{\mathbf{c}}(\mu,\nu)$.

*Proof.* Let $(\boldsymbol{\varepsilon}_l = (\varepsilon_{l,1},\dots,\varepsilon_{l,N}))_l$ a sequence converging to 0. Let $\gamma_l = (\gamma_{l,1},\dots,\gamma_{l,N})$ be the optimum of $\mathrm{EOT}_{\mathbf{c}}^{\boldsymbol{\varepsilon_l}}(\mu,\nu)$. By Lemma 9, up to an extraction, $\gamma_l \to \gamma^\star = (\gamma_1^\star,\dots,\gamma_N^\star) \in \Gamma^N(\mu,\nu)$. Let now $\gamma = (\gamma_1,\dots,\gamma_N)$ be the optimum of $\mathrm{EOT}_{\mathbf{c}}(\mu,\nu)$. By optimality of $\gamma$ and $\gamma_l$, for all $i$:

$$0 \le \int c_i d\gamma_{l,i} - \int c_i d\gamma_i \le \sum_i \varepsilon_{l,i} \left(\mathrm{KL}(\gamma_i, \mu\otimes\nu) - \mathrm{KL}(\gamma_{l,i}, \mu\otimes\nu)\right)$$

By lower semi continuity of $\mathrm{KL}(.,\mu\otimes\nu)$ and by taking the limit inferior as $l\to\infty$, we get for all $i$, $\liminf_{\ell\to\infty} \int c_i d\gamma_{l,i} = \int c_i d\gamma_i$. Moreover by continuity of $\gamma \to \int c_i d\gamma_i$ we therefore obtain that for all $i$, $\int c_i d\gamma_i^\star \le \int c_i d\gamma_i$. Then by optimality of $\gamma$ the result follows. □

### 8.8.4 Projected Accelerated Gradient Descent

**Proposition 8.8.5.** *Let* $a \in \Delta_N^+$ *and* $b \in \Delta_m^+$ *and* $\mathbf{C} := (C_i)_{1\le i\le N} \in (\mathbb{R}^{n\times m})^N$ *be* $N$ *cost matrices and* $\boldsymbol{\varepsilon} := (\varepsilon,...,\varepsilon)$ *where* $\varepsilon > 0$. *Then by denoting* $K_i^{\lambda_i} = \exp(-\lambda_i C_i/\varepsilon)$, *we have*

$$\mathrm{EOT}_{\mathbf{c}}^{\boldsymbol{\varepsilon}}(\mu,\nu) = \sup_{\lambda\in\Delta_N^+} \sup_{f\in\mathbb{R}^n,\ g\in\mathbb{R}^m} F_{\mathbf{C}}^{\varepsilon}(\lambda,f,g),$$

*where*

$$F_{\mathbf{C}}^{\varepsilon}(\lambda,f,g) := \langle f,a\rangle + \langle g,b\rangle - \varepsilon\left[\log\left(\sum_{i=1}^N \langle e^{\mathbf{f}/\varepsilon}, K_i^{\lambda_i} e^{\mathbf{g}/\varepsilon}\rangle\right) + 1\right]$$

*Moreover,* $F_{\mathbf{C}}^{\varepsilon}$ *is concave, differentiable and* $\nabla F$ *is* $\frac{\max\left(\max_{1\le i\le N}\|C_i\|_\infty^2, 2N\right)}{\varepsilon}$ *Lipschitz-continuous on* $\mathbb{R}^N\times\mathbb{R}^n\times\mathbb{R}^m$.

*Proof.* Let $\mathcal{Q} := \left\{P := (P_1,...,P_N) \in (\mathbb{R}_+^{n\times m})^N : \quad \sum_{k=1}^N \sum_{i,j} P_k^{i,j} = 1\right\}$. Note that

$\Gamma^N_{a,b} \subset \mathcal{Q}$, therefore from the primal formulation of the problem we have that

$$\begin{aligned}
\mathrm{EOT}^{\varepsilon}_{\mathbf{c}}(\mu,\nu) &= \sup_{\lambda\in\Delta_N^+} \inf_{P\in\Gamma^N_{a,b}} \sum_{i=1}^N \lambda_i \langle P_i, C_i\rangle - \varepsilon \mathrm{H}(P_i) \\
&= \sup_{\lambda\in\Delta_N^+} \inf_{P\in\mathcal{Q}} \sup_{f\in\mathbb{R}^n,\ g\in\mathbb{R}^m} \sum_{i=1}^N \lambda_i \langle P_i, C_i\rangle - \varepsilon \mathrm{H}(P_i) \\
&\quad + f^T\left(a - \sum_i P_i \mathbf{1}_m\right) + g^T\left(b - \sum_i P_i^T \mathbf{1}_n\right)
\end{aligned}$$

The constraints are qualified for this convex problem, hence by Slater's sufficient condition [242, Section 5.2.3], strong duality holds. Therefore we have

$$\begin{aligned}
\mathrm{EOT}^{\varepsilon}_{\mathbf{c}}(\mu,\nu) &= \sup_{\lambda\in\Delta_N^+} \sup_{f\in\mathbb{R}^n,\ g\in\mathbb{R}^m} \inf_{P\in\mathcal{Q}} \sum_{i=1}^N \lambda_i \langle P_i, C_i\rangle - \varepsilon \mathrm{H}(P_i) \\
&\quad + f^T\left(a - \sum_i P_i \mathbf{1}_m\right) + g^T\left(b - \sum_i P_i^T \mathbf{1}_n\right) \\
&= \sup_{\lambda\in\Delta_N^+} \sup_{f\in\mathbb{R}^n,\ g\in\mathbb{R}^m} \langle f, a\rangle + \langle g, b\rangle \\
&\quad + \inf_{P\in\mathcal{Q}} \sum_{k=1}^N \sum_{i,j} P_k^{i,j}\left(\lambda_k C_k^{i,j} + \varepsilon\left(\log(P_k^{i,j}) - 1\right) - f_i - g_j\right)
\end{aligned}$$

Let us now focus on the following problem:

$$\inf_{P\in\mathcal{Q}} \sum_{k=1}^N \sum_{i,j} P_k^{i,j}\left(\lambda_k C_k^{i,j} + \varepsilon\left(\log(P_k^{i,j}) - 1\right) - f_i - g_j\right)$$

Note that for all $i, j, k$ and some small $\delta$,

$$P_k^{i,j}\left(\lambda_k C_k^{i,j} - \varepsilon\left(\log(P_k^{i,j}) - 1\right) - f_i - g_j\right) < 0$$

if $P_k^{i,j} \in (0, \delta)$ and this quantity goes to 0 as $P_k^{i,j}$ goes to 0. Therefore $P_k^{i,j} > 0$ and the problem becomes

$$\inf_{P>0} \sup_{\nu\in\mathbb{R}} \sum_{k=1}^N \sum_{i,j} P_k^{i,j}\left(\lambda_k C_k^{i,j} + \varepsilon\left(\log(P_k^{i,j}) - 1\right) - f_i - g_j\right) + \nu\left(\sum_{k=1}^N \sum_{i,j} P_k^{i,j} - 1\right).$$

The solution to this problem is for all $k \in \{1, .., N\}$,

$$P_k = \frac{\exp\left(\frac{f\mathbf{1}_n^T + \mathbf{1}_m g^T - \lambda_k C_k}{\varepsilon}\right)}{\sum_{k=1}^N \sum_{i,j} \exp\left(\frac{f_i + g_j - \lambda_k C_k^{i,j}}{\varepsilon}\right)}$$

Therefore we obtain that

$$\begin{aligned}
\mathrm{EOT}_{\mathbf{c}}^{\boldsymbol{\varepsilon}}(\mu,\nu) = & \sup_{\lambda \in \Delta_N^+} \sup_{f \in \mathbb{R}^n,\ g \in \mathbb{R}^m} \langle f, a\rangle + \langle g, b\rangle \\
& - \varepsilon \sum_{k=1}^N \sum_{i,j} P_k^{i,j} \left[ \log\left( \sum_{k=1}^N \sum_{i,j} \exp\left( \frac{f_i + g_j - \lambda_k C_k^{i,j}}{\varepsilon} \right) \right) + 1 \right] \\
= & \sup_{\lambda \in \Delta_N^+} \sup_{f \in \mathbb{R}^n,\ g \in \mathbb{R}^m} \langle f, a\rangle + \langle g, b\rangle - \varepsilon \left[ \log\left( \sum_{k=1}^N \sum_{i,j} \exp\left( \frac{f_i + g_j - \lambda_k C_k^{i,j}}{\varepsilon} \right) \right) + 1 \right].
\end{aligned}$$

From now on, we denote for all $\lambda \in \Delta_N^+$

$$\mathrm{EOT}_{\mathbf{c}}^{\boldsymbol{\varepsilon},\lambda}(\mu,\nu) := \inf_{P \in \Gamma_{a,b}^N} \sum_{i=1}^N \lambda_i \langle P_i, C_i\rangle - \varepsilon \mathrm{H}(P_i)$$

$$\mathrm{EOT}_{\mathbf{c}}^{\boldsymbol{\varepsilon},\lambda}(\mu,\nu) := \sup_{f \in \mathbb{R}^n,\ g \in \mathbb{R}^m} \langle f, a\rangle + \langle g, b\rangle - \varepsilon \left[ \log\left( \sum_{k=1}^N \sum_{i,j} \exp\left( \frac{f_i + g_j - \lambda_k C_k^{i,j}}{\varepsilon} \right) \right) + 1 \right]$$

which has just been shown to be dual and equal. Thanks to [190, Theorem 1], as for all $\lambda \in \mathbb{R}^N$, $P \in \Gamma_{a,b}^N \to \sum_{i=1}^N \lambda_i \langle P_i, C_i\rangle - \varepsilon \mathrm{H}(P_i)$ is $\varepsilon$-strongly convex, then for all $\lambda \in \mathbb{R}^N$, $(f, g) \to \nabla_{(f,g)} F(\lambda, f, g)$ is $\frac{\|A\|_{1\to 2}^2}{\varepsilon}$ Lipschitz-continuous where $A$ is the linear operator of the equality constraints of the primal problem. Moreover this norm is equal to the maximum Euclidean norm of a column of A. By definition, each column of A contains only $2N$ non-zero elements, which are equal to one. Hence, $\|A\|_{1\to 2} = \sqrt{2N}$. Let us now show that for all $(f, g) \in \mathbb{R}^n \times \mathbb{R}^m$ $\lambda \in \mathbb{R}^N \to \nabla_\lambda F(\lambda, f, g)$ is also Lipschitz-continuous. Indeed we remarks that

$$\frac{\partial^2 F}{\partial \lambda_q \partial \lambda_k} = \frac{1}{\varepsilon \nu^2} \left[ \sigma_{q,1}(\lambda)\sigma_{k,1}(\lambda) - \nu(\sigma_{k,2}(\lambda)\mathbb{1}_{k=q}) \right]$$

where $\mathbb{1}_{k=q} = 1$ iff $k = q$ and 0 otherwise, for all $k \in \{1, ..., N\}$ and $p \geq 1$

$$\sigma_{k,p}(\lambda) = \sum_{i,j} (C_k^{i,j})^p \exp\left( \frac{f_i + g_j - \lambda_k C_k^{i,j}}{\varepsilon} \right)$$

$$\nu = \sum_{k=1}^N \sum_{i,j} \exp\left( \frac{f_i + g_j - \lambda_k C_k^{i,j}}{\varepsilon} \right).$$

Let $v \in \mathbb{R}^N$, and by denoting $\nabla^2_\lambda F$ the Hessian of $F$ with respect to $\lambda$ for fixed $f, g$ we obtain first that

$$\begin{aligned}
v^T \nabla^2_\lambda F v &= \frac{1}{\varepsilon\nu^2}\left[\left(\sum_{k=1}^N v_k \sigma_{q,1}(\lambda)\right)^2 - \nu \sum_{k=1}^N v_k^2 \sigma_{k,2}\right] \\
&\leq \frac{1}{\varepsilon\nu^2}\left(\sum_{k=1}^N v_k \sigma_{q,1}(\lambda)\right)^2 \\
&- \frac{1}{\varepsilon\nu^2}\left(\sum_{k=1}^N |v_k| \sqrt{\sum_{i,j} \exp\left(\frac{f_i + g_j - \lambda_k C_k^{i,j}}{\varepsilon}\right)} \sqrt{\sum_{i,j} (C_k^{i,j})^2 \exp\left(\frac{f_i + g_j - \lambda_k C_k^{i,j}}{\varepsilon}\right)}\right)^2 \\
&\leq \frac{1}{\varepsilon\nu^2}\left[\left(\sum_{k=1}^N v_k \sigma_{q,1}(\lambda)\right)^2 - \left(\sum_{k=1}^N |v_k| \sum_{i,j} |C_k^{i,j}| \exp\left(\frac{f_i + g_j - \lambda_k C_k^{i,j}}{\varepsilon}\right)\right)^2\right] \\
&\leq 0
\end{aligned}$$

Indeed the last two inequalities come from Cauchy Schwartz. Moreover we have

$$\begin{aligned}
\frac{1}{\varepsilon\nu^2}\left[\left(\sum_{k=1}^N v_k \sigma_{q,1}(\lambda)\right)^2 - \nu \sum_{k=1}^N v_k^2 \sigma_{k,2}\right] = v^T \nabla^2_\lambda F v \leq 0 \\
-\frac{\sum_{k=1}^N v_k^2 \sigma_{k,2}}{\varepsilon\nu} \leq \\
-\frac{\sum_{k=1}^N v_k^2 \max_{1\leq i\leq N}(\|C_i\|^2_\infty)}{\varepsilon} \leq
\end{aligned}$$

Therefore we deduce that $\lambda \in \mathbb{R}^N \to \nabla_\lambda F(\lambda, f, g)$ is $\frac{\max_{1\leq i\leq N}(\|C_i\|^2_\infty)}{\varepsilon}$ Lipschitz-continuous, hence $\nabla F(\lambda, f, g)$ is $\frac{\max\left(\max_{1\leq i\leq N}\|C_i\|^2_\infty, 2N\right)}{\varepsilon}$ Lipschitz-continuous on $\mathbb{R}^N \times \mathbb{R}^n \times \mathbb{R}^m$. □

Denote $L := \frac{\max\left(\max_{1\leq i\leq N}\|C_i\|^2_\infty, 2N\right)}{\varepsilon}$ the Lipschitz constant of $F^\varepsilon_{\mathbf{C}}$. Moreover for all $\lambda \in \mathbb{R}^N$, let $\mathrm{Proj}_{\Delta_N^+}(\lambda)$ the unique solution of the following optimization problem

$$\min_{x \in \Delta_N^+} \|x - \lambda\|_2^2. \tag{8.29}$$

Let us now introduce the following algorithm.

[236, 237] give us that the accelerated projected gradient ascent algorithm achieves the optimal rate for first order methods of $\mathcal{O}(1/k^2)$ for smooth functions.

**Algorithm 16** Accelerated Projected Gradient Ascent Algorithm

**Input:** $\mathbf{C} = (C_i)_{1\leq i\leq N}$, $a$, $b$, $\varepsilon$, $L$

**Init:** $f^{-1} = f^0 \leftarrow \mathbf{0}_n$; $g^{-1} = g^0 \leftarrow \mathbf{0}_m$; $\lambda^{-1} = \lambda^0 \leftarrow (1/N, ..., 1/N) \in \mathbb{R}^N$

**for** $k = 1, 2, ...$ **do**

$(v, w, z)^T \leftarrow (\lambda^{k-1}, f^{k-1}, g^{k-1})^T + \frac{k-2}{k+1}\left((\lambda^{k-1}, f^{k-1}, g^{k-1})^T - (\lambda^{k-2}, f^{k-2}, g^{k-2})^T\right)$;

$\lambda^k \leftarrow \mathrm{Proj}_{\Delta_N^+}\left(v + \frac{1}{L}\nabla_\lambda F^\varepsilon_{\mathbf{C}}(v, w, z)\right)$;

$(g^k, f^k)^T \leftarrow (w, z)^T + \frac{1}{L}\nabla_{(f,g)} F^\varepsilon_{\mathbf{C}}(v, w, z)$.

**end**

**Result**: $\lambda, f, g$

To perform the projection we use the algorithm proposed in [238] which finds the solution of (8.29) after $\mathcal{O}(N\log(N))$ algebraic operations [243].

### 8.8.5 Fair cutting cake problem

Let $\mathcal{X}$, be a set representing a cake. The aim of the cutting cake problem is to divide it in $\mathcal{X}_1, \ldots, \mathcal{X}_N$ disjoint sets among the $N$ individuals. The utility for a single individual $i$ for a slice $S$ is denoted $V_i(S)$. It is often assumed that $V_i(\mathcal{X}) = 1$ and that $V_i$ is additive for disjoint sets. There exists many criteria to assess fairness for a partition $\mathcal{X}_1, \ldots, \mathcal{X}_N$ such as proportionality ($V_i(\mathcal{X}_i) \geq 1/N$), envy-freeness ($V_i(\mathcal{X}_i) \geq V_i(\mathcal{X}_j)$) or equitability ($V_i(\mathcal{X}_i) = V_j(\mathcal{X}_j)$). A possible problem to solve equitability and proportionality in the cutting cake problem is the following:

$$\inf_{\substack{\mathcal{X}_1,\ldots,\mathcal{X}_N \\ \sqcup_{i=1}^N \mathcal{X}_i = \mathcal{X}}} \max_i V_i(\mathcal{X}_i) \tag{8.30}$$

Note that here we do not want to solve the problem under equality constraints since the problem might not be well defined. Moreover the existence of the optimum is not immediate. A natural relaxation of this problem is when there is a divisible quantity of each element of the cake ($x \in \mathcal{X}$). In that case, the cake is no more a set but rather a distribution on this set $\mu$. Following the primal formulation of EOT, it is clear that it is a relaxation of the cutting cake problem where the goal is to divide the cake viewed as a distribution. For the cutting cake problem with two cakes $\mathcal{X}$ and $\mathcal{Y}$, the problem can be cast as follows:

$$\inf_{\substack{\mathcal{X}_1,\ldots,\mathcal{X}_N \text{ s.t. } \sqcup_{i=1}^N \mathcal{X}_i = \mathcal{X} \\ \mathcal{Y}_1,\ldots,\mathcal{Y}_N \text{ s.t. } \sqcup_{i=1}^N \mathcal{Y}_i = \mathcal{Y}}} \max_i V_i(\mathcal{X}_i, \mathcal{Y}_i) \tag{8.31}$$

Here EOT is the relaxation of this problem where we split the cakes viewed as distributions instead of sets themselves. Note that in this problem, the utility of the agents are coupled.

## 8.9 Illustrations and Experiments

### 8.9.1 Primal Formulation

Here we show the couplings obtained when we consider three negative costs $\tilde{c}_i$ which corresponds to the situation where we aim to obtain a fair division of goods between three agents. Moreover we show the couplings obtained according to the transport viewpoint where we consider the opposite of these three negative cost functions, i.e. $c_i := -\tilde{c}_i$. We can see that the couplings obtained in the two situations are completely different, which is expected. Indeed in the fair division problem, we aim at finding couplings which maximize the total utility of each agent ($\int c_i d\gamma_i^1$) while ensuring that their are equal while in the other case, we aim at finding couplings which minimize the total transportation cost of each agent ($\int c_i d\gamma_i^2$) while ensuring that their are equal. Obviously we always have that

$$\forall i \quad \int c_i d\gamma_i^2 \leq \int c_i d\gamma_i^1.$$

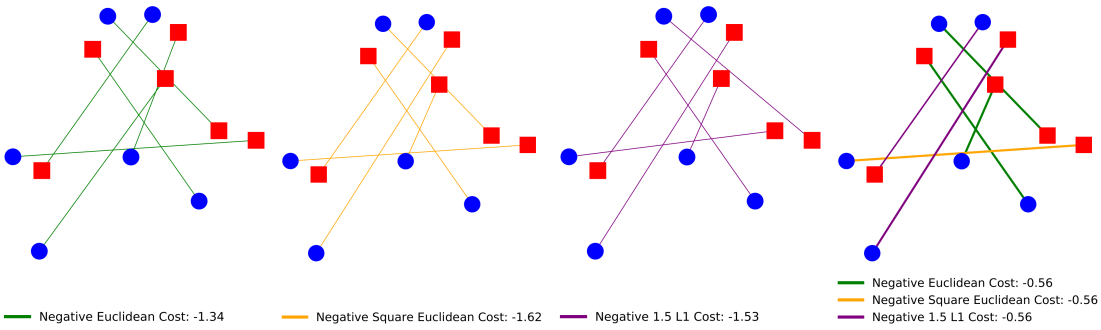



Figure 8.4: Comparison of the optimal couplings obtained from standard OT for three different costs and EOT in case of negative costs (i.e. utilities). Blue dots and red squares represent the locations of two discrete uniform measures. *Left, middle left, middle right*: Kantorovich couplings between the two measures for negative Euclidean cost ($-\|\cdot\|_2$), negative square Euclidean cost ($-\|\cdot\|_2^2$) and negative 1.5 L1 norm ($-\|\cdot\|_1^{1.5}$) respectively. *Right*: Equitable and optimal division of the resources between the $N = 3$ different negative costs (i.e. utilities) given by EOT. Note that the partition between the agents is equitable (i.e. utilities are equal) and proportional (i.e. utilities are larger than $1/N$.

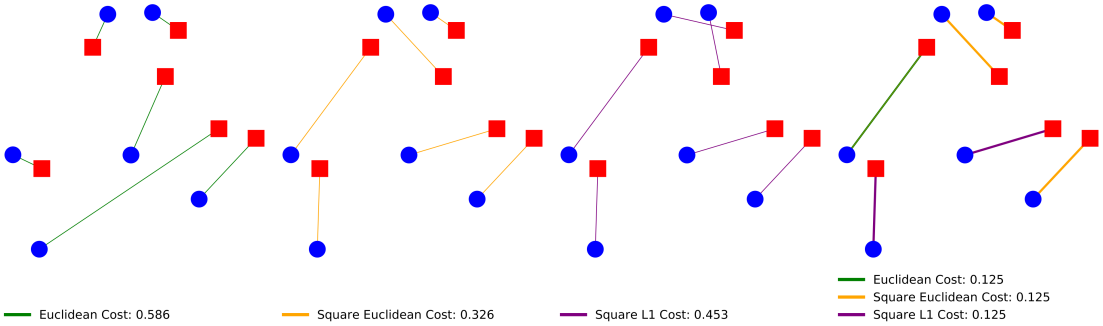



Figure 8.5: Comparison of the optimal couplings obtained from standard OT for three different costs and EOT in case of postive costs. Blue dots and red squares represent the locations of two discrete uniform measures. *Left, middle left, middle right*: Kantorovich couplings between the two measures for Euclidean cost ($\|\cdot\|_2$), square Euclidean cost ($\|\cdot\|_2^2$) and 1.5 L1 norm ($\|\cdot\|_1^{1.5}$) respectively. *Right*: transport couplings of EOT solving Eq. (8.1). Note that each cost contributes equally and its contribution is lower than the smallest OT cost.

### 8.9.2 Dual Formulation

Here we show the dual variables obtained in the exact same settings as in the primal illustrations. Figure 8.6 shows the dual associated to the primal problem exposed in Figure 8.4 and Figure 8.7 shows the dual associated to the primal problem exposed in Figure 8.5.

**Transport viewpoint of the Dual Formulation.** Assume that the $N$ agents are not able to solve the primal problem (8.1) which aims at finding the cheapest equitable partition of the work among the $N$ agents for transporting the distributions of goods $\mu$ to the distributions of stores $\nu$. Moreover assume that there is an external agent who can do the transportation work for them with the following pricing scheme: he or she splits the logistic task into that of collecting and then delivering the goods, and will apply a collection price $\tilde{f}(x)$ for one unit of good located at $x$ (no matter where that unit is sent to), and a delivery price $\tilde{g}(y)$ for one unit to the location $y$ (no matter from which place that unit comes from). Then the external agent for transporting some goods $\mu$ to some stores $\nu$ will charge $\int_{x\in\mathcal{X}} \tilde{f}(x)d\mu(x) + \int_{y\in\mathcal{Y}} \tilde{g}(y)d\nu(y)$. However he or she has the constraint that the pricing must be equitable among the agents and therefore wants to ensure that each agent will pay exactly $\frac{1}{N}\int_{x\in\mathcal{X}} \tilde{f}(x)d\mu(x) + \int_{y\in\mathcal{Y}} \tilde{g}(y)d\nu(y)$. Denote $f = \frac{\tilde{f}}{N}$, $g = \frac{\tilde{g}}{N}$ and therefore the price paid by each agent becomes $\int_{x\in\mathcal{X}} f(x)d\mu(x) + \int_{y\in\mathcal{Y}} g(y)d\nu(y)$. Moreover, to ensure that each agent will not pay more than he would if he was doing the job himself or herself, he or she must guarantee that for all $\lambda \in \Delta_N^+$, the

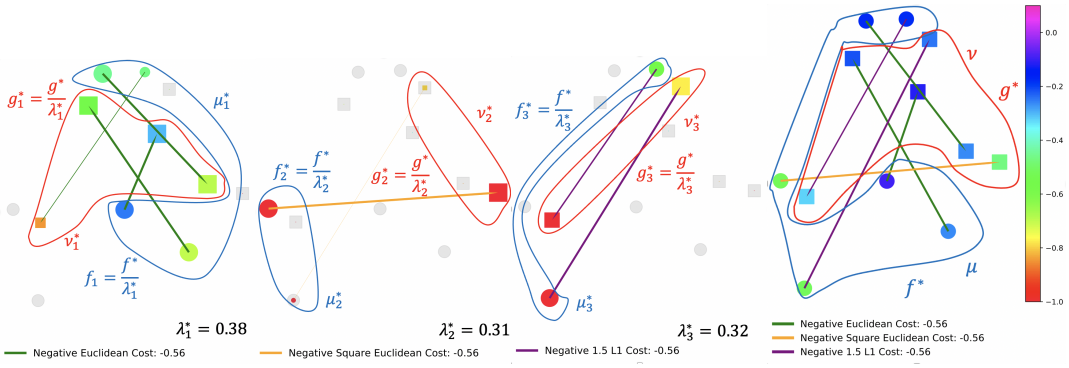



Figure 8.6: *Left, middle left, middle right*: the size of dots and squares is proportional to the weight of their representing atom in the distributions $\mu_k^*$ and $\nu_k^*$ respectively. The utilities $f_k^*$ and $g_k^*$ for each point in respectively $\mu_k^*$ and $\nu_k^*$ are represented by the color of dots and squares according to the color scale on the right hand side. The gray dots and squares correspond to the points that are ignored by agent $k$ in the sense that there is no mass or almost no mass in distributions $\mu_k^*$ or $\nu_k^*$. *Right*: the size of dots and squares are uniform since they correspond to the weights of uniform distributions $\mu$ and $\nu$ respectively. The values of $f^*$ and $g^*$ are given also by the color at each point. Note that each agent gets exactly the same total utility, corresponding exactly to EOT. This value can be computed using dual formulation (8.5) and for each figure it equals the sum of the values (encoded with colors) multiplied by the weight of each point (encoded with sizes).

pricing scheme $(f,g)$ satisfies:

$$f \oplus g \leq \min(\lambda_i c_i).$$

Indeed under this constraint, it is easy for the agents to check that they will never pay more than what they would pay if they were doing the transportation task as we have

$$\int_{x\in\mathcal{X}} f(x)d\mu(x) + \int_{y\in\mathcal{Y}} g(y)d\nu(y) \leq \int_{\mathcal{X}\times\mathcal{Y}} \min_i(\lambda_i c_i)d\gamma$$

which holds for every $\gamma$ in particular for $\gamma^* = \sum_{i=1}^N \gamma_i^*$ optimal solution of the primal problem (8.1) from which follows

$$\begin{aligned}\int_{x\in\mathcal{X}} f(x)d\mu(x) + \int_{y\in\mathcal{Y}} g(y)d\nu(y) &\leq \sum_{i=1}^N \int_{\mathcal{X}\times\mathcal{Y}} \min_i(\lambda_i c_i)d\gamma_i^* \\ &\leq \sum_{i=1}^N \lambda_i \int_{\mathcal{X}\times\mathcal{Y}} c_i d\gamma_i^* \\ &= \mathrm{EOT}_{\mathbf{c}}(\mu,\nu)\end{aligned}$$

Therefore the external agent aims to maximise his or her selling price under the above constraints which is exactly the dual formulation of our problem.

Another interpretation of the dual problem when the cost are non-negative can be expressed as follows. Let us introduce the subset of $(\mathcal{C}_b(\mathcal{X}) \times \mathcal{C}_b(\mathcal{Y}))^N$:

$$\mathcal{G}_{\mathbf{c}}^N := \left\{(f_k, g_k)_{k=1}^N \text{ s.t. } \forall k,\ f_k \oplus g_k \leq c_k\right\}$$

Let us now show the following reformulation of the problem. See Appendix 8.9.2 for the proof.

**Proposition 8.9.1.** *Under the same assumptions of Proposition 8.3.1, we have*

$$\begin{aligned} \mathrm{EOT}_{\mathbf{c}}(\mu,\nu) = \sup_{(f_k,g_k)_{k=1}^N \in \mathcal{G}_{\mathbf{c}}^N} \inf_{\substack{t\in\mathbb{R} \\ (\mu_k,\nu_k)_{k=1}^N \in \Upsilon^N(\mu,\nu)}} & t \\ \text{s.t. } \forall k,\ \int f_k d\mu_k + \int g_k d\nu_k = t & \end{aligned} \tag{8.32}$$

*Proof.* Let us first introduce the following Lemma which guarantees that compacity of $\Upsilon^N(\mu,\nu)$ for the weak topology.

**Lemma 14.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be Polish spaces, and $\mu$ and $\nu$ two probability measures respectively on $\mathcal{X}$ and $\mathcal{Y}$. Then $\Upsilon^N(\mu,\nu)$ is sequentially compact for the weak topology induced by $\|\gamma\| = \max_{i=1,..,N} \|\mu_i\|_{\mathrm{TV}} + \|\nu_i\|_{\mathrm{TV}}$.*

*Proof.* Let $(\gamma^n)_{n\geq 0}$ a sequence in $\Upsilon^N(\mu,\nu)$, and let us denote for all $n \geq 0$, $\gamma^n = (\mu_i^n, \nu_i^n)_{i=1}^N$. We first remarks that for all $i \in \{1,...,N\}$ and $n \geq 0$, $\|\mu_i^n\|_{\mathrm{TV}} \leq 1$ and $\|\nu_i^n\|_{\mathrm{TV}} \leq 1$ therefore for all $i \in \{1,...,N\}$, $(\mu_i^n)_{n\geq 0}$ and $(\nu_i^n)_{n\geq 0}$ are uniformly bounded. Moreover as $\{\mu\}$ and $\{\nu\}$ are tight, for any $\delta > 0$, there exists $K \subset \mathcal{X}$ and $L \subset \mathcal{Y}$ compact such that $\mu(K^c) \leq \delta$ and $\nu(L^c) \leq \delta$. Then, we obtain that for any for all $i \in \{1,...,N\}$, $\mu_i^n(K^c) \leq \delta$ and $\nu_i^n(L^c) \leq \delta$. Therefore, for all $i \in \{1,...,N\}$, $(\mu_i^n)_{n\geq 0}$ and $(\nu_i^n)_{n\geq 0}$ are tight and uniformly bounded and Prokhorov's theorem [239, Theorem A.3.15] guarantees for all $i \in \{1,...,N\}$, $(\mu_i^n)_{n\geq 0}$ and $(\nu_i^n)_{n\geq 0}$ admit a weakly convergent subsequence. By extracting a common convergent subsequence, we obtain that $(\gamma^n)_{n\geq 0}$ admits a weakly convergent subsequence. By continuity of the projection, the limit also lives in $\Upsilon_{\mu,\nu}^N$ and the result follows. □

We can now prove the Proposition. We have that for any $\lambda \in \Delta_N$

$$\begin{aligned} &\sup_{(f,g)\in\mathcal{F}_{\mathbf{c}}^{\lambda}} \int_{x\in\mathcal{X}} f(x)d\mu(x) + \int_{y\in\mathcal{Y}} g(y)d\nu(y) \\ &\leq \sup_{(f_k,g_k)_{k=1}^N \in \mathcal{G}_{\mathbf{c}}^N} \inf_{(\mu_k,\nu_k)_{i=1}^N \in \Upsilon^N(\mu,\nu)} \sum_{k=1}^N \lambda_k \left[\int_{x\in\mathcal{X}} f_k(x)d\mu_k(x) + \int_{y\in\mathcal{Y}} g_k(y)d\nu_k(y)\right] \\ &\leq \mathrm{EOT}_{\mathbf{c}}(\mu,\nu) \end{aligned}$$

Then by taking the supremum over $\lambda \in \Delta_N$, and by applying Theorem 8.3.1 we obtain that

$$\mathrm{EOT}_{\mathbf{c}}(\mu,\nu) = \sup_{\lambda\in\Delta_N} \sup_{(f_k,g_k)_{k=1}^N \in \mathcal{G}_{\mathbf{c}}^N} \inf_{(\mu_k,\nu_k)_{k=1}^N \in \Upsilon^N(\mu,\nu)} \sum_{k=1}^N \lambda_k \left[ \int_{x\in\mathcal{X}} f_k(x) d\mu_k(x) + \int_{y\in\mathcal{Y}} g_k(y) d\nu_k(y) \right]$$

Let $\mathcal{G}_{\mathbf{c}}^N$ and $\Upsilon^N(\mu,\nu)$ be endowed respectively with the uniform norm and the norm defined in Lemma 14. Note that the objective is linear and continuous with respect to $(\mu_k,\nu_k)_{k=1}^N$ and also $(f_k,g_k)_{k=1}^N$. Moreover the spaces $\mathcal{G}_{\mathbf{c}}^N$ and $\Upsilon^N(\mu,\nu)$ are clearly convex. Finally thanks to Lemma 14, $\Upsilon^N(\mu,\nu)$ is compact with respect to the weak topology we can apply Sion's theorem [241] and we obtain that

$$\mathrm{EOT}_{\mathbf{c}}(\mu,\nu) = \sup_{(f_k,g_k)_{k=1}^N \in \mathcal{G}_{\mathbf{c}}^N} \inf_{(\mu_k,\nu_k)_{k=1}^N \in \Upsilon^N(\mu,\nu)} \sup_{\lambda\in\Delta_N} \sum_{k=1}^N \lambda_k \left[ \int_{x\in\mathcal{X}} f_k(x) d\mu_k(x) + \int_{y\in\mathcal{Y}} g_k(y) d\nu_k(y) \right]$$

Let us now fix $(f_k,g_k)_{k=1}^N \in \mathcal{G}_{\mathbf{c}}^N$ and $(\mu_k,\nu_k)_{k=1}^N \in \Upsilon^N(\mu,\nu)$, therefore we have:

$$\begin{aligned}
&\sup_{\lambda\in\Delta_N} \sum_{k=1}^N \lambda_k \left[ \int_{x\in\mathcal{X}} f_k(x) d\mu_k(x) + \int_{y\in\mathcal{Y}} g_k(y) d\nu_k(y) \right] \\
&= \sup_{\lambda} \inf_{t} t \times \left( 1 - \sum_{i=1}^N \lambda_i \right) + \sum_{k=1}^N \lambda_k \left[ \int_{x\in\mathcal{X}} f_k(x) d\mu_k(x) + \int_{y\in\mathcal{Y}} g_k(y) d\nu_k(y) \right] \\
&= \inf_{t} \sup_{\lambda} t + \sum_{k=1}^N \lambda_k \left[ \int_{x\in\mathcal{X}} f_k(x) d\mu_k(x) + \int_{y\in\mathcal{Y}} g_k(y) d\nu_k(y) - t \right] \\
&= \inf_{t} \left\{ t \text{ s.t. } \forall k, \int f_k d\mu_k + \int g_k d\nu_k = t \right\}
\end{aligned}$$

where the inversion is possible as the Slater's conditions are satisfied and the result follows. □

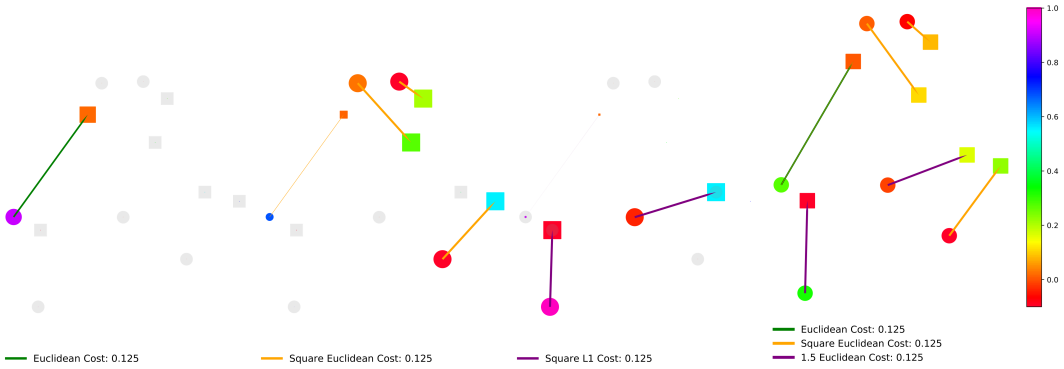



Figure 8.7: *Left, middle left, middle right*: the size of dots and squares is proportional to the weight of their representing atom in the distributions $\mu_k^*$ and $\nu_k^*$ respectively. The collection "cost" $f_k^*$ for each point in $\mu_k^*$, and its delivery counterpart $g_k^*$ in $\nu_k^*$ are represented by the color of dots and squares according to the color scale on the right hand side. The gray dots and squares correspond to the points that are ignored by agent $k$ in the sense that there is no mass or almost no mass in distributions $\mu_k^*$ or $\nu_k^*$. *Right*: the size of dots and squares are uniform since they corresponds to the weights of uniform distributions $\mu$ and $\nu$ respectively. The values of $f^*$ and $g^*$ are given also by the color at each point. Note that each agent earns exactly the same amount of money, corresponding exactly EOT cost. This value can be computed using dual formulation (8.5) or its reformulation (8.32) and for each figure it equals the sum of the values (encoded with colors) multiplied by the weight of each point (encoded with sizes).

### 8.9.3 Approximation of the Dudley Metric

Figure 8.8 illustrates the convergence of the entropic regularization approximation when $\epsilon \to 0$. To do so we plot the relative error from the ground truth defined as $\text{RE} := \frac{\text{EOT}_{\mathbf{c}}^{\epsilon} - \beta_d}{\beta_d}$ for different regularizations where $\beta_d$ is obtained by solving the exact linear program and $\text{EOT}_{\mathbf{c}}^{\epsilon}$ is obtained by our proposed Alg. 15.

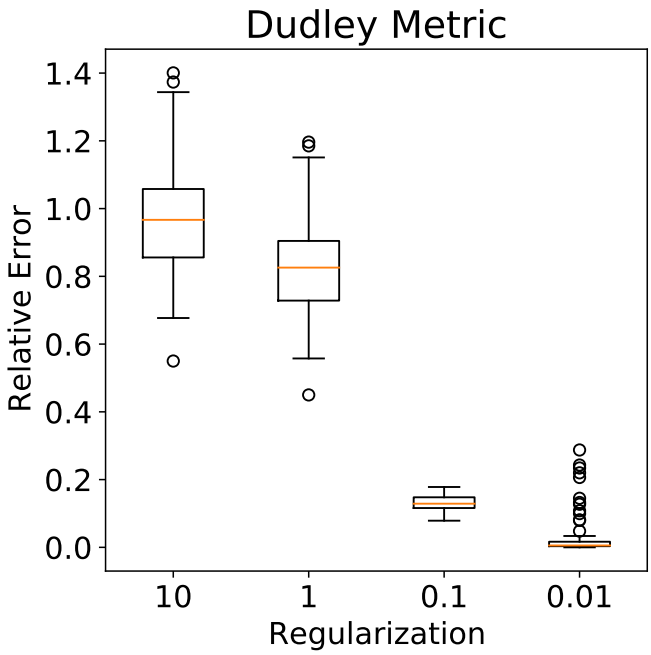



Figure 8.8: In this experiment, we draw 100 samples from two normal distributions and we plot the relative error from ground truth for different regularizations. We consider the case where two costs are involved: $c_1 = 2 \times \mathbf{1}_{x \neq y}$, and $c_2 = d$ where $d$ is the Euclidean distance. This case corresponds exactly to the Dudley metric (see Proposition 8.3.4). We remark that as $\varepsilon \to 0$, the approximation error goes also to 0.

# Chapter 9

# Mixed Nash Equilibria in the Adversarial Examples Game

This paper tackles the problem of adversarial examples from a game theoretic point of view. We study the open question of the existence of mixed Nash equilibria in the zero-sum game formed by the attacker and the classifier. While previous works usually allow only one player to use randomized strategies, we show the necessity of considering randomization for both the classifier and the attacker. We demonstrate that this game has no duality gap, meaning that it always admits approximate Nash equilibria. We also provide the first optimization algorithms to learn a mixture of a finite number of classifiers that approximately realizes the value of this game, *i.e.* procedures to build an optimally robust randomized classifier.

This chapter is based on [4].

## 9.1 Introduction

Adversarial examples [102, 103] are one of the most dizzling problems in machine learning: state of the art classifiers are sensitive to imperceptible perturbations of their inputs that make them fail. Last years, research have concentrated on proposing new defense methods [104, 105, 106] and building more and more sophisticated attacks [107, 108, 109, 110]. So far, most defense strategies proved to be vulnerable to these new attacks or are computationally intractable. This asks the following question: can we build classifiers that are robust against any adversarial attack?

A recent line of research argued that randomized classifiers could help countering adversarial attacks [139, 140, 141, 142]. Along this line, [143] demonstrated, using game theory, that randomized classifiers are indeed more robust than deterministic ones against regularized adversaries. However, the findings of these previous works depends on the definition of considered adversary. In particular, they did not investigate scenarios where the adversary also uses randomized strategies, which is essential to account for if we want to give a principled answer to the above question. Previous works studying adversarial examples from the scope of game theory investigated the randomized framework (for both the classifier and the adversary) in restricted settings where the adversary is either parametric or has a finite number of strategies [144, 145, 146]. Our framework does not assume any constraint on the definition of the adversary, making our conclusions independent on the adversary the classifiers are facing. More precisely, we answer the following questions.

**Question 1.** Is it always possible to reach a Mixed Nash equilibrium in the adversarial example game when both the adversary and the classifier can use randomized strategies?

**Answer 1.** We answer positively to this question. First we motivate in Section 9.2 the necessity for using randomized strategies both with the attacker and the classifier. Then, we extend the work of [151], by rigorously reformulating the adversarial risk as a linear optimization problem over distributions. In fact, we cast the adversarial risk minimization problem as a Distributionally Robust Optimization (DRO) [155] problem for a well suited cost function. This formulation naturally leads us, in Section 9.3, to analyze adversarial risk minimization as a zero-sum game. We demonstrate that, in this game, the duality gap always equals 0, meaning that it always admits approximate mixed Nash equilibria.

**Question 2.** Can we design efficient algorithms to learn an optimally robust randomized classifier?

**Answer 2.** To answer this question, we focus on learning a finite mixture of classifiers. Taking inspiration from robust optimization [152] and subgradient methods [156], we derive in Section 9.4 a first oracle algorithm to optimize a finite mixture. Then, following the line of work of [76], we introduce an entropic regularization to effectively compute an approximation of the optimal mixture. We validate our findings with experiments on simulated and real datasets, namely CIFAR-10 an CIFAR-100 [157].

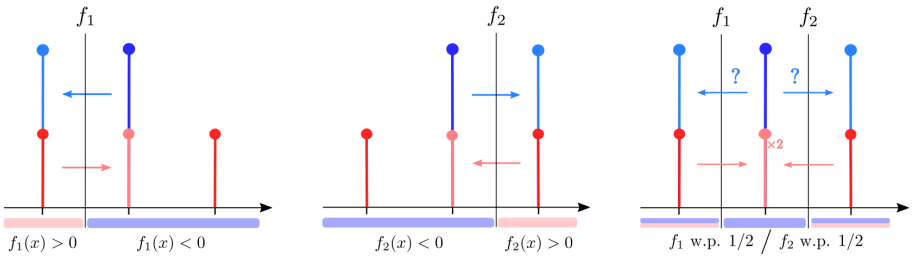



Figure 9.1: Motivating example: blue distribution represents label $-1$ and the red one, label $+1$. The height of columns represents their mass. The red and blue arrows represent the attack on the given classifier. On left: deterministic classifiers ($f_1$ on the left, $f_2$ in the middle) for whose, the blue point can always be attacked. On right: a randomized classifier, where the attacker has a probability $1/2$ of failing, regardless of the attack it selects.

## 9.2 The Adversarial Attack Problem

### 9.2.1 A Motivating Example

Consider the binary classification task illustrated in Figure 9.1. We assume that all input-output pairs $(X, Y)$ are sampled from a distribution $\mathbb{P}$ defined as follows

$$\mathbb{P}(Y = \pm 1) = 1/2 \quad \text{and} \quad \begin{cases} \mathbb{P}(X = 0 \mid Y = -1) = 1 \\ \mathbb{P}(X = \pm 1 \mid Y = 1) = 1/2 \end{cases}$$

Given access to $\mathbb{P}$, the adversary aims to maximize the expected risk, but can only move each point by at most 1 on the real line. In this context, we study two classifiers: $f_1(x) = -x - 1/2$ and $f_2(x) = x - 1/2$[1]. Both $f_1$ and $f_2$ have a standard

[1] $(X, Y) \sim \mathbb{P}$ is misclassified by $f_i$ if and only if $f_i(X)Y \leq 0$

risk of 1/4. In the presence of an adversary, the risk (*a.k.a.* the adversarial risk) increases to 1. Here, using a randomized classifier can make the system more robust. Consider $f$ where $f = f_1$ w.p. 1/2 and $f_2$ otherwise. The standard risk of $f$ remains 1/4 but its adversarial risk is $3/4 < 1$. Indeed, when attacking $f$, any adversary will have to choose between moving points from 0 to 1 or to $-1$. Either way, the attack only works half of the time; hence an overall adversarial risk of 3/4. Furthermore, if $f$ knows the strategy the adversary uses, it can always update the probability it gives to $f_1$ and $f_2$ to get a better (possibly deterministic) defense. For example, if the adversary chooses to always move 0 to 1, the classifier can set $f = f_1$ w.p. 1 to retrieve an adversarial risk of 1/2 instead of 3/4.

Now, what happens if the adversary can use randomized strategies, meaning that for each point it can flip a coin before deciding where to move? In this case, the adversary could decide to move points from 0 to 1 w.p. 1/2 and to $-1$ otherwise. This strategy is still optimal with an adversarial risk of 3/4 but now the classifier cannot use its knowledge of the adversary's strategy to lower the risk. We are in a state where neither the adversary nor the classifier can benefit from unilaterally changing its strategy. In the game theory terminology, this state is called a Mixed Nash equilibrium.

### 9.2.2 General setting

Let us consider a classification task with input space $\mathcal{X}$ and output space $\mathcal{Y}$. Let $(\mathcal{X}, d)$ be a proper (i.e. closed balls are compact) Polish (i.e. completely separable) metric space representing the inputs space[2]. Let $\mathcal{Y} = \{1, \dots, K\}$ be the labels set, endowed with the trivial metric $d'(y, y') = \mathbf{1}_{y \neq y'}$. Then the space $(\mathcal{X} \times \mathcal{Y}, d \oplus d')$ is a proper Polish space. For any Polish space $\mathcal{Z}$, we denote $\mathcal{P}(\mathcal{Z})$ the Polish space of Borel probability measures on $\mathcal{Z}$. Let us assume the data is drawn from $\mathbb{P} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$. Let $(\Theta, d_\Theta)$ be a Polish space (not necessarily proper) representing the set of classifier parameters (for instance neural networks). We also define a loss function: $l : \Theta \times (\mathcal{X} \times \mathcal{Y}) \to [0, \infty)$ satisfying the following set of assumptions.

**Assumption 3** (Loss function)**.** *1) The loss function $l$ is a non negative Borel measurable function. 2) For all $\theta \in \Theta$, $l(\theta, \cdot)$ is upper-semi continuous. 3) There exists $M > 0$ such that for all $\theta \in \Theta$, $(x, y) \in \mathcal{X} \times \mathcal{Y}$, $0 \leq l(\theta, (x, y)) \leq M$.*

It is usual to assume upper-semi continuity when studying optimization over distributions [30, 155]. Furthermore, considering bounded (and positive) loss functions is also very common in learning theory [244] and is not restrictive.

In the adversarial examples framework, the loss of interest is the 0/1 loss, for whose surrogates are misunderstood [149, 148]; hence it is essential that the 0/1

[2]For instance, for any norm $\|\cdot\|$, $(\mathbb{R}^d, \|\cdot\|)$ is a proper Polish metric space.

loss satisfies Assumption 3. In the binary classification setting (*i.e.* $\mathcal{Y} = \{-1, +1\}$) the 0/1 loss writes $l_{0/1}(\theta, (x, y)) = \mathbf{1}_{yf_\theta(x) \leq 0}$. Then, assuming that for all $\theta$, $f_\theta(\cdot)$ is continuous and for all $x$, $f_\cdot(x)$ is continuous, the 0/1 loss satisfies Assumption 3. In particular, it is the case for neural networks with continuous activation functions.

### 9.2.3 Adversarial Risk Minimization

The standard risk for a single classifier $\theta$ associated with the loss $l$ satisfying Assumption 3 writes: $\mathcal{R}(\theta) := \mathbb{E}_{(x,y)\sim\mathbb{P}}\left[l(\theta, (x, y))\right]$. Similarly, the adversarial risk of $\theta$ at level $\varepsilon$ associated with the loss $l$ is defined as[3]

$$\mathcal{R}^{\varepsilon}_{adv}(\theta) := \mathbb{E}_{(x,y)\sim\mathbb{P}}\left[\sup_{x'\in\mathcal{X},\ d(x,x')\leq\varepsilon} l(\theta, (x', y))\right].$$

It is clear that $\mathcal{R}^{0}_{adv}(\theta) = \mathcal{R}(\theta)$ for all $\theta$. We can generalize these notions with distributions of classifiers. In other terms the classifier is then randomized according to some distribution $\mu \in \mathcal{P}(\Theta)$. A classifier is randomized if for a given input, the output of the classifier is a probability distribution. The standard risk of a randomized classifier $\mu$ writes $\mathcal{R}(\mu) = \mathbb{E}_{\theta\sim\mu}\left[\mathcal{R}(\theta)\right]$. Similarly, the adversarial risk of the randomized classifier $\mu$ at level $\varepsilon$ is[4]

$$\mathcal{R}^{\varepsilon}_{adv}(\mu) := \mathbb{E}_{(x,y)\sim\mathbb{P}}\left[\sup_{x'\in\mathcal{X},\ d(x,x')\leq\varepsilon} \mathbb{E}_{\theta\sim\mu}\left[l(\theta, (x', y))\right]\right].$$

For instance, for the 0/1 loss, the inner maximization problem, consists in maximizing the probability of misclassification for a given couple $(x, y)$. Note that $\mathcal{R}(\delta_\theta) = \mathcal{R}(\theta)$ and $\mathcal{R}^{\varepsilon}_{adv}(\delta_\theta) = \mathcal{R}^{\varepsilon}_{adv}(\theta)$. In the remainder of the paper, we study the adversarial risk minimization problems with randomized and deterministic classifiers and denote

$$\mathcal{V}^{\varepsilon}_{rand} := \inf_{\mu\in\mathcal{P}(\Theta)} \mathcal{R}^{\varepsilon}_{adv}(\mu),\ \mathcal{V}^{\varepsilon}_{det} := \inf_{\theta\in\Theta} \mathcal{R}^{\varepsilon}_{adv}(\theta) \tag{9.1}$$

**Remark 13.** *We can show (see Appendix 9.11) that the standard risk infima are equal :* $\mathcal{V}^{0}_{rand} = \mathcal{V}^{0}_{det}$. *Hence, no randomization is needed for minimizing the standard risk. Denoting* $\mathcal{V}$ *this common value, we also have the following inequalities for any* $\varepsilon > 0$, $\mathcal{V} \leq \mathcal{V}^{\varepsilon}_{rand} \leq \mathcal{V}^{\varepsilon}_{det}$.

[3]For the well-posedness, see Lemma 18 in Appendix.

[4]This risk is also well posed (see Lemma 18 in the Appendix).

### 9.2.4 Distributional Formulation of the Adversarial Risk

To account for the possible randomness of the adversary, we rewrite the adversarial attack problem as a convex optimization problem over distributions. Let us first introduce the set of adversarial distributions.

**Definition 5** (Set of adversarial distributions)**.** *Let $\mathbb{P}$ be a Borel probability distribution on $\mathcal{X} \times \mathcal{Y}$ and $\varepsilon > 0$. We define the set of adversarial distributions as*

$$\mathcal{A}_\varepsilon(\mathbb{P}) := \big\{\mathbb{Q} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}) \mid \exists \gamma \in \mathcal{P}\left((\mathcal{X} \times \mathcal{Y})^2\right), \\ d(x, x') \leq \varepsilon,\ y = y' \ \ \gamma\text{-a.s.},\ \pi_{1\sharp}\gamma = \mathbb{P},\ \pi_{2\sharp}\gamma = \mathbb{Q}\big\}$$

*where $\pi_i$ denotes the projection on the $i$-th component, and $g_\sharp$ the push-forward measure by a measurable function $g$.*

An attacker that can move the initial distribution $\mathbb{P}$ anywhere in $\mathcal{A}_\varepsilon(\mathbb{P})$ is not applying a point-wise deterministic perturbation as considered in the standard adversarial risk. In other words, for a point $(x, y) \sim \mathbb{P}$, the attacker could choose a distribution $q(\cdot \mid (x, y))$ whose support is included in $\{(x', y') \mid d(x, x') \leq \epsilon,\ y = y'\}$ from which he will sample the adversarial attack. In this sense, we say the attacker is allowed to be randomized.

**Link with DRO.** Adversarial examples have been studied in the light of DRO by former works [152, 154], but an exact reformulation of the adversarial risk as a DRO problem has not been made yet. When $(\mathcal{Z}, d)$ is a Polish space and $c : \mathcal{Z}^2 \to \mathbb{R}^+ \cup \{+\infty\}$ is a lower semi-continuous function, for $\mathbb{P}, \mathbb{Q} \in \mathcal{P}(\mathcal{Z})$ , the primal Optimal Transport problem is defined as

$$\mathrm{OT}_c(\mathbb{P}, \mathbb{Q}) := \inf_{\gamma \in \Gamma_{\mathbb{P},\mathbb{Q}}} \int_{\mathcal{Z}^2} c(z, z') d\gamma(z, z')$$

with $\Pi(\mathbb{P}, \mathbb{Q}) := \{\gamma \in \mathcal{P}(\mathcal{Z}^2) \mid \ \pi_{1\sharp}\gamma = \mathbb{P},\ \pi_{2\sharp}\gamma = \mathbb{Q}\}$. When $\eta > 0$ and for $\mathbb{P} \in \mathcal{P}(\mathcal{Z})$, the associated Wasserstein uncertainty set is defined as:

$$\mathcal{B}_c(\mathbb{P}, \eta) := \{\mathbb{Q} \in \mathcal{P}(\mathcal{Z}) \mid \mathrm{OT}_c(\mathbb{P}, \mathbb{Q}) \leq \eta\}$$

A DRO problem is a linear optimization problem over Wasserstein uncertainty sets $\sup_{\mathbb{Q} \in \mathcal{B}_c(\mathbb{P},\eta)} \int g(z) d\mathbb{Q}(z)$ for some upper semi-continuous function $g$ [245]. For an arbitrary $\varepsilon > 0$, we define the cost $c_\varepsilon$ as follows

$$c_\varepsilon((x, y), (x', y')) := \begin{cases} 0 & \text{if } d(x, x') \leq \varepsilon \text{ and } y = y' \\ +\infty & \text{otherwise.} \end{cases}$$

This cost is lower semi-continuous and penalizes to infinity perturbations that change the label or move the input by a distance greater than $\varepsilon$. As Proposition 9.2.1 shows, the Wasserstein ball associated with $c_\varepsilon$ is equal to $\mathcal{A}_\varepsilon(\mathbb{P})$.

**Proposition 9.2.1.** *Let $\mathbb{P}$ be a Borel probability distribution on $\mathcal{X} \times \mathcal{Y}$ and $\varepsilon > 0$ and $\eta \geq 0$, then $\mathcal{B}_{c_\varepsilon}(\mathbb{P}, \eta) = \mathcal{A}_\varepsilon(\mathbb{P})$. Moreover, $\mathcal{A}_\varepsilon(\mathbb{P})$ is convex and compact for the weak topology of $\mathcal{P}(\mathcal{X} \times \mathcal{Y})$.*

Thanks to this result, we can reformulate the adversarial risk as the value of a convex problem over $\mathcal{A}_\varepsilon(\mathbb{P})$.

**Proposition 9.2.2.** *Let $\mathbb{P}$ be a Borel probability distribution on $\mathcal{X} \times \mathcal{Y}$ and $\mu$ a Borel probability distribution on $\Theta$. Let $l : \Theta \times (\mathcal{X} \times \mathcal{Y}) \to [0, \infty)$ satisfying Assumption 3. Let $\varepsilon > 0$. Then:*

$$\mathcal{R}^\varepsilon_{adv}(\mu) = \sup_{\mathbb{Q} \in \mathcal{A}_\varepsilon(\mathbb{P})} \mathbb{E}_{(x',y') \sim \mathbb{Q}, \theta \sim \mu} \left[ l(\theta, (x', y')) \right]. \tag{9.2}$$

*The supremum is attained. Moreover $\mathbb{Q}^* \in \mathcal{A}_\varepsilon(\mathbb{P})$ is an optimum of Problem (9.2) if and only if there exists $\gamma^* \in \mathcal{P}\left((\mathcal{X} \times \mathcal{Y})^2\right)$ such that: $\Pi_{1\sharp}\gamma^* = \mathbb{P}$, $\Pi_{2\sharp}\gamma^* = \mathbb{Q}^*$, $d(x, x') \leq \varepsilon$, $y = y'$ and $l(x', y') = \sup_{u \in \mathcal{X}, d(x,u) \leq \varepsilon} l(u, y)$ $\gamma^*$-almost surely.*

The adversarial attack problem is a DRO problem for the cost $c_\varepsilon$. Proposition 9.2.2 means that, against a fixed classifier $\mu$, the randomized attacker that can move the distribution in $\mathcal{A}_\varepsilon(\mathbb{P})$ has exactly the same power as an attacker that moves every single point $x$ in the ball of radius $\varepsilon$. By Proposition 9.2.2, we also deduce that the adversarial risk can be casted as a linear optimization problem over distributions.

**Remark 14.** *In a recent work, [151] proposed a similar adversary using Markov kernels but left as an open question the link with the classical adversarial risk, due to measurability issues. Proposition 9.2.2 solves these issues. The result is similar to [155]. Although we believe its proof might be extended for infinite valued costs, [155] did not treat that case. We provide an alternative proof in this special case.*

## 9.3 Nash Equilibria in the Adversarial Game

### 9.3.1 Adversarial Attacks as a Zero-Sum Game

Thanks to Proposition 9.2, the adversarial risk minimization problem can be seen as a two-player zero-sum game that writes as follows,

$$\inf_{\mu \in \mathcal{P}(\Theta)} \sup_{\mathbb{Q} \in \mathcal{A}_\varepsilon(\mathbb{P})} \mathbb{E}_{(x,y) \sim \mathbb{Q}, \theta \sim \mu} \left[ l(\theta, (x, y)) \right]. \tag{9.3}$$

In this game the classifier objective is to find the best distribution $\mu \in \mathcal{P}(\Theta)$ while the adversary is manipulating the data distribution. For the classifier, solving the

infimum problem in Equation (9.3) simply amounts to solving the adversarial risk minimization problem – Problem (9.1), whether the classifier is randomized or not. Then, given a randomized classifier $\mu \in \mathcal{P}(\Theta)$, the goal of the attacker is to find a new data-set distribution $\mathbb{Q}$ in the set of adversarial distributions $\mathcal{A}_\varepsilon(\mathbb{P})$ that maximizes the risk of $\mu$. More formally, the adversary looks for

$$\mathbb{Q} \in \underset{\mathbb{Q}\in\mathcal{A}_\varepsilon(\mathbb{P})}{\operatorname{argmax}} \mathbb{E}_{(x,y)\sim\mathbb{Q},\theta\sim\mu}\left[l(\theta,(x,y))\right].$$

In the game theoretic terminology, $\mathbb{Q}$ is also called the best response of the attacker to the classifier $\mu$.

**Remark 15.** *Note that for a given classifier $\mu$ there always exists a "deterministic" best response, i.e. every single point $(x,y)$ is mapped to another single point $T(x,y)$. Let $T : \mathcal{X}\times\mathcal{Y} \to \mathcal{X}$ be defined such that for all $(x,y) \in \mathcal{X}\times\mathcal{Y}$, $l(T(x,y),y) = \sup_{x',\ d(x,x')\leq\varepsilon} l(x',y)$. Thanks to [246, Proposition 7.50], $(T,id)$ is $\mathbb{P}$-measurable. Moreover, we get that $\mathbb{Q} = (T,id)_\sharp\mathbb{P}$ belongs to the best response to $\mu$. Therefore, $T$ is the optimal "deterministic" attack against the classifier $\mu$.*

### 9.3.2 Dual Formulation of the Game

Every zero sum game has a dual formulation that allows a deeper understanding of the framework. Here, from Proposition 9.2.2, we can define the dual problem of adversarial risk minimization for randomized classifiers. This dual problem also characterizes a two-player zero-sum game that writes as follows,

$$\sup_{\mathbb{Q}\in\mathcal{A}_\varepsilon(\mathbb{P})} \inf_{\mu\in\mathcal{P}(\Theta)} \mathbb{E}_{(x,y)\sim\mathbb{Q},\theta\sim\mu}\left[l(\theta,(x,y))\right]. \tag{9.4}$$

In this dual game problem, the adversary plays first and seeks an adversarial distribution that has the highest possible risk when faced with an arbitrary classifier. This means that it has to select an adversarial perturbation for every input $x$, without seeing the classifier first. In this case, as pointed out by the motivating example in Section 9.2.1, the attack can (and should) be randomized to ensure maximal harm against several classifiers. Then, given an adversarial distribution, the classifier objective is to find the best possible classifier on this distribution. Let us denote $\mathcal{D}^\varepsilon$ the value of the dual problem. Since the weak duality is always satisfied, we get

$$\mathcal{D}^\varepsilon \leq \mathcal{V}^\varepsilon_{rand} \leq \mathcal{V}^\varepsilon_{det}. \tag{9.5}$$

Inequalities in Equation (9.5) mean that the lowest risk the classifier can get (regardless of the game we look at) is $\mathcal{D}^\varepsilon$. In particular, this means that the primal

version of the game, *i.e.* the adversarial risk minimization problem, will always have a value greater or equal to $\mathcal{D}^\varepsilon$. As we discussed in Section 9.2.1, this lower bound may not be attained by a deterministic classifier. As we will demonstrate in the next section, optimizing over randomized classifiers allows to approach $\mathcal{D}^\varepsilon$ arbitrary closely.

**Remark 16.** *Note that, we can always define the dual problem when the classifier is deterministic,*

$$\sup_{\mathbb{Q}\in\mathcal{A}_\varepsilon(\mathbb{P})} \inf_{\theta\in\Theta} \mathbb{E}_{(x,y)\sim\mathbb{Q}} \left[l(\theta,(x,y))\right].$$

*Furthermore, we can demonstrate that the dual problems for deterministic and randomized classifiers have the same value* [5]*; hence the inequalities in Equation* (9.5)*.*

### 9.3.3 Nash Equilibria for Randomized Strategies

In the adversarial examples game, a Nash equilibrium is a couple $(\mu^*, \mathbb{Q}^*) \in \mathcal{P}(\Theta) \times \mathcal{A}_\varepsilon(\mathbb{P})$ where both the classifier and the attacker have no incentive to deviate unilaterally from their strategies $\mu^*$ and $\mathbb{Q}^*$. More formally, $(\mu^*, \mathbb{Q}^*)$ is a Nash equilibrium of the adversarial examples game if $(\mu^*, \mathbb{Q}^*)$ is a saddle point of the objective function

$$(\mu, \mathbb{Q}) \mapsto \mathbb{E}_{(x,y)\sim\mathbb{Q},\theta\sim\mu} \left[l(\theta,(x,y))\right].$$

Alternatively, we can say that $(\mu^*, \mathbb{Q}^*)$ is a Nash equilibrium if and only if $\mu^*$ solves the adversarial risk minimization problem – Problem (9.1), $\mathbb{Q}^*$ the dual problem – Problem (9.6), and $\mathcal{D}^\varepsilon = \mathcal{V}^\varepsilon_{rand}$. In our problem, $\mathbb{Q}^*$ always exists but it might not be the case for $\mu^*$. Then for any $\delta > 0$, we say that $(\mu_\delta, \mathbb{Q}^*)$ is a $\delta$-approximate Nash equilibrium if $\mathbb{Q}^*$ solves the dual problem and $\mu_\delta$ satisfies $\mathcal{D}^\varepsilon \geq \mathcal{R}^\varepsilon_{adv}(\mu_\delta) - \delta$.

We now state our main result: the existence of approximate Nash equilibria in the adversarial examples game when both the classifier and the adversary can use randomized strategies. More precisely, we demonstrate that the duality gap between the adversary and the classifier problems is zero, which gives as a corollary the existence of Nash equilibria.

**Theorem 9.3.1.** *Let $\mathbb{P} \in \mathcal{P}(\mathcal{X}\times\mathcal{Y})$. Let $\varepsilon > 0$. Let $l : \Theta \times (\mathcal{X}\times\mathcal{Y}) \to [0,\infty)$ satisfying Assumption 3. Then strong duality always holds in the randomized setting:*

$$\begin{aligned} &\inf_{\mu\in\mathcal{P}(\Theta)} \max_{\mathbb{Q}\in\mathcal{A}_\varepsilon(\mathbb{P})} \mathbb{E}_{\theta\sim\mu,(x,y)\sim\mathbb{Q}} \left[l(\theta,(x,y))\right] \\ = &\max_{\mathbb{Q}\in\mathcal{A}_\varepsilon(\mathbb{P})} \inf_{\mu\in\mathcal{P}(\Theta)} \mathbb{E}_{\theta\sim\mu,(x,y)\sim\mathbb{Q}} \left[l(\theta,(x,y))\right] \end{aligned} \tag{9.6}$$

[5] See Appendix 9.11 for more details

*The supremum is always attained. If $\Theta$ is a compact set, and for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$, $l(\cdot, (x, y))$ is lower semi-continuous, the infimum is also attained.*

**Corollary 1.** *Under Assumption 3, for any $\delta > 0$, there exists a $\delta$-approximate Nash-Equibilrium $(\mu_\delta, \mathbb{Q}^*)$. Moreover, if the infimum is attained, there exists a Nash equilibrium $(\mu^*, \mathbb{Q}^*)$ to the adversarial examples game.*

Theorem 9.3.1 shows that $\mathcal{D}^\varepsilon = \mathcal{V}^\varepsilon_{rand}$. From a game theoretic perspective, this means that the minimal adversarial risk for a randomized classifier against any attack (primal problem) is the same as the maximal risk an adversary can get by using an attack strategy that is oblivious to the classifier it faces (dual problem). This suggests that playing randomized strategies for the classifier could substantially improve robustness to adversarial examples. In the next section, we will design an algorithm that efficiently learn a randomized classifier and show improved adversarial robustness over classical deterministic defenses.

**Remark 17.** *Theorem 9.3.1 remains true if one replaces $\mathcal{A}_\varepsilon(\mathbb{P})$ with any other Wasserstein compact uncertainty sets (see [245] for conditions of compactness).*

## 9.4 Finding the Optimal Classifiers

### 9.4.1 An Entropic Regularization

Let $\{(x_i, y_i)\}_{i=1}^N$ samples independently drawn from $\mathbb{P}$ and denote $\widehat{\mathbb{P}} := \frac{1}{N}\sum_{i=1}^N \delta_{(x_i,y_i)}$ the associated empirical distribution. One can show the adversarial empirical risk minimization can be casted as:

$$\widehat{\mathcal{R}}^{\varepsilon,*}_{adv} := \inf_{\mu\in\mathcal{P}(\Theta)} \sum_{i=1}^N \sup_{\mathbb{Q}_i\in\Gamma_{i,\varepsilon}} \mathbb{E}_{(x,y)\sim\mathbb{Q}_i,\theta\sim\mu}\left[l(\theta,(x,y))\right]$$

where $\Gamma_{i,\varepsilon}$ is defined as :

$$\Gamma_{i,\varepsilon} := \Big\{\mathbb{Q}_i \mid \int d\mathbb{Q}_i = \frac{1}{N},\ \int c_\varepsilon((x_i,y_i),\cdot)d\mathbb{Q}_i = 0\Big\}.$$

More details on this decomposition are given in Appendix 9.11. In the following, we regularize the above objective by adding an entropic term to each inner supremum problem. Let $\boldsymbol{\alpha} := (\alpha_i)_{i=1}^N \in \mathbb{R}_+^N$ such that for all $i \in \{1, \dots, N\}$, and let us consider the following optimization problem:

$$\widehat{\mathcal{R}}^{\varepsilon,*}_{adv,\boldsymbol{\alpha}} := \inf_{\mu\in\mathcal{P}(\Theta)} \sum_{i=1}^N \sup_{\mathbb{Q}_i\in\Gamma_{i,\varepsilon}} \mathbb{E}_{\mathbb{Q}_i,\mu}\left[l(\theta,(x,y))\right] - \alpha_i \mathrm{KL}\left(\mathbb{Q}_i \Big|\Big| \frac{1}{N}\mathbb{U}_{(x_i,y_i)}\right)$$

where $\mathbb{U}_{(x,y)}$ is an arbitrary distribution of support equal to:

$$S^{(\varepsilon)}_{(x,y)} := \Big\{(x',y') : \text{ s.t. } c_\varepsilon((x,y),(x',y')) = 0\Big\},$$

and for all $\mathbb{Q}, \mathbb{U} \in \mathcal{M}_+(\mathcal{X} \times \mathcal{Y})$,

$$\mathrm{KL}(\mathbb{Q}, \mathbb{U}) := \begin{cases} \int \log(\frac{d\mathbb{Q}}{d\mathbb{U}}) d\mathbb{Q} + |\mathbb{U}| - |\mathbb{Q}| & \text{if } \mathbb{Q} \ll \mathbb{U} \\ +\infty & \text{otherwise.} \end{cases}$$

Note that when $\boldsymbol{\alpha} = 0$, we recover the problem of interest $\widehat{\mathcal{R}}^{\varepsilon,*}_{adv,\boldsymbol{\alpha}} = \widehat{\mathcal{R}}^{\varepsilon,*}_{adv}$. Moreover, we show the regularized supremum tends to the standard supremum when $\boldsymbol{\alpha} \to 0$.

**Proposition 9.4.1.** *For $\mu \in \mathcal{P}(\Theta)$, one has*

$$\begin{aligned} &\lim_{\alpha_i \to 0} \sup_{\mathbb{Q}_i \in \Gamma_{i,\varepsilon}} \mathbb{E}_{\mathbb{Q}_i,\mu}\left[l(\theta,(x,y))\right] - \alpha_i KL\left(\mathbb{Q}, \frac{1}{N}\mathbb{U}_{(x_i,y_i)}\right) \\ &= \sup_{\mathbb{Q}_i \in \Gamma_{i,\varepsilon}} \mathbb{E}_{(x,y)\sim\mathbb{Q}_i,\theta\sim\mu}\left[l(\theta,(x,y))\right]. \end{aligned}$$

By adding an entropic term to the objective, we obtain an explicit formulation of the supremum involved in the sum: as soon as $\boldsymbol{\alpha} > 0$ (which means that each $\alpha_i > 0$), each sub-problem becomes just the Fenchel-Legendre transform of $\mathrm{KL}(\cdot, \mathbb{U}_{(x_i,y_i)}/N)$ which has the following closed form:

$$\begin{aligned} &\sup_{\mathbb{Q}_i \in \Gamma_{i,\varepsilon}} \mathbb{E}_{\mathbb{Q}_i,\mu}\left[l(\theta,(x,y))\right] - \alpha_i \mathrm{KL}\left(\mathbb{Q}_i, \frac{1}{N}\mathbb{U}_{(x_i,y_i)}\right) \\ &= \frac{\alpha_i}{N} \log\left(\int_{\mathcal{X}\times\mathcal{Y}} \exp\left(\frac{\mathbb{E}_{\theta\sim\mu}\left[l(\theta,(x,y))\right]}{\alpha_i}\right) d\mathbb{U}_{(x_i,y_i)}\right). \end{aligned}$$

Finally, we end up with the following problem:

$$\inf_{\mu\in\mathcal{P}(\Theta)} \sum_{i=1}^{N} \frac{\alpha_i}{N} \log\left(\int \exp \frac{\mathbb{E}_\mu\left[l(\theta,(x,y))\right]}{\alpha_i} d\mathbb{U}_{(x_i,y_i)}\right).$$

In order to solve the above problem, one needs to compute the integral involved in the objective. To do so, we estimate it by randomly sampling $m_i \geq 1$ samples $(u_1^{(i)}, \ldots, u_{m_i}^{(i)}) \in (\mathcal{X} \times \mathcal{Y})^{m_i}$ from $\mathbb{U}_{(x_i,y_i)}$ for all $i \in \{1, \ldots, N\}$ which leads to the following optimization problem

$$\inf_{\mu\in\mathcal{P}(\Theta)} \sum_{i=1}^{N} \frac{\alpha_i}{N} \log\left(\frac{1}{m_i}\sum_{j=1}^{m_i} \exp \frac{\mathbb{E}_\mu\left[l(\theta, u_j^{(i)})\right]}{\alpha_i}\right) \tag{9.7}$$

denoted $\widehat{\mathcal{R}}^{\varepsilon,\boldsymbol{m}}_{adv,\boldsymbol{\alpha}}$ where $\boldsymbol{m} := (m_i)_{i=1}^N$ in the following. Now we aim at controlling the error made with our approximations. We decompose the error into two terms

$$|\widehat{\mathcal{R}}^{\varepsilon,\boldsymbol{m}}_{adv,\boldsymbol{\alpha}} - \widehat{\mathcal{R}}^{\varepsilon,*}_{adv}| \leq |\widehat{\mathcal{R}}^{\varepsilon,*}_{adv,\boldsymbol{\alpha}} - \widehat{\mathcal{R}}^{\varepsilon,\boldsymbol{m}}_{adv,\boldsymbol{\alpha}}| + |\widehat{\mathcal{R}}^{\varepsilon,*}_{adv,\boldsymbol{\alpha}} - \widehat{\mathcal{R}}^{\varepsilon,*}_{adv}|$$

where the first one corresponds to the statistical error made by our estimation of the integral, and the second to the approximation error made by the entropic regularization of the objective. First, we show a control of the statistical error using Rademacher complexities [244].

**Proposition 9.4.2.** *Let $m \geq 1$ and $\alpha > 0$ and denote $\boldsymbol{\alpha} := (\alpha, \dots, \alpha) \in \mathbb{R}^N$ and $\boldsymbol{m} := (m, \dots, m) \in \mathbb{R}^N$. Then by denoting $\tilde{M} = \max(M, 1)$, we have with a probability of at least $1 - \delta$*

$$|\widehat{\mathcal{R}}^{\varepsilon,*}_{adv,\boldsymbol{\alpha}} - \widehat{\mathcal{R}}^{\varepsilon,\boldsymbol{m}}_{adv,\boldsymbol{\alpha}}| \leq \frac{2e^{M/\alpha}}{N} \sum_{i=1}^{N} R_i + 6\tilde{M}e^{M/\alpha}\sqrt{\frac{\log(\frac{4}{\delta})}{2mN}}$$

*where $R_i := \frac{1}{m}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\theta\in\Theta}\sum_{j=1}^m \sigma_j l(\theta, u_j^{(i)})\right]$ and $\boldsymbol{\sigma} := (\sigma_1, \dots, \sigma_m)$ with $\sigma_i$ i.i.d. sampled as $\mathbb{P}[\sigma_i = \pm 1] = 1/2$.*

We deduce from the above Proposition that in the particular case where $\Theta$ is finite such that $|\Theta| = L$, with probability of at least $1 - \delta$

$$|\widehat{\mathcal{R}}^{\varepsilon,*}_{adv,\boldsymbol{\alpha}} - \widehat{\mathcal{R}}^{\varepsilon,\boldsymbol{m}}_{adv,\boldsymbol{\alpha}}| \in \mathcal{O}\left(Me^{M/\alpha}\sqrt{\frac{\log(L)}{m}}\right).$$

This case is of particular interest when one wants to learn the optimal mixture of some given classifiers in order to minimize the adversarial risk. In the following proposition, we control the approximation error made by adding an entropic term to the objective.

**Proposition 9.4.3.** *Denote for $\beta > 0$, $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and $\mu \in \mathcal{P}(\Theta)$, $A^{(x,y)}_{\beta,\mu} := \{u | \sup_{v\in S^{(\varepsilon)}_{(x,y)}} \mathbb{E}_\mu[l(\theta, v)] \leq \mathbb{E}_\mu[l(\theta, u)] + \beta\}$. If there exists $C_\beta$ such that for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and $\mu \in \mathcal{P}(\Theta)$, $\mathbb{U}_{(x,y)}\left(A^{(x,y)}_{\beta,\mu}\right) \geq C_\beta$ then we have*

$$|\widehat{\mathcal{R}}^{\varepsilon,*}_{adv,\boldsymbol{\alpha}} - \widehat{\mathcal{R}}^{\varepsilon,*}_{adv}| \leq 2\alpha|\log(C_\beta)| + \beta.$$

The assumption made in the above Proposition states that for any given random classifier $\mu$, and any given point $(x, y)$, the set of $\beta$-optimal attacks at this point has at least a certain amount of mass depending on the $\beta$ chosen. This assumption is always met when $\beta$ is sufficiently large. However in order to obtain a tight control of the error, a trade-off exists between $\beta$ and the smallest amount of mass $C_\beta$ of $\beta$-optimal attacks. Now that we have shown that solving (9.7) allows to obtain an approximation of the true solution $\widehat{\mathcal{R}}^{\varepsilon,*}_{adv}$, we next aim at deriving an algorithm to compute it.

### 9.4.2 Proposed Algorithms

From now on, we focus on finite class of classifiers. Let $\Theta = \{\theta_1, \dots, \theta_L\}$, we aim to learn the optimal mixture of classifiers in this case. The adversarial empirical risk is therefore defined as:

$$\widehat{\mathcal{R}}^{\varepsilon}_{adv}(\boldsymbol{\lambda}) = \sum_{i=1}^{N} \sup_{\mathbb{Q}_i \in \Gamma_{i,\varepsilon}} \mathbb{E}_{(x,y)\sim\mathbb{Q}_i} \left[ \sum_{k=1}^{L} \lambda_k l(\theta_k, (x,y)) \right]$$

for $\boldsymbol{\lambda} \in \Delta_L := \{\boldsymbol{\lambda} \in \mathbb{R}^L_+ \text{ s.t. } \sum_{i=1}^{L} \lambda_i = 1\}$, the probability simplex of $\mathbb{R}^L$. One can notice that $\widehat{\mathcal{R}}^{\varepsilon}_{adv}(\cdot)$ is a continuous convex function, hence $\min_{\boldsymbol{\lambda}\in\Delta_L} \mathcal{R}^{\varepsilon}_{adv}(\boldsymbol{\lambda})$ is attained for a certain $\boldsymbol{\lambda}^*$. Then there exists a non-approximate Nash equilibrium $(\boldsymbol{\lambda}^*, \mathbb{Q}^*)$ in the adversarial game when $\Theta$ is finite. Here, we present two algorithms to learn the optimal mixture of the adversarial risk minimization problem.

**Algorithm 17** Oracle-based Algorithm

$\boldsymbol{\lambda}_0 = \frac{\mathbf{1}_L}{L}; T;\ \eta = \frac{2}{M\sqrt{LT}}$

**for** $t = 1, \dots, T$ **do**

  $\tilde{\mathbb{Q}}$ s.t. $\exists \mathbb{Q}^* \in \mathcal{A}_\varepsilon(\mathbb{P})$ best response to $\boldsymbol{\lambda}_{t-1}$ and for all $k \in [L]$, $|\mathbb{E}_{\tilde{\mathbb{Q}}}(l(\theta_k, (x,y))) - \mathbb{E}_{\mathbb{Q}^*}(l(\theta_k, (x,y)))| \leq \delta$

  $\boldsymbol{g}_t = \left(\mathbb{E}_{\tilde{\mathbb{Q}}}(l(\theta_1, (x,y)), \dots, \mathbb{E}_{\tilde{\mathbb{Q}}}(l(\theta_L, (x,y))\right)^T$

  $\boldsymbol{\lambda}_t = \Pi_{\Delta_L}\left(\boldsymbol{\lambda}_{t-1} - \eta \boldsymbol{g}_t\right)$

**end**

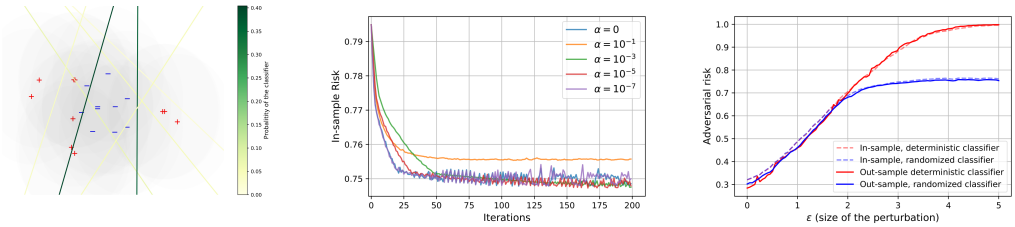



Figure 9.2: On left, 40 data samples with their set of possible attacks represented in shadow and the optimal randomized classifier, with a color gradient representing the probability of the classifier. In the middle, convergence of the oracle ($\alpha = 0$) and regularized algorithm for different values of regularization parameters. On right, in-sample and out-sample risk for randomized and deterministic minimum risk in function of the perturbation size $\varepsilon$. In the latter case, the randomized classifier is optimized with oracle Algorithm 17.

**An Entropic Relaxation.** Using the results from Section 9.4.1, adding an entropic term to the objective allows to have a simple reformulation of the problem, as follows:

$$\inf_{\boldsymbol{\lambda}\in\Delta_L} \sum_{i=1}^{N} \frac{\varepsilon_i}{N} \log\left(\frac{1}{m_i}\sum_{j=1}^{m_i} \exp\left(\frac{\sum_{k=1}^{L}\lambda_k l(\theta_k, u_j^{(i)})}{\varepsilon_i}\right)\right)$$

Note that in $\boldsymbol{\lambda}$, the objective is convex and smooth. One can apply the accelerated PGD [236, 237] which enjoys an optimal convergence rate for first order methods of $\mathcal{O}(T^{-2})$ for $T$ iterations.

**A First Oracle Algorithm.** Indepedently from the entropic regularization,we present an oracle-based algorithm inspired from [152] and the convergence of projected sub-gradient methods [156]. The computation of the inner supremum problem is usually NP-hard [6], but one may assume the existence of an approximate oracle to this supremum. The algorithm is presented in Algorithm 17. We get the following guarantee for this algorithm.

**Proposition 9.4.4.** *Let $l : \Theta \times (\mathcal{X} \times \mathcal{Y}) \to [0, \infty)$ satisfying Assumption 3. Then, Algorithm 17 satisfies:*

$$\min_{t\in[T]} \widehat{\mathcal{R}}^{\varepsilon}_{adv}(\boldsymbol{\lambda}_t) - \widehat{\mathcal{R}}^{\varepsilon,*}_{adv} \leq 2\delta + \frac{2M\sqrt{L}}{\sqrt{T}}$$

The main drawback of the above algorithm is that one needs to have access to an oracle to guarantee the convergence of the proposed algorithm. In the following we present its regularized version in order to approximate the solution and propose a simple algorithm to solve it.

### 9.4.3 A General Heuristic Algorithm

So far, our algorithms are not easily practicable in the case of deep learning. Adversarial examples are known to be easily transferrable from one model to another [247, 248]. So we aim at learning diverse models. To this end, and support our theoretical claims, we propose an heuristic algorithm (see Algorithm 18) to train a robust mixture of $L$ classifiers. We alternatively train these classifiers with adversarial examples against the current mixture and update the probabilities of the mixture according to the algorithms we proposed in Section 9.4.2. More details on this algorithm are available in Appendix 9.10.

[6] See Appendix 9.11 for details.

**Algorithm 18** Adversarial Training for Mixtures

$L$: number of models, $T$: number of iterations,
$T_\theta$: number of updates for the models $\boldsymbol{\theta}$,
$T_\lambda$: number of updates for the mixture $\boldsymbol{\lambda}$,
$\boldsymbol{\lambda}_0 = (\lambda_0^1, \dots \lambda_0^L)$, $\boldsymbol{\theta}_0 = (\theta_0^1, \dots \theta_0^L)$
**for** $t = 1, \dots, T$ **do**
  Let $B_t$ be a batch of data.
  **if** $t \mod (T_\theta L + 1) \neq 0$ **then**
    $k$ sampled uniformly in $\{1, \dots, L\}$
    $\tilde{B}_t \leftarrow$ Attack of images in $B_t$ for the model $(\boldsymbol{\lambda}_t, \boldsymbol{\theta}_t)$
    $\theta_k^t \leftarrow$ Update $\theta_k^{t-1}$ with $\tilde{B}_t$ for fixed $\boldsymbol{\lambda}_t$ with a SGD step
  **else**
    $\boldsymbol{\lambda}_t \leftarrow$Update $\boldsymbol{\lambda}_{t-1}$ on $B_t$ for fixed $\boldsymbol{\theta}_t$ with oracle-based or regularized algorithm with $T_\lambda$ iterations.
  **end**
**end**

## 9.5 Experiments

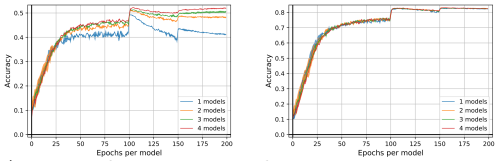
### 9.5.1 Synthetic Dataset

To illustrate our theoretical findings, we start by testing our learning algorithm on the following synthetic two-dimensional problem. Let us consider the distribution $\mathbb{P}$ defined as $\mathbb{P}(Y = \pm 1) = 1/2$, $\mathbb{P}(X \mid Y = -1) = \mathcal{N}(0, I_2)$ and $\mathbb{P}(X \mid Y = 1) = \frac{1}{2}[\mathcal{N}((-3, 0), I_2) + \mathcal{N}((3, 0), I_2)]$. We sample 1000 training points from this distribution and randomly generate 10 linear classifiers that achieves a standard training risk lower than 0.4. To simulate an adversary with budget $\varepsilon$ in $\ell_2$ norm, we proceed as follows. For every sample $(x, y) \sim \mathbb{P}$ we generate 1000 points uniformly at random in the ball of radius $\varepsilon$ and select the one maximizing the risk for the 0/1 loss. Figure 9.2 (left) illustrates the type of mixture we get after convergence of our algorithms. Note that in this toy problem, we are likely to find the optimal adversary with this sampling strategy if we sample enough attack points.

To evaluate the convergence of our algorithms, we compute the adversarial risk of our mixture for each iteration of both the oracle and regularized algorithms. Figure 9.2 illustrates the convergence of the algorithms w.r.t the regularization parameter. We observe that the risk for both algorithms converge. Moreover, they converge towards the oracle minimizer when the regularization parameter $\alpha$ goes to 0.

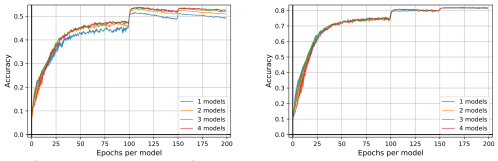
Finally, to demonstrate the improvement randomized techniques offer against deterministic defenses, we plot in Figure 9.2 (right) the minimum adversarial risk for both randomized and deterministic classifiers w.r.t. $\varepsilon$. The adversarial risk is

| Models | Acc. | $\text{APGD}_{\text{CE}}$ | $\text{APGD}_{\text{DLR}}$ | Rob. Acc |
|---|---|---|---|---|
| 1 | 81.9% | 47.6% | 47.7% | 45.6% |
| 2 | 81.9% | 49.0% | 49.6% | 47.0% |
| 3 | 81.7% | 49.0% | 49.3% | 46.9% |
| 4 | **82.6%** | **49.7%** | **49.8%** | **47.2%** |



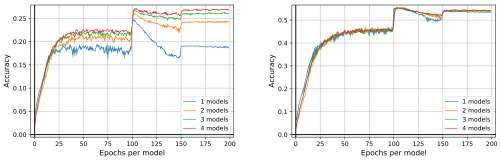

Adversarial Training, CIFAR-10 dataset results

| Models | Acc. | $\text{APGD}_{\text{CE}}$ | $\text{APGD}_{\text{DLR}}$ | Rob. Acc |
|---|---|---|---|---|
| 1 | 79.6% | 50.9% | 48.9% | 48.3% |
| 2 | 80.3% | 52.3% | 51.2% | 50.2% |
| 3 | 80.7% | 52.8% | 51.7% | 50.7% |
| 4 | **80.9%** | **53.0%** | **51.8%** | **50.8%** |




TRADES, CIFAR-10 dataset results

| Models | Acc. | $\text{APGD}_{\text{CE}}$ | $\text{APGD}_{\text{DLR}}$ | Rob. Acc |
|---|---|---|---|---|
| 1 | 55.2% | 24.1% | 23.8% | 22.5% |
| 2 | 55.2% | 25.3% | 26.1% | 23.6% |
| 3 | **55.4%** | 25.7% | 26.8% | 24.2% |
| 4 | 55.3% | **26.0%** | **27.5%** | **24.5%** |




Adversarial Training, CIFAR-100 dataset results

Figure 9.3: Upper plots: Adversarial Training, CIFAR-10 dataset results. Middle plots: TRADES, CIFAR-10 dataset results. Bottom plots: CIFAR-100 dataset results. On left: Comparison of our algorithm with a standard adversarial training (one model). We reported the results for the model with the best robust accuracy obtained over two independent runs because adversarial training might be unstable. Standard and Robust accuracy (respectively in the middle and on right) on CIFAR-10 test images in function of the number of epochs per classifier with 1 to 3 ResNet18 models. The performed attack is PGD with 20 iterations and $\varepsilon = 8/255$.

strictly better for randomized classifier whenever the adversarial budget $\varepsilon$ is bigger than 2. This illustration validates our analysis of Theorem 9.3.1, and motivates a in depth study of a more challenging framework, namely image classification with neural networks.

### 9.5.2 CIFAR Datasets

**Experimental Setup.** We now implement our heuristic algorithm (Alg. 18) on CIFAR-10 and CIFAR-100 datasets for both Adversarial Traning [104] and TRADES [249] loss. To evaluate the performance of Algorithm 18, we trained from 1 to 4 ResNet18 [250] models on 200 epochs per model[7]. We study the

[7] $L \times 200$ epochs in total, where $L$ is the number of models.

robustness with regards to $\ell_\infty$ norm and fixed adversarial budget $\varepsilon = 8/255$. The attack we used in the inner maximization of the training is an adapted (adaptative) version of PGD for mixtures of classifiers with 10 steps. Note that for one single model, Algorithm 18 exactly corresponds to adversarial training [104] or TRADES. For each of our setups, we made two independent runs and select the best one. The training time of our algorithm is around four times longer than a standard Adversarial Training (with PGD 10 iter.) with two models, eight times with three models and twelve times with four models. We trained our models with a batch of size 1024 on 8 Nvidia V100 GPUs. We give more details on implementation in Appendix 9.10.

**Evaluation Protocol.** At each epoch, we evaluate the current mixture on test data against PGD attack with 20 iterations. To select our model and avoid overfitting [251], we kept the most robust against this PGD attack. To make a final evaluation of our mixture of models, we used an adapted version of AutoPGD untargeted attacks [110] for randomized classifiers with both Cross-Entropy (CE) and Difference of Logits Ratio (DLR) loss. For both attacks, we made 100 iterations and 5 restarts.

**Results.** The results are presented in Figure 9.3. We remark our algorithm outperforms a standard adversarial training in all the cases by more 1% on CIFAR-10 and CIFAR-100, without additional loss of standard accuracy as it is attested by the left figures. On TRADES, the gain is even more important by more than 2% in robust accuracy. Moreover, it seems our algorithm, by adding more and more models, reduces the overfitting of adversarial training. It also appears that robustness increases as the number of models increases. So far, experiments are computationally very costful and it is difficult to raise precise conclusions. Further, hyperparameter tuning [252] such as architecture, unlabeled data [253] or activation function may still increase the results.

## 9.6 Related Work and Discussions

**Distributionally Robust Optimization.** Several recent works [152, 153, 154] studied the problem of adversarial examples through the scope of distributionally robust optimization. In these frameworks, the set of adversarial distributions is defined using an $\ell_p$ Wasserstein ball (the adversary is allowed to have an *average* perturbation of at most $\varepsilon$ in $\ell_p$ norm). This however does not match the usual adversarial attack problem, where the adversary cannot move any point by more than $\varepsilon$. In the present work, we introduce a cost function allowing us to cast

the adversarial example problem as a DRO one, without changing the adversary constraints.

**Optimal Transport (OT).** Bhagoji et al. [150] and Pydi and Jog [151] investigated classifier-agnostic lower bounds on the adversarial risk of any deterministic classifier using OT. These works only evaluate lower bounds on the primal deterministic formulation of the problem, while we study the existence of mixed Nash equilibria. Note that Pydi and Jog [151] started to investigate a way to formalize the adversary using Markov kernels, but did not investigate the impact of randomized strategies on the game. We extended this work by rigorously reformulating the adversarial risk as a linear optimization problem over distributions and we study this problem from a game theoretic point of view.

**Game Theory.** Adversarial examples have been studied under the notions of Stackelberg game in [147], and zero-sum game in [144, 145, 146]. These works considered restricted settings (convex loss, parametric adversaries, etc.) that do not comply with the nature of the problem. Indeed, we prove in Appendix 9.9.3 that when the loss is convex and the set $\Theta$ is convex, the duality gap is zero for deterministic classifiers. However, it has been proven that no convex loss can be a good surrogate for the 0/1 loss in the adversarial setting [148, 149], narrowing the scope of this result. If one can show that for sufficiently separated conditional distributions, an optimal deterministic classifier always exists (see Appendix 9.11 for a clear statement), necessary and sufficient conditions for the need of randomization are still to be established. [143] studied partly this question for regularized deterministic adversaries, leaving the general setting of randomized adversaries and mixed equilibria unanswered, which is the very scope of this paper.

# Supplementary material

## 9.7 Notations

Let $(\mathcal{Z}, d)$ be a Polish metric space (i.e. complete and separable). We say that $(\mathcal{Z}, d)$ is proper if for all $z_0 \in \mathcal{Z}$ and $R > 0$, $B(z_0, R) := \{z \mid d(z, z_0) \leq R\}$ is compact. For $(\mathcal{Z}, d)$ a Polish space, we denote $\mathcal{P}(\mathcal{Z})$ the set of Borel probability measures on $\mathcal{Z}$ endowed with $\|\cdot\|_{TV}$ strong topology. We recall the notion of weak topology: we say that a sequence $(\mu_n)_n$ of $\mathcal{P}(\mathcal{Z})$ converges weakly to $\mu \in \mathcal{P}(\mathcal{Z})$ if and only if for every bounded continuous function $f$ on $\mathcal{Z}$, $\int f d\mu_n \to_{n\to\infty} \int f d\mu$. Endowed with its weak topology, $\mathcal{P}(\mathcal{Z})$ is a Polish space. For $\mu \in \mathcal{P}(\mathcal{Z})$, we define $L^1(\mu)$ the set of integrable functions with respect to $\mu$. We denote $\pi_1 : (z, z') \in \mathcal{Z}^2 \mapsto z$ and $\pi_2 : (z, z') \in \mathcal{Z}^2 \mapsto z'$ respectively the projections on the first and second component, which are continuous applications. For a measure $\mu$ and a measurable mapping $\mu$, we denote $g_\sharp \mu$ the pushforward measure of $\mu$ by $g$. Let $L \geq 1$ be an integer and denote $\Delta_L := \{\lambda \in \mathbb{R}_+^L \text{ s.t. } \sum_{k=1}^L \lambda_k = 1\}$, the probability simplex of $\mathbb{R}^L$.

## 9.8 Useful Lemmas

**Lemma 15** (Fubini's theorem)**.** *Let $l : \Theta \times (\mathcal{X} \times \mathcal{Y}) \to [0, \infty)$ satisfying Assumption 3. Then for all $\mu \in \mathcal{P}(\Theta)$, $\int l(\theta, \cdot) d\mu(\theta)$ is Borel measurable; for $\mathbb{Q} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$, $\int l(\cdot, (x, y)) d\mathbb{Q}(x, y)$ is Borel measurable. Moreover: $\int l(\theta, (x, y)) d\mu(\theta) d\mathbb{Q}(x, y) = \int l(\theta, (x, y)) d\mathbb{Q}(x, y) d\mu(\theta)$*

**Lemma 16.** *Let $l : \Theta \times (\mathcal{X} \times \mathcal{Y}) \to [0, \infty)$ satisfying Assumption 3. Then for all $\mu \in \mathcal{P}(\Theta)$, $(x, y) \mapsto \int l(\theta, (x, y)) d\mu(\theta)$ is upper semi-continuous and hence Borel measurable.*

*Proof.* Let $(x_n, y_n)_n$ be a sequence of $\mathcal{X} \times \mathcal{Y}$ converging to $(x, y) \in \mathcal{X} \times \mathcal{Y}$. For all $\theta \in \Theta$, $M - l(\theta, \cdot)$ is non negative and lower semi-continuous. Then by Fatou's Lemma applied:

$$\begin{aligned} \int M - l(\theta, (x, y)) d\mu(\theta) &\leq \int \liminf_{n\to\infty} M - l(\theta, (x_n, y_n)) d\mu(\theta) \\ &\leq \liminf_{n\to\infty} \int M - l(\theta, (x_n, y_n)) d\mu(\theta) \end{aligned}$$

Then we deduce that: $\int M - l(\theta, \cdot) d\mu(\theta)$ is lower semi-continuous and then $\int l(\theta, \cdot) d\mu(\theta)$ is upper-semi continuous. □

**Lemma 17.** *Let $l : \Theta \times (\mathcal{X} \times \mathcal{Y}) \to [0, \infty)$ satisfying Assumption 3 Then for all $\mu \in \mathcal{P}(\Theta)$, $\mathbb{Q} \mapsto \int l(\theta, (x, y))d\mu(\theta)d\mathbb{Q}(x, y)$ is upper semi-continuous for weak topology of measures.*

*Proof.* $-\int l(\theta, \cdot)d\mu(\theta)$ is lower semi-continuous from Lemma 16. Then $M - \int l(\theta, \cdot)d\mu(\theta)$ is lower semi-continuous and non negative. Let denote $v$ this function. Let $(v_n)_n$ be a non-decreasing sequence of continuous bounded functions such that $v_n \to v$. Let $(\mathbb{Q}_k)_k$ converging weakly towards $\mathbb{Q}$. Then by monotone convergence:

$$\int v d\mathbb{Q} = \lim_n \int v_n d\mathbb{Q} = \lim_n \lim_k \int v_n d\mathbb{Q}_k \leq \lim_k \inf \int v d\mathbb{Q}_k$$

Then $\mathbb{Q} \mapsto \int v d\mathbb{Q}$ is lower semi-continuous and then $\mathbb{Q} \mapsto \int l(\theta, (x, y))d\mu(\theta)d\mathbb{Q}(x, y)$ is upper semi-continuous for weak topology of measures. □

**Lemma 18.** *Let $l : \Theta \times (\mathcal{X} \times \mathcal{Y}) \to [0, \infty)$ satisfying Assumption 3. Then for all $\mu \in \mathcal{P}(\Theta)$, $(x, y) \mapsto \sup_{(x',y'),d(x,x')\leq\varepsilon,y=y'} \int l(\theta, (x', y'))d\mu(\theta)$ is universally measurable (i.e. measurable for all Borel probability measures). And hence the adversarial risk is well defined.*

*Proof.* Let $\phi : (x, y) \mapsto \sup_{(x',y'),d(x,x')\leq\varepsilon,y=y'} \int l(\theta, (x', y'))d\mu(\theta)$. Then for $u \in \bar{\mathbb{R}}$:

$$\{\phi(x, y) > u\} = \text{Proj}_1 \left\{ ((x, y), (x', y')) \mid \int l(\theta, (x', y'))d\mu(\theta) - c_\varepsilon((x, y), (x', y')) > u \right\}$$

By Lemma 17: $((x, y), (x', y')) \mapsto \int l(\theta, (x', y'))d\mu(\theta) - c_\varepsilon((x, y), (x', y'))$ is upper-semicontinuous hence Borel measurable. So its level sets are Borel sets, and by [246, Proposition 7.39], the projection of a Borel set is analytic. And then $\{\phi(x, y) > u\}$ universally measurable thanks to [246, Corollary 7.42.1]. We deduce that $\phi$ is universally measurable. □

## 9.9 Proofs

### 9.9.1 Proof of Proposition 9.2.1

*Proof.* Let $\eta > 0$. Let $\mathbb{Q} \in \mathcal{A}_\varepsilon(\mathbb{P})$. There exists $\gamma \in \mathcal{P}((\mathcal{X} \times \mathcal{Y})^2)$ such that, $d(x, x') \leq \varepsilon$, $y = y'$ $\gamma$-almost surely, and $\pi_{1\sharp}\gamma = \mathbb{P}$, and $\pi_{2\sharp}\gamma = \mathbb{Q}$. Then $\int c_\varepsilon d\gamma = 0 \leq \eta$. Then, we deduce that $\text{OT}_{c_\varepsilon}(\mathbb{P}, \mathbb{Q}) \leq \eta$, and $\mathbb{Q} \in \mathcal{B}_{c_\varepsilon}(\mathbb{P}, \eta)$. Reciprocally, let $\mathbb{Q} \in \mathcal{B}_{c_\varepsilon}(\mathbb{P}, \eta)$. Then, since the infimum is attained in the Wasserstein definition, there exists $\gamma \in \mathcal{P}((\mathcal{X} \times \mathcal{Y})^2)$ such that $\int c_\varepsilon d\gamma \leq \eta$. Since $c_\varepsilon((x, x'), (y, y')) = +\infty$ when $d(x, x') > \varepsilon$ and $y \neq y'$, we deduce that, $d(x, x') \leq \varepsilon$ and $y = y'$, $\gamma$-almost surely. Then $\mathbb{Q} \in \mathcal{A}_\varepsilon(\mathbb{P})$. We have then shown that: $\mathcal{A}_\varepsilon(\mathbb{P}) = \mathcal{B}_{c_\varepsilon}(\mathbb{P}, \eta)$.

The convexity of $\mathcal{A}_\varepsilon(\mathbb{P})$ is then immediate from the relation with the Wasserstein uncertainty set.

Let us show first that $\mathcal{A}_\varepsilon(\mathbb{P})$ is relatively compact for weak topology. To do so we will show that $\mathcal{A}_\varepsilon(\mathbb{P})$ is tight and apply Prokhorov's theorem. Let $\delta > 0$, $(\mathcal{X} \times \mathcal{Y}, d \oplus d')$ being a Polish space, $\{\mathbb{P}\}$ is tight then there exists $K_\delta$ compact such that $\mathbb{P}(K_\delta) \geq 1 - \delta$. Let $\tilde{K}_\delta := \{(x', y') \mid \exists (x, y) \in K_\delta,\ d(x', x) \leq \varepsilon,\ y = y'\}$. Recalling that $(\mathcal{X}, d)$ is proper (i.e. the closed balls are compact), so $\tilde{K}_\delta$ is compact. Moreover for $\mathbb{Q} \in \mathcal{A}_\varepsilon(\mathbb{P})$, $\mathbb{Q}(\tilde{K}_\delta) \geq \mathbb{P}(K_\delta) \geq 1 - \delta$. And then, Prokhorov's theorem holds, and $\mathcal{A}_\varepsilon(\mathbb{P})$ is relatively compact for weak topology.

Let us now prove that $\mathcal{A}_\varepsilon(\mathbb{P})$ is closed to conclude. Let $(\mathbb{Q}_n)_n$ be a sequence of $\mathcal{A}_\varepsilon(\mathbb{P})$ converging towards some $\mathbb{Q}$ for weak topology. For each $n$, there exists $\gamma_n \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ such that $d(x, x') \leq \varepsilon$ and $y = y'$ $\gamma_n$-almost surely and $\pi_{1\sharp}\gamma_n = \mathbb{P}$, $\pi_{2\sharp}\gamma_n = \mathbb{Q}_n$. $\{\mathbb{Q}_n, n \geq 0\}$ is relatively compact, then tight, then $\bigcup_n \Gamma_{\mathbb{P},\mathbb{Q}_n}$ is tight, then relatively compact by Prokhorov's theorem. $(\gamma_n)_n \in \bigcup_n \Gamma_{\mathbb{P},\mathbb{Q}_n}$, then up to an extraction, $\gamma_n \to \gamma$. Then $d(x, x') \leq \varepsilon$ and $y = y'$ $\gamma$-almost surely, and by continuity, $\pi_{1\sharp}\gamma = \mathbb{P}$ and by continuity, $\pi_{2\sharp}\gamma = \mathbb{Q}$. And hence $\mathcal{A}_\varepsilon(\mathbb{P})$ is closed.

Finally $\mathcal{A}_\varepsilon(\mathbb{P})$ is a convex compact set for the weak topology. □

### 9.9.2 Proof of Proposition 9.2.2

*Proof.* Let $\mu \in \mathcal{P}(\Theta)$. Let $\tilde{f} : ((x, y), (x', y')) \mapsto \mathbb{E}_{\theta\sim\mu}[l(\theta, (x, y))] - c_\varepsilon((x, y), (x', y'))$. $\tilde{f}$ is upper-semi continuous, hence upper semi-analytic. Then, by upper semi continuity of $\mathbb{E}_{\theta\sim\mu}[l(\theta, \cdot)]$ on the compact $\{(x', y') \mid\ d(x, x') \leq \varepsilon, y = y'\}$ and [246, Proposition 7.50], there exists a universally measurable mapping $T$ such that $\mathbb{E}_{\theta\sim\mu}[l(\theta, T(x, y))] = \sup_{(x',y'),\ d(x,x')\leq\varepsilon, y=y'} \mathbb{E}_{\theta\sim\mu}[l(\theta, (x, y))]$. Let $\mathbb{Q} = T_\sharp\mathbb{P}$, then $\mathbb{Q} \in \mathcal{A}_\varepsilon(\mathbb{P})$. And then

$$\mathbb{E}_{(x,y)\sim\mathbb{P}}\left[\sup_{(x',y'),\ d(x,x')\leq\varepsilon, y=y'} \mathbb{E}_{\theta\sim\mu}[l(\theta, (x', y'))]\right] \leq \sup_{\mathbb{Q}\in\mathcal{A}_\varepsilon(\mathbb{P})} \mathbb{E}_{(x,y)\sim\mathbb{Q}}\left[\mathbb{E}_{\theta\sim\mu}[l(\theta, (x, y))]\right]$$

.

Reciprocally, let $\mathbb{Q} \in \mathcal{A}_\varepsilon(\mathbb{P})$. There exists $\gamma \in \mathcal{P}((\mathcal{X} \times \mathcal{Y})^2)$, such that $d(x, x') \leq \varepsilon$ and $y = y'$ $\gamma$-almost surely, and, $\pi_{1\sharp}\gamma = \mathbb{P}$ and $\pi_{2\sharp}\gamma = \mathbb{Q}$. Then: $\mathbb{E}_{\theta\sim\mu}[l(\theta, (x', y'))] \leq \sup_{(u,v),\ d(x,u)\leq\varepsilon, y=v} \mathbb{E}_{\theta\sim\mu}[l(\theta, (u, v))]$ $\gamma$-almost surely. Then, we deduce that:

$$\begin{aligned}
\mathbb{E}_{(x',y')\sim\mathbb{Q}}\left[\mathbb{E}_{\theta\sim\mu}[l(\theta, (x', y'))]\right] &= \mathbb{E}_{(x,y,x',y')\sim\gamma}\left[\mathbb{E}_{\theta\sim\mu}[l(\theta, (x', y'))]\right] \\
&\leq \mathbb{E}_{(x,y,x',y')\sim\gamma}\left[\sup_{(u,v),\ d(x,u)\leq\varepsilon, y=v} \mathbb{E}_{\theta\sim\mu}[l(\theta, (u, v))]\right] \\
&\leq \mathbb{E}_{(x,y)\sim\mathbb{P}}\left[\sup_{(u,v),\ d(x,u)\leq\varepsilon, y=v} \mathbb{E}_{\theta\sim\mu}[l(\theta, (u, v))]\right]
\end{aligned}$$

Then we deduce the expected result:

$$\mathcal{R}^{\varepsilon}_{adv}(\mu) = \sup_{\mathbb{Q}\in\mathcal{A}_{\varepsilon}(\mathbb{P})} \mathbb{E}_{(x,y)\sim\mathbb{Q}}\left[\mathbb{E}_{\theta\sim\mu}\left[l(\theta,(x,y))\right]\right]$$

Let us show that the optimum is attained. $\mathbb{Q} \mapsto \mathbb{E}_{(x,y)\sim\mathbb{Q}}\left[\mathbb{E}_{\theta\sim\mu}\left[l(\theta,(x,y))\right]\right]$ is upper semi continuous by Lemma 17 for the weak topology of measures, and $\mathcal{A}_{\varepsilon}(\mathbb{P})$ is compact by Proposition 9.2.1, then by [246, Proposition 7.32], the supremum is attained for a certain $\mathbb{Q}^* \in \mathcal{A}_{\varepsilon}(\mathbb{P})$.

□

### 9.9.3 Proof of Theorem 9.3.1

Let us first recall the Fan's Theorem.

**Theorem 9.9.1.** *Let $U$ be a compact convex Haussdorff space and $V$ be convex space (not necessarily topological). Let $\psi : U \times V \to \mathbb{R}$ be a concave-convex function such that for all $v \in V$, $\psi(\cdot, v)$ is upper semi-continuous then:*

$$\inf_{v\in V}\max_{u\in U}\psi(u,v) = \max_{u\in U}\inf_{v\in V}\psi(u,v)$$

We are now set to prove Theorem 9.3.1.

*Proof.* $\mathcal{A}_{\varepsilon}(\mathbb{P})$, endowed with the weak topology of measures, is a Hausdorff compact convex space, thanks to Proposition 9.2.1. Moreover, $\mathcal{P}(\Theta)$ is clearly convex and $(\mathbb{Q},\mu) \mapsto \int l d\mu d\mathbb{Q}$ is bilinear, hence concave-convex. Moreover thanks to Lemma 17, for all $\mu$, $\mathbb{Q} \mapsto \int l d\mu d\mathbb{Q}$ is upper semi-continuous. Then Fan's theorem applies and strong duality holds. □

In the related work (Section 9.6), we mentioned a particular form of Theorem 9.3.1 for convex cases. As mentioned, this result has limited impact in the adversarial classification setting. It is still a direct corollary of Fan's theorem. This theorem can be stated as follows:

**Theorem 9.9.2.** *Let $\mathbb{P} \in \mathcal{P}(\mathcal{X}\times\mathcal{Y})$, $\varepsilon > 0$ and $\Theta$ a convex set. Let $l$ be a loss satisfying Assumption 3, and also, $(x,y) \in \mathcal{X}\times\mathcal{Y}$, $l(\cdot,(x,y))$ is a convex function, then we have the following:*

$$\inf_{\theta\in\Theta}\sup_{\mathbb{Q}\in\mathcal{A}_{\varepsilon}(\mathbb{P})}\mathbb{E}_{\mathbb{Q}}\left[l(\theta,(x,y))\right] = \sup_{\mathbb{Q}\in\mathcal{A}_{\varepsilon}(\mathbb{P})}\inf_{\theta\in\Theta}\mathbb{E}_{\mathbb{Q}}\left[l(\theta,(x,y))\right]$$

*The supremum is always attained. If $\Theta$ is a compact set then, the infimum is also attained.*

### 9.9.4 Proof of Proposition 9.4.1

*Proof.* Let us first show that for $\alpha \geq 0$, $\sup_{\mathbb{Q}_i \in \Gamma_{i,\varepsilon}} \mathbb{E}_{\mathbb{Q}_i,\mu}\left[l(\theta,(x,y))\right] - \alpha \mathrm{KL}\left(\mathbb{Q}_i, \frac{1}{N}\mathbb{U}_{(x_i,y_i)}\right)$ admits a solution. Let $\alpha \geq 0$, $(\mathbb{Q}^n_{\alpha,i})_{n\geq 0}$ a sequence such that

$$\mathbb{E}_{\mathbb{Q}^n_{\alpha,i},\mu}\left[l(\theta,(x,y))\right] - \alpha \mathrm{KL}\left(\mathbb{Q}^n_{\alpha,i}, \frac{1}{N}\mathbb{U}_{(x_i,y_i)}\right) \xrightarrow[n\to+\infty]{} \sup_{\mathbb{Q}_i \in \Gamma_{i,\varepsilon}} \mathbb{E}_{\mathbb{Q}_i,\mu}\left[l(\theta,(x,y))\right] - \alpha \mathrm{KL}\left(\mathbb{Q}_i, \frac{1}{N}\mathbb{U}_{(x_i,y_i)}\right).$$

As $\Gamma_{i,\varepsilon}$ is tight ($(\mathcal{X},d)$ is a proper metric space therefore all the closed ball are compact) and by Prokhorov's theorem, we can extract a subsequence which converges toward $\mathbb{Q}^*_{\alpha,i}$. Moreover, $l$ is upper semi-continuous (u.s.c), thus $\mathbb{Q} \to \mathbb{E}_{\mathbb{Q},\mu}\left[l(\theta,(x,y))\right]$ is also u.s.c.[8] Moreover $\mathbb{Q} \to -\alpha \mathrm{KL}\left(\mathbb{Q}, \frac{1}{N}\mathbb{U}_{(x_i,y_i)}\right)$ is also u.s.c. [9], therefore, by considering the limit superior as $n$ goes to infinity we obtain that

$$\begin{aligned}
&\limsup_{n\to+\infty} \mathbb{E}_{\mathbb{Q}^n_{\alpha,i},\mu}\left[l(\theta,(x,y))\right] - \alpha \mathrm{KL}\left(\mathbb{Q}^n_{\alpha,i}, \frac{1}{N}\mathbb{U}_{(x_i,y_i)}\right) \\
&= \sup_{\mathbb{Q}_i \in \Gamma_{i,\varepsilon}} \mathbb{E}_{\mathbb{Q}_i,\mu}\left[l(\theta,(x,y))\right] - \alpha \mathrm{KL}\left(\mathbb{Q}_i, \frac{1}{N}\mathbb{U}_{(x_i,y_i)}\right) \\
&\leq \mathbb{E}_{\mathbb{Q}^*_{\alpha,i},\mu}\left[l(\theta,(x,y))\right] - \alpha \mathrm{KL}\left(\mathbb{Q}^*_{\alpha,i}, \frac{1}{N}\mathbb{U}_{(x_i,y_i)}\right)
\end{aligned}$$

from which we deduce that $\mathbb{Q}^*_{\alpha,i}$ is optimal.

Let us now show the result. We consider a positive sequence of $(\alpha_i^{(\ell)})_{\ell\geq 0}$ such that $\alpha_i^{(\ell)} \to 0$. Let us denote $\mathbb{Q}^*_{\alpha_i^{(\ell)},i}$ and $\mathbb{Q}^*_i$ the solutions of $\max_{\mathbb{Q}_i \in \Gamma_{i,\varepsilon}} \mathbb{E}_{\mathbb{Q}_i,\mu}\left[l(\theta,(x,y))\right] - \alpha_i^{(\ell)} \mathrm{KL}\left(\mathbb{Q}_i, \frac{1}{N}\mathbb{U}_{(x_i,y_i)}\right)$ and $\max_{\mathbb{Q}_i \in \Gamma_{i,\varepsilon}} \mathbb{E}_{\mathbb{Q}_i,\mu}\left[l(\theta,(x,y))\right]$ respectively. Since $\Gamma_{i,\varepsilon}$ is tight, $(\mathbb{Q}^*_{\alpha_i^{(\ell)},i})_{\ell\geq 0}$ is also tight and we can extract by Prokhorov's theorem a subsequence which converges towards $\mathbb{Q}^*$. Moreover we have

$$\mathbb{E}_{\mathbb{Q}^*_i,\mu}\left[l(\theta,(x,y))\right] - \alpha_i^{(\ell)} \mathrm{KL}\left(\mathbb{Q}^*_i, \frac{1}{N}\mathbb{U}_{(x_i,y_i)}\right) \leq \mathbb{E}_{\mathbb{Q}^*_{\alpha_i^{(\ell)},i},\mu}\left[l(\theta,(x,y))\right] - \alpha_i^{(\ell)} \mathrm{KL}\left(\mathbb{Q}^*_{\alpha_i^{(\ell)},i}, \frac{1}{N}\mathbb{U}_{(x_i,y_i)}\right)$$

[8] Indeed by considering a decreasing sequence of continuous and bounded functions which converge towards $\mathbb{E}_{\mu}\left[l(\theta,(x,y))\right]$ and by definition of the weak convergence the result follows.

[9] for $\alpha = 0$ the result is clear, and if $\alpha > 0$, note that $\mathrm{KL}\left(\cdot, \frac{1}{N}\mathbb{U}_{(x_i,y_i)}\right)$ is lower semi-continuous

from which follows that

$$0 \leq \mathbb{E}_{\mathbb{Q}_i^*,\mu}\left[l(\theta,(x,y))\right] - \mathbb{E}_{\mathbb{Q}^*_{\alpha_i^{(\ell)},i},\mu}\left[l(\theta,(x,y))\right] \leq \alpha_i^{(\ell)}\mathrm{KL}\left(\mathbb{Q}_i^*, \frac{1}{N}\mathbb{U}_{(x_i,y_i)}\right) - \alpha_i^{(\ell)}\mathrm{KL}\left(\mathbb{Q}^*_{\alpha_i^{(\ell)},i}, \frac{1}{N}\mathbb{U}_{(x_i,y_i)}\right)$$

Then by considering the limit superior we obtain that

$$\limsup_{\ell\to+\infty} \mathbb{E}_{\mathbb{Q}^*_{\alpha_i^{(\ell)},i},\mu}\left[l(\theta,(x,y))\right] = \mathbb{E}_{\mathbb{Q}_i^*,\mu}\left[l(\theta,(x,y))\right].$$

from which follows that

$$\mathbb{E}_{\mathbb{Q}_i^*,\mu}\left[l(\theta,(x,y))\right] \leq \mathbb{E}_{\mathbb{Q}^*,\mu}\left[l(\theta,(x,y))\right]$$

and by optimality of $\mathbb{Q}_i^*$ we obtain the desired result. □

### 9.9.5 Proof of Proposition 9.4.2

*Proof.* Let us denote for all $\mu \in \mathcal{P}(\Theta)$,

$$\widehat{\mathcal{R}}^{\boldsymbol{\varepsilon},\mathbf{m}}_{adv,\boldsymbol{\alpha}}(\mu) := \sum_{i=1}^N \frac{\alpha_i}{N}\log\left(\frac{1}{m_i}\sum_{j=1}^{m_i}\exp\frac{\mathbb{E}_\mu\left[l(\theta,u_j^{(i)})\right]}{\alpha_i}\right).$$

Let also consider $(\mu_n^{(\mathbf{m})})_{n\geq 0}$ and $(\mu_n)_{n\geq 0}$ two sequences such that

$$\widehat{\mathcal{R}}^{\boldsymbol{\varepsilon},\mathbf{m}}_{adv,\boldsymbol{\alpha}}(\mu_n^{(\mathbf{m})}) \xrightarrow[n\to+\infty]{} \widehat{\mathcal{R}}^{\boldsymbol{\varepsilon},\mathbf{m}}_{adv,\boldsymbol{\alpha}}, \qquad \widehat{\mathcal{R}}^{\boldsymbol{\varepsilon}}_{adv,\boldsymbol{\alpha}}(\mu_n) \xrightarrow[n\to+\infty]{} \widehat{\mathcal{R}}^{\boldsymbol{\varepsilon},*}_{adv,\boldsymbol{\alpha}}.$$

We first remarks that

$$\begin{aligned}\widehat{\mathcal{R}}^{\boldsymbol{\varepsilon},\mathbf{m}}_{adv,\boldsymbol{\alpha}} - \widehat{\mathcal{R}}^{\boldsymbol{\varepsilon},*}_{adv,\boldsymbol{\alpha}} &\leq \widehat{\mathcal{R}}^{\boldsymbol{\varepsilon},\mathbf{m}}_{adv,\boldsymbol{\alpha}} - \widehat{\mathcal{R}}^{\boldsymbol{\varepsilon},\mathbf{m}}_{adv,\boldsymbol{\alpha}}(\mu_n) + \widehat{\mathcal{R}}^{\boldsymbol{\varepsilon},\mathbf{m}}_{adv,\boldsymbol{\alpha}}(\mu_n) - \widehat{\mathcal{R}}^{\boldsymbol{\varepsilon}}_{adv,\boldsymbol{\alpha}}(\mu_n) + \widehat{\mathcal{R}}^{\boldsymbol{\varepsilon}}_{adv,\boldsymbol{\alpha}}(\mu_n) - \widehat{\mathcal{R}}^{\boldsymbol{\varepsilon},*}_{adv,\boldsymbol{\alpha}} \\ &\leq \sup_{\mu\in\mathcal{P}(\Theta)}\left|\widehat{\mathcal{R}}^{\boldsymbol{\varepsilon},\mathbf{m}}_{adv,\boldsymbol{\alpha}}(\mu) - \widehat{\mathcal{R}}^{\boldsymbol{\varepsilon}}_{adv,\boldsymbol{\alpha}}(\mu)\right| + \widehat{\mathcal{R}}^{\boldsymbol{\varepsilon}}_{adv,\boldsymbol{\alpha}}(\mu_n) - \widehat{\mathcal{R}}^{\boldsymbol{\varepsilon},*}_{adv,\boldsymbol{\alpha}},\end{aligned}$$

and by considering the limit, we obtain that

$$\widehat{\mathcal{R}}^{\boldsymbol{\varepsilon},\mathbf{m}}_{adv,\boldsymbol{\alpha}} - \widehat{\mathcal{R}}^{\boldsymbol{\varepsilon},*}_{adv,\boldsymbol{\alpha}} \leq \sup_{\mu\in\mathcal{P}(\Theta)}\left|\widehat{\mathcal{R}}^{\boldsymbol{\varepsilon},\mathbf{m}}_{adv,\boldsymbol{\alpha}}(\mu) - \widehat{\mathcal{R}}^{\boldsymbol{\varepsilon}}_{adv,\boldsymbol{\alpha}}(\mu)\right|$$

Simarly we have that

$$\begin{aligned}\widehat{\mathcal{R}}^{\boldsymbol{\varepsilon},*}_{adv,\boldsymbol{\alpha}} - \widehat{\mathcal{R}}^{\boldsymbol{\varepsilon},\mathbf{m}}_{adv,\boldsymbol{\alpha}} &\leq \widehat{\mathcal{R}}^{\boldsymbol{\varepsilon},*}_{adv,\boldsymbol{\alpha}} - \widehat{\mathcal{R}}^{\boldsymbol{\varepsilon}}_{adv,\boldsymbol{\alpha}}(\mu_n^{(\boldsymbol{m})}) + \widehat{\mathcal{R}}^{\boldsymbol{\varepsilon}}_{adv,\boldsymbol{\alpha}}(\mu_n^{(\boldsymbol{m})}) - \widehat{\mathcal{R}}^{\boldsymbol{\varepsilon},\mathbf{m}}_{adv,\boldsymbol{\alpha}}(\mu_n^{(\boldsymbol{m})}) \\ &\quad + \widehat{\mathcal{R}}^{\boldsymbol{\varepsilon},\mathbf{m}}_{adv,\boldsymbol{\alpha}}(\mu_n^{(\boldsymbol{m})}) - \widehat{\mathcal{R}}^{\boldsymbol{\varepsilon},\mathbf{m}}_{adv,\boldsymbol{\alpha}}\end{aligned}$$

from which follows that

$$\widehat{\mathcal{R}}^{\boldsymbol{\varepsilon},*}_{adv,\boldsymbol{\alpha}} - \widehat{\mathcal{R}}^{\boldsymbol{\varepsilon},\mathbf{m}}_{adv,\boldsymbol{\alpha}} \leq \sup_{\mu\in\mathcal{P}(\Theta)} \left|\widehat{\mathcal{R}}^{\boldsymbol{\varepsilon},\mathbf{m}}_{adv,\boldsymbol{\alpha}}(\mu) - \widehat{\mathcal{R}}^{\boldsymbol{\varepsilon}}_{adv,\boldsymbol{\alpha}}(\mu)\right|$$

Therefore we obtain that

$$\left|\widehat{\mathcal{R}}^{\boldsymbol{\varepsilon},*}_{adv,\boldsymbol{\alpha}} - \widehat{\mathcal{R}}^{\boldsymbol{\varepsilon},\mathbf{m}}_{adv,\boldsymbol{\alpha}}\right| \leq \sum_{i=1}^{N} \frac{\alpha}{N} \sup_{\mu\in\mathcal{P}(\Theta)} \Big| \log\left(\frac{1}{m_i}\sum_{j=1}^{m_i} \exp\left(\frac{\mathbb{E}_{\theta\sim\mu}\left[l(\theta,u_j^{(i)}))\right]}{\alpha}\right)\right) - \log\left(\int_{\mathcal{X}\times\mathcal{Y}} \exp\left(\frac{\mathbb{E}_{\theta\sim\mu}\left[l(\theta,(x,y))\right]}{\alpha}\right) d\mathbb{U}_{(x_i,y_i)}\right)\Big|.$$

Observe that $l \geq 0$, therefore because the log function is 1-Lipschitz on $[1,+\infty)$, we obtain that

$$\left|\widehat{\mathcal{R}}^{\boldsymbol{\varepsilon},*}_{adv,\boldsymbol{\alpha}} - \widehat{\mathcal{R}}^{\boldsymbol{\varepsilon},\mathbf{m}}_{adv,\boldsymbol{\alpha}}\right| \leq$$

$$\sum_{i=1}^{N} \frac{\alpha}{N} \sup_{\mu\in\mathcal{P}(\Theta)} \left|\frac{1}{m_i}\sum_{j=1}^{m_i} \exp\left(\frac{\mathbb{E}_{\theta\sim\mu}\left[l(\theta,u_j^{(i)}))\right]}{\alpha}\right) - \int_{\mathcal{X}\times\mathcal{Y}} \exp\left(\frac{\mathbb{E}_{\theta\sim\mu}\left[l(\theta,(x,y))\right]}{\alpha}\right) d\mathbb{U}_{(x_i,y_i)}\right|.$$

Let us now denote for all $i = 1,\ldots,N$,

$$\widehat{R}_i(\mu,\boldsymbol{u}^{(i)}) := \sum_{j=1}^{m_i} \exp\left(\frac{\mathbb{E}_{\theta\sim\mu}\left[l(\theta,u_j^{(i)}))\right]}{\alpha}\right)$$

$$R_i(\mu) := \int_{\mathcal{X}\times\mathcal{Y}} \exp\left(\frac{\mathbb{E}_{\theta\sim\mu}\left[l(\theta,(x,y))\right]}{\alpha}\right) d\mathbb{U}_{(x_i,y_i)}.$$

and let us define

$$f(\boldsymbol{u}^{(1)},\ldots,\boldsymbol{u}^{(N)}) := \sum_{i=1}^{N} \frac{\alpha}{N} \sup_{\mu\in\mathcal{P}(\Theta)} \left|\widehat{R}_i(\mu) - R_i(\mu)\right|$$

where $\boldsymbol{u}^{(i)} := (u_1^{(i)},\ldots,u_1^{(m)})$. By denoting $z^{(i)} = (u_1^{(i)},\ldots,u_{k-1}^{(i)},z,u_{k+1}^{(i)},\ldots,u_m^{(i)})$, we have that

$$\begin{aligned}
|f(\boldsymbol{u}^{(1)},\ldots,\boldsymbol{u}^{(N)}) - f(\boldsymbol{u}^{(1)},\ldots,\boldsymbol{u}^{(i-1)},\boldsymbol{z}^{(i)},\boldsymbol{u}^{(i+1)},\ldots,\boldsymbol{u}^{(N)})| & \\
&\leq \frac{\alpha}{N}\Big|\sup_{\mu\in\mathcal{P}(\Theta)}\left|\widehat{R}_i(\mu,\boldsymbol{u}^{(i)}) - R_i(\mu)\right| - \sup_{\mu\in\mathcal{P}(\Theta)}\left|\widehat{R}_i(\mu,\boldsymbol{z}^{(i)}) - R_i(\mu)\right|\Big| \\
&\leq \frac{\alpha}{N}\Big|\frac{1}{m}\left[\exp\left(\frac{\mathbb{E}_{\theta\sim\mu}\left[l(\theta,u_k^{(i)}))\right]}{\alpha}\right) - \exp\left(\frac{\mathbb{E}_{\theta\sim\mu}\left[l(\theta,z^{(i)}))\right]}{\alpha}\right)\right]\Big| \\
&\leq \frac{2\exp(M/\alpha)}{Nm}
\end{aligned}$$

where the last inequality comes from the fact that the loss is upper bounded by $l \leq M$. Then by appling the McDiarmid's Inequality, we obtain that with a probability of at least $1-\delta$,

$$\left|\widehat{\mathcal{R}}^{\boldsymbol{\varepsilon},*}_{adv,\boldsymbol{\alpha}} - \widehat{\mathcal{R}}^{\boldsymbol{\varepsilon},\mathbf{m}}_{adv,\boldsymbol{\alpha}}\right| \leq \mathbb{E}(f(\boldsymbol{u}^{(1)},\ldots,\boldsymbol{u}^{(N)})) + \frac{2\exp(M/\alpha)}{\sqrt{mN}}\sqrt{\frac{\log(2/\delta)}{2}}.$$

Thanks to [254, Lemma 26.2], we have for all $i \in \{1,\ldots,N\}$

$$\mathbb{E}(f(\boldsymbol{u}^{(1)},\ldots,\boldsymbol{u}^{(N)})) \leq 2\mathbb{E}(\mathrm{Rad}(\mathcal{F}_i \circ \mathbf{u}^{(\mathbf{i})}))$$

where for any class of function $\mathcal{F}$ defined on $\mathcal{Z}$ and point $\boldsymbol{z} : (z_1,\ldots,z_q) \in \mathcal{Z}^q$

$$\mathcal{F} \circ \boldsymbol{z} := \Big\{(f(z_1),\ldots,f(z_q)),\ f \in \mathcal{F}\Big\} \quad , \quad \mathrm{Rad}(\mathcal{F} \circ \boldsymbol{z}) := \frac{1}{q}\mathbb{E}_{\boldsymbol{\sigma}\sim\{\pm 1\}}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{q}\sigma_i f(z_i)\right]$$

$$\mathcal{F}_i := \Big\{u \to \exp\left(\frac{\mathbb{E}_{\theta\sim\mu}\left[l(\theta,u))\right]}{\alpha}\right),\ \mu \in \mathcal{P}(\Theta)\Big\}.$$

Moreover as $x \to \exp(x/\alpha)$ is $\frac{\exp(M/\alpha)}{\alpha}$-Lipstchitz on $(-\infty, M]$, by [254, Lemma 26.9], we have

$$\mathrm{Rad}(\mathcal{F}_i \circ \mathbf{u}^{(\mathbf{i})}) \leq \frac{\exp(M/\alpha)}{\alpha}\mathrm{Rad}(\mathcal{H}_i \circ \mathbf{u}^{(\mathbf{i})})$$

where

$$\mathcal{H}_i := \Big\{u \to \mathbb{E}_{\theta\sim\mu}\left[l(\theta,u))\right],\ \mu \in \mathcal{P}(\Theta)\Big\}.$$

Let us now define

$$g(\boldsymbol{u}^{(1)},\ldots,\boldsymbol{u}^{(N)}) := \sum_{j=1}^{N}\frac{2\exp(M/\alpha)}{N}\mathrm{Rad}(\mathcal{H}_j \circ \mathbf{u}^{(\mathbf{j})}).$$

We observe that

$$\begin{aligned}
&|g(\boldsymbol{u}^{(1)},\ldots,\boldsymbol{u}^{(N)}) - g(\boldsymbol{u}^{(1)},\ldots,\boldsymbol{u}^{(i-1)},\boldsymbol{z}^{(i)},\boldsymbol{u}^{(i+1)},\ldots,\boldsymbol{u}^{(N)})| \\
&\leq \frac{2\exp(M/\alpha)}{N}|\mathrm{Rad}(\mathcal{H}_i \circ \mathbf{u}^{(\mathbf{i})}) - \mathrm{Rad}(\mathcal{H}_i \circ \mathbf{z}^{(\mathbf{i})})| \\
&\leq \frac{2\exp(M/\alpha)}{N}\frac{2M}{m}.
\end{aligned}$$

By Applying the McDiarmid's Inequality, we have that with a probability of at least $1-\delta$

$$\mathbb{E}(g(\boldsymbol{u}^{(1)},\ldots,\boldsymbol{u}^{(N)})) \leq g(\boldsymbol{u}^{(1)},\ldots,\boldsymbol{u}^{(N)}) + \frac{4\exp(M/\alpha)M}{\sqrt{mN}}\sqrt{\frac{\log(2/\delta)}{2}}.$$

Remarks also that

$$\begin{aligned}\mathrm{Rad}(\mathcal{H}_i \circ \mathbf{u}^{(\mathbf{i})}) &= \frac{1}{m}\mathbb{E}_{\boldsymbol{\sigma}\sim\{\pm 1\}}\left[\sup_{\mu\in\mathcal{P}(\Theta)}\sum_{j=1}^{m}\sigma_i\mathbb{E}_{\mu}(l(\theta,u_j^{(i)}))\right]\\ &= \frac{1}{m}\mathbb{E}_{\boldsymbol{\sigma}\sim\{\pm 1\}}\left[\sup_{\theta\in\Theta}\sum_{j=1}^{m}\sigma_i l(\theta,u_j^{(i)})\right]\end{aligned}$$

Finally, applying a union bound leads to the desired result.

□

### 9.9.6 Proof of Proposition 9.4.3

*Proof.* Following the same steps than the proof of Proposition 9.4.2, let $(\mu_n^{\varepsilon})_{n\geq 0}$ and $(\mu_n)_{n\geq 0}$ two sequences such that

$$\widehat{\mathcal{R}}^{\varepsilon}_{adv,\boldsymbol{\alpha}}(\mu_n^{\varepsilon}) \xrightarrow[n\to+\infty]{} \widehat{\mathcal{R}}^{\varepsilon,*}_{adv,\boldsymbol{\alpha}}, \qquad \widehat{\mathcal{R}}^{\varepsilon}_{adv}(\mu_n) \xrightarrow[n\to+\infty]{} \widehat{\mathcal{R}}^{\varepsilon,*}_{adv}.$$

Remarks that

$$\begin{aligned}\widehat{\mathcal{R}}^{\varepsilon,*}_{adv,\boldsymbol{\alpha}} - \widehat{\mathcal{R}}^{\varepsilon,*}_{adv} &\leq \widehat{\mathcal{R}}^{\varepsilon,*}_{adv,\boldsymbol{\alpha}} - \widehat{\mathcal{R}}^{\varepsilon}_{adv,\boldsymbol{\alpha}}(\mu_n) + \widehat{\mathcal{R}}^{\varepsilon}_{adv,\boldsymbol{\alpha}}(\mu_n) - \widehat{\mathcal{R}}^{\varepsilon}_{adv}(\mu_n) + \widehat{\mathcal{R}}^{\varepsilon}_{adv}(\mu_n) - \widehat{\mathcal{R}}^{\varepsilon,*}_{adv}\\ &\leq \sup_{\mu\in\mathcal{P}(\Theta)}\left|\widehat{\mathcal{R}}^{\varepsilon}_{adv,\boldsymbol{\alpha}}(\mu) - \widehat{\mathcal{R}}^{\varepsilon}_{adv}(\mu)\right| + \widehat{\mathcal{R}}^{\varepsilon}_{adv}(\mu_n) - \widehat{\mathcal{R}}^{\varepsilon,*}_{adv}\end{aligned}$$

Then by considering the limit we obtain that

$$\widehat{\mathcal{R}}^{\varepsilon,*}_{adv,\boldsymbol{\alpha}} - \widehat{\mathcal{R}}^{\varepsilon,*}_{adv} \leq \sup_{\mu\in\mathcal{P}(\Theta)}\left|\widehat{\mathcal{R}}^{\varepsilon}_{adv,\boldsymbol{\alpha}}(\mu) - \widehat{\mathcal{R}}^{\varepsilon}_{adv}(\mu)\right|.$$

Similarly, we obtain that

$$\widehat{\mathcal{R}}^{\varepsilon,*}_{adv} - \widehat{\mathcal{R}}^{\varepsilon,*}_{adv,\boldsymbol{\alpha}} \leq \sup_{\mu\in\mathcal{P}(\Theta)}\left|\widehat{\mathcal{R}}^{\varepsilon}_{adv,\boldsymbol{\alpha}}(\mu) - \widehat{\mathcal{R}}^{\varepsilon}_{adv}(\mu)\right|,$$

from which follows that

$$\left|\widehat{\mathcal{R}}^{\varepsilon,*}_{adv,\boldsymbol{\alpha}} - \widehat{\mathcal{R}}^{\varepsilon,*}_{adv}\right| \leq \frac{1}{N}\sum_{i=1}^{N}\sup_{\mu\in\mathcal{P}(\Theta)}\left|\alpha\log\left(\int_{\mathcal{X}\times\mathcal{Y}}\exp\left(\frac{\mathbb{E}_{\mu}[l(\theta,(x,y))]}{\alpha}\right)d\mathbb{U}_{(x_i,y_i)}\right) - \sup_{u\in S^{\varepsilon}_{(x_i,y_i)}}\mathbb{E}_{\mu}[l(\theta,u)]\right|.$$

Let $\mu \in \mathcal{P}(\Theta)$ and $i \in \{1, \dots, N\}$, then we have

$$
\begin{aligned}
&\left|\alpha \log\left(\int_{\mathcal{X}\times\mathcal{Y}} \exp\left(\frac{\mathbb{E}_\mu[l(\theta,(x,y))]}{\alpha}\right) d\mathbb{U}_{(x_i,y_i)}\right) - \sup_{u\in S^\varepsilon_{(x_i,y_i)}} \mathbb{E}_\mu[l(\theta,u)]\right| \\
&\quad = \left|\alpha \log\left(\int_{\mathcal{X}\times\mathcal{Y}} \exp\left(\frac{\mathbb{E}_\mu[l(\theta,(x,y))] - \sup_{u\in S^\varepsilon_{(x_i,y_i)}} \mathbb{E}_\mu[l(\theta,u)]}{\alpha}\right) d\mathbb{U}_{(x_i,y_i)}\right)\right| \\
&\quad = \alpha\Bigg| \log\left(\int_{A^{(x_i,y_i)}_{\beta,\mu}} \exp\left(\frac{\mathbb{E}_\mu[l(\theta,(x,y))] - \sup_{u\in S^\varepsilon_{(x_i,y_i)}} \mathbb{E}_\mu[l(\theta,u)]}{\alpha}\right) d\mathbb{U}_{(x_i,y_i)}\right. \\
&\quad + \left.\int_{(A^{(x_i,y_i)}_{\beta,\mu})^c} \exp\left(\frac{\mathbb{E}_\mu[l(\theta,(x,y))] - \sup_{u\in S^\varepsilon_{(x_i,y_i)}} \mathbb{E}_\mu[l(\theta,u)]}{\alpha}\right) d\mathbb{U}_{(x_i,y_i)}\right)\Bigg| \\
&\quad \leq \alpha\left| \log\left(\exp(-\beta/\alpha)\mathbb{U}_{(x_i,y_i)}\left(A^{(x_i,y_i)}_{\beta,\mu}\right)\right)\right| \\
&\quad + \alpha\left| \log\left(1 + \frac{\exp(\beta/\alpha)}{\mathbb{U}_{(x_i,y_i)}\left(A^{(x_i,y_i)}_{\beta,\mu}\right)} \int_{(A^{(x_i,y_i)}_{\beta,\mu})^c} E d\mathbb{U}_{(x_i,y_i)}\right)\right| \\
&\quad \leq \alpha \log(1/C_\beta) + \beta + \frac{\alpha}{C_\beta} \\
&\quad \leq 2\alpha \log(1/C_\beta) + \beta
\end{aligned}
$$

where

$$
E = \exp\left(\frac{\mathbb{E}_\mu[l(\theta,(x,y))] - \sup_{u\in S^\varepsilon_{(x_i,y_i)}} \mathbb{E}_\mu[l(\theta,u)]}{\alpha}\right),
$$

finally we obtain that

$$
\begin{aligned}
&\left|\alpha \log\left(\int_{\mathcal{X}\times\mathcal{Y}} \exp\left(\frac{\mathbb{E}_\mu[l(\theta,(x,y))]}{\alpha}\right) d\mathbb{U}_{(x_i,y_i)}\right) - \sup_{u\in S^\varepsilon_{(x_i,y_i)}} \mathbb{E}_\mu[l(\theta,u)]\right| \\
&\quad \leq \alpha \log(1/C_\beta) + \beta + \frac{\alpha}{C_\beta} \\
&\quad \leq 2\alpha \log(1/C_\beta) + \beta
\end{aligned}
$$

□

### 9.9.7 Proof of Proposition 9.4.4

*Proof.* Thanks to Danskin theorem, if $\mathbb{Q}^*$ is a best response to $\boldsymbol{\lambda}$, then

$$
\boldsymbol{g}^* := \left(\mathbb{E}_{\mathbb{Q}^*}\left[l(\theta_1,(x,y))\right], \dots, \mathbb{E}_{\mathbb{Q}^*}\left[l(\theta_L,(x,y))\right]\right)^T
$$

is a subgradient of $\boldsymbol{\lambda} \to \mathcal{R}^{\varepsilon}_{adv}(\boldsymbol{\lambda})$. Let $\eta \geq 0$ be the learning rate. Then we have for all $t \geq 1$:

$$\begin{aligned}
\|\boldsymbol{\lambda}_t - \boldsymbol{\lambda}^*\|^2 &\leq \|\boldsymbol{\lambda}_{t-1} - \eta \boldsymbol{g}_t - \boldsymbol{\lambda}^*\|^2 \\
&= \|\boldsymbol{\lambda}_{t-1} - \boldsymbol{\lambda}^*\|^2 - 2\eta \langle \boldsymbol{g}_t, \boldsymbol{\lambda}_{t-1} - \boldsymbol{\lambda}^* \rangle + \eta^2 \|\boldsymbol{g}_t\|^2 \\
&\leq \|\boldsymbol{\lambda}_{t-1} - \boldsymbol{\lambda}^*\|^2 - 2\eta \langle \boldsymbol{g}^*_t, \boldsymbol{\lambda}_{t-1} - \boldsymbol{\lambda}^* \rangle + 2\eta \langle \boldsymbol{g}^*_t - \boldsymbol{g}_t, \boldsymbol{\lambda}_{t-1} - \boldsymbol{\lambda}^* \rangle + \eta^2 M^2 L \\
&\leq \|\boldsymbol{\lambda}_{t-1} - \boldsymbol{\lambda}^*\|^2 - 2\eta \left( \mathcal{R}^{\varepsilon}_{adv}(\boldsymbol{\lambda}_t) - \mathcal{R}^{\varepsilon}_{adv}(\boldsymbol{\lambda}^*) \right) + 4\eta\delta + \eta^2 M^2 L
\end{aligned}$$

We then deduce by summing:

$$2\eta \sum_{t=1}^{T} \mathcal{R}^{\varepsilon}_{adv}(\boldsymbol{\lambda}_t) - \mathcal{R}^{\varepsilon}_{adv}(\boldsymbol{\lambda}^*) \leq 4\delta\eta T + \|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|^2 + \eta^2 M^2 L T$$

Then we have:

$$\min_{t \in [T]} \mathcal{R}^{\varepsilon}_{adv}(\boldsymbol{\lambda}_t) - \mathcal{R}^{\varepsilon}_{adv}(\boldsymbol{\lambda}^*) \leq 2\delta + \frac{4}{\eta T} + M^2 L \eta$$

The left-hand term is minimal for $\eta = \frac{2}{M\sqrt{LT}}$, and for this value:

$$\min_{t \in [T]} \mathcal{R}^{\varepsilon}_{adv}(\boldsymbol{\lambda}_t) - \mathcal{R}^{\varepsilon}_{adv}(\boldsymbol{\lambda}^*) \leq 2\delta + \frac{2M\sqrt{L}}{\sqrt{T}}$$

□

.

## 9.10 Additional Experimental Results

### 9.10.1 Experimental setting.

**Optimizer.** For each of our models, The optimizer we used in all our implementations is SGD with learning rate set to 0.4 at epoch 0 and is divided by 10 at half training then by 10 at the three quarters of training. The momentum is set to 0.9 and the weight decay to $5 \times 10^{-4}$. The batch size is set to 1024.

**Adaptation of Attacks.** Since our classifier is randomized, we need to adapt the attack accordingly. To do so we used the expected loss:

$$\tilde{l}\left((\boldsymbol{\lambda}, \boldsymbol{\theta}), (x, y)\right) = \sum_{k=1}^{L} \lambda_k l(\theta_k, (x, y))$$

to compute the gradient in the attacks, regardless the loss (DLR or cross-entropy). For the inner maximization at training time, we used a PGD attack on the cross-entropy loss with $\varepsilon = 0.03$. For the final evaluation, we used the untargeted $DLR$ attack with default parameters.

**Regularization in Practice.** The entropic regularization in higher dimensional setting need to be adapted to be more likely to find adversaries. To do so, we computed PGD attacks with only 3 iterations with 5 different restarts instead of sampling uniformly 5 points in the $\ell_\infty$-ball. In our experiments in the main paper, we use a regularization parameter $\alpha = 0.001$. The learning rate for the minimization on $\boldsymbol{\lambda}$ is always fixed to 0.001.

**Alternate Minimization Parameters.** Algorithm 18 implies an alternate minimization algorithm. We set the number of updates of $\boldsymbol{\theta}$ to $T_\theta = 50$ and, the update of $\boldsymbol{\lambda}$ to $T_\lambda = 25$.

### 9.10.2 Effect of the Regularization

In this subsection, we experimentally investigate the effect of the regularization. In Figure 9.4, we notice, that the regularization has the effect of stabilizing, reducing the variance and improving the level of the robust accuracy for adversarial training for mixtures (Algorithm 18). The standard accuracy curves are very similar in both cases.

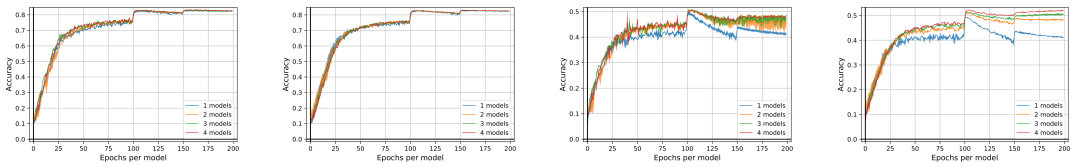



Figure 9.4: On left and middle-left: Standard accuracies over epochs with respectively no regularization and regularization set to $\alpha = 0.001$. On middle right and right: Robust accuracies for the same parameters against PGD attack with 20 iterations and $\varepsilon = 0.03$.

### 9.10.3 Additional Experiments on WideResNet28x10

We now evaluate our algorithm on WideResNet28x10 [255] architecture. Due to computation costs, we limit ourselves to 1 and 2 models, with regularization parameter set to 0.001 as in the paper experiments section. Results are reported in Figure 9.5. We remark this architecture can lead to more robust models, corroborating the results from [252].

| Models | Acc. | $\text{APGD}_{\text{CE}}$ | $\text{APGD}_{\text{DLR}}$ | Rob. Acc |
|---|---|---|---|---|
| 1 | 85.2% | 49.9% | 50.2% | 48.5% |
| 2 | **86.0%** | **51.5%** | **52.1%** | **49.6%** |

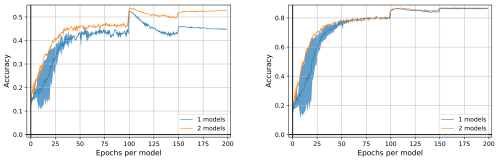



Figure 9.5: On left: Comparison of our algorithm with a standard adversarial training (one model) on WideResNet28x10. We reported the results for the model with the best robust accuracy obtained over two independent runs because adversarial training might be unstable. Standard and Robust accuracy (respectively in the middle and on right) on CIFAR-10 test images in function of the number of epochs per classifier with 1 and 2 WideResNet28x10 models. The performed attack is PGD with 20 iterations and $\varepsilon = 8/255$.

### 9.10.4 Overfitting in Adversarial Robustness

We further investigate the overfitting of our heuristic algorithm. We plotted in Figure 9.6 the robust accuracy on ResNet18 with 1 to 5 models. The most robust mixture of 5 models against PGD with 20 iterations arrives at epoch 198, *i.e.* at the end of the training, contrary to 1 to 4 models, where the most robust mixture occurs around epoch 101. However, the accuracy against AGPD with 100 iterations in lower than the one at epoch 101 with global robust accuracy of 47.6% at epoch 101 and 45.3% at epoch 198. This strange phenomenon would suggest that the more powerful the attacks are, the more the models are subject to overfitting. We leave this question to further works.

## 9.11 Additional Results

### 9.11.1 Equality of Standard Randomized and Deterministic Minimal Risks

**Proposition 9.11.1.** *Let $\mathbb{P}$ be a Borel probability distribution on $\mathcal{X} \times \mathcal{Y}$, and $l$ a loss satisfying Assumption 3, then:*

$$\inf_{\mu \in \mathcal{P}(\Theta)} \mathcal{R}(\mu) = \inf_{\theta \in \Theta} \mathcal{R}(\theta)$$

*Proof.* It is clear that: $\inf_{\mu \in \mathcal{P}(\Theta)} \mathcal{R}(\mu) \leq \inf_{\theta \in \Theta} \mathcal{R}(\theta)$. Now, let $\mu \in \mathcal{P}(\Theta)$, then:

$$\begin{aligned} \mathcal{R}(\mu) = \mathbb{E}_{\theta \sim \mu}(\mathcal{R}(\theta)) &\geq \operatorname*{essinf}_{\mu} \mathbb{E}_{\theta \sim \mu}(\mathcal{R}(\theta)) \\ &\geq \inf_{\theta \in \Theta} \mathcal{R}(\theta). \end{aligned}$$

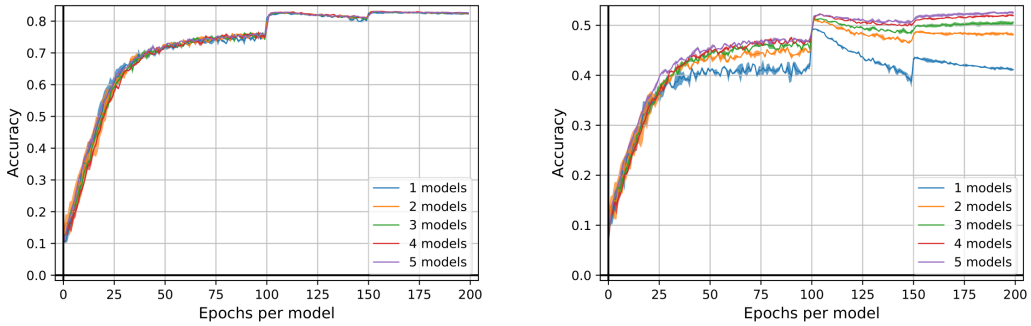



Figure 9.6: Standard and Robust accuracy (respectively on left and on right) on CIFAR-10 test images in function of the number of epochs per classifier with 1 to 5 ResNet18 models. The performed attack is PGD with 20 iterations and $\varepsilon = 8/255$. The best mixture for 5 models occurs at the end of training (epoch 198).

where essinf denotes the essential infimum. □

We can deduce an immediate corollary.

**Corollary 2.** *Under Assumption 3, the dual for randomized and deterministic classifiers are equal.*

### 9.11.2 Decomposition of the Empirical Risk for Entropic Regularization

**Proposition 9.11.2.** *Let* $\hat{\mathbb{P}} := \frac{1}{N}\sum_{i=1}^{N} \delta_{(x_i,y_i)}$. *Let* $l$ *be a loss satisfying Assumption 3. Then we have:*

$$\frac{1}{N}\sum_{i=1}^{N} \sup_{x,\ d(x,x_i)\leq\varepsilon} \mathbb{E}_{\theta\sim\mu}\left[l(\theta,(x,y))\right] = \sum_{i=1}^{N} \sup_{\mathbb{Q}_i\in\Gamma_{i,\varepsilon}} \mathbb{E}_{(x,y)\sim\mathbb{Q}_i,\theta\sim\mu}\left[l(\theta,(x,y))\right]$$

*where* $\Gamma_{i,\varepsilon}$ *is defined as :*

$$\Gamma_{i,\varepsilon} := \left\{\mathbb{Q}_i \mid \int d\mathbb{Q}_i = \frac{1}{N},\ \int c_\varepsilon((x_i,y_i),\cdot)d\mathbb{Q}_i = 0\right\}.$$

*Proof.* This proposition is a direct application of Proposition 9.2.2 for diracs $\delta_{(x_i,y_i)}$. □

### 9.11.3 On the NP-Hardness of Attacking a Mixture of Classifiers

In general, the problem of finding a best response to a mixture of classifiers is in general NP-hard. Let us justify it on a mixture of linear classifiers in binary classification: $f_{\theta_k}(x) = \langle \theta, x \rangle$ for $k \in [L]$ and $\boldsymbol{\lambda} = \mathbf{1}_L / L$. Let us consider the $\ell_2$ norm and $x = 0$ and $y = 1$. Then the problem of attacking $x$ is the following:

$$\sup_{\tau,\ \|\tau\| \leq \varepsilon} \frac{1}{L} \sum_{k=1}^{L} \mathbf{1}_{\langle \theta_k, \tau \rangle \leq 0}$$

This problem is equivalent to a linear binary classification problem on $\tau$, which is known to be NP-hard.

### 9.11.4 Case of Separated Conditional Distribtions

**Proposition 9.11.3.** *Let $\mathcal{Y} = \{-1, +1\}$. Let $\mathbb{P} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$. Let $\varepsilon > 0$. For $i \in \mathcal{Y}$, let us denote $\mathbb{P}_i$ the distribution of $\mathbb{P}$ conditionally to $y = i$. Let us assume that $d_{\mathcal{X}}(\mathrm{supp}(\mathbb{P}_{1+1}), \mathrm{supp}(\mathbb{P}_{-1})) > 2\varepsilon$. Let us consider the nearest neighbor deterministic classifier : $f(x) = d(x, \mathrm{supp}(\mathbb{P}_{+1})) - d(x, \mathrm{supp}(\mathbb{P}_{-1}))$ and the $0/1$ loss $l(f, (x, y)) = \mathbf{1}_{yf(x) \leq 0}$. Then $f$ satisfies both optimal standard and adversarial risks: $\mathcal{R}(f) = 0$ and $\mathcal{R}^{\varepsilon}_{adv}(f) = 0$.*

*Proof.* Let Let denote $p_i = \mathbb{P}(y = i)$. Then we have

$$\mathcal{R}^{\varepsilon}_{adv}(f) = p_{+1} \mathbb{E}_{\mathbb{P}_{+1}} \left[ \sup_{x',\ d(x,x') \leq \varepsilon} \mathbf{1}_{f(x') \leq 0} \right] + p_{-1} \mathbb{E}_{\mathbb{P}_{-1}} \left[ \sup_{x',\ d(x,x') \leq \varepsilon} \mathbf{1}_{f(x') \geq 0} \right]$$

For $x \in \mathrm{supp}(\mathbb{P}_{+1})$, we have, for all $x'$ such that $d(x, x') \neq 0$, $f(x') > 0$, then: $\mathbb{E}_{\mathbb{P}_{+1}} \left[ \sup_{x',\ d(x,x') \leq \varepsilon} \mathbf{1}_{f(x') \leq 0} \right] = 0$. Similarly, we have $\mathbb{E}_{\mathbb{P}_{-1}} \left[ \sup_{x',\ d(x,x') \leq \varepsilon} \mathbf{1}_{f(x') \geq 0} \right] = 0$. We then deduce the result. □

# Conclusion

In this thesis, we proposed new regularization schemes of the OT problem using low-rank methods and studied two ML problems in fairness and robustness using the OT formalism. More precisely, we proposed contributions lying at the intersection of the entropic OT and low-rank methods, and two applications of the Kantorovich relaxation for the fair division problem and the adversarial attack problem, which we summarize below.

## Low-rank Optimal Transport

In chapter 4, we proposed a new approach to speeding up the resolution of entropy regularized OT with the Sinkhorn algorithm by considering a low nonnegative rank factorization of the kernel matrix. By incorporating parameterized feature maps, we are able to approximate the entropic OT with common cost functions while maintaining the positiveness of the factorization. Furthermore, we showed that our approach is highly versatile and can be used as a successful extension to the OT-GAN framework for training GANs at scale with linear time iterations. Our contributions represent a significant step towards improving the computational efficiency of optimal transport methods, with potential applications in a range of fields such as computer vision and natural language processing.

In chapter 5, we introduced a new regularization scheme for optimal transport problems, called Low-rank Optimal Transport (LOT), which imposes a low nonnegative rank constraint on the feasible set of couplings. By directly constraining the couplings rather than approximating the kernel, our approach can solve the OT problem under low-rank constraints with arbitrary costs. This is achieved through a generic approach that optimizes jointly on sub-couplings and a common marginal distribution using a mirror-descent approach. We showed that our algorithm is guaranteed to converge and can achieve linear time complexity when low rank assumptions are exploited on the cost matrix. Overall, LOT provides a promising new direction for optimal transport regularization that can lead to more efficient and scalable solutions for a wide range of applications.

In chapter 6, we assembled theoretical and practical arguments to support low-rank factorizations for OT. We have presented two controls: one concerning the approximation error to the true optimal transport and another concerning the statistical rates of the plug-in estimator. The latter is showed to be independent of the dimension, which is of particular interest when studying OT in ML settings. We have motivated further the use of LOT as a loss by introducing its debiased version and showed that it possesses desirable properties: positivity and metrization of the convergence in law. We have also presented the links between the bias induced by such regularization and clustering methods, and studied empirically the effects of hyperparameters involved in the practical estimation of LOT. The strong theoretical foundations provided in this paper motivate further studies of the empirical behaviour of LOT estimator, notably on finding suitable local minima and on improvements on the convergence of the MD scheme using other adaptive choices for step sizes.

In chapter 7, we showed that the factorization introduced in [3] to speed up classic OT delivers an even larger impact when applied to GW: indeed, the combination of low-rank couplings with-low rank cost matrices is the only one, to our knowledge, that achieves linear time/memory complexity for the Gromov-Wasserstein problem. By adding low-rank constraints, our goal is no longer to approach the optimal coupling, but rather to promote low-rank solutions among many that have a low GW cost. We showed in experiments that low-rank couplings can reach low GW costs with similar performance as the entropic regularization, the current default approach, while being much faster to compute and that they are directly useful in real-world tasks. Our assumptions to reach linearity mostly rest on two important assumptions: the rank of distance matrices (the intrinsic dimensionality of data points) must be dominated by the number of points and that a small enough rank $r$ is able to capture the configuration of the input measures. Pending these constraints, which are valid in most relevant experimental setups we know of, we have demonstrated that our approach is versatile, remains faithful to the original GW formulation, and scales to sizes that are out of reach for the SoTA entropic solver.

## Applications of OT in Machine Learning

In chapter 8, we proposed a relaxed version of the fair division problem, called EOT (Equitable and Optimal Transport), using the fundamental idea of Kantorovich in order to relax the optimal transport problem intially defined by Monge. By considering resources as distributions rather than sets, we showed that EOT

exploits the divisibility of the resources and provides an equitable, optimal, and proportional partition of the resources by maximizing the minimum of individual utilities. The dual formulation of EOT is derived, along with strong duality results. We established the relationship between EOT and some common Integral Probability Metrics and proposed an entropic regularized version of the problem, which can be approximated using an efficient algorithm similar to the Sinkhorn algorithm. Our work contributes to the growing literature on fair resource allocation and opens new perspectives on the application of optimal transport for applied problems.

In chapter 9, we have presented a novel game-theoretic perspective of the adversarial risk minimization problem using the fundamental principle introduced by Kantorovich [32] to relax the Monge formulation of optimal transport. By viewed the adversary as a coupling rather than a deterministic map, we showed that the adversarial risk minimization problem can be reformulated as a distributionally robust optimization problem over Wasserstein balls. We the studied the existence of Nash equilibria in this two-players zero-sum game and showed its existence when we allow the classifier to be random. We also proposed a framework for learning a robust mixture of classifiers which leads to improved robustness against adversarial attacks. The results demonstrate the effectiveness of our method and its potential for practical applications in the field of adversarial machine learning.

# Bibliography


[1] Meyer Scetbon and Marco Cuturi. Low-rank optimal transport: Approximation, statistics and debiasing, 2022. URL `https://arxiv.org/abs/2205.12365`.

[2] Meyer Scetbon, Gabriel Peyré, and Marco Cuturi. Linear-time gromov wasserstein distances using low rank couplings and costs. *ICML*, 2022.

[3] Meyer Scetbon, Marco Cuturi, and Gabriel Peyré. Low-rank sinkhorn factorization, 2021.

[4] Laurent Meunier*, Meyer Scetbon*, Rafael B Pinot, Jamal Atif, and Yann Chevaleyre. Mixed nash equilibria in the adversarial examples game. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 7677–7687. PMLR, 18–24 Jul 2021.

[5] Meyer Scetbon*, Laurent Meunier*, Jamal Atif, and Marco Cuturi. Equitable and optimal transport with multiple agents. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*. PMLR, 13–15 Apr 2021.

[6] Meyer Scetbon and Marco Cuturi. Linear time sinkhorn divergences using positive features. In *Advances in Neural Information Processing Systems*, 2020.

[7] Meyer Scetbon, Laurent Meunier, and Yaniv Romano. An asymptotic test for conditional independence using analytic kernel embeddings. In *Proceedings of the 39th International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 2022.

[8] Nicholas J. Irons, Meyer Scetbon, Soumik Pal, and Zaid Harchaoui. Triangular flows for generative modeling: Statistical consistency, smoothness classes, and fast rates. In *Proceedings of The 25th International Conference on Artificial*

*Intelligence and Statistics*, Proceedings of Machine Learning Research. PMLR, 2022.

[9] Meyer Scetbon, Michael Elad, and Peyman Milanfar. Deep k-svd denoising. *IEEE Transactions on Image Processing*, 30:5944–5955, 2021. doi: 10.1109/TIP.2021.3090531.

[10] Meyer Scetbon and Zaid Harchaoui. A spectral analysis of dot-product kernels. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research. PMLR, 2021.

[11] Meyer Scetbon and Zaid Harchaoui. Harmonic decompositions of convolutional networks. In *Proceedings of the 37th International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 2020.

[12] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[13] Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: A survey of some recent advances. *ESAIM: probability and statistics*, 9:323–375, 2005.

[14] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[15] Ronald DeVore, Gerard Kerkyacharian, Dominique Picard, and Vladimir Temlyakov. Approximation methods for supervised learning. *Foundations of Computational Mathematics*, 6:3–58, 2006.

[16] Stephen M Stigler. The epic story of maximum likelihood. *Statistical Science*, pages 598–620, 2007.

[17] Frank J Massey Jr. The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78, 1951.

[18] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

[19] Judea Pearl. Causal inference in statistics: An overview. 2009.

[20] Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. *arXiv preprint arXiv:1202.3775*, 2012.

[21] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[22] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[23] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015.

[24] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[26] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[27] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[28] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[29] Imre Csiszár. I-divergence geometry of probability distributions and minimization problems. *The annals of probability*, pages 146–158, 1975.

[30] Cédric Villani. *Topics in optimal transportation*. Number 58. American Mathematical Soc., 2003.

[31] Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences*, pages 666–704, 1781.

[32] Leonid Kantorovich. On the transfer of masses (in russian). *Doklady Akademii Nauk*, 37(2):227–229, 1942.

[33] Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5-6), 2019. ISSN 1935-8245.

[34] Ashok Makkuva, Amirhossein Taghvaei, Sewoong Oh, and Jason Lee. Optimal transport mapping via input convex neural networks. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6672–6681. PMLR, 13–18 Jul 2020.

[35] Alexander Korotin, Lingxiao Li, Justin Solomon, and Evgeny Burnaev. Continuous wasserstein-2 barycenter estimation without minimax optimization. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=3tFAs5E-Pe`.

[36] Alexander Tong, Jessie Huang, Guy Wolf, David Van Dijk, and Smita Krishnaswamy. TrajectoryNet: A dynamic optimal transport network for modeling cellular dynamics. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9526–9536. PMLR, 13–18 Jul 2020.

[37] Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, November 2000.

[38] Julien Rabin, Gabriel Peyré, Julie Delon, and Marc Bernot. Wasserstein barycenter and its application to texture mixing. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 435–446. Springer, 2011.

[39] Nicolas Bonneel, Michiel Van De Panne, Sylvain Paris, and Wolfgang Heidrich. Displacement interpolation using lagrangian mass transport. *ACM Transactions on Graphics*, 30(6):158, 2011.

[40] Steven Haker, Lei Zhu, Allen Tannenbaum, and Sigurd Angenent. Optimal mass transport for registration and warping. *International Journal of Computer Vision*, 60(3):225–240, 2004.

[41] Geoffrey Schiebinger, Jian Shu, Marcin Tabaka, Brian Cleary, Vidya Subramanian, Aryeh Solomon, Joshua Gould, Siyan Liu, Stacie Lin, Peter Berube, et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943, 2019.

[42] Morgan A Schmitz, Matthieu Heitz, Nicolas Bonneel, Fred Ngole, David Coeurjolly, Marco Cuturi, Gabriel Peyré, and Jean-Luc Starck. Wasserstein dictionary learning: Optimal transport-based unsupervised nonlinear dictionary learning. *SIAM Journal on Imaging Sciences*, 11(1):643–678, 2018.

[43] Matthieu Heitz, Nicolas Bonneel, David Coeurjolly, Marco Cuturi, and Gabriel Peyré. Ground metric learning on graphs. *Journal of Mathematical Imaging and Vision*, pages 1–19, 2020.

[44] Hicham Janati, Thomas Bazeille, Bertrand Thirion, Marco Cuturi, and Alexandre Gramfort. Multi-subject meg/eeg source imaging with sparse multi-task regression. *NeuroImage*, page 116847, 2020.

[45] Sunil Koundal, Rena Elkin, Saad Nadeem, Yuechuan Xue, Stefan Constantinou, Simon Sanggaard, Xiaodan Liu, Brittany Monte, Feng Xu, William Van Nostrand, et al. Optimal mass transport with lagrangian workflow reveals advective and diffusion driven solute transport in the glymphatic system. *Scientific reports*, 10(1):1–18, 2020.

[46] Nicolas Bonneel, Gabriel Peyré, and Marco Cuturi. Wasserstein barycentric coordinates: histogram regression using optimal transport. *ACM Transactions on Graphics*, 35(4):71:1–71:10, 2016.

[47] Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov-wasserstein averaging of kernel and distance matrices. In *International Conference on Machine Learning*, pages 2664–2672, 2016.

[48] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5099–5108. Curran Associates, Inc., 2017.

[49] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. *Proceedings of the 34th International Conference on Machine Learning*, 70:214–223, 2017.

[50] Tim Salimans, Han Zhang, Alec Radford, and Dimitris Metaxas. Improving GANs using optimal transport. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=rkQkBnJAb`.

[51] Aude Genevay, Lénaic Chizat, Francis Bach, Marco Cuturi, and Gabriel Peyré. Sample complexity of sinkhorn divergences. *arXiv preprint arXiv:1810.02733*, 2018.

[52] David Alvarez-Melis and Tommi Jaakkola. Gromov-wasserstein alignment of word embedding spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1881–1890, 2018.

[53] Jean Alaux, Edouard Grave, Marco Cuturi, and Armand Joulin. Unsupervised hyper-alignment for multilingual word embeddings. In *International Conference on Learning Representations*, 2019.

[54] Edouard Grave, Armand Joulin, and Quentin Berthet. Unsupervised alignment of embeddings with wasserstein procrustes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1880–1890, 2019.

[55] Tatsunori Hashimoto, David Gifford, and Tommi Jaakkola. Learning population-level diffusions with generative RNNs. In *International Conference on Machine Learning*, pages 2417–2426, 2016.

[56] Karren Dai Yang, Karthik Damodaran, Saradha Venkatachalapathy, Ali C Soylemezoglu, GV Shivashankar, and Caroline Uhler. Predicting cell lineages using autoencoders and optimal transport. *PLoS computational biology*, 16 (4):e1007828, 2020.

[57] Nicolas Courty, Rémi Flamary, and Devis Tuia. Domain adaptation with regularized optimal transport. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 274–289. Springer, 2014.

[58] Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9):1853–1865, 2017.

[59] Nicolas Bonneel, Gabriel Peyré, and Marco Cuturi. Wasserstein barycentric coordinates: histogram regression using optimal transport. *ACM Trans. Graph.*, 35(4):71–1, 2016.

[60] Facundo Mémoli. Gromov–Wasserstein distances and the metric approach to object matching. *Foundations of Computational Mathematics*, 11(4):417–487, 2011.

[61] Pinar Demetci, Rebecca Santorella, Björn Sandstede, William Stafford Noble, and Ritambhara Singh. Gromov-wasserstein optimal transport to align single-cell multi-omics data. *bioRxiv*, 2020. doi: 10.1101/2020.04.28.066787.

[62] Andrew J Blumberg, Mathieu Carriere, Michael A Mandell, Raul Rabadan, and Soledad Villar. Mrec: a fast and versatile framework for aligning and

matching point clouds with applications to single cell molecular data. *arXiv preprint arXiv:2001.01666*, 2020.

[63] Hongteng Xu, Dixin Luo, Hongyuan Zha, and Lawrence Carin Duke. Gromov-wasserstein learning for graph matching and node embedding. In *International conference on machine learning*, pages 6932–6941. PMLR, 2019.

[64] Charlotte Bunne, David Alvarez-Melis, Andreas Krause, and Stefanie Jegelka. Learning generative models across incomparable spaces. *arXiv preprint arXiv:1905.05461*, 2019.

[65] Laetitia Chapel, Mokhtar Alaya, and Gilles Gasso. Partial optimal transport with applications on positive-unlabeled learning. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, 2020.

[66] Titouan Vayer, Laetita Chapel, Rémi Flamary, Romain Tavenard, and Nicolas Courty. Fused gromov-wasserstein distance for structured objects: theoretical foundations and mathematical properties. *arXiv preprint arXiv:1811.02834*, 2018.

[67] Danielle Ezuz, Justin Solomon, Vladimir G Kim, and Mirela Ben-Chen. Gwcnn: A metric alignment layer for deep shape analysis. In *Computer Graphics Forum*, volume 36, pages 49–57. Wiley Online Library, 2017.

[68] George B. Dantzig. Linear programming. In J. K. Lenstra, A. H. G. Rinnooy Kan, and A. Schrijver, editors, *History of mathematical programming: a collection of personal reminiscences*, pages 257–282. Elsevier Science Publishers, 1991.

[69] Lester Randolph Ford and Delbert Ray Fulkerson. *Flows in Networks*. Princeton University Press, 1962.

[70] Andrew V Goldberg and Robert E Tarjan. Finding minimum-cost circulations by canceling negative cycles. *Journal of the ACM (JACM)*, 36(4):873–886, 1989.

[71] Nicolas Fournier and Arnaud Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3-4):707–738, 2015.

[72] Jonathan Weed, Francis Bach, et al. Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. *Bernoulli*, 25(4A):2620–2648, 2019.

[73] Jonathan Niles-Weed and Philippe Rigollet. Estimation of wasserstein distances in the spiked transport model. *Bernoulli*, 28(4):2663–2688, 2022.

[74] Lenaic Chizat, Pierre Roussillon, Flavien Léger, François-Xavier Vialard, and Gabriel Peyré. Faster wasserstein distance estimation with the sinkhorn divergence. *Advances in Neural Information Processing Systems*, 33, 2020.

[75] Christian Clason, Dirk A. Lorenz, Hinrich Mahler, and Benedikt Wirth. Entropic regularization of continuous optimal transport problems. *Journal of Mathematical Analysis and Applications*, 494(1):124432, 2021. ISSN 0022-247X. doi: https://doi.org/10.1016/j.jmaa.2020.124432.

[76] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300, 2013.

[77] Arnaud Dessein, Nicolas Papadakis, and Jean-Luc Rouas. Regularized optimal transport and the rot mover's distance. *The Journal of Machine Learning Research*, 19(1):590–642, 2018.

[78] Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with sinkhorn divergences, 2017.

[79] Lenaïc Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced optimal transport problems. *Math. Comput.*, 87(314):2563–2609, 2018.

[80] Aude Genevay, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic optimization for large-scale optimal transport. In *Advances in Neural Information Processing Systems*, pages 3440–3448, 2016.

[81] Aude Genevay, Lénaic Chizat, Francis Bach, Marco Cuturi, and Gabriel Peyré. Sample complexity of sinkhorn divergences. 2019.

[82] Gonzalo Mena and Jonathan Niles-Weed. Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem. *Advances in Neural Information Processing Systems*, 32, 2019.

[83] Eliane Maria Loiola, Nair Maria Maia de Abreu, Paulo Oswaldo Boaventura-Netto, Peter Hahn, and Tania Querido. A survey for the quadratic assignment problem. *European Journal Operational Research*, 176(2):657–690, 2007.

[84] Steven Gold and Anand Rangarajan. Softassign versus softmax: Benchmarks in combinatorial optimization. *Advances in neural information processing systems*, pages 626–632, 1996.

[85] Justin Solomon, Gabriel Peyré, Vladimir G Kim, and Suvrit Sra. Entropic metric alignment for correspondence problems. *ACM Transactions on Graphics (TOG)*, 35(4):1–13, 2016.

[86] Facundo Mémoli. Gromov–wasserstein distances and the metric approach to object matching. *Foundations of computational mathematics*, 11(4):417–487, 2011.

[87] Ryoma Sato, Marco Cuturi, Makoto Yamada, and Hisashi Kashima. Fast and robust comparison of probability measures in heterogeneous spaces. *arXiv preprint arXiv:2002.01615*, 2020.

[88] Titouan Vayer, Rémi Flamary, Romain Tavenard, Laetitia Chapel, and Nicolas Courty. Sliced gromov-wasserstein. *arXiv preprint arXiv:1905.10124*, 2019.

[89] Tam Le, Nhat Ho, and Makoto Yamada. Flow-based alignment approaches for probability measures in different spaces. In *International Conference on Artificial Intelligence and Statistics*, pages 3934–3942. PMLR, 2021.

[90] Samir Chowdhury, David Miller, and Tom Needham. Quantized gromov-wasserstein. *arXiv preprint arXiv:2104.02013*, 2021.

[91] Hongteng Xu, Dixin Luo, and Lawrence Carin. Scalable gromov-wasserstein learning for graph partitioning and matching. *arXiv preprint arXiv:1905.07645*, 2019.

[92] Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1853–1865, 2016.

[93] Guillaume Carlier, Alfred Galichon, and Filippo Santambrogio. From Knöthe's transport to Brenier's map and a continuation method for optimal transport. *SIAM Journal on Mathematical Analysis*, 41(6):2554–2576, 2010.

[94] H. Steinhaus. Sur la division pragmatique. *Econometrica*, 17:315–319, 1949. ISSN 00129682, 14680262. URL `http://www.jstor.org/stable/1907319`.

[95] Tor Lattimore, Koby Crammer, and Csaba Szepesvári. Linear multi-resource allocation with semi-bandit feedback. In *Advances in Neural Information Processing Systems*, pages 964–972, 2015.

[96] Hervé Moulin. *Fair division and collective welfare*. MIT press, 2004.

[97] Lester E Dubins and Edwin H Spanier. How to cut a cake fairly. *The American Mathematical Monthly*, 68(1P1):1–17, 1961.

[98] Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang, and Ariel D Procaccia. *Handbook of computational social choice*. Cambridge University Press, 2016.

[99] John Cloutier, Kathryn L Nyman, and Francis Edward Su. Two-player envy-free multi-cake division. *Mathematical Social Sciences*, 59(1):26–37, 2010.

[100] Erika Mackin and Lirong Xia. Allocating indivisible items in categorized domains. *arXiv preprint arXiv:1504.05932*, 2015.

[101] Haibin Wang, Sujoy Sikdar, Xiaoxi Guo, Lirong Xia, Yongzhi Cao, and Hanpin Wang. Multi-type resource allocation with partial preferences. *arXiv preprint arXiv:1906.06836*, 2019.

[102] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 387–402. Springer, 2013.

[103] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

[104] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[105] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Jonathan Uesato, and Pascal Frossard. Robustness via curvature regularization, and vice versa. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9078–9086, 2019.

[106] Jeremy M. Cohen, Elan Rosenfeld, and J. Zico Kolter. Certified adversarial robustness via randomized smoothing. *arXiv preprint arXiv:1902.02918*.

[107] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

[108] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016.

[109] Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pages 3–14, 2017.

[110] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International Conference on Machine Learning*, 2020.

[111] Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94, 2015.

[112] Marco Cuturi. Sinkhorn distances: lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems 26*, pages 2292–2300, 2013.

[113] Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouve, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2681–2690, 2019.

[114] Alexandre Gramfort, Gabriel Peyré, and Marco Cuturi. Fast optimal transport averaging of neuroimaging data. In *Information Processing in Medical Imaging - 24th International Conference, IPMI 2015*, pages 261–272, 2015.

[115] Justin Solomon, Leonidas Guibas, and Adrian Butscher. Dirichlet energy for analysis and synthesis of soft maps. In *Computer Graphics Forum*, volume 32, pages 197–206. Wiley Online Library, 2013.

[116] Justin Solomon, Raif Rustamov, Leonidas Guibas, and Adrian Butscher. Earth mover's distances on discrete surfaces. *Transaction on Graphics*, 33(4), 2014.

[117] Jason Altschuler, Jonathan Weed, and Philippe Rigollet. Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. *arXiv preprint arXiv:1705.09634*, 2017.

[118] Pavel Dvurechensky, Alexander Gasnikov, and Alexey Kroshnin. Computational optimal transport: Complexity by accelerated gradient descent is better than by sinkhorn's algorithm. *arXiv preprint arXiv:1802.04367*, 2018.

[119] Tianyi Lin, Nhat Ho, and Michael Jordan. On efficient optimal transport: An analysis of greedy and accelerated mirror descent algorithms. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine*

*Learning Research*, pages 3982–3991, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL `http://proceedings.mlr.press/v97/lin19a.html`.

[120] Gabriel Peyré and Marco Cuturi. Metric learning: a survey. *Foundations and Trends in Machine Learning*, 11(5-6), 2019.

[121] Justin Solomon, Fernando De Goes, Gabriel Peyré, Marco Cuturi, Adrian Butscher, Andy Nguyen, Tao Du, and Leonidas Guibas. Convolutional Wasserstein distances: efficient optimal transportation on geometric domains. *ACM Transactions on Graphics*, 34(4):66:1–66:11, 2015.

[122] Jason Altschuler, Francis Bach, Alessandro Rudi, and Jonathan Weed. Massively scalable sinkhorn distances via the nystr\" om method. *arXiv preprint arXiv:1812.05189*, 2018.

[123] Jason M. Altschuler and Enric Boix-Adsera. Polynomial-time algorithms for multimarginal optimal transport problems with structure, 2020.

[124] Aden Forrow, Jan-Christian Hütter, Mor Nitzan, Philippe Rigollet, Geoffrey Schiebinger, and Jonathan Weed. Statistical optimal transport via factored couplings, 2018.

[125] Martial Agueh and Guillaume Carlier. Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.

[126] Marco Cuturi and Arnaud Doucet. Fast computation of Wasserstein barycenters. In *Proceedings of ICML*, volume 32, pages 685–693, 2014.

[127] Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. Iterative Bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

[128] Aden Forrow, Jan-Christian Hütter, Mor Nitzan, Philippe Rigollet, Geoffrey Schiebinger, and Jonathan Weed. Statistical optimal transport via factored couplings. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2454–2465. PMLR, 2019.

[129] Richard M. Dudley. The speed of mean Glivenko-Cantelli convergence. *Annals of Mathematical Statistics*, 40(1):40–50, 1969.

[130] Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-Ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. *arXiv preprint arXiv:1810.08278*, 2018.

[131] Rainer E Burkard, Eranda Cela, Panos M Pardalos, and Leonidas S Pitsoulis. The quadratic assignment problem. In *Handbook of combinatorial optimization*, pages 1713–1809. Springer, 1998.

[132] Ton Steerneman. On the total variation and hellinger distance between signed measures; an application to product measures. *Proceedings of the American Mathematical Society*, 88(4):684–688, 1983.

[133] Gintare Karolina Dziugaite, Daniel M Roy, and Zoubin Ghahramani. Training generative neural networks via maximum mean discrepancy optimization. *arXiv preprint arXiv:1505.03906*, 2015.

[134] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

[135] Youssef Mroueh and Tom Sercu. Fisher gan. In *Advances in Neural Information Processing Systems*, pages 2513–2523, 2017.

[136] Hisham Husain, Richard Nock, and Robert C Williamson. A primal-dual link between gans and autoencoders. In *Advances in Neural Information Processing Systems*, pages 413–422, 2019.

[137] M. Scetbon and G. Varoquaux. Comparing distributions: $\ell_1$ geometry improves kernel two-sample testing, 2019.

[138] Bharath K Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, Gert RG Lanckriet, et al. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6:1550–1599, 2012.

[139] Guneet S. Dhillon, Kamyar Azizzadenesheli, Jeremy D. Bernstein, Jean Kossaifi, Aran Khanna, Zachary C. Lipton, and Animashree Anandkumar. Stochastic activation pruning for robust adversarial defense. In *International Conference on Learning Representations*, 2018.

[140] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. Mitigating adversarial effects through randomization. In *International Conference on Learning Representations*, 2018.

[141] Rafael Pinot, Laurent Meunier, Alexandre Araujo, Hisashi Kashima, Florian Yger, Cédric Gouy-Pailler, and Jamal Atif. Theoretical evidence for adversarial robustness through randomization. In *Advances in Neural Information Processing Systems*, pages 11838–11848, 2019.

[142] Bao Wang, Zuoqiang Shi, and Stanley Osher. Resnets ensemble via the feynman-kac formalism to improve natural and robust accuracies. In *Advances in Neural Information Processing Systems 32*, pages 1655–1665. Curran Associates, Inc., 2019.

[143] Rafael Pinot, Raphael Ettedgui, Geovani Rizk, Yann Chevaleyre, and Jamal Atif. Randomization matters. how to defend against strong adversarial attacks. *International Conference on Machine Learning*, 2020.

[144] S. Rota Bulò, B. Biggio, I. Pillai, M. Pelillo, and F. Roli. Randomized prediction games for adversarial machine learning. *IEEE Transactions on Neural Networks and Learning Systems*, 28(11):2466–2478, 2017.

[145] Juan C. Perdomo and Yaron Singer. Robust attacks against multiple classifiers. *arXiv preprint arXiv:1906.02816*, 2019.

[146] Avishek Joey Bose, Gauthier Gidel, Hugo Berard, Andre Cianflone, Pascal Vincent, Simon Lacoste-Julien, and William L. Hamilton. Adversarial example games, 2021.

[147] Michael Brückner and Tobias Scheffer. Stackelberg games for adversarial prediction problems. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, page 547–555, New York, NY, USA, 2011. Association for Computing Machinery. doi: 10.1145/2020408.2020495.

[148] Han Bao, Clay Scott, and Masashi Sugiyama. Calibrated surrogate losses for adversarially robust classification. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 408–451. PMLR, 09–12 Jul 2020. URL `http://proceedings.mlr.press/v125/bao20a.html`.

[149] Zac Cranko, Aditya Menon, Richard Nock, Cheng Soon Ong, Zhan Shi, and Christian Walder. Monge blunts bayes: Hardness results for adversarial training. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1406–1415. PMLR, 09–15 Jun 2019. URL `http://proceedings.mlr.press/v97/cranko19a.html`.

[150] Arjun Nitin Bhagoji, Daniel Cullina, and Prateek Mittal. Lower bounds on adversarial robustness from optimal transport. In *Advances in Neural Information Processing Systems 32*, pages 7496–7508. Curran Associates, Inc., 2019.

[151] Muni Sreenivas Pydi and Varun Jog. Adversarial risk via optimal transport and optimal couplings. In *International Conference on Machine Learning*. 2020.

[152] Aman Sinha, Hongseok Namkoong, and John Duchi. Certifying some distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*, 2017.

[153] Jaeho Lee and Maxim Raginsky. Minimax statistical learning with wasserstein distances. In *Advances in Neural Information Processing Systems 31*, pages 2687–2696. Curran Associates, Inc., 2018.

[154] Zhuozhuo Tu, Jingwei Zhang, and Dacheng Tao. Theoretical analysis of adversarial learning: A minimax approach. *arXiv preprint arXiv:1811.05232*, 2018.

[155] Jose Blanchet and Karthyek Murthy. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600, 2019.

[156] Stephen Boyd. Subgradient methods. 2003.

[157] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

[158] Cedric Villani. *Optimal Transport: Old and New*, volume 338. Springer Verlag, 2009.

[159] Filippo Santambrogio. *Optimal transport for applied mathematicians*. Birkhauser, 2015.

[160] Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on Pure and Applied Mathematics*, 44(4): 375–417, 1991.

[161] Ravindra K Ahuja, Thomas L Magnanti, James B Orlin, and MR Reddy. Applications of network optimization. *Handbooks in Operations Research and Management Science*, 7:1–83, 1995.

[162] Richard Sinkhorn and Paul Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal Mathematics*, 21:343–348, 1967.

[163] Lev M Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3):200–217, 1967.

[164] Joel Franklin and Jens Lorenz. On the scaling of multidimensional matrices. *Linear Algebra and its Applications*, 114:717–735, 1989.

[165] Facundo Memoli. On the use of Gromov-Hausdorff Distances for Shape Comparison. In M. Botsch, R. Pajarola, B. Chen, and M. Zwicker, editors, *Eurographics Symposium on Point-Based Graphics*. The Eurographics Association, 2007. ISBN 978-3-905673-51-7. doi: 10.2312/SPBG/SPBG07/081-090.

[166] Karl-Theodor Sturm. The space of spaces: curvature bounds and gradient flows on the space of metric measure spaces. *arXiv preprint arXiv:1208.0434*, 2012.

[167] Samir Chowdhury and Facundo Mémoli. The gromov–wasserstein distance between networks and stable network invariants. *Information and Inference: A Journal of the IMA*, 8(4):757–787, 2019.

[168] Tjalling C Koopmans and Martin Beckmann. Assignment problems and the location of economic activities. *Econometrica: journal of the Econometric Society*, pages 53–76, 1957.

[169] Anand Rangarajan, Alan L Yuille, Steven Gold, and Eric Mjolsness. Convergence properties of the softassign quadratic assignment algorithm. *Neural Computation*, 11(6):1455–1474, August 1999.

[170] Steven Gold and Anand Rangarajan. A graduated assignment algorithm for graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(4):377–388, April 1996.

[171] Cedric Villani. *Topics in Optimal Transportation*. Graduate Studies in Mathematics Series. American Mathematical Society, 2003. ISBN 9780821833124.

[172] Robert E. Tarjan. Dynamic trees as search trees via euler tours, applied to the network simplex algorithm. *Mathematical Programming*, 78(2):169–177, 1997.

[173] Dimitris Bertsimas and John N Tsitsiklis. *Introduction to Linear Optimization*. Athena Scientific, 1997.

[174] Aaditya Ramdas, Nicolás García Trillos, and Marco Cuturi. On Wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19 (2):47, 2017.

[175] Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with Sinkhorn divergences. In *Proceedings of AISTATS*, pages 1608–1617, 2018.

[176] Pavel Dvurechensky, Alexander Gasnikov, and Alexey Kroshnin. Computational optimal transport: Complexity by accelerated gradient descent is better than by sinkhorn's algorithm. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1367–1376. PMLR, 10–15 Jul 2018.

[177] Vladimir Oliker. Embedding $S_n$ into $\mathbb{R}^{n+1}$ with given integral gauss curvature and optimal mass transport on $S_n$. *Advances in Mathematics*, 213(2):600 – 620, 2007.

[178] Cameron Musco and Christopher Musco. Recursive sampling for the nyström method, 2016.

[179] Sergey Guminov, Pavel Dvurechensky, Nazarii Tupitsa, and Alexander Gasnikov. Accelerated alternating minimization, accelerated sinkhorn's algorithm and accelerated iterative bregman projections, 2019.

[180] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in neural information processing systems*, pages 1177–1184, 2008.

[181] Alex J. Smola, Zoltán L. Óvári, and Robert C Williamson. Regularization with dot-product kernels. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 308–314. MIT Press, 2001.

[182] Youngmin Cho and Lawrence K. Saul. Kernel methods for deep learning. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 342–350. Curran Associates, Inc., 2009.

[183] Francis Bach. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(19):1–53, 2017.

[184] Julien Mairal, Piotr Koniusz, Zaid Harchaoui, and Cordelia Schmid. Convolutional kernel networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2627–2635. Curran Associates, Inc., 2014. URL `http://papers.nips.cc/paper/5348-convolutional-kernel-networks.pdf`.

[185] Aude Genevay, Gabriel Peyré, and Marco Cuturi. GAN and VAE from an optimal transport point of view. (arXiv preprint arXiv:1706.01807), 2017.

[186] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks, 2015.

[187] Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabás Póczos. Mmd gan: Towards deeper understanding of moment matching network. *arXiv preprint arXiv:1705.08584*, 2017.

[188] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research). URL `http://www.cs.toronto.edu/~kriz/cifar.html`.

[189] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[190] Yu Nesterov. Smooth minimization of non-smooth functions. *Mathematical programming*, 103(1):127–152, 2005.

[191] Tamer Başar and Pierre Bernhard. *H-infinity optimal control and related minimax design problems: a dynamic game approach*. Springer Science & Business Media, 2008.

[192] Pierre Baldi, Peter Sadowski, and Daniel Whiteson. Searching for exotic particles in high-energy physics with deep learning. *Nature communications*, 5:4308, 2014.

[193] Bernhard Schmitzer. Stabilized sparse scaling algorithms for entropy regularized transport problems. *SIAM Journal on Scientific Computing*, 41(3): A1443–A1481, 2019.

[194] Yujia Xie, Xiangfeng Wang, Ruijia Wang, and Hongyuan Zha. A fast proximal point method for computing exact wasserstein distance. In *Uncertainty in Artificial Intelligence*, pages 433–453. PMLR, 2020.

[195] Joel E. Cohen and Uriel G. Rothblum. Nonnegative ranks, decompositions, and factorizations of nonnegative matrices. *Linear Algebra and its Applications*, 190:149 – 168, 1993. ISSN 0024-3795. doi: https://doi.org/10.1016/0024-3795(93)90224-C. URL `http://www.sciencedirect.com/science/article/pii/002437959390224C`.

[196] Richard L Dykstra. An algorithm for restricted least squares regression. *Journal of the American Statistical Association*, 78(384):837–842, 1983.

[197] Heinz H Bauschke and Adrian S Lewis. Dykstras algorithm with bregman projections: A convergence proof. *Optimization*, 48(4):409–427, 2000.

[198] Saeed Ghadimi, Guanghui Lan, and Hongchao Zhang. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization, 2013.

[199] Heinz H Bauschke, Jérôme Bolte, and Marc Teboulle. A descent lemma beyond lipschitz gradient continuity: first-order methods revisited and applications. *Mathematics of Operations Research*, 42(2):330–348, 2017.

[200] Ainesh Bakshi and David P. Woodruff. Sublinear time low-rank approximation of distance matrices, 2018.

[201] Piotr Indyk, Ali Vakilian, Tal Wagner, and David Woodruff. Sample-optimal low-rank approximation of distance matrices, 2019.

[202] John Adrian Bondy, Uppaluri Siva Ramachandra Murty, et al. *Graph theory with applications*, volume 290. Macmillan London, 1976.

[203] Haihao Lu, Robert M. Freund, and Yurii Nesterov. Relatively-smooth convex optimization by first-order methods, and applications, 2017.

[204] Kelvin Shuangjian Zhang, Gabriel Peyré, Jalal Fadili, and Marcelo Pereyra. Wasserstein control of mirror langevin monte carlo, 2020.

[205] Pinar Demetci, Rebecca Santorella, Björn Sandstede, William Stafford Noble, and Ritambhara Singh. Gromov-wasserstein optimal transport to align single-cell multi-omics data. *BioRxiv*, 2020.

[206] Xinye Zheng, Jianbo Ye, James Z Wang, and Jia Li. Scott: Shape-location combined tracking with optimal transport. *SIAM Journal on Mathematics of Data Science*, 2(2):284–308, 2020.

[207] Weijie Liu, Chao Zhang, Nenggan Zheng, and Hui Qian. Approximating optimal transport via low-rank and sparse factorization. *arXiv preprint arXiv:2111.06546*, 2021.

[208] Samuel Kutin. Extensions to mcdiarmid's inequality when differences are bounded with high probability. *Dept. Comput. Sci., Univ. Chicago, Chicago, IL, USA, Tech. Rep. TR-2002-04*, 2002.

[209] Aude Genevay, Gabriel Peyre, and Marco Cuturi. Learning generative models with sinkhorn divergences. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research. PMLR, 09–11 Apr 2018.

[210] Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *The Annals of Statistics*, pages 2263–2291, 2013.

[211] Ryan D'Orazio, Nicolas Loizou, Issam Laradji, and Ioannis Mitliagkas. Stochastic mirror descent: Convergence analysis and adaptive variants via the mirror stochastic polyak stepsize. *arXiv preprint arXiv:2110.15412*, 2021.

[212] Anastasia Bayandina, Pavel Dvurechensky, Alexander Gasnikov, Fedor Stonyakin, and Alexander Titov. Mirror descent and convex optimization problems with non-smooth inequality constraints. In *Large-scale and distributed optimization*, pages 181–213. Springer, 2018.

[213] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[214] Marco Cuturi, Laetitia Meng-Papaxanthos, Yingtao Tian, Charlotte Bunne, Geoff Davis, and Olivier Teboul. Optimal transport tools (ott): A jax toolbox for all things wasserstein. *arXiv preprint arXiv:2201.12324*, 2022.

[215] Richard Sinkhorn. A relationship between arbitrary positive matrices and doubly stochastic matrices. *Ann. Math. Statist.*, 35:876–879, 1964.

[216] Tianyi Lin, Nhat Ho, and Michael Jordan. On efficient optimal transport: An analysis of greedy and accelerated mirror descent algorithms. In *International Conference on Machine Learning*, pages 3982–3991. PMLR, 2019.

[217] Jason Altschuler, Francis Bach, Alessandro Rudi, and Jonathan Weed. Approximating the quadratic transportation metric in near-linear time. *arXiv preprint arXiv:1810.10046*, 2018.

[218] Hiroshi Konno. Maximization of a convex quadratic function under linear constraints. *Mathematical programming*, 11(1):117–127, 1976.

[219] Richard Mansfield Dudley et al. Weak convergence of probabilities on nonseparable metric spaces and empirical measures on euclidean spaces. *Illinois Journal of Mathematics*, 10(1):109–126, 1966.

[220] Song Chen, Blue B Lake, and Kun Zhang. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nature biotechnology*, 37(12):1452–1457, 2019.

[221] Luke Zappia, Belinda Phipson, and Alicia Oshlack. Splatter: simulation of single-cell rna sequencing data. *bioRxiv*, 2017. doi: 10.1101/133173.

[222] Jie Liu, Yuanhao Huang, Ritambhara Singh, Jean-Philippe Vert, and William Stafford Noble. Jointly embedding multiple single-cell omics measurements. *BioRxiv*, page 644310, 2019.

[223] Blue B Lake, Song Chen, Brandon C Sos, Jean Fan, Gwendolyn E Kaeser, Yun C Yung, Thu E Duong, Derek Gao, Jerold Chun, Peter V Kharchenko, et al. Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nature biotechnology*, 36(1):70–80, 2018.

[224] Grace XY Zheng, Jessica M Terry, Phillip Belgrader, Paul Ryvkin, Zachary W Bent, Ryan Wilson, Solongo B Ziraldo, Tobias D Wheeler, Geoff P McDermott, Junjie Zhu, et al. Massively parallel digital transcriptional profiling of single cells. *Nature communications*, 8(1):1–12, 2017.

[225] F Alexander Wolf, Philipp Angerer, and Fabian J Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology*, 19(1):1–5, 2018.

[226] David Alvarez-Melis, Stefanie Jegelka, and Tommi S Jaakkola. Towards optimal transport with global invariances. *arXiv preprint 1806.09277*, 2018.

[227] Eric Wong, Frank R. Schmidt, and J. Zico Kolter. Wasserstein adversarial examples via projected sinkhorn iterations, 2019.

[228] Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997.

[229] François-Pierre Paty and Marco Cuturi. Subspace robust wasserstein distances. *arXiv preprint arXiv:1901.08949*, 2019.

[230] Ruilin Li, Xiaojing Ye, Haomin Zhou, and Hongyuan Zha. Learning to match via inverse optimal transport. *J. Mach. Learn. Res.*, 20:80–1, 2019.

[231] Haodong Sun, Haomin Zhou, Hongyuan Zha, and Xiaojing Ye. Learning cost functions for optimal transport. *arXiv preprint arXiv:2002.09650*, 2020.

[232] Arnaud Dupuy, Alfred Galichon, and Yifei Sun. Estimating matching affinity matrix under low-rank constraints. *Arxiv:1612.09585*, 2016.

[233] Mathis Petrovich, Chao Liang, Yanbin Liu, Yao-Hung Hubert Tsai, Linchao Zhu, Yi Yang, Ruslan Salakhutdinov, and Makoto Yamada. Feature robust optimal transport for high-dimensional data. *arXiv preprint arXiv:2005.12123*, 2020.

[234] Marilda Sotomayor and Alvin Roth. Two-sided matching: A study in game-theoretic modelling and analysis. *Econometric Society Monographs*, (18), 1990.

[235] Lénaïc Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Unbalanced optimal transport: Dynamic and kantorovich formulations. *Journal of Functional Analysis*, 274(11):3090–3123, 2018.

[236] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1): 183–202, 2009.

[237] Paul Tseng. On accelerated proximal gradient methods for convex-concave optimization. *submitted to SIAM Journal on Optimization*, 1, 2008.

[238] Shai Shalev-Shwartz and Yoram Singer. Efficient learning of label ranking by soft projections onto polyhedra. *Journal of Machine Learning Research*, 7 (Jul):1567–1599, 2006.

[239] Paul Dupuis and Richard S Ellis. *A weak convergence approach to the theory of large deviations*, volume 902. John Wiley & Sons, 2011.

[240] Haim Brezis. *Functional analysis, Sobolev spaces and partial differential equations*. Springer Science & Business Media, 2010.

[241] Maurice Sion. On general minimax theorems. *Pacific J. Math.*, 8(1):171–176, 1958. URL `https://projecteuclid.org:443/euclid.pjm/1103040253`.

[242] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[243] Weiran Wang and Miguel A. Carreira-Perpinan. Projection onto the probability simplex: An efficient algorithm with a simple proof, and an application, 2013.

[244] Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

[245] Man-Chung Yue, Daniel Kuhn, and Wolfram Wiesemann. On linear optimization over wasserstein balls. *arXiv preprint arXiv:2004.07162*, 2020.

[246] Dimitir P Bertsekas and Steven Shreve. *Stochastic optimal control: the discrete-time case*. 2004.

[247] Florian Tramèr, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. The space of transferable adversarial examples. *arXiv preprint arXiv:1704.03453*, 2017.

[248] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*, 2016.

[249] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P Xing, Laurent El Ghaoui, and Michael I Jordan. Theoretically principled trade-off between robustness and accuracy. *International conference on Machine Learning*, 2019.

[250] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[251] Leslie Rice, Eric Wong, and Zico Kolter. Overfitting in adversarially robust deep learning. In *International Conference on Machine Learning*, pages 8093–8104. PMLR, 2020.

[252] Sven Gowal, Chongli Qin, Jonathan Uesato, Timothy Mann, and Pushmeet Kohli. Uncovering the limits of adversarial training against norm-bounded adversarial examples. *arXiv preprint arXiv:2010.03593*, 2020.

[253] Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, Percy Liang, and John C Duchi. Unlabeled data improves adversarial robustness. *arXiv preprint arXiv:1905.13736*, 2019.

[254] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[255] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 87.1–87.12. BMVA Press, 2016. ISBN 1-901725-59-6. doi: 10.5244/C.30.87.

ECOLE
DOCTORALE
DE MATHEMATIQUES
HADAMARD

**Titre :** Avancées en Transport Optimal : Structures de Faible Rang et Applications à l'Apprentissage Automatique

**Mots clés :** Transport Optimal, Optimisation, Statistiques, Apprentissage Automatique

**Résumé :** Le transport optimal (TO) joue un rôle de plus en plus important en apprentissage automatique (AA) pour comparer des mesures de probabilités. Le problème du TO a été utilisé dans de nombreuses applications et formulé de plusieurs manières. Parmi ces formulations, le problème de Monge et le programme linéaire de Kantorovich se démarquent. La première implique de trouver une transformation efficace pour envoyer une mesure sur une autre, tandis que la seconde relâche la contrainte qu'impose Monge pour faire correspondre des mesures en autorisant la division des masses. Le TO de Kantorovich est beaucoup plus accessible aux calculs et a été la formulation la plus exploitée en sciences des données. Cependant, elle pose, dans sa forme originale, plusieurs défis lorsqu'elle est utilisée pour des problèmes appliqués : (i) calculer le TO entre des distributions discrètes équivaut à résoudre un programme linéaire large et coûteux qui nécessite une complexité super-cubique par rapport aux nombre de points; (ii) estimer le TO en utilisant des mesures échantillonnées est voué à l'échec en raison de la malédiction de la dimensionnalité. Ces problèmes peuvent être atténués en utilisant une régularisation entropique, résolue avec l'algorithme Sinkhorn, qui améliore à la fois les aspects statistiques et computationnels. Bien que beaucoup plus rapide, le TO entropique nécessite toujours une complexité quadratique par rapport au nombre de points et reste donc prohibitif pour les problèmes à grande échelle. Profitant de cette opportunité, j'ai consacré une partie importante de ma thèse à travailler sur des nouvelles approches de calculs pour le TO, ce qui a conduit à ma ligne de travail sur l'introduction du transport optimal de faible rang. J'ai également réalisé que l'idée fondamentale proposée par Kantorovich pour relaxer le TO pouvait être appliquée dans d'autres contextes, et j'ai proposé de nouvelles approches en utilisant cette même idée pour aborder le problème de la division équitable et le problème des attaques adverses à travers le prisme du TO. Cette thèse est donc divisée en deux parties principales. Dans la première partie, je présente de nouvelles approches de régularisation pour le problème du TO, ainsi que son extension quadratique, le problème de Gromov-Wasserstein (GW), en imposant des structures de faible rang sur les couplages. Les algorithmes obtenus possèdent une complexité linéaire à la fois en temps et en mémoire par rapport au nombre de points et permettent donc l'application du transport et ses extensions dans le regime d'un très grands nombre de points. Dans ma première tentative vers cet objectif, je propose d'approcher les itérations de Sinkhorn résolvant le TO entropique en imposant une factorisation de faible rang spécifique du noyau associé, ce qui donne une factorisation de rang non négatif faible du couplage optimal. Ensuite, je propose de généraliser cette idée et de résoudre le problème du TO ainsi que le problème GW en imposant directement une contrainte de rang non négatif faible sur les couplages admissibles dans le problème d'optimisation du transport. Nous montrons que ces nouveaux schémas de régularisation ont de meilleures performances computationnelles et statistiques que l'approche entropique et qu'ils peuvent même atteindre une complexité linéaire sous des hypothèses de rang faible sur les matrices de coûts associés au problème de transport. Ces nouveaux schémas de calcul ouvrent la voie à l'utilisation du TO à grande échelle. Dans une deuxième partie, je présente deux contextes où l'idée fondamentale proposée par Kantorovich pour résoudre le problème de l'OT peut également être appliquée, offrant ainsi une nouvelle perspective sur des problèmes de ML de longue date. Plus précisément, nous proposons de relaxer le problème de division équitable entre plusieurs agents dans l'espace des distributions en permettant la division des masses de ressources dans leur repartition. Ce faisant, nous montrons qu'il est toujours possible d'obtenir une partition équitable des ressources et nous obtenons une généralisation du problème du TO lorsqu'il y a plusieurs coûts impliqués. Nous abordons également le problème des exemples adverses à l'aide du TO. Dans ce problème, l'attaquant est représenté sous forme d'une fonction déterministe qui projette la distribution des données vers une distribution adverse visant à maximiser le risque du classificateur. En relaxant la définition de l'attaquant pour qu'il soit non plus une fonction mais un couplage, nous obtenons une formulation variationnelle du risque adverse qui nous permet d'interpréter le problème de minimisation du risque adverse comme un jeu à somme nulle à deux joueurs et nous étudions la question de l'existence d'équilibres de Nash dans ce jeu.

**Title :** Advances in Optimal Transport: Low-Rank Structures and Applications in Machine Learning

**Keywords :** Optimal Transport, Optimization, Statistics, Machine Learning

**Abstract :** Optimal transport (OT) plays an increasingly important role in machine learning (ML) to compare probability distributions. The OT problem has been used in many applications, and stated with a wide variety of formulations. Among these the Monge ansatz and the Kantorovich linear program stand out. The former involves finding an efficient push-forward map that can morph a measure onto another, while the latter relax the matching of the measures by allowing the splitting of masses. Kantorovich OT is far more amenable to computations and has been the main focus in data sciences. Yet, it poses, in its original form, several challenges when used for applied problems: (i) computing OT between discrete distributions amounts to solving a large and expensive network flow problem which requires a supercubic complexity in the number of points; (ii) estimating OT using sampled measures is doomed by the curse of dimensionality. These issues can be mitigated using an entropic regularization, solved with the Sinkhorn algorithm, which improves on both statistical and computational aspects. While much faster, entropic OT still requires a quadratic complexity with respect to the number of points and therefore remains prohibitive for large-scale problems. Seizing this opportunity, I devoted a significant part of my thesis to work on scalable approaches to OT, which led to my line of work on the introduction of low-rank optimal transport (LOT). I also realized that the fundamental idea proposed by Kantorovich to relax OT could be applied in other settings, and I proposed new approaches using this very same idea to tackle the fair division problem and the adversarial attacks problem through the lens of OT. This thesis is therefore divided in two main parts. In the first part, I present new regularization approaches for the OT problem, as well as its quadratic extension, the Gromov-Wasserstein (GW) problem, by imposing low-rank structures on couplings. This yields a linear complexity both in time and memory with respect to the number of points. In my first attempt towards that goal, I proposed to approximate the iterations of the Sinkhorn algorithm solving entropic OT by forcing a specific low-rank factorization of the kernel involved, resulting in a low non-negative rank factorization of the optimal coupling. Then I propose to generalize this idea and to directly solve the OT problem as well as the GW problem under low non-negative rank constraints on the admissible couplings. We show that these new regularization schemes have better computational and statistical performances compared to the entropic approach and that they can even reach a linear complexity under low-rank assumptions on the ground cost matrices. These new computational schemes pave the way for the use of OT in the large-scale setting. In a second part, I present two settings where the fundamental idea proposed by Kantorovich to relax the OT problem can also be applied, offering new perspective on longstanding ML problems. More precisely, we propose to relax and lift the fair division problem between multiple agents into the space of distributions by allowing the splitting of resource masses in the partition. By doing so, we show that it is always possible to obtain a fair partition of the resources and we obtain a generalization of the OT problem when multiple costs are involved. We also tackle the problem of adversarial examples using OT. In this problem, the attacker can be represented as a deterministic map that push forward the data distribution towards an adversarial one that aims at maximizing the risk of the classifier. By relaxing the definition of the attacker to be a coupling, we obtain a variational formulation of the adversarial risk which allows us to interpret the adversarial risk minimization problem as a two-player zero-sum game and we study the question of the existence of Nash equilibria in this game.

**Institut Polytechnique de Paris**
91120 Palaiseau, France