Leveraging Regularization, Projections and Elliptical Distributions in Optimal Transport

Boris Muzellec

Abstract

The ability to compare and manipulate probability distributions is a crucial component of numerous machine learning (ML) algorithms. The statistics literature provides a rich class of divergence functions to measure the discrepancy between two probability distributions, such as the Kullback-Leibler (KL) divergence, the total variation (TV) distance, or more generally the family of f-divergences. Yet, these divergences rely on comparing density functions pointwise, and saturate or diverge when the supports of the probability measures are disjoint. This fact can be a major drawback in ML tasks involving discrete or high-dimensional measures, and calls for more geometry-aware discrepancies. Optimal transport (OT) has proven to be a well-suited alternative: starting from a cost function (e.g. a distance) on the space on which measures are supported, OT consists in finding a mapping or coupling (i.e. a joint law) 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 probability distributions, or point histograms, in the form of an optimization problem. The fact that OT highly depends on the geometry of the distributions' ground space makes it particularly well suited to numerous ML applications, notably those that consist in fitting a probability measure, such as generative modeling. Further, as a consequence of its strong geometric component, OT is the object of a rich mathematical theory regarding its metric and topological properties, on which ML practitioners can rely to build and study their models.

Despite those advantages, the applications of OT in data sciences have long been hindered by the mathematical and computational complexities of the underlying optimization problem. Indeed, computing OT between discrete distributions amounts to solving a large and expensive linear program, and results in quantities that are not differentiable, which is impractical for ML gradient-based algorithms. Worse still, in the general non-discrete case, there are no known efficient methods for estimating optimal transport for moderate to high dimensions – the existing methods relying on approximating PDEs, which is only tractable in low dimension. In particular, the approach that consists in sampling from distributions and estimating OT using sampled measures is doomed by the curse of dimensionality: the sample convergence rate of OT is exponentially slow w.r.t. the dimension of the ambient space.

To alleviate those issues, two main approaches have been considered. The first consists in regularizing the optimization problem in order to obtain more favorable properties, such as smoothness or strict convexity. Approximations of OT divergences can then be obtained at a much lower cost from those regularized problems. In particular, entropic regularization yields couplings and discrepancies that are smooth and differentiable, and that can be obtained efficiently using Sinkhorn's algorithm or stochastic optimization. Hence, it has become the prevailing choice of regularization. The second approach consists in keeping the optimization problem as such, focusing on particular cases that admit closed-form solutions or that can be efficiently solved. A notable example is given by the optimal transport problem for 1D distributions, which can be solved in closed form from the quantile functions of the distributions under mild assumptions on the ground cost. In particular, discrete one-dimensional OT has a much lower computational complexity, since it can be solved with a simple call to a sorting procedure. For this reason, variants of OT relying on 1D projections such as sliced Wasserstein (SW) distances have recently gained popularity in the ML community. Likewise, in the multidimensional setting, Gaussian measures and their elliptical generalizations are one of the very few instances for which OT is available in closed form. In this particular case, OT defines the so-called Bures-Wasserstein geometry, due to its links with the Riemannian Bures geometry on positive semi-definite (PSD) matrices.

Even though closed-form instances of OT have been leveraged in recent works, the guiding principle of this thesis is that there remains many research opportunities to develop new algorithmic tools that can leverage or extend such closed forms.

Our thesis builds extensively on the Bures-Wasserstein geometry, with the aim to use it as basic tool in data science applications. To do so, we consider settings in

which the Bures-Wasserstein geometry is alternatively employed as a basic tool for representation learning, enhanced using subspace projections, and smoothed further using entropic regularization. In a first contribution, the Bures-Wasserstein geometry is used to define embeddings as elliptical probability distributions. Our work extends on the classical representation of data as vectors, i.e. points in $\mathbb{R}^d$, to naturally encode a notion of spread or uncertainty. To train those embeddings, we propose numerical tools that leverage the underlying Riemannian structure of the Bures metric. In the second contribution, we propose a new approach that exploits "classical" (unregularized) OT, the Bures-Wasserstein geometry and projected OT. Indeed, we prove the existence of transportation maps and plans that extrapolate Monge maps restricted to lower-dimensional projections, and a characterization of such subspace-optimal plans. We then show that subspace-optimal plans admit closed forms in the case of Gaussian measures, that are linked to properties of the Bures metric. Our third contribution consists in deriving closed forms for entropic OT, as well as unbalanced entropic OT, between Gaussian measures scaled with a varying total mass. These expressions constitute the first non-trivial closed forms for entropic OT, providing the first continuous test case for the study of entropic OT and shedding some light on the mass transportation/creation trade-off in unbalanced OT. Finally, in a last contribution, entropic OT is leveraged to tackle missing data imputation in a non-parametric and distribution-preserving way. Although this imputation is performed according to a very intuitive criterion, we show in extensive experiments that our algorithms are competitive with state-of-the-art methods.

Résumé

Pouvoir manipuler et comparer des mesures de probabilité est essentiel pour de nombreuses applications en apprentissage automatique (*machine learning*). Il existe dans la littérature statistique une vaste classe de divergences permettant de mesurer la différence entre deux distributions, comprenant par exemple la divergence de Kullback-Leibler (KL), la distance de variation totale (VT), ou plus généralement la famille des f-divergences. Cependant, ces divergences reposent sur la comparaison point-à-point des fonctions de densité, et saturent ou divergent lorsque les supports des mesures sont disjoints. Ceci peut être un inconvénient majeur pour les applications d'apprentissage automatique qui nécessitent de comparer des mesures discrètes ou en haute dimension, et appelle à l'emploi de divergences reposant sur des liens plus forts avec la géométrie des espaces sous-jacents. Le transport optimal (TO) s'est avéré constituer une alternative adaptée : partant d'une fonction de coût (e.g. une distance) définie sur l'espace dans lequel les mesures sont supportées, le TO consiste à trouver une application ou un couplage (i.e. une loi jointe) entre les deux mesures qui soit optimal par rapport à ce coût. En d'autres termes, le TO est une extension naturelle de la fonction de coût de base en une divergence entre mesures de probabilité, ou entre histogrammes de points, sous la forme d'un problème d'optimisation. Du fait que le TO dépende fortement de la géométrie de l'espace de base des distributions, il est particulièrement bien adapté à de nombreuses applications en machine learning, notamment celles qui consistent à apprendre une mesure de probabilité, tel qu'en apprentissage génératif. De plus, en conséquence de son fort aspect géométrique, le transport optimal est l'objet d'une riche théorie mathématique concernant ses propriétés métriques et topologiques, sur laquelle la communauté de l'apprentissage automatique peut s'appuyer pour construire et étudier ses modèles.

En dépit de ces avantages, l'emploi du TO pour les sciences des données a longtemps été limité par les difficultés mathématiques et computationnelles liées au problème d'optimisation sous-jacent. En effet, calculer le transport optimal entre deux mesures discrètes revient à résoudre un coûteux programme linéaire de grande taille, et résulte en des quantités qui ne sont pas différentiables, ce qui est inadapté aux algorithmes de machine learning reposant sur la descente de gradient. Pire encore, dans le cas général non discret, il n'existe pas de méthode efficace pour estimer le TO dans des dimensions modérées ou élevées – les méthodes existantes s'appuyant sur l'approximation d'EDP, ce qui n'est praticable qu'en basse dimension. En particulier, l'approche qui consiste à échantillonner les distributions et à estimer le TO à partir des mesures empiriques résultantes souffre du fléau de la dimension : la vitesse de convergence rapportée au nombre d'échantillons est exponentiellement faible par rapport à la dimension de l'espace ambiant.

Pour contourner ces problèmes, deux approches ont été proposées. La première consiste à régulariser le problème d'optimisation afin de lui garantir de nouvelles propriétés, telles qu'une meilleure régularité ou encore la stricte convexité. Des approximations des divergences du TO peuvent ensuite être obtenues à partir de ces problèmes régularisés à un plus faible coût. Tout particulièrement, la régularisation entropique fournit des couplages et des divergences réguliers et différentiables qui peuvent être calculés efficacement à l'aide de l'algorithme de Sinkhorn, ou grâce à des méthodes d'optimisation stochastique. De ce fait, l'entropie est devenue le choix de régularisation le plus répandu. La seconde approche consiste quant à elle à conserver le problème d'optimisation dans sa forme initiale, en se concentrant sur des cas particuliers admettant des solutions en forme close ou pouvant se résoudre efficacement. Un exemple primordial est le cas du transport optimal en une dimension, qui peut être explicitement résolu à partir des fonctions quantile des distributions, sous des hypothèses modérées portant sur la fonction de coût utilisée. En particulier, le transport 1D entre distributions discrètes a une faible complexité puisqu'il peut être calculé à l'aide d'un algorithme de tri. Pour cette raison, des variantes du TO reposant sur des projections 1D telles que les distances Wasserstein "tranchées" (*sliced Wasserstein*) ont récemment gagné en popularité dans la communauté du ML. Dans le cas multi-dimensionnel, un second exemple est celui des mesures gaussiennes et de leurs généralisations elliptiques qui

constituent l'un des rares cas particuliers pour lesquels le TO admet une forme close. Dans ce second cas, le TO définit la géométrie de Bures-Wasserstein et possède de forts liens avec la géométrie riemannienne de Bures sur les matrices positives semi-definies (PSD).

Bien que certains travaux récents se soient appuyés sur des formes closes du transport optimal, le principe directeur de cette thèse est que de nombreuses pistes de recherche restent à explorer afin de développer de nouveaux outils algorithmiques permettant d'exploiter ou d'étendre de telles formes closes.

Cette thèse s'appuie tout particulièrement sur la géométrie du Bures-Wasserstein, dans le but de l'utiliser comme outil de base pour des applications en science des données. Pour ce faire, nous considérons des situations dans lesquelles la géométrie de Bures-Wasserstein est tantôt utilisée comme un outil pour l'apprentissage de représentations, étendue à partir de projections sur des sous-espaces, ou régularisée par un terme entropique. Dans une première contribution, la géométrie de Bures-Wasserstein est utilisée pour définir des plongements sous la forme de distributions elliptiques. Nos travaux étendent la représentation classique sous forme de vecteurs, i.e. de points dans $\mathbb{R}^d$, pour encoder de manière naturelle une notion d'étendue ou d'incertitude. Pour apprendre ces plongements, nous proposons de nouveaux outils numériques qui exploitent la structure riemannienne sous-jacente de la métrique de Bures. Dans une deuxième contribution, nous proposons une nouvelle approche qui exploite le transport optimal non régularisé "classique", la géométrie de Bures-Wasserstein et le TO projeté. Plus précisément, nous prouvons l'existence de fonctions et couplages de transport qui extrapolent des applications de Monge restreintes à des projections en faible dimension, et fournissons une caractérisation de ces plans de transport "sous-espace optimaux". Nous montrons que ces plans sous-espace optimaux admettent des formes closes dans le cas de mesures gaussiennes, liés à des propriétés de la métrique de Bures. La troisième contribution de cette thèse consiste à obtenir des formes closes pour le transport entropique ainsi que pour le transport entropique déséquilibré entre des mesures gaussiennes non-normalisées. Ces formes closes constituent les premières expressions non triviales pour le transport entropique, fournissent le premier exemple dans le cas continu pour l'étude du transport entropique et illustrent l'arbitrage entre transport et création de masse dans la transport déséquilibré. Finalement, dans une dernière contribution nous utilisons le transport entropique pour imputer des données manquantes de manière non-paramétrique et en préservant les distributions. Bien que cette imputation soit effectuée selon un critère très intuitif, nous montrons dans des expériences exhaustives que nos algorithmes sont compétitifs par rapport à l'état de l'art.

Contents

Outline and Contributions

Optimal transport (OT) is a two-century-old problem that has given birth to a rich mathematical theory and to numerous applications, that are both still being very actively developed to this date. OT was first formalized by Monge in his 1781 treatise. Motivated by his observation of workers moving earth to build fortifications, Monge raised the problem of optimally mapping two measures μ and ν of equal mass onto each other, according to a cost that is equal to the distance traveled by the workers per unit of mass. Due to its mathematical difficulty – most notably, the absence of guarantees regarding the existence of a solution – very limited progress was made on Monge's problem until the 1940s, when Kantorovich proposed a relaxation: instead of optimizing on one-to-one maps that push forward μ to ν, Kantorovich [1942] considers *couplings*, i.e. joint laws between μ and ν. This new formulation has allowed the OT theory to flourish, as Kantorovich's problem admits a solution under much less restrictive conditions than Monge's. In particular, it encompasses the case of discrete distributions, which can be interpreted as a resource allocation problem such as considered in [Tolstoi, 1930, Hitchcock, 1941]. This discrete version of Kantorovich's problem was numerically solved by Dantzig [1949], with further algorithmic refinements starting from the 1950s with the development of the linear programming literature [Dantzig, 1951] and min-cost flow problems [Ford and Fulkerson, 1962, Goldberg and Tarjan, 1989, Ahuja et al., 1993], closing a fecund phase in which OT became one of the foundational problems of mathematical programming.

OT's renaissance in mathematics. Starting from the late 1980s and succeeding to the preluding works of Rachev and Rüschendorf [see Rachev and Rüschendorf, 1998, and references therein], the mathematical aspects of OT were progressively better understood – including the challenging Monge problem. In his seminal paper, Brenier [1987] proved the existence of an optimal Monge map between measures that admit a density in the case of a quadratic ground cost, and characterized this map as the unique transportation map that is the gradient of a convex function. This fundamental result served as a building block for many theoretical works on Monge maps. In particular, it allowed to reformulate Monge's problem as the Monge-Ampère PDE, which Caffarelli [1991] used to prove regularity properties of the solutions in the quadratic case. McCann [1997] then introduced measure interpolants that now bear his name and which constitute the optimal transport geodesic between two measures according to the Wasserstein distance, defined by OT when the ground cost is a distance to a power $p \geq 1$. Observing that the space of measures endowed with the Wasserstein distances shares key properties with manifolds has paved the way to the seminal work of Jordan et al. [1998], who showed that the Fokker-Plank equation can be recast as a Wasserstein proximal minimization scheme – known as the JKO scheme – of a functional taking measures as arguments. This construction was perfected in [Ambrosio et al., 2006], where a gradient flow theory generalizing that of Euclidean spaces was built on the Wasserstein space. Further links with PDEs and fluid mechanics were developed in [Benamou and Brenier, 2000], defining the so-called *dynamic* formulation of OT. These works paved the way for decisive contributions by both Villani [2008] and Figalli et al.

[2010] whose respective works on the Ricci curvature and isoperimetric inequalities, among others, were recognized with Fields medals.

Optimal transport in data sciences. In parallel, in the early 2000s OT has begun to appear in more applied domains such image processing, computer vision and machine learning. Indeed, discrete OT was "rediscovered" in [Rubner et al., 2000] for image retrieval tasks under the name of the *earth mover's distance* (EMD). From then, it was put to application in image processing and computer graphics [Rabin et al., 2011, Bonneel et al., 2011, Haker et al., 2004], but its usage remained limited by its $O(n^3 \log(n))$ complexity despite specialized solvers [Pele and Werman, 2009]. This issue was alleviated by the addition of an entropic regularization term to Kantorovich's problem by Cuturi [2013]. Entropic regularization not only ensures the uniqueness of the solution by strict convexity, but also allows to solve the corresponding problem in $O(n^2)$ time using Sinkhorn's algorithm [Sinkhorn, 1964], and results in a differentiable divergence. Further, Solomon et al. [2015] showed that for some domains and cost functions resulting in a separable kernel (e.g. for measures on a 2D or 3D grid with a squared norm cost), fast convolution techniques could be used to bring down the complexity to $O(n^{1+1/D})$. In turn, these results opened the way to a more widespread use in data sciences and machine learning. In particular, Frogner et al. [2015] used entropic OT with relaxed marginal constraints as a loss function for multilabel classification, building on a contribution of Kusner et al. [2015] who proposed to compare documents by representing them as bags-of-words and using OT between word embeddings in $\mathbb{R}^d$. Remarkably, the renewed interest of the machine learning community for optimal transport has led to applications not necessarily relying on a regularized formulation, notably for domain adaptation [Courty et al., 2014, 2017], generative modeling [Arjovsky et al., 2017], and distributionally robust learning [Esfahani and Kuhn, 2018].

Modern OT challenges in machine learning. Yet, applications of OT to data sciences are still hindered by several issues. In particular, the unfavorable statistical properties of OT linked to its high sample complexity have been the object of much work lately. Weed and Bach [2017] proved a sharp bound that shows that estimating Wasserstein distances requires an exponential number of samples w.r.t. the intrinsic dimension of the set on which measures are supported. Entropic regularization was shown to not only alleviate computational issues, but also to yield better sample rates [Genevay et al., 2019]. Alternatively, further refinements on Weed and Bach's bound can be obtained by assuming that measures differ in a low-dimensional subspace [Niles-Weed and Rigollet, 2019]. In the unregularized setting, those results theoretically justify a recent trend that consists in using OT between low-dimensional projections of measures to define measure discrepancies [Rabin et al., 2011, Bonneel et al., 2015, Paty and Cuturi, 2019] that benefit from lower computational costs, and hopefully better sample complexity. More generally, leveraging closed forms of transportation maps and OT distances in particular cases is a promising approach to reduce the computational and sample complexity, even the more so as methods for OT between continuous measures are scarce. As an example, Flamary et al. [2019] proved favorable sample complexity bounds for linear transportation maps, which encompass (but do not restrict to) the case of Gaussian and elliptical distributions. The issue of the statistical and computational complexity of OT is one of the aspects that the OT community is currently addressing, but other promising directions regarding applications of OT are also being investigated. As an example, it appeared in several works that the marginals constraints of OT could be too restrictive for some applications [Schiebinger et al., 2019, Frogner et al., 2015], which has led to the development of unbalanced optimal transport [Chizat, 2017], where the constraints are replaced with penalties. Further, OT gradient flows were shown

to constitute a key tool in analyzing the behavior of over-parameterized models [Chizat and Bach, 2018, Chizat et al., 2020], which is a burning topic in ML.

Contributions of this thesis. This thesis, started in 2017, makes a few contributions towards helping optimal transport theory overcome some of its well-documented computational and statistical drawbacks, and gain applicability in machine learning.

(i) In an initial work [Muzellec and Cuturi, 2018], we leveraged the fact that OT admits a closed form between elliptical distributions (defining the so-called Bures-Wasserstein geometry), to propose a new tool to embed complex data: rather than embed words as vectors in $\mathbb{R}^d$ [Borg and Groenen, 2005, Maaten and Hinton, 2008], we proposed to represent them as elliptical probability measures. In particular, this representation allows to naturally encode the notion of uncertainty, which we showed to be of particular interest for natural language processing (NLP) tasks. Proposing these algorithms required investigating numerical methods to perform optimization using the Riemannian structure of the Bures metric on PSD matrices.

From this starting point, we further investigated the use of the Bures-Wasserstein geometry in conjunction with other approaches that were currently being considered in the ML community to obtain better complexity.

(iii) We studied the problem of extrapolating transportation plans from maps defined between the projections of measures on lower-dimensional subspaces [Muzellec and Cuturi, 2019]. We showed the existence of such plans and provided a theoretical characterization, from which we exhibited two particular instances that generalize the Knothe-Rosenblatt transport [Knothe, 1957, Rosenblatt, 1952], and proved that they admit closed forms between Gaussian measures that are linked to properties of the Bures metric.

(iv) We proposed a last contribution on the topic of OT between elliptical distributions in [Janati and Muzellec et al., 2020], in which we provided the first closed forms for entropic OT and unbalanced entropic OT between Gaussian measures. Remarkably, these are, to our knowledge, the first example of closed-form expressions for entropic OT, and they can now be used as a testbed for researchers wishing to investigate numerical algorithms for entropic OT (and more generally variants of Sinkhorn's algorithm). They also provide a case in which the mass transportation/destruction trade-off in unbalanced OT can be characterized exactly.

(v) Finally, the last contribution in this thesis focuses on an application of entropic OT to imputing missing data [Muzellec et al., 2020]. This work relies on the simple intuition that two random batches from the same dataset should have similar distributions. We turned this criterion into a loss function using Sinkhorn divergences, and proposed flexible methods that can alternatively fit a parametric imputation model, or perform imputation without any parametric assumption on the underlying data distribution.

We now turn to a more detailed presentation of the chapters constituting this thesis. For each chapter, we present related work, and sketch the contributions of this thesis.

Chapter 1: Optimal Transport Geometries

This chapter introduces the key concepts and results on optimal transport on which this thesis builds upon, and, as such, does not introduce original contributions. Because of our focus on ML applications, we state these results for measures supported on $\mathbb{R}^d$ that are

either discrete or absolutely continuous (a.c.). Three OT "geometries" are introduced: the original Monge-Kantorovich OT geometry, the Bures-Wasserstein geometry on elliptical distributions, and the geometry of entropy-regularized OT.

	Chapter 2	Chapter 3	Chapter 4	Chapter 5
Unregularized OT	✓	✓		
Bures-Wasserstein	✓	✓	✓	
Entropy-regularized OT			✓	✓

Table 1: Summary of the OT geometries used in the main chapters of this thesis.

Monge-Kantorovich Optimal Transport. This chapter starts with the presentation of the optimal transport problem, which was initially introduced in Monge's 1781 memoir. Monge studied the problem of optimally mapping masses of earth represented by measures μ and ν, according to the ground cost $c(x, y) = \|x - y\|$:

$$\inf_{T:T_\sharp\mu=\nu} \int_{\mathbb{R}^d} c(x, T(x))\mathrm{d}\mu(x), \tag{$\mathcal{M}$}$$

where we denote $T_\sharp\mu = \nu$ the fact that T pushes forward μ to ν, i.e. that $\nu(A) = \mu(T^{-1}(A))$ for all measurable sets A.

Because this problem is mathematically challenging (in particular, the existence of solutions is not guaranteed), Kantorovich introduced in 1942 the relaxed problem

$$\inf_{\gamma\in\Pi(\mu,\nu)} \iint_{\mathbb{R}^d\times\mathbb{R}^d} c(x, y)\mathrm{d}\gamma(x, y), \tag{$\mathcal{K}$}$$

where the transportation maps from ($\mathcal{M}$) are replaced with couplings $\gamma \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$, i.e. probability measures having μ and ν as marginals. In particular, when the ground cost is a distance to a power $p \geq 1$, i.e. $c(x, y) = d^p(x, y)$, ($\mathcal{K}$) defines the celebrated Wasserstein distances.

After introducing problems ($\mathcal{M}$) and ($\mathcal{K}$), a collection of results based on [Santambrogio, 2015] concerning the existence of solutions to ($\mathcal{M}$) and ($\mathcal{K}$) and their links is recalled. In particular, the celebrated Brenier theorem [Brenier, 1987] for the existence and characterization of Monge maps with a quadratic cost as the gradient of a convex function will play a central role in the case of elliptical measures introduced in Section 2.

To conclude this section, the computational aspects of OT are presented. Those aspects, which are crucial in a machine learning perspective, are discussed depending on the type of measures that are involved - discrete, or absolutely continuous (a.c.). In particular, the discrete case boils down to the linear program

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

where $\mathbf{a} \in \Delta_n, \mathbf{b} \in \Delta_m$ are probability weight vectors and $\mathbf{C} = [c(x_i, y_j)]_{i=1,\dots,n,j=1,\dots,m}$ is the ground cost matrix. (D-OT) can be solved using the network simplex algorithm with complexity $O(nm(n + m)\log(nm))$ [see Ahuja et al., 1993, Peyré et al., 2019]. This high computational cost can be mitigated in some particular cases and variants based on 1D transport. Indeed, in 1D the optimal transport map can be written as a monotone map involving the cumulative distribution functions F_ν, F_μ and their inverses, the quantile functions:

$$T : x \mapsto F_\mu^{[-1]} \circ F_\nu(x).$$

As a consequence, in the discrete setting 1D OT can be solved in $O(n \log n)$ time by sorting the supporting points of the distributions. Building on those properties, sliced OT [Rabin et al., 2011] is defined as the expectation of OT on random 1D projections, and Knothe-Rosenblatt couplings [Knothe, 1957, Rosenblatt, 1952] are built using a recursive 1D matching between conditional distributions.

	Chapter 2	Chapter 3	Chapter 4	Chapter 5
Discrete-Discrete				✓
Continuous-Continuous	✓	✓	✓	
1D & KR transport		✓		

Table 2: Summary of the OT settings used in the main chapters of this thesis.

The two following sections are dedicated to settings in which OT enjoys particularly favorable computational properties, namely OT for elliptical distributions, and entropic regularization of OT.

The Bures-Wasserstein Geometry. The case of OT between Gaussian measures with a quadratic cost is one of the very few settings in which Wasserstein distances and Monge maps are available in closed form. This fact was independently discovered in several seminal works [Dowson and Landau, 1982, Olkin and Pukelsheim, 1982, Givens et al., 1984]. This exception is essentially due to the fact that the Brenier theorem proves the existence of linear Monge maps, and serves as an essential tool to prove that the Wasserstein space of Gaussian measures defines a Riemannian manifold [Takatsu, 2011].

In fact, most of these properties can be extended to the more general class of *elliptical distributions*. Elliptically-contoured distributions can be seen as a generalization of Gaussian distributions, either defined as having characteristic functions of the form $e^{it^T\mathbf{c}}g(e^{t^T\mathbf{C}t})$ ($g = \exp(-\cdot/2)$ corresponding to Gaussian measures) as in [Cambanis et al., 1981], or using a less compact definition based on density functions with an elliptical symmetry as in [Gelbrich, 1990]. In his seminal paper, Gelbrich [1990] proves that the Wasserstein distances between distributions from the same elliptical family $\mu_{\mathbf{a},\mathbf{A}}, \mu_{\mathbf{b},\mathbf{B}}$ has the same expression as for Gaussian measures (involving mean vectors $\mathbf{a}, \mathbf{b}$ and covariance matrices $\mathbf{A}, \mathbf{B}$)

$$W_2^2(\mu_{\mathbf{a},\mathbf{A}}, \mu_{\mathbf{b},\mathbf{B}}) = \|\mathbf{a}-\mathbf{b}\|^2 + \mathfrak{B}^2(\mathbf{A},\mathbf{B}), \tag{1}$$

where $\mathfrak{B}^2(\mathbf{A},\mathbf{B}) \stackrel{\text{def}}{=} \text{Tr}\mathbf{A} + \text{Tr}\mathbf{B} - 2\text{Tr}\left(\mathbf{A}^{1/2}\mathbf{B}\mathbf{A}^{1/2}\right)^{1/2}$ is the Bures metric for PSD matrices [Bures, 1969, Bhatia et al., 2018], and so do Monge maps: $T_\sharp \mu_{\mathbf{a},\mathbf{A}} = \mu_{\mathbf{b},\mathbf{B}}$ with

$$T : x \to \mathbf{A}^{-\frac{1}{2}}\left(\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}}\right)^{\frac{1}{2}}\mathbf{A}^{-\frac{1}{2}}(x-\mathbf{a}) + \mathbf{b}. \tag{2}$$

The Bures metric is linked to a "maximum correlation" optimization problem [Olkin and Pukelsheim, 1982] that allows to prove the joint convexity of the Bures metric and a lower bound on the 2-Wasserstein distance between any two distributions with finite second moments [Dowson and Landau, 1982]. The Riemannian structure of the Bures metric on the PSD cone [Bhatia et al., 2018, Malagò et al., 2018] allows to derive Wasserstein geodesics for elliptical distributions, and to characterize Wasserstein barycentres from a fixed-point equation on PSD matrices [Agueh and Carlier, 2011, Bhatia et al., 2018], from which an algorithm converging to this barycenter can be obtained [Álvarez-Esteban et al., 2016].

Entropic Regularization of Optimal Transport. In the general case, closed forms for OT distances and couplings are not available. In the prevalent discrete setting, the computational costs associated with solving (D-OT) along with the fact that $\mathrm{OT}(\mu, \nu)$ is not differentiable can be prohibitive in many ML applications. Starting from [Cuturi, 2013], the prevailing approach to accommodate for these issues has been to add an entropic regularization term to the Kantorovich problem:

$$\mathrm{OT}_\varepsilon(\mu, \nu) \stackrel{\text{def}}{=} \inf_{\gamma \in \Pi(\mu,\nu)} \iint_{\mathbb{R}^d \times \mathbb{R}^d} c(x, y) \mathrm{d}\gamma(x, y) + \varepsilon \, \mathrm{KL}(\gamma \| \mu \otimes \nu). \tag{Ent-OT}$$

In the discrete setting, OT_ε defines a differentiable discrepancy, which can be efficiently computed using Sinkhorn's algorithm [Sinkhorn, 1964], at the price of no longer defining a positive divergence. This discrepancy can be turned into a positive definite divergence by subtracting debiasing terms to OT_ε. This defines the Sinkhorn divergence [Genevay et al., 2018]

$$S_\varepsilon(\mu, \nu) \stackrel{\text{def}}{=} \mathrm{OT}_\varepsilon(\mu, \nu) - \frac{1}{2}(\mathrm{OT}_\varepsilon(\mu, \nu) + \mathrm{OT}_\varepsilon(\mu, \nu)). \tag{3}$$

When the ground cost c is positive definite, S_ε is a differentiable, convex (but not jointly) positive definite divergence which metrizes weak star convergence and retains the favorable computational complexity of OT_ε [Feydy et al., 2019].

Alternative regularizations of optimal transport were considered in Blondel et al. [2018], allowing to obtain sparse but differentiable OT plans - at the price of Sinkhorn's algorithm no longer applying. Further, Chizat [2017] extends regularization of OT to the unbalanced OT problem, in which the constraints on coupling marginals are replaced with penalization terms.

In Chapter 4, we elaborate on entropy-regularized OT and (entropic) unbalanced OT, proving closed forms in the case of Gaussian measures.

Chapter 2: Generalizing Point Embeddings using the Wasserstein Space of Elliptical Distributions

This chapter is based on [Muzellec and Cuturi, 2018].

Learning mathematical representations that can be conveniently manipulated for complex objects is a challenging task with numerous applications in ML. While these representations have traditionally been in the form of vectors, i.e. points in $\mathbb{R}^k$, we propose in this work to extend these points to representations as elliptical probability measures, in the Bures-Wasserstein geometry.

Related work. There exists a vast literature on the problem of obtaining low-dimensional representations $y_1, y_2, ..., y_n \in \mathbb{R}^k$ of complex, high-dimensional objects $x_1, x_2, ..., x_n$ living in a space $\mathcal{X}$. When the objects to be represented are themselves vectors in $\mathbb{R}^d$, a prevalent method, often used as a pre-processing step, is *principal component analysis* (PCA) [Pearson, 1901]. More generally, when these objects are equipped with a distance $d_{\mathcal{X}}(x_i, x_j)$, embeddings are naturally sought so that the distances $\|y_i - y_j\|$ are as close as possible to $d_{ij} \stackrel{\text{def}}{=} d_{\mathcal{X}}(x_i, x_j)$. Closeness criteria include *distortion*[1] [Johnson and Lindenstrauss, 1984, Bourgain, 1985] or the *stress* $\left(\sum_{i \neq j} (d_{ij} - \|y_i - y_j\|)^2 / d_{ij}^2\right)^{1/2}$ as in metric multidimensional scaling [De Leeuw, 1977, Borg and Groenen, 2005]. Several

[1] An embedding has distortion α if there exists $r > 0$ such that $\forall i, j, r \leq \frac{d_{ij}}{\|y_i - y_j\|} \leq \alpha r$.

approches have then refined these methods, departing from the original goal of finding isometric embeddings to focus on notions of intrinsic dataset geometry [Tenenbaum et al., 2000, Roweis and Saul, 2000, Hinton and Roweis, 2003, Maaten and Hinton, 2008]. Finally, some tasks require to compute embeddings without the guidance of a ground distance or similarity measure. This is notably the case in NLP, where word embeddings are computed based on the co-occurence of similar words [Mikolov et al., 2013b, Pennington et al., 2014, Bojanowski et al., 2017], for lack of a natural distance between words.

More recently, two distinct trends have emerged. The first (i) learns representations in a latent space by minimizing reconstruction error [Hinton and Salakhutdinov, 2006, Kingma and Welling, 2014, Tolstikhin et al., 2018]. The second (ii) seeks embeddings into more "exotic" geometries, e.g. generalized MDS on the sphere [Maron et al., 2010], or in hyperbolic spaces [Nickel and Kiela, 2017].

As part of the second trend (ii), probabilistic embeddings were proposed by Vilnis and McCallum [2015]. This approach consists in representing objects as parametric probability distributions over $\mathbb{R}^d$, which extends the traditional representation in $\mathbb{R}^k$ as points that can be seen as Dirac distributions. Vilnis and McCallum propose to embed words as Gaussian measures in the geometry of the Kullback-Leibler divergence (KL), or of the expected likelihood (ℓ_2) kernel [Jebara et al., 2004]. However, these geometries cannot naturally extend point embeddings, as they saturate when measures are Diracs (to infinity or to a constant value). Moreover, due to numerical stability issues linked to the KL divergence between Gaussian measures, only Gaussian distributions with diagonal covariance matrices have been considered in [Vilnis and McCallum, 2015]. In a concurrent work, Singh et al. [2020] considered representing words as histograms over context words, based on pre-computed word embeddings such as `glove` [Pennington et al., 2014]. Subsequent work to ours considered embeddings in $\mathcal{P}(\mathbb{R}^d)$ in the form of empirical distributions with fixed support cardinality using entropy-regularized OT [Frogner et al., 2019]. Finally, let us mention that our use of OT metrics to learn embeddings was inspired by the theoretical results of Andoni et al. [2015], who showed that $\mathcal{P}(\mathbb{R}^3)$ equipped with the Wasserstein distances is snowflake-universal[2].

Contributions. The main contributions of this chapter concern the benefits of representing objects as elliptical distributions in the Bures-Wasserstein geometry, along with practical tools and guidelines for optimization within this geometry.

(i) **Representing objects as elliptical distributions in the Bures-Wasserstein geometry:** We propose to represent each object as an elliptical distribution $\mu_{\mathbf{a},\mathbf{A}}$ using a mean vectors $\mathbf{a}$ and a PSD covariance matrix $\mathbf{A}$, and endow these representations with the Bures-Wasserstein distance

$$W_2(\mu_{\mathbf{a},\mathbf{A}}, \mu_{\mathbf{b},\mathbf{B}}) = \|\mathbf{a}-\mathbf{b}\|^2 + \mathfrak{B}^2(\mathbf{A},\mathbf{B}).$$

This representation has several benefits compared with Gaussian measures in the KL or ℓ_2 geometry:

a. First, it seamlessly includes point embeddings as Dirac measures, which can alternatively be seen as degenerate elliptical distributions with a $\mathbf{0}$ covariance matrix. In particular, the Bures-Wasserstein distance between two degenerate Dirac elliptical distributions is simply the Euclidean distance between their means: $W_2(\mu_{\mathbf{a},\mathbf{0}}, \mu_{\mathbf{b},\mathbf{0}}) = \|\mathbf{a}-\mathbf{b}\|^2$;

[2] i.e., it embedds $d_{\mathcal{X}}^{\theta}, \theta \in (0,1)$ with arbitrarily low distortion.

b. Next, the proposed methods remain valid for any choice of representing elliptical family, and not only Gaussian measures. In particular, this allows to represent objects as uniform distributions over ellipsoids, which have compact supports and are therefore more amenable to visualization;

c. Finally, used with the numerical tools we propose, the Bures distance is numerically stable, which allows to use full covariance matrices (as opposed to diagonal covariance matrices is previous works). This allows to make a fuller use of the dimension d of the ambient space: with full covariance matrices, elliptical embeddings can use up to $d + d(d+1)/2$ scalar parameters, but diagonal elliptical embeddings are limited to $2d$.

(ii) **Numerical tools and methods for optimization with Bures distances**: We provide numerical tools to optimize models based on Bures distances with gradient-based methods. More precisely, we address two issues: (a.) computing and differentiating the Bures distance and (b.) ensuring that matrices remain PSD throughout gradient descent.

a. We leverage the fact that Newton-Schulz (NS) iterations [Higham, 2008] with a suitable initialization simultaneously yield Monge maps $\mathbf{T}^{\mathbf{AB}}$ and their inverses $\mathbf{T}^{\mathbf{BA}}$ to minimize the amount of NS runs required to compute and differentiate Bures distances. Our method relies on the following Bures identities:

$$\mathfrak{B}^2(\mathbf{A},\mathbf{B}) = \|\mathbf{A} - \mathbf{T}^{\mathbf{AB}}\mathbf{A}\|_F^2 \quad \text{and} \quad \nabla_{\mathbf{A}}\mathfrak{B}^2(\mathbf{A},\mathbf{B}) = \mathbf{I}_\mathrm{d} - \mathbf{T}^{\mathbf{AB}}.$$

By keeping the maps $\mathbf{T}^{\mathbf{AB}}$ in memory, this allows to compute gradients without recomputing any matrix roots or inverses. In comparison, automatic differentiation has a complexity equivalent to computing the distances again. An important practical point is that all proposed manipulations are easily parallelizable on GPUs.

b. We avoid any projection on the PSD cone by using a $\mathbf{A} = \mathbf{L}\mathbf{L}^T$ parameterization and optimizing on the $\mathbf{L}$ factor, which is free to take any value in $\mathbb{R}^{d\times d}$. Remarkably, we show that Euclidean gradient descent on the $\mathbf{L}$ factor,

$$\mathbf{L} \leftarrow \mathbf{L} - \eta\nabla_{\mathbf{L}}\tfrac{1}{2}\mathfrak{B}^2(\mathbf{L}\mathbf{L}^T,\mathbf{B}),$$

is equivalent to taking a step of size η along the geodesic from $\mathbf{A} = \mathbf{L}\mathbf{L}^T$ to $\mathbf{B}$:

$$\mathbf{C}_{\mathbf{A}\to\mathbf{B}}(\eta) = [(1-\eta)\mathbf{I}_\mathrm{d} + \eta\mathbf{T}^{\mathbf{AB}}]\mathbf{A}[(1-\eta)\mathbf{I}_\mathrm{d} + \eta\mathbf{T}^{\mathbf{AB}}].$$

In other words, a $\mathbf{A} = \mathbf{L}\mathbf{L}^T$ parameterization is projection-free and allows to emulate Riemannian optimization in the Bures geometry.

(iii) **Applications to similarity and hypernymy representation with word embeddings**: In large-scale experiments, we compute word embeddings from the `ukWac` and `WaCkypedia` corpora [Baroni et al., 2009] by minimizing the Bures-Wasserstein equivalent of the hinge loss [Vilnis and McCallum, 2015]:

$$\sum_{(w,c)\in\mathcal{R}} \left[M - [\mu_w : \nu_c] + \tfrac{1}{n} \sum_{c'\in N(w)} [\mu_w : \nu_{c'}] \right]_+ ,$$

where $\mathcal{R}$ is the set of word/context pairs co-occuring in a sliding window of a given size and $N(w)$ a random set of negative contexts for the word w, and

$$[\mu_{\mathbf{a},\mathbf{A}} : \mu_{\mathbf{b},\mathbf{B}}] \stackrel{\text{def}}{=} \langle \mathbf{a},\, \mathbf{b}\rangle + F(\mathbf{A},\mathbf{B}),$$

where $F(\mathbf{A}, \mathbf{B}) \stackrel{\text{def}}{=} \mathrm{Tr}(\mathbf{A}^{1/2}\mathbf{B}\mathbf{A}^{1/2})^{1/2}$ is the Bures fidelity (see Chapter 1). The resulting 250K embeddings, trained on a 3 billion token dataset, are competitive with then state-of-the-art skipgram embeddings [Mikolov et al., 2013b] and diagonal Gaussian embeddings [Vilnis and McCallum, 2015] on similarity and entailment benchmarks.

In a second experiment, we train embeddings on the `WordNet` dataset [Miller and Charles, 1991] to encode hypernymy[3] relations (constituting a DAG on nouns), consistently beating the then state-of-the-art Poincaré embeddings [Nickel and Kiela, 2017] in link prediction tasks.

Chapter 3: Building Optimal Transport Plans on Subspace Projections

This chapter is based on [Muzellec and Cuturi, 2019].

OT suffers from the curse of dimensionality. For this reason, discrepancies relying on OT between lower-dimensional projections of measures have recently been considered. In this chapter, we show how global transport maps and couplings can be extrapolated from a Monge map between projected measures.

Related Work. In $\mathbb{R}^d$, the Wasserstein distance between empirical measures over n samples converges at speed $O(n^{-1/d})$ to the distance between the original distributions [Dudley, 1969, Fournier and Guillin, 2015]. At best, this rate can be improved if the distribution is actually supported on a lower-dimensional surface [Weed and Bach, 2017] – in which case the dimension parameter in the rate can be replaced with this intrinsic dimension parameter – or can be turned to $O(n^{-2/d})$ under some additional hypothesis [Chizat et al., 2020]. This unfavorable sample complexity associated with a $O(n^3 \log n)$ computational complexity has led to approaches consisting in first projecting measures on lower-dimensional subspaces before computing OT between projected measures. Most notably, sliced Wasserstein (SW) distances [Rabin et al., 2011, Bonneel et al., 2015] average Wasserstein distances between 1D projections (see Section 1.2):

$$\mathrm{SW}_p^p(\mu, \nu) \stackrel{\text{def}}{=} \int_{S^d} W_p^p((p_\theta)_\sharp \mu, (p_\theta)_\sharp \nu) \mathrm{d}\theta,$$

where p_θ is the projection on the line of direction $\theta \in \mathbb{R}^d$. In the discrete setting, each projected distance (and coupling) can be obtained via sorting in $O(n \log n)$ time. These favorable runtimes, along with the fact that SW defines a metric between measures (although distinct from the Wasserstein metric), has led to a recent spark of interest for VAE and GAN applications [Deshpande et al., 2018, Wu et al., 2019]. Paty and Cuturi [2019] extend projections to subspaces of dimension $1 \leq k < d$ that are adversarially selected. Extrapolating transportation maps defined in few dimensions is linked to Knothe-Rosenblatt (KR) transport [Rosenblatt, 1952, Knothe, 1957], which defines a coupling between two measures by recursively extending 1D transport maps. Carlier et al. [2009] shows that KR transport can be obtained as the limit map with re-weighted quadratic costs, a result we extend to extrapolations of k-dimensional maps.

Contributions. The previously cited approaches that rely on subspace projections allow to define OT-based discrepancies, but do not provide transportation maps between the

[3] A is a hypernym of B if every B is a A, e.g. "mammal" is a hypernym of "dog".

original measures. In this chapter, we study how transportation maps and plans that coincide with a given map S defined on a linear subspace E (with projection operator p_E) can be obtained. That is, we are interested in transportation plans γ (resp. maps T) whose projections $\gamma_E = (p_E, p_E)_\sharp \gamma$ on that subspace E coincide with the optimal transportation plan $(\mathbf{I}_{\mathrm{d}E}, S)_\sharp \mu_E$ (resp. $p_E \circ T = S \circ p_E$).

(i) **Subspace-optimal plans and maps**: Given a Monge map S between two measures μ_E and ν_E projected on a linear subspace E of $\mathbb{R}^d$, we define global plans between the original measures μ and ν that coincide with S on E : $\Pi_E(\mu,\nu) \stackrel{\text{def}}{=} \{\gamma \in \Pi(\mu,\nu) : \gamma_E = (\mathbf{I}_{\mathrm{d}E}, S)_\sharp \mu_E\}$. We prove the existence of such *subspace-optimal* plans, and further characterize them using their disintegrations (i.e. their conditionals) on $E \times E$: denoting μ_{x_E} the disintegration of μ on $E^\perp \times \{x_E\}$, any plan $\gamma \in \Pi_E(\mu,\nu)$ is fully characterized by the conditional couplings on the graph of S between μ_{x_E} and $\nu_{S(x_E)}$ for $x_E \in E$, i.e. $\gamma_{(x_E, S(x_E))}, x_E \in E$.

(ii) **Monge-Independent plans and Monge-Knothe maps**: We focus on two particular instances of E-optimal plans. Monge-Independent (MI) plans are obtained by extending γ_E using independent couplings between μ_{x_E} and $\nu_{S(x_E)}$,

$$\pi_{\mathrm{MI}} \stackrel{\text{def}}{=} (\mu_{x_E} \otimes \nu_{S(x_E)}) \otimes (\mathbf{I}_{\mathrm{d}E}, S)_\sharp \mu_E,$$

and Monge-Knothe (MK) maps can be seen as a generalization of Knothe-Rosenblatt transport that extend γ_E using optimal couplings:

$$T_{\mathrm{MK}}(x_E, x_{E^\perp}) \stackrel{\text{def}}{=} (S(x_E), \hat{T}(x_E; x_{E^\perp})) \in E \oplus E^\perp,$$

where $\hat{T}(x_E; \cdot) : E^\perp \to E^\perp$ denotes the Monge map from μ_{x_E} to $\nu_{S(x_E)}$. Further, we prove the following properties for MI and MK transport:

(a) Discrete subspace-optimal transport converges to MI transport as sample size goes to infinity;

(b) MK transport is the subspace-optimal transport with the smallest transportation cost;

(c) Similarly to Knothe-Rosenblatt transport [Carlier et al., 2009], MK transport can be obtained as the limit transportation map with the re-weighted quadratic cost $c(x,y) = \sum_{i=1}^k (x_i - y_i)^2 + \varepsilon \sum_{j=1}^{d-k} (x_{j+k} - y_{j+k})^2$ when ε goes to 0;

(iii) **Closed forms for Gaussian measures**: Similarly to 2-Wasserstein distances and Monge maps, we prove that MI and MK transports admit closed-form expressions for Gaussian measures. More precisely, MI transport can be written as a degenerate Gaussian coupling, and MK transport as a block-triangular map. Incidentally, we give a closed form for the Knothe-Rosenblatt transport between Gaussian measures involving the Cholesky factors of the covariance matrices.

(iv) **Experiments on synthetic data, elliptical word embeddings, and for domain adaptation**: We show on synthetic data that MI and MK transports are more robust than classical transport in situations where the signal in distributions is concentrated on a lower-dimensional subspace. We show how MK transport can be used to distort the geometry of elliptical word embeddings in the case of polysemous words. Finally, we provide an algorithm for selecting a mediating subspace E when it is not prescribed, which we illustrate on a domain adaptation task with Gaussian mixture models.

Chapter 4: Entropic Optimal Transport between (Unbalanced) Gaussian Measures

This chapter is based on [Janati and Muzellec et al., 2020].

Entropic regularization has not only proved to be an efficient method to make OT more easily computable in a discrete setting, but also to alleviate the unfavorable sample complexity of OT [Genevay et al., 2019]. Yet, as of recently no closed-form solution for entropy-regularized OT between continuous distributions was known, in neither balanced nor unbalanced settings. This absence of closed-form formulas for a fixed regularization strength posed an important practical problem to evaluate the performance of stochastic algorithms that try to approximate regularized OT. The purpose of this chapter is to fill this gap, and provide closed-form expressions for balanced and unbalanced OT for Gaussian measures, which can then be used as test cases, or as a principled regularization of the Bures-Wasserstein distances.

Related Work. That Wasserstein distances and Monge maps have a closed form between Gaussian measures is a well-known fact [Dowson and Landau, 1982, Olkin and Pukelsheim, 1982, Givens et al., 1984, Bhatia et al., 2018], which has been extended to elliptical distributions from the same family [Gelbrich, 1990]. Yet, despite being widely used in practice, no similar results were known in the case of entropy-regularized OT [Cuturi et al., 2007, Peyré et al., 2019], with the exception of centered, one-dimensional normal distributions [Gerolin et al., 2020].

Shortly after the publication of our work [Janati and Muzellec et al., 2020], several works with partially overlapping contributions were made public [Mallasto et al., 2020, del Barrio and Loubes, 2020], with slightly different approaches. However, their results do not cover the unbalanced case.

Contributions. In this chapter, we present the first non-trivial closed forms for entropy-regularized OT between continuous measures:

(i) **A closed form for** (Ent-OT) **between Gaussian measures**: We show that the optimal entropic transportation plan between Gaussian measures is a Gaussian measure itself. This result is obtained by proving the convergence of Sinkhorn iterations, which lead to a fixed-point equation on symmetric matrices. We derive the solution of this fixed-point equation to obtain a closed form for entropic OT. This closed form is proven to remain well-defined, convex and differentiable even for singular covariance matrices, unlike the Bures metric (which loses differentiability). Finally, we derive its gradients and minimizers.

(ii) **Debiased Sinkhorn barycenters between Gaussian measures**: Using the definition of debiased Sinkhorn barycenters [Luise et al., 2019, Janati et al., 2020a], we show that the debiased entropic barycenter of Gaussian measures is Gaussian and that its covariance verifies a fixed-point equation similar to that of [Agueh and Carlier, 2011].

(iii) **A closed form for regularized unbalanced OT between Gaussian measures**: We provide a closed-form expression of the unbalanced transport plan between unnormalized Gaussian measures, with entropic regularization and KL marginal penalties. This transport plan is proven to be an unnormalized Gaussian measure itself. We provide a closed form for the cost of unbalanced OT as a function of the measure masses, and of the mass of the optimal plan (whose expression we provide).

The formula we obtain sheds some light on the link between mass destruction and the distance between the means in unbalanced OT.

Chapter 5: Missing Data Imputation using Optimal Transport

This chapter is based on [Muzellec et al., 2020].

Missing data is a fundamental issue in data sciences. Even with a moderate dimension and missing rate, ignoring data points with missing values quickly ceases to be a valid option [Zhu et al., 2019]. Hence, prior to performing downstream tasks (such as fitting a classification or regression model) it is often necessary to define a method to replace missing data with reasonable values. In this chapter, we describe an OT-based method to impute missing values that can rely or not on parametric assumptions on the underlying data distribution.

Related Work. The missing data problem is the object of a rich literature in the statistics community. The predominant nomenclature is that of Rubin [1976]: it distinguishes between three settings, namely missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) data. Most of the literature is devoted to methods for MCAR and MAR data, which are statistically ignorable, meaning they allow to impute without having to model the missingness mechanism itself [see Little and Rubin, 2002, van Buuren, 2018]. Imputation methods generally aim to preserve the distribution of the data, in order to limit the bias they introduce when performing downstream tasks. From a bird's eye view, imputation methods can be divided in two categories, depending on the type of assumptions on the data distribution they rely on:

1. **Methods relying on conditional models**: e.g. [van Buuren and Groothuis-Oudshoorn, 2011, MICE] which perform iterative regression, or iterative random forests [Stekhoven and Buhlmann, 2011]. These methods model conditional distributions by imputing variables one by one, in a round-robin fashion.

2. **Methods relying on joint models**: e.g. methods assuming a low-rank matrix model [Hastie et al., 2015, Josse et al., 2016], Gaussian joint models estimated via the EM algorithm [Dempster et al., 1977], or Bayesian joint models [Murray and Reiter, 2016]. More recently, deep learning (DL) models based on variational autoencoders [Kingma and Welling, 2014, VAE] such as [Mattei and Frellsen, 2019, MIWAE], [Ivanov et al., 2019, VAEAC] or generative adversarial networks [Goodfellow et al., 2014, GAN] such as [Yoon et al., 2018, GAIN] have emerged.

Contributions. In this chapter, we leverage OT to propose flexible missing value imputation methods that can operate either with our without parametric assumptions on the data distribution.

(i) **An OT-based imputation criterion**: Our methods stem from the simple observation that two randomly-sampled batches from the same dataset should have similar distributions. Using Sinkhorn divergences to measure the discrepancy between the distributions of two batches, we turn this criterion into a loss for missing data imputation:

$$\mathcal{L}_m(\mathbf{X}) \stackrel{\text{def}}{=} \sum_{\substack{K:0\leq k_1<\ldots<k_m\leq n \\ L:0\leq \ell_1<\ldots<\ell_m\leq n}} S_\varepsilon(\mu_m(\mathbf{X}_K), \mu_m(\mathbf{X}_L)), \tag{4}$$

where S_ε is the Sinkhorn divergence [Genevay et al., 2018], $\mathbf{X}_K$ denotes the batch constituted of points with indices in $K = \{k_1, k_2..., k_m\}$, and $\mu_m(\mathbf{X}_K) = \frac{1}{m}\sum_{i=1}^{m} \delta_{X_{k_i}}$ is the empirical measure supported on this batch. Minimizing this loss with respect to imputed values allows to perform missing value imputation in a distribution-preserving way.

(ii) **Sinkhorn-based imputation algorithms**: We design two imputation algorithms to minimize (4) that either rely (b.) or not (a.) on parametric models for the data distribution:

a. **Direct Sinkhorn imputation**: the first algorithm makes no parametric assumption on the data distribution. It minimizes (4) using gradient descent w.r.t. missing values directly. In other words, it optimizes as many parameters as there are missing values, without additional constraints. Hence, it can be applied to any dataset with quantitative variables without further assumptions;

b. **Sinkhorn round-robin imputation**: the second algorithm adapts the round-robin imputation scheme to the Sinkhorn batch loss (4). This method can be used to fit any differentiable parametric model, such as linear models or multi-layer perceptrons (MLP). A key advantage of this second method is that it allows to perform out-of-sample imputation once the model has been fitted, without running the training algorithm again.

(iii) **Large-scale experimental validation**: We show that our methods are competitive against baselines and state-of-the-art methods (including DL-based ones) on 23 UCI datasets. We consider MCAR, MAR and MNAR settings with different mechanisms, and a wide range of missing rates (10%, 30% and 50%).

Contributions de cette Thèse

Le transport optimal (TO) est un problème vieux de deux siècles qui a donné naissance à une riche théorie mathématique ainsi qu'à de nombreuses applications, encore activement développées à ce jour. Le TO fut initialement formalisé par Monge dans son traité de 1781. Motivé par l'observation de travaux de terrassement militaire, Monge s'interrogea quant à la manière optimale de transformer une mesure μ en une mesure ν de masse égale sous l'action d'une application, par rapport à un coût égal à la distance parcourue par les travailleurs par unité de masse. Du fait de sa difficulté mathématique – et tout particulièrement de l'absence de garanties quant à l'existence d'une solution – les progrès accomplis sur le problème de Monge furent très limités jusqu'aux années 1940, quand Kantorovich en proposa une version relâchée : au lieu d'optimiser par rapport à des applications point-à-point qui "poussent" μ sur ν, Kantorovich [1942] considéra des *couplages*, c'est à dire des lois jointes entre μ et ν. Cette nouvelle formulation a permis à la théorie du TO de s'épanouir, car le problème de Kantorovich admet une solution sous des hypothèses beaucoup moins restrictives que le problème de Monge. En particulier, il comprend le cas des distributions discrètes, qui peut être interprété comme un problème d'allocation de ressources tel que posé par [Tolstoi, 1930, Hitchcock, 1941]. La version discrète du problème de Kantorovich fut résolue numériquement par Dantzig [1949], avant de connaître des raffinemeents algorithmiques dans les années 1950 avec le développement de la programmation linéaire [Dantzig, 1951] et des problèmes de flots de coût minimum [Ford and Fulkerson, 1962, Goldberg and Tarjan, 1989, Ahuja et al., 1993], refermant une phase féconde durant laquelle le TO est devenu l'un des problèmes fondamentaux de la programmation mathématique.

La renaissance du transport optimal en mathématiques. À partir de la fin des années 1980 et succédant aux travaux de Rachev and Rüschendorf [voir Rachev and Rüschendorf, 1998, et références à l'intérieur], les aspects mathématiques du TO furent progressivement mieux compris – y compris ceux relevant du problème de Monge. Dans son article précurseur, Brenier [1987] prouva l'existence d'une application de Monge optimale entre mesures admettant une densité et dans le cas d'une fonction de coût quadratique, et caractérisa cette application comme l'unique transport étant le gradient d'une fonction convexe. Ce résultat fondamental a été un outil essentiel pour de nombreux travaux théoriques sur les applications de Monge. En particulier, il permit de reformuler le problème de Monge sous la forme de l'EDP de Monge-Ampère, sur laquelle Caffarelli [1991] s'appuya pour prouver des propriétés de régularité des solutions du cas quadratique. McCann [1997] introduisit ensuite les interpolations de mesures qui portent désormais son nom, et qui constituent la géodésique de transport optimal entre deux mesures selon la distance de Wasserstein, qui est définie par le TO dans le cas où le coût de base est une distance élevée à une puissance $p \geq 1$. En observant que l'espace des mesures doté de la distance de Wasserstein partage des propriétés clé avec les variétés, McCann a ouvert la voie aux travaux fondateurs de Jordan et al. [1998], qui montrèrent que l'équation de Fokker-Plank peut s'interprêter comme un schéma proximal en distance de Wasserstein – connu comme le schéma JKO – d'une fonctionnelle prenant des mesures en argument. Cette construction fut complétée par [Ambrosio et al., 2006], qui construisirent une théorie des flots de gradients en

distance de Wasserstein généralisant les flots euclidiens. Des liens ultérieurs avec les EDP et la mécanique des fluides furent développés par [Benamou, 2003], définissant la formulation dite *dynamique* du TO. Ces travaux ouvrirent la voie aux contributions essentielles de Villani [2008] et Figalli et al. [2010] dont les travaux respectifs sur la courbure de Ricci et les inégalités isopérimétriques, entre autres, furent récompensés par deux médailles Fields.

Transport optimal et sciences des données. Parallèlement, le TO apparut dès le début des années 2000 dans des domaines plus appliqués tels que le traitement d'images, la vision par ordinateur et l'apprentissage automatique. En effet, le transport discret fut "redécouvert" par [Rubner et al., 2000] pour des tâches d'extraction d'images sous le nom de "distance de terrassement" (*earth mover's distance*, EMD). Dès lors, il fut employé en traitement d'images et en programmation graphique [Rabin et al., 2011, Bonneel et al., 2011, Haker et al., 2004], mais ses usages demeurèrent limités par sa complexité en $O(n^3 \log(n))$ malgré des solveurs spécialisés [Pele and Werman, 2009]. Cette difficulté fut contournée par l'ajout d'un terme de régularisation entropique au problème de Kantorovich par Cuturi [2013]. En effet, la régularisation entropique permet non seulement d'assurer l'unicité de la solution par stricte convexité, mais permet aussi de résoudre le problème correspondant en complexité $O(n^2)$ à l'aide de l'algorithme de Sinkhorn [Sinkhorn, 1964], et produit une divergence différentiable. Qui plus est, Solomon et al. [2015] a montré que pour certains coûts et domaines correspondant à un noyau séparable (e.g. pour des mesures sur une grille 2D ou 3D avec un coût égal à une norme au carré), des techniques de convolution rapide pouvaient être employées pour réduire cette complexité à $O(n^{1+1/D})$. À leur tour, ces résultats ont ouvert la voie à un usage plus répandu du TO en science des données et en apprentissage automatique. En particulier, Frogner et al. [2015] emploie le transport entropique avec des contraintes marginales relâchées comme fonction de perte pour la classification multi-label, s'appuyant sur une contribution de Kusner et al. [2015] qui avait proposé de comparer des documents en les représentant comme des histogrammes de mots, en utilisant le TO entre plongements de mots dans $\mathbb{R}^d$. Remarquablement, cet intérêt renouvelé de la communauté du machine learning pour le transport optimal a mené à des applications qui ne s'appuient pas nécessairement sur une formulation régularisée, notamment en adaptation de domaine [Courty et al., 2014, 2017], en apprentissage génératif [Arjovsky et al., 2017] et pour l'apprentissage robuste au sens des distributions [Esfahani and Kuhn, 2018].

Challenges modernes du TO en apprentissage automatique. Malgré ces progrès, les applications du TO en sciences des données restent limitées par certaines difficultés. En particulier, les propriétés statistiques peu favorables du TO liées à sa complexité d'échantillonage élevée ont dernièrement fait l'objet de nombreux travaux. Weed and Bach [2017] ont prouvé une borne précise montrant qu'estimer la distance de Wasserstein requiert un nombre d'échantillons exponentiel en la dimension de l'ensemble sur lequel les mesures sont supportées. La régularisation entropique s'est avérée permettre non seulement de diminuer la complexité calculatoire du TO, mais aussi sa complexité d'échantillonage [Genevay et al., 2019]. De manière alternative, des raffinements par rapport à la borne de Weed and Bach peuvent être obtenus en supposant que les mesures diffèrent sur un sous-espace de faible dimension [Niles-Weed and Rigollet, 2019]. Dans le cadre non-régularisé, ces résultats justifient une tendance récente consistant à utiliser le TO entre des projections en basse dimension des mesures, pour définir des divergences entre distributions [Rabin et al., 2011, Bonneel et al., 2015, Paty and Cuturi, 2019] qui bénéficient de coûts de calcul plus faibles, et potentiellement d'une meilleure complexité d'échantillonage. Plus généralement, exploiter les cas particuliers pour lesquels les applications de transport et les distances de TO sont en forme close constitue une approche prometteuse pour réduire les complexités de calcul et d'échantillonage, d'autant plus que les méthodes permettant de résoudre le

TO entre mesures continues sont rares. Par exemple, Flamary et al. [2019] a prouvé que le cas des applications de transport linéaire, comprenant (mais ne se limitant pas à ces cas) les cas gaussien et elliptique, bénéficie de meilleures bornes de complexité statistique. Les difficultés liées aux complexités computationnelles et statistiques du TO sont un des aspects sur lesquels la communauté du transport optimal travaille actuellement, mais d'autres directions de recherche concernant les applications du TO sont aussi en cours d'exploration. Par exemple, il est apparu dans plusieurs travaux que les contraintes marginales du transport optimal pouvaient être trop restrictives pour certains usages [Schiebinger et al., 2019, Frogner et al., 2015], ce qui a donné lieu au développement du transport optimal déséquilibré [Chizat, 2017], où les contraintes sont remplacées par des pénalités. En outre, les flots de gradient de Wasserstein se sont avéré constituer un outil clé pour l'analyse des modèles sur-paramétrés [Chizat and Bach, 2018, Chizat et al., 2020], qui constituent un sujet de recherche de pointe en apprentissage automatique.

Contributions de cette thèse. Cette thèse, qui a débuté en 2017, propose quelques contributions dans le but d'aider le transport optimal à dépasser ses difficultés computationnelles et statistiques bien connues, et à gagner en employabilité pour l'apprentissage automatique.

(i) Dans un premier projet [Muzellec and Cuturi, 2018], l'expression en forme close du TO entre distributions elliptiques (qui définit la géométrie de Bures-Wasserstein) est exploité pour proposer un nouvel outil de plongement de données complexes: plutôt que de représenter les mots comme des vecteurs dans $\mathbb{R}^d$ [Borg and Groenen, 2005, Maaten and Hinton, 2008], nous proposons de les représenter à l'aide de mesures de probabilité elliptiques. En particulier, cette représentation permet d'encoder de manière naturelle la notion d'incertitude, que nous prouvons être bénéfique dans le cadre d'applications en traitement du langage naturel. Afin de concevoir ces algorithmes, nous avons développé des méthodes numériques d'optimisation qui tirent profit de la structure riemannienne de la métrique de Bures pour les matrices PSD.

Depuis ce point de départ, nous avons approfondi l'usage de la géométrie de Bures-Wasserstein en conjonction avec d'autres approches qui étaient alors explorées par la communauté de l'apprentissage automatique, pour obtenir une meilleur complexité.

(iii) Nous étudions l'extrapolation de plans de transport à partir d'applications définies entre les projections de mesures sur des sous-espaces de faible dimension [Muzellec and Cuturi, 2019]. Nous montrons l'existence de plans extrapolés et en fournissons une caractérisation théorique, à partir de laquelle nous exhibons deux instances particulières qui généralisent le transport de Knote-Rosenblatt [Knothe, 1957, Rosenblatt, 1952], et prouvons qu'elles admettent des formes closes pour les mesures gaussiennes, liées aux propriétés de la métrique de Bures.

(iv) Nous proposons une dernière contribution portant sur le TO entre distributions elliptiques dans [Janati and Muzellec et al., 2020], dans laquelle nous fournissons la première forme close pour le transport optimal entropique entre mesures gaussiennes. Remarquablement, ces expressions constituent à notre connaissance le premier exemple de forme closes pour le transport entropique, et pourront désormais être utilisées comme cas de test par les chercheurs qui conçoivent et étudient les algorithmes pour le TO entropique (et plus généralement les variantes de l'algorithme de Sinkhorn). Elles fournissent en outre un exemple dans lequel l'arbitrage entre création et transport de masse en transport optimal déséquilibré peut être caractérisé de manière exacte.

(v) Finalement, la dernière contribution de cette thèse porte sur une application du TO entropique à l'imputation de données manquantes [Muzellec et al., 2020]. Ce travail s'appuie sur le fait intuitif que deux batches aléatoires tirés du même jeu de données devraient avoir des distributions similaires. Partant, nous transformons ce critère en une fonction de perte utilisant la divergence de Sinkhorn, et proposons des méthodes d'imputation flexibles qui peuvent au choix servir à entraîner un modèle d'imputation paramétrique, ou à effectuer une imputation sans faire d'hypothèse paramétrique sur la distribution sous-jacente des données.

Notation

Ambiant Spaces.

$\mathcal{M}(\mathbb{R}^d)$: the set of positive measures over $\mathbb{R}^d$

$\mathcal{P}(\mathbb{R}^d)$: the set of probability measures over $\mathbb{R}^d$

$\mathcal{P}_p(\mathbb{R}^d)$: the set of probability measures over $\mathbb{R}^d$ with finite p first moments $\int_{\mathbb{R}^d} \|x\|^p \mathrm{d}\mu(x)$

$\mathcal{C}(\mathcal{X})$: the set of continuous real-valued functions on $\mathcal{X}$

$\mathcal{C}_b(\mathcal{X})$: the set of continuous, bounded real-valued functions on $\mathcal{X}$

Measures.

$T_\sharp \mu$: the pushforward measure of μ by T s.t. for all $A \subset \mathbb{R}^d$, $T_\sharp \mu(A) = \mu(T^{-1}(A))$

$\mu_n \rightharpoonup \mu$: $\mu_n \rightharpoonup \mu$ iff $\forall f \in \mathcal{C}_b(\mathbb{R}^d)$, $\int_{\mathbb{R}^d} f \mathrm{d}\mu_n \to \int_{\mathbb{R}^d} f \mathrm{d}\mu$ (weak convergence)

λ_V : The Lebesgue measure on V

Norms and Matrices.

$\mathcal{S}^d$: the set of symmetric square matrices in $\mathbb{R}^{d\times d}$

S_+^d : the set of symmetric positive semi-definite matrices in $\mathbb{R}^{d\times d}$

S_{++}^d : the set of symmetric positive definite matrices in $\mathbb{R}^{d\times d}$

$\mathbf{A} \geq \mathbf{B}$: $\mathbf{A} \geq \mathbf{B}$ (resp. $\mathbf{A} > \mathbf{B}$) iff $\mathbf{A} - \mathbf{B} \in S_+^d$ (resp. S_{++}^d) (Loewner partial order)

$\langle \mathbf{A}, \mathbf{B} \rangle$: $\langle \mathbf{A}, \mathbf{B} \rangle \stackrel{\text{def}}{=} \mathrm{Tr}\mathbf{A}^T\mathbf{B}$ (Frobenius inner product)

$\|\mathbf{A}\|$: $\|\mathbf{A}\| \stackrel{\text{def}}{=} (\mathrm{Tr}\mathbf{A}^T\mathbf{A})^{1/2}$ (Frobenius norm)

$\|\mathbf{A}\|_{\text{op}}$: $\|\mathbf{A}\|_{\text{op}} = \sup_{x\neq 0} \frac{\|\mathbf{A}x\|}{\|x\|}$ (operator norm, also equal to the leading singular value of $\mathbf{A}$)

$|\mathbf{A}|$: the determinant of $\mathbf{A}$ (also $\det \mathbf{A}$)

$\|x - y\|_{\mathbf{C}}$: $\|x - y\|_{\mathbf{C}}^2 \stackrel{\text{def}}{=} (x - y)^T\mathbf{C}(x - y)$ (Mahalanobis norm induced by $\mathbf{C}$)

$\mathbf{C}^\dagger$: the Moore-Penrose pseudo-inverse of $\mathbf{C}$ [Penrose, 1955]

Others.

$[\![1, n]\!]$: $[\![1, n]\!] \stackrel{\text{def}}{=} [1, n] \cap \mathbb{N}$

$\mathfrak{S}_n$: the set of permutations over $[\![1, n]\!]$

Chapter 1

Optimal Transport Geometries

In this chapter, we introduce the key results and concepts from the optimal transport (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 machine learning (ML) problems.

We start by presenting the original Monge formulation of OT and its Kantorovich relaxation in Section 1, with an emphasis on the case where the ground cost is a distance to a power, which defines the Wasserstein distances. The links between both formulations and their practical aspects are discussed, depending on whether discrete or continuous distributions are considered. The numerical challenges associated with OT will lead us to investigate particular cases or variants based on 1D OT that can be solved in closed form, or easily approximated.

In Section 2, we delve into the case of elliptical distributions, for which optimal transport has links with the Bures geometry on PSD matrices. Elliptical distributions can be defined as generalizations of Gaussian distributions, and correspond to one of the very few cases were transport maps and Wasserstein distances are available in closed form. This geometry will play a role in Chapters 2 to 4.

Finally, we present entropy-regularized OT (Ent-OT) in Section 3. Initially introduced as an approximation of OT that can be easily computed using Sinkhorn's algorithm, Ent-OT is now widely used in the ML community as it is smooth and differentiable. The recent introduction of Sinkhorn divergences, which inherit from the numerical advantages of Ent-OT and its differentiability while defining divergences for probability distributions in a rigourous sense, has reinforced the use of entropic regularization for data science. We conclude this section by mentioning alternative regularizations of OT, and the unbalanced OT (UOT) problem between measures with different total masses. Ent-OT is considered in Chapters 4 and 5, in a theoretical and applied perspective respectively.

	Chapter 2	Chapter 3	Chapter 4	Chapter 5
Unregularized OT	✓	✓		
Bures-Wasserstein	✓	✓	✓	
Entropy-regularized OT			✓	✓

Table 1.1: Summary of OT geometries used in the main chapters of this thesis.

1 Monge-Kantorovich Optimal Transport

Comparing and mapping distributions is an recurring task in machine learning, in both supervised and unsupervised settings. As will be shown in this chapter, the optimal transport theory provides robust criteria for quantifying differences between measures, and defining mappings between them. In this thesis, measures will be assumed to be supported on $\mathbb{R}^d$. While the full scope of optimal transport allows for much more generality, this thesis will essential consider two types of measures: (i) discrete measures and (ii) absolutely continuous (a.c.) measures, i.e. measures that admit a density w.r.t. the Lebesgue measure.

1.1 Monge and Kantorovich formulations

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

$$\inf_{T:T_\sharp\mu=\nu} \int_{\mathbb{R}^d} c(x, T(x))\mathrm{d}\mu(x), \tag{$\mathcal{M}$}$$

where $T_\sharp\mu$ is the *pushforward* measure of μ by T, defined by $T_\sharp\mu(A) = \mu(T^{-1}(A))$ for all μ-measurable sets A.[1] When it exists, an optimal map in $(\mathcal{M})$ is called a *Monge* map. Although it is intuitive, Monge's formulation is mathematically challenging: in particular, the existence of a Monge map is not guaranteed. As an example, consider the case where μ is a Dirac distribution. Then, $T_\sharp\mu$ is necessarily also a Dirac distribution, hence there can be no transport in Monge's sense if ν is not a Dirac distribution as well. This also highlights the intrinsic asymmetry of $(\mathcal{M})$, as conversely, it is always possible to find a Monge map going *to* a Dirac measure δ_y, by setting $\forall x, T(x) = y$.

Kantorovich formulation. To alleviate this issue, Kantorovich [1942] introduced a generalization of Monge's problem. Instead of considering maps, Kantorovich proposed to optimize over couplings, i.e. measures over the product space $\mathbb{R}^d \times \mathbb{R}^d$ that have μ and ν as marginals:

$$\inf_{\gamma\in\Pi(\mu,\nu)} \iint_{\mathbb{R}^d\times\mathbb{R}^d} c(x, y)\mathrm{d}\gamma(x, y), \tag{$\mathcal{K}$}$$

where $\Pi(\mu, \nu) \stackrel{\text{def}}{=} \{\gamma \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d) : \pi_{1\sharp}\gamma = \mu, \pi_{2\sharp}\gamma = \nu\}$ is the set of transportation plans, and $\pi_1 : (x, y) \mapsto x, \pi_2 : (x, y) \mapsto y$ are the canonical projections. A key advantage of this formulation is that a solution to $(\mathcal{K})$ exists under weak conditions on the cost function c.

Theorem 1.1 (Santambrogio [2015, Theorem 1.7])**.** *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 $(\mathcal{K})$ admits a solution.*

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 μ, ν have moments of order p, the Wasserstein distances can be defined from $(\mathcal{K})$.

Definition 1.2 (Wasserstein Distances)**.** *Let $p \geq 1$ and $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$. The p-Wasserstein distance is defined as*

$$W_p(\mu, \nu) \stackrel{\text{def}}{=} \inf_{\gamma\in\Pi(\mu,\nu)} \Big(\iint_{\mathbb{R}^d\times\mathbb{R}^d} d(x, y)^p \mathrm{d}\gamma(x, y)\Big)^{1/p}. \tag{1.1}$$

[1] Equivalently, if X is a random variable with law μ, then $T_\sharp\mu$ is the law of $T(X)$.

Wasserstein distances satisfy all three metric axioms on $\mathcal{P}_p(\mathbb{R}^d)$ [Santambrogio, 2015, Prop 5.1], and metrize weak convergence plus convergence of moments of order p [Santambrogio, 2015, Thm 5.11]. A sequence of measures μ_n converges weakly to a measure μ (denoted $\mu_n \rightharpoonup \mu$) *i.f.f.* for any continuous bounded function $f \in \mathcal{C}_b(\mathbb{R}^d)$, the integrals $\int_{\mathbb{R}^d} f \mathrm{d}\mu_n$ converge to $\int_{\mathbb{R}^d} f \mathrm{d}\mu$. 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.

Within the scope of this thesis, the ground distance will always be the Euclidean distance $d(x, y) = \|x - y\|_2$. In particular, the $p = 2$ case will play a crucial role, as the 2-Wasserstein distance satisfies particular properties (most of which are consequences of Brenier's theorem below). Therefore, unless stated otherwise, Wasserstein distances will designate the 2-Wasserstein distance W_2.

Bridging Monge and Kantorovich: the continuous setting. In light of the previous considerations, it is natural to ask under which conditions a Monge map might exist, and what links exist between Monge and Kantorovich formulations. For an absolutely continuous measure μ, Theorems 1.3 and 1.4 below show that under conditions on the cost function and/or compactness assumptions, the Kantorovich formulation $(\mathcal{K})$ generalizes Monge's $(\mathcal{M})$, in the sense that the solution to $(\mathcal{M})$ coincides with the solution of $(\mathcal{K})$ in the coupling formalism.

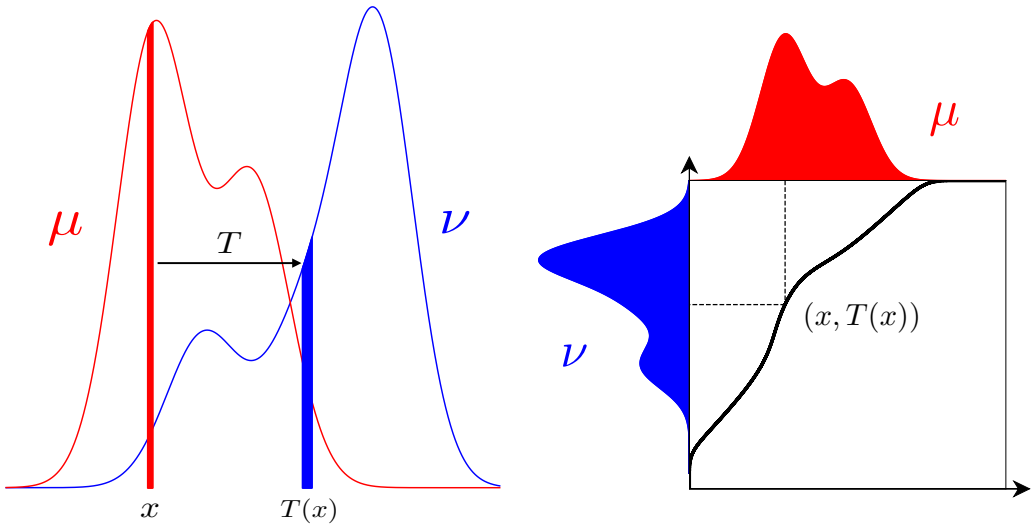


Figure 1.1: For a.c. measures, an optimal transport map (left) has an equivalent coupling supported on its graph (right).

Theorem 1.3 (Santambrogio [2015, Theorem 1.17.]). *Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ be compactly supported, and such that μ 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 γ, and T and γ are related by $\gamma = (id, T)_\sharp \mu$.*

Hence, under the conditions of Theorem 1.3, an optimal Monge map exists and can equivalently be described as an optimal transportation plan supported on its graph (Figure 1.1). In particular, for a.c. and compactly supported μ and ν, Theorem 1.3 holds when $c(x, y) = \|x - y\|^p$ with $p > 1$ as is the case for the Wasserstein distances (Definition 1.2, excluding the $p = 1$ case). The $p = 2$ case holds a particular place in the optimal transport theory, as shown by Brenier in his seminal paper [Brenier, 1987].

Theorem 1.4 (Brenier [1987]). *Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ such that μ is a.c., and $c(x,y) = \|x-y\|^2$. Then, problem $(\mathcal{M})$ admits a unique solution, which is characterized (among all transport maps) as being the gradient of a convex function ϕ: $\forall x \in \mathbb{R}^d, T^\star(x) = \nabla\phi(x)$.*

Note that contrary to Theorem 1.3, Theorem 1.4 no longer requires compact supports. Compared to Theorem 1.3, the major contribution of Theorem 1.4 is the unique characterization of the transport map as the gradient of a convex function. It will play a key role in Section 2. As an example, it implies the following immediate corollary.

Corollary 1.5 (Theorem 1.4). *Let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ be a.c., $c(x,y) = \|x-y\|^2$ and $\phi : \mathbb{R}^d \to \mathbb{R}$ a convex function. Then, $\nabla\phi : \mathbb{R}^d \to \mathbb{R}^d$ is the unique optimal Monge map from μ to $\nabla\phi_\sharp\mu$.*

Theorems 1.3 and 1.4 also imply that for compactly supported a.c. measures, or when $p = 2$, Wasserstein distances can also be formulated from a Monge point of view:

$$W_p(\mu,\nu) = \inf_{T:T_\sharp\mu=\nu} \left(\int_{\mathbb{R}^d} \|x - T(x)\|^p \mathrm{d}\mu(x) \right)^{1/p}. \tag{1.2}$$

Monge and Kantorovich: the discrete setting. When μ is a discrete distribution of the form $\sum_{i=1}^n a_i \delta_{x_i}$ with $\mathbf{a} \in \Delta_n$ and $\forall i \in [\![1,n]\!], x_i \in \mathbb{R}^d$, the existence of a Monge map occurs in few specific cases, the most notable being when μ and ν are discrete distributions with uniform weights and equal number of points.

Proposition 1.6. *Let $\mu = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}, \nu = \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$ with $n \in \mathbb{N}^\star$ and $\forall i \in [\![1,n]\!], x_i, y_i \in \mathbb{R}^d$. Then there exists a (not necessarily unique) Monge map from μ to ν. It takes the form of a permutation $\sigma \in \mathfrak{S}_n$ mapping each x_i to $y_{\sigma(i)}$, and has an equivalent optimal Kantorovich plan $\gamma^\star = \frac{1}{n}\sum_{i=1}^n \delta_{(x_i, y_{\sigma(i)})}$.*

Proposition 1.6 is a consequence of the Birkhoff–von Neumann theorem [Birkhoff, 1946, Von Neumann, 1953] and is sometimes referred to as the optimal matching problem.

Remark 1.7. *It is sometimes said for short that whenever a transport map exists, $(\mathcal{M})$ and $(\mathcal{K})$ coincide. This is a false statement: as a counter-example, consider as in Figure 1.2 two measures consisting of two Diracs each with weights $(1/4, 3/4)$, and a $\|\cdot\|^2$ ground cost. Although a transport map exists (mapping between points with equal weights), by varying the positions of the Diracs it can be made arbitrarily sub-optimal compared to the optimal coupling. See [Santambrogio, 2015, §1.4] for other counter-examples.*

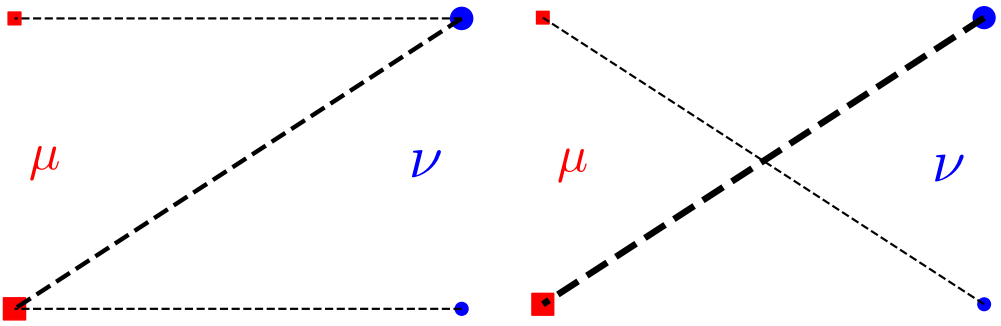


Figure 1.2: Optimal coupling (left) v.s. matching (right): although there exists a unique one-to-one mass-preserving matching, it is clearly sub-optimal compared to the Kantorovich plan.

In this thesis, both Monge and Kantorovich formulations will be used: Chapter 2 is based on the Monge point of view and Chapter 5 on the Kantorovich version, while Chapter 3 makes heavy use of the interplay between both formulations.

1.2 Computational aspects

Optimal transport quantities, such as the Wasserstein distances, are intrinsically defined through optimization problems. Hence, the computational aspects of solving these optimization problems are key in determining whether OT can be of practical use in different machine learning settings. In this section, the numerical aspects of OT are presented according to the different possible settings, and easily computable OT variants are introduced.

The one-dimensional setting. In the one-dimensional setting, optimal transport can be computed from the cumulative distribution functions (CDF) $F_\mu(x) \overset{\text{def}}{=} \int_{-\infty}^{x} \mathrm{d}\mu$ and their generalized inverses $F_\mu^{[-1]}(x) \overset{\text{def}}{=} \inf\{t \in \mathbb{R} : F_\mu(t) \geq x\}$ (also called quantile functions). Therefore, whenever those functions are easily computable, so is OT. This fact is the building block of the Knothe-Rosenblatt transport and sliced Wasserstein distances, which are introduced later. The following result describes one-dimensional OT in the continuous setting.

Proposition 1.8 (Santambrogio [2015, Theorem 2.9])**.** *Let $\mu, \nu \in \mathcal{P}(\mathbb{R})$ such that μ is a.c. and $c(x, y) = h(x - y)$ where $h : \mathbb{R} \to \mathbb{R}_+$ is a convex (resp. strictly convex) function. Then, there exists a (resp. exists a unique) Monge map from μ to ν. This map is monotone, and can be written as*

$$T : x \mapsto F_\mu^{[-1]} \circ F_\nu(x).$$

Moreover, the value of the objectives of problems $(\mathcal{M})$ and $(\mathcal{K})$ is given by

$$\int_0^1 h(F_\nu^{[-1]}(x) - F_\mu^{[-1]}(x))\mathrm{d}x.$$

In particular, one-dimensional Wasserstein distances are equal to the L_p distance between quantile functions:

$$\begin{aligned} W_p^p(\mu, \nu) &= \int_0^1 |F_\nu^{[-1]}(x) - F_\mu^{[-1]}(x)|^p \mathrm{d}x \\ &\triangleq \|F_\nu^{[-1]} - F_\mu^{[-1]}(x)\|_{L_p}^p. \end{aligned}$$

A similar result holds when μ and ν are discrete measures.

Proposition 1.9. *Let $\mu = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ and $\nu = \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$ with $x_1 \leq x_2 \leq ... \leq x_n$ and $y_1 \leq y_2 \leq ... \leq y_n$ and c, h as in Proposition 1.8. Then, there exists an optimal transport map given by*

$$\forall i \in [\![1, n]\!], T(x_i) = y_i,$$

and its corresponding transport cost is $\frac{1}{n}\sum_{i=1}^n h(y_i - x_i)$.

In particular, if μ, ν have sorted support points as in Proposition 1.9, it holds that $W_p^p(\mu, \nu) = |x_i - y_i|^p$. This implies that in the discrete and uniform setting, optimal transport and Wasserstein distances can be obtained by sorting supporting points, in $O(n \log n)$ time. If μ and ν are discrete but with non-uniform weights or a different number of points in their supports, it is still possible to compute an optimal transport plan that relies on sorting, as illustrated in Figure 1.3.

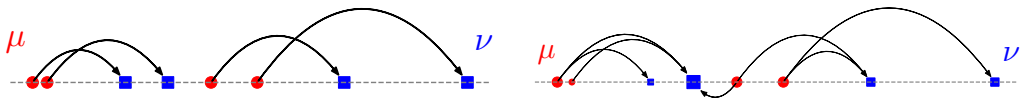


Figure 1.3: One-dimensional discrete transport. *Left:* uniform weights, *right:* non-uniform weights.

This introductory one-dimensional example already hints that the computational challenges induced by OT are quite different depending on whether the distributions μ and ν are discrete, or continuous[2]. This yields three broad settings, which are now introduced.

Discrete-discrete transport. When both distributions are discrete and can be written as $\mu = \sum_{i=1}^{n} a_i \delta_{x_i}$ and $\nu = \sum_{i=j}^{m} b_j \delta_{y_j}$ with $\mathbf{a} \in \Delta_n, \mathbf{b} \in \Delta_m$, $(\mathcal{K})$ is equivalent to the following linear program:

$$\min_{\mathbf{P} \in U(\mathbf{a},\mathbf{b})} \langle \mathbf{P}, \mathbf{C} \rangle, \tag{D-OT}$$

with $U(\mathbf{a}, \mathbf{b}) \stackrel{\text{def}}{=} \{\mathbf{P} \in \mathbb{R}_+^{n \times m} : \mathbf{P}\mathbb{1}_m = \mathbf{a}, \mathbf{P}^T \mathbb{1}_n = \mathbf{b}\}$ and $\mathbf{C} = [c(x_i, y_j)]_{1 \leq i \leq n, 1 \leq j \leq m}$. (D-OT) can be algorithmically solved using the network simplex algorithm, in $O(n + m)nm \log(n + m))$ time [see Ahuja et al., 1993]. Hence, although it is tractable, discrete optimal transport can be computationally expensive, and has the additional inconvenient of not being differentiable w.r.t. $\mathbf{a}$ and $\mathbf{b}$ due to the non-uniqueness of an optimal plan $\mathbf{P}^\star$. However, discrete OT plans are sparse, which is a valuable property in matching-based applications such as domain adaptation [Courty et al., 2014]. This sparsity comes from the fact that there always exists an optimal plan lying on a vertex of $U(\mathbf{a}, \mathbf{b})$: such a plan has at most $n + m$ nonzero entries.

Discrete-continuous transport. The case where μ is discrete and ν a.c. (often referred to as the *semi-discrete* setting) is already more challenging. It can be solved using quasi-Newton solvers relying on the computation of Laguerre cells and making piecewise constant approximations of the density [Mérigot, 2011] in a low-dimensional setting, or approximated using stochastic optimization [Genevay et al., 2016] (which requires however being able to sample from ν).

Continuous-continuous transport. When both μ and ν are a.c., closed forms or scalable methods for optimal transport are scarce. Thanks to the Brenier theorem (Theorem 1.4), the case of a quadratic cost $\|\cdot - \cdot\|^2$ enjoys additional properties that make it tractable in some cases. A first consequence of Theorem 1.4 is that $(\mathcal{M})$ with a quadratic cost is equivalent to the Monge-Ampère equation. Indeed, let p (resp. q) denote the density function of μ (resp. ν). Then, $(\mathcal{M})$ is equivalent to finding a convex function f such that

$$|\nabla f|^2 = \frac{p}{q \circ \nabla f}. \tag{1.3}$$

Secondly, optimal transport maps and 2-Wasserstein distances are available in closed form for the class of *elliptical* distributions, which is the subject of Section 2.

[2]Of course, in all generality a probability measure need not be either discrete or continuous. More complex settings could also be considered, but fall out of the scope of this thesis.

1.3 Variants of optimal transport.

As the above considerations show, solving optimal transport can be very computationally challenging, even so in the discrete-discrete setting if the sample size is large. A notable exception is 1D transport, which can be conveniently solved through sorting or when knowledge of quantile functions is available (see Section 1.2). This exception has motivated variants of optimal transport that enjoy favorable computational properties.

Sliced Wasserstein Distances. Rabin et al. [2011] propose to average the Wasserstein distance between projections on sampled one-dimensional directions, which defines the Sliced Wasserstein (SW) distances:

$$\mathrm{SW}_p^p(\mu,\nu) \overset{\text{def}}{=} \int_{S^d} W_p^p(p_{\theta\sharp}\mu, p_{\theta\sharp}\nu)\mathrm{d}\theta, \tag{SW}$$

where $\forall x \in \mathbb{R}^d, p_\theta(x) = \langle x, \theta\rangle$. Like Wasserstein distances, sliced Wasserstein distances satisfy all three metric axioms. However, Wasserstein and sliced Wasserstein distances are not equal. In practice, SW distances are estimated by averaging the projected Wasserstein distances along a fixed number of random directions, using Proposition 1.9. Moreover, SW distances are differentiable (even in the discrete setting) [Bonneel et al., 2015]. For instance,

$$\partial_{x_i}\mathrm{SW}_2^2\left(\tfrac{1}{n}\sum_{i=1}^n \delta_{x_i}, \tfrac{1}{n}\sum_{j=1}^n \delta_{y_j}\right) = \frac{2}{n}\int_{S^d}(\langle x_i, \theta\rangle - \langle y_{\sigma_\theta(i)}, \theta\rangle)\theta\mathrm{d}\theta, \tag{1.4}$$

where σ_θ is the permutation corresponding to the optimal map on the direction $\theta \in S^d$ (see Proposition 1.9).

The convenience of SW distances has lead to a recent interest in the ML community, in the GAN/VAE literature in particular [Deshpande et al., 2018, Wu et al., 2019]. Note however that even though SW provide a cheap way of comparing distributions, they have no associated pushforward mapping. They can however be associated to a coupling that corresponds to the average of the 1D couplings:

$$\mathrm{d}\gamma_{\mathrm{SW}}(x,y) \overset{\text{def}}{=} \int_{S^d} \mathrm{d}\gamma_\theta(x,y)\mathrm{d}\theta.$$

Knothe-Rosenblatt (KR) transport. In independent works, Knothe [1957] and Rosenblatt [1952] proposed a method for defining a transport map between two a.c. measures. It consists in a recursive scheme, relying on 1D monotone transport maps between conditional distributions. More precisely, let $f(x_1, x_2, ..., x_d)$ and $g(y_1, y_2, ..., y_d)$ denote the density functions of two a.c. probability measures $\mu, \nu \in \mathbb{R}^d$. Let f_1 and g_1 denote the marginal density functions of μ, ν on the first coordinate. Then, there exists a monotone map T_1 (as in Section 1.2), mapping f_1 to g_1. The broad idea is to then map the marginals on the first two directions, $f_2(x_1, x_2)$ and $g_2(y_1, y_2)$, in a way that conserves the transport of the first marginal. This implies in particular that (x_1, x_2) can only be mapped to a point of the form $(T_1(x_1), y_2)$: in other words, the conditional density $f_{x_1}(x_2)$ of f_2 given x_1 must be mapped to $g_{T(x_1)}(y_2)$ of g_2 given $T(x_1)$. Again, there exists a monotone optimal map $T_2(x_1, \cdot) : x_2 \mapsto T_2(x_1, x_2)$ mapping f_{x_1} to $g_{T(x_1)}$ optimally, and one can see that $(x_1, x_2) \mapsto (T_1(x_1), T_2(x_1, x_2)$ maps f_2 to g_2 (although this time not optimally). Applying this method recursively yields a map of the form

$$T_{\mathrm{KR}}(x_1, x_2, ..., x_d) = (T_1(x_1), T_2(x_1, x_2), ..., T_d(x_1, x_2, ...x_d)),$$

which verifies $T_{\mathrm{KR}\sharp}\mu = \nu$, and is monotone for the lexicographic order. Although this map T_{KR} is not in general an optimal map, it defines an accessible instance of a transport map. A more precise presentation of KR transport is given in Santambrogio [2015, Chapter 2]. Chapter 3 introduces generalizations of the KR transport, with closed forms for Gaussian distributions.

	Chapter 2	Chapter 3	Chapter 4	Chapter 5
Discrete-Discrete				✓
Continuous-Continuous	✓	✓	✓	
1D & KR transport		✓		

Table 1.2: Summary of the OT settings used in the main chapters of this thesis.

2 The Bures-Wasserstein Geometry

Out of the different settings presented in Section 1.2, the continuous-continuous one seems to be the most numerically challenging, as the only general methods available rely on approximating the solutions of PDEs [Benamou and Brenier, 2000], which does not scale well with the dimension of the ambient space. A noticeable exception are elliptical distributions, which can be seen as generalizations of Gaussian distributions or multivariate generalizations of location-scale families, for which closed-form solutions exist.

Unless stated otherwise, we consider the Frobenius inner product and norm on matrices in the following.

2.1 Elliptical distributions

Several concurrent definitions of elliptical distributions (also known as elliptically-contoured distributions) coexist. The most intuitive definition would be to see elliptical distributions as distributions on $\mathbb{R}^d$ having a density function that has elliptical level sets, i.e. density functions of the form $x \mapsto f(\|x - \mathbf{c}\|^2_{\mathbf{C}^{-1}})/\sqrt{|\mathbf{C}|}$, where $\mathbf{c} \in \mathbb{R}^d$ is the mean (or location) parameter, $\mathbf{C} \in S^d_{++}$ is the scale parameter, and $f : \mathbb{R}^d \to \mathbb{R}$ satisfies $\int_{\mathbb{R}^d} f(\|x\|^2)\mathrm{d}x = 1$. As an example, Gaussian distributions correspond to $f \propto \exp(-\cdot/2)$. However, this definition lacks in generality as it requires $\mathbf{C}$ to be invertible and therefore does not encompass degenerate distributions, supported on lower-dimensional subspaces. To address this issue, Gelbrich [1990] proposed a more general definition, which is stated here in a simplified version.

Definition 1.10 (Elliptical Distributions, [Gelbrich, 1990])**.** *Let $\mathbf{c} \in \mathbb{R}^d, \mathbf{C} \in S^d_+$. Let $\lambda_{\mathrm{Im}\mathbf{C}}$ denote the Lebesgue measure over the image of $\mathbf{C}$. An elliptical distribution with mean $\mathbf{c}$ and scale parameter $\mathbf{C}$ is a probability measure of the form*

$$\mathrm{d}\mu_{g,\mathbf{c},\mathbf{C}}(x) = g(\|x - \mathbf{c}\|^2_{\mathbf{C}^\dagger})\mathrm{d}\lambda_{\mathrm{Im}\mathbf{C}}(x), \tag{1.5}$$

where $g : \mathbb{R}^d \to \mathbb{R}_+$ satisfies $\int_{\mathrm{Im}\mathbf{C}} g(\|x\|^2_{\mathbf{C}^\dagger})\mathrm{d}\lambda_{\mathrm{Im}\mathbf{C}}(x) = 1$ and $\mathbf{C}^\dagger$ is the Moore-Penrose pseudo-inverse of $\mathbf{C}$.

A predating definition, less intuitive but more compact, relies on the characteristic function of a random vector: $\phi_X \overset{\text{def}}{=} t \in \mathbb{R}^d \mapsto \mathbb{E}_X[e^{it^T X}]$. Recalling that the characteristic function of a centered multivariate Gaussian random vector is $e^{it^T\mathbf{c}}g^{t^T\mathbf{C}t}$ with $g = \exp(-\cdot/2)$, the intuition behind this definition is to allow the function g to be picked in a broader class.

Definition 1.11 (Elliptical Distributions, [Cambanis et al., 1981])**.** *A random vector X is* elliptically-contoured *if there exist $\mathbf{c} \in \mathbb{R}^d$, $\mathbf{C} \in S_+^d$ and a function $g : \mathbb{R}_+ \to \mathbb{R}$ such that its characteristic function is of the form $\phi_X(t) = e^{it^T\mathbf{c}} g(e^{t^T\mathbf{C}t})$.*

Both Definitions 1.10 and 1.11 generalize the initial intuition by encompassing degenerate measures. Hence, an elliptical distribution is fully characterized by its mean $\mathbf{c} \in \mathbb{R}^d$, scale parameter $\mathbf{C} \in S_+^d$, and generating function g (defined in either Definition 1.10's or Definition 1.11's formalism). When two elliptical distributions share the same generating function g, they are said to belong to the same *family* of elliptical distributions. Examples of families include (multivariate) Gaussian distributions, (multivariate) t-distributions, or uniform distributions supported on ellipsoids.

Remark 1.12. *From the analogy with Gaussian measures, one could expect the covariance matrix $\Sigma_{g,\mathbf{C}}$ of an elliptically-contoured random vector to be equal to its scale parameter $\mathbf{C}$. It is in fact equal to $\tau_g\mathbf{C}$, where $\tau_g > 0$ depends on the generating function g only. In the setting of Definition 1.11, one has $\tau_g = -2g'(0)$ [Cambanis et al., 1981, Theorem 4.]. τ_g can also be written in Definition 1.10's setting, but has a less compact formulation [Gelbrich, 1990, Equation (14)]. As examples, in Gelbrich's formalism $g(x) \propto \exp(-x/2)$ yields $\tau_g = 1$ and corresponds to Gaussian measures, whereas $g(x) \propto \mathbb{1}_{x\leq 1}$ yields $\tau_g = \frac{1}{d+2}$ and corresponds to d-dimensional ellipsoids of radius 1 endowed with a uniform measure.*

2.2 The Bures-Wasserstein distance

In independent seminal works, Dowson and Landau [1982], Olkin and Pukelsheim [1982] and Givens et al. [1984] showed that the 2-Wasserstein distance between multivariate Gaussian distributions admits a closed form, known as the *Bures-Wasserstein* distance or also the *Fréchet* distance. Although not stated in those terms, their results also provide a closed form for the Monge map between two Gaussian measures. All three proofs rely on a version of Lemma 1.19, which expresses the maximal possible covariance between two random vectors. Gelbrich [1990] then extended these results to any two (potentially degenerate) elliptical distributions from the same family.

Theorem 1.13 (Gelbrich [1990])**.** *Let $g : \mathbb{R}_+ \to \mathbb{R}$ as in Definition 1.10. Then for any two members of the same elliptical family, the 2-Wasserstein distance has a closed form: $\forall \mathbf{a}, \mathbf{b} \in \mathbb{R}^d, \forall \mathbf{A}, \mathbf{B} \in S_+^d$,*

$$W_2^2(\mu_{g,\mathbf{a},\mathbf{A}}, \mu_{g,\mathbf{b},\mathbf{B}}) = \|\mathbf{a} - \mathbf{b}\|^2 + \mathfrak{B}^2(\Sigma_{g,\mathbf{A}}, \Sigma_{g,\mathbf{B}}), \tag{1.6}$$

where $\Sigma_{g,\mathbf{A}} = \tau_g\mathbf{A}$ (Remark 1.12), and

$$\mathfrak{B}^2(\mathbf{A}, \mathbf{B}) \stackrel{\text{def}}{=} \mathrm{Tr}\mathbf{A} + \mathrm{Tr}\mathbf{B} - 2\mathrm{Tr}(\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}})^{\frac{1}{2}} \tag{1.7}$$

is the Bures *[Bures, 1969, Bhatia et al., 2018] metric on the cone of PSD matrices.*

By homogeneity of the Bures metric and following Remark 1.12, the Wasserstein-Bures distance can alternatively be formulated in terms of scale parmeters as

$$W_2^2(\mu_{g,\mathbf{a},\mathbf{A}}, \mu_{g,\mathbf{b},\mathbf{B}}) = \|\mathbf{a} - \mathbf{b}\|^2 + \tau_g\mathfrak{B}^2(\mathbf{A}, \mathbf{B}).$$

Remark 1.14 (Particular cases)**.** *When $\mathbf{A}$ and $\mathbf{B}$ commute,* (1.7) *further simplifies to the Frobenius distance between matrix roots: $\mathfrak{B}^2(\mathbf{A}, \mathbf{B}) = \|\mathbf{A}^{1/2} - \mathbf{B}^{1/2}\|_F^2$. When they are both diagonal matrices, this quantity is called the Hellinger distance. When the covariance matrices go to 0 (and distributions converge to Dirac distributions), one recovers the L_2 distance between the means.*

Proposition 1.15. *Let $\mu_{g,\mathbf{a},\mathbf{A}}$ and $\mu_{g,\mathbf{b},\mathbf{B}}$ be two elliptical distributions from the same family, such that* $\mathrm{Im}\mathbf{B} \subset \mathrm{Im}\mathbf{A}$. *Then, the map $T : x \mapsto \mathbf{T}^{\mathbf{AB}}(x - \mathbf{a}) + \mathbf{b}$ with*

$$\mathbf{T}^{\mathbf{AB}} \stackrel{\text{def}}{=} \mathbf{A}^{\frac{\dagger}{2}}(\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}})^{\frac{1}{2}}\mathbf{A}^{\frac{\dagger}{2}} \tag{1.8}$$

is the optimal Monge map from $\mu_{g,\mathbf{a},\mathbf{A}}$ to $\mu_{g,\mathbf{b},\mathbf{B}}$, where $\mathbf{A}^{\frac{\dagger}{2}}$ is the Moore-Penrose pseudo-inverse of $\mathbf{A}^{\frac{1}{2}}$.

Proof. This is a direct consequence of Brenier's theorem [Brenier, 1987] and of Lemma 1.20 below. □

Note that contrary to the Bures distance, the Monge map (1.8) is scale invariant and can interchangeably be formulated using scale parameters or covariance matrices. In the remainder of this thesis, the dagger notation will be dropped and $\mathbf{A}^{-1}$ will denote the inverse of $\mathbf{A}$ when it exists, and its pseudo-inverse otherwise.

Remark 1.16 (Matrix square roots)**.** *For symmetric positive semi-definite (PSD) matrices, square roots can be defined using the eigenvalue decomposition: if $\mathbf{A} = \mathbf{P}\,\mathrm{diag}(\lambda_1, ..., \lambda_d)\mathbf{P}^T$, then $\mathbf{A}^{1/2} \stackrel{\text{def}}{=} \mathbf{P}\,\mathrm{diag}(\sqrt{\lambda_1}, ..., \lambda_d)\mathbf{P}^T$. In this thesis, square roots of non-symmetric matrices $\mathbf{A}$ with no eigenvalues in $\mathbb{R}_-$ will sometimes be considered: in that case, they will always be the unique square root of $\mathbf{A}$ with all eigenvalues in $\mathbb{R}_+$ [Higham, 2008, Theorem 1.29].*

Remark 1.17. *The* $\mathrm{Im}\mathbf{B} \subset \mathrm{Im}\mathbf{A}$ *assumption is required in Proposition 1.15 but not in Theorem 1.13. Indeed, consider two cases. First, if* $\mathrm{rk}\mathbf{B} > \mathrm{rk}\mathbf{A}$*, then no transport map going from $\mu_{g,\mathbf{a},\mathbf{A}}$ to $\mu_{g,\mathbf{b},\mathbf{B}}$ exists, since it is informally impossible to create mass in more dimensions than covered by $\mu_{g,\mathbf{a},\mathbf{A}}$ through the action of a map. Secondly, when* $\mathrm{rk}\mathbf{B} = \mathrm{rk}\mathbf{A}$ *but* $\mathrm{Im}\mathbf{B} \not\subset \mathrm{Im}\mathbf{A}$*, transport maps exist but take other forms, as* $\mathrm{Im}\mathbf{T}^{\mathbf{AB}} \subset \mathrm{Im}\mathbf{A}$*. However, Theorem 1.13 remarkably remains valid in either case, corresponding to the cost of the optimal coupling.*

The Bures-Wasserstein distance corresponds to the equality case of a lower bound on the Wasserstein distance, as originally proven by Dowson and Landau [1982].

Proposition 1.18 (Dowson and Landau [1982])**.** *Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ be two centered probability measures with covariance matrices $\mathbf{A}, \mathbf{B} \in S_+^d$. Then,*

$$\mathfrak{B}^2(\mathbf{A}, \mathbf{B}) \leq \mathbb{E}_{\substack{X\sim\mu \\ Y\sim\nu}}\|X - Y\|^2 \leq \mathrm{Tr}(\mathbf{A} + \mathbf{B} + 2(\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}})^{\frac{1}{2}}). \tag{1.9}$$

An important fact is that Proposition 1.18 is not restricted to elliptical distributions, but is applicable to any pair of probability measures with finite second order moments. In particular, it implies that if μ (resp. ν) has mean vector $\mathbf{a}$ (resp. $\mathbf{b}$) and covariance matrix $\mathbf{A}$ (resp. $\mathbf{B}$), then

$$\|\mathbf{a} - \mathbf{b}\|^2 + \mathfrak{B}^2(\mathbf{A}, \mathbf{B}) \leq W_2^2(\mathbf{A}, \mathbf{B}).$$

Hence, even for distributions that are not elliptically-contoured, the Bures-Wasserstein distance is a quantity of interest, as it provides a lower bound on the transport cost. Under the lens of Lemma 1.19, this lower bound can be seen as the cost of optimally matching the first two moments of μ and ν. The RHS of (1.9) gives information on the worst possible coupling (in a quadratic cost sense) between two distributions. However, from an optimal transport perspective it is not so informative, as the Wasserstein distance can always be bounded from above by the cost of the indepedent coupling $\mu \otimes \nu$:

$$W_2^2(\mathbf{A}, \mathbf{B}) \leq \|\mathbf{a} - \mathbf{b}\|^2 + \mathrm{Tr}\mathbf{A} + \mathrm{Tr}\mathbf{B}.$$

2.2.1 The Bures distance on PSD matrices

The definition of the Bures distance originates from quantum information theory, where it is used to measure the distance between two *states* represented by PSD density matrices with trace 1 [Bures, 1969, Bengtsson and Życzkowski, 2017]. In the context of quantum information theory, the quantity $F(\mathbf{A}, \mathbf{B})$ such that $\mathfrak{B}^2(\mathbf{A}, \mathbf{B}) = \mathrm{Tr}\mathbf{A} + \mathrm{Tr}\mathbf{B} - 2F(\mathbf{A}, \mathbf{B})$,

$$F(\mathbf{A}, \mathbf{B}) \stackrel{\text{def}}{=} \mathrm{Tr}(\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}})^{\frac{1}{2}} = \mathrm{Tr}(\mathbf{A}\mathbf{B})^{\frac{1}{2}},$$

is called the *fidelity* between states $\mathbf{A}$ and $\mathbf{B}$. In an optimal transport perspective, the fidelity represents the maximal attainable covariance.

Lemma 1.19 (Olkin and Pukelsheim [1982], Bhatia et al. [2018])**.** *Let $\mathbf{A}, \mathbf{B} \in S^d_{++}$. Then,*

$$F(\mathbf{A}, \mathbf{B}) = \max_{\mathbf{C}:\left(\begin{smallmatrix}\mathbf{A} & \mathbf{C}\\ \mathbf{C}^T & \mathbf{B}\end{smallmatrix}\right)\succeq 0} \mathrm{Tr}\mathbf{C}, \tag{1.10}$$

and the maximum is attained at $\mathbf{C} = \mathbf{A}\mathbf{T}^{\mathbf{AB}} = (\mathbf{A}\mathbf{B})^{1/2}$.

An alternative characterization of the Monge map $\mathbf{T}^{\mathbf{AB}}$ is provided by the following lemma.

Lemma 1.20. *Let $\mathbf{A}, \mathbf{B} \in S^d_{++}$ (resp. $\mathbf{A}, \mathbf{B} \in S^d_{+}$ s.t. $\mathrm{Im}\mathbf{B} \subset \mathrm{Im}\mathbf{A}$). Then*

$$\begin{aligned}\mathbf{T}^{\mathbf{AB}} &= \mathbf{A}^{-\frac{1}{2}}(\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}})^{\frac{1}{2}}\mathbf{A}^{-\frac{1}{2}}\\ &= \mathbf{B}^{\frac{1}{2}}(\mathbf{B}^{\frac{1}{2}}\mathbf{A}\mathbf{B}^{\frac{1}{2}})^{-\frac{1}{2}}\mathbf{B}^{\frac{1}{2}}\end{aligned}$$

is the unique symmetric positive (resp. semi-)definite solution of $\mathbf{T}\mathbf{A}\mathbf{T}^T = \mathbf{B}$.

Proof. One can check that $\mathbf{T}^{\mathbf{AB}}$ satisfies $\mathbf{T}^{\mathbf{AB}}\mathbf{A}\mathbf{T}^{\mathbf{AB}} = \mathbf{B}$ in either formulation. The uniqueness can be proven from the existence of a unique symmetric positive definite root of $\mathbf{A}^{1/2}\mathbf{B}\mathbf{A}^{1/2}$ or $\left(\mathbf{B}^{1/2}\mathbf{A}\mathbf{B}^{1/2}\right)^{-1}$, and incidentally proves that both expressions of $\mathbf{T}^{\mathbf{AB}}$ are indeed equal. □

Lemma 1.20 proves that $\mathbf{T}^{\mathbf{AB}}$ is the Monge map from $\mathcal{N}(0, \mathbf{A})$ to $\mathcal{N}(0, \mathbf{B})$ with a quadratic cost. Indeed, let $\mathbf{T}$ be a linear map, and X a random vector with covariance matrix $\mathbf{A}$. Then, $\mathbf{T}X$ has covariance matrix $\mathbf{T}\mathbf{A}\mathbf{T}^T$. Hence, Lemma 1.20 shows that $\mathbf{T}^{\mathbf{AB}}$ is a transport map from $\mathcal{N}(0, \mathbf{A})$ to $\mathcal{N}(0, \mathbf{B})$. According to Theorem 1.4, it is a Monge map *i.f.f.* it is the gradient of a convex function, which is the case as $\mathbf{T}^{\mathbf{AB}}$ is symmetric and positive definite.

Further, Lemma 1.19 provides a direct proof of the LHS inequality of Proposition 1.18: indeed, if μ and ν are centered measures with covariance matrices $\mathbf{A}, \mathbf{B}$ and $\gamma \in \Pi(\mu, \nu,)$, then

$$\mathbb{E}_{(X,Y)\sim\gamma}\|X - Y\|^2 = \mathrm{Tr}\mathbf{A} + \mathrm{Tr}\mathbf{B} - 2\mathrm{Tr}\,\mathrm{Cov}_{(X,Y)\sim\gamma}(X, Y),$$

and $\mathrm{Tr}\,\mathrm{Cov}_{(X,Y)\sim\gamma}(X, Y)$ can be bounded from above using Lemma 1.19. Further, Lemma 1.19 shows that $\mathbb{E}_{(X,Y)\sim\gamma}\|X - Y\|^2 = \mathfrak{B}^2(\mathbf{A}, \mathbf{B})$ if and only if the covariance matrix of γ is $\left(\begin{smallmatrix}\mathbf{A} & \mathbf{A}\mathbf{T}^{\mathbf{AB}}\\ \mathbf{T}^{\mathbf{AB}}\mathbf{A} & \mathbf{B}\end{smallmatrix}\right)$, i.e. i.f.f. γ is the law of $(X, \mathbf{T}^{\mathbf{AB}}X)$. Hence, given that if $\mathrm{Cov}(X) = \mathbf{A}$ then $\mathrm{Cov}(\mathbf{T}X) = \mathbf{T}\mathbf{A}\mathbf{T}^T$, Lemmas 1.19 and 1.20 yield another interpretation of the Bures-Wasserstein distance: it is the minimal quadratic transportation cost associated with matching the first two moments of two measures μ and ν through the action of a map. Under this perspective, it is thus natural that the Bures-Wasserstein distance coincides with the Wasserstein distance for elliptical distributions of a given family, which are uniquely characterized by their means and covariances.

Another consequence of Lemma 1.19 is the joint convexity of the (squared) Bures distance, which can be obtained from writing problem (1.10) in a dual formulation.

Proposition 1.21. *The squared Bures distance* $\mathfrak{B}^2(\mathbf{A}, \mathbf{B})$ *is jointly convex in* $\mathbf{A}$ *and* $\mathbf{B}$.

Proof. Problem (1.10) can be equivalently rewritten using the variable $\mathbf{X} = \left(\begin{smallmatrix} \mathbf{X}_1 & \mathbf{X}_2 \\ \mathbf{X}_3 & \mathbf{X}_4 \end{smallmatrix}\right)$ as

$$F(\mathbf{A}, \mathbf{B}) = \frac{1}{2} \max_{\substack{\mathbf{X} \geq 0 \\ \mathbf{X}_1 = \mathbf{A}, \mathbf{X}_4 = \mathbf{B}}} \langle \mathbf{X}, \left(\begin{smallmatrix} 0 & \mathbf{I}_\mathrm{d} \\ \mathbf{I}_\mathrm{d} & 0 \end{smallmatrix}\right) \rangle.$$

Strong duality holds, and this problem can be shown to admit the following dual formulation:

$$F(\mathbf{A}, \mathbf{B}) = \frac{1}{2} \min_{\left(\begin{smallmatrix} \mathbf{F} & -\mathbf{I}_\mathrm{d} \\ -\mathbf{I}_\mathrm{d} & \mathbf{G} \end{smallmatrix}\right) \succeq 0} \langle \mathbf{F}, \mathbf{A} \rangle + \langle \mathbf{G}, \mathbf{B} \rangle. \tag{1.11}$$

Hence, the fidelity $F(\mathbf{A}, \mathbf{B})$ can be written as the pointwise infimum of linear functionals in $(\mathbf{A}, \mathbf{B})$. Therefore, it is jointly concave in $\mathbf{A}$ and $\mathbf{B}$, which makes $\mathfrak{B}^2(\mathbf{A}, \mathbf{B})$ jointly convex.

□

Proposition 1.22. *Let* $\mathbf{A}, \mathbf{B} \in S^d_{++}$. *Then* $\nabla_{\mathbf{A}} \mathfrak{B}^2(\mathbf{A}, \mathbf{B}) = \mathbf{I}_\mathrm{d} - \mathbf{T}^{\mathbf{AB}}$.

Proof. This can be proven by direct calculus, as in Section 5. Alternatively, we can use problem (1.11) to obtain $\nabla_{\mathbf{A}} F(\mathbf{A}, \mathbf{B}) = \frac{1}{2}\mathbf{F}^\star = \frac{1}{2}\mathbf{T}^{\mathbf{AB}}$ [Bhatia et al., 2018]. □

Riemannian structure. The Bures distance is actually a Riemmanian[3] metric on PSD matrices. From this fact it can be shown that the Wasserstein space of Gaussian measures is itself a Riemmanian manifold [Takatsu, 2011].

Proposition 1.23 (Bhatia et al. [2018], Malagò et al. [2018])**.** *The Bures distance defines a Riemannian metric over the cone of PSD matrices, with associated metric tensor*

$$G_{\mathbf{A}}(\mathbf{U}, \mathbf{V}) = \mathrm{Tr}(\mathcal{L}_{\mathbf{A}}(\mathbf{U})\mathbf{V}),$$

where $\mathcal{L}_{\mathbf{A}}(\mathbf{U})$ *is the solution of the Lyapunov equation* $\mathbf{X}\mathbf{A} + \mathbf{A}\mathbf{X} = \mathbf{U}$. *When* $\mathbf{A}, \mathbf{B} \in S^d_+$ *satisfy* $\mathrm{Im}\mathbf{B} \subset \mathrm{Im}\mathbf{A}$*, the Bures geodesic from* $\mathbf{A}$ *to* $\mathbf{B}$ *is given by*

$$\mathbf{C}_{\mathbf{A} \to \mathbf{B}}(t) = [(1-t)\mathbf{I}_\mathrm{d} + t\mathbf{T}^{\mathbf{AB}}]\mathbf{A}[(1-t)\mathbf{I}_\mathrm{d} + t\mathbf{T}^{\mathbf{AB}}], \quad t \in [0, 1], \tag{1.12}$$

and the exp *and* log *maps of the Riemannian Bures metric are given by*

$$\exp_{\mathbf{C}}(\mathbf{V}) = (\mathcal{L}_{\mathbf{C}}(\mathbf{V}) + \mathbf{I}_\mathrm{d})\, \mathbf{C}\, (\mathcal{L}_{\mathbf{C}}(\mathbf{V}) + \mathbf{I}_\mathrm{d}) \tag{1.13}$$

$$\log_{\mathbf{C}}(\mathbf{B}) = \left(\mathbf{T}^{\mathbf{CB}} - \mathbf{I}_\mathrm{d}\right) \mathbf{C} + \mathbf{C} \left(\mathbf{T}^{\mathbf{CB}} - \mathbf{I}_\mathrm{d}\right). \tag{1.14}$$

Bures-Wasserstein barycenters. We conclude this section by mentioning the Bures-Wasserstein barycenter problem, which is characterized by a fixed-point equation. In Chapter 4, we extend this result to (debiased) entropy-regularized OT between Gaussian measures.

Theorem 1.24 (Agueh and Carlier [2011, Theorem 6.1])**.** *Let* $n > 0$ *and* $\forall i \in [\![1, n]\!], \lambda_i \in \mathbb{R}_+, \mathbf{A}_i \in S^d_{++}, \mu_i = \mathcal{N}(0, \mathbf{A}_i)$ *with* $\sum_i \lambda_i = 1$. *Then, the Wasserstein barycenter problem*

$$\min_{\nu \in \mathcal{P}_2(\mathbb{R}^d)} \sum_{i=1}^n \lambda_i W_2^2(\mu_i, \nu) \tag{1.15}$$

[3] We refer to the textbook [Lee, 1997] for an introduction to Riemmanian geometry.

admits a unique solution $\nu^\star = \mathcal{N}(0, \mathbf{B})$*, where* $\mathbf{B}$ *is the unique solution of the Bures barycenter problem*

$$\min_{\mathbf{B}\in S_{++}^d} \sum_{i=1}^n \lambda_i \mathfrak{B}^2(\mathbf{A}_i, \mathbf{B}), \tag{1.16}$$

which is characterized by the equation

$$\sum_{i=1}^n \lambda_i (\mathbf{B}\mathbf{A}_i\mathbf{B})^{\frac{1}{2}} = \mathbf{B},$$

or equivalently

$$\sum_{i=1}^n \lambda_i \mathbf{T}^{\mathbf{B}\mathbf{A}_i} = \mathbf{I}_\mathrm{d}.$$

Theorem 1.24 can be extended to elliptical distribution, with the same relations on the covariance matrices or scale parameters. Further, Álvarez-Esteban et al. [2016] show that the solution of (1.16) can be obtained by the fixed-point iteration

$$\mathbf{B}_{n+1} = \mathbf{B}_n^{-1/2} \left(\sum_{i=1}^n \lambda_i (\mathbf{B}_n^{1/2}\mathbf{A}_i\mathbf{B}_n^{1/2})^{\frac{1}{2}} \right)^2 \mathbf{B}_n^{-1/2}.$$

3 Entropic Regularization of Optimal Transport

As mentioned in Section 1.2, OT distances can be expensive to compute, even in the relatively simple discrete setting. Besides, they suffer from a lack of differentiability that is detrimental to applications in machine learning. Starting from Cuturi [2013], the prevalent approach has consisted in adding an entropic regularization term to the optimal transport problem, which ensures its differentiability and allows the use of efficient algorithms. Let $\varepsilon > 0$ be the regularization strength, the entropy-regularized optimal transport problem is defined as

$$\mathrm{OT}_\varepsilon(\mu, \nu) \stackrel{\text{def}}{=} \inf_{\gamma\in\Pi(\mu,\nu)} \iint_{\mathbb{R}^d\times\mathbb{R}^d} c(x,y)\mathrm{d}\gamma(x,y) + \varepsilon\,\mathrm{KL}(\gamma\|\mu\otimes\nu), \tag{Ent-OT}$$

where $\mathrm{KL}(\gamma\|\mu\otimes\nu) \stackrel{\text{def}}{=} \iint_{\mathbb{R}^d\times\mathbb{R}^d} \log\left(\frac{\mathrm{d}\gamma(x,y)}{\mathrm{d}\mu(x)\mathrm{d}\nu(y)}\right)\mathrm{d}\gamma(x,y) + \iint_{\mathbb{R}^d\times\mathbb{R}^d} (\mathrm{d}\mu(x)\mathrm{d}\nu(y) - \mathrm{d}\gamma(x,y))$ is the Kullback-Leibler (KL) divergence. As the KL divergence is strictly convex in its first argument, this regularization term turns $(\mathcal{K})$ into the strictly convex problem (Ent-OT). In particular, strong duality holds. The dual problem of (Ent-OT) plays an important role in characterizing the additional properties induced by entropic regularization.

Proposition 1.25. *Strong duality holds, and* (Ent-OT) *has the following dual form*

$$\max_{f,g\in\mathcal{C}(\mathbb{R}^d)} \int_{\mathbb{R}^d} f\mathrm{d}\mu + \int_{\mathbb{R}^d} g\mathrm{d}\nu - \varepsilon \iint_{\mathbb{R}^d\times\mathbb{R}^d} e^{\frac{f(x)+g(y)-c(x,y)}{\varepsilon}}\mathrm{d}\mu(x)\mathrm{d}\nu(y). \tag{1.17}$$

At optimality, the dual variables f, g *are linked to the optimal transport plan* π *for* (Ent-OT) *via the following relation:*

$$\mathrm{d}\pi(x,y) = \exp\left(\frac{f(x)+g(y)-c(x,y)}{\varepsilon}\right)\mathrm{d}\mu(x)\mathrm{d}\nu(y), \tag{1.18}$$

and f, g satisfy

$$\begin{aligned} f(x) &= -\varepsilon \log \int_{\mathbb{R}^d} e^{\frac{g(y)-c(x,y)}{\varepsilon}} \mathrm{d}\nu(y) \qquad \mu - a.e. \\ g(y) &= -\varepsilon \log \int_{\mathbb{R}^d} e^{\frac{f(x)-c(x,y)}{\varepsilon}} \mathrm{d}\mu(x) \qquad \nu - a.e. \end{aligned} \tag{1.19}$$

A detailed proof of Proposition 1.25 (generalized to alternative regularization terms) can be found in [Chizat, 2017, Genevay, 2019].

Discrete entropic OT. Entropic regularization was initially introduced by Cuturi [2013] in the discrete setting:

$$\min_{\mathbf{P} \in U(\mathbf{a},\mathbf{b})} \langle \mathbf{P}, \mathbf{C} \rangle + \varepsilon \sum_{i,j} p_{ij} \log\left(\frac{p_{ij}}{a_i b_j}\right), \tag{1.20}$$

with two main motivations: (i) allowing a fast approximation of OT and (ii) ensuring smoothness and differentiability of OT [Peyré et al., 2019]. Propositions 1.26 and 1.27 show that this convexification allows both objectives to be attained.

Proposition 1.26 (Cuturi [2013], Peyré et al. [2019])**.** *Let $\alpha = \sum_{i=1}^n a_i \delta_{x_i}$ and $\beta = \sum_{j=1}^m b_j \delta_{y_j}$ be two discrete distributions, and $\varepsilon > 0$. Then,* (Ent-OT) *admits a unique solution $\mathbf{P}$ which is of the form*

$$\mathbf{P} = \mathrm{diag}(\mathbf{u}) \mathbf{K} \, \mathrm{diag}(\mathbf{v}), \tag{1.21}$$

with $\mathbf{K} \overset{\text{def}}{=} [\exp(-\frac{c(x_i,y_j)}{\varepsilon})]_{i,j} \in \mathbb{R}_+^{n \times m}$ and $\mathbf{u} \in \mathbb{R}_+^n, \mathbf{v} \in \mathbb{R}_+^m$ can be obtained using the fixed-point iterations

$$\mathbf{u}^{(l+1)} = \frac{1}{\mathbf{K}(\mathbf{v}^{(l)} \odot \mathbf{b})} \quad \text{and} \quad \mathbf{v}^{(l+1)} = \frac{1}{\mathbf{K}^T(\mathbf{u}^{(l+1)} \odot \mathbf{a})}. \tag{1.22}$$

The iterations (1.22) are known as Sinkhorn's algorithm, after Sinkhorn [1964] who first proved their convergence. With the notations $\mathbf{f}_i = f(x_i), \mathbf{g}_j = g(y_j)$ and $c_{ij} = c(x_i, y_j)$, the variables $\mathbf{u}, \mathbf{v}$ in Equation (1.21) are linked to f, g in Equation (1.17) through the relations

$$\mathbf{u} = [\exp(\mathbf{f}_i/\varepsilon)]_{i \in [\![1,n]\!]}, \qquad \mathbf{v} = [\exp(\mathbf{g}_j/\varepsilon)]_{j \in [\![1,m]\!]}, \tag{1.23}$$

and are sometimes called the *exponential scalings*. Hence, with this parameterization Sinkhorn's algorithm is equivalent to iteratively enforcing (1.19). Sinkhorn's algorithm is easy to implement and can be efficiently parallelized using graphics processing units (GPU) [Cuturi, 2013], but is numerically unstable for small values of ε. In that case, it can be run in the log-domain: this yields the iterations

$$\forall i \in [\![1,n]\!], \quad \mathbf{f}_i^{(l+1)} = -\varepsilon \log \sum_{j=1}^m b_j \exp((\mathbf{g}_j^{(l)} - c_{ij})/\varepsilon), \tag{1.24}$$

$$\forall j \in [\![1,m]\!], \quad \mathbf{g}_j^{(l+1)} = -\varepsilon \log \sum_{i=1}^n a_i \exp((\mathbf{f}_i^{(l+1)} - c_{ij})/\varepsilon). \tag{1.25}$$

The KL regularization term encourages the optimal plan to put mass on the whole support of $\mu \otimes \nu$. As shown in Figure 1.4 in the discrete case, this yields transportation plans with strictly positive entries everywhere, whereas unregularized transportation plans are sparse, with at most $n + m$ non-zero entries (see Section 1.2). Moreover, KL regularization ensures the uniqueness of a solution, and hence (by Danskin's theorem [Danskin, 1967]) the differentiability of (Ent-OT).

Proposition 1.27 (Gradients, Cuturi and Doucet [2014]). *Let* $\alpha = \sum_{i=1}^{n} a_i \delta_{x_i}$ *and* $\beta = \sum_{j=1}^{m} b_j \delta_{y_j}$. *Then,* OT_ε *is jointly convex w.r.t.* $(\mathbf{a}, \mathbf{b})$ *and differentiable, with gradients*

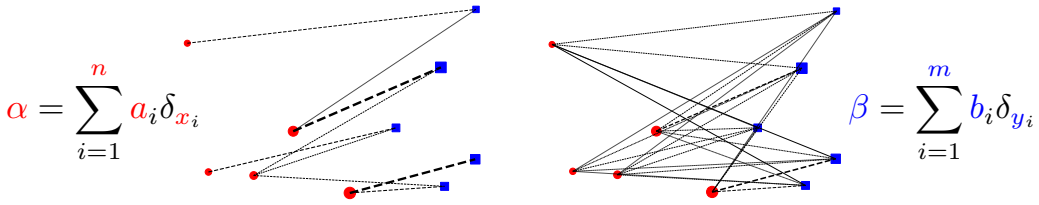
$$\nabla_{(\mathbf{a},\mathbf{b})} \mathrm{OT}_\varepsilon(\alpha, \beta) = (\mathbf{f}, \mathbf{g}), \tag{1.26}$$

where $(\mathbf{f}, \mathbf{g})$ *satisfies* (1.19). *If* $c(x, y) = \frac{1}{2}\|x - y\|^2$, *the gradients w.r.t. the supports are given by*

$$\forall i \in [\![1, n]\!], \quad \nabla_{x_i} \mathrm{OT}_\varepsilon(\alpha, \beta) = \frac{1}{a_i} \sum_{j=1}^{m} p_{ij}(x_i - y_j),$$

$$\forall j \in [\![1, m]\!], \quad \nabla_{y_j} \mathrm{OT}_\varepsilon(\alpha, \beta) = \frac{1}{b_j} \sum_{i=1}^{n} p_{ij}(y_j - x_i),$$

where $\mathbf{P}$ *is the optimal plan in* (Ent-OT).



(a) (D-OT) (b) (Ent-OT)

Figure 1.4: Effect of regularization on transportation plan density. *Left*: unregularized (sparse) OT plan. *Right*: regularized (dense) entropic OT plan.

Proposition 1.27 shows that entropy-regularized OT constitutes a suitable loss function for machine learning [Frogner et al., 2015], contrary to classical unregularized OT which is not differentiable. In practice, two strategies can be used to compute gradients: the first consists in using the dual potentials given by Sinkhorn's algorithm (in virtue of Proposition 1.27), while the second consists in performing automatic differentiation on the Sinkhorn iterations, which is the approach suggested in Genevay et al. [2018]. The latter method has a computational overhead equivalent to computing the (forward) Sinkhorn iterations, but recent research [Ablin et al., 2020] shows that better approximations of the gradients can be obtained that way.

Finally, Proposition 1.26 shows that entropic regularization allows to compute fast and differentiable transportation plans. However, a remaining question concerns which quantity to use to measure the difference between two distributions based on those entropic plans. Indeed, OT_ε is symmetric and has the advantage of having easily computable gradients, but it is no longer a distance as it does not satisfy the triangle inequality, nor even a divergence as it is not positive.[4] To alleviate the positivity issue, [Cuturi, 2013] propose to use $\mathrm{OT}_\varepsilon^{(sharp)} \overset{\text{def}}{=} \langle \mathbf{P}_\varepsilon, \mathbf{C} \rangle$, where $\mathbf{P}_\varepsilon$ is the solution of Equation (1.20). Luise et al. [2018] name this quantity *sharp Sinkhorn* and provide an algorithm to compute its gradients. $\mathrm{OT}_\varepsilon^{(sharp)}$ is positive, but $\mathrm{OT}_\varepsilon^{(sharp)}(\alpha, \alpha)$ can be strictly positive and hence sharp Sinkhorn is not a divergence.

Genevay et al. [2018] proposed to subtract debiasing terms from OT_ε, defining the Sinkhorn divergence:

$$S_\varepsilon(\mu, \nu) \overset{\text{def}}{=} \mathrm{OT}_\varepsilon(\mu, \nu) - \tfrac{1}{2}(\mathrm{OT}_\varepsilon(\mu, \mu) + \mathrm{OT}_\varepsilon(\nu, \nu)).$$

[4] In particular, for $\varepsilon > 0$, $\mathrm{OT}_\varepsilon(\alpha, \alpha) \leq 0$ and can be strictly negative.

Feydy et al. [2019] then proved that the Sinkhorn divergence defines a suitable loss function.

Proposition 1.28 (Feydy et al. [2019] (Simplified))**.** *Let $c(x, y) = \|x - y\|^p, p \geq 1$. Then for all compactly supported $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$, $S_\varepsilon(\mu, \nu)$ defines a symmetric positive definite divergence, which is convex in μ or ν (but not jointly), and metrizes weak convergence.*

Hence, short of the triangle inequality, Sinkhorn divergences possess all the properties that make Wasserstein distances suitable losses between distributions in a machine learning context. Further, Feydy et al. [2019] show that the computational overhead induced by computing $\mathrm{OT}_\varepsilon(\mu, \mu)$ and $\mathrm{OT}_\varepsilon(\nu, \nu)$ terms is limited, as Sinkhorn iterations can be adapted in a symmetric variant to obtain faster convergence. Chapter 5 will make heavy use of the favorable properties of Sinkhorn divergences.

Semi-discrete and continuous entropic transport. Although its computational advantages are most apparent in a fully discrete setting, entropic regularization has also been used to develop methods for the semi-discrete and continuous settings. In the semi-discrete setting, entropic regularization leads to replacing the indicator functions of Laguerre cells with a smoothed version [Peyré et al., 2019], resulting in a stochastic optimization problem which is amenable for stochastic gradient methods [Genevay et al., 2016]. In the continuous setting however, (Ent-OT) can no longer directly be cast as a stochastic optimization problem. A stochastic formulation can be obtained again by approximating the dual form of (4.1) using a kernel representation, which allows to use stochastic gradient methods [Genevay et al., 2016], an approach which was refined by Mensch and Peyré [2020]. In Chapter 4, closed forms for entropy-regularized optimal transport between Gaussian measures are proven, which constitute the first non-trivial closed forms in the continuous setting.

Alternative regularizations. As discussed in this section, entropic regularization allows to define approximations of OT distances that are differentiable, and to compute them efficiently using Sinkhorn's algorithm. However, the differentiability can be achieved using a wider range of strictly convex regularization functions $\mathcal{R}$ on the transportation plans [Blondel et al., 2018, Dessein et al., 2018, Muzellec et al., 2017]:

$$\mathrm{ROT}_\varepsilon^{\mathcal{R}}(\mu, \nu) \stackrel{\text{def}}{=} \min_{\gamma \in \Pi(\mu, \nu)} \iint_{\mathbb{R}^d \times \mathbb{R}^d} c(x, y) \mathrm{d}\gamma(x, y) + \varepsilon \mathcal{R}(\gamma). \tag{1.27}$$

While only entropy-regularized OT can be solved using Sinkhorn's algorithm, solutions to (1.27) are usually computed using dual ascent methods. This implies that ROT is in general less practical to compute or approximate than entropic OT. The main motivation behind ROT is rather to consider regularization functions which are sparsity-preserving in a discrete setting. In particular, Blondel et al. [2018] show that squared-norm regularization allows to retain most of the sparsity of unregularized OT plans, while leading to differentiable quantities.

Unbalanced optimal transport. So far, only regularization of couplings with exact marginals have been considered, i.e. with the constraint $\gamma \in \Pi(\mu, \nu)$. An additional step in relaxing OT consists in replacing couplings with positive measures that have free marginals and total mass, and penalizing the difference between the marginals and the original measures according to some divergence. Chizat [2017] considers in particular an unbalanced problem with entropic regularization on the transportation plan, and penalization of the

marginals with ϕ-divergences [Csiszár, 1975]:

$$\begin{aligned}\mathrm{UOT}_{\varepsilon,\tau}(\mu,\nu) \stackrel{\text{def}}{=} \min_{\gamma\in\mathcal{M}(\mathbb{R}^d\times\mathbb{R}^d)} \Big\{ \iint_{\mathbb{R}^d\times\mathbb{R}^d} c(x,y)\mathrm{d}\gamma(x,y) + \varepsilon\,\mathrm{KL}(\gamma\|\mu\otimes\nu) \\ + \tau D_{\phi_2}(\pi_{1\sharp}\gamma\|\mu) + \tau D_{\phi_1}(\pi_{2\sharp}\gamma\|\nu)\Big\},\end{aligned} \tag{U-OT}$$

where $D_\phi(\mu\|\nu) \stackrel{\text{def}}{=} \int_{\mathbb{R}^d}\phi\left(\frac{\mathrm{d}\mu(x)}{\mathrm{d}\nu(x)}\right)\mathrm{d}\nu(x)$, and the KL divergence is generalized to positive measures: $\mathrm{KL}(\mu\|\nu) = \int_{\mathbb{R}^d}\log\left(\frac{\mathrm{d}\mu(x)}{\mathrm{d}\nu(x)}\right)\mathrm{d}\mu(x) - \int_{\mathbb{R}^d}\mathrm{d}\mu + \int_{\mathbb{R}^d}\mathrm{d}\nu$, i.e. $\mathrm{KL} = D_\phi$ with $\phi(x) = x\log x - x + 1$. (U-OT) can be seen as (Ent-OT) where mass creation or deletion is allowed along with mass transportation. This intuition is formalized from a dynamical transport point of view in [Chizat et al., 2018a].

In the particular case where $D_{\phi_1} = D_{\phi_2} = \mathrm{KL}$, the optimal plan is of the form $\mathrm{d}\pi(x,y) = e^{\frac{f(x)+g(y)-c(x,y)}{\varepsilon}}\mathrm{d}\mu(x)\mathrm{d}\nu(y)$, where $f, g \in \mathcal{C}(\mathbb{R}^d)$ satisfy

$$\begin{aligned} f(x) &= -\rho\tau\log\int_{\mathbb{R}^d} e^{\frac{g(y)-c(x,y)}{\varepsilon}}\mathrm{d}\nu(y) \qquad \mu-a.e. \\ g(y) &= -\rho\tau\log\int_{\mathbb{R}^d} e^{\frac{f(x)-c(x,y)}{\varepsilon}}\mathrm{d}\mu(x) \qquad \nu-a.e. \end{aligned} \tag{1.28}$$

with $\rho \stackrel{\text{def}}{=} \frac{\tau}{\tau+\varepsilon}$. Hence, in the discrete setting, (U-OT) can be solved using the generalized Sinkhorn iterations

$$\mathbf{u}^{(l+1)} = \left(\frac{1}{\mathbf{K}(\mathbf{v}^{(l)}\odot\mathbf{b})}\right)^\rho \quad \text{and} \quad \mathbf{v}^{(l+1)} = \left(\frac{1}{\mathbf{K}^T(\mathbf{u}^{(l+1)}\odot\mathbf{a})}\right)^\rho, \tag{1.29}$$

with $\mathbf{K} = \exp\left(-\mathbf{C}/\varepsilon\right), \mathbf{u} = \exp(\mathbf{f}/\varepsilon)$, and $\mathbf{v} = \exp(\mathbf{g}/\varepsilon)$.

In Chapter 4, closed forms for (Ent-OT) and (U-OT) are proved for Gaussian measures based on Sinkhorn-like fixed-point equations.

Chapter 2

Embeddings in the Wasserstein Space of Elliptical Distributions

Embedding complex objects as vectors in low dimensional spaces is a longstanding problem in machine learning. We propose in this chapter an extension of that approach, which consists in embedding objects as elliptical probability distributions, namely distributions whose densities have elliptical level sets. We endow these measures with the 2-Wasserstein metric, with two important benefits:

(i) For such measures, the squared 2-Wasserstein metric has a closed form, equal to a weighted sum of the squared Euclidean distance between means and the squared Bures metric between covariance matrices. The latter is a Riemannian metric between positive semi-definite matrices, which turns out to be Euclidean on a suitable factor representation of such matrices, which is valid on the entire geodesic between these matrices;

(ii) The 2-Wasserstein distance boils down to the usual Euclidean metric when comparing Diracs, and therefore provides a natural framework to extend point embeddings.

We show that for these reasons Wasserstein elliptical embeddings are more intuitive and yield tools that are better behaved numerically than the alternative choice of Gaussian embeddings with the Kullback-Leibler divergence. In particular, and unlike previous work based on the KL geometry, we learn elliptical distributions that are not necessarily diagonal. We demonstrate the advantages of elliptical embeddings by using them for visualization, to compute embeddings of words, and to reflect entailment or hypernymy.

This chapter is based on [Muzellec and Cuturi, 2018]. In this original work, Newton-Schulz (NS) iterations were utilized to obtain the matrix roots and inverse roots required for the computation of the Bures distance and its gradient. In this updated version, we use NS iterations to directly obtain Monge maps and inverse maps, resulting in a more efficient numerical scheme.

1 Introduction

One of the holy grails of machine learning is to compute meaningful low-dimensional embeddings for high-dimensional complex data. That ability has recently proved crucial to tackle more advanced tasks, such as for instance: inference on texts using word embeddings [Mikolov et al., 2013b, Pennington et al., 2014, Bojanowski et al., 2017], improved image understanding [Norouzi et al., 2014], representations for nodes in large graphs [Grover and Leskovec, 2016].

Such embeddings have been traditionally recovered by seeking *isometric* embeddings in lower dimensional Euclidean spaces, as studied in [Johnson and Lindenstrauss, 1984, Bourgain, 1985]. Given n input points $x_1, \ldots, x_n$, one seeks as many embeddings $\mathbf{y}_1, \ldots, \mathbf{y}_n$ in a target space $\mathcal{Y} = \mathbb{R}^d$ whose pairwise distances $\|\mathbf{y}_i - \mathbf{y}_j\|_2$ do not depart too much from the original distances $d_{\mathcal{X}}(x_i, x_j)$ in the input space. Note that when d is restricted to be 2 or 3, these embeddings $(\mathbf{y}_i)_i$ provide a useful way to visualize the entire dataset. Starting with metric multidimensional scaling (mMDS) [De Leeuw, 1977, Borg and Groenen, 2005], several approaches have refined this intuition [Tenenbaum et al., 2000, Roweis and Saul, 2000, Hinton and Roweis, 2003, Maaten and Hinton, 2008]. More general criteria, such as reconstruction error [Hinton and Salakhutdinov, 2006, Kingma and Welling, 2014]; co-occurence [Globerson et al., 2007]; or relational knowledge, be it in metric learning [Weinberger and Saul, 2009] or between words [Mikolov et al., 2013b] can be used to obtain vector embeddings. In such cases, distances $\|\mathbf{y}_i - \mathbf{y}_j\|_2$ between embeddings, or alternatively their dot-products $\langle \mathbf{y}_i, \mathbf{y}_j \rangle$ must comply with sophisticated desiderata. Naturally, more general and flexible approaches in which the embedding space $\mathcal{Y}$ needs not be Euclidean can be considered, for instance in generalized MDS on the sphere [Maron et al., 2010], on surfaces [Bronstein et al., 2006], in spaces of trees [Bădoiu et al., 2007, Fakcharoenphol et al., 2003] or, more recently, computed in the Poincaré hyperbolic space [Nickel and Kiela, 2017].

Probabilistic Embeddings. Our work belongs to a recent trend, pioneered by Vilnis and McCallum, who proposed to embed data points as *probability measures* in $\mathbb{R}^d$ [2015], and therefore generalize point embeddings. Indeed, point embeddings can be regarded as a very particular—and degenerate—case of probabilistic embedding, in which the uncertainty is infinitely concentrated on a single point (a Dirac). Probability measures can be more spread-out, or event multimodal, and provide therefore an opportunity for additional flexibility. Naturally, such an opportunity can only be exploited by defining a metric, divergence or dot-product on the space (or a subspace thereof) of probability measures. Vilnis and McCallum proposed to embed words as *Gaussians* endowed either with the Kullback-Leibler (KL) divergence or the expected likelihood kernel [Jebara et al., 2004]. The Kullback-Leibler and expected likelihood kernel on measures have, however, an important drawback: these geometries do not coincide with the usual Euclidean metric between point embeddings when the variances of these Gaussians collapse. Indeed, the KL divergence and the ℓ_2 distance between two Gaussians diverges to ∞ or saturates when the variances of these Gaussians become small. To avoid numerical instabilities arising from this degeneracy, Vilnis and McCallum must restrict their work to diagonal covariance matrices. In a concurrent approach, Singh et al. represent words as distributions over their contexts in the optimal transport geometry [Singh et al., 2020].

Contributions. We propose in this work a new framework for probabilistic embeddings, in which point embeddings are seamlessly handled as a particular case. We consider arbitrary families of elliptical distributions, which subsume Gaussians, and also include uniform elliptical distributions, which are arguably easier to visualize because of their compact support. Our approach uses the 2-Wasserstein distance to compare elliptical distributions. The latter can handle degenerate measures, and both its value and its gradients admit

closed forms [Gelbrich, 1990], either in their natural Riemannian formulation, as well as in a more amenable local Euclidean parameterization. We provide numerical tools to carry out the computation of elliptical embeddings in different scenarios, both to optimize them with respect to metric requirements (as is done in multidimensional scaling) or with respect to dot-products (as shown in our applications to word embeddings for entailment, similarity and hypernymy tasks) for which we introduce a proxy using a polarization identity.

2 The Geometry of Elliptical Distributions in the Wasserstein Space

We recall in this section basic facts about elliptical distributions in $\mathbb{R}^d$. We adopt a general formulation that can handle measures supported on subspaces of $\mathbb{R}^d$ as well as Dirac (point) measures. That level of generality is needed to provide a seamless connection with usual vector embeddings, seen in the context of this chapter as Dirac masses. We recall results from the literature showing that the squared 2-Wasserstein distance between two distributions from the same family of elliptical distributions is equal to the squared Euclidean distance between their means plus the squared Bures metric between their scale parameter scaled by a suitable constant.

Elliptically-contoured densities. In their simplest form, elliptical distributions can be seen as generalizations of Gaussian multivariate densities in $\mathbb{R}^d$: their level sets describe concentric ellipsoids, shaped following a scale parameter $\mathbf{C} \in S^d_{++}$, and centered around a mean parameter $\mathbf{c} \in \mathbb{R}^d$ [Cambanis et al., 1981]. The density at a point $\mathbf{x}$ of such distributions is $f(\|\mathbf{x}-\mathbf{c}\|_{\mathbf{C}^{-1}})/\sqrt{|\mathbf{C}|}$ where the generator function f is such that $\int_{\mathbb{R}^d} f(\|\mathbf{x}\|^2)\mathrm{d}\mathbf{x} = 1$. Gaussians are recovered with $f = g, g(\cdot) \propto e^{-\cdot/2}$ while uniform distributions on full rank ellipsoids result from $f = u, u(\cdot) \propto \mathbf{1}_{\cdot \leq 1}$.

Because the norm induced by $\mathbf{C}^{-1}$ appears in formulas above, the scale parameter $\mathbf{C}$ must have full rank for these definitions to be meaningful. Cases where $\mathbf{C}$ does not have full rank can however appear when a probability measure is supported on an affine subspace[1] of $\mathbb{R}^d$, such as lines in $\mathbb{R}^2$, or even possibly a space of null dimension when the measure is supported on a single point (a Dirac measure), in which case its scale parameter $\mathbf{C}$ is $\mathbf{0}$. We provide in what follows a more general approach to handle these degenerate cases.

Elliptical distributions. To lift this limitation, several reformulations of elliptical distributions have been proposed to handle degenerate scale matrices $\mathbf{C}$ of rank $\mathrm{rk}\mathbf{C} < d$. Gelbrich [1990, Theorem 2.4] defines elliptical distributions as measures with a density w.r.t the Lebesgue measure of dimension $\mathrm{rk}\mathbf{C}$, in the affine space $\mathbf{c} + \mathrm{Im}\mathbf{C}$, where the image of $\mathbf{C}$ is $\mathrm{Im}\mathbf{C} \stackrel{\text{def}}{=} \{\mathbf{C}\mathbf{x}, \mathbf{x} \in \mathbb{R}^d\}$. This approach is intuitive, in that it reduces to describing densities in their relevant subspace. A more elegant approach uses the parameterization provided by characteristic functions [Cambanis et al., 1981, Fang et al., 1990]. In a nutshell, recall that the characteristic function of a multivariate Gaussian is equal to $\phi(\mathbf{t}) = e^{i\mathbf{t}^T\mathbf{c}} g(\mathbf{t}^T\mathbf{C}\mathbf{t})$ where, as in the paragraph above, $g(\cdot) = e^{-\cdot/2}$. A natural generalization to consider other elliptical distributions is therefore to consider for g other functions h of positive type [Ushakov, 1999, Theo.1.8.9], such as the indicator function u above, and still apply them to the same argument $\mathbf{t}^T\mathbf{C}\mathbf{t}$. Such functions are called *characteristic generators* and fully determine, along with a mean $\mathbf{c}$ and a scale parameter $\mathbf{C}$, an elliptical measure. This parameterization does not require the scale parameter $\mathbf{C}$ to be invertible, and therefore allows to define probability distributions that do not have necessarily a density w.r.t to

[1] For instance, the random variable Y in $\mathbb{R}^2$ obtained by duplicating the same normal random variable X in $\mathbb{R}$, $Y = [X, X]$, is supported on a line in $\mathbb{R}^2$ and has no density w.r.t the Lebesgue measure in $\mathbb{R}^2$.

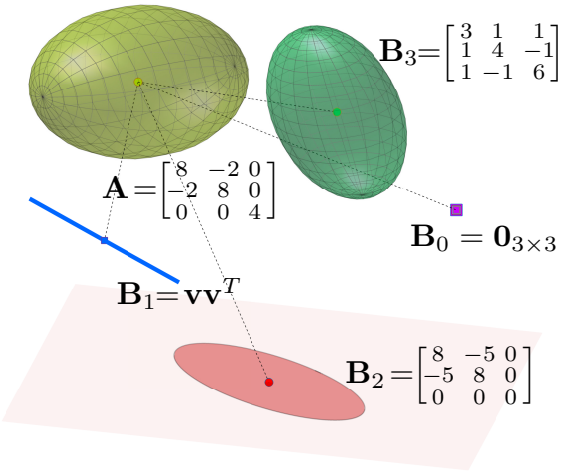


Figure 2.1: Five measures from the family of uniform elliptical distributions in $\mathbb{R}^3$. Each measure has a mean (location) and scale parameter. In this carefully selected example, the reference measure (with scale parameter $\mathbf{A}$) is equidistant (according to the 2-Wasserstein metric) to the four remaining measures, whose scale parameters $\mathbf{B}_0, \mathbf{B}_1, \mathbf{B}_2, \mathbf{B}_3$ have ranks equal to their indices (here, $\mathbf{v} = [3, 7, -2]^T$).

the Lebesgue measure in $\mathbb{R}^d$. Both constructions are relatively complex, and we refer the interested reader to these references for a rigorous treatment.

Rank-deficient elliptical distributions and their variances. For the purpose of this work, we will only require the following result: the variance of an elliptical measure is equal to its scale parameter $\mathbf{C}$ multiplied by a scalar that only depends on its characteristic generator. Indeed, given a mean vector $\mathbf{c} \in \mathbb{R}^d$, a scale *semi*-definite matrix $\mathbf{C} \in S_+^d$ and a characteristic generator function h, we define $\mu_{h,\mathbf{c},\mathbf{C}}$ to be the measure with characteristic function $\mathbf{t} \mapsto e^{i\mathbf{t}^T\mathbf{c}} h(\mathbf{t}^T\mathbf{C}\mathbf{t})$. In that case, one can show that the covariance matrix of $\mu_{h,\mathbf{c},\mathbf{C}}$ is equal to its scale parameter $\mathbf{C}$ times a constant τ_h that only depends on h, namely

$$\operatorname{var}(\mu_{h,\mathbf{c},\mathbf{C}}) = \tau_h \mathbf{C} \ . \tag{2.1}$$

For Gaussians, the scale parameter $\mathbf{C}$ and its covariance matrice coincide, that is $\tau_g = 1$. For uniform elliptical distributions, one has $\tau_u = 1/(d+2)$: the covariance of a uniform distribution on the volume $\{\mathbf{c} + \mathbf{C}\mathbf{x}, \mathbf{x} \in \mathbb{R}^d, \|\mathbf{x}\| \leq 1\}$, such as those represented in Figure 2.1, is equal to $\mathbf{C}/(d+2)$.

The 2-Wasserstein Bures metric. A natural metric for elliptical distributions arises from optimal transport (OT) theory. We refer interested readers to [Santambrogio, 2015, Peyré et al., 2019] for exhaustive surveys on OT. Recall that for two arbitrary probability measures $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$, their squared 2-Wasserstein distance is equal to

$$W_2^2(\mu, \nu) \stackrel{\text{def}}{=} \inf_{X\sim\mu, Y\sim\nu} \mathbb{E}_{\|X-Y\|_2^2}.$$

This formula rarely has a closed form. However, in the footsteps of Dowson and Landau [1982] who proved it for Gaussians, Gelbrich [1990] showed that for $\alpha \stackrel{\text{def}}{=} \mu_{h,\mathbf{a},\mathbf{A}}$ and $\beta \stackrel{\text{def}}{=} \mu_{h,\mathbf{b},\mathbf{B}}$ in the *same* family $\mathcal{P}_h = \{\mu_{h,\mathbf{c},\mathbf{C}}, \mathbf{c} \in \mathbb{R}^d, \mathbf{C} \in S_+^d\}$, one has

$$\begin{aligned} W_2^2(\alpha, \beta) &= \|\mathbf{a} - \mathbf{b}\|_2^2 + \mathfrak{B}^2(\operatorname{var}\alpha, \operatorname{var}\beta) \\ &= \|\mathbf{a} - \mathbf{b}\|_2^2 + \tau_h \mathfrak{B}^2(\mathbf{A}, \mathbf{B}), \end{aligned} \tag{2.2}$$

where $\mathfrak{B}^2$ is the (squared) Bures metric on S_+^d, proposed in quantum information geometry [1969] and studied recently in [Bhatia et al., 2018, Malagò et al., 2018],

$$\mathfrak{B}^2(\mathbf{X}, \mathbf{Y}) \stackrel{\text{def}}{=} \mathrm{Tr}(\mathbf{X} + \mathbf{Y} - 2(\mathbf{X}^{\frac{1}{2}}\mathbf{Y}\mathbf{X}^{\frac{1}{2}})^{\frac{1}{2}}). \tag{2.3}$$

The factor τ_h next to the rightmost term $\mathfrak{B}^2$ in (2.2) arises from homogeneity of $\mathfrak{B}^2$ in its arguments (2.3), which is leveraged using the identity in (2.1).

A few remarks.

(i) When both scale matrices $\mathbf{A} = \mathrm{diag}\,\mathbf{d_A}$ and $\mathbf{B} = \mathrm{diag}\,\mathbf{d_B}$ are diagonal, $W_2^2(\alpha, \beta)$ is the sum of two terms: the usual squared Euclidean distance between their means, plus τ_h times the squared *Hellinger* metric between the diagonals $\mathbf{d_A}, \mathbf{d_B}$:

$$\mathfrak{H}^2(\mathbf{d_A}, \mathbf{d_B}) \stackrel{\text{def}}{=} \|\sqrt{\mathbf{d_A}} - \sqrt{\mathbf{d_B}}\|_2^2.$$

(ii) The distance W_2 between two Diracs $\delta_\mathbf{a}, \delta_\mathbf{b}$ is equal to the usual distance between vectors $\|\mathbf{a} - \mathbf{b}\|_2$.

(iii) The squared distance W_2^2 between a Dirac $\delta_\mathbf{a}$ and a measure $\mu_{h,\mathbf{b},\mathbf{B}}$ in $\mathcal{P}_h$ reduces to $\|\mathbf{a} - \mathbf{b}\|^2 + \tau_h \mathrm{Tr}\mathbf{B}$. The distance between a point and an ellipsoid distribution therefore always *increases* as the scale parameter of the latter increases. Although this point makes sense from the quadratic viewpoint of W_2^2 (in which the quadratic contribution $\|\mathbf{a} - \mathbf{x}\|_2^2$ of points $\mathbf{x}$ in the ellipsoid that stand further away from $\mathbf{a}$ than $\mathbf{b}$ will dominate that brought by points $\mathbf{x}$ that are closer, see Figure 2.3) this may be counterintuitive for applications to visualization, an issue that will be addressed in Section 4.

(iv) The W_2 distance between two elliptical distributions in the same family $\mathcal{P}_h$ is always finite, no matter how degenerate they are. This is illustrated in Figure 2.1 in which a uniform measure $\mu_{\mathbf{a},\mathbf{A}}$ is shown to be exactly equidistant to four other uniform elliptical measures, some of which are degenerate. However, as can be hinted by the simple example of the Hellinger metric, that distance may not be differentiable for degenerate measures (in the same sense that $(\sqrt{x} - \sqrt{y})^2$ is defined at $x = 0$ but not differentiable w.r.t x).

(v) Although we focus in this chapter on uniform elliptical distributions, notably because they are easier to plot and visualize, considering any other elliptical family simply amounts to changing the constant τ_h next to the Bures metric in (2.2). Alternatively, increasing (or tuning) that parameter τ_h simply amounts to considering elliptical distributions with increasingly heavier tails.

3 Optimizing over the Space of Elliptical Embeddings

Our goal in this chapter is to use the set of elliptical distributions endowed with the W_2 distance as an embedding space. To optimize objective functions involving W_2 terms, we study in this section several parameterizations of the parameters of elliptical distributions. Location parameters only appear in the computation of W_2 through their Euclidean metric, and offer therefore no particular challenge. Scale parameters are more tricky to handle since they are constrained to lie in S_+^d. Rather than keeping track of scale parameters, we advocate optimizing directly on factors of such parameters, which results in simple Euclidean (unconstrained) updates reviewed below.

Geodesics for elliptical distributions. When $\mathbf{A}$ and $\mathbf{B}$ have full rank, the geodesic from α to β is a curve of measures in the same family of elliptic distributions, characterized by location and scale parameters $\mathbf{c}(t), \mathbf{C}(t)$, where

$$\mathbf{c}(t) = (1-t)\mathbf{a} + t\mathbf{b}; \quad \mathbf{C}(t) = \left((1-t)\mathbf{I} + t\mathbf{T}^{\mathbf{AB}}\right) \mathbf{A} \left((1-t)\mathbf{I} + t\mathbf{T}^{\mathbf{AB}}\right), \tag{2.4}$$

and where the matrix $\mathbf{T}^{\mathbf{AB}}$ is such that $\mathbf{x} \to \mathbf{T}^{\mathbf{AB}}(\mathbf{x} - \mathbf{a}) + \mathbf{b}$ is the so-called Monge or Brenier optimal transportation map [1987] from α to β, given in closed form as

$$\mathbf{T}^{\mathbf{AB}} \stackrel{\text{def}}{=} \mathbf{A}^{-\frac{1}{2}}(\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}})^{\frac{1}{2}}\mathbf{A}^{-\frac{1}{2}}, \tag{2.5}$$

and is the unique PSD matrix such that $\mathbf{B} = \mathbf{T}^{\mathbf{AB}}\mathbf{A}\mathbf{T}^{\mathbf{AB}}$ (Lemma 1.20). When $\mathbf{A}$ is degenerate, such a curve still exists as long as $\text{Im}\mathbf{B} \subset \text{Im}\mathbf{A}$, in which case the expression above is still valid using pseudo-inverse square roots $\mathbf{A}^{\dagger/2}$ in place of the usual inverse square-root (Proposition 1.15).

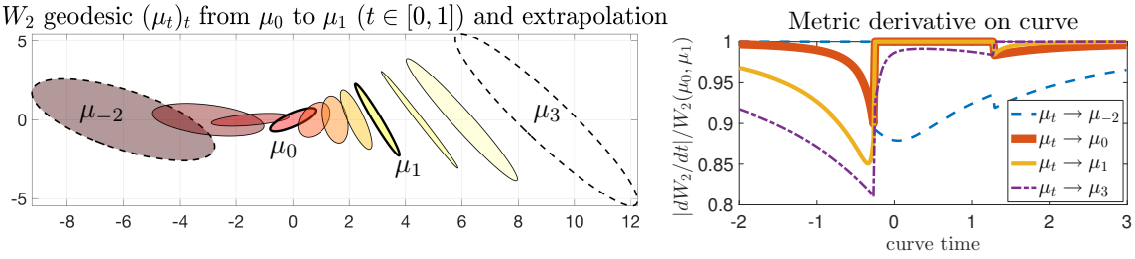


Figure 2.2: (left) Interpolation $(\mu_t)_t$ between two measures μ_0 and μ_1 following the geodesic equation (2.4). The same formula can be used to interpolate on the left and right of times $0, 1$. Displayed times are $[-2, -1, -.5, 0, .25, .5, .75, 1, 1.5, 2, 3]$. Note that geodesicity is not ensured outside of the boundaries $[0, 1]$. This is illustrated in the right plot displaying normalized metric derivatives of the curve μ_t to four relevant points: $\mu_0, \mu_1, \mu_{-2}, \mu_3$. The curve μ_t is not always locally geodesic, as can be seen by the fact that the metric derivative is strictly smaller than 1 in several cases.

Differentiability in Riemannian parameterization. Scale parameters are restricted to lie on the cone S_+^d. For such problems, it is well known that a direct gradient-and-project based optimization on scale parameters would prove too expensive. A natural remedy to this issue is to perform manifold optimization [Absil et al., 2009]. Indeed, as in any Riemannian manifold, the Riemannian gradient $\text{grad}_x \frac{1}{2}d^2(x, y)$ is given by $-\log_x y$ [Lee, 1997]. Using the expressions of the exp and log given in Proposition 1.23, we can show that minimizing $\frac{1}{2}\mathfrak{B}^2(\mathbf{A}, \mathbf{B})$ using Riemannian gradient descent with step length η corresponds to making updates of the form

$$\mathbf{A}' = \left((1-\eta)\mathbf{I} + \eta\mathbf{T}^{\mathbf{AB}}\right) \mathbf{A} \left((1-\eta)\mathbf{I} + \eta\mathbf{T}^{\mathbf{AB}}\right). \tag{2.6}$$

When $0 \leq \eta \leq 1$, this corresponds to considering a new point $\mathbf{A}'$ closer to $\mathbf{B}$ along the Bures geodesic between $\mathbf{A}$ and $\mathbf{B}$. When η is negative or larger than 1, $\mathbf{A}'$ no longer lies on this geodesic but is guaranteed to remain PSD, as can be seen from (2.6). Figure 2.2 shows a W_2 geodesic between two measures μ_0 and μ_1, as well as its extrapolation following exactly the formula given in (2.4). This figure illustrates that μ_t is not necessarily geodesic outside of the boundaries $[0, 1]$ w.r.t. three relevant measures, because its metric derivative is smaller than 1 [Ambrosio et al., 2006, Theorem 1.1.2]. When negative steps are taken (for instance when the W_2^2 distance needs to be increased), this lack of geodisicity has proved difficult to handle numerically for a simple reason: such updates may lead to degenerate

scale parameters $\mathbf{A}'$, as illustrated around time $t = 1.5$ of the curve in Figure 2.2. Another obvious drawback of Riemannian approaches is that they are not as well studied as simpler non-constrained Euclidean problems, for which a plethora of optimization techniques are available. This observations motivates an alternative Euclidean parameterization, detailed in the next paragraph.

Differentiability in Euclidean parameterization. A canonical way to handle a PSD constraint for $\mathbf{A}$ is to rewrite it in factor form $\mathbf{A} = \mathbf{L}\mathbf{L}^T$. In the particular case of the Bures metric, we show that this simple parametrization comes without losing the geometric interest of manifold optimization, while benefiting from simpler additive updates. Indeed, one can (see Section 5) that the gradient of the squared Bures metric has the following gradient:

$$\nabla_{\mathbf{L}} \frac{1}{2}\mathfrak{B}^2(\mathbf{A}, \mathbf{B}) = \left(\mathbf{I} - \mathbf{T}^{\mathbf{AB}}\right)\mathbf{L}, \quad \text{with updates } \mathbf{L}' = \left((1-\eta)\mathbf{I} + \eta\mathbf{T}^{\mathbf{AB}}\right)\mathbf{L}. \tag{2.7}$$

Links between Euclidean and Riemannian parameterizations. The factor updates in (2.7) are exactly equivalent to the Riemannian ones (2.6) in the sense that $\mathbf{A}' = \mathbf{L}'\mathbf{L}'^T$. Therefore, by using a factor parameterization we carry out updates that stay on the Riemannian geodesic and yet only require linear updates on $\mathbf{L}$, independently of the factor $\mathbf{L}$ chosen to represent $\mathbf{A}$ (given a factor $\mathbf{L}$ of $\mathbf{A}$, any right-side multiplication of that matrix by a unitary matrix remains a factor of $\mathbf{A}$).

When considering a general loss function $\mathcal{L}$ that takes as arguments squared Bures distances, one can also show that $\mathcal{L}$ is geodesically convex w.r.t. to scale matrices $\mathbf{A}$ if and only if it is convex in the usual sense with respect to $\mathbf{L}$, where $\mathbf{A} = \mathbf{L}\mathbf{L}^T$. Write now $\mathbf{L}_{\mathbf{B}} = \mathbf{T}^{\mathbf{AB}}\mathbf{L}$. One can recover that $\mathbf{L}_{\mathbf{B}}\mathbf{L}_{\mathbf{B}}^T = \mathbf{B}$. Therefore, expanding the expression $\mathfrak{B}^2$ for the right term we obtain

$$\begin{aligned}
\mathfrak{B}^2(\mathbf{A}, \mathbf{B}) &= \mathfrak{B}^2\left(\mathbf{L}\mathbf{L}^T, \mathbf{L}_{\mathbf{B}}\mathbf{L}_{\mathbf{B}}^T\right) \\
&= \mathfrak{B}^2\left(\mathbf{L}\mathbf{L}^T, \mathbf{T}^{\mathbf{AB}}\mathbf{L}\left(\mathbf{T}^{\mathbf{AB}}\mathbf{L}\right)^T\right) \\
&= \|\mathbf{L} - \mathbf{T}^{\mathbf{AB}}\mathbf{L}\|_F^2
\end{aligned} \tag{2.8}$$

Indeed, the Bures distance simply reduces to the Frobenius distance between two factors of $\mathbf{A}$ and $\mathbf{B}$. However these factors need to be carefully chosen: given $\mathbf{L}$ for $\mathbf{A}$, the factor for $\mathbf{B}$ must be computed according to an optimal transport map $\mathbf{T}^{\mathbf{AB}}$. In fact, the Bures distance is equal to the minimal Frobenius norm between factors of $\mathbf{A}$ and $\mathbf{B}$ [Bhatia et al., 2018]:

$$\mathfrak{B}(\mathbf{A}, \mathbf{B}) = \min_{\substack{\mathbf{M},\mathbf{N} \in \mathbb{R}^{d\times d} \\ \mathbf{M}\mathbf{M}^T = \mathbf{A}, \mathbf{N}\mathbf{N}^T = \mathbf{B}}} \|\mathbf{M} - \mathbf{N}\|_F.$$

Polarization between elliptical distributions. Some of the applications we consider, such as the estimation of word embeddings, are inherently based on dot-products. By analogy with the polarization identity, $\langle \mathbf{x}, \mathbf{y}\rangle = (\|\mathbf{x}-\mathbf{0}\|^2 + \|\mathbf{y}-\mathbf{0}\|^2 - \|\mathbf{x}-\mathbf{y}\|^2)/2$, we define a Wasserstein-Bures *pseudo*-dot-product based on the quantum fidelity $F(\mathbf{A}, \mathbf{B})$ [Bures, 1969] (see Section 2), where $\delta_{\mathbf{0}} = \mu_{\mathbf{0}_d, \mathbf{0}_{d\times d}}$ is the Dirac mass at $\mathbf{0}$,

$$[\mu_{\mathbf{a},\mathbf{A}} : \mu_{\mathbf{b},\mathbf{B}}] \stackrel{\text{def}}{=} \tfrac{1}{2}\left(W_2^2(\mu_{\mathbf{a},\mathbf{A}}, \delta_{\mathbf{0}}) + W_2^2(\mu_{\mathbf{b},\mathbf{B}}, \delta_{\mathbf{0}}) - W_2^2(\mu_{\mathbf{a},\mathbf{A}}, \mu_{\mathbf{b},\mathbf{B}})\right) \tag{2.9}$$

$$= \langle \mathbf{a}, \mathbf{b}\rangle + \mathrm{Tr}(\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}})^{\frac{1}{2}}. \tag{2.10}$$

Note that $[\cdot : \cdot]$ is not an actual inner product since the Bures metric is not Hilbertian, unless we restrict ourselves to diagonal covariance matrices, in which case it is the the inner

product between $(\mathbf{a}, \sqrt{\mathbf{d_A}})$ and $(\mathbf{b}, \sqrt{\mathbf{d_B}})$. We use $[\mu_{\mathbf{a},\mathbf{A}} : \mu_{\mathbf{b},\mathbf{B}}]$ as a similarity measure which has, however, some regularity: one can show that when $\mathbf{a}, \mathbf{b}$ are constrained to have equal norms and $\mathbf{A}$ and $\mathbf{B}$ equal traces, then $[\mu_{\mathbf{a},\mathbf{A}} : \mu_{\mathbf{b},\mathbf{B}}]$ is maximal when $\mathbf{a} = \mathbf{b}$ and $\mathbf{A} = \mathbf{B}$. Differentiating all three terms in that sum, the gradient of this pseudo dot-product w.r.t. $\mathbf{A}$ reduces to $\nabla_{\mathbf{A}}[\mu_{\mathbf{a},\mathbf{A}} : \mu_{\mathbf{b},\mathbf{B}}] = \mathbf{T}^{\mathbf{AB}}$.

3.1 Computational aspects

The computational bottleneck of gradient-based Bures optimization lies in the matrix square roots and inverse square roots operations that arise when instantiating transport maps $\mathbf{T}$ as in (2.5). A naive method using eigenvector decomposition is far too time-consuming, and there is not yet, to the best of our knowledge, a straightforward way to perform it in batches on a GPU. We propose to use Newton-Schulz iterations (Algorithms 1 and 2, see [Higham, 2008, Theorem 5.2 and Ch. 6]) to directly compute Monge maps $\mathbf{T}^{\mathbf{AB}}$ and $\mathbf{T}^{\mathbf{BA}}$. These iterations rely exclusively on matrix-matrix multiplications, and stream efficiently on GPUs.

Algorithm 1 NS Root Iterations

Input: PSD matrix $\mathbf{A}$, $\varepsilon > 0$
  Initialization: $\mathbf{Y} \leftarrow \frac{\mathbf{A}}{(1+\varepsilon)\|\mathbf{A}\|}, \mathbf{Z} \leftarrow \mathbf{I}$
  while not converged **do**
    $\mathbf{T} \leftarrow (3\mathbf{I} - \mathbf{ZY})/2$
    $\mathbf{Y} \leftarrow \mathbf{YT}$
    $\mathbf{Z} \leftarrow \mathbf{TZ}$
  end while
  $\mathbf{Y} \leftarrow \sqrt{(1+\varepsilon)\|\mathbf{A}\|}\mathbf{Y}$
  $\mathbf{Z} \leftarrow \frac{\mathbf{Z}}{\sqrt{(1+\varepsilon)\|\mathbf{A}\|}}$
Output: $\mathbf{Y} = \mathbf{A}^{1/2}$, $\mathbf{Z} = \mathbf{A}^{-1/2}$

Algorithm 2 NS Monge Iterations

Input: PSD matrices $\mathbf{A}, \mathbf{B}$, $\epsilon > 0$
  $\mathbf{Y} \leftarrow \frac{\mathbf{B}}{(1+\epsilon)\|\mathbf{B}\|}, \mathbf{Z} \leftarrow \frac{\mathbf{A}}{(1+\epsilon)\|\mathbf{A}\|}$
  while not converged **do**
    $\mathbf{T} \leftarrow (3\mathbf{I} - \mathbf{ZY})/2$
    $\mathbf{Y} \leftarrow \mathbf{YT}$
    $\mathbf{Z} \leftarrow \mathbf{TZ}$
  end while
  $\mathbf{Y} \leftarrow \sqrt{\|\mathbf{B}\|/\|\mathbf{A}\|}\mathbf{Y}$
  $\mathbf{Z} \leftarrow \sqrt{\|\mathbf{A}\|/\|\mathbf{B}\|}\mathbf{Z}$
Output: $\mathbf{Y} = \mathbf{T}^{\mathbf{AB}}$, $\mathbf{Z} = \mathbf{T}^{\mathbf{BA}}$

In a gradient update, both the loss and the gradient of the metric are needed. A naive computation of $\mathfrak{B}^2(\mathbf{A}, \mathbf{B})$, $\nabla_{\mathbf{A}}\mathfrak{B}^2(\mathbf{A}, \mathbf{B})$ and $\nabla_{\mathbf{B}}\mathfrak{B}^2(\mathbf{A}, \mathbf{B})$ would require the knowledge of 6 roots:

$$\mathbf{A}^{\frac{1}{2}}, \mathbf{B}^{\frac{1}{2}}, (\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}})^{\frac{1}{2}}, (\mathbf{B}^{\frac{1}{2}}\mathbf{A}\mathbf{B}^{\frac{1}{2}})^{\frac{1}{2}}, \mathbf{A}^{-\frac{1}{2}}, \text{ and } \mathbf{B}^{-\frac{1}{2}},$$

to compute the following transport maps:

$$\mathbf{T}^{\mathbf{AB}} = \mathbf{A}^{-\frac{1}{2}}(\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}})^{\frac{1}{2}}\mathbf{A}^{-\frac{1}{2}}, \qquad \mathbf{T}^{\mathbf{BA}} = \mathbf{B}^{-\frac{1}{2}}(\mathbf{B}^{\frac{1}{2}}\mathbf{A}\mathbf{B}^{\frac{1}{2}})^{\frac{1}{2}}\mathbf{B}^{-\frac{1}{2}},$$

namely four matrix roots and two matrix inverse roots, which can be computed using SVD or Algorithm 1. However, we can avoid computing those six matrices using Algorithm 2, i.e. Newton-Schulz iterations with a different initialization, which directly yields $\mathbf{T}^{\mathbf{AB}}$ and $\mathbf{T}^{\mathbf{BA}}$ [Higham, 2008, §5.3]. From there, Bures distances and gradients can directly be computed using (2.8) and (2.7).

When computing the gradients of $n \times m$ squared Wasserstein distances $W_2^2(\alpha_i, \beta_j)$ in parallel, one only needs to run Algorithm 2 $n \times m$ times (in parallel) to compute matrices $(\mathbf{T}^{\mathbf{A}_i\mathbf{B}_j}, \mathbf{T}^{\mathbf{B}_j\mathbf{A}_i})_{i \leq n, j \leq m}$. On the other hand, using an automatic differentiation framework would require an additional backward computation of the same complexity as the forward pass, hence requiring roughly twice as many operations per batch.

Avoiding rank deficiency at optimization time. Although $\mathfrak{B}^2(\mathbf{A}, \mathbf{B})$ is defined for rank-deficient matrices $\mathbf{A}$ and $\mathbf{B}$, it is not differentiable with respect to these matrices

if they are rank-deficient. Indeed, as mentioned earlier, this can be compared to the non-differentiability of the Hellinger metric, $(\sqrt{x} - \sqrt{y})^2$ when x or y becomes 0, at which point if becomes *not* differentiable. If $\mathrm{Im}\mathbf{B} \not\subset \mathrm{Im}\mathbf{A}$, which is notably the case if $\mathrm{rk}\mathbf{B} > \mathrm{rk}\mathbf{A}$, then $\nabla_{\mathbf{A}} \mathfrak{B}^2(\mathbf{A}, \mathbf{B})$ no longer exists. However, even in that case, $\nabla_{\mathbf{B}} \mathfrak{B}^2(\mathbf{A}, \mathbf{B})$ exists iff $\mathrm{Im}\mathbf{A} \subset \mathrm{Im}\mathbf{B}$. Since it would be cumbersome to account for these subtleties in a large scale optimization setting, we propose to add a small common regularization term to all the factor products considered for our embeddings, and set $\mathbf{A}_\varepsilon = \mathbf{L}\mathbf{L}^T + \varepsilon \mathbf{I}$ were $\varepsilon > 0$ is a hyperparameter. This ensures that all matrices are full rank, and thus that all gradients exist. Most importantly, all our derivations still hold with this regularization, and can be shown to leave the method to compute the gradients w.r.t $\mathbf{L}$ unchanged, namely remain equal to $\left(\mathbf{I} - \mathbf{T}^{\mathbf{A}_\varepsilon \mathbf{B}}\right) \mathbf{L}$.

4 Experiments

We discuss in this section several applications of elliptical embeddings. We first consider a simple mMDS type visualization task, in which elliptical distributions in $d = 2$ are used to embed isometrically points in high dimension. We argue that for such purposes, a more natural way to visualize ellipses is to use their precision matrices. This is due to the fact that the human eye somewhat acts in the opposite direction to the Bures metric, as discussed in Figure 2.3. We follow with more advanced experiments in which we consider the task of computing word embeddings on large corpora as a testing ground, and equal or improve on the state-of-the-art.

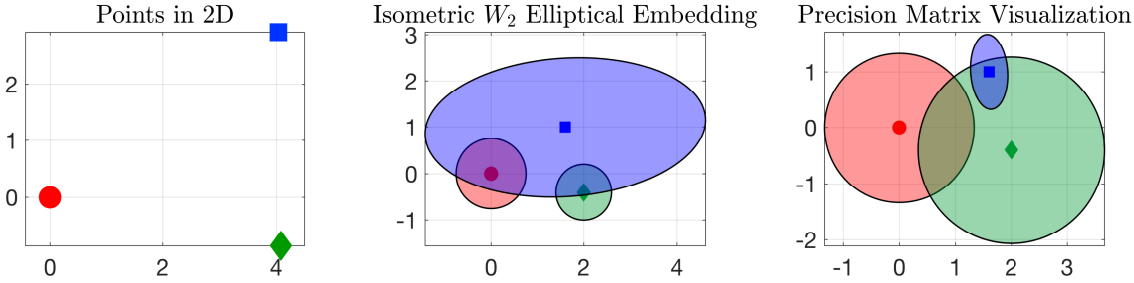


Figure 2.3: (left) three points on the plane. (middle) *isometric* elliptic embedding with the Bures metric: ellipses of a given color have the same respective distances as points on the left. Although the mechanics of optimal transport indicate that the blue ellipsoid is far from the two others, in agreement with the left plot, the human eye tends to focus on those areas that overlap (below the ellipsoid center) rather than those far away areas (north-east area) that contribute more significantly to the W_2 distance. (right) the precision matrix visualization, obtained by considering ellipses with the same axes but inverted eigenvalues, agree better with intuition, since they emphasize that overlap and extension of the ellipse means on the contrary that those axis contribute less to the increase of the metric.

Visualizing datasets using ellipsoids. Multidimensional scaling [De Leeuw, 1977] aims at embedding points $\mathbf{x}_1, \ldots, \mathbf{x}_n$ in a finite metric space in a lower dimensional one by minimizing the *stress* $\sum_{ij} (\|\mathbf{x}_i - \mathbf{x}_j\| - \|\mathbf{y}_i - \mathbf{y}_j\|)^2$. In our case, this translates to the minimization of $\mathcal{L}_{\mathrm{MDS}}(\mathbf{a}_1, \ldots \mathbf{a}_n, \mathbf{A}_1, \ldots, \mathbf{A}_n) = \sum_{ij} (\|\mathbf{x}_i - \mathbf{x}_j\| - W_2(\mu_{\mathbf{a}_i, \mathbf{A}_i}, \mu_{\mathbf{a}_j, \mathbf{A}_j}))^2$. This objective can be crudely minimized with a simple gradient descent approach operating on factors as advocated in Section 3, as illustrated in a toy example carried out using data from OECD's PISA study[2].

[2] `http://pisadataexplorer.oecd.org/ide/idepisa/`

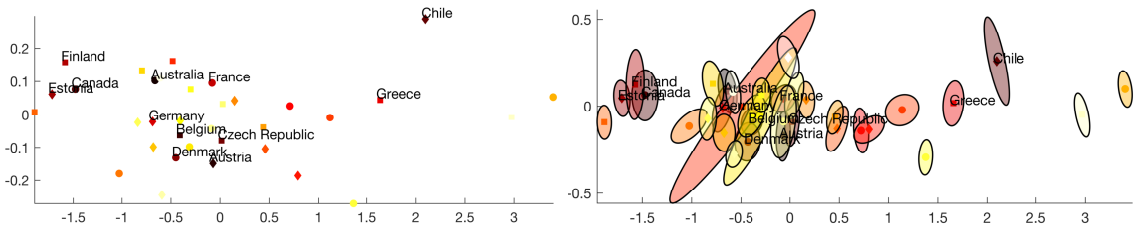


Figure 2.4: Toy experiment: visualization of a dataset of 10 PISA scores for 35 countries in the OECD. (left) MDS embeddings of these countries on the plane (right) elliptical embeddings on the plane using the precision visualization discussed in Figure 2.3. The normalized stress with standard MDS is 0.62. The stress with elliptical embeddings is close to $5e-3$ after 1000 gradient iterations, with random initializations for scale matrices (following a Standard Wishart with 4 degrees of freedom) and initial means located on the MDS solution.

Word embeddings. The skipgram model [Mikolov et al., 2013a] computes word embeddings in a vector space by maximizing the log-probability of observing surrounding context words given an input central word. Vilnis and McCallum [2015] extended this approach to *diagonal* Gaussian embeddings using an energy whose overall principles we adopt here, adapted to elliptical distributions with *full* covariance matrices in the 2-Wasserstein space. For every word w, we consider an input (as a word) and an ouput (as a context) representation as an elliptical measure, denoted respectively μ_w and ν_w, both parameterized by a location vector and a scale parameter (stored in factor form). Given a set $\mathcal{R}$ of positive

Table 2.1: Results for elliptical embeddings (evaluated using our cosine mixture) compared to diagonal Gaussian embeddings trained with the `seomoz` package (evaluated using expected likelihood cosine similarity as recommended by Vilnis and McCallum).

Dataset	W2G/45/C	Ell/12/CM
SimLex	**25.09**	24.09
WordSim	53.45	**66.02**
WordSim-R	61.70	**71.07**
WordSim-S	48.99	**60.58**
MEN	65.16	**65.58**
MC	59.48	**65.95**
RG	**69.77**	65.58
YP	**37.18**	25.14
MT-287	**61.72**	59.53
MT-771	**57.63**	56.78
RW	**40.14**	29.04

word/context pairs of words (w, c), and for each input word a set $N(w)$ of n negative contexts words sampled randomly, we adapt Vilnis and McCallum's loss function to the W_2^2 distance to minimize the following hinge loss:

$$\sum_{(w,c)\in\mathcal{R}} \left[M - [\mu_w : \nu_c] + \frac{1}{n} \sum_{c'\in N(w)} [\mu_w : \nu_{c'}] \right]_+ ,$$

where $M > 0$ is a margin parameter. We train our embeddings on the concatenated ukWaC and WaCkypedia corpora [Baroni et al., 2009], consisting of about 3 billion tokens, on which we keep only the tokens appearing more than 100 times in the text (for a total

number of 261583 different words). We train our embeddings using adagrad [Duchi et al., 2011], sampling one negative context per positive context and, in order to prevent the norms of the embeddings to be too highly correlated with the corresponding word frequencies (see Figure in supplementary material), we use two distinct sets of embeddings for the input and context words.

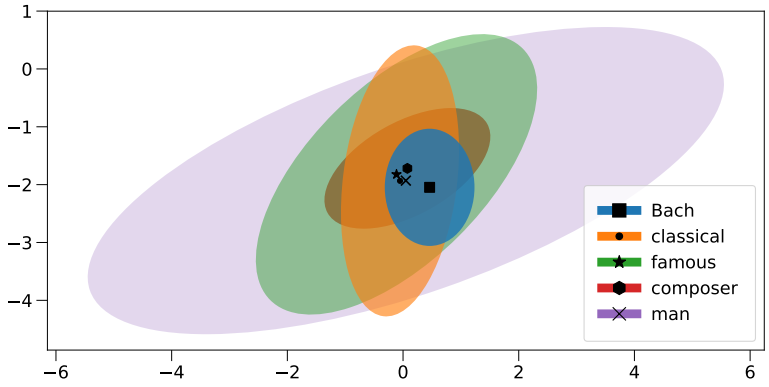


Figure 2.5: Precision matrix visualization of trained embeddings of a set of words on the plane spanned by the two principal eigenvectors of the covariance matrix of "Bach".

We compare our full elliptical to diagonal Gaussian embeddings trained using the methods described in [Vilnis and McCallum, 2015] on a collection of similarity datasets by computing the Spearman rank correlation between the similarity scores provided in the data and the scores we compute based on our embeddings. Note that these results are obtained using context (ν_w) rather than input (μ_w) embeddings. For a fair comparison across methods, we set dimensions by ensuring that the number of free parameters remains the same: because of the symmetry in the covariance matrix, elliptical embeddings in dimension d have $d + d(d+1)/2$ free parameters (d for the means, $d(d+1)/2$ for the covariance matrices), as compared with $2d$ for diagonal Gaussians. For elliptical embeddings, we use the common practice of using some form of normalized quantity (a cosine) rather than the direct dot product. We implement this here by computing the mean of two cosine terms, each corresponding separately to mean and covariance contributions:

$$\mathfrak{S}_{\mathfrak{B}}[\mu_{\mathbf{a},\mathbf{A}}, \mu_{\mathbf{b},\mathbf{B}}] \overset{\text{def}}{=} \frac{\langle \mathbf{a}, \mathbf{b} \rangle}{\|\mathbf{a}\|\|\mathbf{b}\|} + \frac{\mathrm{Tr}(\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}})^{\frac{1}{2}}}{\sqrt{\mathrm{Tr}\mathbf{A}\mathrm{Tr}\mathbf{B}}}.$$

Using this similarity measure rather than the Wasserstein-Bures dot product is motivated by the fact that the norms of the embeddings show some dependency with word frequencies (see figures in supplementary) and become dominant when comparing words with different frequencies scales. An alternative could have been obtained by normalizing the Wasserstein-Bures dot product in a more standard way that pools together means and covariances. However, as discussed in the supplementary material, this choice makes it harder to deal with the variations in scale of the means and covariances, therefore decreasing performance. We also evaluate our embeddings on the Entailment dataset ([Baroni et al., 2012]), on which we obtain results roughly comparable to those of [Vilnis and McCallum, 2015]. Note that contrary to the similarity experiments, in this framework using the (unsymmetrical) KL divergence makes sense and possibly gives an advantage, as it is possible to choose the order of the arguments in the KL divergence between the entailing and entailed words.

Hypernymy. In this experiment, we use the framework of [Nickel and Kiela, 2017] on hypernymy relationships to test our embeddings. A word A is said to be a *hypernym* of

Table 2.2: Entailment benchmark: we evaluate our embeddings on the Entailment dataset using average precision (AP) and F1 scores. The threshold for F1 is chosen to be the best at test time.

Model	AP	F1
W2G/45/Cosine	0.70	0.74
W2G/45/KL	0.72	0.74
Ell/12/CM	0.70	0.73

a word B if any B is a type of A, e.g. any *dog* is a type of *mammal*, thus constituting a tree-like structure on nouns. The `WORDNET` dataset [Miller, 1995] features a transitive closure of 743,241 hypernymy relations on 82,115 distinct nouns, which we consider as an undirected graph of relations $\mathcal{R}$. Similarly to the skipgram model, for each noun u we sample a fixed number n of negative examples and store them in set $\mathcal{N}(u)$ to optimize the following loss:

$$\sum_{(u,v)\in\mathcal{R}} \log \frac{e^{[\mu_u,\mu_v]}}{e^{[\mu_u,\mu_v]} + \sum_{v'\in\mathcal{N}(u)} e^{[\mu_u,\mu_{v'}]}}.$$

We train the model using SGD with only one set of embeddings. The embeddings are then

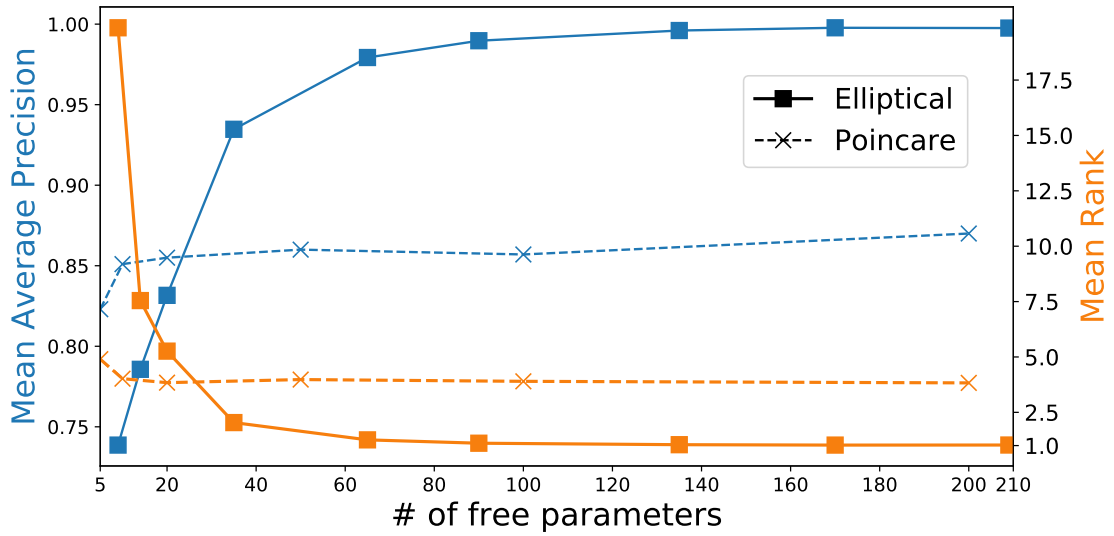


Figure 2.6: Reconstruction performance of our embeddings against Poincare embeddings (reported from [Nickel and Kiela, 2017], as we were not able to reproduce scores comparable to these values) evaluated by mean retrieved rank (lower=better) and MAP (higher=better).

evaluated on a link reconstruction task: we embed the full tree and rank the similarity of each positive hypernym pair (u, v) among all negative pairs (u, v') and compute the mean rank thus achieved as well as the mean average precision (MAP), using the Wasserstein-Bures dot product as the similarity measure. Elliptical embeddings consistently outperform Poincare embeddings for dimensions above a small threshold, as shown in Figure 2.6, which confirms our intuition that the addition of a notion of variance or uncertainty to point embeddings allows for a richer and more significant representation of words.

4.1 Model Hyperparameters and Training Details

Word Embeddings. We train our embeddings on the concatenated ukWaC and WaCkypedia corpora [Baroni et al., 2009], consisting of about 3 billion tokens, on which we keep

only the tokens appearing more than 100 times in the text after lowercasing and removal of all punctuation (for a total number of 261583 different words). We optimize 5 epoches using adagrad [Duchi et al., 2011] with $\epsilon = 10^{-8}$ with a learning rate of 0.01. We use a window size of 10 (i.e. positive examples consist of the first 5 preceding and first 5 succeeding words), set the margin to 10, sample one negative context per positive context and, in order to prevent the norms of the embeddings to be too highly correlated with the corresponding word frequencies (see Figure 2.7), we use two distinct sets of embeddings for the input and context words. In order to use as much parallelization as possible, we use batches of size 10000, but believe that smaller batches would lead to improved performances. We limit matrix square root approximations to 6 Newton-Schulz iterations and add $0.01\mathbf{I}_\mathrm{d}$ to the covariances to ensure non-singularity.

To generate batches, we use the same sampling tricks as in [Mikolov et al., 2013b], namely subsampling the frequent terms (using a threshold of 10^{-5} as recommended for large datasets) and smoothing the negative distribution by using probabilities $\{f_i^{3/4}/Z\}$ where f_i is the frequency of word i for sampling negative contexts $\{c'_i\}$.

We then evaluate our embeddings on the following datasets:

- Simlex [Hill et al., 2015],
- WordSim [Finkelstein et al., 2002],
- MEN [Bruni et al., 2014],
- MC [Miller and Charles, 1991],
- RG [Rubenstein and Goodenough, 1965],
- YP [Yang and Powers, 2005],
- MTurk [Radinsky et al., 2011, Halawi et al., 2012],
- RW [thang Luong et al., 2013],

using the context embeddings and the Wasserstein-Bures cosine as a similarity measure.

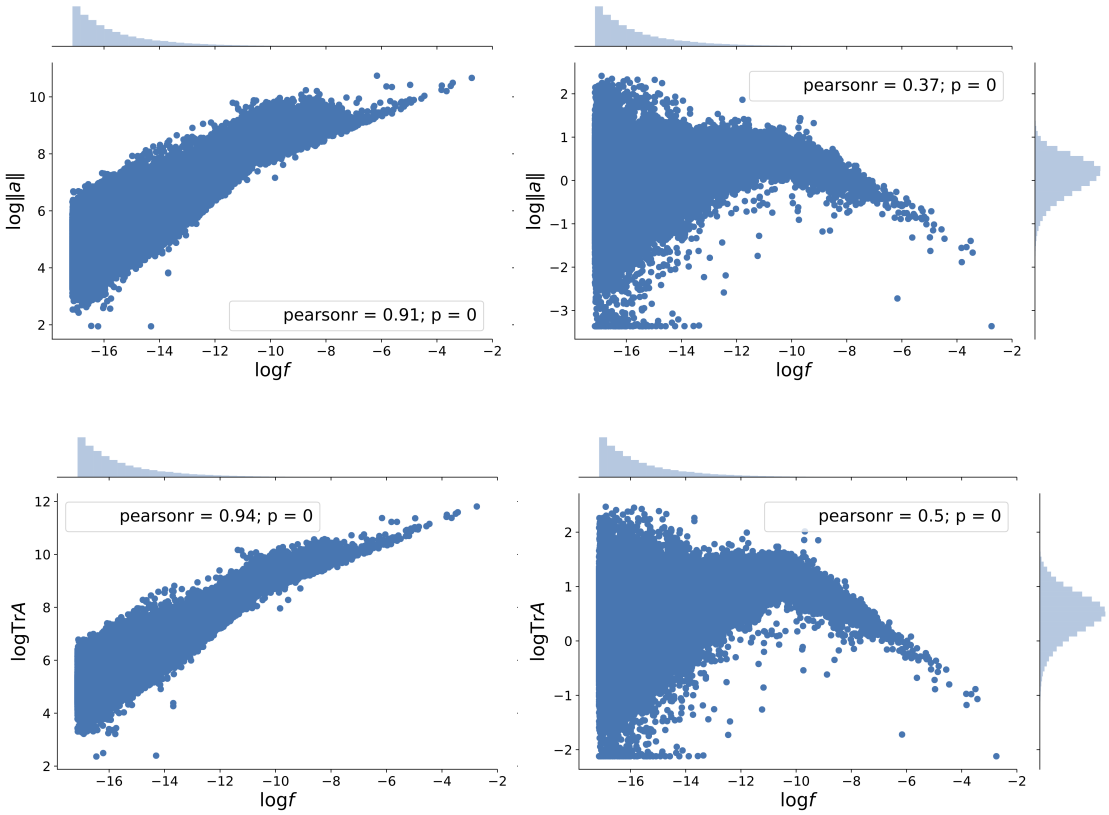
Hypernymy. We train our embeddings on the transitive closure of the `WORDNET` dataset [Miller, 1995] which features 743,241 hypernymy relations on 82,115 distinct nouns. For disambiguation, note that if (u, v) is a hypernymy relation with $u \neq v$, (v, u) is in general *not* a positive relation, but (u, u) is as a noun is always its own hypernym.

We perform our optimization using SGD with batches of 1000 relations, a learning rate 0.02 for dimensions 3 and 4 and 0.01 for higher dimensions, sample 50 negative examples per positive relation, use 6 square root iterations and add $0.01\mathbf{I}_\mathrm{d}$ to the covariances. Contrary to the skipgram experiment, we use a single set of embeddings and use the Wasserstein-Bures dot product as a similarity measure.

4.2 The Wasserstein-Bures cosine

As discussed in Section 4, a natural choice of similarity measure would be the Wasserstein-Bures cosine, obtained by normalizing the Wasserstein-Bures dot product with the means' norms and covariances' root traces jointly:

$$\cos_{\mathfrak{B}}[\mu_{\mathbf{a},\mathbf{A}}, \mu_{\mathbf{b},\mathbf{B}}] \stackrel{\text{def}}{=} \frac{\langle \mathbf{a}, \mathbf{b} \rangle + \mathrm{Tr}[\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}}]^{\frac{1}{2}}}{(\|\mathbf{a}\|^2 + \mathrm{Tr}\mathbf{A})^{\frac{1}{2}}(\|\mathbf{b}\|^2 + \mathrm{Tr}\mathbf{B})^{\frac{1}{2}}}.$$



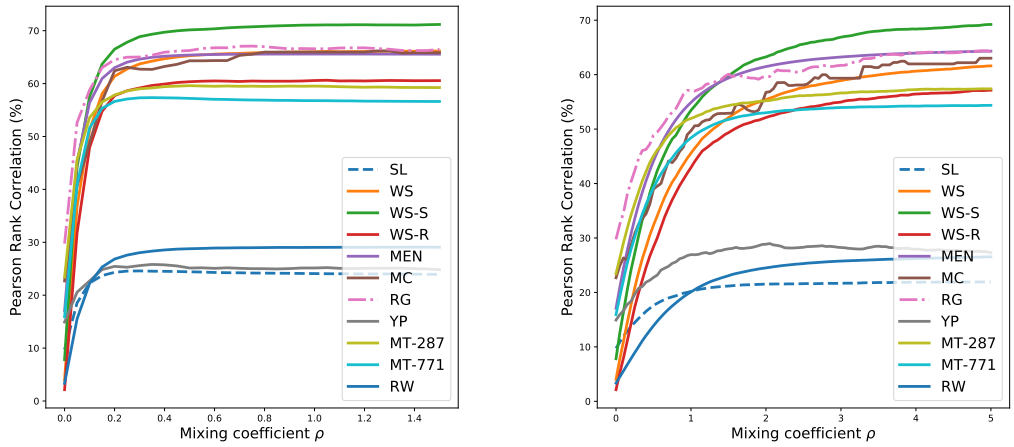
(inputs) (contexts)

Figure 2.7: log-log plot of the norms (top) and traces (bottom) of the embeddings' means vs. word frequency: the sizes of the input embeddings (left) follow a power law, whereas context embeddings (right) give less importance to very frequent words and emphasize on medium frequency words.

However, we have found that in some applications (and notably in our skipgram experiments) such a joint normalization can result in either the means or the covariances to have a negligible contribution if the scales of the parameters differ too much. To circumvent this problem, we introduce another similarity measure, which is a mixture of two cosine terms:

$$\mathfrak{S}_{\mathfrak{B}}[\mu_{\mathbf{a},\mathbf{A}}, \mu_{\mathbf{b},\mathbf{B}}] := \frac{\langle \mathbf{a}, \mathbf{b} \rangle}{\|\mathbf{a}\|\|\mathbf{b}\|} + \frac{\mathrm{Tr}[\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}}]^{\frac{1}{2}}}{\sqrt{\mathrm{Tr}\mathbf{A}\mathrm{Tr}\mathbf{B}}}.$$

This latter similarity measure allows to gather information from the means and the covariances independently. Note that while the term corresponding to the covariances is obtained in a cosine-like normalization, it takes values between 0 and 1 as it only involve traces of PSD matrices, whereas the means term is a regular Euclidean cosine and therefore takes values between -1 and 1. We compare the behaviors of these two measures on the word similarity evaluation task by introducing a mixing coefficient ρ, and defining



(a) $\mathfrak{S}_{\mathfrak{B}}$ (b) $\cos_{\mathfrak{B}}$

Figure 2.8: Pearson rank correlation scores on similarity benchmarks as a function of the mixing coefficient: $\mathfrak{S}_{\mathfrak{B}}$ smoothly attains a maximum in performance around $\rho = 1$, whereas $\cos_{\mathfrak{B}}$ has a less regular behavior.

$$\cos_{\mathfrak{B}}[\mu_{\mathbf{a},\mathbf{A}}, \mu_{\mathbf{b},\mathbf{B}}; \rho] \stackrel{\text{def}}{=} \frac{\langle \mathbf{a}, \mathbf{b} \rangle + \rho \mathrm{Tr}[\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}}]^{\frac{1}{2}}}{(\|\mathbf{a}\|^2 + \rho \mathrm{Tr}\mathbf{A})^{\frac{1}{2}}(\|\mathbf{b}\|^2 + \rho \mathrm{Tr}\mathbf{B})^{\frac{1}{2}}}$$

$$\mathfrak{S}_{\mathfrak{B}}[\mu_{\mathbf{a},\mathbf{A}}, \mu_{\mathbf{b},\mathbf{B}}; \rho] \stackrel{\text{def}}{=} \frac{\langle \mathbf{a}, \mathbf{b} \rangle}{\|\mathbf{a}\|\|\mathbf{b}\|} + \rho \frac{\mathrm{Tr}[\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}}]^{\frac{1}{2}}}{\sqrt{\mathrm{Tr}\mathbf{A}\mathrm{Tr}\mathbf{B}}}.$$

As can be seen from Figure 2.8, the Wasserstein-Bures cosine is less well behaved and makes it difficult to find an optimal mixing value. On the other hand, the mixture of cosines similarity measure varies more smoothly and seems to reach a performance maximum around $\rho = 1$, and achieves better performance than the Wasserstein-Bures cosine on most datasets.

Conclusion

We have proposed to use the space of elliptical distributions endowed with the W_2 metric to embed complex objects. This latest iteration of probabilistic embeddings, in which a point an object is represented as a probability measure, can consider elliptical measures (including Gaussians) with arbitrary covariance matrices. Using the W_2 metric we can provides a natural and seamless generalization of point embeddings in $\mathbb{R}^d$. Each embedding is described with a location $\mathbf{c}$ and a scale $\mathbf{C}$ parameter, the latter being represented in practice using a factor matrix $\mathbf{L}$, where $\mathbf{C}$ is recovered as $\mathbf{L}\mathbf{L}^T$. The visualization part of work is still subject to open questions. One may seek a different method than that proposed here using precision matrices, and ask whether one can include more advanced constraints on these embeddings, such as inclusions or the presence (or absence) of intersections across ellipses. Handling multimodality using mixtures of Gaussians could be pursued. In that case a natural upper bound on the W_2 distance can be computed by solving the OT problem between these mixtures of Gaussians using a simpler proxy: consider them as discrete measures putting Dirac masses in the space of Gaussians endowed with the W_2

metric as a ground cost, and use the optimal cost of that proxy as an upper bound of their Wasserstein distance. Finally, note that the set of elliptical measures $\mu_{\mathbf{c},\mathbf{C}}$ endowed with the Bures metric can also be interpreted, given that $\mathbf{C} = \mathbf{L}\mathbf{L}^T, \mathbf{L} \in \mathbb{R}^{d\times k}$, and writing $\tilde{\mathbf{l}}_i = \mathbf{l}_i - \bar{\mathbf{l}}$ for the centered column vectors of $\mathbf{L}$, as a discrete point cloud $(\mathbf{c} + \frac{1}{\sqrt{k}}\tilde{\mathbf{l}}_i)_i$ endowed with a W_2 metric only looking at their first and second order moments. These k points, whose mean and covariance matrix match $\mathbf{c}$ and $\mathbf{C}$, can therefore fully characterize the geometric properties of the distribution $\mu_{\mathbf{c},\mathbf{C}}$, and may provide a simple form of multimodal embedding.

5 Appendix: Derivation of the Euclidean Gradient of the Bures metric

Let $\otimes$ denote the Kronecker product of matrices. Recall [see Fackler, 2005] that

$$[\mathbf{B}^\top \otimes \mathbf{A}]\mathrm{vec}(\mathbf{X}) = \mathrm{vec}(\mathbf{AXB}) \quad \text{and} \quad [\mathbf{A}\otimes\mathbf{B}][\mathbf{C}\otimes\mathbf{D}] = [\mathbf{AC}\otimes\mathbf{BD}].$$

In the following, we will often omit the $\mathrm{vec}(\cdot)$ and treat matrices as vectors when the context makes it clear. We will make use of the following identities:

$$\begin{aligned}\partial_{\mathbf{X}}\ f\circ g(\mathbf{X}) &= \partial_{\mathbf{X}} f(g(\mathbf{X}))\partial_{\mathbf{X}} g(\mathbf{X})\\ \partial_{\mathbf{X}}\ (fg)(\mathbf{X}) &= [g(\mathbf{X})^\top \otimes \mathbf{I}_\mathrm{d}]\partial_{\mathbf{X}} f(\mathbf{X}) + [\mathbf{I}_\mathrm{d}\otimes g(\mathbf{X})]\partial_{\mathbf{X}} g(\mathbf{X}).\end{aligned}$$

Likewise, we will write the solution $\mathcal{L}_{\mathbf{A}}(\mathbf{B})$ of the Lyapunov equation $\mathbf{XA}+\mathbf{AX}=\mathbf{B}$ using Kronecker notations:

$$\begin{aligned}\partial_{\mathbf{X}} X^{1/2}[\mathbf{H}] &= L_{\mathbf{X}^{1/2}}(\mathbf{H})\\ &= [\mathbf{X}^{1/2}\otimes\mathbf{I}_\mathrm{d} + \mathbf{I}_\mathrm{d}\otimes\mathbf{X}^{1/2}]^{-1}\mathbf{H}.\end{aligned}$$

Gradient of $\mathfrak{B}^2(\mathbf{A},\mathbf{B})$. Let $F(\mathbf{A},\mathbf{B}) = \mathrm{Tr}(\mathbf{B}^{\frac12}\mathbf{AB}^{\frac12})^{\frac12}$ denote the fidelity, let us differentiate F w.r.t $\mathbf{A}$ for the Frobenius inner product:

$$\begin{aligned}\nabla_{\mathbf{A}} F(\mathbf{A},\mathbf{B}) &= \left[\partial_{\mathbf{A}}(\mathbf{B}^{\frac12}\mathbf{AB}^{\frac12})^{\frac12}\right]^\top \mathbf{I}_\mathrm{d}\\ &= \left[\left[\mathbf{B}^{\frac12}\mathbf{AB}^{\frac12})^{\frac12}\otimes\mathbf{I}_\mathrm{d} + \mathbf{I}_\mathrm{d}\otimes(\mathbf{B}^{\frac12}\mathbf{AB}^{\frac12})^{\frac12}\right]^{-1}\partial_{\mathbf{A}}(\mathbf{B}^{\frac12}\mathbf{AB}^{\frac12})\right]^\top \mathbf{I}_\mathrm{d}\\ &= \left[\mathbf{B}^{\frac12}\otimes\mathbf{B}^{\frac12}\right]\left[(\mathbf{B}^{\frac12}\mathbf{AB}^{\frac12})^{\frac12}\otimes\mathbf{I}_\mathrm{d} + \mathbf{I}_\mathrm{d}\otimes(\mathbf{B}^{\frac12}\mathbf{AB}^{\frac12})^{\frac12}\right]^{-1}\mathbf{I}_\mathrm{d}\\ &= \left[\mathbf{B}^{\frac12}\otimes\mathbf{B}^{\frac12}\right]\frac12(\mathbf{B}^{\frac12}\mathbf{AB}^{\frac12})^{-\frac12}\\ &= \frac12\mathbf{B}^{\frac12}(\mathbf{B}^{\frac12}\mathbf{AB}^{\frac12})^{-\frac12}\mathbf{B}^{\frac12}\\ &= \frac12\mathbf{T}^{\mathbf{AB}},\end{aligned}$$

where the fourth line comes from the fact that $\forall\mathbf{A}\in S^d_{++}, \mathcal{L}_{\mathbf{A}}(\mathbf{I}_\mathrm{d}) = \frac12\mathbf{A}^{-1/2}$.

Gradient of $\mathfrak{B}^2(\mathbf{LL}^T,\mathbf{B})$. Let now $\mathbf{A}=\mathbf{LL}^\top$, let us differentiate w.r.t $\mathbf{L}$:

$$\begin{aligned}\nabla_{\mathbf{L}} f(\mathbf{LL}^\top,\mathbf{B}) &= \left[\partial_{\mathbf{L}}(\mathbf{B}^{\frac12}\mathbf{AB}^{\frac12})^{\frac12}\right]^\top \mathbf{I}_\mathrm{d}\\ &= \partial_{\mathbf{L}}\mathbf{A}^\top\left[\partial_{\mathbf{A}}(\mathbf{B}^{\frac12}\mathbf{AB}^{\frac12})^{\frac12}\right]^\top \mathbf{I}_\mathrm{d}\\ &= \left[\mathbf{L}^\top\otimes\mathbf{I}_\mathrm{d}\right]\left[\mathbf{I}_\mathrm{d}+\mathbf{T}_{n,n}\right]\frac12\mathbf{B}^{\frac12}(\mathbf{B}^{\frac12}\mathbf{AB}^{\frac12})^{-\frac12}\mathbf{B}^{\frac12}\\ &= \mathbf{B}^{\frac12}(\mathbf{B}^{\frac12}\mathbf{AB}^{\frac12})^{-\frac12}\mathbf{B}^{\frac12}\mathbf{L}\\ &= \mathbf{T}^{\mathbf{AB}}\mathbf{L},\end{aligned}$$

where $\mathbf{T}_{n,n}$ is the transposition tensor, such that $\forall\mathbf{X}\in\mathbb{R}^{n\times n}, \mathbf{T}_{n,n}\mathrm{vec}(\mathbf{X}) = \mathrm{vec}(\mathbf{X}^\top)$.

Finally, using the same calculations and the fact that $\partial_{\mathbf{L}}\left[\mathbf{LL}^\top+\varepsilon\mathbf{I}_\mathrm{d}\right] = \partial_{\mathbf{L}}\left[\mathbf{LL}^\top\right]$, one can see that if $\mathbf{A}=\mathbf{LL}^\top+\varepsilon\mathbf{I}_\mathrm{d}$, then we still have

$$\nabla_{\mathbf{L}} F(\mathbf{LL}^\top+\varepsilon\mathbf{I}_\mathrm{d},\mathbf{B}) = \mathbf{T}^{\mathbf{AB}}\mathbf{L}.$$

Chapter 3

Building Optimal Transport Plans on Subspace Projections

Computing optimal transport (OT) between measures in high dimensions is doomed by the curse of dimensionality. A popular approach to avoid this curse is to project input measures on lower-dimensional subspaces (1D lines in the case of sliced Wasserstein distances), solve the OT problem between these reduced measures, and settle for the Wasserstein distance between these reductions, rather than that between the original measures. This approach is however difficult to extend to the case in which one wants to compute an OT map (a Monge map) between the original measures. Since computations are carried out on lower-dimensional projections, classical map estimation techniques can only produce maps operating in these reduced dimensions. We propose in this work two methods to extrapolate, from an transport map that is optimal on a subspace, one that is nearly optimal in the entire space. We prove that the best optimal transport plan that takes such "subspace detours" is a generalization of the Knothe-Rosenblatt transport. We show that these plans can be explicitly formulated when comparing Gaussian measures (between which the Wasserstein distance is commonly referred to as the Bures or Fréchet distance). We provide an algorithm to select optimal subspaces given pairs of Gaussian measures, and study scenarios in which that mediating subspace can be selected using prior information. We consider applications to semantic mediation between elliptical word embeddings and domain adaptation with Gaussian mixture models.

This chapter is based on [Muzellec and Cuturi, 2019].

1 Introduction

Minimizing the transport cost between two probability distributions [Villani, 2008] results in two useful quantities: the minimum cost itself, often cast as a loss or a metric (the Wasserstein distance), and the minimizing solution, a function known as the Monge map [Monge, 1781] that pushes forward the first measure onto the second with least expected cost. While the former has long attracted the attention of the machine learning community, the latter is playing an increasingly important role in data sciences. Indeed, important problems such as domain adaptation [Courty et al., 2014], generative modelling [Goodfellow et al., 2014, Arjovsky et al., 2017, Genevay et al., 2018], reconstruction of cell trajectories in biology Schiebinger et al. [2019] and auto-encoders [Kingma and Welling, 2014, Tolstikhin et al., 2018] among others can be recast as the problem of finding a map, preferably optimal, which transforms a reference distribution into another. However, accurately estimating an OT map from data samples is a difficult problem, plagued by the well documented instability of OT in high-dimensional spaces [Dudley, 1969, Fournier and Guillin, 2015] and its high computational cost.

Optimal transport on subspaces. Several approaches, both in theory and in practice, aim at bridging this gap. Theory [Weed and Bach, 2017] supports the idea that sample complexity can be improved when the measures are supported on lower-dimensional manifolds of high-dimensional spaces. Practical insights [Cuturi, 2013] supported by theory [Genevay et al., 2019] advocate using regularizations to improve both computational and sample complexity. Some regularity in OT maps can also be encoded by looking at specific families of maps [Seguy et al., 2018, Paty et al., 2020]. Another trend relies on lower-dimensional projections of measures before computing OT. In particular, sliced Wasserstein (SW) distances [Bonneel et al., 2015] leverage the simplicity of OT between 1D measures to define distances and barycentres, by averaging the optimal transport between projections onto several random directions. This approach has been applied to alleviate training complexity in the GAN/VAE literature [Deshpande et al., 2018, Wu et al., 2019] and was generalized very recently in [Paty and Cuturi, 2019] who considered projections on k-dimensional subspaces that are adversarially selected. However, these subspace approaches only carry out half of the goal of OT: by design, they do result in more robust measures of OT costs, but they can only provide maps in subspaces that are optimal (or nearly so) between the *projected* measures, not transportation maps in the original, high-dimensional space in which the original measures live. For instance, the closest thing to a map one can obtain from using several SW univariate projections is an average of several permutations, which is not a map but a transport plan or coupling [Rowland et al., 2019][Rabin et al., 2011, p.6].

Our approach. Whereas the approaches cited above focus on OT maps and plans in projection subspaces only, we consider here plans and maps on the original space that are constrained to be optimal when projected on a given subspace E. This results in the definition of a class of transportation plans that figuratively need to make an optimal "detour" in E. We propose two constructions to recover such maps corresponding respectively (i) to the independent product between conditioned measures, and (ii) to the optimal conditioned map.

Chapter structure. After recalling background material on OT in Section 2, we introduce in Section 3 the class of *subspace-optimal* plans that satisfy projection constraints on a given subspace E. We characterize the degrees of freedom of E-optimal plans using their disintegrations on E and introduce two extremal instances: *Monge-Independent* plans,

which assume independence of the conditionals, and *Monge-Knothe* maps, in which the conditionals are optimally coupled. We give closed forms for the transport between Gaussian distributions in Section 4, respectively as a degenerate Gaussian distribution, and a linear map with block-triangular matrix representation. We provide guidelines and a minimizing algorithm for selecting a subspace E when it is not prescribed *a priori* in Section 5. Finally, in section 6 we showcase the behavior of MK and MI transports on (noisy) synthetic data, show how using a mediating subspace can be applied to selecting meanings for polysemous elliptical word embeddings, and experiment using MK maps with the minimizing algorithm on a domain adaptation task with Gaussian mixture models.

Notations. For E a linear subspace of $\mathbb{R}^d$, $E^\perp$ is its orthogonal complement, $\mathbf{V}_E \in \mathbb{R}^{d\times k}$ (resp. $\mathbf{V}_{E^\perp} \in \mathbb{R}^{d\times d-k}$) the matrix of orthonormal basis vectors of E (resp $E^\perp$). $p_E : x \to \mathbf{V}_E^\top x$ is the orthogonal projection operator onto E. $\mathcal{P}_2(\mathbb{R}^d)$ is the space of probability distributions over $\mathbb{R}^d$ with finite second moments. $\mathcal{B}(\mathbb{R}^d)$ is the Borel algebra over $\mathbb{R}^d$. $\rightharpoonup$ denotes the weak convergence of measures. $\otimes$ is the product of measures, and is used in measure disintegration by abuse of notation.

2 Reminders on Optimal Transport Plans, Maps and Disintegration of Measure

Let us start by recalling basic facts on Monge-Kantorovich optimal transport.

Kantorovich plans. For two probability measures $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, we refer to the set of couplings

$$\Pi(\mu,\nu) \stackrel{\text{def}}{=} \{\gamma \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d) : \forall A, B \in \mathcal{B}(\mathbb{R}^d), \gamma(A \times \mathbb{R}^d) = \mu(A), \gamma(\mathbb{R}^d \times B) = \nu(B)\}$$

as the set of *transportation plans* between μ, ν. The 2-Wasserstein distance between μ and ν is defined as

$$W_2^2(\mu,\nu) \stackrel{\text{def}}{=} \min_{\gamma\in\Pi(\mu,\nu)} \mathbb{E}_{(X,Y)\sim\gamma}\left[\|X-Y\|^2\right].$$

Conveniently, transportation problems with quadratic cost can be reduced to transportation between centered measures. Indeed, let $\mathbf{m}_\mu$ (resp. $\mathbf{m}_\nu$) denote first moment of μ (resp. ν). Then,

$$\forall\gamma \in \Pi(\mu,\nu), \mathbb{E}_{(X,Y)\sim\gamma}[\|X-Y\|^2] = \|\mathbf{m}_\mu - \mathbf{m}_\nu\|^2 + \mathbb{E}_{(X,Y)\sim\gamma}[\|(X-\mathbf{m}_\mu)-(Y-\mathbf{m}_\nu)\|^2].$$

Therefore, in the following all probability measures are assumed to be centered, unless stated otherwise.

Monge maps. For a Borel-measurable map T, the push-forward of μ by T is defined as the measure $T_\sharp\mu$ satisfying for all $A \in \mathcal{B}(\mathbb{R}^d), T_\sharp\mu(A) = \mu(T^{-1}(A))$. A map such that $T_\sharp\mu = \nu$ is called a *transportation map* from μ to ν. When an optimal transportation map exists, the Wasserstein distance can be written in the form of the Monge problem

$$W_2^2(\mu,\nu) = \min_{T:T_\sharp\mu=\nu} \mathbb{E}_{X\sim\mu}[\|X-T(X)\|^2]. \tag{3.1}$$

When it exists, the optimal transportation map $T^\star$ in the Monge problem is called the *Monge map* from μ to ν. It is then related to the optimal transportation plan $\gamma^\star$ by the relation $\gamma^\star = (\mathbf{I}_\mathrm{d}, T^\star)_\sharp\mu$. When μ and ν are absolutely continuous (a.c.), a Monge map always exists [Santambrogio, 2015, Theorem 1.22].

Global maps or plans that are locally optimal. Considering the projection operator on E, p_E, we write $\mu_E = (p_E)_\sharp \mu$ for the marginal distribution of μ on E. Suppose that we are given a Monge map S between the two projected measures μ_E and ν_E. One of the contributions of this chapter is to propose extensions of this map S as a transportation plan γ (resp. a new map T) whose projection $\gamma_E = (p_E, p_E)_\sharp \gamma$ on that subspace E coincides with the optimal transportation plan $(\mathbf{I}_{\mathrm{d}E}, S)_\sharp \mu_E$ (resp. $p_E \circ T = S \circ p_E$). Formally, the transports introduced in Section 3 only require that S be a transport map from μ_E to ν_E, but optimality is required in the closed forms given in section 4 for Gaussian distributions. In either case, this constraint implies that γ is built "assuming that" it is equal to $(\mathbf{I}_{\mathrm{d}E}, S)_\sharp \mu_E$ on E. This is rigorously defined using the notion of measure disintegration.

Disintegration of measures. The disintegration of μ on a subspace E is the collection of measures $(\mu_{x_E})_{x_E \in E}$ supported on the fibers $\{x_E\} \times E^\perp$ such that any test function ϕ can be integrated against μ as $\int_{\mathbb{R}^d} \phi \mathrm{d}\mu = \int_E \left(\int_{E^\perp} \phi(y) \mathrm{d}\mu_{x_E}(y) \right) \mathrm{d}\mu_E(x_E)$. In particular, if $X \sim \mu$, then the law of X given x_E is μ_{x_E}. By abuse of the measure product notation $\otimes$, measure disintegration is denoted as $\mu = \mu_{x_E} \otimes \mu_E$. A more general description of disintegration can be found in [Ambrosio et al., 2006, Ch. 5.5].

3 Lifting Transport from Subspaces to the Full Space

Given two distributions $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, it is often easier to compute a Monge map S between their marginals μ_E, ν_E on a k-dimensional subspace E rather than in the whole space $\mathbb{R}^d$. When $k = 1$, this fact is at the heart of sliced wasserstein approaches [Bonneel et al., 2015], which have recently sparked interest in the GAN/VAE literature [Deshpande et al., 2018, Wu et al., 2019]. However, when $k < d$, there is in general no straightforward way of extending S to a transportation map or plan between μ and ν. In this section, we prove the existence of such extensions and characterize them.

Subspace-optimal plans. A transportation plan between μ_E and ν_E is a coupling living in $\mathcal{P}(E \times E)$. In general, it cannot be cast directly as a transportation plan between μ and ν taking values in $\mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$. However, the existence of such a "lifted" plan is given by the following result, which is used in OT theory to prove that W_p is a metric:

Lemma 3.1 (The Gluing Lemma, Villani [2008])**.** *Let $\mu_1, \mu_2, \mu_3 \in \mathcal{P}(\mathbb{R}^d)$. If γ_{12} is a coupling of (μ_1, μ_2) and γ_{23} is a coupling of (μ_2, μ_3), then one can construct a triple of random variables (Z_1, Z_2, Z_3) such that $(Z_1, Z_2) \sim \gamma_{12}$ and $(Z_2, Z_3) \sim \gamma_{23}$.*

By extension of the lemma, if we define (i) a coupling between μ and μ_E, (ii) a coupling between ν and ν_E, and (iii) the optimal coupling between μ_E and ν_E, $(\mathbf{I}_{\mathrm{d}}, S)_\sharp \mu_E$ (where S stands for the Monge map from μ_E to ν_E), we get the existence of four random variables (with laws μ, μ_E, ν and ν_E) which follow the desired joint laws. However, the lemma does not imply the uniqueness of those random variables, nor does it give a closed form for the corresponding coupling between μ and ν.

Definition 3.2 (Subspace-Optimal Plans)**.** *Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ and E be a k-dimensional subspace of $\mathbb{R}^d$. Let S be a Monge map from μ_E to ν_E. We define the set of* E-optimal plans *between μ and ν as $\Pi_E(\mu, \nu) \stackrel{\text{def}}{=} \{\gamma \in \Pi(\mu, \nu) : \gamma_E = (\mathbf{I}_{\mathrm{d}E}, S)_\sharp \mu_E\}$.*

Degrees of freedom in $\Pi_E(\mu, \nu)$. When $k < d$, there can be infinitely many E-optimal plans. However, we can further characterize the degrees of freedom available to define plans in $\Pi_E(\mu, \nu)$. Indeed, let $\gamma \in \Pi_E(\mu, \nu)$. Then, disintegrating γ on $E \times E$, we get $\gamma = \gamma_{(x_E, y_E)} \otimes \gamma_E$, i.e. plans in $\Pi_E(\mu, \nu)$ only differ on their disintegrations on $E \times E$.

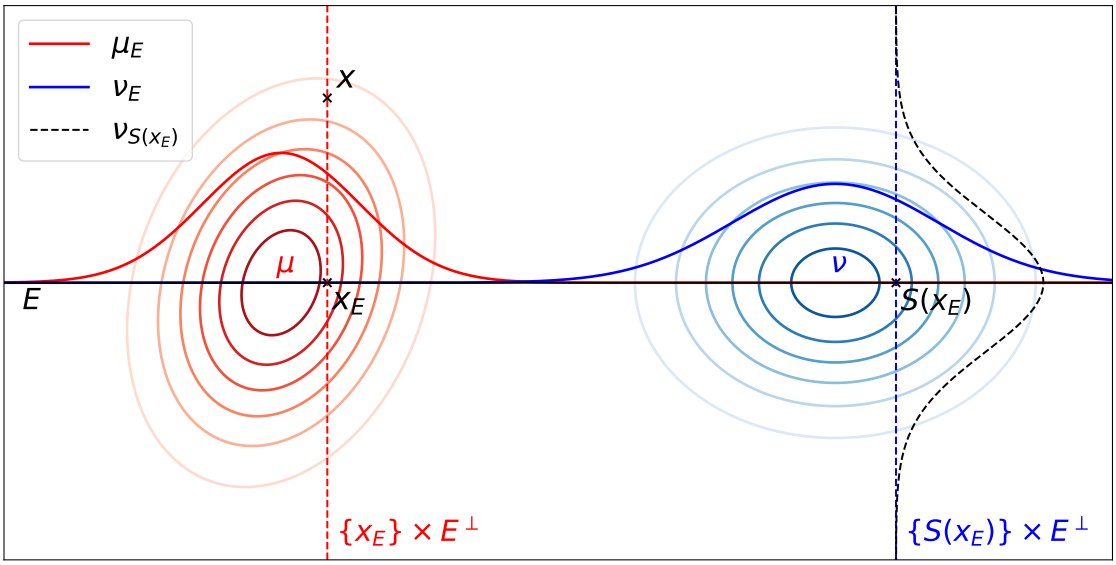


Figure 3.1: A $d = 2, k = 1$ illustration. Any $\gamma \in \Pi_E(\mu, \nu)$ being supported on $\mathcal{G}(S) \times (E^\perp)^2$, all the mass from x is transported on the fiber $\{S(x_E)\} \times E^\perp$. Different γ's in $\Pi_E(\mu, \nu)$ correspond to different couplings between the fibers $\{x_E\} \times E^\perp$ and $\{S(x_E)\} \times E^\perp$.

Further, since γ_E stems from a transport (Monge) map S, it is supported on the graph of S on E, $\mathcal{G}(S) = \{(x_E, S(x_E)) : x_E \in E\} \subset E \times E$. This implies that γ puts zero mass when $y_E \neq S(x_E)$ and thus that γ is fully characterized by $\gamma_{(x_E, S(x_E))}, x_E \in E$, i.e. by the couplings between μ_{x_E} and $\nu_{S(x_E)}$ for $x_E \in E$. This is illustrated in Figure 3.1. Two such couplings are presented: the first, *Monge-Independent* (MI) transport (Definition 3.3) corresponds to independent couplings between the conditionals, while the second *Monge-Knothe* (MK) transport (Definition 3.4) corresponds to optimal couplings between the conditionals.

Definition 3.3 (Monge-Independent Plans)**.** *The Monge-Independent plan disintegrates as the product of the independent couplings between μ_{x_E} and $\nu_{S(x_E)}$ for $x_E \in E$, and the coupling corresponding to S on E:*

$$\pi^{MI} \overset{\text{def}}{=} (\mu_{x_E} \otimes \nu_{S(x_E)}) \otimes (\mathbf{I}_{\mathrm{d}E}, S)_\sharp \mu_E.$$

Monge-Independent transport only requires that there exists a Monge map S between μ_E and ν_E (and not on the whole space), but extends S as a transportation plan and not a map. Since it couples disintegrations with the independent law, it is particularly suited to settings where all the information is contained in E, as shown in section 6.

When there exists a Monge map between disintegrations μ_{x_E} to $\nu_{S(x_E)}$ for all $x_E \in E$ (e.g. when μ and ν are a.c.), it is possible to extend S as a transportation map between μ and ν using those maps. The Monge-Knothe transport corresponds to the E-optimal plan with optimal couplings between the disintegrations.

Definition 3.4 (Monge-Knothe Transport)**.** *For all $x_E \in E$, let $\hat{T}(x_E; \cdot) : E^\perp \to E^\perp$ denote the Monge map from μ_{x_E} to $\nu_{S(x_E)}$. The Monge-Knothe transportation map is defined as*

$$\begin{aligned} T_{MK} : & E \oplus E^\perp \to E \oplus E^\perp \\ & (x_E, x_{E^\perp}) \mapsto (S(x_E), \hat{T}(x_E; x_{E^\perp})). \end{aligned}$$

The proof that T_{MK} defines a transport map from μ to ν is a direct adaptation of the proof for the Knothe-Rosenblatt transport [Santambrogio, 2015, Section 2.3]. When it is not possible to define a Monge map between the disintegrations, one can still consider the optimal couplings $\pi^{\text{OT}}(\mu_{x_E}, \nu_{S(x_E)})$ and define $\pi^{\text{MK}} = \pi^{\text{OT}}(\mu_{x_E}, \nu_{S(x_E)}) \otimes (\mathbf{I}_{\mathrm{d}E}, S)_\sharp \mu_E$, which we still call Monge-Knothe plan by abuse. In either case, π^{MK} is the E-optimal plan with lowest global cost:

Proposition 3.5. *The Monge-Knothe plan is optimal in $\Pi_E(\mu, \nu)$, namely*

$$\pi^{MK} \in \underset{\gamma \in \Pi_E(\mu,\nu)}{\arg\min}\, \mathbb{E}_{(X,Y)\sim\gamma}[\|X - Y\|^2].$$

Proof. E-optimal plans only differ in the couplings they induce between μ_{x_E} and $\nu_{S(x_E)}$ for $x_E \in E$. Since π^{MK} corresponds to the case when these couplings are optimal, disintegrating γ over $E \times E$ in $\int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 d\gamma(x, y)$ shows that $\gamma = \pi^{\text{MK}}$ has the lowest cost. □

Relation with the Knothe-Rosenblatt (KR) transport. These definitions are related to the KR transport [Santambrogio, 2015, section 2.3], which consists in defining a transport map between two a.c. measures by recursively (i) computing the Monge map T_1 between the first two one-dimensional marginals of μ and ν and (ii) repeating the process between the disintegrated measures μ_{x_1} and $\nu_{T_1(x_1)}$. MI and MK marginalize on the $k \geq 1$ dimensional subspace E, and respectively define the transport between disintegrations μ_{x_E} and $\nu_{S(x_E)}$ as the product measure and the optimal transport instead of recursing.

MK as a limit of optimal transport with re-weighted quadratic costs. Similarly to KR [Carlier et al., 2009], MK transport maps can intuitively be obtained as the limit of optimal transport maps, when the costs on $E^\perp$ become negligible compared to the costs on E.

Proposition 3.6. *Let $\mathbb{R}^d = E \oplus E^\perp$, $\begin{pmatrix}\mathbf{V}_E & \mathbf{V}_{E^\perp}\end{pmatrix}$ an orthonormal basis of $E \oplus E^\perp$ and $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ be two a.c. probability measures. Define*

$$\forall \varepsilon > 0, \quad \mathbf{P}_\varepsilon \overset{\text{def}}{=} \mathbf{V}_E \mathbf{V}_E^\top + \varepsilon \mathbf{V}_{E^\perp} \mathbf{V}_{E^\perp}^\top \quad \textit{and} \quad d^2_{\mathbf{P}_\varepsilon}(x, y) \overset{\text{def}}{=} (x - y)^\top \mathbf{P}_\varepsilon (x - y). \tag{3.2}$$

Let T_ε be the optimal transport map for the cost $d^2_{\mathbf{P}_\varepsilon}$. Then $T_\varepsilon \to T_{MK}$ in $\mathrm{L}_2(\mu)$.

Proof. The proof is a simpler, two-step variation of that of [Carlier et al., 2009], which we refer to for additional details. For all $\varepsilon \geq 0$, let π_ε be the optimal plan for $d^2_{\mathbf{P}_\varepsilon}$, and suppose there exists π such that $\pi_\varepsilon \rightharpoonup \pi$ (which is possible up to subsequences). By definition of π_ε, we have that

$$\forall \varepsilon \geq 0, \int d^2_{\mathbf{P}_\varepsilon} \mathrm{d}\pi_\varepsilon \leq \int d^2_{\mathbf{P}_\varepsilon} \mathrm{d}\pi^{\text{MK}}.$$

Since $d^2_{\mathbf{P}_\varepsilon}$ converges locally uniformly to $d^2_{\mathbf{V}_E} \overset{\text{def}}{=} (x, y) \to (x - y)^\top \mathbf{V}_E \mathbf{V}_E^\top (x - y)$, we get $\int d^2_{\mathbf{V}_E} \mathrm{d}\pi \leq \int d^2_{\mathbf{V}_E} \mathrm{d}\pi^{\text{MK}}$. But by definition of π^{MK}, $\pi^{\text{MK}}_E \overset{\text{def}}{=} (p_E, p_E)_\sharp \pi^{\text{MK}}$ is the optimal transport plan on E, therefore the last inequality implies that both marginals on E coincide, i.e. $\pi_E = \pi^{\text{MK}}_E$.

Next, notice that the π_ε's all have the same marginals μ_E, ν_E on E and hence cannot perform better on E than π^{MK}. Therefore,

$$\begin{aligned}\int_{E\times E} d^2_{\mathbf{V}_E}\mathrm{d}\pi^{\mathrm{MK}} + \varepsilon\int d^2_{\mathbf{V}_{E^\perp}}\mathrm{d}\pi_\varepsilon &\leq \int d^2_{\mathbf{P}_\varepsilon}\mathrm{d}\pi_\varepsilon \\ &\leq \int d^2_{\mathbf{P}_\varepsilon}\mathrm{d}\pi^{\mathrm{MK}} \\ &= \int_{E\times E} d^2_{\mathbf{V}_E}\mathrm{d}(\pi^{\mathrm{MK}})_E + \varepsilon\int d^2_{\mathbf{V}_{E^\perp}}\mathrm{d}\pi^{\mathrm{MK}}.\end{aligned}$$

Hence, passing to the $\varepsilon \to 0$ limit, we have

$$\int d^2_{\mathbf{V}_{E^\perp}}\mathrm{d}\pi \leq \int d^2_{\mathbf{V}_{E^\perp}}\mathrm{d}\pi^{\mathrm{MK}}.$$

Let us now disintegrate this inequality on $E \times E$ (using the equality $\pi_E = (\pi^{\mathrm{MK}})_E$):

$$\int\Big(\int_{E^\perp\times E^\perp} d^2_{\mathbf{V}_{E^\perp}}\mathrm{d}\pi_{(x_E,y_E)}\Big)\mathrm{d}\pi^{\mathrm{MK}}_E(x_E,y_E) \leq \int\Big(\int_{E^\perp\times E^\perp} d^2_{\mathbf{V}_{E^\perp}}\mathrm{d}\pi^{\mathrm{MK}}_{(x_E,y_E)}\Big)\mathrm{d}\pi^{\mathrm{MK}}_E(x_E,y_E).$$

Again, by definition, for (x_E, y_E) in the support of π^{MK}_E, $\pi^{\mathrm{MK}}_{(x_E,y_E)}$ is the optimal transportation plan between μ_{x_E} and ν_{y_E}, and the previous inequality implies $\pi_{(x_E,y_E)} = \pi^{\mathrm{MK}}_{(x_E,y_E)}$ for π^{MK}_E-a.e. (x_E, y_E), and finally $\pi = \pi^{\mathrm{MK}}$. Finally, by the a.c. hypothesis, all transport plans π_ε come from transport maps T_ε, which implies $T_\varepsilon \to T_{\mathrm{MK}}$ in $\mathrm{L}_2(\mu)$. □

MI as a limit of the discrete case. When μ and ν are a.c., for $n \in \mathbb{N}$ let μ_n, ν_n denote the uniform distribution over n i.i.d. samples from μ and ν respectively, and let π_n be an optimal transportation plan between $(p_E)_\sharp\mu_n$ and $(p_E)_\sharp\nu_n$ given by a Monge map (which is possible assuming uniform weights and non-overlapping projections). We have that $\mu_n \rightharpoonup \mu$ and $\nu_n \rightharpoonup \mu$. From [Santambrogio, 2015, Th 1.50, 1.51], we have that $\pi_n \in \mathcal{P}_2(E \times E)$ converges weakly, up to subsequences, to a coupling $\pi \in \mathcal{P}_2(E \times E)$ that is optimal for μ_E and ν_E. On the other hand, up to points having the same projections, the discrete plans π_n can also be seen as plans in $\mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$. A natural question is then whether the sequence $\pi_n \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$ has a limit in $\mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$.

Proposition 3.7. *Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ be a.c. and compactly supported, $\mu_n, \nu_n, n \geq 0$ be uniform distributions over n i.i.d. samples, and $\pi_n \in \Pi_E(\mu_n, \nu_n), n \geq 0$. Then $\pi_n \rightharpoonup \pi^{MI}(\mu, \nu)$.*

Proof. Let $\mathbf{X} \subset \mathbb{R}^d$ be a compact set, and consider two a.c. probabilty measures μ and ν supported on $\mathbf{X}$. Let E be a k-dimensional subspace which we identify w.l.o.g. with $\mathbb{R}^k$ and $\pi^{\mathrm{MI}} \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$ as in Definition 3.3. For $n \in \mathbb{N}$, denote n-sample empirical measures of μ and ν by $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ and $\nu_n = \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$ where the x_i (resp. y_i) are i.i.d. samples from μ (resp. ν). Let $S_n : \mathbb{R}^k \to \mathbb{R}^k$ be the Monge map from the projection on E $(p_E)_\sharp\mu_n$ of μ_n to that of ν_n, and $\pi_n \overset{\mathrm{def}}{=} (\mathbf{I}_\mathrm{d}, S_n)_\sharp[(p_E)_\sharp\mu_n]$.

Since μ and ν are supposed absolutely continuous, almost surely no two points have the same projection on E. Hence, t_n can be extended to a transport between μ_n and ν_n, whose transport plan we will denote γ_n.

Let $f \in C_b(\mathbf{X} \times \mathbf{X})$. Since $\mathbf{X}$ is compact, by density (given by the Stone-Weierstrass theorem) it is sufficient to consider functions of the form

$$f(x_1, ..., x_d; y_1, ..., y_d) = g(x_1, ..., x_k; y_1, ..., y_k)h(x_{k+1}, ..., x_d; y_{k+1}, ..., y_d).$$

We will use this along with the disintegrations of γ_n on $E \times E$ (denoted $(\gamma_n)_{x_{1:k},y_{1:k}}$ for $(x_{1:k}, y_{1:k}) \in E \times E$) to prove convergence:

$$\begin{aligned}\int_{\mathbf{X}\times\mathbf{X}} f d\gamma_n &= \int_{\mathbf{X}\times\mathbf{X}} g(x_{1:k}, y_{1:k}) h(x_{k+1:d}, y_{k+1:d}) d\gamma_n \\ &= \int_{E\times E} g(x_{1:k}, y_{1:k}) d\pi_n \int h(x_{k+1:d}, y_{k+1:d}) d(\gamma_n)_{x_{1:k},y_{1:k}} \\ &= \int_{E\times E} g(x_{1:k}, y_{1:k}) d\pi_n \int h(x_{k+1:d}, y_{k+1:d}) d(\mu_n)_{x_{1:k}} d(\nu_n)_{t_n(x_{1:k})}.\end{aligned}$$

Then, we use (i) the Arzela-Ascoli theorem to get uniform convergence of t_n to T_E to get $d(\nu_n)_{t_n(x_{1:k})} \rightharpoonup d(\nu)_{T_E(x_{1:k})}$ and (ii) the convergence $\pi_n \rightharpoonup \pi_E^{\mathrm{MI}} \stackrel{\text{def}}{=} (p_E, p_E)_\sharp \pi^{\mathrm{MI}}$ to get

$$\begin{aligned}&\int_{E\times E} g(x_{1:k}, y_{1:k}) d\pi_n \int h(x_{k+1:d}, y_{k+1:d}) d(\mu_n)_{x_{1:k}} d(\nu_n)_{t_n(x_{1:k})} \\ &\quad\to \int_{E\times E} g(x_{1:k}, y_{1:k}) d\pi_E^{\mathrm{MI}} \int h(x_{k+1:d}, y_{k+1:d}) d\mu_{x_{1:k}} d\nu_{T_E(x_{1:k})} \\ &\quad= \int_{\mathbf{X}\times\mathbf{X}} f d\pi^{\mathrm{MI}},\end{aligned}$$

which concludes the proof in the compact case. □

We conjecture that under additional assumptions, the compactness hypothesis can be relaxed. In particular, we empirically observe convergence for Gaussians.

4 Explicit Formulas for Subspace Detours in the Bures Metric

Multivariate Gaussian measures are a specific case of continuous distributions for which Wasserstein distances and Monge maps are available in closed form. We first recall basic facts from Section 2 about optimal transport between Gaussian measures, and then show that the E-optimal transports MI and MK introduced in section Section 3 are also in closed form. For two Gaussians μ, ν, one has

$$W_2^2(\mu, \nu) = \|\mathbf{m}_\mu - \mathbf{m}_\nu\|^2 + \mathfrak{B}^2(\operatorname{var}\mu, \operatorname{var}\nu)$$

where $\mathfrak{B}^2$ is the *Bures* metric [Bhatia et al., 2018] between PSD matrices:

$$\mathfrak{B}^2(\mathbf{A}, \mathbf{B}) \stackrel{\text{def}}{=} \mathrm{Tr}\mathbf{A} + \mathrm{Tr}\mathbf{B} - 2\mathrm{Tr}(\mathbf{A}^{1/2}\mathbf{B}\mathbf{A}^{1/2})^{1/2}.$$

The Monge map from a centered Gaussian distribution μ with covariance matrix $\mathbf{A}$ to one ν with covariance matrix $\mathbf{B}$ is linear and is represented by the matrix

$$\mathbf{T}^{\mathbf{AB}} \stackrel{\text{def}}{=} \mathbf{A}^{-1/2}(\mathbf{A}^{1/2}\mathbf{B}\mathbf{A}^{1/2})^{1/2}\mathbf{A}^{-1/2}.$$

For any linear transport map, $\mathbf{T}_\sharp\mu$ has covariance $\mathbf{T}\mathbf{A}\mathbf{T}^\top$, and the transportation cost from μ to ν is

$$\mathbb{E}_{X\sim\mu}[\|X - \mathbf{T}X\|^2] = \mathrm{Tr}\mathbf{A} + \mathrm{Tr}\mathbf{B} - \mathrm{Tr}(\mathbf{T}\mathbf{A} + \mathbf{A}\mathbf{T}^\top).$$

In the following, μ (resp. ν) will denote the centered Gaussian distribution with covariance matrix $\mathbf{A}$ (resp. $\mathbf{B}$). We write $\mathbf{A} = \begin{pmatrix} \mathbf{A}_E & \mathbf{A}_{EE^\perp} \\ \mathbf{A}_{EE^\perp}^\top & \mathbf{A}_{E^\perp} \end{pmatrix}$ when $\mathbf{A}$ is represented in an orthonormal basis $\begin{pmatrix} \mathbf{V}_E & \mathbf{V}_{E^\perp} \end{pmatrix}$ w.r.t. $E \oplus E^\perp$.

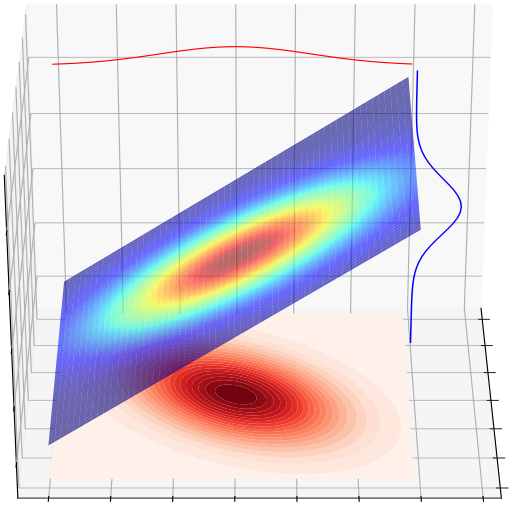


Figure 3.2: MI transport from a 2D Gaussian (red) to a 1D Gaussian (blue), projected on the x-axis. The two 1D distributions represent the projections of both Gaussians on the x-axis, the blue one being already originally supported on the x-axis. The oblique hyperplane is the support of π^{MI}, onto which its density is represented.

Monge-Independent transport between Gaussian measures. The MI transport between Gaussian measures is given by a degenerate Gaussian, i.e. a measure with Gaussian density over the image of its covariance matrix Σ.

Proposition 3.8 (Monge-Independent (MI) transport for Gaussian measures)**.**

$$\textit{Let } \mathbf{C} \overset{\text{def}}{=} \left(\mathbf{V}_E\mathbf{A}_E + \mathbf{V}_{E^\perp}\mathbf{A}_{EE^\perp}^\top\right)\mathbf{T}^{\mathbf{A}_E\mathbf{B}_E}\left(\mathbf{V}_{E^\top} + (\mathbf{B}_E)^{-1}\mathbf{B}_{EE^\perp}\mathbf{V}_{E^\perp}^\top\right) \tag{3.3}$$

and $\Sigma \overset{\text{def}}{=} \left(\begin{smallmatrix}\mathbf{A} & \mathbf{C}\\ \mathbf{C}^\top & \mathbf{B}\end{smallmatrix}\right)$. *Then* $\pi^{MI}(\mu,\nu) = \mathcal{N}(0_{2d},\Sigma) \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$.

Due to its being lengthy and merely technical, the proof of Proposition 3.8 is deferred to Section 7.

Knothe-Rosenblatt and Monge-Knothe transport between Gaussian measures. Before giving the closed-form MK map for Gaussian measures, we derive the KR map [Santambrogio, 2015, §2.3] with successive marginalization[1] on $x_1, x_2, ..., x_d$. When $d = 2$ and the basis is orthonormal for $E \oplus E^\perp$, those two notions coincide.

Proposition 3.9 (Knothe-Rosenblatt (KR) transport between Gaussian measures)**.** *Let* $\mathbf{L}_A$ *(resp.* $\mathbf{L}_B$*) be the Cholesky factor of* $\mathbf{A}$ *(resp.* $\mathbf{B}$*). The KR transport from* μ *to* ν *is a linear map whose matrix is given by* $\mathbf{T}_{KR}^{\mathbf{AB}} = \mathbf{L}_B(\mathbf{L}_A)^{-1}$*. Its cost is the squared Frobenius distance between the Cholesky factors* $\mathbf{L}_A$ *and* $\mathbf{L}_B$*:*

$$\mathbb{E}_{X\sim\mu}[\|X - T_{KR}^{\mathbf{AB}}X\|^2] = \|\mathbf{L}_A - \mathbf{L}_B\|^2.$$

Proof. The KR transport with successive marginalization on $x_1, x_2, ..., x_d$ between two a.c. distributions has a lower triangular Jacobian with positive entries on the diagonal. Further, since the one-dimensional disintegrations of Gaussian measures are Gaussian measures themselves, and since Monge maps between Gaussian measures are linear, the KR transport

[1]Note that compared to Santambrogio [2015], this is the reversed marginalization order, which is why the KR map here has *lower* triangular Jacobian.

between two centered Gaussians is a linear map, hence its matrix representation equals its Jacobian and is lower triangular.

Let $\mathbf{T} = \mathbf{L}_B(\mathbf{L}_A)^{-1}$. We have

$$\begin{aligned}\mathbf{T}\mathbf{A}\mathbf{T}^\top &= \mathbf{L}_B\mathbf{L}_A^{-1}\mathbf{L}_A\mathbf{L}_A^\top\mathbf{L}_A^{-\top}\mathbf{L}_B^\top \\ &= \mathbf{L}_B\mathbf{L}_B^\top \\ &= \mathbf{B},\end{aligned}$$

i.e. $\mathbf{T}_\sharp\mu = \nu$. Further, since $\mathbf{T}\mathbf{L}_A$ is the Cholesky factor for $\mathbf{B}$, and since $\mathbf{A}$ is supposed non-singular, by unicity of the Cholesky decomposition $\mathbf{T}$ is the only lower triangular matrix satisfying $\mathbf{T}_\sharp\mu = \nu$. Hence, it is the KR transport map from μ to ν.

Finally, we have that

$$\begin{aligned}\mathbb{E}_{X\sim\mu}[\|X - \mathbf{T}_{\text{KR}}X\|^2] &= \operatorname{Tr}(\mathbf{A} + \mathbf{B} - (\mathbf{A}(\mathbf{T}_{\text{KR}})^\top + \mathbf{T}_{\text{KR}}\mathbf{A})) \\ &= \operatorname{Tr}(\mathbf{L}_A\mathbf{L}_A^\top + \mathbf{L}_B\mathbf{L}_B^\top - (\mathbf{L}_A\mathbf{L}_B^\top + \mathbf{L}_B\mathbf{L}_A^\top)) \\ &= \|\mathbf{L}_A - \mathbf{L}_B\|^2.\end{aligned}$$

□

Corollary 3.10. *The (square root) cost of the Knothe-Rosenblatt transport* $(\mathbb{E}_{X\sim\mu}[\|X - \mathbf{T}_{KR}X\|^2])^{1/2}$ *between centered gaussians defines a distance (i.e. it satisfies all three metric axioms).*

Proof. This comes from the fact that $\left(\mathbb{E}_{X\sim\mu}[\|X - \mathbf{T}_{\text{KR}}X\|^2]\right)^{1/2} = \|\mathbf{L}_A - \mathbf{L}_B\|$. □

As can be expected from the fact that MK can be seen as a generalization of KR, the MK transportation map is linear and has a block-triangular structure. The next proposition shows that the MK transport map can be expressed as a function of the Schur complements

$$\mathbf{A}/\mathbf{A}_E \overset{\text{def}}{=} \mathbf{A}_{E^\perp} - \mathbf{A}_{EE^\perp}^\top\mathbf{A}_E^{-1}\mathbf{A}_{EE^\perp} \quad \text{and} \quad \mathbf{B}/\mathbf{B}_E \overset{\text{def}}{=} \mathbf{B}_{E^\perp} - \mathbf{B}_{EE^\perp}^\top\mathbf{B}_E^{-1}\mathbf{B}_{EE^\perp}$$

of $\mathbf{A}$ w.r.t. $\mathbf{A}_E$, and $\mathbf{B}$ w.r.t. $\mathbf{B}_E$, which are the covariance matrices of μ (resp. ν) conditioned on E.

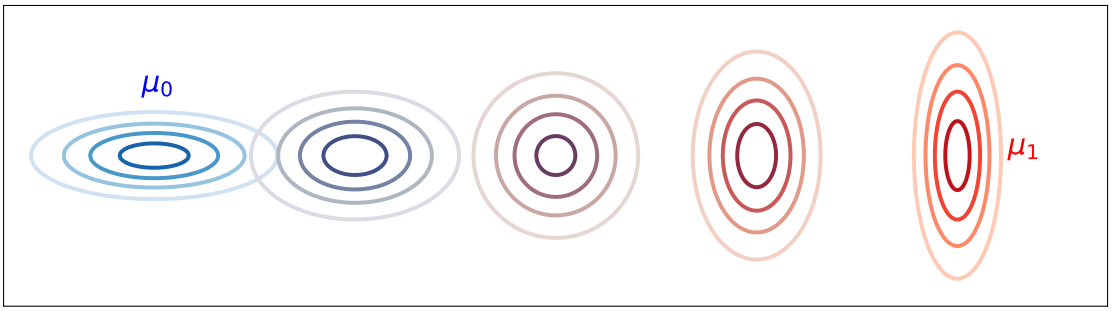
Proposition 3.11 (Monge-Knothe (MK) Transport for Gaussians)**.** *Let* $\mathbf{A}$ *and* $\mathbf{B}$ *be represented in an orthonormal basis w.r.t.* $E \oplus E^\perp$*. The MK transport map on* E *between* $\mu = \mathcal{N}(0_d, \mathbf{A})$ *and* $\nu = \mathcal{N}(0_d, \mathbf{B})$ *is linear, and represent by the following matrix:*

$$\mathbf{T}_{\text{MK}} = \begin{pmatrix} \mathbf{T}^{\mathbf{A}_E\mathbf{B}_E} & 0_{k\times(d-k)} \\ \left[\mathbf{B}_{EE^\perp}^\top(\mathbf{T}^{\mathbf{A}_E\mathbf{B}_E})^{-1} - \mathbf{T}^{(\mathbf{A}/\mathbf{A}_E)(\mathbf{B}/\mathbf{B}_E)}\mathbf{A}_{EE^\perp}^\top\right](\mathbf{A}_E)^{-1} & \mathbf{T}^{(\mathbf{A}/\mathbf{A}_E)(\mathbf{B}/\mathbf{B}_E)} \end{pmatrix}.$$

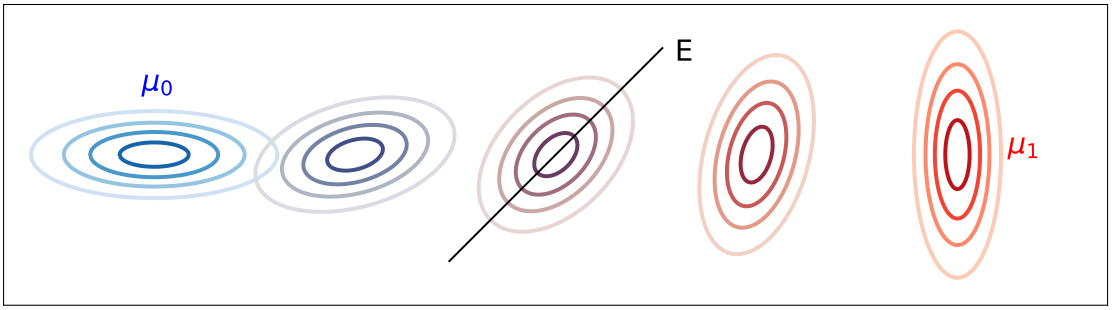
Proof. As can be seen from the structure of the MK transport map in Definition 3.4, T_{MK} has a lower block-triangular Jacobian (with block sizes k and $d-k$), with PSD matrices on the diagonal (corresponding to the Jacobians of the Monge maps (i) between marginals and (ii) between conditionals). Further, since μ and ν are Gaussian measures, their disintegrations are Gaussian as well. Hence, all Monge maps from the disintegrations of μ to that of ν are linear, and therefore the matrix representing $\mathbf{T}$ is equal to its Jacobian. One can check that the map $\mathbf{T}$ in the proposition verifies $\mathbf{T}\mathbf{A}\mathbf{T}^\top = \mathbf{B}$ and is of the right form. Finally, one can verify that it is the unique such matrix, hence it is the MK transport map. □

5 Selecting the Supporting Subspace

Both MI and MK transports are highly dependent on the chosen subspace E. Depending on applications, E can either be prescribed (e.g. if one has access to a transport map between the marginals in a given subspace) or has to be selected. In the latter case, we give guidelines on how prior knowledge can be used, and alternatively propose an algorithm for minimizing the MK distance.



(a) Usual Monge Interpolation of Gaussians



(b) Monge-Knothe Interpolation through E

Figure 3.3: (a) Wasserstein-Bures geodesic and (b) Monge-Knothe interpolation through $E = \{(x, y) : x = y\}$ from μ_0 to μ_1, at times $t = 0, 0.25, 0.5, 0.75, 1$.

Subspace selection using prior knowledge. When prior knowledge is available, one can choose a mediating subspace E to enforce specific criteria when comparing two distributions. Indeed, if the directions in E are known to correspond to given properties of the data, then MK or MI transport privileges those properties when matching distributions over those not encoded by E. In particular, if one has access to features $\mathbf{X}$ from a reference dataset, one can use principal component analysis (PCA) and select the first k principal directions to compare datasets $\mathbf{X}_1$ and $\mathbf{X}_2$. MK and MI then allow comparing $\mathbf{X}_1$ and $\mathbf{X}_2$ using the most significant features from the reference $\mathbf{X}$ with higher priority. In section 6, we experiment this method on word embeddings.

Minimal Monge-Knothe subspace. Alternatively, in the absence of prior knowledge, it is natural to aim at finding the subspace which minimizes MK. Unfortunately, optimization on the Grassmann manifold is quite hard in general, which makes direct optimization of MK w.r.t. E impractical. Optimizing with respect to an orthonormal matrix $\mathbf{V}$ of basis vectors of $\mathbb{R}^d$ is a more practical parameterization, which allows to perform projected gradient descent (Algorithm 3). The projection step consists in computing a polar decomposition, as the projection of a matrix $\mathbf{V}$ onto the set of unitary matrices is the unitary matrix in the polar decomposition of $\mathbf{V}$. The proposed initialization is $V = \text{Polar}(\mathbf{AB})$, as this is the optimal solution when $\mathbf{A}$ and $\mathbf{B}$ are co-diagonalizable. Note that since the function being minimized is non-convex, Algorithm 3 is only guaranteed to converge to a local minimum. In section 6, experimental evaluation of Algorithm 3 is carried out on noise-contaminated synthetic data (Figure 3.6) and on a domain adaptation task with Gaussian mixture models on the Office Home dataset [Venkateswara et al., 2017] with inception features (Figure 3.7).

Algorithm 3 MK Subspace Selection

Input: $\mathbf{A}, \mathbf{B} \in \mathrm{PSD}, k \in [\![1, d]\!], \eta$
  $\mathbf{V} \leftarrow \mathrm{Polar}(\mathbf{AB})$
  while not converged **do**
    $\mathcal{L} \leftarrow \mathrm{MK}(\mathbf{V}^\top \mathbf{A}\mathbf{V}, \mathbf{V}^\top \mathbf{B}\mathbf{V}; k)$
    $\mathbf{V} \leftarrow \mathbf{V} - \eta \nabla_{\mathbf{V}} \mathcal{L}$
    $\mathbf{V} \leftarrow \mathrm{Polar}(\mathbf{V})$
  end while
Output: $E = \mathrm{Span}\{\mathbf{v}_1, .., \mathbf{v}_k\}$

6 Experiments

(a) Gray Source (b) Gray OT (c) Gray Target

Figure 3.4: OT color transfer between gray projections.

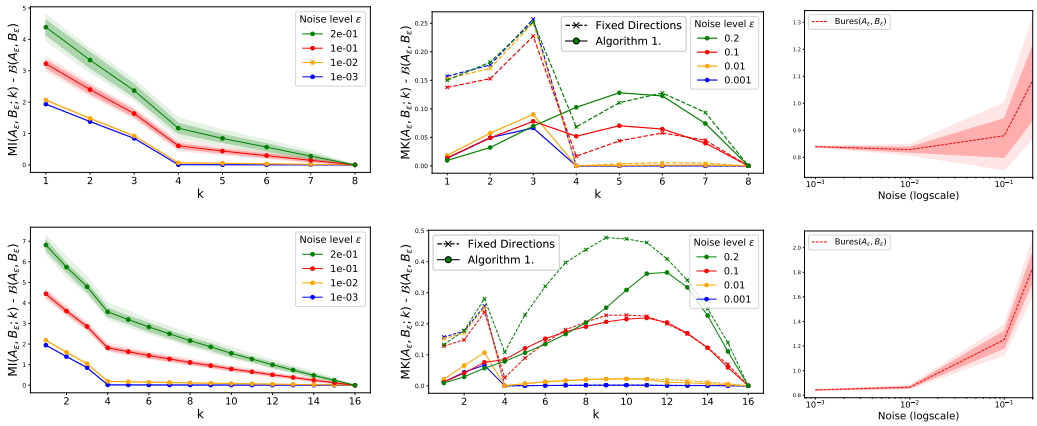
(a) Source (b) Full OT (2.67s) (c) MK (0.052s) (d) Sliced (0.057s) (e) Target

Figure 3.5: Color transfer, after quantization using 3000 k-means clusters, with corresponding runtimes.

Color transfer. Given a source and a target image, the goal of color transfer is to map the color palette of the source image (represented by its RGB histogram) into that of the target image. A natural toolbox for such a task is optimal transport, see *e.g.* Bonneel et al. [2015], Ferradans et al. [2014], Rabin et al. [2014]. First, a k-means quantization of both images is computed. Then, the colors of the pixels within each source cluster are modified according to the optimal transport map between both color distributions. In Figure 3.5, we illustrate discrete MK transport maps for color transfer. In this setting, we project images on the 1D space of grayscale images, relying on the 1D OT sorting-based algorithm (Figure 3.4). Then, we solve small 2D OT problems on the corresponding disintegrations. We compare this approach with classic full OT maps and a sliced OT approach (with 100 random projections). As can be seen in Figure 3.5, MK results are visually very similar to that of full OT, with a x50 speedup allowed by the fast 1D OT sorting-based algorithm that is comparable to sliced OT.

Synthetic data. We test the behavior of MK and MI in a noisy environment, where the signal is supported in a subspace of small dimension. We represent the signal using

two normalized PSD matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{d_1 \times d_1}$ and sample noise $\Sigma_1, \Sigma_2 \in \mathbb{R}^{d_2 \times d_2}, d_2 \geq d_1$ from a Wishart distribution with parameter $\mathbf{I}$. We then build the noisy covariance $\mathbf{A}_\varepsilon = \begin{pmatrix} \mathbf{A} & 0 \\ 0 & 0 \end{pmatrix} + \varepsilon\Sigma_1 \in \mathbb{R}^{d_2 \times d_2}$ (and likewise $\mathbf{B}_\varepsilon$) for different noise levels ε and compute MI and MK distances along the first k directions, $k = 1, ..., d_2$. As can be seen in Figure 3.6, both MI and MK curves exhibit a local minimum or an "elbow" when $k = d_1$, i.e. when E corresponds to the subspace where the signal is located. However, important differences in the behaviors of MI and MK can be noticed. Indeed, MI has a steep decreasing curve from 1 to d_1 and then a slower decreasing curve. This is explained by the fact that MI transport computes the OT map along the k directions of E only, and treats the conditionals as being independent. Therefore, if $k \geq d_1$, all the signal has been fitted and for increasing values of k MI starts fitting the noise as well. On the other hand, MK transport computes the optimal transport on both E and the corresponding $(d_2 - k)$-dimensional conditionals. Therefore, if $k \neq d_1$, either or both maps fit a mixture of signal and noise. Local maxima correspond to cases where the signal is the most contaminated by noise, and minima $k = d_1$, $k = d_2$ to cases where either the marginals or the conditionals are unaffected by noise. Using Algorithm 3 instead of the principle directions allows to find better subspaces than the first k directions when $k \leq d_1$, and then behaves similarly (up to the gradient being stuck in local minima and thus being occasionally less competitive). Overall, the differences in behavior of MI and MK show that MI is more adapted to noisy environments, and MK to applications where all directions are meaningful, but where one wishes to prioritize fitting on a subset of those directions, as shown in the next experiment.



(d) Monge-Independent (e) Monge-Knothe (f) Bures

Figure 3.6: (a)-(b): Difference between (a) MI and Bures and (b) MK and Bures metrics for different noise levels ε and subspace dimensions k. (c): Corresponding Bures values. For each ϵ, 100 different noise matrices are sampled. Points show mean values, and shaded areas the 25%-75% and 10%-90% percentiles. Top row: $d_1 = 4, d_2 = 8$. Bottom row: $d_1 = 4, d_2 = 16$.

Semantic mediation. We experiment using reference features for comparing distributions with elliptical word embeddings [Muzellec and Cuturi, 2018], which represent each word from a given corpus using a mean vector and a covariance matrix. For a given embedding, we expect the principal directions of its covariance matrix to be linked to its semantic content. Therefore, the comparison of two words w_1, w_2 based on the principal eigenvectors of a context word c should be impacted by the semantic relations of w_1 and w_2 with respect to c, e.g. if w_1 is polysemous and c is related to a specific meaning. To test this intuition, we compute the nearest neighbors of a given word w according to the MK

distance with E taken as the subspace spanned by the principal directions of two different contexts c_1 and c_2. We exclude means and compute MK based on covariances only, and look at the symmetric difference of the returned sets of words (i.e. words in KNN$(w|c_1)$ but not in KNN$(w|c_2)$, and inversely). Table 3.1 shows that specific contexts affect the nearest neighbors of ambiguous words.

Table 3.1: Symmetric differences of the 20-NN sets of w given c_1 minus w given c_2 using MK. Embeddings are 12×12 pretrained normalized covariance matrices from [Muzellec and Cuturi, 2018]. E is spanned by the 4 principal directions of the contexts. Words are printed in increasing distance order.

Word	Context 1	Context 2	Difference
instrument	monitor	oboe	cathode, monitor, sampler, rca, watts, instrumentation, telescope, synthesizer, ambient
	oboe	monitor	tuned, trombone, guitar, harmonic, octave, baritone, clarinet, saxophone, virtuoso
windows	pc	door	netscape, installer, doubleclick, burner, installs, adapter, router, cpus
	door	pc	screwed, recessed, rails, ceilings, tiling, upvc, profiled, roofs
fox	media	hedgehog	Penny, quiz, Whitman, outraged, Tinker, ads, Keating, Palin, show
	hedgehog	media	panther, reintroduced, kangaroo, Harriet, fair, hedgehog, bush, paw, bunny

MK domain adaptation with Gaussian mixture models. Given a *source* dataset of labeled data, domain adaptation (DA) aims at finding labels for a *target* dataset by transfering knowledge from the source. Such a problem has been successfully tackled using OT-based techniques [Courty et al., 2014]. We illustrate using MK Gaussian maps on a domain adaptation task where both source and target distributions are modeled by a Gaussian mixture model (GMM). We use the Office Home dataset [Venkateswara et al., 2017], which comprises 15000 images from 65 different classes across 4 domains: `Art`, `Clipart`, `Product` and `Real World`. For each image, we consider 2048-dimensional features taken from the coding layer of an inception model, as with Fréchet inception distances [Heusel et al., 2017]. For each source/target pair, we represent the source as a GMM by fitting one Gaussian per source class and defining mixture weights proportional to class frequencies, and we fit a GMM with the same number of components on the target. Since label information is not available for the target dataset, data from different classes may be assigned to the same component. We then compute pairwise MK distances between all source and target components, and solve for the discrete OT plan P using those distances as costs and mixture weights as marginals (as in Chen et al. [2019] with Bures distances). Finally, we map the source distribution on the target by computing the P-barycentric projection of the component-wise MK maps $\frac{1}{\sum_j P_{ij}} \sum_j P_{ij} T^{ij}_{\mathrm{MK}}$, and assign target labels using 1-NN prediction over the mapped source data. The same procedure is applied using Bures distances between the projections on E. We use Algorithm 3 between the empirical covariance matrices of the source and target datasets to select the supporting subspace E, for different values of the supporting dimension k (Figure 3.7).

Several facts can be observed from Figure 3.7. First, using the full 2048-dimensional Bures maps is regularly sub-optimal compared to Bures (resp. MK) maps on a lower-dimensional subspace, even though this is dependent on the source/target combination. This shows the interest of not using all available features equally in transport problems. Secondly, when E is chosen using the minimizing algorithm 3, in most cases MK maps yield equivalent or better classification accuracy that the corresponding Bures maps on the projections, even though they have the same projections on E. However, as can be expected, this does not hold for an arbitrary choice of E (not shown in the figure). Due to the relative simplicity of this DA method (which models the domains as GMMs), we do not aim at comparing with state-of-the-art OT DA methods Courty et al. [2014, 2017] (which compute transportation plans between the discrete distributions directly). The goal is rather to illustrate how MK maps can be used to compute maps which put higher priority

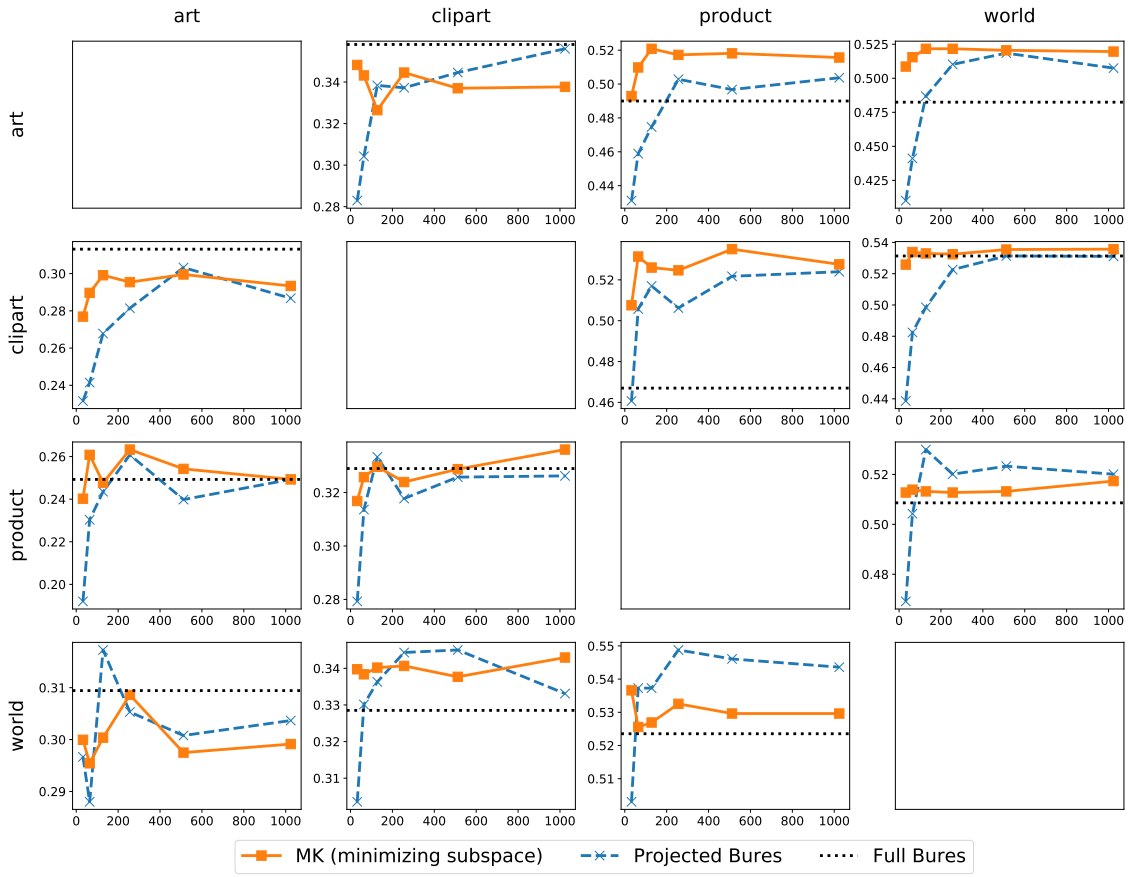


Figure 3.7: Domain Adaptation: 1-NN accuracy scores on the Office Home dataset v.s. dimension k. We compare the k-dimensional projected Bures maps with the E-MK maps and the 2048-D Bures baseline. E is selected using Algorithm 3 between the source and target covariance matrices for $k = 32, 64, 128, 256, 512, 1024$. Rows: sources, Columns: targets.

on the most meaningful feature dimensions. Note also that the mapping between source and target distributions used here is piecewise linear, and is therefore more regular.

Conclusion and Future Work

We have proposed in this chapter a new class of transport plans and maps that are built using optimality constraints on a subspace, but defined over the whole space. We have presented two particular instances, MI and MK, with different properties, and derived closed formulations for Gaussian distributions. Future work includes exploring other applications of OT to machine learning relying on low-dimensional projections, from which subspace-optimal transport could be used to recover full-dimensional plans or maps.

7 Appendix: Proof of Proposition 3.8

Proof. Let $\mathbf{T}_E : \mathbf{A}_E^{-1/2}(\mathbf{A}_E^{1/2}\mathbf{B}_E\mathbf{A}_E^{1/2})^{1/2}\mathbf{A}_E^{-1/2}$ be the Monge map from $\mu_E \overset{\text{def}}{=} (p_E)_\sharp\mu$ to $\nu_E \overset{\text{def}}{=} (p_E)_\sharp\nu$. Let

$$V = \begin{pmatrix} | & & | & | & & | \\ v_1 & \dots & v_k & v_{k+1} & \dots & v_d \\ | & & | & | & & | \end{pmatrix} = \begin{pmatrix} \mathbf{V}_E & \mathbf{V}_{E^\perp} \end{pmatrix} \in \mathbb{R}^{d\times d},$$

where $(v_1 \dots v_k)$ is an orthonormal basis of E and $(v_{k+1} \dots v_d)$ an orthonormal basis of $E^\perp$. Let us denote $X_E \overset{\text{def}}{=} p_E(X) \in \mathbb{R}^k$ and $X_{E^\perp} \overset{\text{def}}{=} p_{E^\perp}(X) \in \mathbb{R}^k$ (and likewise for Y). Denote

$$\mathbf{A}_E = p_E\mathbf{A}p_E^\top, \mathbf{A}_{E^\perp} = p_{E^\perp}\mathbf{A}p_{E^\perp}^\top, \text{ and } \mathbf{A}_{EE^\perp} = p_E\mathbf{A}p_{E^\perp}^\top.$$

With these notations, let us decompose $\mathbb{E}[XY^\top]$ along E and $E^\perp$:

$$\begin{aligned}\mathbb{E}[XY^\top] = \mathbb{E}[\mathbf{V}_E X_E(\mathbf{V}_E Y_E)^\top] &+ \mathbb{E}[\mathbf{V}_{E^\perp} X_{E^\perp}(\mathbf{V}_{E^\perp} Y_{E^\perp})^\top] \\ &+ \mathbb{E}[\mathbf{V}_{E^\perp} X_{E^\perp}(\mathbf{V}_E Y_E)^\top] \\ &+ \mathbb{E}[\mathbf{V}_E X_E(\mathbf{V}_{E^\perp} Y_{E^\perp})^\top].\end{aligned}$$

We can condition all four terms on X_E, and use independence given coordinates on E which implies $(Y_E|X_E) = X_E$. The constraint $Y_E = \mathbf{T}_E X_E$ allows us to derive $\mathbb{E}\left[Y_{E^\perp}|X_E\right]$: indeed, it holds that

$$\begin{pmatrix} Y_E \\ Y_{E^\perp} \end{pmatrix} \sim \mathcal{N}\left(0_d, \begin{pmatrix} \mathbf{B}_E & \mathbf{B}_{EE^\perp} \\ \mathbf{B}_{EE^\perp}^\top & \mathbf{B}_{E^\perp} \end{pmatrix}\right),$$

which, using standard Gaussian conditioning properties, implies that

$$\mathbb{E}\left[Y_{E^\perp}|Y_E = \mathbf{T}_E X_E\right] = \mathbf{B}_{EE^\perp}^\top \mathbf{B}_E^{-1}\mathbf{T}_E X_E,$$

and therefore

$$\mathbb{E}\left[Y_{E^\perp}|\mathbf{P}_E(Y) = \mathbf{T}_E X_E\right] = V_{E^\perp}\mathbf{B}_{EE^\perp}^\top \mathbf{B}_E^{-1}\mathbf{V}_E^\top\mathbf{T}_E X_E.$$

Likewise,

$$\mathbb{E}\left[X_{E^\perp}|\mathbf{P}_E(X)\right] = \mathbf{V}_{E^\perp}\mathbf{A}_{EE^\perp}^\top \mathbf{A}_E^{-1}\mathbf{V}_E^\top X_E.$$

We now have all the ingredients necessary to the derivation of the four terms of $\mathbb{E}[XY^\top]$:

(i)

$$\begin{aligned}\mathbb{E}[\mathbf{V}_E X_E(\mathbf{V}_E Y_E)^\top] &= \mathbf{V}_E\mathbb{E}_{X_E}\left[\mathbb{E}\left[X_E Y_E^\top|X_E\right]\right]\mathbf{V}_E^\top \\ &= \mathbf{V}_E\mathbb{E}_{X_E}\left[X_E\mathbb{E}\left[Y_E^\top|X_E\right]\right]\mathbf{V}_E^\top \\ &= \mathbf{V}_E\mathbb{E}_{X_E}\left[X_E X_E^\top\mathbf{T}_E^\top\right]\mathbf{V}_E^\top \\ &= \mathbf{V}_E\mathbb{E}_{X_E}\left[X_E X_E^\top\right]\mathbf{T}_E^\top\mathbf{V}_E^\top \\ &= \mathbf{V}_E\mathbf{A}_E\mathbf{T}_E\mathbf{V}_E^\top;\end{aligned}$$

(ii)

$$\begin{aligned}
\mathbb{E}[\mathbf{V}_E X_E (\mathbf{V}_{E^\perp} Y_{E^\perp})^\top] &= \mathbf{V}_E \mathbb{E}_{X_E}\left[\mathbb{E}[X_E Y_{E^\perp}^\top | X_E\right] \mathbf{V}_{E^\perp}^\top \\
&= \mathbf{V}_E \mathbb{E}_{X_E}\left[X_E \mathbb{E}\left[Y_{E^\perp}^\top | X_E = \mathbf{T}_E X_E\right]\right] \mathbf{V}_{E^\perp}^\top \\
&= \mathbf{V}_E \mathbb{E}_{X_E}\left[X_E \left(V_{E^\perp} \mathbf{B}_{EE^\perp}^\top \mathbf{B}_E^{-1} \mathbf{V}_E^\top \mathbf{T}_E X_E)\right)^\top\right] \mathbf{V}_{E^\perp}^\top \\
&= \mathbf{V}_E \mathbb{E}_{X_E}\left[X_E X_E^\top\right] \mathbf{T}_E^\top \mathbf{V}_E \mathbf{B}_E^{-\top} \mathbf{B}_{V_{EE^\perp}} \mathbf{V}_{E^\perp}^\top \\
&= \mathbf{V}_E \mathbf{A}_E \mathbf{T}_E \mathbf{V}_E \mathbf{B}_E^{-1} \mathbf{B}_{EE^\perp} \mathbf{V}_{E^\perp}^\top \\
&= \mathbf{V}_E \mathbf{A}_E \mathbf{T}_E \mathbf{V}_E \mathbf{B}_E^{-1} \mathbf{V}_E^\top \mathbf{B}_{EE^\perp} \mathbf{V}_{E^\perp}^\top;
\end{aligned}$$

(iii)

$$\begin{aligned}
\mathbb{E}[\mathbf{V}_{E^\perp} X_{E^\perp} (\mathbf{V}_E Y_E)^\top] &= \mathbf{V}_{E^\perp} \mathbb{E}_{X_E}\left[\mathbb{E}[X_{E^\perp} Y_E^\top | X_E\right] \mathbf{V}_E^\top \\
&= \mathbf{V}_{E^\perp} \mathbb{E}_{X_E}\left[\mathbb{E}\left[X_{E^\perp} | X_E\right] X_E^\top \mathbf{T}_E^\top\right] \mathbf{V}_E^\top \\
&= \mathbf{V}_{E^\perp} \mathbb{E}_{X_E}\left[\mathbf{A}_{EE^\perp}^\top \mathbf{A}_E^{-1} X_E X_E^\top \mathbf{T}_E^\top\right] \mathbf{V}_E^\top \\
&= \mathbf{V}_{E^\perp} \mathbf{V}_{E^\perp} \mathbf{A}_{EE^\perp}^\top \mathbf{A}_E^{-1} \mathbf{V}_E^\top \mathbf{A} \mathbf{T}_E \mathbf{V}_E^\top \\
&= \mathbf{V}_{E^\perp} \mathbf{V}_{E^\perp} \mathbf{A}_{EE^\perp}^\top \mathbf{T}_E \mathbf{V}_E^\top \\
&= \mathbf{V}_{E^\perp} \mathbf{A}_{EE^\perp}^\top \mathbf{T}_E \mathbf{V}_E^\top;
\end{aligned}$$

(iv)

$$\begin{aligned}
\mathbb{E}[\mathbf{V}_{E^\perp} X_{E^\perp} (\mathbf{V}_{E^\perp} Y_{E^\perp})^\top] &= \mathbf{V}_{E^\perp E} \mathbb{E}_{X_E}\left[\mathbb{E}[X_{E^\perp} | X_E \mathbb{E}[Y_{E^\perp}^\top | X_E]\right] \mathbf{V}_{E^\perp}^\top \\
&= \mathbf{V}_{E^\perp} \mathbb{E}_{X_E}\left[\mathbf{V}_{E^\perp} \mathbf{A}_{EE^\perp}^\top \mathbf{A}_E^{-1} \mathbf{V}_E^\top X_E X_E^\top \mathbf{T}_E^\top \mathbf{V}_E \mathbf{B}_E^{-\top} \mathbf{B}_{EE^\perp}\right] \mathbf{V}_{E^\perp}^\top \\
&= \mathbf{V}_{E^\perp} \mathbf{A}_{EE^\perp}^\top \mathbf{A}_E^{-1} \mathbf{V}_E^\top \mathbf{A}_E \mathbf{T}_E \mathbf{V}_E \mathbf{B}_E^{-1} \mathbf{B}_{EE^\perp} \mathbf{V}_{E^\perp}^\top \\
&= \mathbf{V}_{E^\perp} \mathbf{A}_{EE^\perp}^\top \mathbf{T}_E \mathbf{B}_E^{-1} \mathbf{B}_{EE^\perp} \mathbf{V}_{E^\perp}^\top \\
&= V_{E^\perp} \mathbf{A}_{EE^\perp}^\top \mathbf{T}_E \mathbf{V}_E \mathbf{B}_E^{-1} \mathbf{V}_E^\top \mathbf{B}_{EE^\perp}.
\end{aligned}$$

Let $\gamma \overset{\text{def}}{=} \mathcal{N}(0_{2d}, \Sigma_{\pi_E})$. γ, is well defined, since Σ_{π_E} is the covariance matrix of π_E and is thus PSD. From then, γ clearly has marginals $\mathcal{N}(0_d, \mathbf{A})$ and $\mathcal{N}(0_d, \mathbf{B})$, and $(p_E, p_E)_\sharp \gamma$ is a centered Gaussian distribution with covariance matrix

$$\begin{pmatrix} p_E & 0_{d\times d} \\ 0_{d\times d} & p_E \end{pmatrix} \begin{pmatrix} \mathbf{A} & \mathbb{E}_\pi[XY^\top] \\ \mathbb{E}_\pi[YX^\top] & \mathbf{B} \end{pmatrix} \begin{pmatrix} p_E & 0_{d\times d} \\ 0_{d\times d} & p_E \end{pmatrix} = \begin{pmatrix} \mathbf{A}_E & \mathbf{A}_E \mathbf{T}_E \\ \mathbf{T}_E \mathbf{A}_E & \mathbf{B}_E \end{pmatrix},$$

where we use that $p_E p_E = p_E$ and $p_E p_{E^\perp} = 0$. From the $k = d$ case, we recognise the covariance matrix of the optimal transport between centered Gaussians with covariance matrices $\mathbf{A}_E$ and $\mathbf{B}_E$, which proves that the marginal of γ over $E \times E$ is the optimal transport between μ_E and ν_E.

To complete the proof, there remains to show that the disintegration of γ on $E \times E$ is the product law. Denote

$$\begin{aligned}
&\mathbf{C} \overset{\text{def}}{=} \mathbb{E}[XY^\top] \\
&= \mathbf{V}_E \mathbf{A}_E \mathbf{T}_E \left(\mathbf{V}_E^\top + (\mathbf{B}_E)^{-1} \mathbf{V}_E^\top \mathbf{B}_{EE^\perp}\right) + \mathbf{V}_{E^\perp} \mathbf{A}_{E^\perp E} \mathbf{T}_{V_E} \left(\mathbf{V}_E^\top + (\mathbf{B}_{V_E})^{-1} \mathbf{V}_E^\top \mathbf{B}_{EE^\perp}\right) \\
&= \left(\mathbf{V}_E \mathbf{A}_E + \mathbf{V}_{E^\perp} \mathbf{A}_{E^\perp E}\right) \mathbf{T}_E \left(\mathbf{V}_E^\top + (\mathbf{B}_E)^{-1} \mathbf{B}_{EE^\perp} \mathbf{V}_{E^\perp}^\top\right),
\end{aligned}$$

and let $\Sigma_{\pi_{\mathrm{MI}}} = \begin{pmatrix} \mathbf{A} & \mathbb{E}[XY^\top] \\ \mathbb{E}[YX^\top] & \mathbf{B} \end{pmatrix}$ as in Proposition 3.8. It holds that

$$\begin{aligned}
\mathbf{C}_E &\stackrel{\text{def}}{=} \mathbf{V}_E^\top \mathbf{C} \mathbf{V}_E = \mathbf{A}_E \mathbf{T}_E, \\
\mathbf{C}_{E^\perp} &\stackrel{\text{def}}{=} \mathbf{V}_{E^\perp}^\top \mathbf{C} \mathbf{V}_E = \mathbf{A}_{E^\perp E} \mathbf{T}_E (\mathbf{B}_E)^{-1} \mathbf{B}_{EE^\perp}, \\
\mathbf{C}_{EE^\perp} &\stackrel{\text{def}}{=} \mathbf{V}_E^\top \mathbf{C} \mathbf{V}_{E^\perp} = \mathbf{A}_E \mathbf{T}_E (\mathbf{B}_E)^{-1} \mathbf{B}_{EE^\perp}, \\
\mathbf{C}_{E^\perp E} &\stackrel{\text{def}}{=} \mathbf{V}_{E^\perp}^\top \mathbf{C} \mathbf{V}_E = \mathbf{A}_{E^\perp E} \mathbf{T}_E.
\end{aligned}$$

Therefore, if $(X, Y) \sim \gamma$, then

$$\mathrm{Cov}_{(X,Y)\sim\gamma} \begin{pmatrix} X_{E^\perp} \\ Y_{E^\perp} \\ X_E \\ Y_E \end{pmatrix} = \begin{pmatrix} \mathbf{A}_{E^\perp} & \mathbf{C}_{E^\perp} & \mathbf{A}_{E^\perp E} & \mathbf{C}_{E^\perp E} \\ \mathbf{C}_{E^\perp} & \mathbf{B}_{E^\perp} & \mathbf{C}_{EE^\perp}^\top & \mathbf{B}_{E^\perp E} \\ \mathbf{A}_{EE^\perp} & \mathbf{C}_{EE^\perp} & \mathbf{A}_E & \mathbf{C}_E \\ \mathbf{C}_{E^\perp E}^\top & \mathbf{B}_{EE^\perp} & \mathbf{C}_E & \mathbf{B}_E \end{pmatrix},$$

and therefore

$$\mathrm{Cov} \begin{pmatrix} X_{E^\perp} | X_E \\ Y_{E^\perp} | Y_E \end{pmatrix} = \begin{pmatrix} \mathbf{A}_{E^\perp} & \mathbf{C}_{E^\perp} \\ \mathbf{C}_{E^\perp} & \mathbf{B}_{E^\perp} \end{pmatrix} - \begin{pmatrix} \mathbf{A}_{E^\perp E} & \mathbf{C}_{E^\perp E} \\ \mathbf{C}_{EE^\perp}^\top & \mathbf{B}_{E^\perp E} \end{pmatrix} \begin{pmatrix} \mathbf{A}_E & \mathbf{C}_E \\ \mathbf{C}_E & \mathbf{B}_E \end{pmatrix}^\dagger \begin{pmatrix} \mathbf{A}_{EE^\perp} & \mathbf{C}_{EE^\perp} \\ \mathbf{C}_{E^\perp E}^\top & \mathbf{B}_{EE^\perp} \end{pmatrix},$$

where $\mathbf{M}^\dagger$ denotes the Moore-Penrose pseudo-inverse of $\mathbf{M}$. In the present case, one can check that

$$\begin{pmatrix} \mathbf{A}_E & \mathbf{C}_E \\ \mathbf{C}_E & \mathbf{B}_E \end{pmatrix}^\dagger = \frac{1}{4} \begin{pmatrix} \mathbf{A}_E^{-1} & \mathbf{A}_E^{-1} \mathbf{T}_E^{-1} \\ \mathbf{T}_E^{-1} \mathbf{A}_E^{-1} & \mathbf{B}_E^{-1} \end{pmatrix},$$

which gives after simplification

$$\begin{pmatrix} \mathbf{A}_{E^\perp E} & \mathbf{C}_{E^\perp E} \\ \mathbf{C}_{EE^\perp}^\top & \mathbf{B}_{E^\perp E} \end{pmatrix} \begin{pmatrix} \mathbf{A}_E & \mathbf{C}_E \\ \mathbf{C}_E & \mathbf{B}_E \end{pmatrix}^\dagger \begin{pmatrix} \mathbf{A}_{EE^\perp} & \mathbf{C}_{EE^\perp} \\ \mathbf{C}_{E^\perp E}^\top & \mathbf{B}_{EE^\perp} \end{pmatrix} = \begin{pmatrix} \mathbf{A}_{E^\perp E} \mathbf{A}_E^{-1} \mathbf{A}_{EE^\perp} & \mathbf{C}_{E^\perp} \\ \mathbf{C}_{E^\perp} & \mathbf{B}_{E^\perp E} \mathbf{B}_E^{-1} \mathbf{B}_{EE^\perp} \end{pmatrix},$$

and thus

$$\begin{aligned}
\mathrm{Cov} \begin{pmatrix} X_{E^\perp} | X_E \\ Y_{E^\perp} | Y_E \end{pmatrix} &= \begin{pmatrix} \mathbf{A}_{E^\perp} - \mathbf{A}_{E^\perp E} (\mathbf{A}_E)^{-1} \mathbf{A}_{EE^\perp} & 0_d \\ 0_d & \mathbf{B}_{E^\perp} - \mathbf{B}_{E^\perp E} (\mathbf{B}_E)^{-1} \mathbf{B}_{EE^\perp} \end{pmatrix} \\
&= \begin{pmatrix} \mathrm{Cov}(X_{E^\perp} | X_E) & 0_d \\ 0_d & \mathrm{Cov}(Y_{E^\perp} | Y_E) \end{pmatrix},
\end{aligned}$$

that is, the conditional laws of $X_{E^\perp}$ given X_E and $Y_{E^\perp}$ given Y_E are independent under γ.

□

Chapter 4

Entropic Optimal Transport between (Unbalanced) Gaussian Measures

Optimal transport (OT) problems admit closed-form solutions in very few cases, such as in 1D or for Gaussian measures. Yet, these closed forms have proved extremely fecund for practitioners to define tools inspired from the OT geometry. On the other hand, the numerical resolution of OT problems using entropic regularization has given rise to many applications, but because there are no known closed-form solutions for entropic regularized OT problems, these approaches are mostly algorithmic and rely on numerical approximation. In this chapter, we propose to fill the void at the intersection between these two schools of thought in OT by proving that the entropy-regularized optimal transport problem between two Gaussian measures admits a closed form. Contrary to the unregularized case, for which the explicit form is given by the Bures-Wasserstein distance, the closed form we obtain is differentiable everywhere, even for measures with degenerate covariance matrices. We obtain this closed-form solution by solving the fixed-point equation behind Sinkhorn's algorithm, the default method for computing entropic regularized OT. Remarkably, this approach extends to the generalized *unbalanced* case – where Gaussian measures are scaled by positive constants. This extension leads to a closed form expression for unnormalized Gaussian measures as well, and highlights the mass transportation/destruction trade-off seen in unbalanced optimal transport. Moreover, in both settings, we show that the optimal transportation plans are (scaled) Gaussian measures and provide analytical formulas of their parameters. While formulas for regularization with a Shannon entropic term (as opposed Kullback-Leibler) were already proposed by Bojilov and Galichon [2016], we show additional properties of the resulting quantity based on its links with Sinkhorn's algorithm. More importantly, our formulas constitute the first non-trivial multivariate closed forms for unbalanced entropy-regularized optimal transport, thus providing a ground truth for the analysis of unbalanced entropic OT and Sinkhorn's algorithm.

This chapter is based on [Janati and Muzellec et al., 2020].

1 Introduction

Optimal transport (OT) theory [Villani, 2008, Figalli, 2017] has recently inspired several works in data science, where dealing with and comparing probability distributions, and more generally positive measures, is an important staple (see [Peyré et al., 2019] and references therein). For these applications of OT to be successful, a belief now widely shared in the community is that some form of regularization is needed for OT to be both scalable and avoid the curse of dimensionality [Dereich et al., 2013, Fournier and Guillin, 2015]. Two approaches have emerged in recent years to achieve these goals: either regularize directly the measures themselves, by looking at them through a simplified lens; or regularize the original OT problem using various modifications. The first approach exploits well-known closed-form identities for OT when comparing two univariate measures or two multivariate Gaussian measures. In this approach, one exploits those formulas and operates by summarizing complex measures as one or possibly many univariate or multivariate Gaussian measures. The second approach builds on the fact that for arbitrary measures, regularizing the OT problem, either in its primal or dual form, can result in simpler computations and possibly improved sample complexity. The latter approach can offer additional benefits for data science: because the original marginal constraints of the OT problem can also be relaxed, regularized OT can also yield useful tools to compare measures with different total mass — the so-called "unbalanced" case [Benamou, 2003]— which provides a useful additional degree of freedom. Our work in this chapter stands at the intersection of these two approaches. To our knowledge, that intersection was so far empty: no meaningful closed-form formulation was known for regularized optimal transport. We provide closed-form formulas of entropic (OT) of two Gaussian measures for balanced and unbalanced cases.

Summarizing measures *vs.* regularizing OT. Closed-form identities to compute OT distances (or more generally recover Monge maps) are known when either (1) both measures are univariate and the ground cost is submodular [Santambrogio, 2015, §2]: in that case evaluating OT only requires integrating that submodular cost w.r.t. the quantile distributions of both measures; or (2) both measures are Gaussian, in a Hilbert space, and the ground cost is the squared Euclidean metric [Dowson and Landau, 1982, Gelbrich, 1990], in which case the OT cost is given by the Wasserstein-Bures metric [Bhatia et al., 2018, Malagò et al., 2018]. These two formulas have inspired several works in which data measures are either projected onto 1D lines [Rabin et al., 2011, Bonneel et al., 2015], with further developments in [Paty and Cuturi, 2019, Kolouri et al., 2019, Titouan et al., 2019]; or represented by Gaussians, to take advantage of the simpler computational possibilities offered by the Wasserstein-Bures metric [Heusel et al., 2017, Muzellec and Cuturi, 2018, Chen et al., 2019].

Various schemes have been proposed to regularize the OT problem in the primal [Cuturi, 2013, Frogner et al., 2015] or the dual [Shirdhonkar and Jacobs, 2008, Arjovsky et al., 2017, Cuturi and Peyré, 2016]. We focus in this work on the formulation obtained by Chizat et al. [2018b], which combines entropic regularization [Cuturi, 2013] with a more general formulation for unbalanced transport [Chizat et al., 2018a, Liero et al., 2016, 2018]. The advantages of unbalanced entropic transport are numerous: it comes with favorable sample complexity regimes compared to unregularized OT [Genevay et al., 2019], can be cast as a loss with favorable properties [Genevay et al., 2018, Feydy et al., 2019], and can be evaluated using variations of the Sinkhorn algorithm [Genevay et al., 2016].

On the absence of closed-form formulas for regularized OT. Despite its appeal, one of the shortcomings of entropic regularized OT lies in the absence of simple test-cases that admit closed-form formulas. While it is known that regularized OT can be related, in

the limit of infinite regularization, to the energy distance [Ramdas et al., 2017], the absence of closed-form formulas for a fixed regularization strength poses an important practical problem to evaluate the performance of stochastic algorithms that try to approximate regularized OT: we do not know of any setup for which the ground truth value of entropic OT between continuous densities is known. The purpose of this chapter is to fill this gap, and provide closed form expressions for balanced and unbalanced OT for Gaussian measures. We hope these formulas will prove to be useful in two different ways: as a solution to the problem outlined above, to facilitate the evaluation of new methodologies building on entropic OT, and more generally to propose a more robust yet well-grounded replacement to the Bures-Wasserstein metric.

Contributions. Our contributions can be summarized as follows:

- Theorem 4.2 provides a closed form expression of the entropic (OT) plan π, which is shown to be a Gaussian measure itself. The closed form of the objective remains well defined, convex and differentiable even for singular covariance matrices unlike the Bures metric.
- Using the definition of debiased Sinkhorn barycenters [Luise et al., 2019, Janati et al., 2020a], Theorem 4.12 shows that the entropic barycenter of Gaussian measures is Gaussian and its covariance verifies a fixed point equation similar to that of Agueh and Carlier [2011].
- As in the balanced case, Theorem 4.14 provides a closed form expression of the unbalanced Gaussian transport plan. The obtained formula sheds some light on the link between mass destruction and the distance between the means of α, β in unbalanced OT.

Notations. Let $\mathcal{N}(\mathbf{a}, \mathbf{A})$ denote the multivariate Gaussian distribution with mean $\mathbf{a} \in \mathbb{R}^d$ and variance $\mathbf{A} \in \mathcal{S}^d_{++}$. $f = \mathcal{Q}(\mathbf{a}, \mathbf{A})$ denotes the quadratic form $f : x \mapsto -\frac{1}{2}(x^\top \mathbf{A} x - 2\mathbf{a}^\top x)$ with $\mathbf{A} \in \mathcal{S}^d$. For short, we denote $\mathcal{Q}(\mathbf{A}) = \mathcal{Q}(0, \mathbf{A})$. Whenever relevant, we follow the convention $0 \log 0 = 0$. $\mathcal{M}^+_p$ denotes the set of non-negative measures in $\mathbb{R}^d$ with a finite p-th order moment and its subset of probablity measures $\mathcal{P}_p$. For a non-negative measure $\alpha \in \mathcal{M}^+_p(\mathbb{R}^d)$, $\mathcal{L}_2(\alpha)$ denotes the set of functions $f : \mathbb{R}^d \to \mathbb{R}$ such that $\mathbb{E}_\alpha(|f|^2) = \int_{\mathbb{R}^d} |f|^2 \mathrm{d}\alpha < +\infty$.

2 Reminders on Optimal Transport

The Kantorovich problem. Let $\alpha, \beta \in \mathcal{P}_2$ and let $\Pi(\alpha, \beta)$ denote the set of probability measures in $\mathcal{P}_2$ with marginal distributions equal to α and β. The 2-Wasserstein distance is defined as

$$W_2^2(\alpha, \beta) \stackrel{\text{def}}{=} \min_{\pi \in \Pi(\alpha,\beta)} \iint_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 \mathrm{d}\pi(x, y). \tag{4.1}$$

This is known as the *Kantorovich* formulation of optimal transport. When α is absolutely continuous with respect to the Lebesgue measure (i.e. when α has a density), Equation (4.1) can be equivalently rewritten using the *Monge* formulation, where $T_\sharp \alpha = \beta$ *i.f.f.* for all Borel sets A, $\nu(T(A)) = \alpha(A)$:

$$W_2^2(\alpha, \beta) = \min_{T : T_\sharp \alpha = \beta} \int_{\mathbb{R}^d} \|x - T(x)\|^2 \mathrm{d}\alpha(x). \tag{4.2}$$

The optimal map T^* in Equation (4.2) is called the Monge map.

The Wasserstein-Bures metric. Let $\mathcal{N}(m, \Sigma)$ denote the Gaussian distribution on $\mathbb{R}^d$ with mean $m \in \mathbb{R}^d$ and covariance matrix $\Sigma \in S^d_{++}$. A well-known fact [Dowson and Landau, 1982, Takatsu, 2011] is that Equation (4.1) admits a closed form for Gaussian distributions, called the Wasserstein-Bures distance (a.k.a. the *Fréchet* distance):

$$W_2^2(\mathcal{N}(a, \mathbf{A}), \mathcal{N}(b, \mathbf{B})) = \|a - b\|^2 + \mathfrak{B}^2(\mathbf{A}, \mathbf{B}), \tag{4.3}$$

where $\mathfrak{B}$ is the *Bures* distance [Bhatia et al., 2018] between positive matrices:

$$\mathfrak{B}^2(\mathbf{A}, \mathbf{B}) \stackrel{\text{def}}{=} \mathrm{Tr}\mathbf{A} + \mathrm{Tr}\mathbf{B} - 2\mathrm{Tr}(\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}})^{\frac{1}{2}}. \tag{4.4}$$

Moreover, the Monge map between two Gaussian distributions admits a closed form: $T^\star : x \to \mathbf{T}^{\mathbf{AB}}(x - \mathbf{a}) + \mathbf{b}$, with

$$\mathbf{T}^{\mathbf{AB}} \stackrel{\text{def}}{=} \mathbf{A}^{-\frac{1}{2}}(\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}})^{\frac{1}{2}}\mathbf{A}^{-\frac{1}{2}} = \mathbf{B}^{\frac{1}{2}}(\mathbf{B}^{\frac{1}{2}}\mathbf{A}\mathbf{B}^{\frac{1}{2}})^{-\frac{1}{2}}\mathbf{B}^{\frac{1}{2}}, \tag{4.5}$$

which is related to the Bures gradient:

$$\nabla_{\mathbf{A}}\mathfrak{B}^2(\mathbf{A}, \mathbf{B}) = \mathbf{I}_\mathrm{d} - \mathbf{T}^{\mathbf{AB}}. \tag{4.6}$$

$\mathfrak{B}(\mathbf{A}, \mathbf{B})$ and its gradient can be computed efficiently on GPUs using Newton-Schulz iterations which are provided in Algorithm 1 along with numerical experiments in the appendix.

3 Entropy-Regularized Optimal Transport between Gaussian Measures

Solving (4.1) can be quite challenging, even in a discrete setting [Peyré et al., 2019]. Adding an entropic regularization term to (4.1) results in a problem which can be solved efficiently using Sinkhorn's algorithm [Cuturi, 2013]. Let $\sigma > 0$. This corresponds to solving the following problem:

$$\mathrm{OT}_\sigma(\alpha, \beta) \stackrel{\text{def}}{=} \min_{\pi \in \Pi(\alpha,\beta)} \iint_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 \mathrm{d}\pi(x, y) + 2\sigma^2 \,\mathrm{KL}(\pi \| \alpha \otimes \beta), \tag{4.7}$$

where $\mathrm{KL}(\pi\|\alpha \otimes \beta) \stackrel{\text{def}}{=} \iint_{\mathbb{R}^d \times \mathbb{R}^d} \log\left(\frac{\mathrm{d}\pi}{\mathrm{d}\alpha \mathrm{d}\beta}\right) \mathrm{d}\pi$ is the Kullback-Leibler divergence (or relative entropy). As in the Kantorovich case (4.1), OT_σ can be studied with centered measures with no loss of generality.

Lemma 4.1. *Let $\alpha, \beta \in \mathcal{P}$ and $\bar{\alpha}, \bar{\beta}$ their respective centered transformations. It holds that*

$$\mathrm{OT}_\sigma(\alpha, \beta) = \mathrm{OT}_\sigma(\bar{\alpha}, \bar{\beta}) + \|\mathbf{a} - \mathbf{b}\|^2. \tag{4.8}$$

Proof. Let $\mathrm{d}\bar{\alpha}(x) = \mathrm{d}\alpha(x + \mathbf{a})$ (resp. $\mathrm{d}\bar{\beta}(y) = \mathrm{d}\beta(y + \mathbf{b})$, $\mathrm{d}\bar{\pi}(x, y) = \mathrm{d}\pi(x + \mathbf{a}, y + \mathbf{b})$, such that $\bar{\alpha}, \bar{\beta}$ and $\bar{\pi}$ are centered. Then, $\forall \pi \in \Pi(\alpha, \beta)$,

(i) $\bar{\pi} \in \Pi(\bar{\alpha}, \bar{\beta})$,

(ii) $\mathrm{KL}(\pi\|\alpha \otimes \beta) = \mathrm{KL}(\bar{\pi}\|\bar{\alpha} \otimes \bar{\beta})$

(iii)
$$\begin{aligned}\iint_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 \mathrm{d}\bar{\pi}(x, y) &= \iint_{\mathbb{R}^d \times \mathbb{R}^d} \|(x - \mathbf{a}) - (y - \mathbf{b})\|^2 \mathrm{d}\pi(x, y) \\ &= \|\mathbf{a} - \mathbf{b}\|^2 + \iint_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 \mathrm{d}\pi(x, y)\end{aligned}$$

Plugging (i)-(iii) into (4.7), we get $\mathrm{OT}_\sigma(\alpha, \beta) = \mathrm{OT}_\sigma(\bar{\alpha}, \bar{\beta}) + \|\mathbf{a} - \mathbf{b}\|^2$. □

Dual problem and Sinkhorn's algorithm. Compared to (4.1), (4.7) enjoys additional properties, such as the uniqueness of the solution π^*. Moreover, problem (4.7) has the following dual formulation:

$$\mathrm{OT}_\sigma(\alpha,\beta) = \max_{\substack{f\in\mathcal{L}_2(\alpha),\\ g\in\mathcal{L}_2(\beta)}} \mathbb{E}_\alpha(f) + \mathbb{E}_\beta(g) - 2\sigma^2\Big(\iint_{\mathbb{R}^d\times\mathbb{R}^d} e^{\frac{f(x)+g(y)-\|x-y\|^2}{2\sigma^2}}\mathrm{d}\alpha(x)\mathrm{d}\beta(y) - 1\Big). \tag{4.9}$$

If α and β have finite second order moments, a pair of dual potentials (f,g) is optimal if and only they verify the following optimality conditions β-a.s and α-a.s respectively [Mena and Niles-Weed, 2019]:

$$e^{\frac{f(x)}{2\sigma^2}}\left(\int_{\mathbb{R}^d} e^{-\frac{\|x-y\|^2+g(y)}{2\sigma^2}}\mathrm{d}\beta(y)\right) = 1, \quad e^{\frac{g(x)}{2\sigma^2}}\left(\int_{\mathbb{R}^d} e^{-\frac{\|x-y\|^2+f(y)}{2\sigma^2}}\mathrm{d}\alpha(y)\right) = 1. \tag{4.10}$$

Moreover, given a pair of optimal dual potentials (f,g), the optimal transportation plan is given by

$$\frac{\mathrm{d}\pi^\star}{\mathrm{d}\alpha\mathrm{d}\beta}(x,y) = e^{\frac{f(x)+g(y)-\|x-y\|^2}{2\sigma^2}}. \tag{4.11}$$

Starting from a pair of potentials (f_0,g_0), the optimality conditions (4.10) lead to an alternating dual ascent algorithm, which is equivalent to Sinkhorn's algorithm in log-domain:

$$\begin{aligned} g_{n+1} &= \left(y\in\mathbb{R}^d \to -2\sigma^2\log\int_{\mathbb{R}^d} e^{-\frac{\|x-y\|^2+f_n(x)}{2\sigma^2}}\mathrm{d}\alpha(x)\right), \\ f_{n+1} &= \left(x\in\mathbb{R}^d \to -2\sigma^2\log\int_{\mathbb{R}^d} e^{-\frac{\|x-y\|^2+g_{n+1}(y)}{2\sigma^2}}\mathrm{d}\beta(y)\right). \end{aligned} \tag{4.12}$$

Séjourné et al. [2019] showed that when the support of the measures is compact, Sinkhorn's algorithm converges to a pair of dual potentials. Here in particular, we study Sinkhorn's algorithm when α and β are Gaussian measures.

3.1 Closed form expression for Gaussian measures.

Theorem 4.2. *Let $\mathbf{A},\mathbf{B}\in S^d_{++}$ and $\alpha\sim\mathcal{N}(\mathbf{a},\mathbf{A})$ and $\beta\sim\mathcal{N}(\mathbf{b},\mathbf{B})$. Define $\mathbf{D}_\sigma = (4\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}}+\sigma^4\mathbf{I}_\mathrm{d})^{\frac{1}{2}}$. Then,*

$$\mathrm{OT}_\sigma(\alpha,\beta) = \|\mathbf{a}-\mathbf{b}\|^2 + \mathcal{B}^2_\sigma(\mathbf{A},\mathbf{B}), \text{ where} \tag{4.13}$$

$$\mathcal{B}^2_\sigma(\mathbf{A},\mathbf{B}) = \mathrm{Tr}(\mathbf{A}+\mathbf{B}-\mathbf{D}_\sigma) + d\sigma^2(1-\log(2\sigma^2)) + \sigma^2\log\det\left(\mathbf{D}_\sigma+\sigma^2\mathbf{I}_\mathrm{d}\right). \tag{4.14}$$

Moreover, with $\mathbf{C}_\sigma = \frac{1}{2}\mathbf{A}^{\frac{1}{2}}\mathbf{D}_\sigma\mathbf{A}^{-\frac{1}{2}} - \frac{\sigma^2}{2}\mathbf{I}_\mathrm{d}$, the Sinkhorn optimal transportation plan is also a Gaussian measure over $\mathbb{R}^d\times\mathbb{R}^d$ given by

$$\pi^\star \sim \mathcal{N}\left(\begin{pmatrix}\mathbf{a}\\ \mathbf{b}\end{pmatrix}, \begin{pmatrix}\mathbf{A} & \mathbf{C}_\sigma\\ \mathbf{C}_\sigma^\top & \mathbf{B}\end{pmatrix}\right). \tag{4.15}$$

Remark 4.3. *While for our proof it is necessary to assume that $\mathbf{A}$ and $\mathbf{B}$ are positive definite in order for them to have a Lebesgue density, notice that the closed form formula given by Theorem 4.2 remains well-defined for positive semi-definite matrices. Moreover, unlike the Bures-Wasserstein metric, OT_σ is differentiable even when $\mathbf{A}$ or $\mathbf{B}$ are singular.*

A simplified version of Theorem 4.2 was concurrently proven by Gerolin et al. [2020] for univariate centered Gaussians. The proof we provide is more general and is broken down into smaller results, Propositions 4.4 to 4.6 and lemma 4.8. Using Lemma 4.1, we can focus in the rest of this section on centered Gaussians without loss of generality.

Sinkhorn's algorithm and quadratic potentials. We obtain a closed form solution of OT_σ by considering quadratic solutions of (4.10). The following key proposition characterizes the obtained potential after a pair of Sinkhorn iterations with quadratic forms.

Proposition 4.4. *Let $\alpha \sim \mathcal{N}(0, \mathbf{A})$ and $\beta \sim \mathcal{N}(0, \mathbf{B})$ and the Sinkhorn transform $T_\alpha : \mathbb{R}^{\mathbb{R}^d} \to \mathbb{R}^{\mathbb{R}^d}$:*

$$T_\alpha(h)(x) \stackrel{\text{def}}{=} -\log \int_{\mathbb{R}^d} e^{\frac{-\|x-y\|^2}{2\sigma^2}+h(y)} \mathrm{d}\alpha(y). \tag{4.16}$$

Let $\mathbf{X} \in \mathcal{S}_d$. If $h = m + \mathcal{Q}(\mathbf{X})$ i.e $h(x) = m - \frac{1}{2}x^\top \mathbf{X} x$ for some $m \in \mathbb{R}$, then $T_\alpha(h)$ is well-defined if and only if $\mathbf{X}' \stackrel{\text{def}}{=} \sigma^2 \mathbf{X} + \sigma^2 \mathbf{A}^{-1} + \mathbf{I}_\mathrm{d} \succ 0$. In that case,

(i) $T_\alpha(h) = \mathcal{Q}(\mathbf{Y}) + m'$ where $\mathbf{Y} = \frac{1}{\sigma^2}(\mathbf{X}'^{-1} - \mathbf{I}_\mathrm{d})$ and $m' \in \mathbb{R}$ is an additive constant,

(ii) $T_\beta(T_\alpha(h))$ is well-defined and is also a quadratic form up to an additive constant, observing that $\mathbf{Y}' \stackrel{\text{def}}{=} \sigma^2 \mathbf{Y} + \sigma^2 \mathbf{B}^{-1} + \mathbf{I}_\mathrm{d} = \mathbf{X}'^{-1} + \sigma^2 \mathbf{B}^{-1} \succ 0$ and using (i).

Proof. The integrand of $T_\alpha(h)(x)$ can be written as

$$\begin{aligned} e^{\frac{-\|x-y\|^2}{2\sigma^2}+h(y)} \mathrm{d}\alpha(y) &\propto e^{\frac{-\|x-y\|^2}{2\sigma^2} - \frac{1}{2}(y^\top \mathbf{X} y - y^\top \mathbf{A}^{-1} y)} \mathrm{d}y \\ &\propto e^{-\frac{1}{2}(y^\top (\frac{\mathbf{I}_\mathrm{d}}{\sigma^2} + \mathbf{X} + \mathbf{A}^{-1})y) + \frac{x^\top y}{\sigma^2}} \mathrm{d}y \end{aligned}$$

which is integrable if and only if $\mathbf{X} + \mathbf{A}^{-1} + \frac{1}{\sigma^2}\mathbf{I}_\mathrm{d} \succ 0$. Moreover, up to a multiplicative factor, the exponentiated Sinkhorn transform is equivalent to a Gaussian convolution of an exponentiated quadratic form. Lemma 4.17 applies:

$$\begin{aligned} e^{-T_\alpha(h)} &= \int_{\mathbb{R}^d} e^{\frac{-\|x-y\|^2}{2\sigma^2}+f(y)} \mathrm{d}\alpha(y) \\ &\propto \int_{\mathbb{R}^d} e^{\frac{-\|x-y\|^2}{2\sigma^2} + \mathcal{Q}(\mathbf{X})(y) + \mathcal{Q}(\mathbf{A}^{-1})(y)} \mathrm{d}y \\ &\propto \exp\left(\mathcal{Q}\left(\tfrac{\mathbf{I}_\mathrm{d}}{\sigma^2}\right)\right) \star \exp\left(\mathcal{Q}(\mathbf{X}) + \mathcal{Q}(\mathbf{A}^{-1})\right) \\ &\propto \exp\left(\mathcal{Q}\left(\tfrac{\mathbf{I}_\mathrm{d}}{\sigma^2}\right)\right) \star \exp\left(\mathcal{Q}(\mathbf{X} + \mathbf{A}^{-1})\right) \\ &\propto \exp\left(\mathcal{Q}((\mathbf{I}_\mathrm{d} + \sigma^2 \mathbf{X} + \sigma^2 \mathbf{A}^{-1})^{-1}(\mathbf{X} + \mathbf{A}^{-1}))\right). \\ &\propto \exp\left(\mathcal{Q}(\frac{1}{\sigma^2}\mathbf{X}'^{-1}(\mathbf{X}' - \mathbf{I}_\mathrm{d}))\right). \\ &\propto \exp\left(\mathcal{Q}(\frac{1}{\sigma^2}(\mathbf{I}_\mathrm{d} - \mathbf{X}'^{-1}))\right). \end{aligned}$$

Therefore $T_\alpha(h)$ is equal to $\mathcal{Q}(\frac{1}{\sigma^2}(\mathbf{X}'^{-1} - \mathbf{I}_\mathrm{d}))$, up to an additive constant.

Finally, since $\mathbf{B}$ and $\mathbf{X}'$ are positive definite, the positivity condition of $\mathbf{Y}'$ holds and T_β can be applied again to get $T_\beta(T_\alpha(h))$. □

Consider the null inialization $f_0 = 0 = \mathcal{Q}(0)$. Since $\sigma^2 \mathbf{A}^{-1} + \mathbf{I}_\mathrm{d} \succ 0$, Proposition 4.4 applies with $\mathbf{X} = 0$ and a simple induction shows that (f_n, g_n) remain quadratic forms for all n. Sinkhorn's algorithm can thus be written as an algorithm on positive definite matrices.

Proposition 4.5. *Starting with null potentials, Sinkhorn's algorithm is equivalent to the followig iterations:*

$$\mathbf{F}_{n+1} = \sigma^2 \mathbf{A}^{-1} + \mathbf{G}_n^{-1}, \qquad \mathbf{G}_{n+1} = \sigma^2 \mathbf{B}^{-1} + \mathbf{F}_{n+1}^{-1}, \tag{4.17}$$

with $\mathbf{F}_0 = \sigma^2\mathbf{A}^{-1}+\mathbf{I}_\mathrm{d}$ *and* $\mathbf{G}_0 = \sigma^2\mathbf{B}^{-1}+\mathbf{I}_\mathrm{d}$. *Moreover, the sequence* $(\mathbf{F}_n, \mathbf{G}_n)$ *is contractive in the matrix operator norm and converges towards a pair of positive definite matrices* $(\mathbf{F}, \mathbf{G})$.

At optimality, the dual potentials are determined up to additive constants f_0 and g_0: $\frac{f}{2\sigma^2} = \mathcal{Q}(\mathbf{U}) + f_0$ and $\frac{g}{2\sigma^2} = \mathcal{Q}(\mathbf{V}) + g_0$ where $\mathbf{U}$ and $\mathbf{V}$ are given by

$$\mathbf{F} = \sigma^2\mathbf{U} + \sigma^2\mathbf{A}^{-1} + \mathbf{I}_\mathrm{d}, \qquad \mathbf{G} = \sigma^2\mathbf{V} + \sigma^2\mathbf{B}^{-1} + \mathbf{I}_\mathrm{d}. \tag{4.18}$$

Proof.

(i) Deriving the iterations. Let $\mathbf{U}_0 = \mathbf{V}_0 = 0$. Applying Proposition 4.4 to the initial pair of potentials $\mathcal{Q}(\mathbf{U}_0), \mathcal{Q}(\mathbf{V}_0)$ leads to the sequence of quadratic Sinkhorn potentials $\frac{f_n}{2\sigma^2} = \mathcal{Q}(\mathbf{U}_n)$ and $\frac{f_n}{2\sigma^2} = \mathcal{Q}(\mathbf{V}_n)$ where

$$\begin{aligned}\mathbf{V}_{n+1} &= \frac{1}{\sigma^2}((\sigma^2\mathbf{U}_n + \sigma^2\mathbf{A}^{-1} + \mathbf{I}_\mathrm{d})^{-1} - \mathbf{I}_\mathrm{d})\\ \mathbf{U}_{n+1} &= \frac{1}{\sigma^2}((\sigma^2\mathbf{V}_{n+1} + \sigma^2\mathbf{B}^{-1} + \mathbf{I}_\mathrm{d})^{-1} - \mathbf{I}_\mathrm{d}).\end{aligned}$$

The change of variable

$$\begin{aligned}\mathbf{F}_n &= \sigma^2\mathbf{U}_n + \sigma^2\mathbf{A}^{-1} + \mathbf{I}_\mathrm{d}\\ \mathbf{G}_n &= \sigma^2\mathbf{V}_n + \sigma^2\mathbf{B}^{-1} + \mathbf{I}_\mathrm{d}\end{aligned}$$

leads to (4.17).

(ii) Contractivity of the iterations. We now turn to show that this algorithm converges. First, note that since $\mathbf{F}_0, \mathbf{G}_0 \in S^d_{++}$, a straightforward induction shows that $\forall n \geq 0, \mathbf{F}_n, \mathbf{G}_n \in S^d_{++}$. Next, let us write the decoupled iteration on $\mathbf{F}$:

$$\mathbf{F} \leftarrow \sigma^2\mathbf{A}^{-1} + (\sigma^2\mathbf{B}^{-1} + \mathbf{F}^{-1})^{-1}. \tag{4.19}$$

Let $\forall \mathbf{X} \in S^d_{++}, \phi(\mathbf{X}) \stackrel{\text{def}}{=} \sigma^2\mathbf{A}^{-1} + (\sigma^2\mathbf{B}^{-1} + \mathbf{X}^{-1})^{-1} \in S^d_{++}$. For $\mathbf{X} \in S^d_{++}$ and $\mathbf{H} \in \mathbb{R}^{d\times d}$, he first differential of ϕ w.r.t. the Frobenius inner product admits the following expression:

$$D\phi(\mathbf{X})[\mathbf{H}] = (\mathbf{I}_\mathrm{d} + \sigma^2\mathbf{X}\mathbf{B}^{-1})^{-1}\mathbf{H}(\sigma^2\mathbf{B}^{-1}\mathbf{X} + \mathbf{I}_\mathrm{d})^{-1}.$$

Hence, $\|D\phi(\mathbf{X})[\mathbf{H}]\|_\mathrm{op} \leq \|(\mathbf{I}_\mathrm{d} + \sigma^2\mathbf{X}\mathbf{B}^{-1})^{-1}\|^2_\mathrm{op}\|\mathbf{H}\|_\mathrm{op}$. Plugging $\mathbf{H} = \mathbf{I}_\mathrm{d}$, we get that $\|D\phi(\mathbf{X})\|_\mathrm{op} = \|(\mathbf{I}_\mathrm{d} + \sigma^2\mathbf{X}\mathbf{B}^{-1})^{-1}\|^2_\mathrm{op}$. Finally, by matrix similarity

$$\|(\mathbf{I}_\mathrm{d} + \sigma^2\mathbf{X}\mathbf{B}^{-1})^{-1}\|_\mathrm{op} = \|(\mathbf{I}_\mathrm{d} + \sigma^2\mathbf{B}^{-\frac{1}{2}}\mathbf{X}\mathbf{B}^{-\frac{1}{2}})^{-1}\|_\mathrm{op}.$$

Hence, to bound $\|D\phi(\mathbf{X})[\mathbf{H}]\|_\mathrm{op}$ from above we need a lower bound on the smallest eigenvalue of the iterates. For a matrix $\mathbf{M}$, let $\lambda_d(\mathbf{M})$ and $\lambda_A(\mathbf{M})$ denote the smallest (resp. largest) eigenvalue of $\mathbf{M}$. From (4.19) and using Weyl's inequality, we can bound the smallest eigenvalue of $\mathbf{F}_n$ from under:

$$\forall n \geq 1, \lambda_d(\mathbf{F}_n) \geq \frac{\sigma^2}{\lambda_1(\mathbf{A})}.$$

Hence, the iterates live in $\mathcal{A} \stackrel{\text{def}}{=} S^d_{++} \cap \{\mathbf{X} : \lambda_d(\mathbf{X}) \geq \frac{\sigma^2}{\lambda_1(\mathbf{A})}\}$. Finally, for all $\mathbf{X} \in \mathcal{A}$,

$$\begin{aligned}\|(\mathbf{I}_\mathrm{d} + \sigma^2 \mathbf{B}^{-\frac{1}{2}} \mathbf{X} \mathbf{B}^{-\frac{1}{2}})^{-1}\|_{\mathrm{op}} &= \frac{1}{\lambda_d(\mathbf{I}_\mathrm{d} + \sigma^2 \mathbf{B}^{-1/2} \mathbf{X} \mathbf{B}^{-1/2})} \\ &= \frac{1}{1 + \sigma^2 \lambda_d(\mathbf{B}^{-1/2} \mathbf{X} \mathbf{B}^{-1/2})} \\ &\leq \frac{1}{1 + \sigma^2 \lambda_d(\mathbf{B}^{-1}) \lambda_d(\mathbf{X})} \\ &\leq \left(1 + \frac{\sigma^4}{\lambda_1(\mathbf{B}) \lambda_1(\mathbf{A})}\right)^{-1}\end{aligned}$$

which proves that $\|D\phi(\mathbf{X})\|_{\mathrm{op}} \leq \left(1 + \frac{\sigma^4}{\lambda_1(\mathbf{B})\lambda_1(\mathbf{A})}\right)^{-1} < 1$ for $\mathbf{X} \in \mathcal{A}$ and $\sigma^2 > 0$. The same arguments hold for the iterates $(\mathbf{G}_n)_{n \geq 0}$, and show that the iterations (4.17) are contractive, and thus convergent.

□

Closed-form solution of the fixed-point equation. Taking the limit of Sinkhorn's equations (4.17) along with the change of variable (4.18), there exists a pair of optimal potentials determined up to an additive constant:

$$\frac{f}{2\sigma^2} = \mathcal{Q}(\mathbf{U}) = \mathcal{Q}\left(\frac{1}{\sigma^2}(\mathbf{G}^{-1} - \mathbf{I}_\mathrm{d})\right), \qquad \frac{g}{2\sigma^2} = \mathcal{Q}(\mathbf{V}) = \mathcal{Q}\left(\frac{1}{\sigma^2}(\mathbf{F}^{-1} - \mathbf{I}_\mathrm{d})\right), \tag{4.20}$$

where $(\mathbf{F}, \mathbf{G})$ is the solution of the fixed point equations

$$\mathbf{F} = \sigma^2 \mathbf{A}^{-1} + \mathbf{G}^{-1}, \qquad \mathbf{G} = \sigma^2 \mathbf{B}^{-1} + \mathbf{F}^{-1}. \tag{4.21}$$

Let $\mathbf{C} \stackrel{\text{def}}{=} \mathbf{A}\mathbf{G}^{-1}$. Combining both equations of (4.21) in one leads to

$$\mathbf{G} = \sigma^2 \mathbf{B}^{-1} + (\mathbf{G}^{-1} + \sigma^2 \mathbf{A}^{-1})^{-1},$$

which can be shown to be equivalent to

$$\mathbf{C}^2 + \sigma^2 \mathbf{C} - \mathbf{A}\mathbf{B} = 0. \tag{4.22}$$

Notice that since $\mathbf{A}$ and $\mathbf{G}^{-1}$ are positive definite, their product $\mathbf{C} = \mathbf{A}\mathbf{G}^{-1}$ is similar to $\mathbf{A}^{\frac{1}{2}} \mathbf{G}^{-1} \mathbf{A}^{\frac{1}{2}}$. Thus, $\mathbf{C}$ has positive eigenvalues. Proposition 4.6 provides the only feasible solution of (4.22).

Proposition 4.6. *Let $\sigma^2 \geq 0$ and $\mathbf{C}$ satisfying Equation (4.22). Then,*

$$\begin{aligned}\mathbf{C} &= \left(\mathbf{A}\mathbf{B} + \frac{\sigma^4}{4}\mathbf{I}_\mathrm{d}\right)^{\frac{1}{2}} - \tfrac{\sigma^2}{2}\mathbf{I}_\mathrm{d} \\ &= \mathbf{A}^{\frac{1}{2}}(\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}} + \tfrac{\sigma^4}{4}\mathbf{I}_\mathrm{d})^{\frac{1}{2}}\mathbf{A}^{-\frac{1}{2}} - \tfrac{\sigma^2}{2}\mathbf{I}_\mathrm{d}\end{aligned} . \tag{4.23}$$

Proof. Combining the two equations in (4.21) yields

$$\begin{aligned}&\mathbf{G} = \sigma^2 \mathbf{B}^{-1} + (\mathbf{G}^{-1} + \sigma^2 \mathbf{A}^{-1})^{-1} \\ &\Leftrightarrow \mathbf{G}\mathbf{A}^{-1} = \sigma^2 \mathbf{B}^{-1}\mathbf{A}^{-1} + (\mathbf{A}\mathbf{G}^{-1} + \sigma^2 \mathbf{I}_\mathrm{d})^{-1} \\ &\Leftrightarrow \mathbf{C}^{-1} = \sigma^2 (\mathbf{A}\mathbf{B})^{-1} + (\mathbf{C} + \sigma^2 \mathbf{I}_\mathrm{d})^{-1} \\ &\Leftrightarrow \mathbf{C}^{-1}(\mathbf{C} + \sigma^2 \mathbf{I}_\mathrm{d}) = \sigma^2 (\mathbf{A}\mathbf{B})^{-1}(\mathbf{C} + \sigma^2 \mathbf{I}_\mathrm{d}) + \mathbf{I}_\mathrm{d} \\ &\Leftrightarrow \mathbf{I}_\mathrm{d} + \sigma^2 \mathbf{C}^{-1} = \sigma^2 (\mathbf{A}\mathbf{B})^{-1}(\mathbf{C} + \sigma^2 \mathbf{I}_\mathrm{d}) + \mathbf{I}_\mathrm{d} \\ &\Leftrightarrow \mathbf{C} + \sigma^2 \mathbf{I}_\mathrm{d} = \sigma^2 (\mathbf{A}\mathbf{B})^{-1}(\mathbf{C} + \sigma^2 \mathbf{I}_\mathrm{d})\mathbf{C} + \mathbf{C} \\ &\Leftrightarrow \mathbf{C}^2 + \sigma^2 \mathbf{C} - \mathbf{A}\mathbf{B} = 0.\end{aligned} \tag{4.24}$$

Let us now plug (4.23) in (4.22):

$$\begin{aligned}\mathbf{C}^2 &= \mathbf{AB} + \frac{\sigma^4}{2}\mathbf{I}_\mathrm{d} - \sigma^2\left(\mathbf{AB} + \frac{\sigma^4}{4}\mathbf{I}_\mathrm{d}\right)^{\frac{1}{2}} \\ &= \mathbf{AB} - \sigma^2\mathbf{C},\end{aligned}$$

which proves that (4.23) is indeed the solution of (4.22).

Finally, the second expression of $\mathbf{C}$ is obtained by observing that

$$(\mathbf{A}^{\frac{1}{2}}(\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}} + \tfrac{\sigma^4}{4}\mathbf{I}_\mathrm{d})^{\frac{1}{2}}\mathbf{A}^{-\frac{1}{2}})^2 = \mathbf{A}^{\frac{1}{2}}(\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}} + \tfrac{\sigma^4}{4}\mathbf{I}_\mathrm{d})\mathbf{A}^{-\frac{1}{2}} = \mathbf{AB} + \tfrac{\sigma^4}{4}\mathbf{I}_\mathrm{d},$$

i.e. that

$$\left(\mathbf{AB} + \tfrac{\sigma^4}{4}\mathbf{I}_\mathrm{d}\right)^{\frac{1}{2}} = \mathbf{A}^{\frac{1}{2}}(\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}} + \tfrac{\sigma^4}{4}\mathbf{I}_\mathrm{d})^{\frac{1}{2}}\mathbf{A}^{-\frac{1}{2}}.$$

□

Corollary 4.7. *The optimal dual potentials of* (4.20) *can be written in closed form as*

$$\mathbf{U} = \frac{\mathbf{B}}{\sigma^2}(\mathbf{C} + \sigma^2\mathbf{I}_\mathrm{d})^{-1} - \frac{\mathbf{I}_\mathrm{d}}{\sigma^2}, \qquad \mathbf{V} = (\mathbf{C} + \sigma^2\mathbf{I}_\mathrm{d})^{-1}\frac{\mathbf{A}}{\sigma^2} - \frac{\mathbf{I}_\mathrm{d}}{\sigma^2}. \tag{4.25}$$

Moreover, $\mathbf{U}$ *and* $\mathbf{V}$ *remain well-defined even for singular matrices* $\mathbf{A}$ *and* $\mathbf{B}$.

Optimal transportation plan and OT_σ. Using Corollary 4.7 and (4.20), Equation (4.11) leads to a closed-form expression of π. To conclude the proof of Theorem 4.2, we introduce lemma 4.8 that computes the OT_σ loss at optimality. Detailed technical proofs are provided in the appendix.

Lemma 4.8. *Let* $\mathbf{A}, \mathbf{B}, \mathbf{C}$ *be invertible matrices such that* $\mathbf{H} = \left(\begin{smallmatrix}\mathbf{A} & \mathbf{C} \\ \mathbf{C}^\top & \mathbf{B}\end{smallmatrix}\right) \succ 0$. *Let* $\alpha = \mathcal{N}(0, \mathbf{A}), \beta = \mathcal{N}(0, \mathbf{B})$, *and* $\pi = \mathcal{N}(0, \mathbf{H})$. *Then,*

(i)
$$\iint_{\mathbb{R}^d\times\mathbb{R}^d} \|x - y\|^2 \mathrm{d}\pi(x, y) = \mathrm{Tr}(\mathbf{A}) + \mathrm{Tr}(\mathbf{B}) - 2\mathrm{Tr}(\mathbf{C}); \tag{4.26}$$

(ii)
$$\mathrm{KL}\left(\pi \| \alpha \otimes \beta\right) = \tfrac{1}{2}\left(\log\det\mathbf{A} + \log\det\mathbf{B} - \log\det\left(\begin{smallmatrix}\mathbf{A} & \mathbf{C} \\ \mathbf{C}^T & \mathbf{B}\end{smallmatrix}\right)\right). \tag{4.27}$$

Proof. It follows from elementary properties of Gaussian measures that the first and second marginals of π are respectively α and β. Hence,

$$\begin{aligned}\iint_{\mathbb{R}^d\times\mathbb{R}^d} \|x - y\|^2 \mathrm{d}\pi(x, y) &= \iint_{\mathbb{R}^d\times\mathbb{R}^d} \|x\|^2 \mathrm{d}\pi(x, y) + \iint_{\mathbb{R}^d\times\mathbb{R}^d} \|y\|^2 \mathrm{d}\pi(x, y) - 2\iint_{\mathbb{R}^d\times\mathbb{R}^d} \langle x,\, y\rangle \mathrm{d}\pi(x, y) \\ &= \int_{\mathbb{R}^d} \|x\|^2 \mathrm{d}\alpha(x) + \int_{\mathbb{R}^d} \|y\|^2 \mathrm{d}\beta(y) - 2\iint_{\mathbb{R}^d\times\mathbb{R}^d} \langle x,\, y\rangle \mathrm{d}\pi(x, y) \\ &= \mathrm{Tr}(\mathbf{A}) + \mathrm{Tr}(\mathbf{B}) - 2\mathrm{Tr}(\mathbf{C}).\end{aligned}$$

Next, using the closed-form expression of the Kullback-Leibler divergence between Gaussian measures, we have

$$\begin{aligned}\mathrm{KL}\left(\pi \| \alpha \otimes \beta\right) &= \tfrac{1}{2}\left(\mathrm{Tr}\left[\left(\begin{smallmatrix}\mathbf{A} & 0 \\ 0 & \mathbf{B}\end{smallmatrix}\right)^{-1}\left(\begin{smallmatrix}\mathbf{A} & \mathbf{C} \\ \mathbf{C}^T & \mathbf{B}\end{smallmatrix}\right)\right] - 2n + \log\det\left(\begin{smallmatrix}\mathbf{A} & 0 \\ 0 & \mathbf{B}\end{smallmatrix}\right) - \log\det\left(\begin{smallmatrix}\mathbf{A} & \mathbf{C} \\ \mathbf{C}^T & \mathbf{B}\end{smallmatrix}\right)\right) \\ &= \tfrac{1}{2}\left(\log\det\mathbf{A} + \log\det\mathbf{B} - \log\det\left(\begin{smallmatrix}\mathbf{A} & \mathbf{C} \\ \mathbf{C}^T & \mathbf{B}\end{smallmatrix}\right)\right).\end{aligned}$$

□

Closed-form expressions of the optimal transport plan and OT_σ. We are now ready to conclude the proof of Theorem 4.2. Using (4.11) and (4.20), we have

$$\begin{aligned}
\frac{\mathrm{d}\pi}{\mathrm{d}x\mathrm{d}y}(x,y) &= \exp\left(\frac{f(x)+g(y)-\|x-y\|^2}{2\sigma^2}\right)\frac{\mathrm{d}\alpha}{\mathrm{d}x}(x)\frac{\mathrm{d}\beta}{\mathrm{d}y}(y)\\
&\propto \exp\left(\mathcal{Q}(\mathbf{A}^{-1})(x)+\frac{f(x)+g(y)-\|x-y\|^2}{2\sigma^2}+\mathcal{Q}(\mathbf{B}^{-1})(y)\right)\\
&\propto \exp\left(\mathcal{Q}(\mathbf{U}+\mathbf{A}^{-1})(x)+\mathcal{Q}(\mathbf{V}+\mathbf{B}^{-1})(y)+\mathcal{Q}(\begin{smallmatrix}\frac{\mathbf{I}_\mathrm{d}}{\sigma^2} & -\frac{\mathbf{I}_\mathrm{d}}{\sigma^2}\\ -\frac{\mathbf{I}_\mathrm{d}}{\sigma^2} & \frac{\mathbf{I}_\mathrm{d}}{\sigma^2}\end{smallmatrix})(x,y)\right)\\
&= \exp\left(\mathcal{Q}(\begin{smallmatrix}\mathbf{U}+\mathbf{A}^{-1} & 0\\ 0 & \mathbf{V}+\mathbf{B}^{-1}\end{smallmatrix})(x,y)+\mathcal{Q}(\begin{smallmatrix}\frac{\mathbf{I}_\mathrm{d}}{\sigma^2} & -\frac{\mathbf{I}_\mathrm{d}}{\sigma^2}\\ -\frac{\mathbf{I}_\mathrm{d}}{\sigma^2} & \frac{\mathbf{I}_\mathrm{d}}{\sigma^2}\end{smallmatrix})(x,y)\right)\\
&= \exp\left(\mathcal{Q}(\begin{smallmatrix}\frac{\mathbf{I}_\mathrm{d}}{\sigma^2}+\mathbf{U}+\mathbf{A}^{-1} & -\frac{\mathbf{I}_\mathrm{d}}{\sigma^2}\\ -\frac{\mathbf{I}_\mathrm{d}}{\sigma^2} & \frac{\mathbf{I}_\mathrm{d}}{\sigma^2}+\mathbf{V}+\mathbf{B}^{-1}\end{smallmatrix})(x,y)\right)\\
&= \exp\left(\mathcal{Q}(\begin{smallmatrix}\frac{\mathbf{F}}{\sigma^2} & -\frac{\mathbf{I}_\mathrm{d}}{\sigma^2}\\ -\frac{\mathbf{I}_\mathrm{d}}{\sigma^2} & \frac{\mathbf{G}}{\sigma^2}\end{smallmatrix})(x,y)\right)\\
&= \exp\left(\mathcal{Q}(\Gamma)(x,y)\right)
\end{aligned}$$

with $\Gamma \stackrel{\text{def}}{=} \begin{pmatrix}\mathbf{F}/\sigma^2 & -\mathbf{I}_\mathrm{d}/\sigma^2\\ -\mathbf{I}_\mathrm{d}/\sigma^2 & \mathbf{G}/\sigma^2\end{pmatrix}$. Moreover, since $\frac{\mathbf{G}}{2\sigma^2} \succ 0$, and the Schur complement of Γ satisfies $(\mathbf{F}-\mathbf{G}^{-1})/\sigma^2 = \mathbf{A}^{-1} \succ 0$, we have that $\Gamma \succ 0$. Therefore π is a Gaussian $\mathcal{N}(\mathbf{H})$ where $\mathbf{H} = \Gamma^{-1}$ can be obtained using the block inverse formula:

$$\begin{aligned}
\mathbf{H} &= \Gamma^{-1}\\
&= \sigma^2\begin{pmatrix}(\mathbf{F}-\mathbf{G}^{-1})^{-1} & (\mathbf{G}\mathbf{F}-\mathbf{I}_\mathrm{d})^{-1}\\ (\mathbf{F}\mathbf{G}-\mathbf{I}_\mathrm{d})^{-1} & (\mathbf{G}-\mathbf{F}^{-1})^{-1}\end{pmatrix}\\
&= \begin{pmatrix}\mathbf{A} & \mathbf{C}\\ \mathbf{C}^\top & \mathbf{B}\end{pmatrix},
\end{aligned}$$

where we used the optimality equations (4.21) and the definition of $\mathbf{C} = \mathbf{A}\mathbf{G}^{-1}$.

Let us now finally compute$\mathrm{OT}_\sigma(\alpha,\beta)$ using Lemma 4.8. Let $\mathbf{R} = \mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}}$. Using the closed form expression of $\mathbf{C}$ in (4.23), it first holds that

$$\mathbf{Z} \stackrel{\text{def}}{=} \mathbf{A}^{-\frac{1}{2}}\mathbf{C}\mathbf{A}^{\frac{1}{2}} = (\mathbf{R}+\tfrac{\sigma^4}{4}\mathbf{I}_\mathrm{d})^{\frac{1}{2}} - \tfrac{\sigma^2}{2}\mathbf{I}_\mathrm{d}. \tag{4.28}$$

Moreover, since $\mathbf{R} = \mathbf{R}^\top$, it holds that $\mathbf{Z} = \mathbf{Z}^\top$. Hence,

$$\begin{aligned}
\det\left(\begin{smallmatrix}\mathbf{A} & \mathbf{C}\\ \mathbf{C}^T & \mathbf{B}\end{smallmatrix}\right) &= \det(\mathbf{A})\det(\mathbf{B}-\mathbf{C}^\top\mathbf{A}^{-1}\mathbf{C})\\
&= \det(\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}} - \mathbf{A}^{\frac{1}{2}}\mathbf{C}^\top\mathbf{A}^{-1}\mathbf{C}\mathbf{A}^{\frac{1}{2}})\\
&= \det(\mathbf{R}-\mathbf{Z}^\top\mathbf{Z})\\
&= \det(\mathbf{R}-\mathbf{Z}^2)\\
&= \det(\sigma^2(\mathbf{R}+\frac{\sigma^4}{4}\mathbf{I}_\mathrm{d})^{\frac{1}{2}} - \frac{\sigma^4}{2}\mathbf{I}_\mathrm{d})\\
&= (\sigma^2/2)^d\det((4\mathbf{R}+\sigma^4\mathbf{I}_\mathrm{d})^{\frac{1}{2}} - \sigma^2\mathbf{I}_\mathrm{d}).
\end{aligned} \tag{4.29}$$

Since the matrices inside the determinant commute, we can use the identity $\mathbf{P}-\mathbf{Q} = (\mathbf{P}^2-\mathbf{Q}^2)(\mathbf{P}+\mathbf{Q})^{-1}$ to get rid of the negative sign. Equation (4.29) then becomes

$$\begin{aligned}
(\sigma^2/2)^d\det((4\mathbf{R}+\sigma^4\mathbf{I}_\mathrm{d})^{\frac{1}{2}} - \sigma^2\mathbf{I}_\mathrm{d}) &= (\sigma^2/2)^d\det(4\mathbf{R})\det\left(((4\mathbf{R}+\sigma^4\mathbf{I}_\mathrm{d})^{\frac{1}{2}}+\sigma^2\mathbf{I}_\mathrm{d})^{-1}\right)\\
&= (2\sigma^2)^d\det(\mathbf{A}\mathbf{B})\det\left(((4\mathbf{R}+\sigma^4\mathbf{I}_\mathrm{d})^{\frac{1}{2}}+\sigma^2\mathbf{I}_\mathrm{d})^{-1}\right).
\end{aligned}$$

Plugging this expression in (4.27), the determinants of $\mathbf{A}$ and $\mathbf{B}$ cancel out and we finally get

$$\begin{aligned}\mathfrak{B}_{\sigma^2}(\mathbf{A},\mathbf{B}) = \operatorname{Tr}(\mathbf{A}) + \operatorname{Tr}(\mathbf{B}) - \operatorname{Tr}(4\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}} + \sigma^4\mathbf{I}_{\mathrm{d}})^{\frac{1}{2}} \\ + \sigma^2 \log\det\Big((4\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}} + \sigma^4\mathbf{I}_{\mathrm{d}})^{\frac{1}{2}} + \sigma^2\mathbf{I}_{\mathrm{d}}\Big) + \sigma^2 d(1-\log(2\sigma^2)).\end{aligned}$$

3.2 Properties of $\mathfrak{B}_{\sigma^2}$.

Theorem 4.2 shows that π has a Gaussian density. Proposition 4.9 allows to reformulate this optimization problem over couplings in $\mathbb{R}^{d\times d}$ with a positivity constraint.

Proposition 4.9. *Let $\alpha = \mathcal{N}(0,\mathbf{A}), \beta = \mathcal{N}(0,\mathbf{B})$, and $\sigma^2 > 0$. Then,*

$$\begin{aligned}\mathrm{OT}_\sigma(\alpha,\beta) &= \mathfrak{B}_{\sigma^2}(\mathbf{A},\mathbf{B}) \\ &= \min_{\mathbf{C}:\left(\begin{smallmatrix}\mathbf{A} & \mathbf{C}\\ \mathbf{C}^T & \mathbf{B}\end{smallmatrix}\right)\geq 0} \Big\{\operatorname{Tr}(\mathbf{A}+\mathbf{B}-2\mathbf{C}) + \sigma^2(\log\det\mathbf{A}\mathbf{B} - \log\det\left(\begin{smallmatrix}\mathbf{A} & \mathbf{C}\\ \mathbf{C}^T & \mathbf{B}\end{smallmatrix}\right))\Big\} \qquad (4.30)\\ &= \min_{\mathbf{K}\in\mathbb{R}^{d\times d}:\|\mathbf{K}\|_{op}\leq 1} \operatorname{Tr}\mathbf{A} + \operatorname{Tr}\mathbf{B} - 2\operatorname{Tr}\mathbf{A}^{\frac{1}{2}}\mathbf{K}\mathbf{B}^{\frac{1}{2}} - \sigma^2\ln\det(\mathbf{I}_{\mathrm{d}} - \mathbf{K}\mathbf{K}^\top). \qquad (4.31)\end{aligned}$$

Moreover, both (4.30) *and* (4.31) *are convex problems.*

Proof. Using Lemma 4.8, (4.7) becomes

$$\begin{aligned}\mathfrak{B}^2_\sigma(\mathbf{A},\mathbf{B}) = \min_{\mathbf{C}:\left(\begin{smallmatrix}\mathbf{A} & \mathbf{C}\\ \mathbf{C}^T & \mathbf{B}\end{smallmatrix}\right)\geq 0} \Big\{&\operatorname{Tr}(\mathbf{A}) + \operatorname{Tr}(\mathbf{B}) - 2\operatorname{Tr}(\mathbf{C}) \\ &+ \sigma^2(\log\det\mathbf{A} + \log\det\mathbf{B} - \log\det\left(\begin{smallmatrix}\mathbf{A} & \mathbf{C}\\ \mathbf{C}^T & \mathbf{B}\end{smallmatrix}\right))\Big\},\end{aligned}$$

which gives (4.30).

Let us now prove (4.31). A necessary and sufficient condition for $\left(\begin{smallmatrix}\mathbf{A} & \mathbf{C}\\ \mathbf{C}^T & \mathbf{B}\end{smallmatrix}\right) \geq 0$ is that there exists a contraction $\mathbf{K}$ (i.e. $\mathbf{K}\in\mathbb{R}^{d\times d}$ s.t. $\|\mathbf{K}\|_{\mathrm{op}}\leq 1$) such that $\mathbf{C} = \mathbf{A}^{\frac{1}{2}}\mathbf{K}\mathbf{B}^{\frac{1}{2}}$ [Bhatia, 2007, Ch. 1]. With this parameterization, we have (using Schur complements) that

$$\begin{aligned}\det\left(\begin{smallmatrix}\mathbf{A} & \mathbf{C}\\ \mathbf{C}^T & \mathbf{B}\end{smallmatrix}\right) &= \det\mathbf{B}\det(\mathbf{A} - \mathbf{C}\mathbf{B}^{-1}\mathbf{C}^\top) \\ &= \det\mathbf{B}\det\mathbf{A}\det(\mathbf{I}_{\mathrm{d}} - \mathbf{K}\mathbf{K}^\top).\end{aligned}$$

Hence, injecting this in Equation (4.30), we have the following equivalent problem:

$$\mathfrak{B}^2_\sigma(\mathbf{A},\mathbf{B}) = \min_{\mathbf{K}\in\mathbb{R}^{d\times d}:\|\mathbf{K}\|_{\mathrm{op}}\leq 1} \operatorname{Tr}\mathbf{A} + \operatorname{Tr}\mathbf{B} - 2\operatorname{Tr}\mathbf{A}^{\frac{1}{2}}\mathbf{K}\mathbf{B}^{\frac{1}{2}} - \sigma^2\ln\det(\mathbf{I}_{\mathrm{d}} - \mathbf{K}\mathbf{K}^\top). \qquad (4.32)$$

Let's prove that both problems are convex.

- (4.30): The set $\{\mathbf{C} : \left(\begin{smallmatrix}\mathbf{A} & \mathbf{C}\\ \mathbf{C}^T & \mathbf{B}\end{smallmatrix}\right) \geq 0\}$ is convex, since $\left(\begin{smallmatrix}\mathbf{A} & \mathbf{C}_1\\ \mathbf{C}_1^T & \mathbf{B}\end{smallmatrix}\right) \geq 0$ and $\left(\begin{smallmatrix}\mathbf{A} & \mathbf{C}_2\\ \mathbf{C}_2^T & \mathbf{B}\end{smallmatrix}\right) \geq 0$ implies that

  $$\left(\begin{smallmatrix}\mathbf{A} & (1-\theta)\mathbf{C}_1+\theta\mathbf{C}_2\\ (1-\theta)\mathbf{C}_1^T+\theta\mathbf{C}_2^T & \mathbf{B}\end{smallmatrix}\right) = (1-\theta)\left(\begin{smallmatrix}\mathbf{A} & \mathbf{C}_1\\ \mathbf{C}_1^T & \mathbf{B}\end{smallmatrix}\right) + \theta\left(\begin{smallmatrix}\mathbf{A} & \mathbf{C}_2\\ \mathbf{C}_2^T & \mathbf{B}\end{smallmatrix}\right) \geq 0.$$

  Following the same decomposition, the concavity of the $\log\det$ function implies that $\mathbf{C} \to \log\det\left(\begin{smallmatrix}\mathbf{A} & \mathbf{C}\\ \mathbf{C}^T & \mathbf{B}\end{smallmatrix}\right)$ is concave, and hence that the objective function of (4.30) is convex;

- (4.31): The ball $\mathcal{B}_{\text{op}} \stackrel{\text{def}}{=} \{\mathbf{K} \in \mathbb{R}^{d\times d} : \|\mathbf{K}\|_{\text{op}} \leq 1\}$ is obviously convex. Hence, there remains to prove that $f(\mathbf{K}) : \mathbf{K} \in \mathcal{B}_{\text{op}} \to \log\det(\mathbf{I}_\text{d} - \mathbf{K}\mathbf{K}^\top)$ is concave. Indeed, it holds that $f(\mathbf{K}) = \log\det \left(\begin{smallmatrix} \mathbf{I}_\text{d} & \mathbf{K} \\ \mathbf{K}^T & \mathbf{I}_\text{d} \end{smallmatrix}\right)$. Hence, $\forall \mathbf{K}, \mathbf{H} \in \mathcal{B}_{\text{op}}, \forall t \in [0,1]$,

$$\begin{aligned} f((1-t)\mathbf{K} + t\mathbf{H}) &= \log\det\left\{(1-t)\left(\begin{smallmatrix} \mathbf{I}_\text{d} & \mathbf{K} \\ \mathbf{K}^T & \mathbf{I}_\text{d} \end{smallmatrix}\right) + t\left(\begin{smallmatrix} \mathbf{I}_\text{d} & \mathbf{H} \\ \mathbf{H}^T & \mathbf{I}_\text{d} \end{smallmatrix}\right)\right\} \\ &\geq (1-t)\log\det\left(\begin{smallmatrix} \mathbf{I}_\text{d} & \mathbf{K} \\ \mathbf{K}^T & \mathbf{I}_\text{d} \end{smallmatrix}\right) + t\log\det\left(\begin{smallmatrix} \mathbf{I}_\text{d} & \mathbf{H} \\ \mathbf{H}^T & \mathbf{I}_\text{d} \end{smallmatrix}\right) \\ &= (1-t)f(\mathbf{K}) + tf(\mathbf{H}), \end{aligned}$$

  where the second line follows from the concavity of $\log\det$.

□

We now study the convexity and differentiability of $\mathfrak{B}_{\sigma^2}$, which are more conveniently derived from the dual problem of (4.30) given as a positive definite program:

Proposition 4.10. *The dual problem of* (4.30) *can be written with no duality gap as*

$$\begin{aligned} \max_{\mathbf{F},\mathbf{G}\succ 0}\Big\{ &-\langle \mathbf{F}, \mathbf{A}\rangle - \langle \mathbf{G}, \mathbf{B}\rangle + \sigma^2 \log\det(\mathbf{F}\mathbf{G} - \mathbf{I}_\text{d}) \\ &+ \operatorname{Tr}(\mathbf{A}) + \operatorname{Tr}(\mathbf{B}) + \sigma^2\log\det\mathbf{A}\mathbf{B} + 2d\sigma^2(1-\log\sigma^2))\Big\}. \end{aligned} \tag{4.33}$$

Proof. By Proposition 4.9, (4.30) is convex, hence strong duality holds. Ignoring the terms not depending on $\mathbf{C}$, problem (4.30) can be written using the redundant parameterization $\mathbf{X} = \left(\begin{smallmatrix} \mathbf{X}_1 & \mathbf{X}_2 \\ \mathbf{X}_3 & \mathbf{X}_4 \end{smallmatrix}\right)$:

$$\mathfrak{D}(\mathbf{A},\mathbf{B}) \stackrel{\text{def}}{=} \min_{\substack{\mathbf{X}\succ 0 \\ \mathbf{X}_1=\mathbf{A},\mathbf{X}_4=\mathbf{B}}} -\operatorname{Tr}(\mathbf{X}_2) - \operatorname{Tr}(\mathbf{X}_3) - \sigma^2\log\det(\mathbf{X}) \tag{4.34}$$

$$= \min_{\substack{\mathbf{X}\succ 0 \\ \mathbf{X}_1=\mathbf{A},\mathbf{X}_4=\mathbf{B}}} -\langle \mathbf{X}, \left(\begin{smallmatrix} 0 & \mathbf{I}_\text{d} \\ \mathbf{I}_\text{d} & 0 \end{smallmatrix}\right)\rangle - \sigma^2\log\det(\mathbf{X}) \tag{4.35}$$

$$= \min_{\substack{\mathbf{X}\succ 0 \\ \mathbf{X}_1=\mathbf{A},\mathbf{X}_4=\mathbf{B}}} \mathcal{F}(\mathbf{X}), \tag{4.36}$$

where the functional $\mathcal{F}$ is convex. Moreover, its Legendre transform is given by

$$\begin{aligned} \mathcal{F}^\star(\mathbf{Y}) &= \max_{\mathbf{X}\succ 0}\langle \mathbf{X}, \mathbf{Y} + \left(\begin{smallmatrix} 0 & \mathbf{I}_\text{d} \\ \mathbf{I}_\text{d} & 0 \end{smallmatrix}\right)\rangle + \sigma^2\log\det(\mathbf{X}) \\ &= \left(-\sigma^2\log\det\right)^\star\left(\mathbf{Y} + \left(\begin{smallmatrix} 0 & \mathbf{I}_\text{d} \\ \mathbf{I}_\text{d} & 0 \end{smallmatrix}\right)\right) \\ &= \sigma^2(-\log\det)^\star\left(\frac{1}{\sigma^2}\left(\mathbf{Y} + \left(\begin{smallmatrix} 0 & \mathbf{I}_\text{d} \\ \mathbf{I}_\text{d} & 0 \end{smallmatrix}\right)\right)\right) \\ &= -\sigma^2\log\det\left(-\frac{1}{\sigma^2}\left(\mathbf{Y} + \left(\begin{smallmatrix} 0 & \mathbf{I}_\text{d} \\ \mathbf{I}_\text{d} & 0 \end{smallmatrix}\right)\right)\right) - 2\sigma^2 d \\ &= -\sigma^2\log\det\left(-\left(\mathbf{Y} + \left(\begin{smallmatrix} 0 & \mathbf{I}_\text{d} \\ \mathbf{I}_\text{d} & 0 \end{smallmatrix}\right)\right)\right) - 2d(\sigma^2 - \sigma^2\log(\sigma^2)). \end{aligned}$$

Let $\mathcal{H}$ be the linear operator $\mathcal{H} : \mathbf{X} \mapsto (\mathbf{X}_1, \mathbf{X}_4)$. Its conjugate operator is defined on $\mathcal{S}^d_{++} \times \mathcal{S}^d_{++}$ and is given by $\mathcal{H}^\star(\mathbf{F},\mathbf{G}) = \left(\begin{smallmatrix} \mathbf{F} & 0 \\ 0 & \mathbf{G} \end{smallmatrix}\right)$. Therefore, Fenchel's duality theorem leads to

$$\begin{aligned} \mathfrak{D}(\mathbf{A},\mathbf{B}) &= \max_{\mathbf{F},\mathbf{G}\succ 0} -\langle \mathbf{F}, \mathbf{A}\rangle - \langle \mathbf{G}, \mathbf{B}\rangle - \mathcal{F}^\star\left(-\mathcal{H}^\star(\mathbf{F},\mathbf{G})\right) \\ &= \max_{\mathbf{F},\mathbf{G}\succ 0} -\langle \mathbf{F}, \mathbf{A}\rangle - \langle \mathbf{G}, \mathbf{B}\rangle + \sigma^2\log\det\left(\begin{smallmatrix} \mathbf{F} & -\mathbf{I}_\text{d} \\ -\mathbf{I}_\text{d} & \mathbf{G} \end{smallmatrix}\right) + 2\sigma^2 d(1-\log(\sigma^2)) \\ &= \max_{\mathbf{F},\mathbf{G}\succ 0} -\langle \mathbf{F}, \mathbf{A}\rangle - \langle \mathbf{G}, \mathbf{B}\rangle + \sigma^2\log\det(\mathbf{F}\mathbf{G} - \mathbf{I}_\text{d}) + 2\sigma^2 d(1-\log(\sigma^2)), \end{aligned} \tag{4.37}$$

where the last equality follows from the fact that $\mathbf{I}_\mathrm{d}$ and $\mathbf{G}$ commute. Therefore, reinserting the discarded trace terms, the dual problem of (4.30) can be written as

$$\max_{\mathbf{F},\mathbf{G}\succ 0}\Big\{ -\langle \mathbf{F},\, \mathbf{A}\rangle - \langle \mathbf{G},\, \mathbf{B}\rangle + \sigma^2 \log\det\left(\mathbf{F}\mathbf{G} - \mathbf{I}_\mathrm{d}\right) \\ + \mathrm{Tr}(\mathbf{A}) + \mathrm{Tr}(\mathbf{B}) + \sigma^2 \log\det \mathbf{A}\mathbf{B} + 2d\sigma^2(1 - \log\sigma^2))\Big\}.$$

□

Feydy et al. [2019] showed that on compact spaces, the gradient of OT_σ is given by the optimal dual potentials. This result was later extended by Janati et al. [2020a] to sub-Gaussian measures with unbounded supports. The following proposition re-establishes this statement for Gaussian measures.

Proposition 4.11. *Assume $\sigma > 0$ and consider the pair $\mathbf{U}, \mathbf{V}$ of Corollary 4.7. Then*

(i) The optimal pair $(\mathbf{F}^*, \mathbf{G}^*)$ of (4.33) is a solution to the fixed point problem (4.21);

(ii) $\mathfrak{B}_{\sigma^2}$ is differentiable and $\nabla\mathfrak{B}_{\sigma^2}(\mathbf{A},\mathbf{B}) = -(\sigma^2\mathbf{U}, \sigma^2\mathbf{V})$. Thus,

$$\nabla_{\mathbf{A}}\mathfrak{B}_{\sigma^2}(\mathbf{A},\mathbf{B}) = \mathbf{I}_\mathrm{d} - \mathbf{B}^{\frac{1}{2}}\left((\mathbf{B}^{\frac{1}{2}}\mathbf{A}\mathbf{B}^{\frac{1}{2}} + \frac{\sigma^4}{4}\mathbf{I}_\mathrm{d})^{\frac{1}{2}} + \frac{\sigma^2}{2}\mathbf{I}_\mathrm{d}\right)^{-1}\mathbf{B}^{\frac{1}{2}};$$

(iii) $(\mathbf{A},\mathbf{B}) \mapsto \mathfrak{B}_{\sigma^2}(\mathbf{A},\mathbf{B})$ is convex in $\mathbf{A}$ and in $\mathbf{B}$, but not jointly;

(iv) For a fixed $\mathbf{B}$ with spectral decomposition $\mathbf{B} = \mathbf{P}\Sigma\mathbf{P}^\top$, the function $\phi_{\mathbf{B}} : \mathbf{A} \mapsto \mathfrak{B}_{\sigma^2}(\mathbf{A},\mathbf{B})$ is minimized at $\mathbf{A}_0 = \mathbf{P}(\Sigma - \sigma^2\mathbf{I}_\mathrm{d})_+\mathbf{P}^\top$ where the thresholding operator $_+$ is defined by $x_+ = \max(x, 0)$ for any $x \in \mathbb{R}$ and extended element-wise to diagonal matrices.

Proof.

(i) *Optimality:* Canceling out the gradients in (4.33) leads to the following optimality conditions:

$$\begin{aligned} -\mathbf{A} + \sigma^2\mathbf{G}(\mathbf{F}\mathbf{G} - \mathbf{I}_\mathrm{d})^{-1} &= 0 \\ -\mathbf{B} + \sigma^2(\mathbf{F}\mathbf{G} - \mathbf{I}_\mathrm{d})^{-1}\mathbf{F} &= 0, \end{aligned} \tag{4.38}$$

i.e.

$$\begin{aligned} \mathbf{F} &= \sigma^2\mathbf{A}^{-1} + \mathbf{G}^{-1} \\ \mathbf{G} &= \sigma^2\mathbf{B}^{-1} + \mathbf{F}^{-1}. \end{aligned} \tag{4.39}$$

Thus $(\mathbf{F}, \mathbf{G})$ is a solution of the Sinkhorn fixed point equation (4.21).

(ii) *Differentiabilty:* Using Danskin's theorem [Danskin, 1967] on problem (4.33) leads to the formula of the gradient as a function of the optimal dual pair $(\mathbf{F}, \mathbf{G})$. Indeed, keeping in mind that $\nabla_{\mathbf{A}} \log\det(\mathbf{A}) = -\mathbf{A}^{-1}$ and using the change of variable of Proposition 4.5, we recover the dual potentials of Corollary 4.7:

$$\begin{aligned} \nabla\mathfrak{B}_{\sigma^2}(\mathbf{A},\mathbf{B}) &= \left(\mathbf{I}_\mathrm{d} - \mathbf{F}^* + \sigma^2\mathbf{A}^{-1}, \mathbf{I}_\mathrm{d} - \mathbf{G}^* + \sigma^2\mathbf{B}^{-1}\right) \\ &= -\sigma^2(\mathbf{U},\mathbf{V}). \end{aligned}$$

Using Corollary 4.7, it holds that

$$\begin{aligned}\nabla_{\mathbf{A}}\mathfrak{B}_{\sigma^2}(\mathbf{A},\mathbf{B}) &= -\sigma^2\mathbf{U}\\ &= \mathbf{I}_\mathrm{d} - \mathbf{B}(\mathbf{C}+\sigma^2\mathbf{I}_\mathrm{d})^{-1}\\ &= \mathbf{I}_\mathrm{d} - \mathbf{B}\left((\mathbf{A}\mathbf{B}+\frac{\sigma^4}{4}\mathbf{I}_\mathrm{d})^{\frac{1}{2}}+\frac{\sigma^2}{2}\mathbf{I}_\mathrm{d}\right)^{-1}\\ &= \mathbf{I}_\mathrm{d} - \mathbf{B}^{\frac{1}{2}}\left((\mathbf{B}^{\frac{1}{2}}\mathbf{A}\mathbf{B}^{\frac{1}{2}}+\frac{\sigma^4}{4}\mathbf{I}_\mathrm{d})^{\frac{1}{2}}+\frac{\sigma^2}{2}\mathbf{I}_\mathrm{d}\right)^{-1}\mathbf{B}^{\frac{1}{2}}\\ &= \mathbf{I}_\mathrm{d} - \mathbf{B}^{\frac{1}{2}}\left(\mathbf{D}^{\frac{1}{2}}+\frac{\sigma^2}{2}\mathbf{I}_\mathrm{d}\right)^{-1}\mathbf{B}^{\frac{1}{2}},\end{aligned}$$

where $\mathbf{D} \overset{\text{def}}{=} \mathbf{B}^{\frac{1}{2}}\mathbf{A}\mathbf{B}^{\frac{1}{2}} + \frac{\sigma^4}{4}\mathbf{I}_\mathrm{d}$.

(iii) *Convexity:* Assume without loss of generality that $\mathbf{B}$ is fixed and let $G : \mathbf{B} \mapsto \nabla_{\mathbf{A}}\mathfrak{B}_{\sigma^2}(\mathbf{A},\mathbf{B})$. As long as $\sigma > 0$, G is differentiable as a composition of differentiable functions. Let's show that the Hessian of $\psi : \mathbf{A} \mapsto \mathfrak{B}_{\sigma^2}(\mathbf{A},\mathbf{B})$ is a positive quadratic form. Take a direction $\mathbf{H} \in \mathcal{S}_+^d$. It holds that

$$\begin{aligned}\nabla^2_{\mathbf{A}}\mathfrak{B}_{\sigma^2}(\mathbf{A},\mathbf{B})(\mathbf{H},\mathbf{H}) &= \langle \mathbf{H}, \mathrm{Jac}_G(\mathbf{A})(\mathbf{H})\rangle\\ &= \mathrm{Tr}(\mathbf{H}\,\mathrm{Jac}_G(\mathbf{A})(\mathbf{H})).\end{aligned}$$

For the sake of clarity, let us write $G(\mathbf{A}) = \mathbf{I}_\mathrm{d} - L(W(\phi(\mathbf{A})))$ with the following intermediary functions:

$$\begin{aligned}L &: \mathbf{A} \mapsto \mathbf{B}^{\frac{1}{2}}\mathbf{A}\mathbf{B}^{\frac{1}{2}}\\ Q &: \mathbf{A} \mapsto \mathbf{A}^{\frac{1}{2}}\\ \phi &: \mathbf{A} \mapsto Q(L(\mathbf{A}) + \frac{\sigma^4}{4}\mathbf{I}_\mathrm{d})\\ W &: \mathbf{A} \mapsto (\mathbf{A}+\frac{\sigma^2}{2}\mathbf{I}_\mathrm{d})^{-1}.\end{aligned}$$

Moreover, their derivatives are given by

$$\begin{aligned}\mathrm{Jac}_L(\mathbf{A})(\mathbf{H}) &= \mathbf{B}^{\frac{1}{2}}\mathbf{H}\mathbf{B}^{\frac{1}{2}}\\ \mathrm{Jac}_W(\mathbf{A})(\mathbf{H}) &= -(\mathbf{A}+\frac{\sigma^2}{2}\mathbf{I}_\mathrm{d})^{-1}\mathbf{H}(\mathbf{A}+\frac{\sigma^2}{2}\mathbf{I}_\mathrm{d})^{-1}\\ \mathrm{Jac}_Q(\mathbf{A})(\mathbf{H}) &= \mathbf{Z},\end{aligned}$$

where $\mathbf{Z} \in \mathcal{S}_+^d$ is the unique solution of the Sylvester equation $\mathbf{Z}\mathbf{A}^{\frac{1}{2}} + \mathbf{A}^{\frac{1}{2}}\mathbf{Z} = \mathbf{H}$.

Using the chain rule:

$$\begin{aligned}\mathrm{Jac}_G(\mathbf{A})(\mathbf{H}) &= -\,\mathrm{Jac}_L(W(\phi(\mathbf{A})))(\mathrm{Jac}_W(\phi(\mathbf{A}))(\mathrm{Jac}_\phi(\mathbf{A})(\mathbf{H})))\\ &= -\mathbf{B}^{\frac{1}{2}}\,\mathrm{Jac}_W(\phi(\mathbf{A}))(\mathrm{Jac}_\phi(\mathbf{A})(\mathbf{H}))\mathbf{B}^{\frac{1}{2}}\\ &= \mathbf{B}^{\frac{1}{2}}\left(\phi(\mathbf{A})+\frac{\sigma^2}{2}\mathbf{I}_\mathrm{d}\right)^{-1}\mathrm{Jac}_\phi(\mathbf{A})(\mathbf{H})\left(\phi(\mathbf{A})+\frac{\sigma^2}{2}\mathbf{I}_\mathrm{d}\right)^{-1}\mathbf{B}^{\frac{1}{2}}\\ &= \mathbf{B}^{\frac{1}{2}}\left(\mathbf{D}^{\frac{1}{2}}+\frac{\sigma^2}{2}\mathbf{I}_\mathrm{d}\right)^{-1}\mathrm{Jac}_\phi(\mathbf{A})(\mathbf{H})\left(\mathbf{D}^{\frac{1}{2}}+\frac{\sigma^2}{2}\mathbf{I}_\mathrm{d}\right)^{-1}\mathbf{B}^{\frac{1}{2}}.\end{aligned}$$

Again using the chain rule:

$$\begin{aligned}\mathbf{Y} \overset{\text{def}}{=} \mathrm{Jac}_\phi(\mathbf{A})(\mathbf{H}) &= \mathrm{Jac}_Q(L(\mathbf{A}) + \frac{\sigma^4}{4}\mathbf{I}_\mathrm{d})((\mathrm{Jac}_L(\mathbf{A}))(\mathbf{H})) \\ &= \mathrm{Jac}_Q(L(\mathbf{A}) + \frac{\sigma^4}{4}\mathbf{I}_\mathrm{d})(\mathbf{B}^{\frac{1}{2}}\mathbf{H}\mathbf{B}^{\frac{1}{2}}) \\ &= \mathrm{Jac}_Q(\mathbf{D})(\mathbf{B}^{\frac{1}{2}}\mathbf{H}\mathbf{B}^{\frac{1}{2}}).\end{aligned}$$

Therefore, $\mathbf{Y} \succ 0$ is the unique solution of the Sylvester equation:

$$\mathbf{Y}\mathbf{D}^{\frac{1}{2}} + \mathbf{D}^{\frac{1}{2}}\mathbf{Y} = \mathbf{B}^{\frac{1}{2}}\mathbf{H}\mathbf{B}^{\frac{1}{2}}.$$

Combining everything, we get

$$\begin{aligned}\nabla^2_{\mathbf{A}}\mathfrak{B}_{\sigma^2}(\mathbf{A},\mathbf{B})(\mathbf{H},\mathbf{H}) &= \langle \mathbf{H}, \mathrm{Jac}_G(\mathbf{A})(\mathbf{H})\rangle \\ &= \mathrm{Tr}\left(\mathbf{H}\,\mathrm{Jac}_G(\mathbf{A})(\mathbf{H})\right) \\ &= \mathrm{Tr}\left(\mathbf{H}\mathbf{B}^{\frac{1}{2}}\left(\mathbf{D}^{\frac{1}{2}} + \frac{\sigma^2}{2}\mathbf{I}_\mathrm{d}\right)^{-1}\mathbf{Y}\left(\mathbf{D}^{\frac{1}{2}} + \frac{\sigma^2}{2}\mathbf{I}_\mathrm{d}\right)^{-1}\mathbf{B}^{\frac{1}{2}}\right) \\ &= \mathrm{Tr}\left(\mathbf{B}^{\frac{1}{2}}\mathbf{H}\mathbf{B}^{\frac{1}{2}}\left(\mathbf{D}^{\frac{1}{2}} + \frac{\sigma^2}{2}\mathbf{I}_\mathrm{d}\right)^{-1}\mathbf{Y}\left(\mathbf{D}^{\frac{1}{2}} + \frac{\sigma^2}{2}\mathbf{I}_\mathrm{d}\right)^{-1}\right).\end{aligned}$$

Since $\mathbf{H}$ and $\mathbf{Y}$ are positive, the matrices

$$\mathbf{B}^{\frac{1}{2}}\mathbf{H}\mathbf{B}^{\frac{1}{2}} \quad \text{and} \quad \left(\mathbf{D}^{\frac{1}{2}} + \frac{\sigma^2}{2}\mathbf{I}_\mathrm{d}\right)^{-1}\mathbf{Y}\left(\mathbf{D}^{\frac{1}{2}} + \frac{\sigma^2}{2}\mathbf{I}_\mathrm{d}\right)^{-1}$$

are positive semi-definite as well. Their product is similar to a positive semi-definite matrix, therefore the trace above is non-negative.

Hence, given that $\mathbf{A}$ and $\mathbf{H}$ are arbitrary positive semi-definite matrices, it holds that

$$\nabla^2_{\mathbf{A}}\mathfrak{B}_{\sigma^2}(\mathbf{A},\mathbf{B})(\mathbf{H},\mathbf{H}) \geq 0.$$

Therefore, $\mathbf{A} \mapsto \mathfrak{B}_{\sigma^2}(\mathbf{A},\mathbf{B})$ is convex.

Counter-example of joint convexity: If $\mathfrak{B}_{\sigma^2}$ were jointly convex, then $\delta \overset{\text{def}}{=}: \mathbf{A} \to \mathfrak{B}_{\sigma^2}(\mathbf{A},\mathbf{A})$ would be a convex function. In the 1-dimensional case with $\sigma = 1$, one can see that this would be equivalent to $x \to \ln((x^2+1)^{1/2}+1) - (x^2+1)^{1/2}$ being convex, whereas it is in fact strictly concave.

(iv) *Minimizer of $\phi_{\mathbf{B}}$:* With fixed $\mathbf{B}$, cancelling the gradient of $\phi_{\mathbf{B}} \overset{\text{def}}{=}: \mathbf{A} \mapsto \mathfrak{B}_{\sigma^2}(\mathbf{A},\mathbf{B})$ leads to $\mathbf{A} = \mathbf{B} - \sigma^2\mathbf{I}_\mathrm{d}$ which is well defined if and only if $\mathbf{B} \succeq \sigma^2\mathbf{I}_\mathrm{d}$. However, if $\mathbf{B} - \sigma^2\mathbf{I}_\mathrm{d}$ is not positive semi-definite, let us write the eigenvalue decomposition $\mathbf{B} = \mathbf{P}\Sigma\mathbf{P}^\top$ and define $\mathbf{A}_0 \overset{\text{def}}{=} \mathbf{P}(\Sigma - \sigma^2\mathbf{I}_\mathrm{d})_+\mathbf{P}^\top$ where the operator $x_+ = \max(x,0)$ is applied element-wise. Then,

$$\begin{aligned}\nabla_{\mathbf{A}}\phi_{\mathbf{B}}(\mathbf{A}_0) &= \mathbf{I}_\mathrm{d} - \mathbf{P}\Sigma^{\frac{1}{2}}\mathbf{P}^\top\left((\mathbf{P}(\Sigma^2 - \sigma^2\Sigma)_+\mathbf{P}^\top + \frac{\sigma^4}{4}\mathbf{I}_\mathrm{d})^{\frac{1}{2}} + \frac{\sigma^2}{2}\mathbf{I}_\mathrm{d}\right)^{-1}\mathbf{P}\Sigma^{\frac{1}{2}}\mathbf{P}^\top \\ &= \mathbf{I}_\mathrm{d} - \mathbf{P}\Sigma^{\frac{1}{2}}\left(((\Sigma^2 - \sigma^2\Sigma)_+ + \frac{\sigma^4}{4}\mathbf{I}_\mathrm{d})^{\frac{1}{2}} + \frac{\sigma^2}{2}\mathbf{I}_\mathrm{d}\right)^{-1}\Sigma^{\frac{1}{2}}\mathbf{P}^\top \\ &= \mathbf{I}_\mathrm{d} - \mathbf{P}\Sigma^{\frac{1}{2}}\left((\Sigma - \sigma^2\mathbf{I}_\mathrm{d})_+ + \sigma^2\mathbf{I}_\mathrm{d}\right)^{-1}\Sigma^{\frac{1}{2}}\mathbf{P}^\top \\ &= \mathbf{P}(\mathbf{I}_\mathrm{d} - \Sigma^{\frac{1}{2}}\left((\Sigma - \sigma^2\mathbf{I}_\mathrm{d})_+ + \sigma^2\mathbf{I}_\mathrm{d}\right)^{-1}\Sigma^{\frac{1}{2}})\mathbf{P}^\top \\ &= \frac{1}{\sigma^2}\mathbf{P}(\sigma^2\mathbf{I}_\mathrm{d} - \Sigma)_+\mathbf{P}^\top.\end{aligned}$$

Thus, given that $(\Sigma - \sigma^2\mathbf{I}_\mathrm{d})_+(\sigma^2\mathbf{I}_\mathrm{d} - \Sigma)_+ = 0$, for any $\mathbf{H} \in \mathcal{S}_+^d$ it holds that

$$\begin{aligned}\langle \mathbf{H} - \mathbf{A}_0, \nabla_\mathbf{A}\phi_\mathbf{B}(\mathbf{A}_0)\rangle &= \langle \mathbf{P}^\top\mathbf{H}\mathbf{P} - (\Sigma - \sigma^2\mathbf{I}_\mathrm{d})_+, (\sigma^2\mathbf{I}_\mathrm{d} - \Sigma)_+\rangle \\ &= \langle \mathbf{P}^\top\mathbf{H}\mathbf{P}, (\sigma^2\mathbf{I}_\mathrm{d} - \Sigma)_+\rangle \\ &= \mathrm{Tr}(\mathbf{P}^\top\mathbf{H}\mathbf{P}(\sigma^2\mathbf{I}_\mathrm{d} - \Sigma)_+) \geq 0,\end{aligned}$$

where the last inequality holds since both matrices are positive semi-definite. Given that $\phi_\mathbf{B}$ is convex, the first order optimality condition holds so $\phi_\mathbf{B}$ is minimized at $\mathbf{A}_0$.

□

3.3 Debiased Sinkhorn Barycenters

When $\mathbf{A}$ and $\mathbf{B}$ are not singular, we recover the gradient of the Bures metric given in (4.6) by letting $\sigma \to 0$ in $\nabla_\mathbf{A}\mathfrak{B}_{\sigma^2}(\mathbf{A}, \mathbf{B})$. Moreover, (iv) illustrates the entropy bias of $\mathfrak{B}_{\sigma^2}$. Feydy et al. [2019] showed that it can be circumvented by considering the Sinkhorn divergence:

$$S_\sigma : (\alpha, \beta) \mapsto \mathrm{OT}_\sigma(\alpha, \beta) - \frac{1}{2}(\mathrm{OT}_\sigma(\alpha, \alpha) + \mathrm{OT}_\sigma(\beta, \beta)), \tag{4.40}$$

which is non-negative and equals 0 if and only if $\alpha = \beta$. Using the differentiability and convexity of S_σ on sub-Gaussian measures [Janati et al., 2020a], we conclude this section by showing that the debiased Sinkhorn barycenter of Gaussian measures remains Gaussian.

Theorem 4.12. *Consider the restriction of* OT_σ *to the set of sub-Gaussian measures*

$$\mathcal{G} \stackrel{\text{def}}{=} \{\mu \in \mathcal{P}_2 | \exists q > 0,\, \mathbb{E}_\mu(e^{q\|X\|^2}) < +\infty\}$$

and K *Gaussian measures* $\alpha_k \sim \mathcal{N}(\mathbf{a}_k, \mathbf{A}_k)$ *with a sequence of positive weights* $(w_k)_{k=1,2,\dots,K}$ *summing to 1. Then, the weighted debiased barycenter defined by*

$$\beta \stackrel{\text{def}}{=} \mathrm{argmin}_{\beta\in\mathcal{G}} \sum_{k=1} w_k S_\sigma(\alpha_k, \beta) \tag{4.41}$$

is a Gaussian measure given by $\mathcal{N}\left(\sum_{k=1}^K w_k\mathbf{a}_k, \mathbf{B}\right)$ *where* $\mathbf{B} \in \mathcal{S}_+^d$ *is a solution of the equation*

$$\sum_{k=1}^K w_k(\mathbf{B}^{\frac{1}{2}}\mathbf{A}_k\mathbf{B}^{\frac{1}{2}} + \frac{\sigma^4}{4}\mathbf{I}_\mathrm{d})^{\frac{1}{2}} = (\mathbf{B}^2 + \frac{\sigma^4}{4}\mathbf{I}_\mathrm{d})^{\frac{1}{2}}. \tag{4.42}$$

Proof. This theorem is a generalization of [Janati et al., 2020a, Thm 3] for multivariate Gaussian measures. First, we are going to break it down using the centering Lemma 4.1. For any probability measure μ, let $\bar{\mu}$ denote its centered counterpart. The debiased barycenter problem is equivalent to

$$\begin{aligned}&\min_{\beta\in\mathcal{G}} \sum_{k=1}^K w_k S_\sigma(\alpha_k, \beta) \\ &= \min_{\beta\in\mathcal{G}} \sum_{k=1}^K w_k \mathrm{OT}_\sigma(\alpha_k, \beta) - \frac{1}{2}(\mathrm{OT}_\sigma(\alpha_k, \alpha_k) + \mathrm{OT}_\sigma(\beta, \beta)) \\ &= \min_{\beta\in\mathcal{G}} \sum_{k=1}^K w_k\|\mathbf{a}_k - \mathbb{E}_\beta(X)\|^2 + w_k\mathrm{OT}_\sigma(\bar{\alpha_k}, \bar{\beta}) - \frac{1}{2}(w_k\mathrm{OT}_\sigma(\bar{\alpha_k}, \bar{\alpha_k}) + \mathrm{OT}_\sigma(\bar{\beta}, \bar{\beta})) \\ &= \min_{\substack{\mathbf{b}\in\mathbb{R}^d \\ \beta\in\mathcal{G}}} \sum_{k=1}^K w_k\|\mathbf{a}_k - \mathbf{b}\|^2 + w_k\mathrm{OT}_\sigma(\bar{\alpha_k}, \bar{\beta}) - \frac{1}{2}(w_k\mathrm{OT}_\sigma(\bar{\alpha_k}, \bar{\alpha_k}) + \mathrm{OT}_\sigma(\bar{\beta}, \bar{\beta}))\end{aligned} \tag{4.43}$$

Therefore, since both arguments are independent, we can first minimize over $\mathbf{b}$ to obtain $\mathbf{b} = \sum_{k=1}^K w_k \mathbf{a}_k$. Without loss of generality, we assume from now on that $\mathbf{a}_k = 0$ for all k.

The rest of this proof is adapted from [Janati et al., 2020a, Thm 3], to $d \geq 1$. Janati et al. [2020a] showed that for sub-Gaussian measures S_σ is convex (w.r.t. one measure at a time) and admits first variations: a function $F : \mathcal{G} \to \mathbb{R}$ has a first variation at α if there exists $\frac{\delta F(\alpha)}{\delta \alpha} \in \mathcal{C}(\mathbb{R}^d)$ such that for any displacement $t\chi$ with $t > 0$ and $\chi = \tilde{\alpha} - \alpha$ with $\tilde{\alpha} \in \mathcal{G}$ verifying

$$F(\alpha + t\chi) = F(\alpha) + t\langle \chi, \frac{\delta F(\alpha)}{\delta \alpha} \rangle + o(t) \ , \tag{4.44}$$

where $\langle \chi, \frac{\delta F(\alpha)}{\delta \alpha}(\alpha) \rangle = \int_{\mathbb{R}^d} \frac{\delta F(\alpha)}{\delta \alpha} \mathrm{d}\chi$ (see [Santambrogio, 2015, §7.2]).

Moreover, F is convex if and only if for any $\alpha, \alpha' \in \mathcal{G}$:

$$F(\alpha) \geq F(\alpha') + \langle \alpha - \alpha', \frac{\delta F(\alpha')}{\delta \alpha} \rangle \ , \tag{4.45}$$

Let (f_k, g_k) denote the potentials associated with $\mathrm{OT}_\sigma(\alpha_k, \beta)$, and h_β the autocorrelation potential associated with $\mathrm{OT}_\sigma(\beta, \beta)$. If β is sub-Gaussian, it holds that $\frac{\delta S_\sigma(\alpha_k, \beta)}{\delta \beta} = g_k - h_\beta$. Therefore, from (4.45) a probability measure β is the debiased barycenter if and only if for any direction $\mu \in \mathcal{G}$, the following optimality condition holds:

$$\begin{aligned} &\langle \sum_{k=1}^K w_k \frac{\delta S_\sigma(\alpha_k, \beta)}{\delta \beta}, \mu - \beta \rangle \geq 0 \\ \Leftrightarrow &\sum_{k=1}^K w_k \langle g_k - h_\beta, \mu - \beta \rangle \geq 0 \end{aligned} \tag{4.46}$$

Moreover, the potentials $(f_k), (g_k)$ and h must verify the Sinkhorn optimality conditions (4.10) for all k and for all x β-a.s and y α-a.s:

$$\begin{aligned} e^{\frac{f_k(x)}{2\sigma^2}} \left(\int_{\mathbb{R}^d} e^{-\frac{\|x-y\|^2 + g_k(y)}{2\sigma^2}} \mathrm{d}\beta(y) \right) &= 1, \\ e^{\frac{g_k(x)}{2\sigma^2}} \left(\int_{\mathbb{R}^d} e^{-\frac{\|x-y\|^2 + f_k(y)}{2\sigma^2}} \mathrm{d}\alpha_k(y) \right) &= 1, \\ e^{\frac{h_\beta(x)}{2\sigma^2}} \left(\int_{\mathbb{R}^d} e^{-\frac{\|x-y\|^2 + h_\beta(y)}{2\sigma^2}} \mathrm{d}\beta(y) \right) &= 1. \end{aligned} \tag{4.47}$$

Let us now show that the Gaussian measure β given in the statement of the theorem is well-defined and verifies all the optimality conditions (4.47). Indeed, assume that β is a Gaussian measure given by $\mathcal{N}(\mathbf{B})$ for some unknown $\mathbf{B} \in S_+^d$ (remember that β is necessarily centered, following the developments (4.43)). The Sinkhorn equations can then be written as a system on positive definite matrices:

$$\begin{aligned} \mathbf{F}_k &= \sigma^2 \mathbf{A}_k^{-1} + \mathbf{G}_k^{-1} \\ \mathbf{G}_k &= \sigma^2 \mathbf{B} + \mathbf{F}_k^{-1} \\ \mathbf{H} &= \sigma^2 \mathbf{B} + \mathbf{H}^{-1}, \end{aligned}$$

where for all k

$$\begin{aligned} \frac{f_k}{2\sigma^2} &= \mathcal{Q}(\frac{1}{\sigma^2}(\mathbf{G}_k^{-1} - \mathbf{I}_\mathrm{d})) + f_k(0) \\ \frac{g_k}{2\sigma^2} &= \mathcal{Q}(\frac{1}{\sigma^2}(\mathbf{F}_k^{-1} - \mathbf{I}_\mathrm{d})) + g_k(0) \\ \frac{h_\beta}{2\sigma^2} &= \mathcal{Q}(\frac{1}{\sigma^2}(\mathbf{H}^{-1} - \mathbf{I}_\mathrm{d})) + h_\beta(0). \end{aligned} \tag{4.48}$$

Moreover, provided $\mathbf{B}$ exists and is positive definite, the system (4.48) has a unique set of solutions $(\mathbf{F}_k)_k, (\mathbf{G}_k)_k, \mathbf{H}$ given by

$$\mathbf{F}_k = \mathbf{B}\mathbf{C}_k^{-1}, \quad \mathbf{G}_k = \mathbf{C}_k^{-1}\mathbf{A}_k, \quad \text{and} \quad \mathbf{H} = \mathbf{B}^{-1}\mathbf{J}, \tag{4.49}$$

where $\mathbf{C}_k = (\mathbf{A}_k\mathbf{B} + \frac{\sigma^4}{4}\mathbf{I}_\mathrm{d})^{\frac{1}{2}} - \frac{\sigma^2}{2}\mathbf{I}_\mathrm{d}$ and $\mathbf{J} = (\mathbf{B}^2 + \frac{\sigma^4}{4}\mathbf{I}_\mathrm{d})^{\frac{1}{2}} + \frac{\sigma^2}{2}\mathbf{I}_\mathrm{d}$. Therefore, the first variation in the LHS of (4.46) can be written as

$$\begin{aligned}
\sum_{k=1}^K w_k \frac{\delta S_\sigma(\alpha_k, \beta)}{\delta \beta} &= \sum_{k=1}^K w_k (g_k - h_\beta) \\
&= \mathcal{Q}(\frac{1}{\sigma^2}(\sum_{k=1}^K w_k \mathbf{F}_k^{-1} - \mathbf{H}^{-1})) + \sum_{w=1}^K w_k (g_k(0) - h_\beta(0)) \\
&\propto \mathcal{Q}(\sum_{k=1}^K w_k \mathbf{C}_k \mathbf{B}^{-1} - \mathbf{J}^{-1}\mathbf{B}) + \sum_{w=1}^K w_k (g_k(0) - h_\beta(0))
\end{aligned} \tag{4.50}$$

and

$$\begin{aligned}
&\sum_{k=1}^K w_k \mathbf{C}_k \mathbf{B}^{-1} - \mathbf{J}^{-1}\mathbf{B} \\
&= \sum_{k=1}^K w_k \mathbf{B}^{-\frac{1}{2}} (\mathbf{B}^{\frac{1}{2}}\mathbf{A}_k\mathbf{B}^{\frac{1}{2}} + \frac{\sigma^4}{4}\mathbf{I}_\mathrm{d})^{\frac{1}{2}}\mathbf{B}^{-\frac{1}{2}} - \mathbf{B}^{-1}(\mathbf{B}^2 + \frac{\sigma^4}{4}\mathbf{I}_\mathrm{d})^{\frac{1}{2}} \\
&= \sum_{k=1}^K w_k \mathbf{B}^{-\frac{1}{2}} (\mathbf{B}^{\frac{1}{2}}\mathbf{A}_k\mathbf{B}^{\frac{1}{2}} + \frac{\sigma^4}{4}\mathbf{I}_\mathrm{d})^{\frac{1}{2}}\mathbf{B}^{-\frac{1}{2}} - \mathbf{B}^{-\frac{1}{2}}(\mathbf{B}^2 + \frac{\sigma^4}{4}\mathbf{I}_\mathrm{d})^{\frac{1}{2}}\mathbf{B}^{-\frac{1}{2}} \\
&= \mathbf{B}^{-\frac{1}{2}} \left(\sum_{k=1}^K w_k (\mathbf{B}^{\frac{1}{2}}\mathbf{A}_k\mathbf{B}^{\frac{1}{2}} + \frac{\sigma^4}{4}\mathbf{I}_\mathrm{d})^{\frac{1}{2}} - (\mathbf{B}^2 + \frac{\sigma^4}{4}\mathbf{I}_\mathrm{d})^{\frac{1}{2}} \right) \mathbf{B}^{-\frac{1}{2}},
\end{aligned} \tag{4.51}$$

which is null if and only if $\mathbf{B}$ is a solution of the equation

$$\sum_{k=1}^K w_k (\mathbf{B}^{\frac{1}{2}}\mathbf{A}_k\mathbf{B}^{\frac{1}{2}} + \frac{\sigma^4}{4}\mathbf{I}_\mathrm{d})^{\frac{1}{2}} = (\mathbf{B}^2 + \frac{\sigma^4}{4}\mathbf{I}_\mathrm{d})^{\frac{1}{2}}. \tag{4.52}$$

Therefore, provided (4.52) holds, for any probability measure $\mu \in \mathcal{G}$:

$$\begin{aligned}
\langle \sum_{k=1}^K w_k \frac{\delta S_\sigma(\alpha_k, \beta)}{\delta \beta}, \mu - \beta \rangle &= \langle \sum_{k=1}^K w_k g_k - h_\beta, \mu - \beta \rangle \\
&= \langle \sum_{w=1}^K w_k g_k(0) - h_\beta(0), \mu - \beta \rangle \\
&= \left(\sum_{w=1}^K w_k g_k(0) - h_\beta(0) \right) \int (\mathrm{d}\mu - \mathrm{d}\beta) \\
&= 0,
\end{aligned} \tag{4.53}$$

since both measures integrate to 1. Therefore, the optimality condition holds.

To end the proof, there remains to show that (4.52) admits a positive definite solution. To show the existence of a solution, the same proof as in [Agueh and Carlier, 2011] applies. Indeed, let $\lambda_d(\mathbf{A}_k)$ and $\lambda_1(\mathbf{A}_k)$ denote respectively the smallest and largest eigenvalue of

$\mathbf{A}_k$. Let $\lambda = \min_k \lambda_d(\mathbf{A}_k)$ and $\Lambda = \max_k \lambda_1(\mathbf{A}_k)$. Let $K_{\lambda,\Lambda}$ be the convex compact subset of positive definite matrices $\mathbf{B}$ such that $\Lambda \mathbf{I}_\mathrm{d} \succeq \mathbf{B} \succeq \lambda \mathbf{I}_\mathrm{d}$. Define the map:

$$\begin{aligned} T : & K_{\lambda,\Lambda} \to \mathcal{S}^d_{++} \\ & \mathbf{B} \mapsto \left(\left(\sum_{k=1}^K w_k (\mathbf{B}^{\frac{1}{2}} \mathbf{A}_k \mathbf{B}^{\frac{1}{2}} + \frac{\sigma^4}{4} \mathbf{I}_\mathrm{d})^{\frac{1}{2}} \right)^2 - \frac{\sigma^4}{4} \mathbf{I}_\mathrm{d} \right)^{1/2}. \end{aligned}$$

Now for any $\mathbf{B} \in K_{\lambda,\Lambda}$, it holds that

$$\lambda \mathbf{I}_\mathrm{d} \preceq T(\mathbf{B}) \preceq \Lambda \mathbf{I}_\mathrm{d}. \tag{4.54}$$

T is therefore a continuous map that maps $K_{\lambda,\Lambda}$ to itself, thus Brouwer's fixed-point theorem guarantees the existence of a solution. □

4 Entropy Regularized OT between Unbalanced Gaussian Measures

We proceed by considering a more general setting, in which measures $\alpha, \beta \in \mathcal{M}_2^+(\mathbb{R}^d)$ have finite integration masses $m_\alpha = \alpha(\mathbb{R}^d)$ and $m_\beta = \beta(\mathbb{R}^d)$ that are not necessarily the same. Following [Chizat et al., 2018b], we define entropy-regularized unbalanced OT as

$$\begin{aligned} \mathrm{UOT}_\sigma(\alpha, \beta) \stackrel{\text{def}}{=} \inf_{\pi \in \mathcal{M}_2^+} \Big\{ & \iint_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 \mathrm{d}\pi(x, y) \\ & + 2\sigma^2 \,\mathrm{KL}(\pi \| \alpha \otimes \beta) + \gamma \,\mathrm{KL}(\pi_1 \| \alpha) + \gamma \,\mathrm{KL}(\pi_2 \| \beta) \Big\}, \end{aligned} \tag{4.55}$$

where $\gamma > 0$ and π_1, π_2 are the marginal distributions of the coupling $\pi \in \mathcal{M}_2^+(\mathbb{R}^2 \times \mathbb{R}^d)$.

Duality and optimality conditions. The KL divergence in (4.55) is finite if and only if π admits a density with respect to $\alpha \otimes \beta$. Moreover, the objective is finite if and only if $\frac{\mathrm{d}\pi}{\mathrm{d}\alpha \mathrm{d}\beta} \in \mathcal{L}_2(\alpha \otimes \beta)$. Therefore (4.55) can be formulated as a variational problem:

$$\begin{aligned} \mathrm{UOT}_\sigma(\alpha, \beta) = \inf_{r \in \mathcal{L}_2(\alpha \otimes \beta)} \Big\{ & \iint_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 r(x, y) \mathrm{d}\alpha(x) \mathrm{d}\beta(y) \\ & + 2\sigma^2 \,\mathrm{KL}(r \| \alpha \otimes \beta) + \gamma \,\mathrm{KL}(r_1 \| \alpha) + \gamma \,\mathrm{KL}(r_2 \| \beta) \Big\}, \end{aligned} \tag{4.56}$$

where $r_1 \stackrel{\text{def}}{=} \int_{\mathbb{R}^d} r(., y) \mathrm{d}\beta(y)$ and $r_2 \stackrel{\text{def}}{=} \int_{\mathbb{R}^d} r(x, .) \mathrm{d}\alpha(x)$ correspond to the marginal density functions and the Kullback-Leibler divergence is defined for $f \in \mathcal{L}^2(\mu)$ as $\mathrm{KL}(f \| \mu) = \int_{\mathbb{R}^d} (f \log(f) + f - 1) \mathrm{d}\mu$. As in Chizat et al. [2018b], Fenchel-Rockafellar duality holds and (4.56) admits the following dual problem:

$$\begin{aligned} \mathrm{UOT}_\sigma(\alpha, \beta) = \sup_{\substack{f \in \mathcal{L}_2(\alpha) \\ g \in \mathcal{L}_2(\beta)}} \Big\{ & \gamma \int_{\mathbb{R}^d} (1 - e^{-\frac{f}{\gamma}}) \mathrm{d}\alpha + \gamma \int_{\mathbb{R}^d} (1 - e^{-\frac{g}{\gamma}}) \mathrm{d}\beta \\ & - 2\sigma^2 \iint_{\mathbb{R}^d \times \mathbb{R}^d} (e^{\frac{-\|x-y\|^2 + f(x) + g(y)}{2\sigma^2}} - 1) \mathrm{d}\alpha(x) \mathrm{d}\beta(y) \Big\}, \end{aligned} \tag{4.57}$$

for which the necessary optimality conditions read, with $\tau \stackrel{\text{def}}{=} \frac{\gamma}{\gamma + 2\sigma^2}$, as

$$\frac{f(x)}{2\sigma^2} \stackrel{a.s}{=} -\tau \log \int_{\mathbb{R}^d} e^{\frac{g(y) - \|x-y\|^2}{2\sigma^2}} \mathrm{d}\beta(y) \;\text{ and }\; \frac{g(x)}{2\sigma^2} \stackrel{a.s}{=} -\tau \log \int_{\mathbb{R}^d} e^{\frac{f(y) - \|x-y\|^2}{2\sigma^2}} \mathrm{d}\alpha(y). \tag{4.58}$$

Moreover, if such a pair of dual potentials exists, then the optimal transportation plan is given by

$$\frac{\mathrm{d}\pi}{\mathrm{d}\alpha \otimes \mathrm{d}\beta}(x,y) = e^{\frac{f(x)+g(y)-\|x-y\|^2}{2\sigma^2}}. \tag{4.59}$$

The following proposition provides a simple formula to compute UOT_σ at optimality. It shows that it is sufficient to know the total transported mass $\pi(\mathbb{R}^d \times \mathbb{R}^d)$.

Proposition 4.13. *Assume there exists an optimal transportation plan π^*, solution of (4.55). Then*

$$\mathrm{UOT}_\sigma(\alpha,\beta) = \gamma(m_\alpha + m_\beta) + 2\sigma^2 m_\alpha m_\beta - 2(\sigma^2+\gamma)\pi^*(\mathbb{R}^d \times \mathbb{R}^d). \tag{4.60}$$

Proof. Using Fubini-Tonelli along with the optimality conditions (4.58), we can write the total mass of the optimal transportation plan as

$$\begin{aligned}
\pi(\mathbb{R}^d \times \mathbb{R}^d) &= \iint_{\mathbb{R}^d \times \mathbb{R}^d} e^{\frac{-\|x-y\|^2+f(x)+g(y)}{2\sigma^2}} \mathrm{d}\alpha(x)\mathrm{d}\beta(y) \\
&= \int_{\mathbb{R}^d} \left(\int_{\mathbb{R}^d} e^{\frac{-\|x-y\|^2+f(x)}{2\sigma^2}} \mathrm{d}\alpha(x) \right) e^{\frac{g(y)}{2\sigma^2}} \mathrm{d}\beta(y) \\
&= \int_{\mathbb{R}^d} e^{\frac{g(y)}{2\sigma^2}(1-\frac{1}{\tau})} \mathrm{d}\beta(y) \\
&= \int_{\mathbb{R}^d} e^{-\frac{g(y)}{\gamma}} \mathrm{d}\beta(y)
\end{aligned}$$

And similarly: $\pi(\mathbb{R}^d \times \mathbb{R}^d) = \int_{\mathbb{R}^d} e^{-\frac{f(x)}{\gamma}} \mathrm{d}\alpha(x)$. Plugging this in the dual objective (4.57), we get

$$\begin{aligned}
&\gamma \int_{\mathbb{R}^d} (1 - e^{-\frac{f}{\gamma}}) \mathrm{d}\alpha + \gamma \int_{\mathbb{R}^d} (1 - e^{-\frac{g}{\gamma}}) \mathrm{d}\beta - 2\sigma^2 \iint_{\mathbb{R}^d \times \mathbb{R}^d} (e^{\frac{-\|x-y\|^2+f(x)+g(y)}{2\sigma^2}} - 1) \mathrm{d}\alpha(x)\mathrm{d}\beta(y) \\
&\quad = \gamma(m_\alpha - m_\pi) + \gamma(m_\beta - m_\pi) - 2\sigma^2(m_\pi - m_\alpha m_\beta),
\end{aligned}$$

which yields the desired expression. □

Unbalanced OT for scaled Gaussians. Let α and β be unnormalized Gaussian measures. Formally, $\alpha = m_\alpha \mathcal{N}(\mathbf{a}, \mathbf{A})$ and $\beta = m_\beta \mathcal{N}(\mathbf{b}, \mathbf{B})$ with $m_\alpha, m_\beta > 0$. Unlike for balanced OT, α and β cannot be assumed to be centered without loss of generality. However, we can still derive a closed form formula for $\mathrm{UOT}_\sigma(\alpha,\beta)$ by considering quadratic potentials of the form

$$\frac{f(\mathbf{x})}{2\sigma^2} = -\frac{1}{2}(x^\top \mathbf{U}\mathbf{x} - 2x^\top \mathbf{u}) + \log(m_u) \;\text{ and }\; \frac{g(x)}{2\sigma^2} = -\frac{1}{2}(x^\top \mathbf{V}\mathbf{x} - 2x^\top \mathbf{v}) + \log(m_v). \tag{4.61}$$

Let σ and γ be the regularization parameters as in (4.56), and $\tau \overset{\text{def}}{=} \frac{\gamma}{2\sigma^2+\gamma}$, $\lambda \overset{\text{def}}{=} \frac{\sigma^2}{1-\tau} = \sigma^2 + \frac{\gamma}{2}$. Let us define the following useful quantities:

$$\mu = \begin{pmatrix} \mathbf{a} + \mathbf{A}\mathbf{X}^{-1}(\mathbf{b}-\mathbf{a}) \\ \mathbf{b} + \mathbf{B}\mathbf{X}^{-1}(\mathbf{a}-\mathbf{b}) \end{pmatrix} \tag{4.62}$$

$$\mathbf{H} = \begin{pmatrix} (\mathbf{I}_\mathrm{d} + \frac{1}{\lambda}\mathbf{C})(\mathbf{A} - \mathbf{A}\mathbf{X}^{-1}\mathbf{A}) & \mathbf{C} + (\mathbf{I}_\mathrm{d} + \frac{1}{\lambda}\mathbf{C})\mathbf{A}\mathbf{X}^{-1}\mathbf{B} \\ \mathbf{C}^\top + (\mathbf{I}_\mathrm{d} + \frac{1}{\lambda}\mathbf{C}^\top)\mathbf{B}\mathbf{X}^{-1}\mathbf{A} & (\mathbf{I}_\mathrm{d} + \frac{1}{\lambda}\mathbf{C}^\top)(\mathbf{B} - \mathbf{B}\mathbf{X}^{-1}\mathbf{B}) \end{pmatrix} \tag{4.63}$$

$$m_\pi = \sigma^{\frac{d\sigma^2}{\gamma+\sigma^2}} \left(m_\alpha m_\beta \det(\mathbf{C}) \sqrt{\frac{\det(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}})^\tau}{\det(\mathbf{A}\mathbf{B})}} \right)^{\frac{1}{\tau+1}} \frac{e^{-\frac{\|\mathbf{a}-\mathbf{b}\|^2_{\mathbf{X}^{-1}}}{2(\tau+1)}}}{\sqrt{\det(\mathbf{C} - \frac{2}{\gamma}\widetilde{\mathbf{A}}\widetilde{\mathbf{B}})}}, \tag{4.64}$$

with

$$\mathbf{X} = \mathbf{A} + \mathbf{B} + \lambda \mathbf{I}_{\mathrm{d}}, \qquad \widetilde{\mathbf{A}} = \frac{\gamma}{2}(\mathbf{I}_{\mathrm{d}} - \lambda(\mathbf{A} + \lambda \mathbf{I}_{\mathrm{d}})^{-1}),$$

$$\widetilde{\mathbf{B}} = \frac{\gamma}{2}(\mathbf{I}_{\mathrm{d}} - \lambda(\mathbf{B} + \lambda \mathbf{I}_{\mathrm{d}})^{-1}), \qquad \mathbf{C} = \left(\frac{1}{\tau}\widetilde{\mathbf{A}}\widetilde{\mathbf{B}} + \frac{\sigma^4}{4}\mathbf{I}_{\mathrm{d}}\right)^{\frac{1}{2}} - \frac{\sigma^2}{2}\mathbf{I}_{\mathrm{d}}.$$

Theorem 4.14. *Let $\alpha = m_\alpha \mathcal{N}(\mathbf{a}, \mathbf{A})$ and $\beta = m_\beta \mathcal{N}(\mathbf{b}, \mathbf{B})$ be two unnormalized Gaussian measures. Let $\tau = \frac{\gamma}{2\sigma^2+\gamma}$ and $\lambda = \frac{\sigma^2}{1-\tau} = \sigma^2 + \frac{\gamma}{2}$ and μ, $\mathbf{H}$, and m_π be as above. Then*

(i) The unbalanced optimal transport plan, minimizer of (4.55), is also an unnormalized Gaussian over $\mathbb{R}^d \times \mathbb{R}^d$ given by $\pi = m_\pi \mathcal{N}(\mu, \mathbf{H})$,

(ii) UOT_σ can be obtained in closed form using Proposition 4.13 with $\pi(\mathbb{R}^d \times \mathbb{R}^d) = m_\pi$.

As in the balanced setting, the proof of Theorem 4.14 relies on proving the contractivity and stability of generalized Sinkhorn iterations [Chizat, 2017, Peyré et al., 2019] w.r.t. quadratic potentials. As it is quite long and at times technical, we defer it to Section 6.2.

Remark 4.15. *The exponential term in the closed-form formula above provides some intuition on how transportation occurs in unbalanced OT. When the difference between the means is too large, the transported mass $m_\pi^\star$ goes to 0 and thus no transport occurs. However for fixed means $\mathbf{a}, \mathbf{b}$, when $\gamma \to +\infty$, we have $\mathbf{X}^{-1} \to 0$ and the exponential term approaches 1.*

5 Numerical Experiments

5.1 Empirical validation of the closed-form formulas

Figure 4.1 illustrates the convergence towards the closed form formulas of both theorems. For each dimension d in [5, 10], we select a pair of Gaussian measuress $\alpha \sim \mathcal{N}(\mathbf{a}, \mathbf{A})$ and $\beta \sim m_\beta \mathcal{N}(\mathbf{b}, \mathbf{B})$ where m_β equals 1 (resp. 2) in the balanced (resp. unbalanced) setting, and randomly generated means $\mathbf{a}, \mathbf{b}$ uniformly in $[-1, 1]^d$ and covariance matrices $\mathbf{A}, \mathbf{B} \in S_{++}^d$ following the Wishart distribution $W_d(0.2 * \mathbf{I}_{\mathrm{d}}, d)$. We generate empirical distributions α_n and β_n with n i.i.d. samples from $\mathcal{N}(\mathbf{a}, \mathbf{A})$ and $\mathcal{N}(\mathbf{b}, \mathbf{B})$ respectively (with total masses 1 and m_β) and compute OT_σ / UOT_σ. We report means and ± shaded standard-deviation areas over 20 independent trials for each value of n.

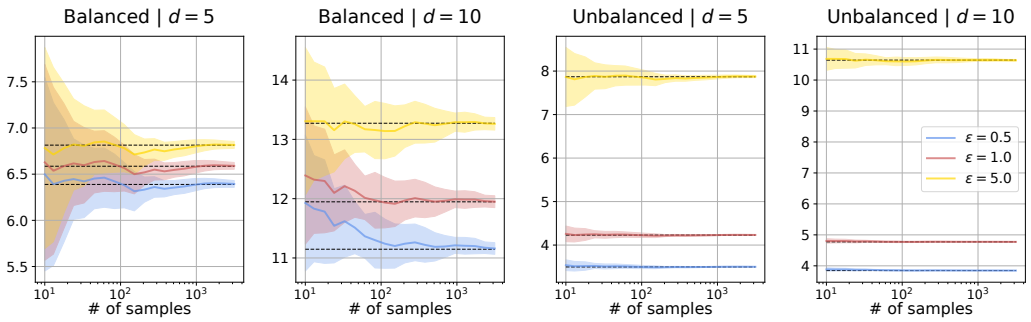


Figure 4.1: Numerical convergence of discrete OT between empirical distributions, $\mathrm{OT}_\sigma(\alpha_n, \beta_n)$ and $\mathrm{UOT}_\sigma(\alpha_n, \beta_n)$, towards the closed form of $\mathrm{OT}_\sigma(\alpha, \beta)$ and $\mathrm{UOT}_\sigma(\alpha, \beta)$ (dashed) given by Theorem 4.2 and Theorem 4.14 for random Gaussians α, β. For unbalanced OT, $\gamma = 1$.

5.2 Transport plan visualization with $d = 1$

Figure 4.2 confronts the expected theoretical plans (contours in black) given by Theorems 4.2 and 4.14 to empirical ones (weights in shades of red) obtained with Sinkhorn's algorithm using 2000 Gaussian samples. The density functions (black) and the empirical histograms (red) of α (resp. β) with 200 bins are displayed on the left (resp. top) of each transport plan. The red weights are computed via a 2d histogram of the transport plan returned by Sinkhorn's algorithm with (200 x 200) bins. Notice the blurring effect of ε and increased mass transportation of the Gaussian tails in unbalanced transport with larger γ.

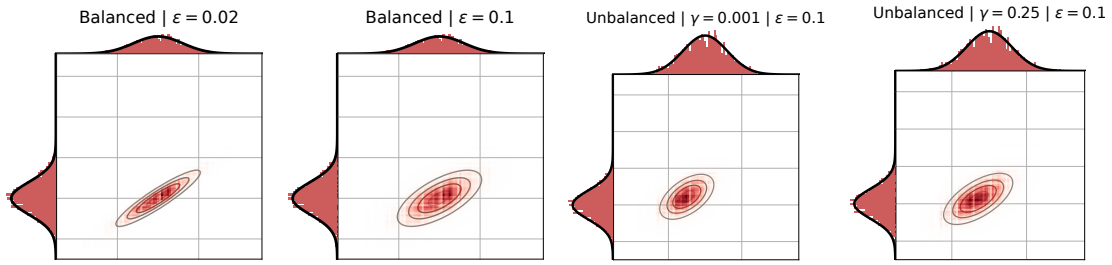


Figure 4.2: Effect of ε in balanced OT and γ in unbalanced OT. Empirical plans (red) correspond to the expected Gaussian contours depicted in black. Here $\alpha = \mathcal{N}(0, 0.04)$ and $\beta = m_\beta \mathcal{N}(0.5, 0.09)$ with $m_\beta = 1$ (balanced) and $m_\beta = 2$ (unbalanced). In unbalanced OT, the right tail of β is not transported, and the mean of the transportation plan is shifted compared to that of the balanced case – as expected from Theorem 4.14 specially for low γ.

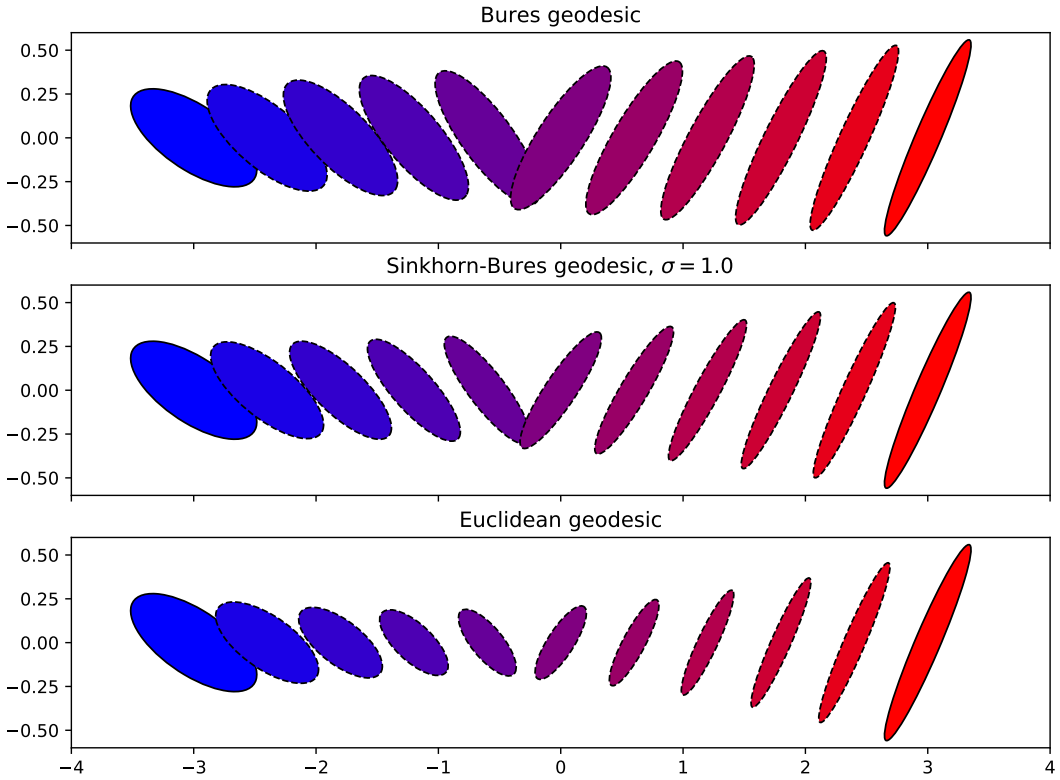


Figure 4.3: Bures, Sinkhorn-Bures, and Euclidean geodesics. Sinkhorn-Bures trajectories converge to Bures geodesics as σ goes to 0, and to Euclidean geodesics as σ goes to infinity.

5.3 Effects of regularization strength

We provide numerical experiments in Figures 4.4 and 4.5 to illustrate the behaviour of transportation plans and corresponding distances as σ goes to 0 or to infinity. As can be seen from (4.14), when $\sigma \to 0$ we recover the Wasserstein-Bures distance (4.3), and the optimal transportation plan converges to the Monge map (4.5). When on the contrary $\sigma \to \infty$, Sinkhorn divergences $\mathfrak{S}_\varepsilon(\alpha, \beta) \stackrel{\text{def}}{=} \mathrm{OT}_\varepsilon(\alpha, \beta) - \frac{1}{2}(\mathrm{OT}_\varepsilon(\alpha, \alpha) + \mathrm{OT}_\varepsilon(\beta, \beta))$ convergence to MMD with a $-c$ kernel (where c is the optimal transport ground cost) [Genevay et al., 2018]. With a $-\ell_2$ kernel, MMD is degenerate and equals 0 for centered measures.

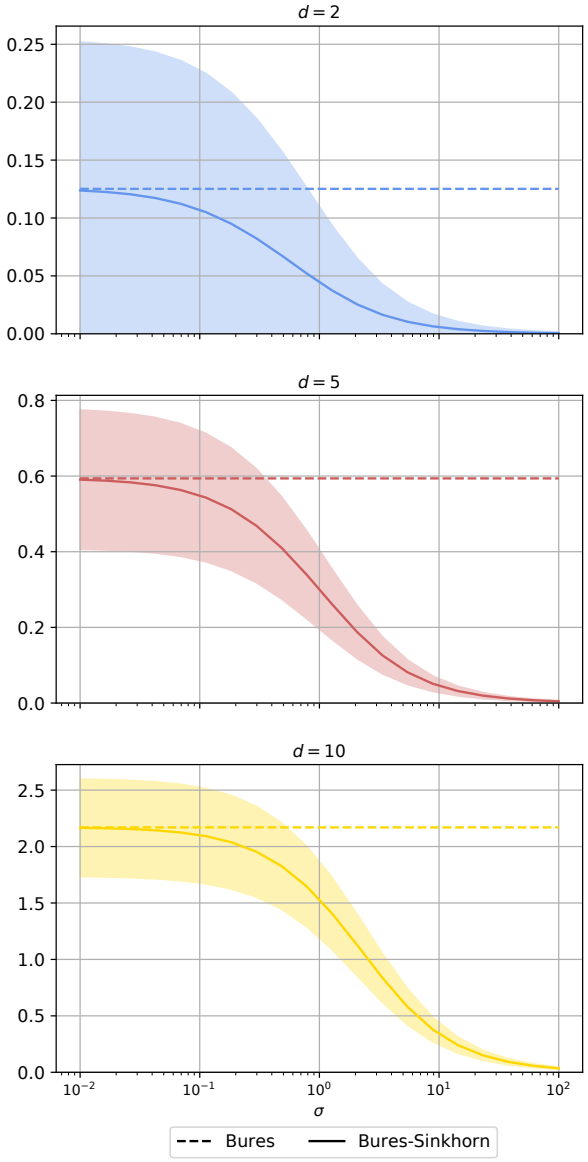


Figure 4.4: Numerical convergence of $\mathfrak{B}_{\sigma^2}(\mathbf{A}, \mathbf{B}) - \frac{1}{2}(\mathfrak{B}_{\sigma^2}(\mathbf{A}, \mathbf{A}) + \mathfrak{B}_{\sigma^2}(\mathbf{B}, \mathbf{B}))$ to $\mathfrak{B}(\mathbf{A}, \mathbf{B})$ as σ goes to 0 and to 0 as σ goes to infinity.

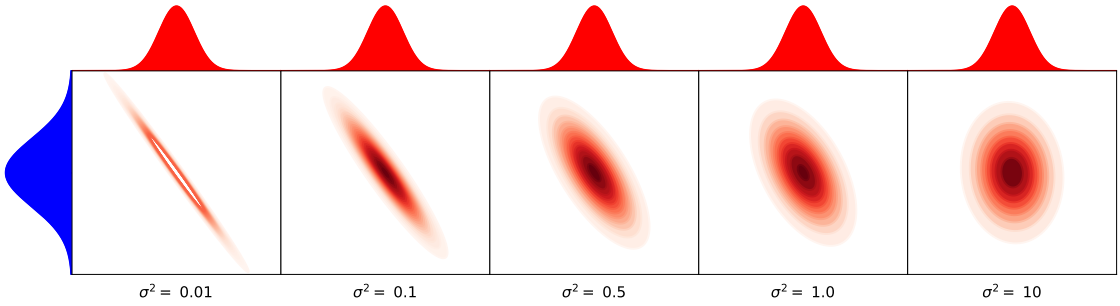


Figure 4.5: Effect of regularization on transportation plans. When σ goes to 0 (left), the transportation plan concentrates on the graph of the linear Monge map. When σ goes to infinity (right), the transportation plan converges to the independent coupling.

5.4 The Newton-Schulz algorithm

Algorithm 4 NS Monge Iterations

Input: PSD matrix $\mathbf{A}, \mathbf{B}$, $\epsilon > 0$

$\mathbf{Y} \leftarrow \frac{\mathbf{B}}{(1+\epsilon)\|\mathbf{B}\|}$, $\mathbf{Z} \leftarrow \frac{\mathbf{A}}{(1+\epsilon)\|\mathbf{A}\|}$

while not converged **do**

$\mathbf{T} \leftarrow (3\mathbf{I} - \mathbf{ZY})/2$

$\mathbf{Y} \leftarrow \mathbf{YT}$

$\mathbf{Z} \leftarrow \mathbf{TZ}$

end while

$\mathbf{Y} \leftarrow \sqrt{\frac{\|\mathbf{B}\|}{\|\mathbf{A}\|}}\mathbf{Y}$, $\mathbf{Z} \leftarrow \sqrt{\frac{\|\mathbf{A}\|}{\|\mathbf{B}\|}}\mathbf{Z}$

Output: $\mathbf{Y} = \mathbf{T}^{\mathbf{AB}}$, $\mathbf{Z} = \mathbf{T}^{\mathbf{BA}}$

The main bottleneck in computing $\mathbf{T}^{\mathbf{AB}}$ is that of computing matrix square roots. This can be performed using singular value decomposition (SVD) or, as suggested in [Muzellec and Cuturi, 2018], using Newton-Schulz (NS) iterations [Higham, 2008, §5.3]. In particular, Newton-Schulz iterations have the advantage of yielding both roots, and inverse roots. Hence, to compute $\mathbf{T}^{\mathbf{AB}}$, one would run NS a first time to obtain $\mathbf{A}^{1/2}$ and $\mathbf{A}^{-1/2}$, and a second time to get $(\mathbf{A}^{1/2}\mathbf{B}\mathbf{A}^{1/2})^{1/2}$ (c.f. Chapter 2).

In fact, as a direct application of [Higham, 2008, Theorem 5.2], one can even compute both $\mathbf{T}^{\mathbf{AB}}$ and $\mathbf{T}^{\mathbf{BA}} = \left(\mathbf{T}^{\mathbf{AB}}\right)^{-1}$ in a single run by initializing the Newton-Schulz algorithm with $\mathbf{A}$ and $\mathbf{B}$, as in Algorithm 4. Using (4.6), and noting that $\mathfrak{B}(\mathbf{A}, \mathbf{B}) = \mathrm{Tr}\mathbf{A} + \mathrm{Tr}\mathbf{B} - 2\mathrm{Tr}(\mathbf{T}^{\mathbf{AB}}\mathbf{A})$, this implies that a single run of NS is sufficient to compute $\mathfrak{B}(\mathbf{A}, \mathbf{B})$, $\nabla_{\mathbf{A}}\mathfrak{B}(\mathbf{A}, \mathbf{B})$ and $\nabla_{\mathbf{B}}\mathfrak{B}(\mathbf{A}, \mathbf{B})$ using basic matrix operations. The main advantage of Newton-Schultz over SVD is that it its efficient scalability on GPUs, as illustrated in Figure 4.6.

Newton-Schulz iterations are quadratically convergent under the condition

$$\|\mathbf{I}_{\mathrm{d}} - \left(\begin{smallmatrix} \mathbf{A} & 0 \\ 0 & \mathbf{B} \end{smallmatrix}\right)^2\| < 1,$$

as shown in [Higham, 2008, Theorem 5.8]. To meet this condition, it is sufficient to rescale $\mathbf{A}$ and $\mathbf{B}$ so that their norms equal $(1+\varepsilon)^{-1}$ for some $\varepsilon > 0$, as in the first step of Algorithm 4 (which can be skipped if $\|\mathbf{A}\| < 1$ (resp. $\|\mathbf{B}\| < 1$)). Finally, the output of the iterations are scaled back, using the homogeneity (resp. inverse homogoneity) of eq. (4.5) w.r.t. $\mathbf{A}$ (resp. $\mathbf{B}$).

A rough theoretical analysis shows that both Newton-Schulz and SVD have a $O(d^3)$ complexity in the dimension. Figure 4.6 compares the running times of Newton-Schulz

iterations and SVD on CPU or GPU used to compute both $\mathbf{A}^{\frac{1}{2}}$ and $\mathbf{A}^{-\frac{1}{2}}$. We simulate a batch of positive definite matrices $\mathbf{A}$ following the Wishart distribution $W(\mathbf{I}_{\mathrm{d}d}, d)$ to which we add $0.1\mathbf{I}_{\mathrm{d}}$ to avoid numerical issues when computing inverse square roots. We display the average run-time of 50 different trials along with its $\pm$ std interval. Notice the different magnitudes between CPUs and GPUs. As a termination criterion, we first run EVD to obtain $\mathbf{A}_{evd}^{1/2}$ and $\mathbf{A}_{evd}^{-1/2}$ and stop the Newton-Schultz algorithm when its n-th running estimate $\mathbf{A}_n^{1/2}$ verifies: $\|\mathbf{A}_n^{1/2} - \mathbf{A}_{evd}^{1/2}\|_1 \leq 10^{-4}$. Notice the different order of magnitude between CPUs and GPUs. Moreover, the computational advantage of Newton-Schultz on GPUs can be further increased when computing multiple square roots in parallel.

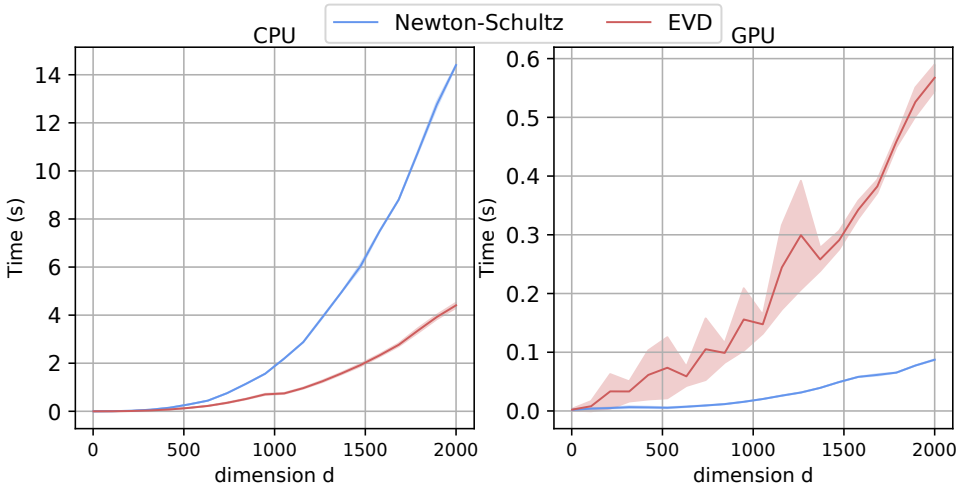


Figure 4.6: Average run-time of Newton-Schulz and EVD to compute on CPUs and GPUs.

Conclusion

In this chapter we have provided – to the best of our knowledge – the first nontrivial closed form expressions of entropy-regularized optimal transport for both balanced and unbalanced measures. While our contributions are mostly theoretical and would certainly promote new theoretical findings in entropic OT, the entropy-regularized Bures-Wasserstein distance obtained here is better suited for real data applications where covariance matrices are prone to be ill-conditioned.

6 Appendix: Technical Lemmas and Proof of Theorem 4.14

6.1 Proofs of technical lemmas

We provide in this appendix the statement of the lemmas used in this chapter along with their proofs.

Lemma 4.16. *[Sum of factorized quadratic forms] Let $\mathbf{A}, \mathbf{B} \in S_d$ such that $\mathbf{A} \neq \mathbf{B}$ and $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$. Denote $\alpha = (\mathbf{A}, \mathbf{a})$ and $\beta = (\mathbf{B}, \mathbf{b})$. Let $P_\alpha(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\mathbf{a})^\top \mathbf{A}(\mathbf{x}-\mathbf{a})$ and $P_\beta(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\mathbf{b})^\top \mathbf{B}(\mathbf{x}-\mathbf{b})$. Then*

$$P_\alpha(x) + P_\beta(x) = -\frac{1}{2}\left((\mathbf{x}-\mathbf{c})^\top \mathbf{C}(\mathbf{x}-\mathbf{c}) + q_{\alpha,\beta}\right), \tag{4.65}$$

where

$$\begin{cases} \mathbf{C} & = \mathbf{A}+\mathbf{B} \\ (\mathbf{A}+\mathbf{B})\mathbf{c} & = (\mathbf{A}\mathbf{a}+\mathbf{B}\mathbf{b}) \\ q_{\alpha,\beta} & = \mathbf{a}^\top\mathbf{A}\mathbf{a} + \mathbf{b}^\top\mathbf{B}\mathbf{b} - c^\top\mathbf{C}\mathbf{c}. \end{cases} \tag{4.66}$$

In particular, if $\mathbf{C} = \mathbf{A}+\mathbf{B}$ is invertible, then

$$\begin{cases} \mathbf{c} = \mathbf{C}^{-1}(\mathbf{A}\mathbf{a}+\mathbf{B}\mathbf{b}) \\ \mathbf{c}^\top\mathbf{C}\mathbf{c} = (\mathbf{A}\mathbf{a}+\mathbf{B}\mathbf{b})^\top\mathbf{C}^{-1}(\mathbf{A}\mathbf{a}+\mathbf{B}\mathbf{b}). \end{cases} \tag{4.67}$$

Proof. On the one hand, we have

$$\begin{aligned} P_\alpha(x) + P_\beta(x) &= -\frac{1}{2}\left((\mathbf{x}-\mathbf{a})^\top\mathbf{A}(\mathbf{x}-\mathbf{a}) + (\mathbf{x}-\mathbf{b})^\top\mathbf{B}(\mathbf{x}-\mathbf{b})\right) \\ &= -\frac{1}{2}\left(\mathbf{x}^\top(\mathbf{A}+\mathbf{B})\mathbf{x} - 2\mathbf{x}^\top(\mathbf{A}\mathbf{a}+\mathbf{B}\mathbf{b}) + \mathbf{a}^\top\mathbf{A}\mathbf{a} + \mathbf{b}^\top\mathbf{B}\mathbf{b}\right). \end{aligned}$$

On the other hand, for an arbitrary $\gamma = (\mathbf{c}, \mathbf{C})$ and $q \in \mathbb{R}$, we have

$$\begin{aligned} P_\gamma(x) - \frac{q}{2} &= -\frac{1}{2}\left((\mathbf{x}-\mathbf{c})^\top\mathbf{C}(\mathbf{x}-\mathbf{C}) + q\right) \\ &= -\frac{1}{2}\left(x^\top\mathbf{C}x - 2x^\top\mathbf{C}\mathbf{c} + \mathbf{c}^\top\mathbf{C}\mathbf{c} + q\right). \end{aligned}$$

If $\mathbf{A} \neq \mathbf{B}$, identification of the parameters of both quadratic forms leads to (4.66). □

Lemma 4.17. *[Gaussian convolution of factorized quadratic forms] Let $\mathbf{A} \in S_d$, $\mathbf{a} \in \mathbb{R}^d$ and $\sigma > 0$ such that $\sigma^2\mathbf{A} + \mathbf{I}_\mathrm{d} \succ 0$. Let $Q_\alpha(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\mathbf{a})^\top\mathbf{A}(\mathbf{x}-\mathbf{a})$. Then the convolution of $e^{\mathcal{Q}_\alpha}$ by the Gaussian kernel $\mathcal{N}\left(0, \frac{\mathbf{I}_\mathrm{d}}{\sigma^2}\right)$ is given by*

$$\begin{aligned} \mathcal{N}\left(0, \frac{\mathbf{I}_\mathrm{d}}{\sigma^2}\right) \star \exp\left(\mathcal{Q}_\alpha\right) &\overset{\text{def}}{=} \int_{\mathbb{R}^d} \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left(-\frac{1}{2\sigma^2}\|. - y\|^2 + \mathcal{Q}_\alpha(y)\right) \mathrm{d}y \\ &= c_\alpha \exp(\mathcal{Q}(\mathbf{a}, \mathbf{J})), \end{aligned} \tag{4.68}$$

where

$$\begin{aligned} \mathbf{J} &= (\sigma^2\mathbf{A} + \mathbf{I}_\mathrm{d})^{-1}\mathbf{A} \\ c_\alpha &= \frac{1}{\sqrt{\det(\sigma^2\mathbf{A} + \mathbf{I}_\mathrm{d})}}. \end{aligned}$$

Proof. Using Lemma 4.16 one can write for any fixed $x \in \mathbb{R}^d$

$$\begin{aligned}-\frac{1}{2\sigma^2}\|x-y\|^2 + \mathcal{Q}_\alpha(y) &= \mathcal{Q}(x, \frac{\mathbf{I}_\mathrm{d}}{\sigma^2})(y) + \mathcal{Q}(\mathbf{a}, \mathbf{A})(y) \\ &= \mathcal{Q}(\mathbf{A}\mathbf{a} + \frac{x}{\sigma^2}, \mathbf{A} + \frac{\mathbf{I}_\mathrm{d}}{\sigma^2})(y) + h(x),\end{aligned}$$

with $h(x) = -\frac{1}{2}\left(\mathbf{a}^\top\mathbf{A}\mathbf{a} + \frac{1}{\sigma^2}\|x\|^2 - \frac{1}{\sigma^2}(\sigma^2\mathbf{A}\mathbf{a}+x)^\top(\sigma^2\mathbf{A}+\mathbf{I}_\mathrm{d})^{-1}(\sigma^2\mathbf{A}\mathbf{a}+x)\right)$. Therefore, the convolution integral is finite if and only if $\mathbf{A} + \frac{\mathbf{I}_\mathrm{d}}{\sigma^2} \succ 0$, in which case we get the integral of a Gaussian density:

$$\begin{aligned}\frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}}\int_{\mathbb{R}^d} \exp\left(\mathcal{Q}(\mathbf{A}\mathbf{a} + \frac{x}{\sigma^2}, \mathbf{A} + \frac{\mathbf{I}_\mathrm{d}}{\sigma^2})(y) + h(x)\right)\mathrm{d}(y) &= \sqrt{\frac{\det(2\pi(\mathbf{A}+\frac{\mathbf{I}_\mathrm{d}}{\sigma^2})^{-1})}{(2\pi\sigma^2)^n}}e^{h(x)} \\ &= \frac{e^{h(x)}}{\sqrt{\det(\sigma^2\mathbf{A}+\mathbf{I}_\mathrm{d})}}\end{aligned}$$

For the sake of clarity, let's separate the terms of h depending on their order in x: $h(x) = -\frac{1}{2}\left(h_2(x) + h_1(x) + h_0\right)$ where

$$\begin{aligned}h_2(x) &= \frac{1}{\sigma^2}(\|x\|^2 - x^\top(\sigma^2\mathbf{A}+\mathbf{I}_\mathrm{d})^{-1}x \\ h_1(x) &= -2x^\top(\sigma^2\mathbf{A}+\mathbf{I}_\mathrm{d})^{-1}\mathbf{A}\mathbf{a} \\ h_0 &= \mathbf{a}\mathbf{A}\mathbf{a} - \sigma^2\mathbf{a}^\top\mathbf{A}(\sigma^2\mathbf{A}+\mathbf{I}_\mathrm{d})^{-1}\mathbf{A}\mathbf{a}.\end{aligned}$$

Finally, we can factorize h_2 and h_0 using Woodbury's matrix identity which holds even for a singular matrix $\mathbf{A}$:

$$(\sigma^2\mathbf{A}+\mathbf{I}_\mathrm{d})^{-1} = \mathbf{I}_\mathrm{d} - \sigma^2(\sigma^2\mathbf{A}+\mathbf{I}_\mathrm{d})^{-1}\mathbf{A}. \tag{4.69}$$

Let $\mathbf{J} = (\sigma^2\mathbf{A}+\mathbf{I}_\mathrm{d})^{-1}\mathbf{A}$. Then,

$$\begin{aligned}h_2(x) &= \frac{1}{\sigma^2}(\|x\|^2 - x^\top(\mathbf{I}_\mathrm{d} - \sigma^2(\sigma^2\mathbf{A}+\mathbf{I}_\mathrm{d})^{-1}\mathbf{A})x \\ &= x^\top(\sigma^2\mathbf{A}+\mathbf{I}_\mathrm{d})^{-1}\mathbf{A}x \\ &= x^\top\mathbf{J}x,\end{aligned}$$

$$h_1(x) = -2x^\top\mathbf{J}\mathbf{a},$$

$$\begin{aligned}h_0 &= \mathbf{a}\mathbf{A}\mathbf{a} - \sigma^2\mathbf{a}^\top\mathbf{A}(\sigma^2\mathbf{A}+\mathbf{I}_\mathrm{d})^{-1}\mathbf{A}\mathbf{a} \\ &= \mathbf{a}^\top\mathbf{A}(\mathbf{I}_\mathrm{d} - \sigma^2(\sigma^2\mathbf{A}+\mathbf{I}_\mathrm{d})^{-1}\mathbf{A})\mathbf{a} \\ &= \mathbf{a}^\top\mathbf{A}(\sigma^2\mathbf{A}+\mathbf{I}_\mathrm{d})^{-1}\mathbf{a} \\ &= \mathbf{a}^\top(\sigma^2\mathbf{A}+\mathbf{I}_\mathrm{d})^{-1}\mathbf{A}\mathbf{a} \\ &= \mathbf{a}^\top\mathbf{J}\mathbf{a}.\end{aligned}$$

Therefore, $h(x) = -\frac{1}{2}\left(x^\top\mathbf{J}x - 2x^\top\mathbf{J}\mathbf{a} + \mathbf{a}^\top\mathbf{J}\mathbf{a}\right) = -\frac{1}{2}(x-\mathbf{a})^\top\mathbf{J}(x-\mathbf{a}) = \mathcal{Q}(\mathbf{a},\mathbf{J})(x)$. □

Lemma 4.18. *[Gaussian convolution of generic quadratic forms] Let $\mathbf{A} \in S_d$ and $\mathbf{a} \in \mathbb{R}^d$ and $\sigma > 0$ such that $\sigma^2\mathbf{A} + \mathbf{I}_\mathrm{d} \succ 0$. Let $Q_\alpha(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}^\top\mathbf{A}\mathbf{x} - 2\mathbf{x}^\top\mathbf{a})$. Then the convolution of $e^{\mathcal{Q}_\alpha}$ by the Gaussian kernel $\mathcal{N}\left(0, \frac{\mathbf{I}_\mathrm{d}}{\sigma^2}\right)$ is given by:*

$$\begin{aligned}\mathcal{N}\left(0, \frac{\mathbf{I}_\mathrm{d}}{\sigma^2}\right) \star \exp\left(\mathcal{Q}_\alpha\right) &\overset{\text{def}}{=} \int_{\mathbb{R}^d} \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}}\exp\left(-\frac{1}{2\sigma^2}\|.-y\|^2 + \mathcal{Q}_\alpha(y)\right)\mathrm{d}y \\ &= c_\alpha \exp(\mathcal{Q}(\mathbf{G}\mathbf{a}, \mathbf{G}\mathbf{A})),\end{aligned} \tag{4.70}$$

where

$$\mathbf{G} = (\sigma^2\mathbf{A} + \mathbf{I}_\mathrm{d})^{-1}$$
$$c_\alpha = \frac{e^{\frac{\sigma^2\mathbf{a}^\top\mathbf{G}\mathbf{a}}{2}}}{\sqrt{\det(\sigma^2\mathbf{A} + \mathbf{I}_\mathrm{d})}}.$$

Proof. Using Lemma 4.16 one can write for any fixed $x \in \mathbb{R}^d$, we have

$$\begin{aligned}
-\frac{1}{2\sigma^2}\|x-y\|^2 + \mathcal{Q}_\alpha(y) &= \mathcal{Q}(x, \frac{\mathbf{I}_\mathrm{d}}{\sigma^2})(y) + \mathcal{Q}(\mathbf{a}, \mathbf{A})(y) \\
&= \mathcal{Q}(\mathbf{a} + \frac{x}{\sigma^2}, \mathbf{A} + \frac{\mathbf{I}_\mathrm{d}}{\sigma^2})(y) - \frac{1}{2\sigma^2}\|x\|^2 \\
&= \mathcal{Q}f((\sigma\mathbf{a} + \frac{x}{\sigma^2}, \mathbf{A} + \frac{\mathbf{I}_\mathrm{d}}{\sigma^2})(y) + h(x),
\end{aligned}$$

with $h(x) = -\frac{1}{2}\left(\frac{1}{\sigma^2}\|x\|^2 - \frac{1}{\sigma^2}(\sigma^2\mathbf{a}+x)^\top(\sigma^2\mathbf{A}+\mathbf{I}_\mathrm{d})^{-1}(\sigma^2\mathbf{a}+x)\right)$. Therefore, the convolution integral is finite if and only if $\mathbf{A} + \frac{\mathbf{I}_\mathrm{d}}{\sigma^2} \succ 0$, in which case we get the integral of a Gaussian density:

$$\begin{aligned}
\frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}}\int_{\mathbb{R}^d} \exp\left(\mathcal{Q}f(\mathbf{a} + \frac{x}{\sigma^2}, \mathbf{A} + \frac{\mathbf{I}_\mathrm{d}}{\sigma^2})(y) + h(x)\right)\mathrm{d}(y) &= \sqrt{\frac{\det(2\pi(\mathbf{A} + \frac{\mathbf{I}_\mathrm{d}}{\sigma^2})^{-1})}{(2\pi\sigma^2)^n}}e^{h(x)} \\
&= \frac{e^{h(x)}}{\sqrt{\det(\sigma^2\mathbf{A} + \mathbf{I}_\mathrm{d})}}
\end{aligned}$$

For the sake of clarity, let's separate the terms of h depending on their order in x: $h(x) = -\frac{1}{2}\left(h_2(x) + h_1(x) + h_0\right)$ where:

$$\begin{aligned}
h_2(x) &= \frac{1}{\sigma^2}(\|x\|^2 - x^\top(\sigma^2\mathbf{A} + \mathbf{I}_\mathrm{d})^{-1}x \\
h_1(x) &= -2x^\top(\sigma^2\mathbf{A} + \mathbf{I}_\mathrm{d})^{-1}\mathbf{a} \\
h_0 &= -\sigma^2\mathbf{a}^\top(\sigma^2\mathbf{A} + \mathbf{I}_\mathrm{d})^{-1}\mathbf{a}
\end{aligned}$$

Finally, we can factorize h_2 and h_0 using Woodbury's matrix identity (4.69) which holds even for a singular matrix $\mathbf{A}$. Let $\mathbf{G} = (\sigma^2\mathbf{A} + \mathbf{I}_\mathrm{d})^{-1}$, then

$$\begin{aligned}
h_2(x) &= \frac{1}{\sigma^2}(\|x\|^2 - x^\top(\mathbf{I}_\mathrm{d} - \sigma^2(\sigma^2\mathbf{A} + \mathbf{I}_\mathrm{d})^{-1}\mathbf{A})x \\
&= x^\top(\sigma^2\mathbf{A} + \mathbf{I}_\mathrm{d})^{-1}\mathbf{A}x \\
&= x^\top\mathbf{G}\mathbf{A}x,
\end{aligned}$$

$$h_1(x) = -2x^\top\mathbf{G}\mathbf{a},$$

$$\begin{aligned}
h_0 &= -\sigma^2\mathbf{a}^\top(\sigma^2\mathbf{A} + \mathbf{I}_\mathrm{d})^{-1}\mathbf{a} \\
&= -\sigma^2\mathbf{a}^\top\mathbf{G}\mathbf{a}.
\end{aligned}$$

Therefore, $h(x) = -\frac{1}{2}\left(x^\top\mathbf{G}\mathbf{A}x - 2x^\top\mathbf{G}\mathbf{a} - \sigma^2\mathbf{a}^\top\mathbf{G}\mathbf{a}\right) = \mathcal{Q}(\mathbf{G}\mathbf{a}, \mathbf{G}\mathbf{A})(x) + \frac{\sigma^2\mathbf{a}^\top\mathbf{G}\mathbf{a}}{2}$. □

6.2 Proof of Theorem 4.14

In the balanced case, we showed that Sinkhorn's transform is stable for quadratic potentials and that the resulting sequence is a contraction. Similarly, the following proposition shows that the unbalanced Sinkhorn transform is stable for quadratic potentials.

Proposition 4.19. *Let α be an unnormalized Gaussian measure given by $m_\alpha \mathcal{N}(\mathbf{a}, \mathbf{A})$. Let $\tau = \frac{\gamma}{2\sigma^2+\gamma}$. Define the unbalanced Sinkhorn transform $T : \mathbb{R}^{\mathbb{R}^d} \to \mathbb{R}^{\mathbb{R}^d}$:*

$$T_\alpha(h)(x) \stackrel{\text{def}}{=} -\tau \log \int_{\mathbb{R}^d} e^{\frac{-\|x-y\|^2}{2\sigma^2}+h(y)} \mathrm{d}\alpha(y) \tag{4.71}$$

Let $\mathbf{U} \in \mathcal{S}_d$, $\mathbf{u} \in \mathbb{R}^d$ and $m_u > 0$. If $h = \log(m_u) + \mathcal{Q}(\mathbf{u}, \mathbf{U})$ i.e $h(x) = \log(m_u) - \frac{1}{2}(x^\top \mathbf{U} x - 2x^\top \mathbf{u})$, then $T_\alpha(h)$ is well defined if and only if $\mathbf{F} \stackrel{\text{def}}{=} \sigma^2 \mathbf{U} + \sigma^2 \mathbf{A}^{-1} + \mathbf{I}_\mathrm{d} \succ 0$, in which case $T_\alpha(h) = \mathcal{Q}(\mathbf{v}, \mathbf{V}) + \log(m_v)$ with the following parameters:

$$\mathbf{V} = \tau \frac{1}{\sigma^2}(\mathbf{F}^{-1} - \mathbf{I}_\mathrm{d}) \tag{4.72}$$

$$\mathbf{v} = -\tau \mathbf{F}^{-1}(\mathbf{A}^{-1}\mathbf{a} + \mathbf{u}) \tag{4.73}$$

$$m_v = \left(\frac{\sqrt{\det(\mathbf{A})\det(\mathbf{F})}}{m_u m_\alpha e^{\frac{q_{u,\alpha}}{2}} \sigma^{2d}} \right)^\tau, \tag{4.74}$$

where $q_{u,\alpha} = \frac{\sigma^2}{\tau^2}\mathbf{v}^\top \mathbf{F} \mathbf{v} - \mathbf{a}^\top \mathbf{A}^{-1} \mathbf{a}$.

Proof. The exponent inside the integral can be written as

$$\begin{aligned} e^{\frac{-\|x-y\|^2}{2\sigma^2}+h(y)} \mathrm{d}\alpha(y) &\propto e^{\frac{-\|x-y\|^2}{2\sigma^2} - \frac{1}{2}(y^\top \mathbf{X} y - y^\top \mathbf{A}^{-1} y)} \mathrm{d}y \\ &\propto e^{-\frac{1}{2}(y^\top (\frac{\mathbf{I}_\mathrm{d}}{\sigma^2} + \mathbf{X} + \mathbf{A}^{-1})y) + \frac{x^\top y}{\sigma^2}} \mathrm{d}y, \end{aligned}$$

which is integrable if and only if $\mathbf{U} + \mathbf{A}^{-1} + \frac{1}{\sigma^2}\mathbf{I}_\mathrm{d} \succ 0 \Leftrightarrow \mathbf{F} \succ 0$. Moreover, up to a multiplicative factor, the exponentiated Sinkhorn transform is equivalent to a Gaussian convolution of an exponentiated quadratic form. Lemma 4.18 applies:

$$\begin{aligned} e^{-T_\alpha(h)} &= \int_{\mathbb{R}^d} e^{\frac{-\|x-y\|^2}{2\sigma^2}+f(y)} \mathrm{d}\alpha(y) \\ &= m_u m_\alpha \frac{\exp(-\frac{1}{2}\mathbf{a}^\top \mathbf{A}^{-1}\mathbf{a})}{\sqrt{\det(2\pi\mathbf{A})}} \int_{\mathbb{R}^d} e^{\frac{-\|x-y\|^2}{2\sigma^2} + \mathcal{Q}(\mathbf{u},\mathbf{U})(y) + \mathcal{Q}(\mathbf{A}^{-1}\mathbf{a}, \mathbf{A}^{-1})(y)} \mathrm{d}y \\ &= m_u m_\alpha \frac{\exp(-\frac{1}{2}\mathbf{a}^\top \mathbf{A}^{-1}\mathbf{a})}{\sqrt{\det(2\pi\mathbf{A})}} \sqrt{(2\pi\sigma^2)^{2d}} \exp\left(\mathcal{N}(\sigma^2 \mathbf{I}_\mathrm{d})\right) \star \exp\left(\mathcal{Q}(\mathbf{u} + \mathbf{A}^{-1}\mathbf{a}, \mathbf{U} + \mathbf{A}^{-1})\right) \\ &= m_u m_\alpha \frac{\sigma^{2d} \exp(-\frac{1}{2}\mathbf{a}^\top \mathbf{A}^{-1}\mathbf{a})}{\sqrt{\det(\mathbf{A})}} \exp\left(\mathcal{N}(\sigma^2 \mathbf{I}_\mathrm{d})\right) \star \exp\left(\mathcal{Q}(\mathbf{u} + \mathbf{A}^{-1}\mathbf{a}, \mathbf{U} + \mathbf{A}^{-1})\right) \\ &= m_u m_\alpha \frac{\sigma^{2d} \exp(-\frac{1}{2}\mathbf{a}^\top \mathbf{A}^{-1}\mathbf{a})}{\sqrt{\det(\mathbf{A})}} c_\alpha \exp\left(\mathcal{Q}(\mathbf{F}^{-1}(\mathbf{u} + \mathbf{A}^1\mathbf{a}), \mathbf{F}^{-1}(\mathbf{U} + \mathbf{A}^{-1})\right) \\ &= m_u m_\alpha \frac{\sigma^{2d} \exp(-\frac{1}{2}\mathbf{a}^\top \mathbf{A}^{-1}\mathbf{a})}{\sqrt{\det(\mathbf{A})}} c_\alpha \exp\left(\mathcal{Q}(\mathbf{F}^{-1}(\mathbf{u} + \mathbf{A}^1\mathbf{a}), \frac{1}{\sigma^2}\mathbf{F}^{-1}(\mathbf{F} - \mathbf{I}_\mathrm{d})\right) \\ &= m_u m_\alpha \frac{\sigma^{2d} \exp(-\frac{1}{2}\mathbf{a}^\top \mathbf{A}^{-1}\mathbf{a})}{\sqrt{\det(\mathbf{A})}} c_\alpha \exp\left(\mathcal{Q}(\mathbf{F}^{-1}(\mathbf{u} + \mathbf{A}^1\mathbf{a}), \frac{1}{\sigma^2}(\mathbf{I}_\mathrm{d} - \mathbf{F}^{-1})\right), \end{aligned}$$

where $c_\alpha = \frac{\exp(\frac{1}{2}\sigma^2(\mathbf{u}+\mathbf{A}^{-1}\mathbf{a})^\top \mathbf{F}^{-1}(\mathbf{u}+\mathbf{A}^{-1}\mathbf{a}))}{\sqrt{\det(\mathbf{F})}}$.

Therefore, by applying $-\tau \log$ we can identify $\mathbf{V}$ and $\mathbf{v}$. Substituting $\mathbf{u} + \mathbf{A}^{-1}\mathbf{a}$ by $-\frac{1}{\tau}\mathbf{F}\mathbf{v}$ leads to the expression of m_v. □

Unlike in the balanced case, the unbalanced Sinkhorn iterations require 2 more parameters ($\mathbf{v}$ and m_v) with tangled updates. Proving the convergence of the resulting algorithm is more challenging. Instead, we directly solve the optimality conditions and show that a pair of quadratic potentials verifies (4.58).

Proposition 4.20. *The pair of quadratic forms (f, g) of (4.61) verifies the optimality conditions (4.58) if and only if*

$$\begin{aligned} \mathbf{F} &\stackrel{\text{def}}{=} \sigma^2 \mathbf{A}^{-1} + \sigma^2 \mathbf{U} + \mathbf{I}_{\mathrm{d}} \succ 0 \\ \mathbf{G} &\stackrel{\text{def}}{=} \sigma^2 \mathbf{B}^{-1} + \sigma^2 \mathbf{V} + \mathbf{I}_{\mathrm{d}} \succ 0, \end{aligned} \tag{4.75}$$

$$\begin{aligned} m_v \left(\frac{m_u m_\alpha e^{\frac{q_{u,\alpha}}{2}} \sigma^d}{\sqrt{\det(\mathbf{A}) \det(\mathbf{F})}} \right)^\tau &= 1 & m_u \left(\frac{m_v m_\beta e^{\frac{q_{v,\beta}}{2}} \sigma^d}{\sqrt{\det(\mathbf{B}) \det(\mathbf{G})}} \right)^\tau &= 1 \\ \mathbf{v} &= -\tau \mathbf{F}^{-1}(\mathbf{A}^{-1}\mathbf{a} + \mathbf{u}) & \mathbf{u} &= -\tau \mathbf{G}^{-1}(\mathbf{B}^{-1}\mathbf{b} + \mathbf{v}) \\ \mathbf{G} &= \tau \mathbf{F}^{-1} + \sigma^2 \mathbf{B}^{-1} + (1-\tau)\mathbf{I}_{\mathrm{d}} & \mathbf{F} &= \tau \mathbf{G}^{-1} + \sigma^2 \mathbf{A}^{-1} + (1-\tau)\mathbf{I}_{\mathrm{d}} \\ q_{u,\alpha} &= \frac{\sigma^2}{\tau^2} \mathbf{v}^\top \mathbf{F} \mathbf{v} - \mathbf{a}^\top \mathbf{A}^{-1} \mathbf{a} & q_{v,\beta} &= \frac{\sigma^2}{\tau^2} \mathbf{u}^\top \mathbf{G} \mathbf{u} - \mathbf{b}^\top \mathbf{B}^{-1} \mathbf{b} \end{aligned} \tag{4.76}$$

Proof. The equations on $m_u, m_v, \mathbf{u}, \mathbf{v}$ follow immediately from Proposition 4.19. Using the definition of $\mathbf{F}$ and $\mathbf{G}$ and substituting $\mathbf{U}$ and $\mathbf{V}$ leads to the equations in $\mathbf{F}$ and $\mathbf{G}$. □

We now turn to solve the system (4.76). Notice that in general, the dual potentials can only be identified up to a an additive constant. Indeed, if a pair (f, g) is optimal, then $(f + K, g - K)$ is also optimal for any $K \in \mathbb{R}$ (the transportation plan and dual objective do not change). Thus, at optimality, it is sufficient to obtain the product $m_u m_v$. We start by identifying $(\mathbf{F}, \mathbf{G})$ then $(\mathbf{u}, \mathbf{v})$ and finally $m_u m_v$.

Identifying F and G. The equations in $\mathbf{F}$ and $\mathbf{G}$ can be shown to be equivalent to those of the balanced case up to a change of variables. Let $\lambda = \frac{1-\tau}{\sigma^2}$. Then,

$$\begin{aligned} &\begin{cases} \mathbf{F} &= \tau \mathbf{G}^{-1} + \sigma^2 \mathbf{A}^{-1} + (1-\tau)\mathbf{I}_{\mathrm{d}} \\ \mathbf{G} &= \tau \mathbf{F}^{-1} + \sigma^2 \mathbf{B}^{-1} + (1-\tau)\mathbf{I}_{\mathrm{d}} \end{cases} \\ \Leftrightarrow &\begin{cases} \mathbf{F} &= \left(\frac{\mathbf{G}}{\tau}\right)^{-1} + \frac{\sigma^2}{\tau}\tau(\mathbf{A}^{-1} + \frac{1}{\lambda}\mathbf{I}_{\mathrm{d}}) \\ \frac{\mathbf{G}}{\tau} &= \mathbf{F}^{-1} + \frac{\sigma^2}{\tau}(\mathbf{B}^{-1} + \frac{1}{\lambda}\mathbf{I}_{\mathrm{d}}) \end{cases} \\ \Leftrightarrow &\begin{cases} \mathbf{F} &= \widetilde{\mathbf{G}}^{-1} + \sigma^2 (\frac{\widetilde{\mathbf{A}}}{\tau})^{-1} \\ \widetilde{\mathbf{G}} &= \mathbf{F}^{-1} + \sigma^2 \widetilde{\mathbf{B}}^{-1}, \end{cases} \end{aligned}$$

which correspond to the balanced OT fixed point equations (4.21) associated with the pair $(\frac{\widetilde{\mathbf{A}}}{\tau}, \widetilde{\mathbf{B}})$ with the following change of variables:

$$\widetilde{\mathbf{G}} \stackrel{\text{def}}{=} \frac{\mathbf{G}}{\tau} \tag{4.77}$$

$$\widetilde{\mathbf{A}} \stackrel{\text{def}}{=} \tau(\mathbf{A}^{-1} + \frac{1}{\lambda}\mathbf{I}_{\mathrm{d}})^{-1} \tag{4.78}$$

$$\widetilde{\mathbf{B}} \stackrel{\text{def}}{=} \tau(\mathbf{B}^{-1} + \frac{1}{\lambda}\mathbf{I}_{\mathrm{d}})^{-1}. \tag{4.79}$$

Notice that since $0 < \tau < 1$, $\widetilde{\mathbf{A}}$ and $\widetilde{\mathbf{B}}$ are well-defined and positive-definite. Therefore, Proposition 4.6 applies and we obtain the closed form

$$\begin{aligned} \mathbf{C} \stackrel{\text{def}}{=} \widetilde{\mathbf{A}}\widetilde{\mathbf{G}}^{-1} &= \left(\frac{1}{\tau}\widetilde{\mathbf{A}}\widetilde{\mathbf{B}} + \frac{\sigma^4}{4}\mathbf{I}_{\mathrm{d}} \right)^{\frac{1}{2}} - \frac{\sigma^2}{2}\mathbf{I}_{\mathrm{d}} \\ &= \widetilde{\mathbf{A}}^{\frac{1}{2}} \left(\frac{1}{\tau}\widetilde{\mathbf{A}}^{\frac{1}{2}}\widetilde{\mathbf{B}}\widetilde{\mathbf{A}}^{\frac{1}{2}} + \frac{\sigma^4}{4}\mathbf{I}_{\mathrm{d}} \right)^{\frac{1}{2}} \widetilde{\mathbf{A}}^{-\frac{1}{2}} - \frac{\sigma^2}{2}\mathbf{I}_{\mathrm{d}}. \end{aligned} \tag{4.80}$$

Similarly, by symmetry:

$$\widetilde{\mathbf{B}}\mathbf{F}^{-1} = \left(\frac{1}{\tau}\widetilde{\mathbf{B}}\widetilde{\mathbf{A}} + \frac{\sigma^4}{4}\mathbf{I}_{\mathrm{d}}\right)^{\frac{1}{2}} - \frac{\sigma^2}{2}\mathbf{I}_{\mathrm{d}} = \mathbf{C}^\top. \tag{4.81}$$

Therefore we obtain $\mathbf{F}$ and $\mathbf{G}$ in closed form:

$$\mathbf{F} = \widetilde{\mathbf{B}}\mathbf{C}^{-1} \tag{4.82}$$

$$\mathbf{G} = \mathbf{C}^{-1}\widetilde{\mathbf{A}}. \tag{4.83}$$

Finally, to obtain the formulas of $\widetilde{\mathbf{A}}$ and $\widetilde{\mathbf{B}}$ of Theorem 4.14, we use Woodburry's identity to write

$$\begin{aligned}
\widetilde{\mathbf{B}} &= \tau\lambda(\mathbf{I}_{\mathrm{d}} - \lambda(\mathbf{B} + \lambda\mathbf{I}_{\mathrm{d}})^{-1}) \\
&= \frac{\gamma}{\gamma + 2\sigma^2}\frac{2\sigma^2 + \gamma}{2}(\mathbf{I}_{\mathrm{d}} - \lambda(\mathbf{B} + \lambda\mathbf{I}_{\mathrm{d}})^{-1}) \\
&= \frac{\gamma}{2}(\mathbf{I}_{\mathrm{d}} - \lambda(\mathbf{B} + \lambda\mathbf{I}_{\mathrm{d}})^{-1}).
\end{aligned}$$

The same derivation applies for $\widetilde{\mathbf{A}}$.

Identifying u and v. Combining the equations in $\mathbf{u}$ and $\mathbf{v}$ leads to

$$\begin{aligned}
&\mathbf{v} = -\tau\mathbf{F}^{-1}(\mathbf{A}^{-1}\mathbf{a} + \tau\mathbf{u}) \\
&\Leftrightarrow \mathbf{F}\mathbf{v} = -\tau\mathbf{A}^{-1}\mathbf{a} - \tau\mathbf{u} \\
&\Leftrightarrow \mathbf{F}\mathbf{v} = -\tau\mathbf{A}^{-1}\mathbf{a} + \tau^2\mathbf{G}^{-1}(\mathbf{B}^{-1}\mathbf{b} + \mathbf{v}) \\
&\Leftrightarrow \mathbf{G}\mathbf{F}\mathbf{v} = -\tau\mathbf{G}\mathbf{A}^{-1}\mathbf{a} + \tau^2(\mathbf{B}^{-1}\mathbf{b} + \mathbf{v}) \\
&\Leftrightarrow (\mathbf{G}\mathbf{F} - \tau^2\mathbf{I}_{\mathrm{d}})\mathbf{v} = -\tau\mathbf{G}\mathbf{A}^{-1}\mathbf{a} + \tau^2\mathbf{B}^{-1}\mathbf{b}.
\end{aligned}$$

Similarly, $(\mathbf{F}\mathbf{G} - \tau^2\mathbf{I}_{\mathrm{d}})\mathbf{u} = -\tau\mathbf{F}\mathbf{B}^{-1}\mathbf{b} + \tau^2\mathbf{A}^{-1}\mathbf{a}$. Moreover, since $0 < \tau < 1$, it holds that

$$\begin{aligned}
(\mathbf{F} - \tau^2\mathbf{G}^{-1}) &\succ (\mathbf{F} - \tau\mathbf{G}^{-1}) \\
&= \sigma^2\widetilde{\mathbf{A}}^{-1} \succ 0.
\end{aligned}$$

Therefore, $(\mathbf{F}\mathbf{G} - \tau^2\mathbf{I}_{\mathrm{d}}) = (\mathbf{F} - \tau^2\mathbf{G}^{-1}\mathbf{I}_{\mathrm{d}})\mathbf{G}$ is invertible. The same applies for $(\mathbf{G}\mathbf{F} - \tau^2\mathbf{I}_{\mathrm{d}})$.

Finally, both equations can be vectorized:

$$\begin{pmatrix} \mathbf{G}\mathbf{F} - \tau^2\mathbf{I}_{\mathrm{d}} & 0 \\ 0 & \mathbf{F}\mathbf{G} - \tau^2\mathbf{I}_{\mathrm{d}} \end{pmatrix}\begin{pmatrix} \mathbf{v} \\ \mathbf{u} \end{pmatrix} = \begin{pmatrix} -\tau\mathbf{G} & \tau^2\mathbf{I}_{\mathrm{d}} \\ \tau^2\mathbf{I}_{\mathrm{d}} & -\tau\mathbf{F} \end{pmatrix}\begin{pmatrix} \mathbf{A}^{-1} & 0 \\ 0 & \mathbf{B}^{-1} \end{pmatrix}\begin{pmatrix} \mathbf{a} \\ \mathbf{b} \end{pmatrix} \tag{4.84}$$

Identifying $m_u m_v$. Now that $\mathbf{F}, \mathbf{G}, \mathbf{u}$ and $\mathbf{v}$ are given in closed form, $m_u m_v$ is obtained by taking the product of both equations:

$$(m_u m_v)^{\tau+1} = \left(\frac{\sqrt{\det(\mathbf{A}\mathbf{B})\det(\mathbf{F}\mathbf{G})}}{\sigma^{2d} m_\alpha m_\beta}\right)^{\tau} \exp(-\frac{\tau}{2}(q_{u,\alpha} + q_{v,\beta})). \tag{4.85}$$

Transportation plan. Let $\omega \stackrel{\text{def}}{=} \frac{m_\alpha m_\beta}{\sqrt{\det(4\pi^2\mathbf{AB})}} m_u m_v e^{-\frac{1}{2}(\mathbf{a}^\top\mathbf{A}^{-1}\mathbf{a}+\mathbf{b}^\top\mathbf{B}^{-1}\mathbf{b})}$. At optimality, the transport plan π is given by

$$\begin{aligned}
\frac{\mathrm{d}\pi}{\mathrm{d}x\mathrm{d}y}(x,y) &= \exp\left(\frac{f(x)+g(y)-\|x-y\|^2}{2\sigma^2}\right)\frac{\mathrm{d}\alpha}{\mathrm{d}x}(x)\frac{\mathrm{d}\beta}{\mathrm{d}y}(y)\\
&= \omega\exp\left(\mathcal{Q}(\mathbf{A}^{-1}\mathbf{a}+\mathbf{u},\mathbf{A}^{-1}+\mathbf{U})(x) - \frac{\|x-y\|^2}{2\sigma^2} + \mathcal{Q}(\mathbf{B}^{-1}\mathbf{b}+\mathbf{v},\mathbf{B}^{-1}+\mathbf{V})(y)\right)\\
&= \omega\exp\left(\mathcal{Q}(\mathbf{U}+\mathbf{A}^{-1})(x) + \mathcal{Q}(\mathbf{V}+\mathbf{B}^{-1})(y) + \mathcal{Q}(\begin{smallmatrix}\frac{\mathbf{I}_\mathrm{d}}{\sigma^2} & -\frac{\mathbf{I}_\mathrm{d}}{\sigma^2}\\ -\frac{\mathbf{I}_\mathrm{d}}{\sigma^2} & \frac{\mathbf{I}_\mathrm{d}}{\sigma^2}\end{smallmatrix})(x,y)\right)\\
&= \omega\exp\left(\mathcal{Q}\left(\begin{pmatrix}\mathbf{A}^{-1}\mathbf{a}+\mathbf{u}\\ \mathbf{B}^{-1}\mathbf{b}+\mathbf{v}\end{pmatrix}, \begin{pmatrix}\mathbf{U}+\mathbf{A}^{-1}+\frac{\mathbf{I}_\mathrm{d}}{\sigma^2} & 0\\ 0 & \mathbf{V}+\mathbf{B}^{-1}+\frac{\mathbf{I}_\mathrm{d}}{\sigma^2}\end{pmatrix}\right)(x,y)\right)\\
&= \omega\exp\left(\mathcal{Q}\left(\begin{pmatrix}\mathbf{A}^{-1}\mathbf{a}+\mathbf{u}\\ \mathbf{B}^{-1}\mathbf{b}+\mathbf{v}\end{pmatrix}, \frac{1}{\sigma^2}\begin{pmatrix}\mathbf{F} & -\mathbf{I}_\mathrm{d}\\ -\mathbf{I}_\mathrm{d} & \mathbf{G}\end{pmatrix}\right)(x,y)\right)\\
&= \omega\exp\left(\mathcal{Q}(\mu,\Gamma)(x,y)\right),
\end{aligned}$$

with $\mu \stackrel{\text{def}}{=} \begin{pmatrix}\mathbf{A}^{-1}\mathbf{a}+\mathbf{u}\\ \mathbf{B}^{-1}\mathbf{b}+\mathbf{v}\end{pmatrix}$ and $\Gamma \stackrel{\text{def}}{=} \begin{pmatrix}\frac{\mathbf{F}}{\sigma^2} & -\frac{\mathbf{I}_\mathrm{d}}{\sigma^2}\\ -\frac{\mathbf{I}_\mathrm{d}}{\sigma^2} & \frac{\mathbf{G}}{\sigma^2}\end{pmatrix}$. Let us show that Γ is positive definite. Since $\frac{\mathbf{G}}{2\sigma^2} \succ 0$, it is sufficient to show that Schur complement $\frac{\mathbf{F}}{\sigma^2} - \frac{1}{\sigma^2}\mathbf{G}^{-1}$ is positive definite. First, we have

$$\frac{\mathbf{F}-\mathbf{G}^{-1}}{\sigma^2} = \tau\widetilde{\mathbf{A}}^{-1} - \frac{1}{\lambda}\mathbf{G}^{-1}.$$

Next, it holds that $\widetilde{\mathbf{A}} \prec \tau\lambda\mathbf{I}_\mathrm{d}$ and $\widetilde{\mathbf{B}} \prec \tau\lambda\mathbf{I}_\mathrm{d}$. Thus, for any $x \in \mathbb{R}^d$ we have

$$x^\top\frac{\widetilde{\mathbf{A}}^{\frac{1}{2}}\widetilde{\mathbf{B}}\widetilde{\mathbf{A}}^{\frac{1}{2}}}{\tau}x \le \lambda\|\widetilde{\mathbf{A}}^{\frac{1}{2}}x\|^2 = \lambda x^\top\widetilde{\mathbf{A}}x \le \tau\lambda^2\|x\|^2,$$

which implies

$$\left(\frac{\widetilde{\mathbf{A}}^{\frac{1}{2}}\widetilde{\mathbf{B}}\widetilde{\mathbf{A}}^{\frac{1}{2}}}{\tau} + \frac{\sigma^4}{4}\mathbf{I}_\mathrm{d}\right)^{\frac{1}{2}} \prec \sqrt{\tau\lambda^2+\frac{\sigma^4}{4}}\mathbf{I}_\mathrm{d} = \frac{\lambda}{2}(\sqrt{4\tau+(1-\tau)^2})\mathbf{I}_\mathrm{d} = \frac{\lambda(1+\tau)}{2}\mathbf{I}_\mathrm{d}.$$

Therefore, using the second equality of (4.80) and inverting (4.82) to obtain $\mathbf{G}^{-1}$:

$$\begin{aligned}
x^\top\mathbf{G}^{-1}x &= x^\top\widetilde{\mathbf{A}}^{-\frac{1}{2}}\left(\left(\frac{\widetilde{\mathbf{A}}^{\frac{1}{2}}\widetilde{\mathbf{B}}\widetilde{\mathbf{A}}^{\frac{1}{2}}}{\tau}+\frac{\sigma^4}{4}\mathbf{I}_\mathrm{d}\right)^{\frac{1}{2}} - \frac{\sigma^2}{2}\mathbf{I}_\mathrm{d})\right)\widetilde{\mathbf{A}}^{-\frac{1}{2}}x\\
&= (\widetilde{\mathbf{A}}^{-\frac{1}{2}}x)^\top\left(\left(\frac{\widetilde{\mathbf{A}}^{\frac{1}{2}}\widetilde{\mathbf{B}}\widetilde{\mathbf{A}}^{\frac{1}{2}}}{\tau}+\frac{\sigma^4}{4}\mathbf{I}_\mathrm{d}\right)^{\frac{1}{2}} - \frac{\lambda(1-\tau)}{2}\mathbf{I}_\mathrm{d})\right)(\widetilde{\mathbf{A}}^{-\frac{1}{2}}x)\\
&\le (\widetilde{\mathbf{A}}^{-\frac{1}{2}}x)^\top\left(\frac{\lambda(1+\tau)}{2}\mathbf{I}_\mathrm{d} - \frac{\lambda(1-\tau)}{2}\mathbf{I}_\mathrm{d})\right)(\widetilde{\mathbf{A}}^{-\frac{1}{2}}x)\\
&= \tau\lambda x^\top\widetilde{\mathbf{A}}^{-1}x.
\end{aligned}$$

Hence $\mathbf{G}^{-1} \prec \tau\lambda\widetilde{\mathbf{A}}^{-1}$. We can therefore conclude that the Schur complement $\frac{1}{\sigma^2}(\mathbf{F}-\mathbf{G}^{-1})$ is positive definite. By completing the square, we can factor $\frac{\mathrm{d}\pi}{\mathrm{d}x\mathrm{d}x}$ as a Gaussian density.

Let $z \overset{\text{def}}{=} \binom{x}{y}$. Then,

$$\begin{aligned}\frac{\mathrm{d}\pi}{\mathrm{d}x\mathrm{d}y}(x,y) &= \omega \exp\left(\mathcal{Q}(\mu,\Gamma)(x,y)\right)\\ &= \omega \exp\left(-\frac{1}{2}(z^\top \Gamma z - 2z^\top \mu)\right)\\ &= \omega \exp\left(\frac{1}{2}\mu^\top\Gamma^{-1}\mu - \frac{1}{2}(z-\Gamma^{-1}\mu)^\top\Gamma(z-\Gamma^{-1}\mu)\right)\\ &= \omega e^{\frac{1}{2}\mu^\top\Gamma^{-1}\mu}\mathcal{N}(\mathbf{H}\mu,\mathbf{H})(z),\end{aligned}$$

where $\mathbf{H} = \Gamma^{-1}$.

Detailed expressions. To conclude the proof of Theorem 4.14, we need to simplify the formulas of $m, \mathbf{H}\mu$ and $\mathbf{H}$. Let us start with the mean $\mathbf{H}\mu$.

$\mathbf{H}\mu$: Using the optimality conditions of Proposition 4.20 and the closed form formula of $\mathbf{v}$ and $\mathbf{u}$, we get

$$\begin{aligned}\mu &= \begin{pmatrix}\mathbf{A}^{-1}\mathbf{a}+\mathbf{u}\\ \mathbf{B}^{-1}\mathbf{b}+\mathbf{v}\end{pmatrix}\\ &= -\frac{1}{\tau}\begin{pmatrix}\mathbf{F}\mathbf{v}\\ \mathbf{G}\mathbf{u}\end{pmatrix}\\ &= -\frac{1}{\tau}\begin{pmatrix}\mathbf{F} & 0\\ 0 & \mathbf{G}\end{pmatrix}\begin{pmatrix}\mathbf{v}\\ \mathbf{u}\end{pmatrix}\\ &= -\frac{1}{\tau}\begin{pmatrix}\mathbf{F} & 0\\ 0 & \mathbf{G}\end{pmatrix}\begin{pmatrix}\mathbf{G}\mathbf{F}-\tau^2\mathbf{I}_\mathrm{d} & 0\\ 0 & \mathbf{F}\mathbf{G}-\tau^2\mathbf{I}_\mathrm{d}\end{pmatrix}^{-1}\begin{pmatrix}-\tau\mathbf{G} & \tau^2\mathbf{I}_\mathrm{d}\\ \tau^2\mathbf{I}_\mathrm{d} & -\tau\mathbf{F}\end{pmatrix}\begin{pmatrix}\mathbf{A}^{-1} & 0\\ 0 & \mathbf{B}^{-1}\end{pmatrix}\begin{pmatrix}\mathbf{a}\\ \mathbf{b}\end{pmatrix}\\ &= \begin{pmatrix}\mathbf{F} & 0\\ 0 & \mathbf{G}\end{pmatrix}\begin{pmatrix}\mathbf{G}\mathbf{F}-\tau^2\mathbf{I}_\mathrm{d} & 0\\ 0 & \mathbf{F}\mathbf{G}-\tau^2\mathbf{I}_\mathrm{d}\end{pmatrix}^{-1}\begin{pmatrix}\mathbf{G} & -\tau\mathbf{I}_\mathrm{d}\\ -\tau\mathbf{I}_\mathrm{d} & \mathbf{F}\end{pmatrix}\begin{pmatrix}\mathbf{A}^{-1} & 0\\ 0 & \mathbf{B}^{-1}\end{pmatrix}\begin{pmatrix}\mathbf{a}\\ \mathbf{b}\end{pmatrix}\\ &= \begin{pmatrix}\mathbf{F} & 0\\ 0 & \mathbf{G}\end{pmatrix}\begin{pmatrix}(\mathbf{F}-\tau^2\mathbf{G}^{-1})^{-1} & -\tau(\mathbf{G}\mathbf{F}-\tau^2\mathbf{I}_\mathrm{d})^{-1}\\ -\tau(\mathbf{F}\mathbf{G}-\tau^2\mathbf{I}_\mathrm{d})^{-1} & (\mathbf{G}-\tau^2\mathbf{F}^{-1})^{-1}\end{pmatrix}\begin{pmatrix}\mathbf{A}^{-1} & 0\\ 0 & \mathbf{B}^{-1}\end{pmatrix}\begin{pmatrix}\mathbf{a}\\ \mathbf{b}\end{pmatrix}\\ &= \begin{pmatrix}\mathbf{F} & 0\\ 0 & \mathbf{G}\end{pmatrix}\begin{pmatrix}\mathbf{F} & \tau\mathbf{I}_\mathrm{d}\\ \tau\mathbf{I}_\mathrm{d} & \mathbf{G}\end{pmatrix}^{-1}\begin{pmatrix}\mathbf{A}^{-1} & 0\\ 0 & \mathbf{B}^{-1}\end{pmatrix}\begin{pmatrix}\mathbf{a}\\ \mathbf{b}\end{pmatrix}\\ &= \begin{pmatrix}\mathbf{I}_\mathrm{d} & \tau\mathbf{G}^{-1}\\ \tau\mathbf{F}^{-1} & \mathbf{I}_\mathrm{d}\end{pmatrix}^{-1}\begin{pmatrix}\mathbf{A}^{-1} & 0\\ 0 & \mathbf{B}^{-1}\end{pmatrix}\begin{pmatrix}\mathbf{a}\\ \mathbf{b}\end{pmatrix}.\end{aligned} \tag{4.86}$$

Therefore,

$$\begin{aligned}\mathbf{H}\mu &= \sigma^2\begin{pmatrix}\mathbf{F} & -\mathbf{I}_\mathrm{d}\\ -\mathbf{I}_\mathrm{d} & \mathbf{G}\end{pmatrix}^{-1}\begin{pmatrix}\mathbf{I}_\mathrm{d} & \tau\mathbf{G}^{-1}\mathbf{I}_\mathrm{d}\\ \tau\mathbf{F}^{-1}\mathbf{I}_\mathrm{d} & \mathbf{I}_\mathrm{d}\end{pmatrix}^{-1}\begin{pmatrix}\mathbf{A}^{-1} & 0\\ 0 & \mathbf{B}^{-1}\end{pmatrix}\begin{pmatrix}\mathbf{a}\\ \mathbf{b}\end{pmatrix}\\ &= \sigma^2\left(\begin{pmatrix}\mathbf{I}_\mathrm{d} & \tau\mathbf{G}^{-1}\mathbf{I}_\mathrm{d}\\ \tau\mathbf{F}^{-1}\mathbf{I}_\mathrm{d} & \mathbf{I}_\mathrm{d}\end{pmatrix}\begin{pmatrix}\mathbf{F} & -\mathbf{I}_\mathrm{d}\\ -\mathbf{I}_\mathrm{d} & \mathbf{G}\end{pmatrix}\right)^{-1}\begin{pmatrix}\mathbf{A}^{-1} & 0\\ 0 & \mathbf{B}^{-1}\end{pmatrix}\begin{pmatrix}\mathbf{a}\\ \mathbf{b}\end{pmatrix}\\ &= \sigma^2\begin{pmatrix}\mathbf{F}-\tau\mathbf{G}^{-1} & -(1-\tau)\mathbf{I}_\mathrm{d}\\ -(1-\tau)\mathbf{I}_\mathrm{d} & \mathbf{G}-\tau\mathbf{F}^{-1}\end{pmatrix}^{-1}\begin{pmatrix}\mathbf{A}^{-1} & 0\\ 0 & \mathbf{B}^{-1}\end{pmatrix}\begin{pmatrix}\mathbf{a}\\ \mathbf{b}\end{pmatrix}\\ &= \sigma^2\begin{pmatrix}\sigma^2\mathbf{A}^{-1}+(1-\tau)\mathbf{I}_\mathrm{d} & -(1-\tau)\mathbf{I}_\mathrm{d}\\ -(1-\tau)\mathbf{I}_\mathrm{d} & \sigma^2\mathbf{B}^{-1}+(1-\tau)\mathbf{I}_\mathrm{d}\end{pmatrix}^{-1}\begin{pmatrix}\mathbf{A}^{-1} & 0\\ 0 & \mathbf{B}^{-1}\end{pmatrix}\begin{pmatrix}\mathbf{a}\\ \mathbf{b}\end{pmatrix}\\ &= \begin{pmatrix}\mathbf{A}^{-1}+\mathbf{I}_\mathrm{d} & -\lambda\mathbf{I}_\mathrm{d}\\ -\lambda\mathbf{I}_\mathrm{d} & \mathbf{B}^{-1}+\lambda\mathbf{I}_\mathrm{d}\end{pmatrix}^{-1}\begin{pmatrix}\mathbf{A}^{-1} & 0\\ 0 & \mathbf{B}^{-1}\end{pmatrix}\begin{pmatrix}\mathbf{a}\\ \mathbf{b}\end{pmatrix}.\end{aligned} \tag{4.87}$$

Let us now compute the inverse of

$$\mathbf{Z} \stackrel{\text{def}}{=} \begin{pmatrix} \mathbf{A}^{-1} + \frac{1}{\lambda}\mathbf{I}_{\mathrm{d}} & -\frac{1}{\lambda}\mathbf{I}_{\mathrm{d}} \\ -\frac{1}{\lambda}\mathbf{I}_{\mathrm{d}} & \mathbf{B}^{-1} + \frac{1}{\lambda}\mathbf{I}_{\mathrm{d}} \end{pmatrix}. \tag{4.88}$$

Let $\mathbf{S}$ and $\mathbf{S}'$ be the respective Schur complements of $\mathbf{A}^{-1} + \frac{1}{\lambda}\mathbf{I}_{\mathrm{d}}$ and $\mathbf{B}^{-1} + \frac{1}{\lambda}\mathbf{I}_{\mathrm{d}}$ in $\mathbf{Z}$. The block inverse formula yields

$$\mathbf{Z}^{-1} = \begin{pmatrix} \mathbf{S} & \frac{1}{\lambda}\mathbf{S}(\mathbf{B}^{-1} + \frac{1}{\lambda}\mathbf{I}_{\mathrm{d}})^{-1} \\ \frac{1}{\lambda}(\mathbf{A}^{-1} + \frac{1}{\lambda}\mathbf{I}_{\mathrm{d}})^{-1}\mathbf{S} & \mathbf{S}' \end{pmatrix}.$$

Using Woodbury's identity twice and denoting $\mathbf{X} \stackrel{\text{def}}{=} \mathbf{A} + \mathbf{B} + \lambda\mathbf{I}_{\mathrm{d}}$, we get

$$\begin{aligned} \mathbf{S} &= (\mathbf{A}^{-1} + \frac{1}{\lambda}\mathbf{I}_{\mathrm{d}} - \frac{1}{\lambda^2}(\mathbf{B}^{-1} + \frac{1}{\lambda}\mathbf{I}_{\mathrm{d}})^{-1})^{-1} \\ &= (\mathbf{A}^{-1} + (\mathbf{B} + \lambda\mathbf{I}_{\mathrm{d}})^{-1})^{-1} \\ &= (\mathbf{A} - \mathbf{A}(\mathbf{A} + \mathbf{B} + \lambda\mathbf{I}_{\mathrm{d}})^{-1}\mathbf{A}) \\ &= \mathbf{A} - \mathbf{A}\mathbf{X}^{-1}\mathbf{A}, \end{aligned}$$

and similarly: $\mathbf{S}' = \mathbf{B} - \mathbf{B}\mathbf{X}^{-1}\mathbf{B}$. The off-diagonal blocks can be simplified as well:

$$\begin{aligned} \frac{1}{\lambda}\mathbf{S}(\mathbf{B}^{-1} + \frac{1}{\lambda}\mathbf{I}_{\mathrm{d}})^{-1} &= \frac{1}{\lambda}(\mathbf{A}^{-1} + (\mathbf{B} + \lambda\mathbf{I}_{\mathrm{d}})^{-1})^{-1}(\mathbf{B}^{-1} + \frac{1}{\lambda}\mathbf{I}_{\mathrm{d}})^{-1} \\ &= (\mathbf{A}^{-1} + (\mathbf{B} + \lambda\mathbf{I}_{\mathrm{d}})^{-1})^{-1}(\lambda\mathbf{I}_{\mathrm{d}} + \mathbf{B}\mathbf{I}_{\mathrm{d}})^{-1}\mathbf{B} \\ &= \big((\mathbf{B} + \lambda\mathbf{I}_{\mathrm{d}}) - (\mathbf{B} + \lambda\mathbf{I}_{\mathrm{d}})(\mathbf{A} + \mathbf{B} + \lambda\mathbf{I}_{\mathrm{d}})^{-1}(\mathbf{B} + \lambda\mathbf{I}_{\mathrm{d}})\big)\,(\lambda\mathbf{I}_{\mathrm{d}} + \mathbf{B}\mathbf{I}_{\mathrm{d}})^{-1}\mathbf{B} \\ &= \mathbf{B} - (\mathbf{B} + \lambda\mathbf{I}_{\mathrm{d}})\mathbf{X}^{-1}\mathbf{B} \\ &= \mathbf{B} - (\mathbf{X} - \mathbf{A})\mathbf{X}^{-1}\mathbf{B} \\ &= \mathbf{A}\mathbf{X}^{-1}\mathbf{B}. \end{aligned}$$

Similarly, $\frac{1}{\lambda}(\mathbf{A}^{-1} + \frac{1}{\lambda}\mathbf{I}_{\mathrm{d}})^{-1}\mathbf{S} = \mathbf{B}\mathbf{X}^{-1}\mathbf{A}$. Thus, the inverse of $\mathbf{Z}$ is given by

$$\mathbf{Z}^{-1} = \begin{pmatrix} \mathbf{A} - \mathbf{A}\mathbf{X}^{-1}\mathbf{A} & \mathbf{A}\mathbf{X}^{-1}\mathbf{B} \\ \mathbf{B}\mathbf{X}^{-1}\mathbf{A} & \mathbf{B} - \mathbf{B}\mathbf{X}^{-1}\mathbf{B} \end{pmatrix}, \tag{4.89}$$

and finally:

$$\begin{aligned} \mathbf{H}\mu = \mathbf{Z}^{-1}\begin{pmatrix} \mathbf{A}^{-1} & 0 \\ 0 & \mathbf{B}^{-1} \end{pmatrix}\begin{pmatrix} \mathbf{a} \\ \mathbf{b} \end{pmatrix} &= \begin{pmatrix} \mathbf{I}_{\mathrm{d}} - \mathbf{A}\mathbf{X}^{-1} & \mathbf{A}\mathbf{X}^{-1} \\ \mathbf{B}\mathbf{X}^{-1} & \mathbf{I}_{\mathrm{d}} - \mathbf{B}\mathbf{X}^{-1} \end{pmatrix}\begin{pmatrix} \mathbf{a} \\ \mathbf{b} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{a} + \mathbf{A}\mathbf{X}^{-1}(\mathbf{b} - \mathbf{a}) \\ \mathbf{b} + \mathbf{B}\mathbf{X}^{-1}(\mathbf{a} - \mathbf{b}) \end{pmatrix}. \end{aligned}$$

Finding the covariance matrix H: To compute $\mathbf{H} = \left(\frac{1}{\sigma^2}\begin{pmatrix} \mathbf{F} & -\mathbf{I}_{\mathrm{d}} \\ -\mathbf{I}_{\mathrm{d}} & \mathbf{G} \end{pmatrix}\right)^{-1}$ one may use the block inverse formula. However, the Schur complement $(\mathbf{F} - \mathbf{G}^{-1})^{-1}$ is not easy to manipulate. Instead notice that the following holds:

$$\begin{aligned} \frac{1}{\sigma^2}\begin{pmatrix} \mathbf{F} & -\mathbf{I}_{\mathrm{d}} \\ -\mathbf{I}_{\mathrm{d}} & \mathbf{G} \end{pmatrix}\begin{pmatrix} \mathbf{I}_{\mathrm{d}} & \tau\mathbf{F}^{-1} \\ \tau\mathbf{G}^{-1} & \mathbf{I}_{\mathrm{d}} \end{pmatrix} &= \frac{1}{\sigma^2}\begin{pmatrix} \mathbf{F} - \tau\mathbf{G}^{-1} & -(1-\tau)\mathbf{I}_{\mathrm{d}} \\ -(1-\tau)\mathbf{I}_{\mathrm{d}} & \mathbf{G} - \tau\mathbf{F}^{-1} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{A}^{-1} + \frac{1}{\lambda}\mathbf{I}_{\mathrm{d}} & -\frac{1}{\lambda}\mathbf{I}_{\mathrm{d}} \\ -\frac{1}{\lambda}\mathbf{I}_{\mathrm{d}} & \mathbf{B}^{-1} + \frac{1}{\lambda}\mathbf{I}_{\mathrm{d}} \end{pmatrix}, \end{aligned}$$

where the last equality follows from the optimality conditions (4.76). Therefore,

$$\mathbf{H} = \begin{pmatrix} \mathbf{I}_{\mathrm{d}} & \tau\mathbf{F}^{-1} \\ \tau\mathbf{G}^{-1} & \mathbf{I}_{\mathrm{d}} \end{pmatrix} \begin{pmatrix} \mathbf{A}^{-1} + \frac{1}{\lambda}\mathbf{I}_{\mathrm{d}} & -\frac{1}{\lambda}\mathbf{I}_{\mathrm{d}} \\ -\frac{1}{\lambda}\mathbf{I}_{\mathrm{d}} & \mathbf{B}^{-1} + \frac{1}{\lambda}\mathbf{I}_{\mathrm{d}} \end{pmatrix}^{-1}.$$

Notice that we have already computed the inverse matrix on the right side above in the developments of $\mathbf{H}\mu$. Thus,

$$\begin{aligned}
\mathbf{H} &= \begin{pmatrix} \mathbf{I}_{\mathrm{d}} & \tau\mathbf{F}^{-1} \\ \tau\mathbf{G}^{-1} & \mathbf{I}_{\mathrm{d}} \end{pmatrix} \begin{pmatrix} \mathbf{A} - \mathbf{A}\mathbf{X}^{-1}\mathbf{A} & \mathbf{A}\mathbf{X}^{-1}\mathbf{B} \\ \mathbf{B}\mathbf{X}^{-1}\mathbf{A} & \mathbf{B} - \mathbf{B}\mathbf{X}^{-1}\mathbf{B} \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{I}_{\mathrm{d}} & \tau\mathbf{C}\widetilde{\mathbf{B}}^{-1} \\ \mathbf{C}^{\top}\widetilde{\mathbf{A}}^{-1} & \mathbf{I}_{\mathrm{d}} \end{pmatrix} \begin{pmatrix} \mathbf{A} - \mathbf{A}\mathbf{X}^{-1}\mathbf{A} & \mathbf{A}\mathbf{X}^{-1}\mathbf{B} \\ \mathbf{B}\mathbf{X}^{-1}\mathbf{A} & \mathbf{B} - \mathbf{B}\mathbf{X}^{-1}\mathbf{B} \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{I}_{\mathrm{d}} & \mathbf{C}(\mathbf{B}^{-1} + \frac{1}{\lambda}\mathbf{I}_{\mathrm{d}}) \\ \mathbf{C}^{\top}(\mathbf{A}^{-1} + \frac{1}{\lambda}\mathbf{I}_{\mathrm{d}}) & \mathbf{I}_{\mathrm{d}} \end{pmatrix} \begin{pmatrix} \mathbf{A} - \mathbf{A}\mathbf{X}^{-1}\mathbf{A} & \mathbf{A}\mathbf{X}^{-1}\mathbf{B} \\ \mathbf{B}\mathbf{X}^{-1}\mathbf{A} & \mathbf{B} - \mathbf{B}\mathbf{X}^{-1}\mathbf{B} \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{I}_{\mathrm{d}} & \mathbf{C}(\mathbf{B}^{-1} + \frac{1}{\lambda}\mathbf{I}_{\mathrm{d}}) \\ \mathbf{C}^{\top}(\mathbf{A}^{-1} + \frac{1}{\lambda}\mathbf{I}_{\mathrm{d}}) & \mathbf{I}_{\mathrm{d}} \end{pmatrix} \begin{pmatrix} \mathbf{A} - \mathbf{A}\mathbf{X}^{-1}\mathbf{A} & \mathbf{A}\mathbf{X}^{-1}\mathbf{B} \\ \mathbf{B}\mathbf{X}^{-1}\mathbf{A} & \mathbf{B} - \mathbf{B}\mathbf{X}^{-1}\mathbf{B} \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{I}_{\mathrm{d}} & \frac{1}{\lambda}\mathbf{C}(\lambda\mathbf{I}_{\mathrm{d}} + \mathbf{B})\mathbf{B}^{-1} \\ \frac{1}{\lambda}\mathbf{C}^{\top}\mathbf{C}(\lambda\mathbf{I}_{\mathrm{d}} + \mathbf{A})\mathbf{A}^{-1} & \mathbf{I}_{\mathrm{d}} \end{pmatrix} \begin{pmatrix} \mathbf{A} - \mathbf{A}\mathbf{X}^{-1}\mathbf{A} & \mathbf{A}\mathbf{X}^{-1}\mathbf{B} \\ \mathbf{B}\mathbf{X}^{-1}\mathbf{A} & \mathbf{B} - \mathbf{B}\mathbf{X}^{-1}\mathbf{B} \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{I}_{\mathrm{d}} & \frac{1}{\lambda}\mathbf{C}(\mathbf{X} - \mathbf{A})\mathbf{B}^{-1} \\ \frac{1}{\lambda}\mathbf{C}^{\top}(\mathbf{X} - \mathbf{B})\mathbf{A}^{-1} & \mathbf{I}_{\mathrm{d}} \end{pmatrix} \begin{pmatrix} \mathbf{A} - \mathbf{A}\mathbf{X}^{-1}\mathbf{A} & \mathbf{A}\mathbf{X}^{-1}\mathbf{B} \\ \mathbf{B}\mathbf{X}^{-1}\mathbf{A} & \mathbf{B} - \mathbf{B}\mathbf{X}^{-1}\mathbf{B} \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{A} - \mathbf{A}\mathbf{X}^{-1}\mathbf{A} + \frac{1}{\lambda}\mathbf{C}(\mathbf{A} - \mathbf{A}\mathbf{X}^{-1}\mathbf{A}) & \mathbf{A}\mathbf{X}^{-1}\mathbf{B} + \frac{1}{\lambda}\mathbf{C}(\mathbf{X} - \mathbf{A})(\mathbf{I}_{\mathrm{d}} - \mathbf{X}^{-1}\mathbf{B}) \\ \frac{1}{\lambda}\mathbf{C}^{\top}(\mathbf{X} - \mathbf{B})(\mathbf{I}_{\mathrm{d}} - \mathbf{X}^{-1}\mathbf{A}) + \mathbf{B}\mathbf{X}^{-1}\mathbf{A} & \frac{1}{\lambda}\mathbf{C}^{\top}(\mathbf{X} - \mathbf{B})\mathbf{X}^{-1}\mathbf{B} + \mathbf{B} - \mathbf{B}\mathbf{X}^{-1}\mathbf{B} \end{pmatrix} \\
&= \begin{pmatrix} (\mathbf{I}_{\mathrm{d}} + \frac{1}{\lambda}\mathbf{C})(\mathbf{A} - \mathbf{A}\mathbf{X}^{-1}\mathbf{A}) & \mathbf{A}\mathbf{X}^{-1}\mathbf{B} + \frac{1}{\lambda}\mathbf{C}(\mathbf{X} - \mathbf{A} - \mathbf{B} + \mathbf{A}\mathbf{X}^{-1}\mathbf{B}) \\ \lambda\mathbf{C}^{\top}(\lambda\mathbf{I}_{\mathrm{d}} + \mathbf{B}\mathbf{X}^{-1}\mathbf{A}) + \mathbf{B}\mathbf{X}^{-1}\mathbf{A} & \frac{1}{\lambda}\mathbf{C}^{\top}(\mathbf{X} - \mathbf{B})\mathbf{X}^{-1}\mathbf{B} + \mathbf{B} - \mathbf{B}\mathbf{X}^{-1}\mathbf{B} \end{pmatrix} \\
&= \begin{pmatrix} (\mathbf{I}_{\mathrm{d}} + \frac{1}{\lambda}\mathbf{C})(\mathbf{A} - \mathbf{A}\mathbf{X}^{-1}\mathbf{A}) & \mathbf{A}\mathbf{X}^{-1}\mathbf{B} + \frac{1}{\lambda}\mathbf{C}(\lambda\mathbf{I}_{\mathrm{d}} + \mathbf{A}\mathbf{X}^{-1}\mathbf{B}) \\ \mathbf{C}^{\top} + \frac{1}{\lambda}\mathbf{C}^{\top}\mathbf{B}\mathbf{X}^{-1}\mathbf{A} + \mathbf{B}\mathbf{X}^{-1}\mathbf{A} & (\mathbf{I}_{\mathrm{d}} + \frac{1}{\lambda}\mathbf{C}^{\top})(\mathbf{B} - \mathbf{B}\mathbf{X}^{-1}\mathbf{B}) \end{pmatrix} \\
&= \begin{pmatrix} (\mathbf{I}_{\mathrm{d}} + \frac{1}{\lambda}\mathbf{C})(\mathbf{A} - \mathbf{A}\mathbf{X}^{-1}\mathbf{A}) & \mathbf{C} + (\mathbf{I}_{\mathrm{d}} + \frac{1}{\lambda}\mathbf{C})\mathbf{A}\mathbf{X}^{-1}\mathbf{B} \\ \mathbf{C}^{\top} + (\mathbf{I}_{\mathrm{d}} + \frac{1}{\lambda}\mathbf{C}^{\top})\mathbf{B}\mathbf{X}^{-1}\mathbf{A} & (\mathbf{I}_{\mathrm{d}} + \frac{1}{\lambda}\mathbf{C}^{\top})(\mathbf{B} - \mathbf{B}\mathbf{X}^{-1}\mathbf{B}) \end{pmatrix}.
\end{aligned}$$

Finding the mass of the plan π: The optimal transport plan is given by

$$\frac{\mathrm{d}\pi}{\mathrm{d}x\mathrm{d}y}(x, y) = \omega e^{\frac{1}{2}\mu^{\top}\Gamma^{-1}\mu}\sqrt{\det(2\pi\mathbf{H})}\mathcal{N}(\mathbf{H}\mu, \mathbf{H})(z), \tag{4.90}$$

where

$$\begin{aligned}
\omega &= \frac{m_\alpha m_\beta}{\sqrt{\det(4\pi^2\mathbf{A}\mathbf{B})}} m_u m_v e^{-\frac{1}{2}(\mathbf{a}^{\top}\mathbf{A}^{-1}\mathbf{a} + \mathbf{b}^{\top}\mathbf{B}^{-1}\mathbf{b})} \\
&= \frac{m_\alpha m_\beta}{\sqrt{\det(4\pi^2\mathbf{A}\mathbf{B})}} \left(\frac{\sqrt{\det(\mathbf{A}\mathbf{B})\det(\mathbf{F}\mathbf{G})}}{\sigma^{2d} m_\alpha m_\beta}\right)^{\frac{\tau}{\tau+1}} e^{-\frac{\tau}{2(\tau+1)}(q_{u,\alpha} + q_{v,\beta})} e^{-\frac{1}{2}(\mathbf{a}^{\top}\mathbf{A}^{-1}\mathbf{a} + \mathbf{b}^{\top}\mathbf{B}^{-1}\mathbf{b})} \\
&= \frac{1}{(2\pi)^d}\left(\frac{m_\alpha m_\beta}{\sqrt{\det(\mathbf{A}\mathbf{B})}}\right)^{\frac{1}{\tau+1}} \left(\frac{\sqrt{\det(\mathbf{F}\mathbf{G})}}{\sigma^{2d}}\right)^{\frac{\tau}{\tau+1}} e^{-\frac{\tau}{2(\tau+1)}(q_{u,\alpha} + q_{v,\beta})} e^{-\frac{1}{2}(\mathbf{a}^{\top}\mathbf{A}^{-1}\mathbf{a} + \mathbf{b}^{\top}\mathbf{B}^{-1}\mathbf{b})}.
\end{aligned}$$

First, let us simplify the argument of the exponential terms. Isolating the terms that depend only on the input means $\mathbf{a}, \mathbf{b}$ it holds that

$$q_{u,\alpha} + q_{v,\beta} = \frac{\sigma^2}{\tau^2}(\mathbf{v}^{\top}\mathbf{F}\mathbf{v} + \mathbf{u}^{\top}\mathbf{G}\mathbf{u}) + \mathbf{a}^{\top}\mathbf{A}^{-1}\mathbf{a} + \mathbf{b}^{\top}\mathbf{B}^{-1}\mathbf{b}.$$

Therefore, the full exponential argument is given by

$$\phi \stackrel{\text{def}}{=} \mu^{\top}\Gamma^{-1}\mu - \frac{\tau}{\tau+1}\frac{\sigma^2}{\tau^2}(\mathbf{v}^{\top}\mathbf{F}\mathbf{v} + \mathbf{u}^{\top}\mathbf{G}\mathbf{u}) - \frac{1}{\tau+1}(\mathbf{a}^{\top}\mathbf{A}^{-1}\mathbf{a} + \mathbf{b}^{\top}\mathbf{B}^{-1}\mathbf{b}). \tag{4.91}$$

First, using (4.87) we may replace μ with its expression:

$$\begin{aligned}\mu^\top \Gamma^{-1}\mu &= \mu^\top \mathbf{H}\mu \\ &= \sigma^2 \begin{pmatrix}\mathbf{A}^{-1}\mathbf{a}\\ \mathbf{B}^{-1}\mathbf{b}\end{pmatrix}^\top \begin{pmatrix}\mathbf{I}_\mathrm{d} & \tau\mathbf{F}^{-1}\\ \tau\mathbf{G}^{-1} & \mathbf{I}_\mathrm{d}\end{pmatrix}^{-1} \begin{pmatrix}\mathbf{F} & -\mathbf{I}_\mathrm{d}\\ -\mathbf{I}_\mathrm{d} & \mathbf{G}\end{pmatrix}^{-1} \begin{pmatrix}\mathbf{I}_\mathrm{d} & \tau\mathbf{G}^{-1}\\ \tau\mathbf{F}^{-1} & \mathbf{I}_\mathrm{d}\end{pmatrix}^{-1} \begin{pmatrix}\mathbf{A}^{-1}\mathbf{a}\\ \mathbf{B}^{-1}\mathbf{b}\end{pmatrix}.\end{aligned}$$

Next, we have

$$\begin{aligned}\frac{\sigma^2}{\tau^2}(\mathbf{v}^\top\mathbf{F}\mathbf{v} + \mathbf{u}^\top\mathbf{G}\mathbf{u}) &= \sigma^2((\mathbf{A}^{-1}\mathbf{a}+\mathbf{u})^\top\mathbf{F}^{-1}(\mathbf{A}^{-1}\mathbf{a}+\mathbf{u}) + (\mathbf{B}^{-1}\mathbf{b}+\mathbf{v})^\top\mathbf{G}^{-1}(\mathbf{B}^{-1}\mathbf{b}+\mathbf{v})) \\ &= \sigma^2\mu^\top \begin{pmatrix}\mathbf{F}^{-1} & 0\\ 0 & \mathbf{G}^{-1}\end{pmatrix}\mu \\ &= \sigma^2 \begin{pmatrix}\mathbf{A}^{-1}\mathbf{a}\\ \mathbf{B}^{-1}\mathbf{b}\end{pmatrix}^\top \begin{pmatrix}\mathbf{I}_\mathrm{d} & \tau\mathbf{F}^{-1}\\ \tau\mathbf{G}^{-1} & \mathbf{I}_\mathrm{d}\end{pmatrix}^{-1} \begin{pmatrix}\mathbf{F}^{-1} & 0\\ 0 & \mathbf{G}^{-1}\end{pmatrix} \begin{pmatrix}\mathbf{I}_\mathrm{d} & \tau\mathbf{G}^{-1}\\ \tau\mathbf{F}^{-1} & \mathbf{I}_\mathrm{d}\end{pmatrix}^{-1} \begin{pmatrix}\mathbf{A}^{-1}\mathbf{a}\\ \mathbf{B}^{-1}\mathbf{b}\end{pmatrix}.\end{aligned}$$

Let $\mathbf{J} = \begin{pmatrix}\mathbf{I}_\mathrm{d} & \tau\mathbf{G}^{-1}\\ \tau\mathbf{F}^{-1} & \mathbf{I}_\mathrm{d}\end{pmatrix}$ and $\mathbf{K} = \begin{pmatrix}\mathbf{F} & 0\\ 0 & \mathbf{G}\end{pmatrix}$. It holds that

$$\mu^\top\Gamma^{-1}\mu - \frac{\tau}{\tau+1}\frac{\sigma^2}{\tau^2}(\mathbf{v}^\top\mathbf{F}\mathbf{v}+\mathbf{u}^\top\mathbf{G}\mathbf{u}) = \begin{pmatrix}\mathbf{A}^{-1}\mathbf{a}\\ \mathbf{B}^{-1}\mathbf{b}\end{pmatrix}^\top \mathbf{J}^{\top-1}(\mathbf{H} - \frac{\sigma^2\tau}{\tau+1}\mathbf{K}^{-1})\mathbf{J}^{-1}\begin{pmatrix}\mathbf{A}^{-1}\mathbf{a}\\ \mathbf{B}^{-1}\mathbf{b}\end{pmatrix}.$$

Let us compute the matrix $\mathbf{J}^{\top-1}(\mathbf{H} - \frac{\tau\sigma^2}{\tau+1}\mathbf{K}^{-1})\mathbf{J}^{-1}$. First keep in mind that $\mathbf{J}\mathbf{K} = \begin{pmatrix}\mathbf{F} & \tau\mathbf{I}_\mathrm{d}\\ \tau\mathbf{I}_\mathrm{d} & \mathbf{G}\end{pmatrix}$. Now, using Woodburry's identity:

$$\begin{aligned}\left(\mathbf{J}^{\top-1}(\mathbf{H} - \frac{\tau}{\tau+1}\mathbf{K}^{-1})\mathbf{J}^{-1}\right)^{-1} &= \mathbf{J}(\mathbf{H} - \frac{\tau\sigma^2}{\tau+1}\mathbf{K}^{-1})^{-1}\mathbf{J}^\top \\ &= \mathbf{J}\left(-\frac{\tau+1}{\tau\sigma^2}\mathbf{K} - \left(\frac{\tau+1}{\tau\sigma^2}\right)^2\mathbf{K}(\mathbf{H}^{-1} - \frac{\tau+1}{\tau\sigma^2}\mathbf{K})^{-1}\mathbf{K}\right)\mathbf{J}^\top \\ &= \frac{\tau+1}{\tau\sigma^2}\left(-\mathbf{J}\mathbf{K}\mathbf{J}^\top - \frac{\tau+1}{\tau\sigma^2}\mathbf{J}\mathbf{K}(\begin{pmatrix}-\frac{\mathbf{F}}{\tau\sigma^2} & -\frac{1}{\sigma^2}\mathbf{I}_\mathrm{d}\\ -\frac{1}{\sigma^2}\mathbf{I}_\mathrm{d} & -\frac{\mathbf{G}}{\tau\sigma^2}\end{pmatrix}^{-1}(\mathbf{J}\mathbf{K}^\top)^\top\right) \\ &= \frac{\tau+1}{\tau\sigma^2}\left(-\mathbf{J}\mathbf{K}\mathbf{J}^\top + (\tau+1)\mathbf{J}\mathbf{K}(\begin{pmatrix}\mathbf{F} & \tau\mathbf{I}_\mathrm{d}\\ \tau\mathbf{I}_\mathrm{d} & \mathbf{G}\end{pmatrix}^{-1}(\mathbf{J}\mathbf{K}^\top)^\top\right) \\ &= \frac{\tau+1}{\tau\sigma^2}\left(-\begin{pmatrix}\mathbf{F} & \tau\mathbf{I}_\mathrm{d}\\ \tau\mathbf{I}_\mathrm{d} & \mathbf{G}\end{pmatrix}\begin{pmatrix}\mathbf{I}_\mathrm{d} & \tau\mathbf{F}^{-1}\\ \tau\mathbf{G}^{-1} & \mathbf{I}_\mathrm{d}\end{pmatrix} + (\tau+1)\begin{pmatrix}\mathbf{F} & \tau\mathbf{I}_\mathrm{d}\\ \tau\mathbf{I}_\mathrm{d} & \mathbf{G}\end{pmatrix}\right) \\ &= \frac{\tau+1}{\tau\sigma^2}\begin{pmatrix}-\mathbf{F} - \tau^2\mathbf{G}^{-1} + (\tau+1)\mathbf{F} & (-2\tau+\tau(\tau+1))\mathbf{I}_\mathrm{d}\\ (-2\tau+\tau(\tau+1))\mathbf{I}_\mathrm{d} & -\mathbf{G} - \tau^2\mathbf{F}^{-1} + (\tau+1)\mathbf{G}\end{pmatrix} \\ &= \frac{\tau+1}{\sigma^2}\begin{pmatrix}\mathbf{F} - \tau\mathbf{G}^{-1} & -(1-\tau)\mathbf{I}_\mathrm{d}\\ -(1-\tau)\mathbf{I}_\mathrm{d} & \mathbf{G} - \tau\mathbf{F}^{-1}\end{pmatrix} \\ &= (\tau+1)\begin{pmatrix}\mathbf{A}^{-1} + \frac{1}{\lambda}\mathbf{I}_\mathrm{d} & -\frac{1}{\lambda}\mathbf{I}_\mathrm{d}\\ -\frac{1}{\lambda}\mathbf{I}_\mathrm{d} & \mathbf{B}^{-1} + \frac{1}{\lambda}\mathbf{I}_\mathrm{d}\end{pmatrix} \\ &= (\tau+1)\mathbf{Z}.\end{aligned}$$

Therefore,

$$\mu^\top\Gamma^{-1}\mu - \frac{\tau}{\tau+1}\frac{\sigma^2}{\tau^2}(\mathbf{v}^\top\mathbf{F}\mathbf{v}+\mathbf{u}^\top\mathbf{G}\mathbf{u}) = \frac{1}{\tau+1}\begin{pmatrix}\mathbf{A}^{-1}\mathbf{a}\\ \mathbf{B}^{-1}\mathbf{b}\end{pmatrix}^\top \mathbf{Z}^{-1}\begin{pmatrix}\mathbf{A}^{-1}\mathbf{a}\\ \mathbf{B}^{-1}\mathbf{b}\end{pmatrix}. \tag{4.92}$$

The full exponential argument ϕ defined in Equation (4.91) is given by

$$\begin{aligned}
\phi &= \frac{1}{\tau+1}\left(\begin{pmatrix}\mathbf{A}^{-1}\mathbf{a}\\ \mathbf{B}^{-1}\mathbf{b}\end{pmatrix}^\top \mathbf{Z}^{-1}\begin{pmatrix}\mathbf{A}^{-1}\mathbf{a}\\ \mathbf{B}^{-1}\mathbf{b}\end{pmatrix} - \mathbf{a}^\top\mathbf{A}^{-1}\mathbf{a} - \mathbf{b}^\top\mathbf{B}^{-1}\mathbf{b}\right)\\
&= \frac{1}{\tau+1}\begin{pmatrix}\mathbf{a}\\ \mathbf{b}\end{pmatrix}^\top\begin{pmatrix}\mathbf{A}^{-1} & 0\\ 0 & \mathbf{B}^{-1}\end{pmatrix}\left(\mathbf{Z}^{-1} - \begin{pmatrix}\mathbf{A} & 0\\ 0 & \mathbf{B}\end{pmatrix}\right)\begin{pmatrix}\mathbf{A}^{-1} & 0\\ 0 & \mathbf{B}^{-1}\end{pmatrix}\begin{pmatrix}\mathbf{a}\\ \mathbf{b}\end{pmatrix}\\
&= \frac{1}{\tau+1}\begin{pmatrix}\mathbf{a}\\ \mathbf{b}\end{pmatrix}^\top\begin{pmatrix}\mathbf{A}^{-1} & 0\\ 0 & \mathbf{B}^{-1}\end{pmatrix}\begin{pmatrix}-\mathbf{A}\mathbf{X}^{-1}\mathbf{A} & \mathbf{A}\mathbf{X}^{-1}\mathbf{B}\\ \mathbf{B}\mathbf{X}^{-1}\mathbf{A} & -\mathbf{B}\mathbf{X}^{-1}\mathbf{B}\end{pmatrix}\begin{pmatrix}\mathbf{A}^{-1} & 0\\ 0 & \mathbf{B}^{-1}\end{pmatrix}\begin{pmatrix}\mathbf{a}\\ \mathbf{b}\end{pmatrix}\\
&= \frac{1}{\tau+1}\begin{pmatrix}\mathbf{a}\\ \mathbf{b}\end{pmatrix}^\top\begin{pmatrix}-\mathbf{X}^{-1} & \mathbf{X}^{-1}\\ \mathbf{X}^{-1} & -\mathbf{X}^{-1}\end{pmatrix}\begin{pmatrix}\mathbf{a}\\ \mathbf{b}\end{pmatrix}\\
&= -\frac{1}{\tau+1}(\mathbf{a}-\mathbf{b})^\top\mathbf{X}^{-1}(\mathbf{a}-\mathbf{b})\\
&= \frac{1}{\tau+1}\|\mathbf{a}-\mathbf{b}\|^2_{\mathbf{X}^{-1}}.
\end{aligned}$$

Substituting in (4.90) leads to

$$\begin{aligned}
m_\pi &\stackrel{\text{def}}{=} \pi(\mathbb{R}^d\times\mathbb{R}^d)\\
&= \sqrt{\det(\mathbf{H})}\left(\frac{m_\alpha m_\beta}{\sqrt{\det(\mathbf{A}\mathbf{B})}}\right)^{\frac{1}{\tau+1}}\left(\frac{\sqrt{\det(\mathbf{F}\mathbf{G})}}{\sigma^{2d}}\right)^{\frac{\tau}{\tau+1}} e^{-\frac{1}{2(\tau+1)}(\|\mathbf{a}-\mathbf{b}\|^2_{\mathbf{X}^{-1}})}.
\end{aligned}$$

The determinants can be easily expressed as functions of $\mathbf{C}$. First notice that

$$\det(\mathbf{H}) = \frac{1}{\det(\Gamma)} = \frac{\sigma^{4d}}{\det(\mathbf{F}\mathbf{G}-\mathbf{I}_{\mathrm{d}})},$$

and using the definition of $\mathbf{C}$, it holds that

$$\mathbf{F}\mathbf{G} = \widetilde{\mathbf{B}}\mathbf{C}^{-2}\widetilde{\mathbf{A}}.$$

Therefore, $\det(\mathbf{F}\mathbf{G}) = \frac{\det(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}})}{\det(\mathbf{C})^2}$. Keeping in mind that the closed-form expression of $\mathbf{C}$ given in (4.82) is applied to the pair $(\frac{1}{\tau}\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}})$ in the unbalanced case, it holds that $\mathbf{C}^2 + \sigma^2\mathbf{C} = \frac{1}{\tau}\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}$. Hence,

$$\begin{aligned}
\mathbf{F}\mathbf{G}-\mathbf{I}_{\mathrm{d}} &= \widetilde{\mathbf{B}}\mathbf{C}^{-2}\widetilde{\mathbf{A}}(\mathbf{I}_{\mathrm{d}} - \widetilde{\mathbf{A}}^{-1}\mathbf{C}^2\widetilde{\mathbf{B}}^{-1})\\
&= \widetilde{\mathbf{B}}\mathbf{C}^{-2}\widetilde{\mathbf{A}}(\mathbf{I}_{\mathrm{d}} - \widetilde{\mathbf{A}}^{-1}(\frac{1}{\tau}\widetilde{\mathbf{A}}\widetilde{\mathbf{B}} - \sigma^2\mathbf{C})\widetilde{\mathbf{B}}^{-1})\\
&= \widetilde{\mathbf{B}}\mathbf{C}^{-2}\widetilde{\mathbf{A}}(\frac{(1-\tau)}{\tau}\mathbf{I}_{\mathrm{d}} + \sigma^2\widetilde{\mathbf{A}}^{-1}\mathbf{C}\widetilde{\mathbf{B}}^{-1})\\
&= \sigma^2\widetilde{\mathbf{B}}\mathbf{C}^{-2}\widetilde{\mathbf{A}}(-\frac{2}{\gamma}\mathbf{I}_{\mathrm{d}} + \widetilde{\mathbf{A}}^{-1}\mathbf{C}\widetilde{\mathbf{B}}^{-1})\\
&= \sigma^2\widetilde{\mathbf{B}}\mathbf{C}^{-2}(-\frac{2}{\gamma}\widetilde{\mathbf{A}}\widetilde{\mathbf{B}} + \mathbf{C})\widetilde{\mathbf{B}}^{-1},
\end{aligned}$$

and therefore

$$\det(\mathbf{F}\mathbf{G}-\mathbf{I}_{\mathrm{d}}) = \sigma^{2d}\frac{\det((-\frac{2}{\gamma}\widetilde{\mathbf{A}}\widetilde{\mathbf{B}} + \mathbf{C})}{\det(\mathbf{C})^2}.$$

Replacing the determinant formulas of $\mathbf{FG}$ and $\mathbf{FG} - \mathbf{I}_{\mathrm{d}}$ and re-arranging the common terms $\det(\mathbf{C})$ and σ finally leads to

$$\begin{aligned}
\pi(\mathbb{R}^d \times \mathbb{R}^d) &= \frac{\left(m_\alpha m_\beta \sigma^{2d} \det(\mathbf{C}) \sqrt{\frac{\det(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}})^\tau}{\det(\mathbf{AB})}}\right)^{\frac{1}{\tau+1}}}{\sqrt{\frac{\det(\mathbf{C}-\frac{2}{\gamma}\widetilde{\mathbf{A}}\widetilde{\mathbf{B}})}{\sigma^{2d}}}} e^{-\frac{1}{2(\tau+1)}(\|\mathbf{a}-\mathbf{b}\|^2_{\mathbf{X}^{-1}})} \\
&= \sigma^{d(\frac{2}{\tau+1}-1)} \frac{\left(m_\alpha m_\beta \det(\mathbf{C}) \sqrt{\frac{\det(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}})^\tau}{\det(\mathbf{AB})}}\right)^{\frac{1}{\tau+1}}}{\sqrt{\det(\mathbf{C}-\frac{2}{\gamma}\widetilde{\mathbf{A}}\widetilde{\mathbf{B}})}} e^{-\frac{1}{2(\tau+1)}(\|\mathbf{a}-\mathbf{b}\|^2_{\mathbf{X}^{-1}})} \\
&= \sigma^{d\frac{1-\tau}{\tau+1}} \frac{\left(m_\alpha m_\beta \det(\mathbf{C}) \sqrt{\frac{\det(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}})^\tau}{\det(\mathbf{AB})}}\right)^{\frac{1}{\tau+1}}}{\sqrt{\det(\mathbf{C}-\frac{2}{\gamma}\widetilde{\mathbf{A}}\widetilde{\mathbf{B}})}} e^{-\frac{1}{2(\tau+1)}(\|\mathbf{a}-\mathbf{b}\|^2_{\mathbf{X}^{-1}})} \\
&= \sigma^{\frac{d\sigma^2}{\sigma^2+\gamma}} \frac{\left(m_\alpha m_\beta \det(\mathbf{C}) \sqrt{\frac{\det(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}})^\tau}{\det(\mathbf{AB})}}\right)^{\frac{1}{\tau+1}}}{\sqrt{\det(\mathbf{C}-\frac{2}{\gamma}\widetilde{\mathbf{A}}\widetilde{\mathbf{B}})}} e^{-\frac{1}{2(\tau+1)}(\|\mathbf{a}-\mathbf{b}\|^2_{\mathbf{X}^{-1}})}.
\end{aligned} \tag{4.93}$$

Deriving a closed form for UOT_σ**.** Using Equation (4.93), a direct application of Proposition 4.13 yields

$$\mathrm{UOT}_\sigma(\alpha, \beta) = \gamma(m_\alpha + m_\beta) + 2\sigma^2(m_\alpha m_\beta) - 2(\sigma^2 + 2\gamma)m_{\pi^\star}. \tag{4.94}$$

This ends the proof of Theorem 4.14.

Chapter 5

Missing Data Imputation using Optimal Transport

Missing data is a crucial issue when applying machine learning algorithms to real-world datasets. Starting from the simple assumption that two batches extracted randomly from the same dataset should share the same distribution, we leverage optimal transport distances to quantify that criterion and turn it into a loss function to impute missing data values. We propose practical methods to minimize these losses using end-to-end learning, that can exploit or not parametric assumptions on the underlying distributions of values. We evaluate our methods on datasets from the UCI repository, in MCAR, MAR and MNAR settings. These experiments show that OT-based methods match or out-perform state-of-the-art imputation methods, even for high percentages of missing values.

This chapter is based on [Muzellec et al., 2020].

1 Introduction

Data collection is usually a messy process, resulting in datasets that have many missing values. This has been an issue for as long as data scientists have prepared, curated and obtained data, and is all the more inevitable given the vast amounts of data currently collected. The literature on the subject is therefore abundant [Little and Rubin, 2002, van Buuren, 2018]: a recent survey indicates that there are more than 150 implementations available to handle missing data [Mayer et al., 2019]. These methods differ on the objectives of their analysis (estimation of parameters and their variance, matrix completion, prediction), the nature of the variables considered (categorical, mixed, etc.), the assumptions about the data, and the missing data mechanisms. Imputation methods, which consist in filling missing entries with plausible values, are very appealing as they allow to both get a guess for the missing entries as well as to perform (with care) downstream machine learning methods on the completed data. Efficient methods include, among others, methods based on low-rank assumptions [Hastie et al., 2015], iterative random forests [Stekhoven and Buhlmann, 2011] and imputation using variational autoencoders [Mattei and Frellsen, 2019, Ivanov et al., 2019]. A desirable property for imputation methods is that they should preserve the joint and marginal distributions of the data. Non-parametric Bayesian strategies [Murray and Reiter, 2016] or recent approaches based on generative adversarial networks [Yoon et al., 2018] are attempts in this direction. However, they can be quite cumbersome to implement in practice.

We argue in this work that the optimal transport (OT) toolbox constitutes a natural, sound and straightforward alternative. Indeed, optimal transport provides geometrically meaningful distances to compare discrete distributions, and therefore data. Furthermore, thanks to recent computational advances grounded on regularization [Cuturi, 2013], OT-based divergences can be computed in a scalable and differentiable way [Peyré et al., 2019]. Those advances have allowed to successfully use OT as a loss function in many applications, including multi-label classification [Frogner et al., 2015], inference of pathways Schiebinger et al. [2019] and generative modeling [Arjovsky et al., 2017, Genevay et al., 2018, Salimans et al., 2018]. Considering the similarities between generative modeling and missing data imputation, it is therefore quite natural to use OT as a loss for the latter.

Contributions. This chapter presents two main contributions. First, we leverage OT to define a loss function for missing value imputation. This loss function is the mathematical translation of the simple intuition that two random batches from the same dataset should follow the same distribution. Next, we provide algorithms for imputing missing values according to this loss. Two types of algorithms are presented, the first (i) being non-parametric, and the second (ii) defining a class of parametric models. The non-parametric algorithm (i) enjoys the most degrees of freedom, and can therefore output imputations which respect the global shape of the data while taking into account its local features. The parametric algorithm (ii) is trained in a round-robin fashion similar to iterative conditional imputation techniques, as implemented for instance in the `mice` package van Buuren and Groothuis-Oudshoorn [2011]. Compared to the non-parametric method, this algorithm allows to perform out-of-sample imputation. This creates a very flexible framework which can be combined with many imputing strategies, including imputation with Multi-Layer Perceptrons. Finally, these methods are showcased in extensive experiments on a variety of datasets and for different missing values proportions and mechanisms, including the difficult case of informative missing entries. The code to reproduce these experiments is available at `https://github.com/BorisMuzellec/MissingDataOT`.

Notations. Let $\mathbf{\Omega} = (\omega_{ij})_{ij} \in \{0,1\}^{n \times d}$ be a binary mask encoding observed entries, i.e. $\omega_{ij} = 1$ (resp. 0) *iff* the entry (i,j) is observed (resp. missing). We observe the following

incomplete data matrix:

$$\mathbf{X} = \mathbf{X}^{(obs)} \odot \mathbf{\Omega} + \text{NA} \odot (\mathbb{1}_{n\times d} - \mathbf{\Omega}),$$

where $\mathbf{X}^{(obs)} \in \mathbb{R}^{n\times d}$ contains the observed entries, $\odot$ is the elementwise product and $\mathbb{1}_{n\times d}$ is an $n \times d$ matrix filled with ones. Given the data matrix $\mathbf{X}$, our goal is to construct an estimate $\hat{\mathbf{X}}$ filling the missing entries of $\mathbf{X}$, which can be written as

$$\hat{\mathbf{X}} = \mathbf{X}^{(obs)} \odot \mathbf{\Omega} + \hat{\mathbf{X}}^{(imp)} \odot (\mathbb{1}_{n\times d} - \mathbf{\Omega}),$$

where $\hat{\mathbf{X}}^{(imp)} \in \mathbb{R}^{n\times d}$ contains the imputed values. Let $\mathbf{x}_{i:}$ denote the i-th row of the data set $\mathbf{X}$, such that $\mathbf{X} = (\mathbf{x}_{i:}^T)_{1\leq i\leq n}$. Similarly, $\mathbf{x}_{:j}$ denotes the j-th column (variable) of the data set $\mathbf{X}$, such that $\mathbf{X} = (\mathbf{x}_{:1}|\dots|\mathbf{x}_{:d})$, and $\mathbf{X}_{:-j}$ denotes the dataset $\mathbf{X}$ in which the j-th variable has been removed. For $K \subset \{1,\dots,n\}$ a set of m indices, $\mathbf{X}_K = (\mathbf{x}_{k:})_{k\in K}$ denotes the corresponding batch, and by $\mu_m(\mathbf{X}_K)$ the empirical measure associated to $\mathbf{X}_K$, i.e.

$$\mu_m(\mathbf{X}_K) := \tfrac{1}{m}\sum_{k\in K} \delta_{\mathbf{x}_{k:}}.$$

Finally, $\Delta_n \overset{\text{def}}{=} \{\mathbf{a} \in \mathbb{R}_+^n : \sum_{i=1}^n a_i = 1\}$ is the simplex in dimension n.

2 Background

2.1 Missing data

Rubin [1976] defined a widely used - yet controversial [Seaman et al., 2013] - nomenclature for missing values mechanisms. This nomenclature distinguishes between three cases: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). In MCAR, the missingness is independent of the data, whereas in MAR, the probability of being missing depends only on observed values. A subsequent part of the literature, with notable exceptions [Kim and Ying, 2018, Mohan and Pearl, 2019], only consider these "simple" mechanisms and struggles for the harder yet prevalent MNAR case. MNAR values lead to important biases in the data, as the probability of missingness then depends on the unobserved values. On the other hand, MCAR and MAR are "ignorable" mechanisms in the sense that they do not make it necessary to model explicitly the distribution of missing values when maximizing the observed likelihood.

The naive workaround which consists in deleting observations with missing entries is not an alternative in high dimension. Indeed, let us assume as in Zhu et al. [2019] that $\mathbf{X}$ is a $n \times d$ data matrix in which each entry is missing independently with probability 0.01. When $d = 5$, this would result in around 95% of the individuals (rows) being retained, but for $d = 300$, only around 5% of rows have no missing entries. Hence, providing plausible imputations for missing values quickly becomes necessary. Classical imputation methods impute according to a joint distribution which is either explicit, or implicitly defined through a set of conditional distributions. As an example, explicit joint modeling methods include imputation models that assume a Gaussian distribution for the data, whose parameters are estimated using EM algorithms [Dempster et al., 1977]. Missing values are then imputed by drawing from their predictive distribution. A second instance of such joint modeling methods are imputations assuming low-rank structure [Josse et al., 2016]. The conditional modeling approach [van Buuren, 2018], also known as "sequential imputation" or "imputation using chained equations" (ice) consists in specifying one model for each variable. It predicts the missing values of each variable using the other variables as explanatory, and cycles through the variables iterating this procedure to update the imputations until predictions stabilize.

Non-parametric methods like k-nearest neighbors imputation [Troyanskaya et al., 2001] or random forest imputation [Stekhoven and Buhlmann, 2011] have also been developed and account for the local geometry of the data. The herein proposed methods lie at the intersection of global and local approaches and are derived in a non-parametric and parametric version.

2.2 Reminders on Wasserstein distances, entropic regularization and Sinkhorn divergences

Let $\alpha = \sum_{i=1}^{n} a_i \delta_{\mathbf{x}_i}$, $\beta = \sum_{i=1}^{n'} b_i \delta_{\mathbf{y}_i}$ be two discrete distributions, described by their supports $(\mathbf{x}_i)_{i=1}^{n} \in \mathbb{R}^{n \times p}$ and $(\mathbf{y}_i)_{i=1}^{n'} \in \mathbb{R}^{n' \times p}$ and weight vectors $\mathbf{a} \in \Delta_n$ and $\mathbf{b} \in \Delta_{n'}$. Optimal transport compares α and β by considering the most efficient way of transporting the masses $\mathbf{a}$ and $\mathbf{b}$ onto each-other, according to a ground cost between the supports. The (2-)Wasserstein distance corresponds to the case where this ground cost is quadratic:

$$W_2^2(\alpha, \beta) \stackrel{\text{def}}{=} \min_{\mathbf{P} \in U(\mathbf{a}, \mathbf{b})} \langle \mathbf{P}, \mathbf{M} \rangle, \tag{5.1}$$

where $U(\mathbf{a}, \mathbf{b}) \stackrel{\text{def}}{=} \{\mathbf{P} \in \mathbb{R}^{n \times n'} : \mathbf{P}\mathbb{1}_{n'} = \mathbf{a}, \mathbf{P}^T \mathbb{1}_n = \mathbf{b}\}$ is the set of transportation plans, and $\mathbf{M} = \left(\|x_i - y_j\|^2\right)_{ij} \in \mathbb{R}^{n \times n'}$ is the matrix of pairwise squared distances between the supports. W_2 is not differentiable and requires solving a costly linear program via network simplex methods [Peyré et al., 2019, §3]. Entropic regularization alleviates both issues: consider

$$\mathrm{OT}_\varepsilon(\alpha, \beta) \stackrel{\text{def}}{=} \min_{\mathbf{P} \in U(\mathbf{a}, \mathbf{b})} \langle \mathbf{P}, \mathbf{M} \rangle + \varepsilon h(\mathbf{P}), \tag{5.2}$$

where $\varepsilon > 0$ and $h(\mathbf{P}) \stackrel{\text{def}}{=} \sum_{ij} p_{ij} \log p_{ij}$ is the negative entropy. Then, $\mathrm{OT}_\varepsilon(\alpha, \beta)$ is differentiable and can be solved using Sinkhorn iterations [Cuturi, 2013]. However, due to the entropy term, OT_ε is no longer positive. This issue is solved through debiasing, by subtracting auto-correlation terms. Let

$$S_\varepsilon(\alpha, \beta) \stackrel{\text{def}}{=} \mathrm{OT}_\varepsilon(\alpha, \beta) - \frac{1}{2}(\mathrm{OT}_\varepsilon(\alpha, \alpha) + \mathrm{OT}_\varepsilon(\beta, \beta)). \tag{5.3}$$

Equation (5.3) defines the Sinkhorn divergences [Genevay et al., 2018], which are positive, convex, and can be computed with little additional cost compared to entropic OT [Feydy et al., 2019]. Sinkhorn divergences hence provide a differentiable and tractable proxy for Wasserstein distances, and will be used in the following.

OT gradient-based methods. Not only are the OT metrics described above good measures of distributional closeness, they are also well-adapted to gradient-based imputation methods. Indeed, let $\mathbf{X}_K, \mathbf{X}_L$ be two batches drawn from $\mathbf{X}$. Then, gradient updates for $\mathrm{OT}_\varepsilon(\mu_m(\mathbf{X}_K), \mu_m(\mathbf{X}_L)), \varepsilon \geq 0$ w.r.t a point $\mathbf{x}_{k:}$ in $\mathbf{X}_K$ correspond to taking steps along the so-called barycentric transport map. Indeed, with (half) quadratic costs, it holds [Cuturi and Doucet, 2014, §4.3] that

$$\nabla_{\mathbf{x}_{k:}} \mathrm{OT}_\varepsilon(\mu_m(\mathbf{X}_K), \mu_m(\mathbf{X}_L)) = m \sum_{\ell} \mathbf{P}^\star_{k\ell}(\mathbf{x}_{k:} - \mathbf{x}_{\ell:}),$$

where $\mathbf{P}^\star$ is the optimal (regularized) transport plan. Therefore, a gradient based-update is of the form

$$\mathbf{x}_{k:} \leftarrow (1 - t)\mathbf{x}_{k:} + tm \sum_{l} \mathbf{P}^\star_{kl} \mathbf{x}_{l:}. \tag{5.4}$$

In a missing value imputation context, Equation (5.4) thus corresponds to updating values to make them closer to the target points given by transportation plans. Building on this fact, OT gradient-based imputation methods are proposed in the next section.

3 Imputing Missing Values using OT

Let $\mathbf{X}_K$ and $\mathbf{X}_L$ be two batches respectively extracted from the complete rows and the incomplete rows in $\mathbf{X}$, such that only the batch $\mathbf{X}_L$ contains missing values. In this one-sided incomplete batch setting, a good imputation should preserve the distribution from the complete batch, meaning that $\mathbf{X}_K$ should be close to $\mathbf{X}_L$ in terms of distributions. The OT-based metrics described in Section 2 provide natural criteria to catch this distributional proximity and derive imputation methods. However, as observed in Section 2, in high dimension or with a high proportion of missing values, it is unlikely or even impossible to obtain batches from $\mathbf{X}$ with no missing values. Nonetheless, a good imputation method should still ensure that the distributions of any two i.i.d. incomplete batches $\mathbf{X}_K$ and $\mathbf{X}_L$, *both* containing missing values, should be close. This implies in particular that OT-metrics between the distributions $\mu_m(\mathbf{X}_K)$ and $\mu_m(\mathbf{X}_L)$ should have small values. This criterion, which is weaker than the one above with one-sided missing data but is more amenable, will be considered from now on.

Direct imputation. Algorithm 5 is a direct implementation of this criterion, aiming to impute missing values for quantitative variables by minimizing OT distances between batches. First, missing values of any variable are initialized with the mean of observed entries plus a small amount of noise (to preserve the marginals and to facilitate the optimization). Then, batches are sequentially sampled and the Sinkhorn divergence between batches is minimized with respect to the imputed values, using gradient updates (here using RMSprop [Tieleman and Hinton, 2015]).

Algorithm 5 Batch Sinkhorn Imputation

Input: $\mathbf{X} \in (\mathbb{R} \cup \mathrm{NA})^{n \times d}$, $\mathbf{\Omega} \in \{0,1\}^{n \times d}$, $\alpha, \eta, \varepsilon > 0$, $n \geq m > 0$,
Initialization: for $j = 1, \ldots, d$,

- for i s.t. $\omega_{ij} = 0$, $\hat{x}_{ij} \leftarrow \overline{\mathbf{x}_{:j}^{obs}} + \varepsilon_{ij}$, with $\varepsilon_{ij} \sim \mathcal{N}(0, \eta)$ and $\overline{\mathbf{x}_{:j}^{obs}}$ corresponds to the mean of the observed entries in the j-th variable (missing entries)
- for i s.t. $\omega_{ij} = 1$, $\hat{x}_{ij} \leftarrow x_{ij}$ (observed entries)

for t = 1, 2, ...,t_{max} **do**
Sample two sets K and L of m indices
$\mathcal{L}(\hat{\mathbf{X}}_K, \hat{\mathbf{X}}_L) \leftarrow S_\varepsilon(\mu_m(\hat{\mathbf{X}}_K), \mu_m(\hat{\mathbf{X}}_L))$
$\hat{\mathbf{X}}_{K \cup L}^{(imp)} \leftarrow \hat{\mathbf{X}}_{K \cup L}^{(imp)} - \alpha \mathrm{RMSprop}(\nabla_{\hat{\mathbf{X}}_{K \cup L}^{(imp)}} \mathcal{L})$
end for
Output: $\hat{\mathbf{X}}$

OT as a loss for missing data imputation. Taking a step back, one can see that Algorithm 5 essentially uses Sinkhorn divergences between batches as a loss function to impute values for a model in which "one parameter equals one imputed value". Formally, for a fixed batch size m, this loss is defined as

$$\mathcal{L}_m(\mathbf{X}) \stackrel{\text{def}}{=} \sum_{\substack{K: 0 \leq k_1 < \ldots < k_m \leq n \\ L: 0 \leq \ell_1 < \ldots < \ell_m \leq n}} S_\varepsilon(\mu_m(\mathbf{X}_K), \mu_m(\mathbf{X}_L)). \tag{5.5}$$

Equation (5.5) corresponds to the "autocorrelation" counterpart of the minibatch Wasserstein distances described in Fatras et al. [2019], Salimans et al. [2018].

Although Algorithm 5 is straightforward, a downside is that it cannot directly generate imputations for out-of-sample data points with missing values. Hence, a natural extension is

to use the loss defined in Equation (5.5) to fit parametric imputation models, provided they are differentiable with respect to their parameters. At a high level, this method is described by Algorithm 6. Algorithm 6 takes as an input an imputer model with a parameter Θ

Algorithm 6 Meta Sinkhorn Imputation

Input: $\mathbf{X} \in \mathbb{R}^{n \times d}$, $\mathbf{\Omega} \in \{0,1\}^{n \times d}$, Imputer$(\cdot, \cdot, \cdot)$, Θ_0, $\varepsilon > 0$, $n \geq m > 0$,
  $\hat{\mathbf{X}}^0 \leftarrow$ same initialization as in Algorithm 5
  $\hat{\Theta} \leftarrow \Theta_0$
  for $t = 1, 2, ..., t_{\max}$ **do**
    for $k = 1, 2, ..., K$ **do**
      $\hat{\mathbf{X}} \leftarrow \text{Imputer}(\hat{\mathbf{X}}^t, \Omega, \hat{\Theta})$
      Sample two sets K and L of m indices
      $\mathcal{L}(\hat{\mathbf{X}}_K, \hat{\mathbf{X}}_L) \leftarrow S_\varepsilon(\mu_m(\hat{\mathbf{X}}_K), \mu_m(\hat{\mathbf{X}}_L))$
      $\nabla_\Theta \mathcal{L} \leftarrow \text{AutoDiff}(\mathcal{L}(\hat{\mathbf{X}}_K, \hat{\mathbf{X}}_L))$
      $\hat{\Theta} \leftarrow \hat{\Theta} - \alpha \text{Adam}(\nabla_\Theta \mathcal{L})$
    end for
    $\hat{\mathbf{X}}^{t+1} \leftarrow \text{Imputer}(\hat{\mathbf{X}}^t, \Omega, \hat{\Theta})$
  end for
Output: Completed data $\hat{\mathbf{X}} = \hat{\mathbf{X}}^{t_{\max}}$, Imputer$(\cdot, \cdot, \hat{\Theta})$

such that Imputer$(\mathbf{X}, \Omega, \Theta)$ returns imputations for the missing values in $\mathbf{X}$. This imputer has to be differentiable w.r.t. its parameter Θ, so that the batch Sinkhorn loss $\mathcal{L}$ can be back-propagated through $\hat{\mathbf{X}}$ to perform gradient-based updates of Θ. Algorithm 6 does not only return the completed data matrix $\hat{\mathbf{X}}$, but also the trained parameter $\hat{\Theta}$, which can then be re-used to impute missing values in out-of-sample data.

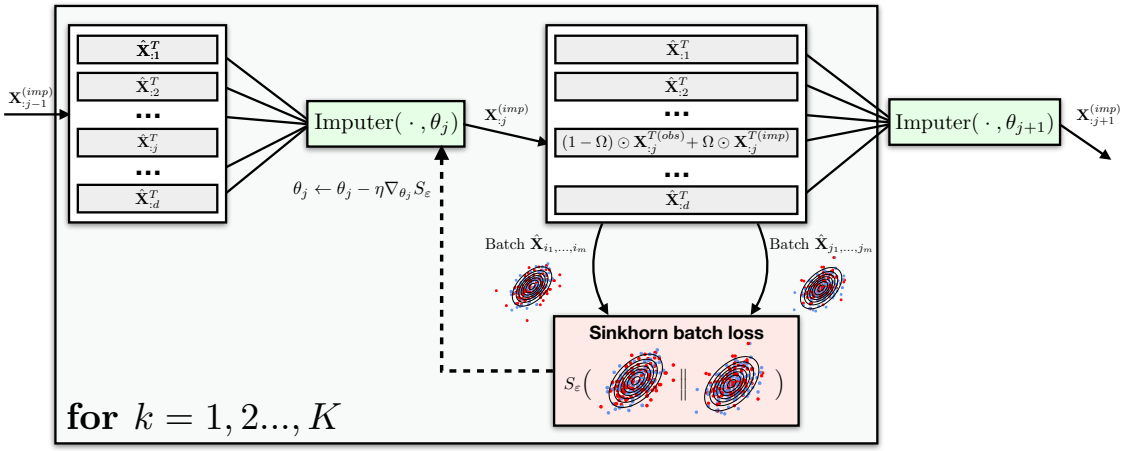


Figure 5.1: Round-robin imputation: illustration of the imputation of the j-th variable in the inner-most loop of Algorithm 7.

Round-robin imputation. A remaining unaddressed point in Algorithm 6 is how to perform the "$\hat{\mathbf{X}} \leftarrow \text{Imputer}(\hat{\mathbf{X}}^t, \Omega, \Theta)$" step in the presence of missing values. A classical method is to perform imputations over variables in a round-robin fashion, i.e. to iteratively predict missing coordinates using other coordinates as features in a cyclical manner. The main advantage of this method is that it decouples variables being used as inputs and those being imputed. This requires having d sets of parameter $(\theta_j)_{1 \leq j \leq d}$, one for each variable, where each θ_j refers to the parameters used to to predict the j-th variable. The j-th variable is iteratively imputed using the $d-1$ remaining variables, according to the chosen

model with parameter θ_j: $\hat{\theta}_j$ is first fitted (using e.g. regression or Bayesian methods), then the j-th variable is imputed. The algorithm then moves to the next variable $j+1$, in a cyclical manner. This round-robin method is implemented for instance in `R`'s `mice` package [van Buuren and Groothuis-Oudshoorn, 2011] or in the `IterativeImputer` method of the `scikit-learn` [Pedregosa et al., 2011] package. When using the Sinkhorn batch loss eq. (5.5) to fit the imputers, this procedure can be seen as a particular case of Algorithm 6 where the imputer parameter Θ is separable with respect to each variable $(\mathbf{x}_{:j})_{1\leq j\leq d}$, i.e. Θ consists in d sets of parameter $(\theta_j)_{1\leq j\leq d}$.

Making this round-robin imputation explicit in the step "$\hat{\mathbf{X}} \leftarrow \text{Imputer}(\hat{\mathbf{X}}^t, \Omega, \Theta)$" of Algorithm 6 leads to Algorithm 7. In Algorithm 7, an imputation $\hat{\mathbf{X}}^t, t = 0, ..., t_{\max}$ is

Algorithm 7 Round-Robin Sinkhorn Imputation

Input: $\mathbf{X} \in \mathbb{R}^{n\times d}$, $\mathbf{\Omega} \in \{0,1\}^{n\times d}$, Imputer$(\cdot,\cdot,\cdot)$, Θ_0, $\varepsilon > 0$, $n \geq m > 0$,
  $\hat{\mathbf{X}}^0 \leftarrow$ same initialization as in Algorithm 5
  $(\hat{\theta}_1, ..., \hat{\theta}_d) \leftarrow \Theta_0$
  for $t = 1, 2, ..., t_{\max}$ **do**
    for $j = 1, 2, ..., d$ **do**
      for $k = 1, 2, ..., K$ **do**
        $\hat{\mathbf{X}}_{:j} \leftarrow \text{Imputer}(\hat{\mathbf{X}}^t_{:-j}, \Omega_{:j}, \hat{\theta}_j)$
        Sample two sets K and L of m indices
        $\mathcal{L} \leftarrow S_\varepsilon(\mu_m(\hat{\mathbf{X}}_K), \mu_m(\hat{\mathbf{X}}_L))$
        $\nabla_{\theta_j}\mathcal{L} \leftarrow \text{AutoDiff}(\mathcal{L})$
        $\hat{\theta}_j \leftarrow \hat{\theta}_j - \alpha\text{Adam}(\nabla_{\theta_j}\mathcal{L})$
      end for
      $\hat{\mathbf{X}}^t_{:j} \leftarrow \text{Imputer}(\hat{\mathbf{X}}^t_{:-j}, \Omega_{:j}, \hat{\theta}_j)$
    end for
    $\hat{\mathbf{X}}^{t+1} \leftarrow \hat{\mathbf{X}}^t$
  end for
Output: Imputations $\hat{\mathbf{X}}^{t_{\max}}$, Imputer$(\cdot,\cdot,\hat{\Theta})$

updated starting from an initial guess $\hat{\mathbf{X}}^0$. The algorithm then consists in three nested loops. (i) The inner-most loop is dedicated to gradient-based updates of the parameter $\hat{\theta}_j$, as illustrated in Figure 5.1. Once this inner-most loop is finished, the j-th variable of $\hat{\mathbf{X}}^t$ is updated using the last update of $\hat{\theta}_j$. (ii) This is performed cyclically over all variables of $\hat{\mathbf{X}}^t$, yielding $\hat{\mathbf{X}}^{t+1}$. (iii) This fitting-and-imputation procedure over all variables is repeated until convergence, or until a given number of iterations is reached.

In practice, several improvements on the generic Algorithms 6 and 7 can be implemented:

1. To better estimate Equation (5.5), one can sample several pairs of batches (instead of a single one) and define $\mathcal{L}$ as the average of S_ε divergences.

2. For Algorithm 7 in a MCAR setting, instead of sampling in each pair two batches from $\hat{\mathbf{X}}$, one of the two batches can be sampled with no missing value on the j-th variable, and the other with missing values on the j-th variable. This allows the imputations for the j-th variable to be fitted on actual non-missing values. This helps ensuring that the imputations for the j-th variable will have a marginal distribution close to that of non-missing values.

3. The order in which the variables are imputed can be adapted. A simple heuristic is to impute variables in increasing order of missing values.

4. During training, the loss can be hard to monitor due to the high variance induced by estimating Equation (5.5) from a few pairs of batches. Therefore, it can be useful to

define a validation set on which fictional additional missing values are sampled to monitor the training of the algorithm, according to the desired accuracy score (e.g. MAE, RMSE or W_2 as in Section 4).

Note that item 2 is *a priori* only legitimate in a MCAR setting. Indeed, under MAR or MNAR assumptions, the distribution of non-missing data is in general not equal to the original (unknown) distribution of missing data.[1] Finally, the use of Adam [Kingma and Ba, 2014] compared to RMSprop in Algorithm 5 is motivated by empirical performance, but does not have a crucial impact on performance. It was observed however that the quality of the imputations given by Algorithm 5 seems to decrease when gradient updates with momentum are used.

4 Experiments

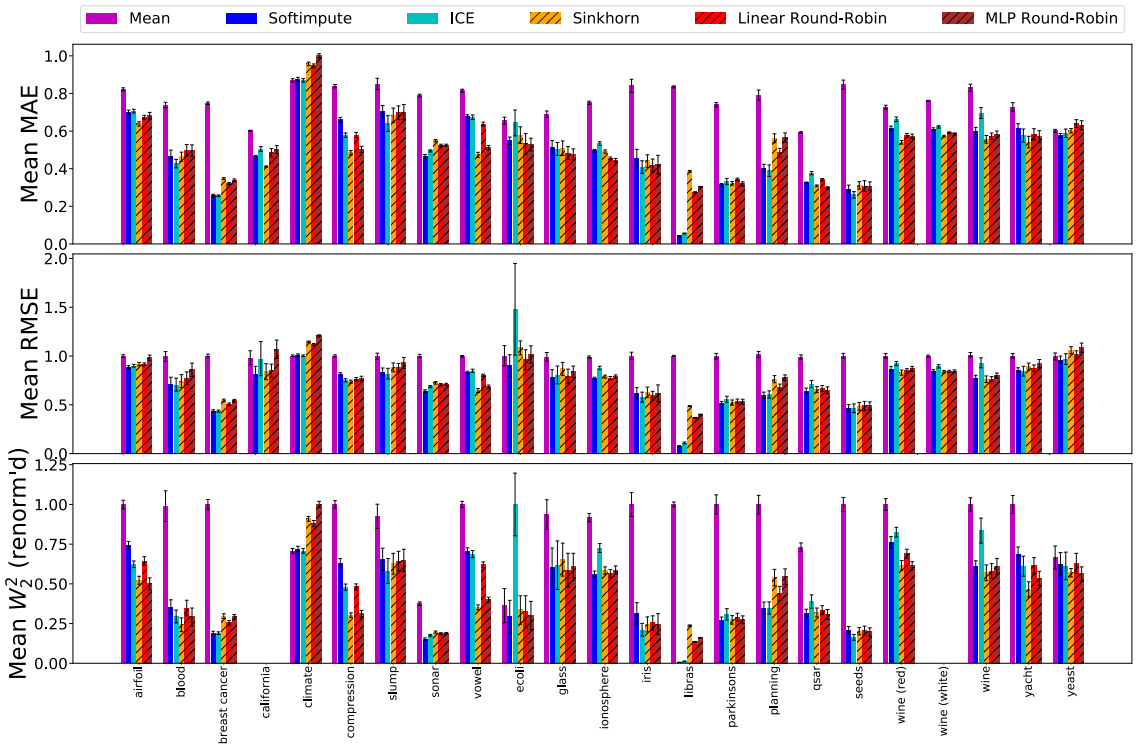


Figure 5.2: (**30% MCAR**) Imputation methods on 23 datasets from the UCI repository (Table 5.1). `Sinkhorn` denotes Algorithm 5 and `Linear RR`, `MLP RR` the two instances of Algorithm 7 precedently described. 30% of the values are missing MCAR. All methods are evaluated on 30 random missing values draws. Error bars correspond to $\pm$ 1 std. For readability we display scaled mean W_2^2, i.e. for each dataset we renormalize the results by the maximum W_2^2. For some datasets W_2 results are not displayed due to their large size, which makes evaluating the unregularized W_2 distance costly.

Baselines. We compare our methods to three baselines:

(i) **mean** is the coordinate-wise mean imputation;

[1] Consider as an example census data in which low/high income people are more likely to fail to answer an income survey than medium income people.

(ii) **ice** (imputation by chained equations) consists in (iterative) imputation using conditional expectation. Here, we use `scikit-learn`'s [Pedregosa et al., 2011] `iterativeImputer` method, which is based on `mice` [van Buuren and Groothuis-Oudshoorn, 2011]. This is one of the most popular methods of imputation as it provides empirically good imputations in many scenario and requires little tuning;

(iii) **softimpute** [Hastie et al., 2015] performs missing values imputation using iterative soft-thresholded SVD's. This method is based on a low-rank assumption for the data and is justified by the fact that many large matrices are well approximated by a low-rank structure [Udell and Townsend, 2019].

Deep learning methods. Additionally, we compare our methods to three DL-based methods:

(iv) **MIWAE** [Mattei and Frellsen, 2019] fits a deep latent variable model (DLVM) [Kingma and Welling, 2014], by optimizing a version of the *importance weighted autoencoder* (IWAE) bound [Burda et al., 2016] adapted to missing data;

(v) **GAIN** [Yoon et al., 2018] is an adaptation of *generative adversarial networks* (GAN) [Goodfellow et al., 2014] to missing data imputation;

(vi) **VAEAC** [Ivanov et al., 2019] are VAEs with easily approximable conditionals that allow to handle missing data.

Transport methods. Three variants of the proposed methods are evaluated:

(vii) **Sinkhorn** designates the direct non-parametric imputation method detailed in Algorithm 5.

For Algorithm 7, two classes of imputers are considered:

(viii) **Linear RR** corresponds to Algorithm 7 where for $1 \leq j \leq d$, $\text{Imputer}(\cdot, \theta_j)$ is a linear model w.r.t. the $d-1$ other variables with weights and biases given by θ_j. This is similar to `mice` or `IterativeImputer`, but fitted with the OT loss eq. (5.5);

(ix) **MLP RR** denotes Algorithm 7 with shallow Multi-Layer Perceptrons (MLP) as imputers. These MLP's have the following architecture: (i) a first $(d-1) \times 2(d-1)$ layer followed by a ReLU layer then (ii) a $2(d-1) \times (d-1)$ layer followed by a ReLU layer and finally (iii) a $(d-1) \times 1$ linear layer. All linear layers have bias terms. Each $\text{Imputer}(\cdot, \theta_j), 1 \leq j \leq d$ is one such MLP with a different set of weights θ_j.

Toy experiments. In Figure 5.3, we generate two-dimensional datasets with strong structures, such as an S-shape, half-moon(s), or concentric circles. A 20% missing rate is introduced (void rows are discarded), and imputations performed using Algorithm 5 or the **ice** method are compared to the ground truth dataset. While the **ice** method is not able to catch the non-linear structure of the distributions at all, **Sinkhorn** performs efficiently by imputing faithfully to the underlying complex data structure (despite the two half-moons and the S-shape being quite challenging). This is remarkable, since Algorithm 5 does not rely on any parametric assumption for the data. This underlines in a low-dimensional setting the flexibility of the proposed method. Finally, note that the trailing points which can be observed for the S shape or the two moons shape come from the fact that Algorithm 5 was used as it is, i.e. with pairs of batches *both* containing missing values, even though these toy examples would have allowed to use batches without missing values. In that case, we obtain imputations that are visually indistinguishable from the ground truth.

4.1 Large-scale experimental setup

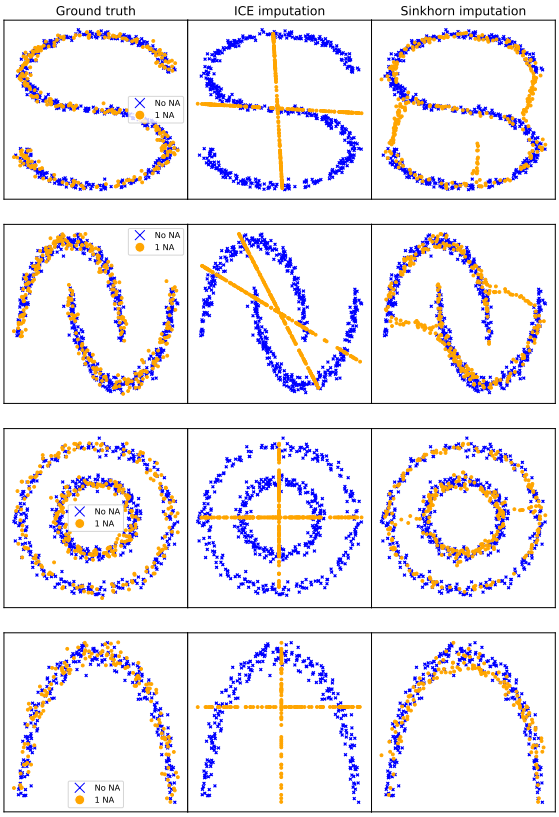


Figure 5.3: Toy examples: 20 % missing values (MCAR) on toy datasets. Blue points have no missing values, orange points have one missing value on either coordinate. **ice** outputs conditional expectation imputations, which are irrelevant due to the high non-linearity of these examples. Since algorithm 5 does not assume a parametric form for the imputations, it is able to satisfyingly impute missing values.

We evaluate each method on 23 datasets from the UCI machine learning repository[2] (see Table 5.1) with varying proportions of missing data and different missing data mechanisms. These datasets only contain quantitative features. Prior to running the experiments, the data is whitened (i.e. centered and scaled to variable-wise unit variance). For each dataset, all methods are evaluated on 30 different draws of missing values masks. For all Sinkhorn-based imputation methods, the regularization parameter ϵ is set to 5% of the median distance between initialization values with no further dataset-dependent tuning. If the dataset has more than 256 points, the batch size is fixed to 128, otherwise to $2^{\lfloor \frac{n}{2} \rfloor}$ where n is the size of the dataset. The noise parameter η in Algorithm 5 is fixed to 0.1. For Sinkhorn round-robin models (**Linear RR** and **MLP RR**), the maximum number of cycles is 10, 10 pairs of batches are sampled per gradient update, and an ℓ^2-weight regularization of magnitude 10^{-5} is applied during training. For all 3 Sinkhorn-based methods, we use gradient methods with adaptive step sizes as per algorithms 5 and 7, with an initial step size fixed to 10^{-2}. For **softimpute**, the hyperparameter is selected at each run through cross-validation on a small grid. This CV is performed by sampling additional missing values. For DL-based methods, the implementations provided in open-access by the authors were used[3,4,5], with the hyperparameter settings recommended in the corresponding papers. In particular, for **GAIN** the α parameter is selected using cross-validation. GPUs are used for Sinkhorn and deep learning methods. The code to reproduce the experiments is available at `https://github.com/BorisMuzellec/MissingDataOT`.

Missing value generation mechanisms. The implementation of a MCAR mechanism is straightfoward. On the contrary, many different mechanisms can lead to a MAR or MNAR setting. We here describe those used in our experiments. In the **MCAR** setting, each value is masked according to the realization of a Bernoulli random variable with a fixed parameter. In the **MAR** setting, for each experiment, a fixed subset of variables that cannot have missing values is sampled. Then, the remaining variables have missing values according to a logistic model with random weights, which takes the non-missing variables

[2] `https://archive.ics.uci.edu/ml/index.php`
[3] `https://github.com/pamattei/miwae`
[4] `https://github.com/jsyoon0823/GAIN`
[5] `https://github.com/tigvarts/vaeac`

Table 5.1: Summary of datasets

dataset	n	d
airfoil_self_noise	1503	5
blood_transfusion	748	4
breast_cancer_diagnostic	569	30
california	20640	8
climate_model_crashes	540	18
concrete_compression	1030	7
concrete_slump	103	7
connectionist_bench_sonar	208	60
connectionist_bench_vowel	990	10
ecoli	336	7
glass	214	9
ionosphere	351	34
iris	150	4
libras	360	90
parkinsons	195	23
planning_relax	182	12
qsar_biodegradation	1055	41
seeds	210	7
wine	178	13
wine_quality_red	1599	10
wine_quality_white	4898	11
yacht_hydrodynamics	308	6
yeast	1484	8

as inputs. A bias term is fitted using line search to attain the desired proportion of missing values. Finally, two different mechanisms are implemented in the **MNAR** setting. The first is identical to the previously described MAR mechanism, but the inputs of the logistic model are then masked by a MCAR mechanism. Hence, the logistic model's outcome now depends on potentially missing values. The second mechanism, 'self masked', samples a subset of variables whose values in the lower and upper p-th percentiles are masked according to a Bernoulli random variable, and the values in-between are left not missing. As detailed in the Section 5, MCAR experiments were performed with 10%, 30% and 50% missing rates, while MAR and both MNAR settings (quantile and logistic masking) were evaluated with a 30% missing rate.

Metrics. Imputation methods are evaluated according to two "pointwise" metrics: mean absolute error (MAE) and root mean square error (RMSE); and one metric on distributions: the squared Wasserstein distance between empirical distributions on points with missing values. Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ be a dataset with missing values. When (i, j) spots a missing entry, recall that $\hat{x}_{ij}$ denotes the corresponding imputation, and let us note x_{ij}^{true} the ground truth. Let $m_0 \stackrel{\text{def}}{=} \#\{(i,j), \omega_{ij} = 0\}$ and $m_1 \stackrel{\text{def}}{=} \#\{i : \exists j,\ \omega_{ij} = 0\}$ respectively denote the total number of missing values and the number of data points with at least one missing value.

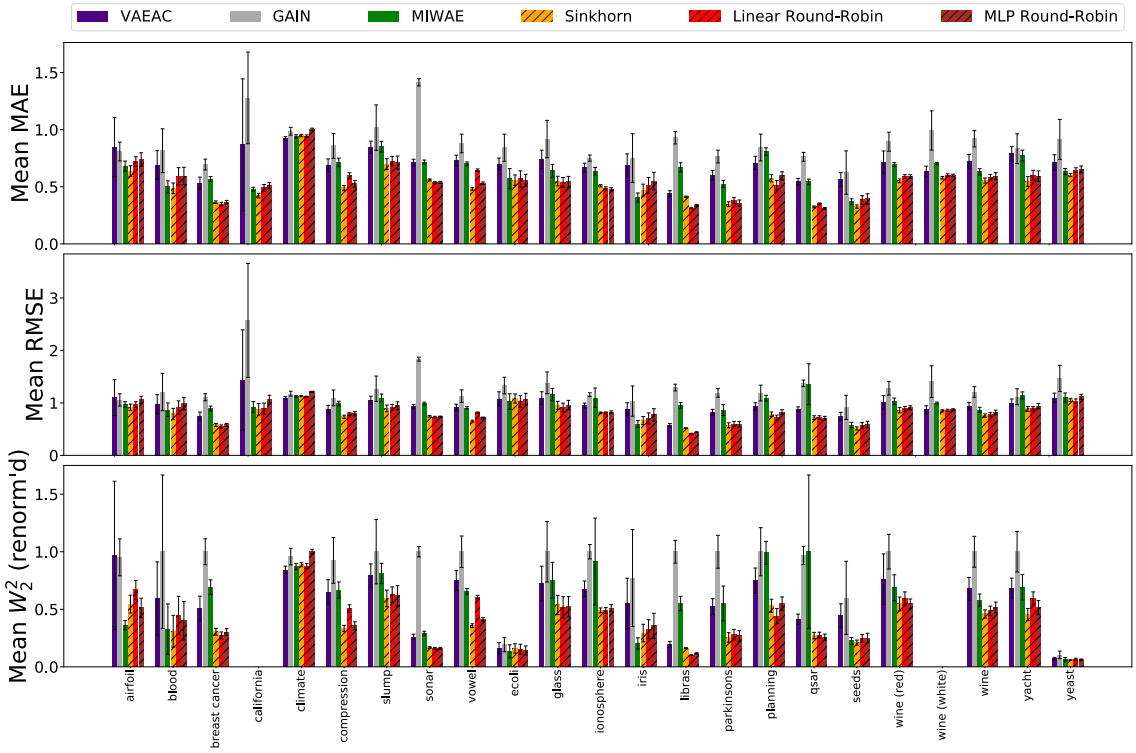


Figure 5.4: (**30% MNAR**) Imputation methods on 23 datasets from the UCI repository (Table 5.1). Values are missing MNAR according to the logistic mechanism described in Section 4, with 30% variables used as inputs of a logistic masking model for the 70% remaining variables. 30% of those input variables are then masked at random. Hence, all variables have 30% missing values. All methods are evaluated on the same 30 random missing values draws. Error bars correspond to $\pm$ 1 std. For readability we display scaled mean W_2^2, i.e. for each dataset we renormalize the results by the maximum W_2^2. For some datasets W_2 results are not displayed due to their large size, which makes evaluating the unregularized W_2 distance costly.

Set $M_1 \overset{\text{def}}{=} \{i : \exists j, \omega_{ij} = 0\}$. We define MAE, RMSE and W_2 imputation metrics as

$$\frac{1}{m_0} \sum_{(i,j):\omega_{ij}=0} |x_{i,j}^{\text{true}} - \hat{x}_{ij}|, \tag{MAE}$$

$$\sqrt{\frac{1}{m_0} \sum_{(i,j):\omega_{ij}=0} (x_{i,j}^{\text{true}} - \hat{x}_{ij})^2}, \tag{RMSE}$$

$$W_2^2 \left(\mu_{m_1}(\hat{\mathbf{X}}_{M_1}), \mu_{m_1}(\mathbf{X}_{M_1}^{(true)}) \right). \tag{W_2}$$

Results. The complete results of the experiments are reported in Section 5. In Figure 5.2 and Figure 5.4, the proposed methods are respectively compared to baselines and Deep Learning (DL) methods in a MCAR and a logistic masking MNAR setting with 30% missing data. As can be seen from Figure 5.2, the linear round-robin model matches or out-performs `scikit`'s iterative imputer (**ice**) on MAE and RMSE scores for most datasets. Since both methods are based on the same cyclical linear imputation model but with different loss functions, this shows that the batched Sinkhorn loss in Equation (5.5) is well-adapted to imputation with parametric models. Comparison with DL methods (Figure 5.4) shows that the proposed OT-based methods consistently outperform DL-based methods, and have the

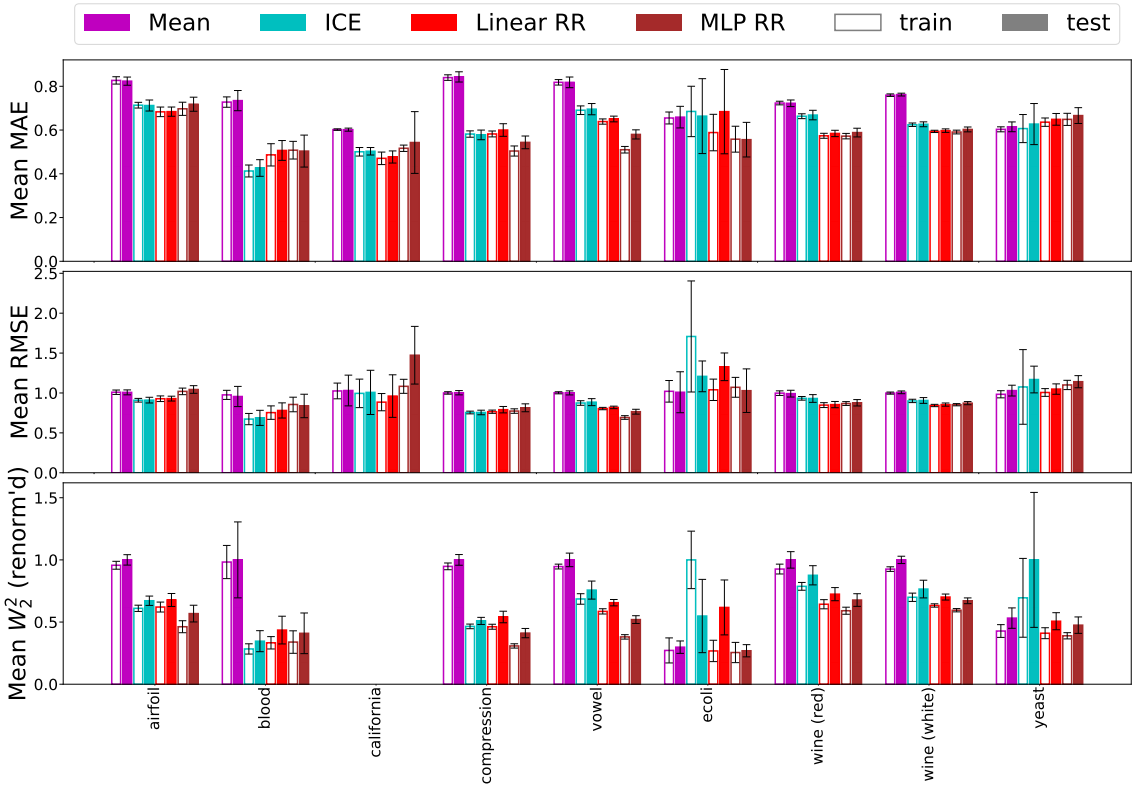


Figure 5.5: (**OOS**) Out of sample imputation: 70% of the data is used for training (filled bars) and 30 % for testing with fixed parameters (dotted bars). 30% of the values are missing MCAR accross both training and testing sets.

additional benefit of having a lower variance in their results overall. Interestingly, while the MAE and RMSE scores of the round-robin MLP model are comparable to that of the linear RR, its W_2 scores are generally better. This suggests that more powerful base imputer models lead to better W_2 scores, from which one can conclude that Equation (5.5) is a good proxy for optimizing the unavailable Equation (5.1) score, and that Algorithm 7 is efficient at doing so. Furthermore, one can observe that the direct imputation method is very competitive over all data and metrics and is in general the best performing OT-based method, as could be expected from the fact that its imputation model is not restricted by a parametric assumption. This favorable behaviour tends to be exacerbated with a growing proportion of missing data, see Figure 5.9 in Section 5.

MAR and MNAR. Figure 5.4 above and Figures 5.10 to 5.12 in Section 5 display the results of our experiments in MAR and MNAR settings, and show that the proposed methods perform well and are robust to difficult missingness mechanisms. This is remarkable, as the proposed methods do not attempt to model those mechanisms. Finally, note that the few datasets on which the proposed methods do not perform as well as baselines – namely `libras` and to a smaller extent `planning_relax` – remain consistently the same across all missingness mechanisms and missing rates. This suggests that this behavior is due to the particular structure of those datasets, rather than to the missingness mechanisms themselves.

Out-of-sample imputation. As mentioned in Section 3, a key benefit of fitting a parametric imputing model with algorithms 6 and 7 is that the resulting model can then be used to impute missing values in out-of-sample (OOS) data. In Figure 5.5, we evaluate the Linear RR and MLP RR models in an OOS imputation experiment. We compare the

training and OOS MAE, RMSE and OT scores on a collection of datasets selected to have a sufficient number of points. At each run, we randomly sample 70% of the data to be used for training, and the remaining 30% to evaluate OOS imputation. 30% of the values are missing MCAR, uniformly over training and testing sets. Out of the methods presented earlier on, we keep those that allow OOS: for the **ice**, **Linear RR** and **MLP RR** methods, OOS imputation is simply performed using the round-robin scheme without further fitting of the parameters on the new data. For the **mean** baseline, missing values in the testing data are imputed using mean observed values from the training data. Figure 5.5 confirms the stability at testing time of the good performance of **Linear RR** and **MLP RR**.

Conclusion

We have shown in this chapter how OT metrics could be used to define a relevant loss for missing data imputation. This loss corresponds to the expectation of Sinkhorn divergences between randomly sampled batches. To minimize it, two classes of algorithms were proposed: one that freely estimates one parameter per imputed value, and one that fits a parametric model. The former class does not rely on making parametric assumptions on the underlying data distribution, and can be used in a very wide range of settings. On the other hand, after training, the latter class allows out-of-sample imputation. To make parametric models trainable, the classical round-robin mechanism was used. Experiments on a variety of datasets, and for numerous missing value settings (MCAR, MAR and MNAR with varying missing values proportions) showed that the proposed models are very competitive, even compared to recent methods based on deep learning. These results confirmed that our loss is a good optimizable proxy for imputation metrics. Future work includes further theoretical study of our loss function (5.5) within the OT framework.

5 Complementary Experimental Results

This appendix contains a full account of our experimental results. These results correspond to the missing value mechanisms described in Section 4:

1. 10% MCAR (Figure 5.7), 30% MCAR (Figure 5.8) and 50% MCAR (Figure 5.9);
2. 30% MAR on 70% of the variables with a logistic masking model (Figure 5.10);
3. 30% MNAR generated with a logistic masking model, whose inputs are then themselves masked (Figure 5.11);
4. 30% MNAR on 30% of the variables, generated by censoring upper and lower quartiles (Figure 5.12).

These experiments follow the setup described in Section 4. In all the following figures, error bars correspond to ± 1 standard deviation across the 30 runs performed on each dataset. For some datasets, the W_2 score is not represented: this is due to their large size, which makes computing unregularized OT computationally intensive.

The results show that the proposed methods, Algorithm 5 and Algorithm 7 with linear and shallow MLP imputers, are very competitive compared to state-of-the-art methods, including those based on deep learning [Mattei and Frellsen, 2019, Yoon et al., 2018, Ivanov et al., 2019], in a wide range of missing data regimes.

Runtimes. Figure 5.6 represents the average runtimes of the methods evaluated in Figure 5.11. These runtimes show that Algorithm 5 has computational running times on par with VAEAC, and faster than the two remaining DL-based methods (GAIN and MIWAE). Round-robin methods are the slowest overall, but the base imputer model being used seems to have nearly no impact on runtimes. This is due to the fact that the computational bottleneck of the proposed methods is the number of Sinkhorn batch divergences that are computed. This number can be made lower by e.g. reducing the number of gradient steps performed for each variable (parameter K in algorithm 7), or the number of cycles t_{max}. This fact suggests that more complex models could be used in round-robin imputation without much additional computational cost.

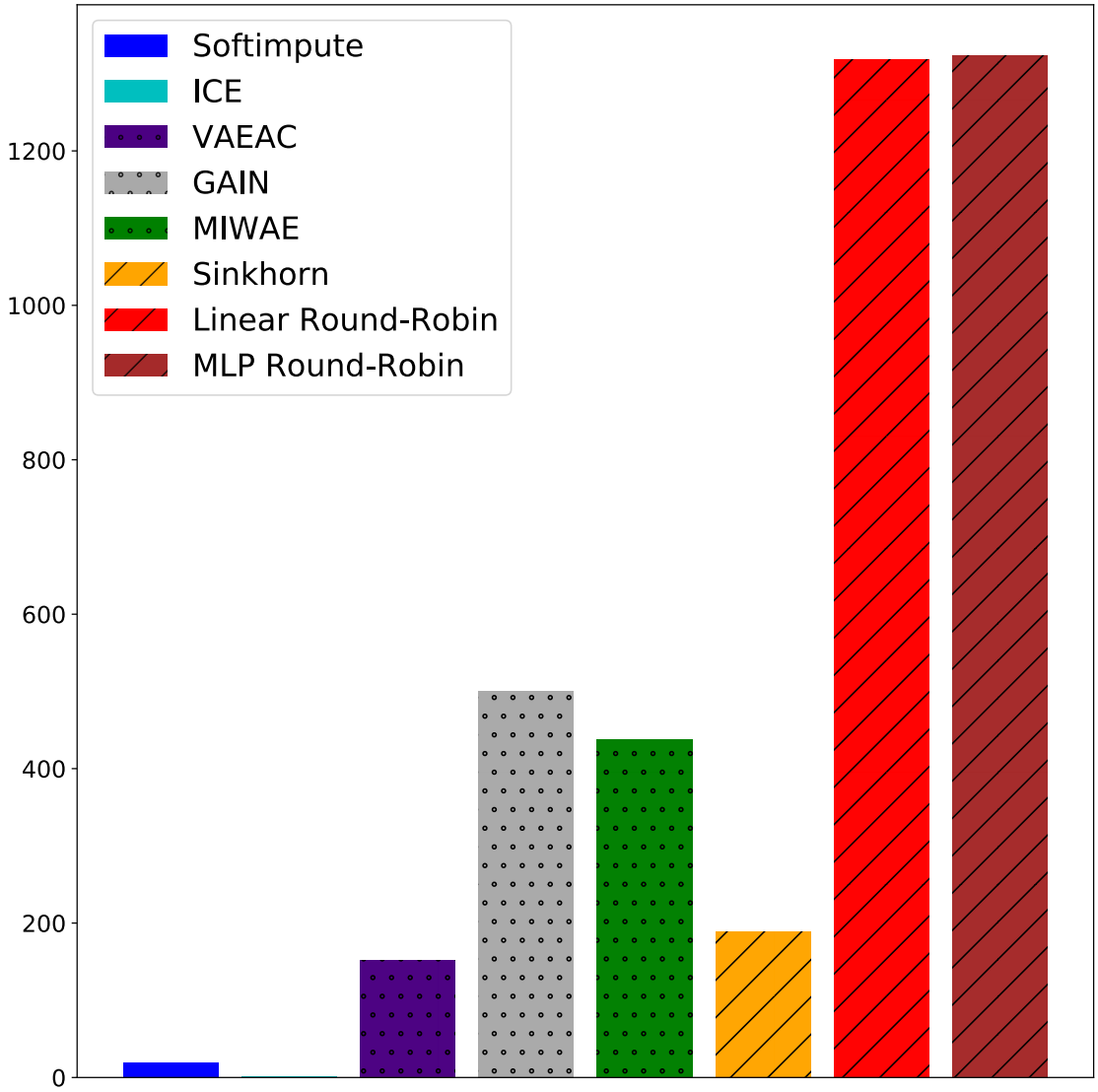


Figure 5.6: Average runtimes (in seconds, over 30 runs and 23 datasets) for the experiment described in fig. 5.11. Note that these times are indicative, as runs where randomly assigned to different GPU models, which may have an impact on runtimes.

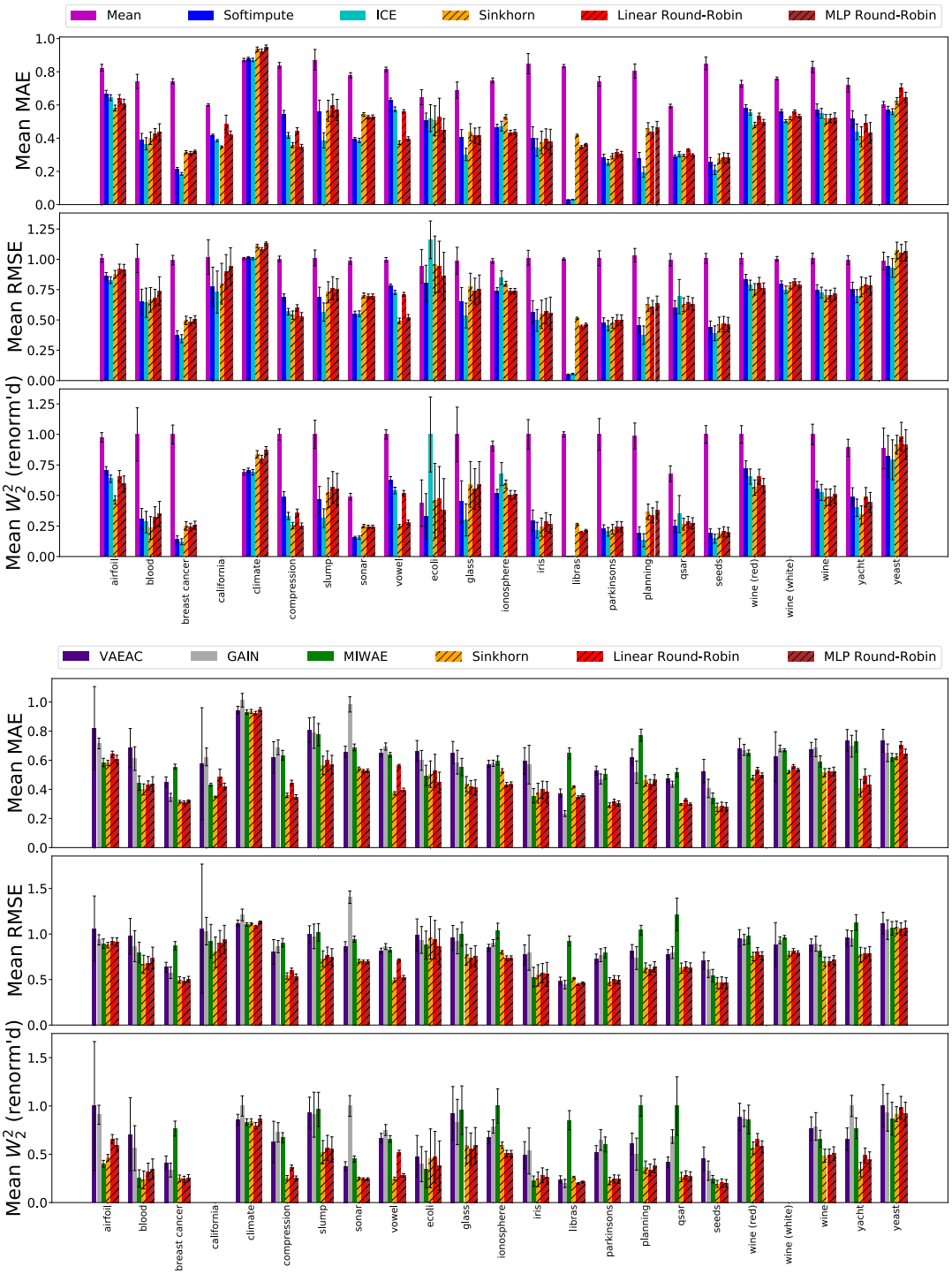


Figure 5.7: (**10 % MCAR**)

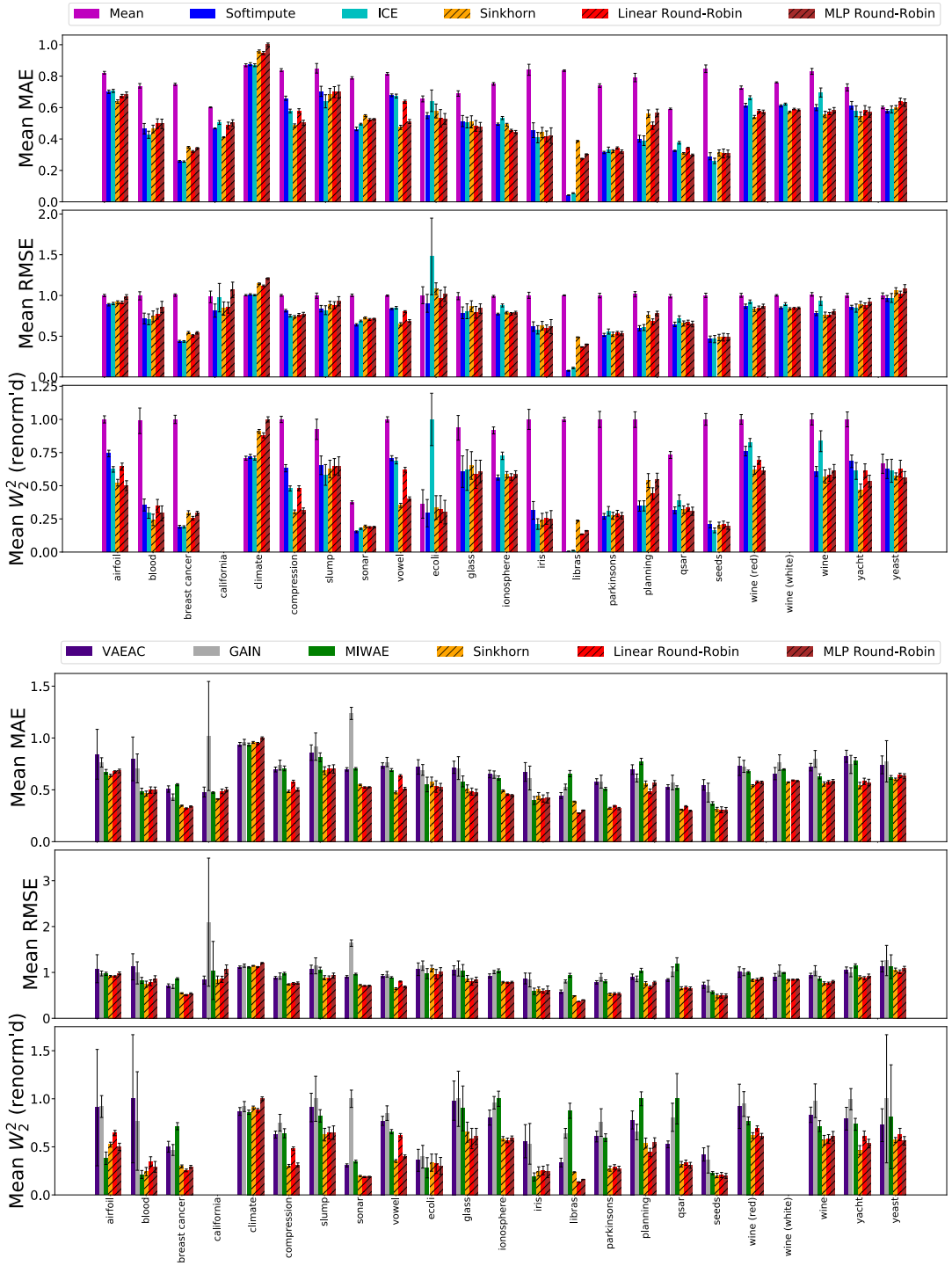


Figure 5.8: (**30 % MCAR**)

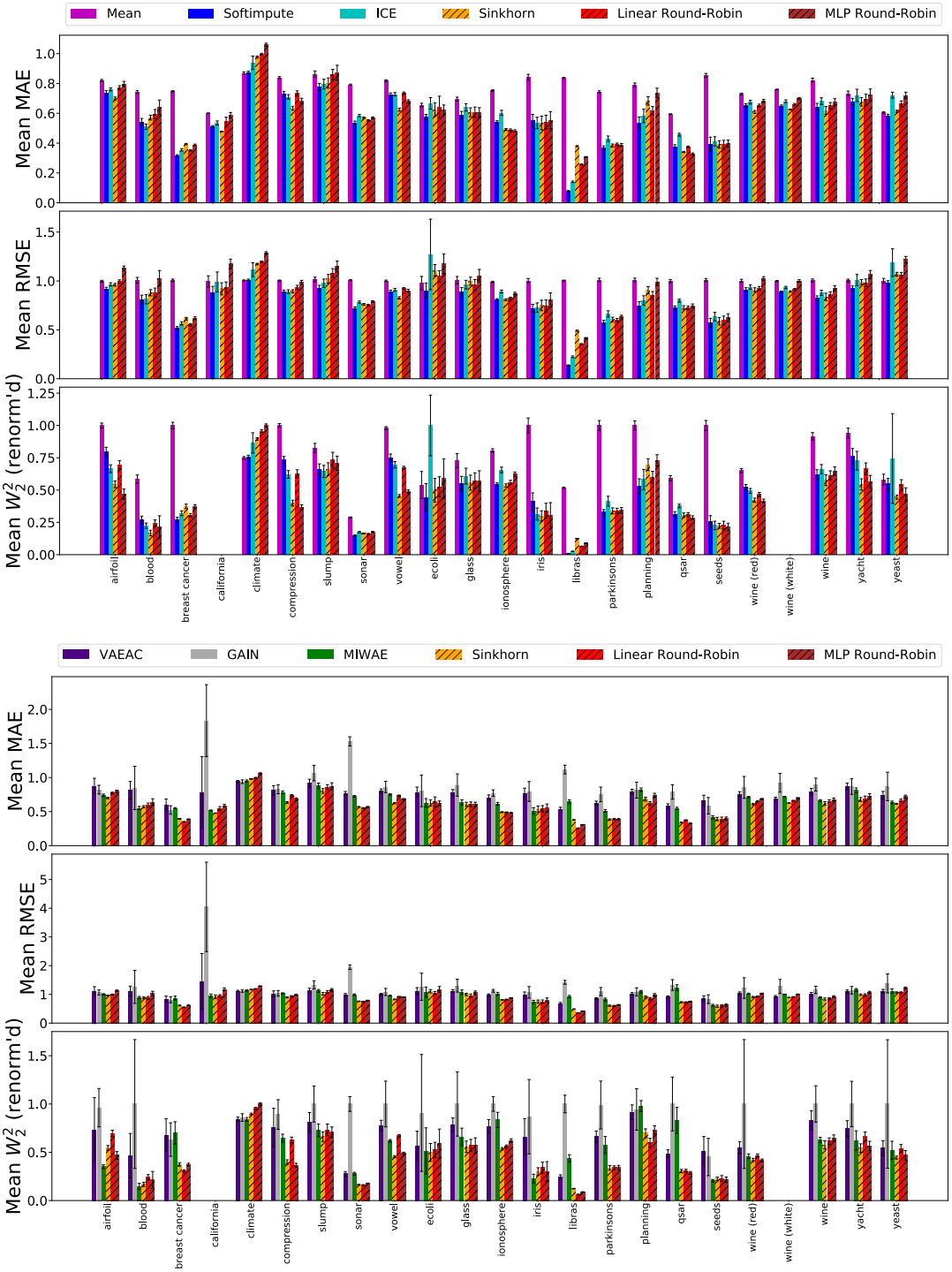


Figure 5.9: (**50 % MCAR**)

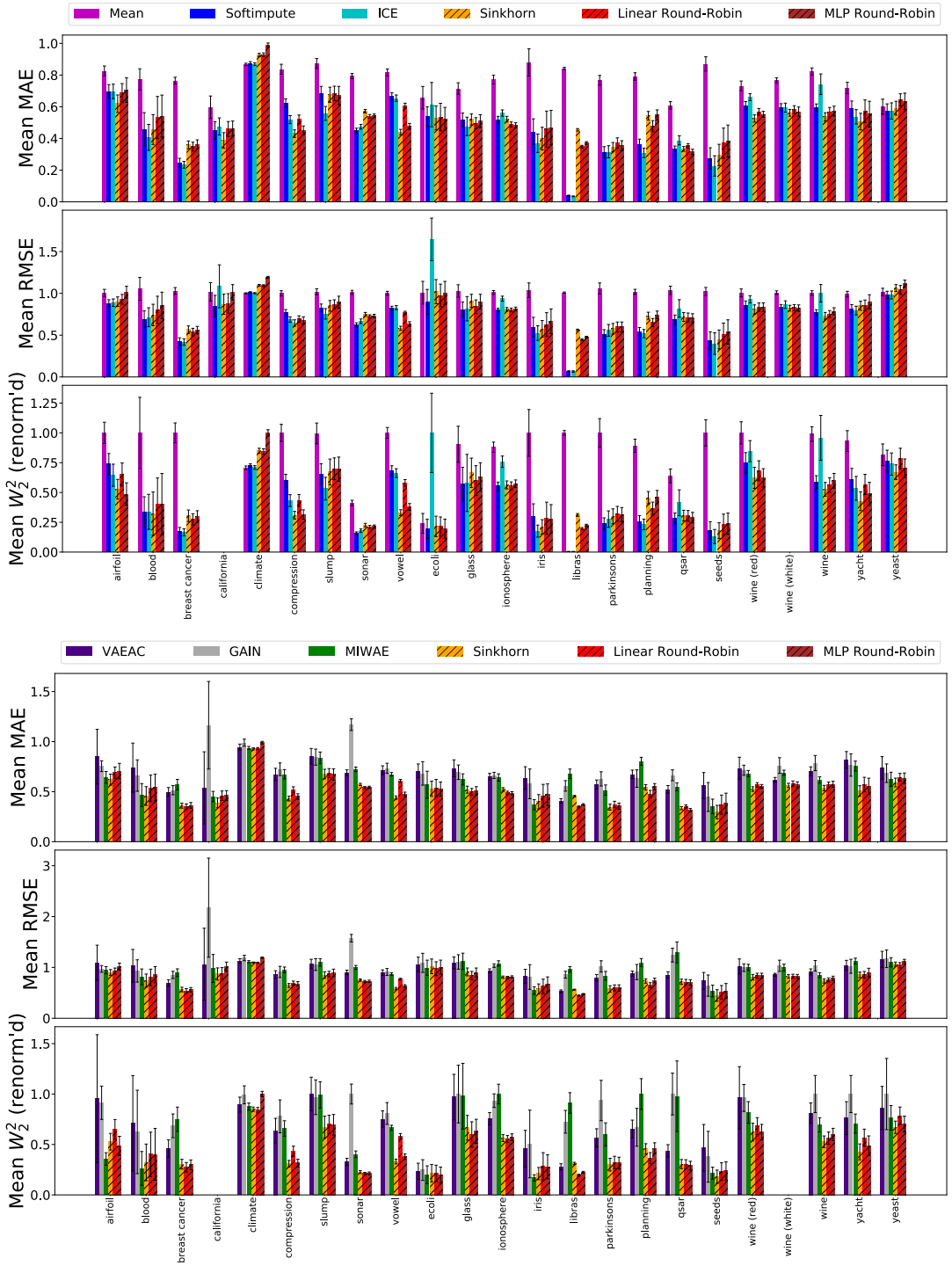


Figure 5.10: (**30 % MAR**)

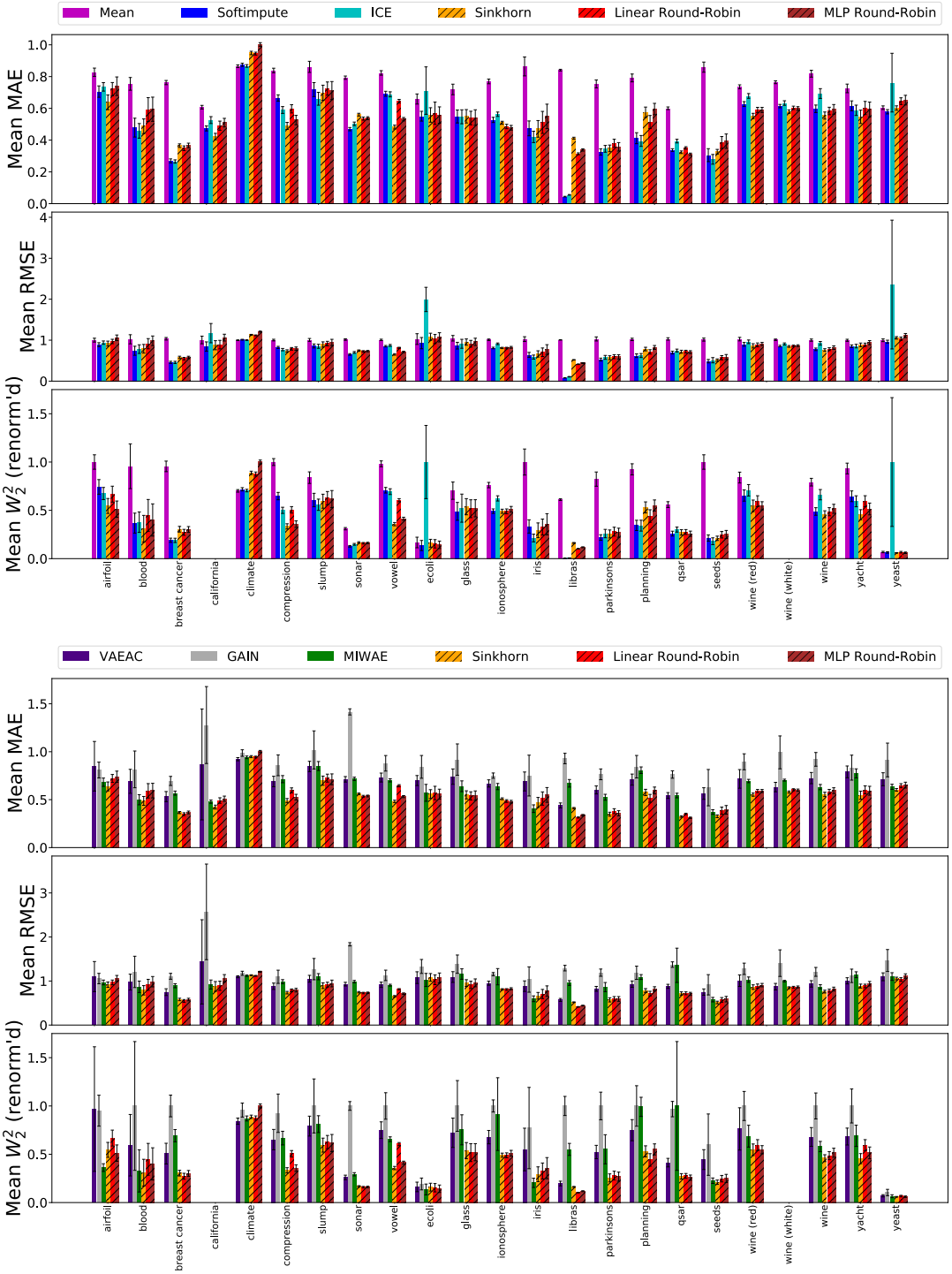


Figure 5.11: (**30 % MNAR, logistic masking**)

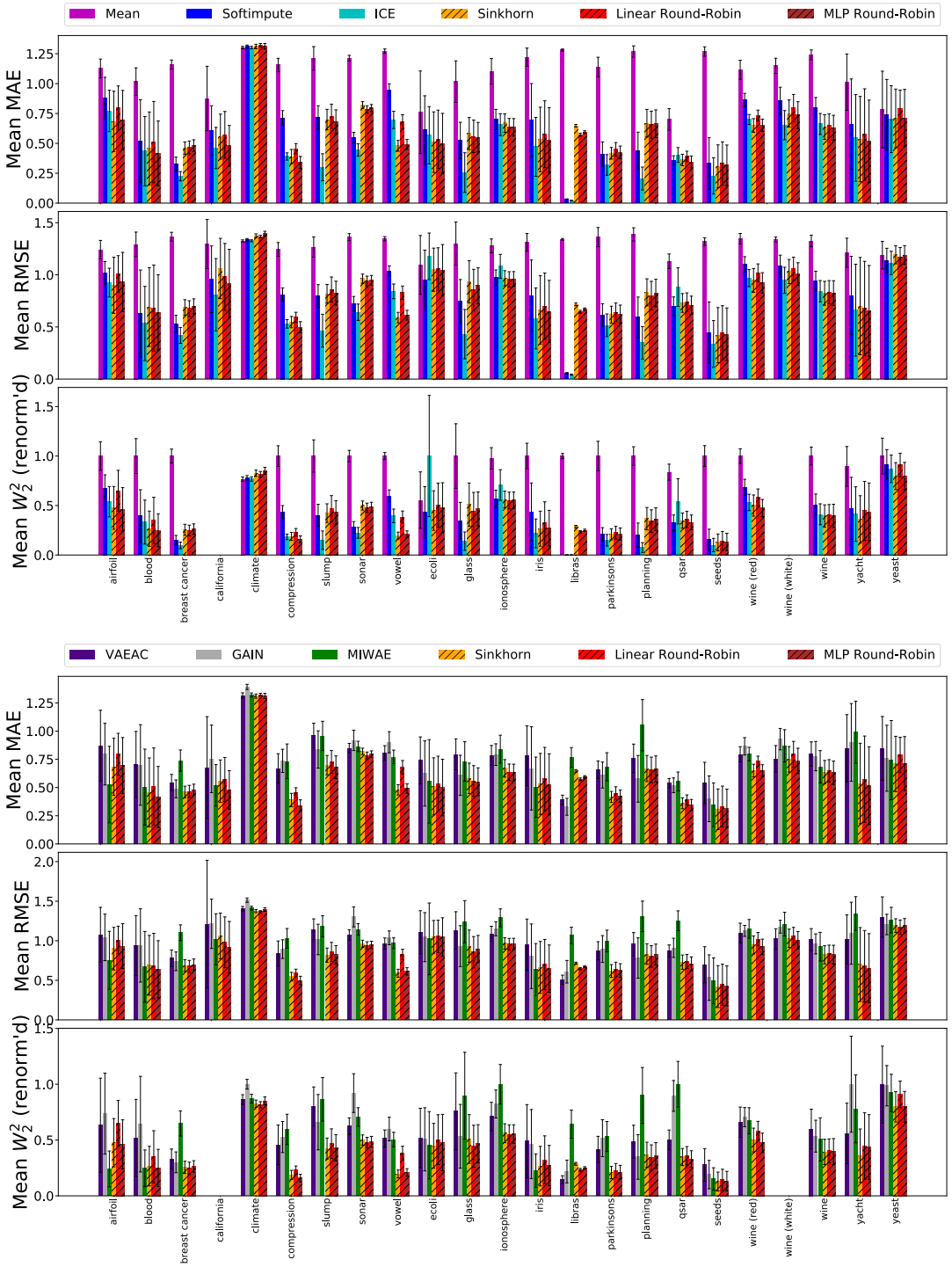


Figure 5.12: (**30 % MNAR, quantile masking**)

Conclusion

In this thesis, we leveraged the specificities of variants of OT relying on closed forms or regularization to provide new theoretical results lying at their intersections, and to carry out applications to machine learning problems. More precisely, we proposed contributions lying at the the intersection of the Bures-Wasserstein geometry and either projected OT or regularized OT, and an application of entropy-regularized OT to missing data imputation, which we summarize below.

Theoretical contributions.

In Chapter 3, we showed that projected OT maps could be extrapolated to the whole space, and that the resulting plans are characterized by their disintegration on the graph of the original lower-dimensional map. Among this class of subspace-optimal plans, we focused on two specific instances motivated by particular properties: Monge-Independent couplings can be obtained as the infinite-samples limit of discrete subspace-optimal plans between empirical measures, and Monge-Knothe plans generalize the Knothe-Rosenblatt transport to k-dimensional subspaces, with similar properties. Making links with the Bures-Wasserstein geometry, closed-form expressions for MI, MK and KR transports between Gaussian measures were proved. Those results complement the recent literature on subspace-projected OT [Rabin et al., 2011, Bonneel et al., 2015, Paty and Cuturi, 2019] by providing maps and couplings in addition to the discrepencies between the original measures.

In Chapter 4, we interfaced between the Bures-Wasserstein geometry and entropy-regularized OT by providing closed-form expressions for entropic OT and unbalanced entropic OT with a quadratic cost between Gaussian measures. These results constitute the first non-trivial closed-form expressions for entropic OT. We showed that entropic balanced (resp. unbalanced) OT between Gaussian measures on $\mathbb{R}^d$ can be written as a balanced (resp. unbalanced) Gaussian measure on $\mathbb{R}^d \times \mathbb{R}^d$. Next, we proved that the debiased barycenters between Gaussian measures satisfy a fixed-point equation that generalizes that of [Agueh and Carlier, 2011]. The proof of these results relies on an adaptation of Sinkhorn's algorithm to quadratic dual potentials, which yields a fixed-point equation that can be put in parallel with the discrete algorithm. Those closed forms will provide a test case for the analysis of entropic OT and Sinkhorn's algorithm, and provide a principled way to circumvent the non-differentiability issues that may arise when the Bures distance is taken between singular covariance matrices. Finally, this first closed form for unbalanced OT allows to get a better understanding of the trade-off between transport and mass creation/deletion.

Applications and numerical tools.

Variants of OT geometries that admit closed forms or tractable computation were put to use in machine learning contexts, developing appropriate numerical tools when required.

In Chapter 2, the Bures-Wasserstein geometry was leveraged to extend the traditional embeddings as points in $\mathbb{R}^d$ to elliptically-contoured probability distributions. Our work extends on recent takes on probabilistic embeddings [Vilnis and McCallum, 2015, Athiwaratkun and Wilson, 2017] that were restricted to diagonal covariance matrices, hindered by numerical constraints due to the choice of the KL geometry. Drawing links between a $\mathbf{L}\mathbf{L}^T$ factorization and the Riemannian structure of the Bures metric, we proposed numerical tools and practical guidelines for gradient descent on the Bures distance, relying notably on the Newton-Schulz algorithm to compute Monge maps and inverse maps, and on the memorization of those maps instead of automatic differentiation. By minimizing skip-gram losses [Mikolov et al., 2013a] or hinge losses [Vilnis and McCallum, 2015] on large corpora and replacing Euclidean dot products with the Bures fidelity, we obtained word embeddings that are competitive with then state-of-the-art methods on similarity, entailment and hypernymy benchmarks. Further, in Chapter 3 we showed how subspace projection on the principal directions of a given elliptical embedding's covariance can be used to influence the similarity results of polysemous words.

Finally, in Chapter 5 we showed how optimal transport can be used as a criterion for missing data imputation, with the intuitive underlying idea that two random batches from the same dataset should have similar distributions. This criterion can be encoded as a loss function using an OT-based discrepancy. In our work, we chose Sinkhorn divergences as they inherit from the smoothness and computational properties of entropic OT while defining positive definite divergences. We proposed two algorithms for minimizing this loss, the first and most flexible not requiring any parametric assumption on the data distribution, and the second allowing to fit parametric models according to our OT batch loss through a round-robin scheme. Extensive experiments with different missing data mechanisms and missing rates showed that our OT criterion and algorithms are very competitive against state-of-the-art methods, including those based on deep learning.

Perspectives

Taking a step back, a common denominator between Bures-Wasserstein transport, subspace-optimal transport, Knothe-Rosenblatt transport and (unbalanced) Gaussian entropic transport is that they leverage some notion of stability to yield closed-form optimal couplings. Indeed, optimal couplings between Gaussian measures are Gaussian distributions themselves in unregularized, entropy-regularized and unbalanced entropy-regularized settings, Gaussian subspace conditionals remain Gaussian in subspace-OT, and the OT maps between conditional distributions in KR recursion are monotone 1D maps. Therefore, it appears that such stability properties can be used as building blocks for closed-form OT expressions or variants: as an example, an OT-based distance for Gaussian mixture models (GMM) was recently studied in [Delon and Desolneux, 2019] by considering couplings that are GMMs themselves. Hence, a future research direction consists in studying problems with couplings or maps restricted to some class, or that only partially match distributions as in [Alfonsi et al., 2019] which considers moment-matching couplings. Alternatively, a promising direction is to first lift distributions to some high-dimensional feature space and approximate them (using e.g. elliptical distributions), before using closed-form couplings and distances. For instance, Fréchet inception distances [Heusel et al., 2017] consist in using the Bures-Wasserstein geometry on inception features [Szegedy et al., 2016], and are now widely used to measure the performance of GANs.

Bibliography

Pierre Ablin, Gabriel Peyré, and Thomas Moreau. Super-efficiency of automatic differentiation for functions defined as a minimum. In *Proceedings of the 37th International Conference on Machine Learning*. PMLR, 2020.

P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009.

M. Agueh and G. Carlier. Barycenters in the Wasserstein space. *SIAM*, 43(2):904–924, 2011.

Ravindra K Ahuja, James B Orlin, and Thomas L Magnanti. *Network flows: theory, algorithms, and applications*. Prentice-Hall, 1993.

Aurélien Alfonsi, Rafaël Coyaud, Virginie Ehrlacher, and Damiano Lombardi. Approximation of optimal transport problems with marginal moments constraints, 2019.

Pedro C. Álvarez-Esteban, E. del Barrio, J.A. Cuesta-Albertos, and C. Matran. A fixed-point approach to barycenters in Wasserstein space. *Journal of Mathematical Analysis and Applications*, 441(2):744 – 762, 2016.

L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows in metric spaces and in the space of probability measures*. Springer, 2006.

Alexandr Andoni, Assaf Naor, and Ofer Neiman. Snowflake universality of Wasserstein spaces. *Annales scientifiques de l'ENS*, 2015. to appear.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

Ben Athiwaratkun and Andrew Wilson. Multimodal word distributions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1645–1656. Association for Computational Linguistics, 2017.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226, September 2009.

Marco Baroni, Raffaella Bernardi, Ngoc-Quynh Do, and Chung-chieh Shan. Entailment above the word level in distributional semantics. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 23–32. Association for Computational Linguistics, 2012.

Jean-David Benamou. Numerical resolution of an "unbalanced" mass transport problem. *ESAIM: Mathematical Modelling and Numerical Analysis - Modélisation Mathématique et Analyse Numérique*, 37:851–868, 2003.

Jean-David Benamou and Yann Brenier. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.

Ingemar Bengtsson and Karol Życzkowski. *Geometry of quantum states: an introduction to quantum entanglement*. Cambridge university press, 2017.

Rajendra Bhatia. *Positive Definite Matrices*. Princeton Series in Applied Mathematics. Princeton University Press, Princeton, NJ, USA, 2007.

Rajendra Bhatia, Tanvi Jain, and Yongdo Lim. On the Bures-Wasserstein distance between positive definite matrices. *Expositiones Mathematicae*, 2018.

Garrett Birkhoff. Three observations on linear algebra. *Univ. Nac. Tacuman, Rev. Ser. A*, 5:147–151, 1946.

Mathieu Blondel, Vivien Seguy, and Antoine Rolet. Smooth and sparse optimal transport. In *International Conference on Artificial Intelligence and Statistics*, pages 880–889, 2018.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

Raicho Bojilov and Alfred Galichon. Matching in closed-form: equilibrium, identification, and comparative statics. *Economic Theory*, 61(4):587–609, 2016.

Nicolas Bonneel, Michiel van de Panne, Sylvain Paris, and Wolfgang Heidrich. Displacement interpolation using lagrangian mass transport. *ACM Transactions on Graphics (TOG)*, 30(6):1–12, 2011.

Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. Sliced and Radon Wasserstein Barycenters of Measures. *Journal of Mathematical Imaging and Vision*, 1 (51):22–45, 2015.

Ingwer Borg and Patrick JF Groenen. *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media, 2005.

Jean Bourgain. On Lipschitz embedding of finite metric spaces in Hilbert space. *Israel Journal of Mathematics*, 52(1):46–52, 1985.

Yann Brenier. Décomposition polaire et réarrangement monotone des champs de vecteurs. *CR Acad. Sci. Paris Sér. I Math.*, 305:805–808, 1987.

Alexander M Bronstein, Michael M Bronstein, and Ron Kimmel. Generalized multidimensional scaling: a framework for isometry-invariant partial surface matching. *Proceedings of the National Academy of Sciences*, 103(5):1168–1172, 2006.

Elia Bruni, Nam Khanh Tran, and Marco Baroni. Multimodal distributional semantics. *J. Artif. Int. Res.*, 49(1):1–47, January 2014.

Mihai Bădoiu, Piotr Indyk, and Anastasios Sidiropoulos. Approximation algorithms for embedding general metrics into trees. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 512–521. Society for Industrial and Applied Mathematics, 2007.

Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. *International Conference on Learning Representations*, 2016.

Donald Bures. An extension of Kakutani's theorem on infinite product measures to the tensor product of semifinite w*-algebras. *Transactions of the American Mathematical Society*, 135:199–212, 1969.

Luis A Caffarelli. Some regularity properties of solutions of monge ampere equation. *Communications on pure and applied mathematics*, 44(8-9):965–969, 1991.

Stamatis Cambanis, Steel Huang, and Gordon Simons. On the theory of elliptically contoured distributions. *Journal of Multivariate Analysis*, 11(3):368 – 385, 1981.

Guillaume Carlier, Alfred Galichon, and Filippo Santambrogio. From Knothe's transport to Brenier's map and a continuation method for optimal transport. *SIAM J. Math. An.*, 2009.

Y. Chen, T. T. Georgiou, and A. Tannenbaum. Optimal transport for Gaussian mixture models. *IEEE Access*, 7:6269–6278, 2019.

Lenaïc Chizat. *Unbalanced optimal transport: Models, numerical methods, applications.* PhD thesis, 2017.

Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in neural information processing systems*, pages 3036–3046, 2018.

Lenaïc Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. An interpolating distance between optimal transport and Fisher–Rao metrics. *Foundations of Computational Mathematics*, 18(1):1–44, 2018a.

Lenaïc Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of Computation*, 87 (314):2563–2609, 2018b.

Lenaic Chizat, Pierre Roussillon, Flavien Léger, François-Xavier Vialard, and Gabriel Peyré. Faster Wasserstein distance estimation with the Sinkhorn divergence. *arXiv preprint arXiv:2006.08172*, 2020.

Nicolas Courty, Remi Flamary, Alain Rakotomamonjy, and Devis Tuia. Optimal transport for domain adaptation. In *NIPS, Workshop on Optimal Transport and Machine Learning*, Montréal, Canada, December 2014.

Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. 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 3730–3739. Curran Associates, Inc., 2017.

Imre Csiszár. I-divergence geometry of probability distributions and minimization problems. *The annals of probability*, pages 146–158, 1975.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, pages 2292–2300, 2013.

Marco Cuturi and Arnaud Doucet. Fast computation of Wasserstein barycenters. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, number 2, pages 685–693, Bejing, China, 22–24 Jun 2014. PMLR.

Marco Cuturi and Gabriel Peyré. A smoothed dual approach for variational Wasserstein problems. *SIAM Journal on Imaging Sciences*, 9(1):320–343, 2016.

Marco Cuturi, Jean-Philippe Vert, Øystein Birkenes, and Tomoko Matsui. A kernel for time series based on global alignments. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing 2007*, 2007. to appear.

John M Danskin. The theory of max-min and its application to weapons allocation problems. 1967.

George B Dantzig. Programming of interdependent activities: Ii mathematical model. *Econometrica*, 17(3/4):200–211, 1949.

George B Dantzig. Application of the simplex method to a transportation problem. *Activity analysis and production and allocation*, 1951.

Jan De Leeuw. Applications of convex analysis to multidimensional scaling. In *Recent Developments in Statistics*, 1977.

Eustasio del Barrio and Jean-Michel Loubes. The statistical effect of entropic regularization in optimal transportation. *arXiv preprint arXiv:2006.05199*, 2020.

Julie Delon and Agnes Desolneux. A Wasserstein-type distance in the space of Gaussian Mixture Models, 2019.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

Steffen Dereich, Michael Scheutzow, and Reik Schottstedt. Constructive quantization: Approximation by empirical measures. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 49, pages 1183–1203, 2013.

Ishan Deshpande, Ziyu Zhang, and Alexander G. Schwing. Generative modeling using the sliced Wasserstein distance. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 3483–3491, 2018.

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.

DC Dowson and BV Landau. The Fréchet distance between multivariate normal distributions. *Journal of multivariate analysis*, 12(3):450–455, 1982.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul): 2121–2159, 2011.

RM Dudley. The speed of mean Glivenko-Cantelli convergence. *The Annals of Mathematical Statistics*, 40(1):40–50, 1969.

Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1-2):115–166, 2018.

Paul L Fackler. Notes on matrix calculus. Technical report, North Carolina State University, 2005.

Jittat Fakcharoenphol, Satish Rao, and Kunal Talwar. A tight bound on approximating arbitrary metrics by tree metrics. In *Proceedings of the thirty-fifth annual ACM symposium on Theory of computing*, pages 448–455. ACM, 2003.

KT Fang, S Kotz, and KW Ng. *Symmetric Multivariate and Related Distributions*. Chapman and Hall/CRC, 1990.

Kilian Fatras, Younes Zine, Rémi Flamary, Rémi Gribonval, and Nicolas Courty. Learning with minibatch Wasserstein : asymptotic and gradient properties. *CoRR*, abs/1910.04091, 2019.

Sira Ferradans, Nicolas Papadakis, Gabriel Peyré, and Jean-François Aujol. Regularized discrete optimal transport. *SIAM Journal on Imaging Sciences*, 7(3):1853–1882, 2014.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and MMD using Sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, pages 2681–2690, 2019.

Alessio Figalli. *The Monge–Ampère equation and its applications*. 2017.

Alessio Figalli, Francesco Maggi, and Aldo Pratelli. A mass transportation approach to quantitative isoperimetric inequalities. *Inventiones mathematicae*, 182(1):167–211, 2010.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. Placing search in context: the concept revisited. *ACM Trans. Inf. Syst.*, 20(1):116–131, 2002.

Rémi Flamary, Karim Lounici, and André Ferrari. Concentration bounds for linear monge mapping estimation and optimal transport domain adaptation. *arXiv preprint arXiv:1905.10155*, 2019.

LR Ford and DR Fulkerson. Flows in networks. 1962.

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.

Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya-Polo, and Tomaso Poggio. Learning with a Wasserstein loss. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, page 2053–2061, Cambridge, MA, USA, 2015. MIT Press.

Charlie Frogner, Farzaneh Mirzazadeh, and Justin Solomon. Learning embeddings into entropic Wasserstein spaces. In *International Conference on Learning Representations*, 2019.

Matthias Gelbrich. On a formula for the l2 Wasserstein metric between measures on Euclidean and Hilbert spaces. *Mathematische Nachrichten*, 147(1):185–203, 1990.

Aude Genevay. *Entropy-regularized optimal transport for machine learning*. PhD thesis, Paris Sciences et Lettres, 2019.

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.

Aude Genevay, Gabriel Peyre, and Marco Cuturi. Learning generative models with Sinkhorn divergences. In Amos Storkey and Fernando Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 1608–1617, Playa Blanca, Lanzarote, Canary Islands, 09–11 Apr 2018. PMLR.

Aude Genevay, Lénaïc Chizat, Francis Bach, Marco Cuturi, and Gabriel Peyré. Sample complexity of Sinkhorn divergences. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 1574–1583. PMLR, 16–18 Apr 2019.

Augusto Gerolin, Juri Grossi, and Paola Gori-Giorgi. Kinetic correlation functionals from the entropic regularization of the strictly correlated electrons problem. *Journal of Chemical Theory and Computation*, 16(1):488–498, 01 2020.

Clark R Givens, Rae Michael Shortt, et al. A class of Wasserstein metrics for probability distributions. *The Michigan Mathematical Journal*, 31(2):231–240, 1984.

Amir Globerson, Gal Chechik, Fernando Pereira, and Naftali Tishby. Euclidean embedding of co-occurrence data. *Journal of Machine Learning Research*, 8(Oct):2265–2295, 2007.

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.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014.

Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864. ACM, 2016.

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.

Guy Halawi, Gideon Dror, Evgeniy Gabrilovich, and Yehuda Koren. Large-scale learning of word relatedness with constraints. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, pages 1406–1414, New York, NY, USA, 2012. ACM.

Trevor Hastie, Rahul Mazumder, Jason D. Lee, and Reza Zadeh. Matrix completion and low-rank SVD via fast alternating least squares. *J. Mach. Learn. Res.*, 16(1):3367–3402, January 2015. ISSN 1532-4435.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. 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 6626–6637. Curran Associates, Inc., 2017.

Nicholas J. Higham. *Functions of Matrices: Theory and Computation (Other Titles in Applied Mathematics)*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2008.

Felix Hill, Roi Reichart, and Anna Korhonen. Simlex-999: Evaluating semantic models with genuine similarity estimation. *Comput. Linguist.*, 41(4):665–695, December 2015. ISSN 0891-2017.

Geoffrey E Hinton and Sam T Roweis. Stochastic neighbor embedding. In *Advances in Neural Information Processing Systems*, pages 857–864, 2003.

Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

Frank L Hitchcock. The distribution of a product from several sources to numerous localities. *Journal of mathematics and physics*, 20(1-4):224–230, 1941.

Oleg Ivanov, Michael Figurnov, and Dmitry Vetrov. Variational autoencoder with arbitrary conditioning. *International Conference on Learning Representations*, 2019.

Hicham Janati, Marco Cuturi, and Alexandre Gramfort. Debiased Sinkhorn barycenters. In *Proceedings of the 34th International Conference on Machine Learning*, 2020a.

Hicham Janati, Boris Muzellec, Gabriel Peyré, and Marco Cuturi. Entropic optimal transport between (unbalanced) Gaussian measures has a closed form. *arXiv preprint arXiv:2006.02572*, 2020b.

Tony Jebara, Risi Kondor, and Andrew Howard. Probability product kernels. *Journal of Machine Learning Research*, 5:819–844, 2004.

William B. Johnson and Joram Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. In *Conference in modern analysis and probability (New Haven, Conn., 1982)*, volume 26 of *Contemp. Math.*, pages 189–206. Amer. Math. Soc., Providence, RI, 1984.

Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the fokker–planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17, 1998.

Julie Josse, François Husson, et al. missMDA: a package for handling missing values in multivariate data analysis. *Journal of Statistical Software*, 70(1):1–31, 2016.

Leonid Vitalievich Kantorovich. On the translocation of masses. In *Dokl. Akad. Nauk. USSR (NS)*, volume 37, pages 199–201, 1942.

J.K. Kim and Z. Ying. *Data Missing Not at Random*. Statistica Sinica. Institute of Statistical Science, Academia Sinica, 2018.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

Herbert Knothe. Contributions to the theory of convex bodies. *The Michigan Mathematical Journal*, 4(1):39–52, 1957.

Soheil Kolouri, Kimia Nadjahi, Umut Simsekli, Roland Badeau, and Gustavo Rohde. Generalized sliced Wasserstein distances. In *Advances in Neural Information Processing Systems*, pages 261–272, 2019.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966, 2015.

J.M. Lee. *Riemannian Manifolds: An Introduction to Curvature*. Graduate Texts in Mathematics. Springer New York, 1997.

Matthias Liero, Alexander Mielke, and Giuseppe Savaré. Optimal transport in competition with reaction: the Hellinger–Kantorovich distance and geodesic curves. *SIAM Journal on Mathematical Analysis*, 48(4):2869–2911, 2016.

Matthias Liero, Alexander Mielke, and Giuseppe Savaré. Optimal entropy-transport problems and a new Hellinger–Kantorovich distance between positive measures. *Inventiones Mathematicae*, 211(3):969–1117, 2018.

Roderick J A Little and Donald B Rubin. *Statistical analysis with missing data*. John Wiley & Sons, 2002.

Giulia Luise, Alessandro Rudi, Massimiliano Pontil, and Carlo Ciliberto. Differential properties of Sinkhorn approximation for learning with Wasserstein distance. In *Advances in Neural Information Processing Systems*, pages 5859–5870, 2018.

Giulia Luise, Saverio Salzo, Massimiliano Pontil, and Carlo Ciliberto. Sinkhorn barycenters with free support via Frank-Wolfe algorithm. In *Advances in Neural Information Processing Systems*, 2019.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

Luigi Malagò, Luigi Montrucchio, and Giovanni Pistone. Wasserstein-Riemannian geometry of positive-definite matrices. *arXiv preprint arXiv:1801.09269*, 2018.

Anton Mallasto, Augusto Gerolin, and Hà Quang Minh. Entropy-regularized 2-Wasserstein distance between Gaussian measures. *arXiv preprint arXiv:2006.03416*, 2020.

Yariv Maron, Michael Lamar, and Elie Bienenstock. Sphere embedding: An application to part-of-speech induction. In *Advances in Neural Information Processing Systems*, pages 1567–1575, 2010.

Pierre-Alexandre Mattei and Jes Frellsen. MIWAE: Deep generative modelling and imputation of incomplete data sets. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, pages 4413–4423, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

Imke Mayer, Julie Josse, Nicholas Tierney, and Nathalie Vialaneix. R-miss-tastic: a unified platform for missing values methods and workflows. *arXiv preprint arXiv:1908.04822*, 2019.

Robert J. McCann. A convexity principle for interacting gases. *Advances in Mathematics*, 128(1):153 – 179, 1997.

Gonzalo Mena and Jonathan Niles-Weed. Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem. In *Advances in Neural Information Processing Systems 32*, pages 4541–4551. Curran Associates, Inc., 2019.

Arthur Mensch and Gabriel Peyré. Online Sinkhorn: optimal transportation distances from sample streams. *arXiv preprint arXiv:2003.01415*, 2020.

Quentin Mérigot. A multiscale approach to optimal transport. In *Computer Graphics Forum*, volume 30, pages 1583–1592. Wiley Online Library, 2011.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *ICLR Workshop*, 2013a.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119, 2013b.

George A. Miller. WordNet: A lexical database for English. *Commun. ACM*, 38(11):39–41, November 1995.

George A. Miller and Walter G. Charles. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28, 1991.

K Mohan and J Pearl. Graphical Models for Processing Missing Data. *Journal of American Statistical Association (JASA)*, 2019.

Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*, 1781.

Jared S. Murray and Jerome P. Reiter. Multiple imputation of missing categorical and continuous values via Bayesian mixture models with local dependence. *Journal of the American Statistical Association*, 111(516):1466–1479, 2016.

Boris Muzellec and Marco Cuturi. Generalizing point embeddings using the Wasserstein space of elliptical distributions. In *Advances in Neural Information Processing Systems 31*, pages 10237–10248. Curran Associates, Inc., 2018.

Boris Muzellec and Marco Cuturi. Subspace detours: Building transport plans that are optimal on subspace projections. In *Advances in Neural Information Processing Systems*, pages 6917–6928, 2019.

Boris Muzellec, Richard Nock, Giorgio Patrini, and Frank Nielsen. Tsallis regularized optimal transport and ecological inference. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

Boris Muzellec, Julie Josse, Claire Boyer, and Marco Cuturi. Missing data imputation using optimal transport. *arXiv preprint arXiv:2002.03860*, 2020.

Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. 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 6341–6350. Curran Associates, Inc., 2017.

Jonathan Niles-Weed and Philippe Rigollet. Estimation of wasserstein distances in the spiked transport model. *arXiv preprint arXiv:1909.07513*, 2019.

Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg Corrado, and Jeffrey Dean. Zero-shot learning by convex combination of semantic embeddings. In *Proceedings of the International Conference on Learning Representations*, 2014.

Ingram Olkin and Friedrich Pukelsheim. The distance between two random vectors with given dispersion matrices. *Linear Algebra and its Applications*, 48:257–263, 1982.

François-Pierre Paty and Marco Cuturi. Subspace robust Wasserstein distances. 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 5072–5081, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

François-Pierre Paty, Alexandre d'Aspremont, and Marco Cuturi. Regularity as regularization: Smooth and strongly convex Brenier potentials in optimal transport. *Proceedings of AISTATS 2020*, 2020.

Karl Pearson. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Ofir Pele and Michael Werman. Fast and robust earth mover's distances. In *2009 IEEE 12th International Conference on Computer Vision*, pages 460–467. IEEE, 2009.

Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

Roger Penrose. A generalized inverse for matrices. In *Mathematical proceedings of the Cambridge philosophical society*, volume 51, pages 406–413. Cambridge University Press, 1955.

Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

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.

Julien Rabin, Sira Ferradans, and Nicolas Papadakis. Adaptive color transfer with relaxed optimal transport. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 4852–4856. IEEE, 2014.

Svetlozar T Rachev and Ludger Rüschendorf. *Mass Transportation Problems: Volume I: Theory*, volume 1. Springer Science & Business Media, 1998.

Kira Radinsky, Eugene Agichtein, Evgeniy Gabrilovich, and Shaul Markovitch. A word at a time: Computing word relatedness using temporal semantic analysis. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 337–346, New York, NY, USA, 2011. ACM.

Aaditya Ramdas, Nicolás García Trillos, and Marco Cuturi. On Wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19(2):47, 2017.

Murray Rosenblatt. Remarks on a multivariate transformation. *The annals of mathematical statistics*, 23(3):470–472, 1952.

Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

Mark Rowland, Jiri Hron, Yunhao Tang, Krzysztof Choromanski, Tamas Sarlos, and Adrian Weller. Orthogonal estimation of Wasserstein distances. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 186–195. PMLR, 16–18 Apr 2019.

Herbert Rubenstein and John B. Goodenough. Contextual correlates of synonymy. *Commun. ACM*, 8(10):627–633, October 1965.

Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

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, 2000.

Tim Salimans, Han Zhang, Alec Radford, and Dimitris Metaxas. Improving GANs using optimal transport. In *International Conference on Learning Representations*, 2018.

Filippo Santambrogio. Optimal transport for applied mathematicians, 2015.

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.

Shaun Seaman, John Galati, Dan Jackson, and John Carlin. What is meant by "missing at random"? *Statistical Science*, pages 257–268, 2013.

Vivien Seguy, Bharath Bhushan Damodaran, Rémi Flamary, Nicolas Courty, Antoine Rolet, and Mathieu Blondel. Large-scale optimal transport and mapping estimation. In *Proceedings of the International Conference in Learning Representations*, 2018.

Thibault Séjourné, Jean Feydy, François-Xavier Vialard, Alain Trouvé, and Gabriel Peyré. Sinkhorn divergences for unbalanced optimal transport. *arXiv preprint arXiv:1910.12958*, 2019.

Sameer Shirdhonkar and David W Jacobs. Approximate earth mover's distance in linear time. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.

Sidak Pal Singh, Andreas Hug, Aymeric Dieuleveut, and Martin Jaggi. Context mover's distance & barycenters: Optimal transport of contexts for building representations. In *International Conference on Artificial Intelligence and Statistics*, pages 3437–3449, 2020.

Richard Sinkhorn. A relationship between arbitrary positive matrices and doubly stochastic matrices. *The annals of mathematical statistics*, 35(2):876–879, 1964.

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 (TOG)*, 34(4):1–11, 2015.

Daniel J. Stekhoven and Peter Buhlmann. MissForest a non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 10 2011. ISSN 1367-4803. doi: 10.1093/bioinformatics/btr597.

C Szegedy, V Vanhoucke, S Ioffe, J Shlens, and Z Wojna. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2016.

Asuka Takatsu. Wasserstein geometry of Gaussian measures. *Osaka J. Math.*, 48(4): 1005–1026, 12 2011.

Joshua B Tenenbaum, Vin De Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

Minh thang Luong, Richard Socher, and Christopher D. Manning. Better word representations with recursive neural networks for morphology. In *In Proceedings of the Thirteenth Annual Conference on Natural Language Learning. Tomas Mikolov, Wen-tau*, 2013.

T. Tieleman and G. Hinton. RMSprop Gradient Optimization. 2015.

Vayer Titouan, Rémi Flamary, Nicolas Courty, Romain Tavenard, and Laetitia Chapel. Sliced Gromov-Wasserstein. In *Advances in Neural Information Processing Systems*, pages 14726–14736, 2019.

Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Scholkopf. Wasserstein auto-encoders. In *International Conference on Learning Representations, ICLR*, 2018.

AN Tolstoi. Methods of finding the minimal total kilometrage in cargo transportation planning in space. *TransPress of the National Commissariat of Transportation*, 1:23–55, 1930.

Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B Altman. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6):520–525, 2001.

Madeleine Udell and Alex Townsend. Why are big data matrices approximately low rank? *SIAM Journal on Mathematics of Data Science*, 1(1):144–160, 2019.

Nikolai G Ushakov. *Selected topics in characteristic functions*. Walter de Gruyter, 1999.

Stef van Buuren. *Flexible Imputation of Missing Data*. Chapman and Hall/CRC, Boca Raton, FL, 2018.

Stef van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software, Articles*, 45(3):1–67, 2011. ISSN 1548-7660.

Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *(IEEE) Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

Luke Vilnis and Andrew McCallum. Word representations via Gaussian embedding. *Proceddings of the International Conference on Learning Representations (ICLR)*, 2015. arXiv preprint arXiv:1412.6623.

John Von Neumann. A certain zero-sum two-person game equivalent to the optimal assignment problem. *Contributions to the Theory of Games*, 2(0):5–12, 1953.

Jonathan Weed and Francis Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *arXiv preprint arXiv:1707.00087*, 2017.

K.Q. Weinberger and L.K. Saul. Distance metric learning for large margin nearest neighbor classification. *The Journal of Machine Learning Research*, 10:207–244, 2009.

Jiqing Wu, Zhiwu Huang, Dinesh Acharya, Wen Li, Janine Thoma, Danda Pani Paudel, and Luc Van Gool. Sliced Wasserstein generative models. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Dongqiang Yang and David M. W. Powers. Measuring semantic similarity in the taxonomy of WordNet. In *Proceedings of the Twenty-eighth Australasian Conference on Computer Science - Volume 38*, ACSC '05, pages 315–322, Darlinghurst, Australia, Australia, 2005. Australian Computer Society, Inc.

Jinsung Yoon, James Jordon, and Mihaela van der Schaar. GAIN: Missing data imputation using generative adversarial nets. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, pages 5689–5698, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

Ziwei Zhu, Tengyao Wang, and Richard J Samworth. High-dimensional principal component analysis with heterogeneous missingness. *arXiv preprint arXiv:1906.12125*, 2019.