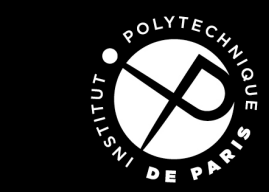


INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2021IPPAG001

Thèse de doctorat

Advances in optimal transport and applications to neuroscience

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à Inria Saclay - l'Ecole nationale de la statistique et de l'administration économique

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

Thèse présentée et soutenue à Palaiseau, le 23/03/2021, par

HICHAM JANATI

Composition du Jury :

Quentin Mérigot Pr, Université Paris-Saclay	Président
Filippo Santambrogio Pr, Université Claude Bernard - Lyon 1	Rapporteur
Stefan Haufe Pr, Universitätsmedizin Berlin	Rapporteur
Dirk Lorenz Pr, Technische Universität Braunschweig	Examinateur
Julie Delon Pr, Université de Paris	Examinatrice
Klaus-Robert Müller Pr, Technische Universität Berlin / Google Brain	Invité
Alexandre Gramfort Pr, Inria Saclay (Parietal)	Directeur de thèse
Marco Cuturi Pr, Google Brain / ENSAE	Co-directeur de thèse

Abstract

Brain imaging devices can provide a glimpse at neural activity in multiple spatial locations and time points. Moreover, neuroimaging studies are usually conducted for multiple individuals undergoing the same experimental protocol. Inferring the underlying sources is a challenging inverse problem that can only be tackled by biasing the solutions with prior domain knowledge. Several prior hypotheses have been pursued in the literature such as promoting sparse over dense solutions or solving the problem for multiple subjects at once. However, none take advantage of the particular spatial geometry of the problem. The purpose of this thesis is to exploit the multi-subject, spatial and temporal aspects of magneto-encephalography data as much as possible to improve the conditioning of the inverse problem. To that end, our contributions revolve around three axes: optimal transport (OT), sparse multi-task regression and time series. Indeed, the ability of OT to capture spatial disparities between measures makes it very well suited to compare and average neural activation patterns based on their shape and location over the cortical surface of the brain. For the sake of scalability, we take advantage of the entropic formulation of optimal transport, which we argue has two important missing pieces. From a theoretical perspective, it has no closed form analytical expressions, and from a practical perspective, entropy leads to a significant increase in variance known as *entropic bias*. We complete this puzzle by studying multivariate Gaussians for which we uncover an entropic OT closed form and propose *debiased* algorithms to compute fast and accurate optimal transport barycenters. Second, we define a multi-task prior based on OT and sparse penalties to jointly solve the inverse problem for multiple subjects to promote spatially coherent solutions. Our real data experiments highlight the benefits of using OT as a prior over classical multi-task regression penalties. Finally, we propose a loss function to compare and average spatio-temporal data that computes temporal alignments across spatially similar observations of the data via a fast GPU friendly algorithm.

Résumé

Les dispositifs d'imagerie cérébrale peuvent donner un aperçu de l'activité neuronale à plusieurs endroits et points dans le temps. En pratique, les études d'imagerie cérébrales sont généralement menées pour plusieurs personnes suivant le même protocole expérimental. L'inférence des régions actives du cerveau est un problème inverse mal posé qui ne peut être résolu qu'en ajoutant des hypothèses a priori sur les solutions. Plusieurs hypothèses préalables ont été poursuivies dans la littérature, comme la favorisation des solutions parcimonieuses ou la résolution du problème pour plusieurs sujets à la fois. Cependant, aucune ne profite de la géométrie spatiale du problème. Le but de cette thèse est d'exploiter au maximum les aspects multisujets, spatiaux et temporels des données de magnétoencéphalographie pour améliorer le conditionnement du problème inverse. À cette fin, nos contributions s'articulent autour de trois axes : le transport optimal (OT), la régression multi-tâches parcimonieuse et les séries temporelles. En effet, la capacité de l'OT à mesurer les disparités spatiales entre les distributions le rend très bien adapté à la comparaison et l'aggrégation des cartes d'activation neurales en fonction de leur forme et de leur emplacement sur la surface du cortex cérébral. Pour des raisons numériques, on utilise la formulation entropique du transport optimal, qui, selon nous, comporte deux pièces manquantes importantes. D'un point de vue théorique, elle n'a aucune expression analytique à ce jour, et d'un point de vue pratique, l'entropie conduit à une augmentation significative de la variance, phénomène connu sous le nom de *biais entropique*. Nous complétons ce puzzle en étudiant les Gaussiennes multivariées pour lesquelles nous découvrons une forme close de l'OT entropique et proposons des algorithmes *debiaisés* pour calculer des barycentres de transport optimal rapides et précis. Ensuite, nous définissons une pénalité multitâche basé sur l'OT et des pénalités de parcimonie pour résoudre le problème inverse pour plusieurs sujets afin de promouvoir des solutions cohérentes sur le plan spatial. Nos résultats sur des données réelles mettent en évidence les avantages de l'utilisation de l'OT comme régularisation par rapport aux pénalités de régression multitâches classiques. Enfin, nous proposons une nouvelle divergence pour comparer et moyenner des données spatio-temporelles basée sur un alignement temporel entre des observations spatialement similaires, le tout via un algorithme rapide et adapté aux GPUs.

Contents

Glossary

Bold lowercase $\mathbf{a}, \mathbf{b}$	vectors $\in \mathbb{R}^p$
$\mathbb{1}$	vector in $\mathbb{R}^p$ with all entries equal to 1
Bold uppercase $\mathbf{A}, \mathbf{B}$	matrices $\in \mathbb{R}^{p \times p}$
$\mathcal{P}(\mathcal{X})$	Probability measures on a space $\mathcal{X}$
$\mathcal{P}_2(\mathcal{X})$	Probability measures with finite second order moment
$\mathcal{M}_+(\mathcal{X})$	Set of non-negative measures on a space $\mathcal{X}$
$\mathcal{S}^p_{++}, \mathcal{S}^p_+$	positive definite (resp. semi-definite) matrices in $\mathbb{R}^{p \times p}$
$\mathcal{N}(\mathbf{a}, \mathbf{A})$	Gaussian distribution with mean $\mathbf{a}$ and covariance matrix $\mathbf{A} \in \mathcal{S}^p_+$
$\|.\|$	Euclidean norm 2 on $\mathbb{R}^p$
$\delta_x \in \mathcal{P}(\mathcal{X})$	Dirac mass at location $x \in \mathcal{X}$
Tr	Trace operator
det	Determinant operator
$\mathbf{E}_\alpha$	Expectation with respect to a distribution α
$\mathbf{V}_\alpha$	Variance with respect to a distribution α
$\Delta_p \subset \mathbb{R}^p$	Probability simplex: non-negative vectors summing to 1
$\mathrm{diag} : \mathbb{R}^p \to \mathbb{R}^{p \times p}$	Creates a diagonal matrix with the entries of a vector
$\mathbf{a} \odot \mathbf{b}, \mathbf{A}, \mathbf{B}$	Element-wise (Hadamard) product between vectors or matrices
$\frac{\mathbf{a}}{\mathbf{b}}, \frac{\mathbf{A}}{\mathbf{B}}$	Element-wise division between vectors or matrices
$\alpha \otimes \beta \in \mathcal{M}_+(\mathcal{X}^2)$	Product measure of α and β
$\mathbf{a} \otimes \mathbf{b}$	Matrix with entries $\mathbf{a}_i \mathbf{b}_j$
$\mathbf{f} \oplus \mathbf{g}$	Matrix with entries $\mathbf{f}_i + \mathbf{g}_j$
$\langle \mathbf{a}, \mathbf{b} \rangle \stackrel{\text{def}}{=} \sum_i \mathbf{a}_i \mathbf{b}_i$	Scalar product on $\mathbb{R}^p$
$\langle \mathbf{A}, \mathbf{C} \rangle \stackrel{\text{def}}{=} \sum_{i,j} \mathbf{A}_{ij} \mathbf{B}_{ij}$	Frobenius product on $\mathbb{R}^{p,p}$
$\langle \alpha, f \rangle \stackrel{\text{def}}{=} \int f \mathrm{d}\alpha$	Weak duality crochet between a measure α and a function f
$\mathbf{A}^{\frac{1}{2}}$	Matrix square root of a positive semi-definite matrix $\mathbf{A}$
Id	Identity matrix
exp and log	always defined element-wise
a-e	almost everywhere
MMD	Maximum mean discrepancy norms
KL	Kullback-Leibler divergence
OT	Optimal transport
UOT	Unbalanced optimal transport
IBP	Iterative Bregman Projections
DTW	Dynamic Time Warping

Chapter 1

Introduction

Ideally, the pursuit of any scientific endeavor starts with a sense of wonder, which, through further reasoning and research branches into a knowledge graph of coarse to fine questions. It may seem obvious that one's ability to provide answers and expand the graph depends very much on how "interesting" the matter is. But could it be the other way around ? A subject becomes "interesting" only after mastering its background leading to a – perhaps unfounded – gut feeling of being able to provide answers to its open questions. Absorbing the required amount of information to reach that state may take days, months or even years. Thus, from an optimistic perspective, anything can be interesting if you look at it long enough. For the subject at hand, we hope that after reading this introduction, "long enough" will not be too long.

1 *Why* optimal transport ?

We start off lightly by motivating optimal transport (OT) from two different perspectives. First, by illustrating its practical use in neuroimaging – which will be the main subject of Chapter 3. Second, by showing how it fits in the statistics landscape.

1.1 Through the lens of the pragmatic: brain imaging data

The purpose of functional brain imaging is to study brain activity. Consider a model of the brain surface given by a triangulated mesh of p vertices. Brain activity can be illustrated by weighting each vertex with a number that may correspond or be proportional to the intensity of the electrical current at that vertex's location.

1.1.1 Comparing neural patterns

Comparing two different activation maps (sets of weights in $\mathbb{R}_+^p$ [1]) can be done using any distance function in $\mathbb{R}^p$. Such a comparison however will not take into account the spatial disparities between the activation maps. Indeed, reducing these maps to pairs of weight vectors disregards all the information in the triangulated structure of their underlying mesh: the order of the vertices matters. Figure 1.1 shows two

[1] Activation maps can be signed vectors, this will be discussed in further detail in Chapter 3.

examples. Keeping in mind that the goal of brain imaging is to highlight the function of individual brain regions, the comparison of the pair **(a)** must take into account the physical distance between the active regions. Provided such measurements, a distance between this pair of maps could simply correspond to the geodesic between their vertices with maximum intensity. This idea however, is not easy to generalize to complex neural patterns (Figure 1.1 **(b)**). Lifting this geodesic to compare such maps is precisely the goal of optimal transport.

Kantorovich OT This generalization requires to see the pair of intensity maps as distributions of mass that must be transported from one to the other in a way that minimizes a cost function, which, in our case, is given by the geodesic. This imposes a first important restriction: the pair of weight vectors must have non-negative entries and add up to the same total mass equal to 1, i.e, they belong to the probability simplex Δ_p. Formally, if we number the vertices from 1 to p and denote $\mathbf{x}, \mathbf{y} \in \Delta_p$, then, the Kantorovich formulation of OT for the cost function c is given by:

$$\mathrm{OT}(\mathbf{x}, \mathbf{y}) \stackrel{\text{def}}{=} \min_{\substack{\pi \in \mathbb{R}_+^{p \times p} \\ \pi \mathbb{1} = \mathbf{x}, \pi^\top \mathbb{1} = \mathbf{y}}} \sum_{i,j}^{p} c(i,j)\pi_{ij} = \langle \mathbf{C}, \pi \rangle \ , \tag{1.1}$$

where $\mathbf{C} \in \mathbb{R}^{p \times p}$ is the matrix with the general entry $\mathbf{C}_{ij} = c(i,j)$. The minimizer π is a discrete joint table with marginals equal to $\mathbf{x}$ and $\mathbf{y}$ that minimizes the transport cost $\langle \mathbf{C}, \pi \rangle$. Therefore, this cost has the same unit as $\mathbf{C}$ and can be seen as the optimal average displacement between the pair of activation maps.

Unbalanced OT The formulation (1.1) can be useful as a validation metric in simulations where activation maps are projected onto the simplex beforehand. However, OT cannot a priori be used to compare the activation maps of two different individuals or different time points: the difference in the overall amplitudes of the activation maps matters. Comparing weight vectors with *unbalanced* masses can be done by relaxing the marginal constraints of (1.1) and replacing them with loose divergences that penalize their violation. Using the Kullback-Leibler as a divergence leads to *unbalanced* OT between $\mathbf{x}, \mathbf{y} \in \mathbb{R}_+^p$ (Liero, Mielke, and Savaré, 2016):

$$\mathrm{UOT}(\mathbf{x}, \mathbf{y}) \stackrel{\text{def}}{=} \min_{\pi \in {\mathbb{R}_+}^{p \times p}} \langle \mathbf{C}, \pi \rangle + \gamma \mathrm{KL}(\pi \mathbb{1} | \mathbf{x}) + \gamma \mathrm{KL}(\pi^\top \mathbb{1} | \mathbf{y}) \ , \tag{1.2}$$

where $\gamma > 0$ is a hyperparameter that controls mass displacement. When γ is small, the marginals of π can be very far from $\mathbf{x}$ and $\mathbf{y}$, thus very little mass is transported. In practice, it should be set relatively to the values of $\mathbf{C}$. Going beyond $\|\mathbf{C}\|_\infty$ leads in practice to transportation plans π almost indistinguishable from each other.

1.1.2 Averaging neural patterns

To understand the function of the healthy Human brain, neuroimaging studies are usually conducted for a large group of subjects that undergo the same experimental protocol. Synthesizing the results of such

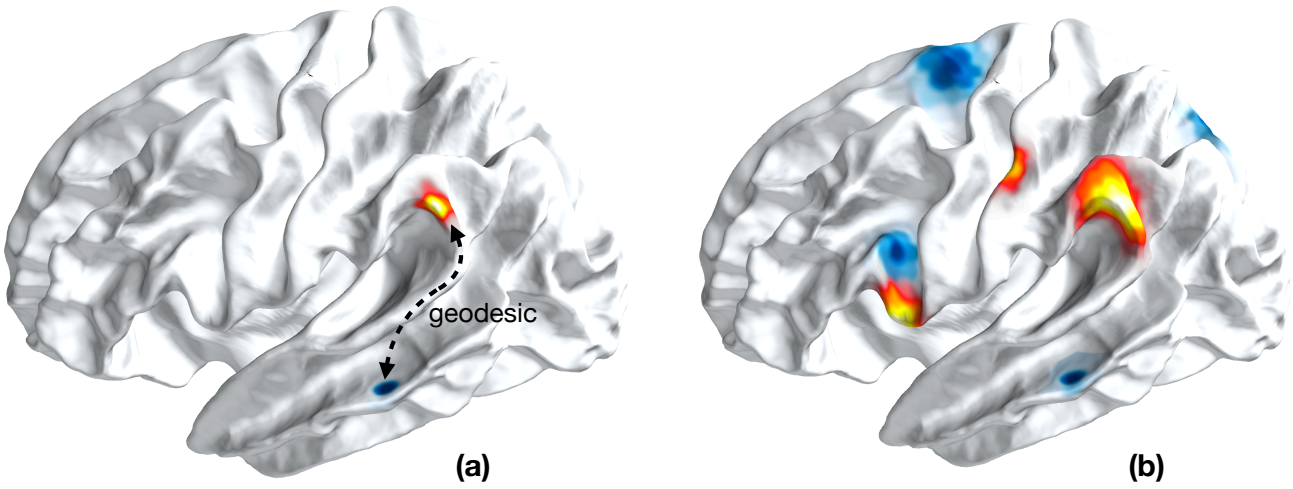


Fig. 1.1. Examples of pairs of brain activation maps. While it easy and intuitive to compare the pair of mono-atomic maps **(a)** by computing the geodesic between their locations, computing such a distance for the pair **(b)** is not as obvious.

studies require a method of aggregating the multiple brain maps. Usually, individual brain anatomies are mapped to a common "brain template" by matching the similar brain convolution patterns[2] to each other. Now that the resulting maps are defined on the same anatomy, any Fréchet mean can used to define the average functional brain (Gramfort, Peyré, and Cuturi, 2015).

Given K activation maps $\mathbf{x}_1, \dots, \mathbf{x}_K$ and a loss function F, their F-Fréchet mean is defined by:

$$\arg\min_{\mathbf{x}} \frac{1}{K} \sum_{k=1}^{K} F(\mathbf{x}, \mathbf{x}_k) \tag{1.3}$$

The most straightforward way of averaging brain maps is undoubtedly via the Euclidean mean, i.e taking $F(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|^2$. However, even when performing the same cognitive task, functional variability across individuals will prevent the different activation maps from perfectly overlapping: *functionally* identical regions are not necessarily *spatially* identical (Poline et al., 2010; Allena et al., 2012). Averaging these maps inevitably leads to a blurred mean. Figure 1.2 compares Fréchet means (a.k.a barycenters) obtained with the quadratic loss and with UOT: leveraging the *ground metric* given by the geodesic is crucial to obtain meaningful averages.

[2] gyri and sulci

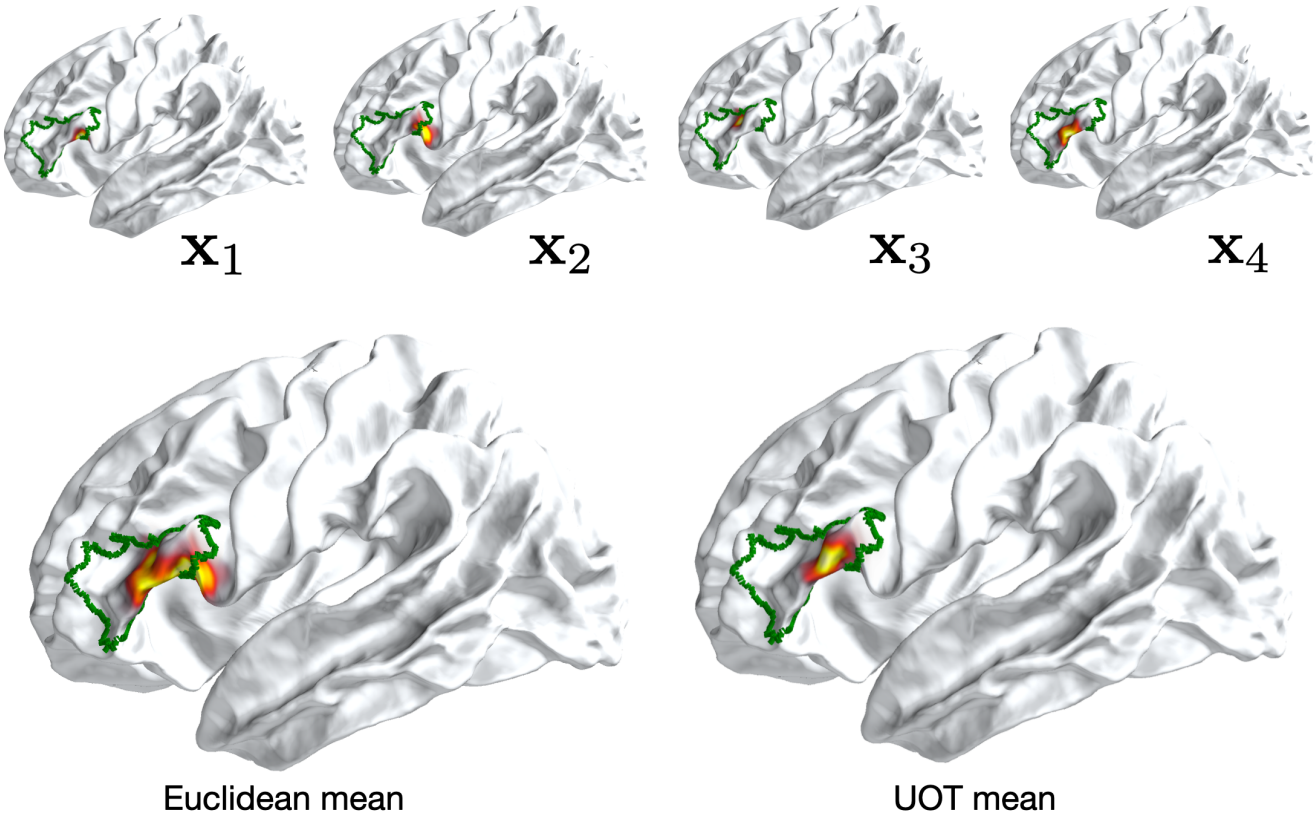


Fig. 1.2. Euclidean and UOT barycenters of 4 simulated activation maps. UOT does not suffer from the averaging blurring artifact.

1.2 Through the lens of the statistician and the geometer

The "geometrical awareness" of OT methods discussed above are possible because we consider the activation maps as distributions over the triangulated mesh of the cortex. So far, we have assumed that the vertices of this mesh are fixed for all activation maps, meaning that they are defined on the same fixed support. This assumption allows for simpler and faster algorithms that operate only on the weights of these measures. However, the theoretical study of OT requires us to let go of this assumption and study OT as a way to compare probability measures with potentially different supports.

1.2.1 f-Divergences

Comparing probability measures on a space $\mathcal{X}$ is a building block of statistics and machine learning models. This role is played by several tools such as the Kullback-Leibler or Total Variation. These functions belong to the larger family of Csiszár-divergences first introduced by Rényi (1961) and later

studied by Csiszár (1963). They can be defined on the set of arbitrary non-negative measures $\mathcal{M}_+(\mathcal{X})$. Csiszár-divergences are also known in the literature as f-divergences, as they are defined through an *entropy* function f.

Definition 1 (f-divergence) *Let $f : \mathbb{R} \to \bar{\mathbb{R}}_+$ be a convex and lower semi-continuous function such that $f(1) = 0$ and $f(\mathbb{R}_-^\star) = +\infty$. Define the constant $f_\infty \stackrel{\text{def}}{=} \lim_{p\to+\infty} \frac{f(p)}{p}$. Adopting the convention $+\infty \times 0 = 0$, the Csiszár divergence associated to f, commonly referred to as f-divergence, is defined on the set of non-negative measures $\mathcal{M}_+(\mathcal{X})$ as:*

$$D_f(\alpha, \beta) \stackrel{\text{def}}{=} \int_{\mathcal{X}} f\left(\frac{\mathrm{d}\alpha}{\mathrm{d}\beta}\right) \mathrm{d}\beta + f_\infty \int_{\mathcal{X}} \mathrm{d}\alpha^\perp \ , \tag{1.4}$$

where $\alpha^\perp$ is the singular component of the Lebesgue decomposition $\alpha = \frac{\mathrm{d}\alpha}{\mathrm{d}\beta}\beta + \alpha^\perp$.

When α admits a Lebesgue density with respect to β, the singular component $\alpha^\perp$ is equal to 0. Thus, the second term in (1.4) disappears. Table 1.1 displays a few examples of Csiszár divergences with their associated entropy functions f. One of the most appealing feature of this family of divergences is their simple formulation with a linear computational cost. However, they are bound to a very limited range of applications due to two major limitations:

1. The Lebesgue decomposition formulation breaks their continuity with respect to a positional displacement of an atom in their support.
2. Even in the case of absolutely continuous measures, the densities of their inputs are compared point-wise thereby neglecting any underlying geometry of $\mathcal{X}$.

More examples and properties of Csiszár divergences can be found in (Liese and Vajda, 2006).

1.2.2 MMD norms

To go beyond this "pointwise" comparison of measures, one must take into account some cross-interaction between the measures. This intuition is particularly accessible when considering a pair of discrete

Divergence	$f(p)$
Kullback-Leibler	$p\log(p) - p + 1$
Total variation	$\frac{1}{2}\lvert p - 1\rvert$
Reverse Kullback-Leibler	$-\log(p)$
Pearson $\mathcal{X}^2$-divergence	$(p-1)^2$
Hellinger distance	$2p - 4\sqrt{p} + 2$

Table 1.1: Examples of Csiszár divergences for different entropy functions.

measures $\alpha = \sum_{i=1}^{p} \alpha_i \delta_{x_i}$ and $\beta = \sum_{j=1}^{p} \beta_j \delta_{y_j}$. If their supports overlap – which is necessary for f-divergences to be well defined – $KL(\alpha, \beta)$ for instance would compare the weights on a one-to-one basis before applying a sum. Summing over all possible pairs (α_i, β_j) would not only be a more comprehensive comparison but also a possibility of including some notion of distance between the positions (x_i, y_j) as well. This inclusion is commonly referred to as "lifting the geometry" of $\mathcal{X}$. For instance, including the positions (x_i, y_j) in this computation through a set of weights $w_{ij} = K(x_i, y_j)$ for some function k leads to the formula: $\sum_{i,j} w_{ij}(\alpha_i - \beta_i)(\alpha_j - \beta_j)$. Notice that this formula does not impose any restriction on $(x_i)_i$ and $(y_j)_j$, thus, it remains well defined even if the supports of α and β are disjoint. This leads to the definition of *maximum mean discrepancy* (MMD) norms (Gretton et al., 2006) or Kernel norms:

Definition 2 (MMD norms) *Let $\mathcal{X}$ be a compact space and K a positive kernel i.e a continuous symmetric function over $\mathcal{X}^2$ such that:*

- $K(x, y) = h(x - y)$ *for some function h*
- $\|\alpha\|_K^2 \stackrel{\text{def}}{=} \int_{\mathcal{X}^2} K \mathrm{d}^2\alpha = \int_{\mathcal{X}^2} K(x, y) \mathrm{d}\alpha(x) \mathrm{d}\alpha(y) \geq 0$ *for any $\alpha \in \mathcal{M}_+(\mathcal{X})$.*

For any $\alpha, \beta \in \mathcal{M}_+(\mathcal{X})$, the MMD distance *between α and β can be defined as:*

$$\mathrm{MMD}_K(\alpha, \beta) \stackrel{\text{def}}{=} \|\alpha - \beta\|_K^2 \tag{1.5}$$

Unlike f-divergences which require the existence of the Lebesgue density $\frac{\mathrm{d}\alpha}{\mathrm{d}\beta}$, MMD norms are well defined for arbitrary measures in $\mathcal{M}_+(\mathcal{X})$. However, even though they formally lift any geometry defined through their kernel, in geometrical applications they do not produce satisfying results. For instance, taking the previous example of averaging neuroimaging data defined on a fixed anatomical support $\mathbf{x}_1, \ldots, \mathbf{x}_N$, the MMD Fréchet loss reads for weight vectors $\mathbf{a}, \mathbf{b}_1, \ldots, \mathbf{b}_N \in \mathbb{R}_+^p$ and a Kernel matrix with the entries $\mathbf{K}_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$:

$$L(\mathbf{a}) \stackrel{\text{def}}{=} \frac{1}{N} \sum_{n=1}^{N} \|\mathbf{a} - \mathbf{b}_n\|_K^2 = \frac{1}{N} \sum_{n=1}^{N} \sum_{k=1}^{N} \left(\langle \mathbf{a}, \mathbf{K}\mathbf{a} \rangle + \langle \mathbf{b}_n, \mathbf{K}\mathbf{b}_n \rangle - 2\langle \mathbf{a}, \mathbf{K}\mathbf{b}_n \rangle \right) \ . \tag{1.6}$$

As long as $\mathbf{K}$ is a positive definite matrix, L is a convex and coercive function. Canceling its gradient leads to the barycenter $\mathbf{a} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{b}_n$, which corresponds to the usual Euclidean mean independently of the choice of K: the geometry of the underlying space is totally ignored. But before making any hasty judgments and condemn MMDs altogether, perhaps taking on free supports would lead to a more "geometrically" aware barycenter ? When restricted to Dirac measures, the MMD acts as a loss on the underlying space as long as $h(0) = 0$:

$$\mathrm{MMD}_k(\delta_x, \delta_y) = K(x, x) + K(y, y) - 2K(x, y) = -2K(x, y) \ . \tag{1.7}$$

This loss can even be a distance on the feature space $\mathcal{X}$. For instance, when k is the kernel of the Energy distance: $k(x, y) = -\|x - y\|$, the MMD corresponds to the ℓ_2 norm between Diracs, for which the *average* dirac would be located at their median locations. How encouraging that may be, taking on point clouds

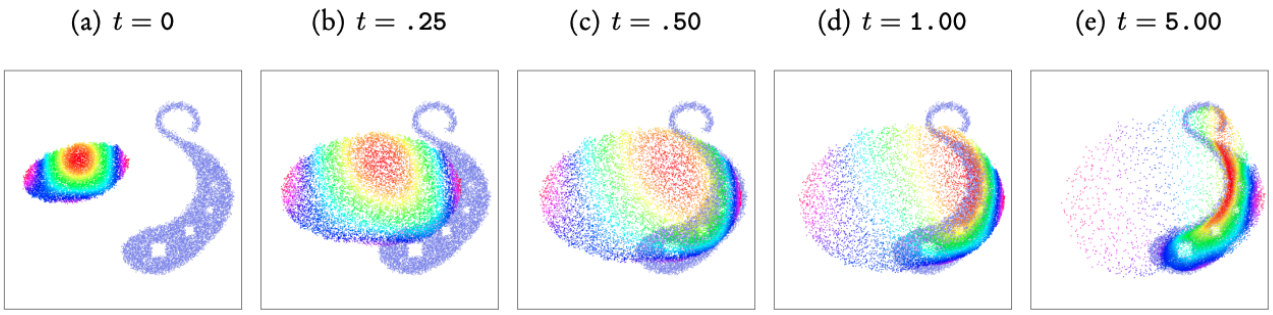


Fig. 1.3. Taken from KeOps's documentation (Charlier et al., 2020). Density fitting of the point cloud on the left to the distribution on the right using a gradient flow with the Energy distance $\mathrm{MMD}_{-\|.\|}$. Particles in the far left are scattered around far from the target distribution due to their dominant repulsive interactions with neighboring particles. The different colors are only for the visual tracking of particle trajectories.

with multiple atoms reveals a major limitation of MMDs known as *electric-field screening*. Similarly to the effect on an electric charge being dominated by interactions with neighboring particles, the MMD gradient of a single particle – when performing density fitting – numerically vanishes outside a short-range radius. Formally, given a target measure $\beta \overset{\text{def}}{=} \frac{1}{N}\sum_{i=1}^{N} \delta_{y_i}$, fitting β corresponds to minimizing over the positions of a measure $\alpha \overset{\text{def}}{=} \frac{1}{M}\sum_{i=1}^{M} \delta_{x_i}$ the quantity $\mathrm{MMD}_k(\alpha(x_1, \dots, x_M), \beta)$. With the kernel $k(x, y) = -2\|x - y\|$ for instance, assuming none of the particles overlap, the descent direct with respect to one particle x_l is given by:

$$-\nabla_{x_l}\mathrm{MMD}_k(\alpha, \beta) = 2\sum_{i \neq l} \frac{x_l - x_i}{\|x_l - x_i\|} - \sum_{j=1}^{N} \frac{x_l - y_j}{\|x_l - y_j\|} \tag{1.8}$$

Under the influence of the first sum, the particles x_l sustain a *repulsive* force that counters the *attractive* pull of β. Figure 1.3 illustrates this dampening effect: the particles in the far left are scattered around their original location.

Since we are mostly interested in comparing measures based on their overall shape, this illustration shows that the geometry of $\mathcal{X}$ intervenes "too late" in the computation of MMDs, acting merely as a weighting function. Instead of computing all-on-all interactions, perhaps this underlying geometry could *first* inform which particles interact with which ?

1.2.3 Optimal transport

If α and β have the same number of Dirac particles with uniform weights, a "good" density fitting loss function L should map each particle δ_{x_i} to its *final* destination $\delta_{y_{\sigma(i)}}$, for some assignment map $\sigma : [\![1, N]\!] \to [\![1, N]\!]$. Ideally, the performed gradient descent steps of each particle should be proportional to the distance they must travel. For instance, with a fixed step-size ω, gradients of the form: $x_i \mapsto \frac{1}{\omega}(x_i - y_{\sigma(i)})$ would lead to convergence in a single descent iteration for all the particles of α. These "ideal" gradients

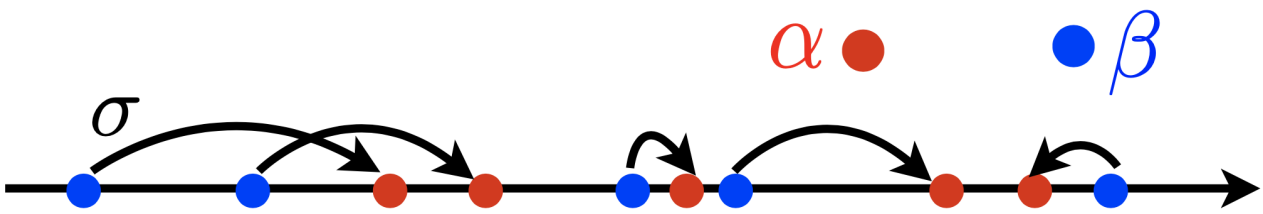


Fig. 1.4. OT on the real line corresponds to a sorting assignment σ. Taken from (Peyré and Cuturi, 2018).

can be obtained with the loss function:

$$\frac{\omega}{2}\sum_{i=1}^{N}\|x_i - y_{\sigma(i)}\|^2 \ . \tag{1.9}$$

For the sake of normalization, take $\omega = \frac{1}{N}$ and define the assignment σ as the greedy optimal permutation in the set of permutations from $[\![1, N]\!]$ to $[\![1, N]\!]$ that minimizes (1.9). The obtained loss function corresponds to the first Optimal transport distance proposed by Monge (1781):

$$\mathrm{OT}(\alpha, \beta) = \min_{\sigma \in G(N)} \frac{1}{2N}\sum_{i=1}^{N}\|x_i - y_{\sigma(i)}\|^2 \ . \tag{1.10}$$

A simple and intuitive example of σ can be retrieved in dimension 1: it corresponds to a sorting operation on the real line of the vector $y_1, \ldots, y_N$ which is illustrated in Figure 1.4. The Monge formulation of OT can thus be seen as a generalization of sorting to multi-dimensional spaces.

In practice however, measures may have different numbers of atoms (non-parametric statistics), with potentially non-uniform weights (functional brain maps):

$$\alpha = \sum_{i=1}^{N} a_i \delta_{x_i} \qquad \beta = \sum_{j=1}^{M} b_j \delta_{y_j} \tag{1.11}$$

In such settings, an assignment function may not exist. A more inclusive formulation of OT consists in seeing the measures not as "particles" to be assigned but as a "volume of fluid" to be transported: an individual mass a_i is not merely transferred to a different location but is *split* and moved across to *fill* multiple target locations. This "non-deterministic" transportation plan can be given by a matrix $\pi \in \mathbb{R}^{N \times M}$ such that $\pi_{i,j}$ corresponds to the fraction of mass transported from $a_i \delta_{x_i}$ to $b_j \delta_{y_j}$. Thus, to guarantee a full transportation, π must verify: $\pi \mathbb{1} = a$ and $\pi^\top \mathbb{1} = b$. Formally, this generalized

formulation of OT corresponds to the problem, introduced by Kantorovich (1942):

$$\mathrm{OT}(\alpha,\beta) = \min_{\substack{\pi\in\mathbb{R}_+^{N\times M}\\ \pi\mathbb{1}=a,\pi^\top\mathbb{1}=b}} \frac{1}{2}\sum_{i=1}^{N}\|x_i-y_j\|^2\pi_{ij} \ . \tag{1.12}$$

Since α and β are probability measures, the constraint set of (1.12) makes π a joint table with marginals α and β. A straightforward generalization to generic probability measures with an arbitrary symmetric cost function $C : \mathcal{X}^2 \to \mathbb{R}_+$ seeks a coupling $\pi \in \mathcal{P}(\mathcal{X}^2)$ with marginals $\pi_1 = \alpha$ and $\pi_2 = \beta$:

$$\mathrm{OT}(\alpha,\beta) = \min_{\substack{\pi\in\mathcal{P}(\mathcal{X}^2)\\ \pi_1=\alpha,\pi_2=\beta}} \int_{\mathcal{X}^2} C\mathrm{d}\pi \ . \tag{1.13}$$

In particular, the cost function $c(x,y) = \|x-y\|^p$ defines the Wasserstein distance of order p:

$$\mathcal{W}_p^p(\alpha,\beta) = \min_{\substack{\pi\in\mathcal{P}(\mathcal{X}^2)\\ \pi_1=\alpha,\pi_2=\beta}} \int_{\mathcal{X}^2} \|x-y\|^p\mathrm{d}\pi(x,y) \ . \tag{1.14}$$

Examples of discrete and continuous transport plans π are illustrated in Figure 1.5. Notice that both formulations of Eqs (1.10) and (1.12) coincide with (1.13) when restricted to their domain of definition.

1.2.4 Statistical and computational complexity

Unlike MMDs, OT gradients do not fade for long-range distances. Moreover, they "lift" the geometry of $\mathcal{X}$ to compare distributions by optimizing "mass transportation" which accounts for the overall shape of the measures. However, these appealing properties are listed with a price tag that is not affordable for most statisticians and machine learning practitioners.

Computational complexity Consider the two discrete measures α, β defined in (1.11). For the sake of simplicity, assume that $N = M$. In practice, the number of atoms N may correspond to the number of bins of a histogram, the number of vertices of a mesh or the number of pixels of an image. As far as machine learning applications are concerned, the complexity in N is of most importance. The MMD distance $\|\alpha-\beta\|_k^2$ can be given by the closed form:

$$\|\alpha-\beta\|_k^2 = \langle a, \mathbf{K}a\rangle + \langle b, \mathbf{K}b\rangle - 2\langle a, \mathbf{K}b\rangle \tag{1.15}$$

which requires an exact number of operations given by $2N^2+3N+3 = O(N^2)$. Computing OT however requires solving the linear programming problem (1.12) which can be done using variants of the network simplex algorithm and thus has a worrisome $O(N^3\log(N))$ complexity. Reducing this complexity through regularization is crucial for most practical uses and will be the subject of Section 2.

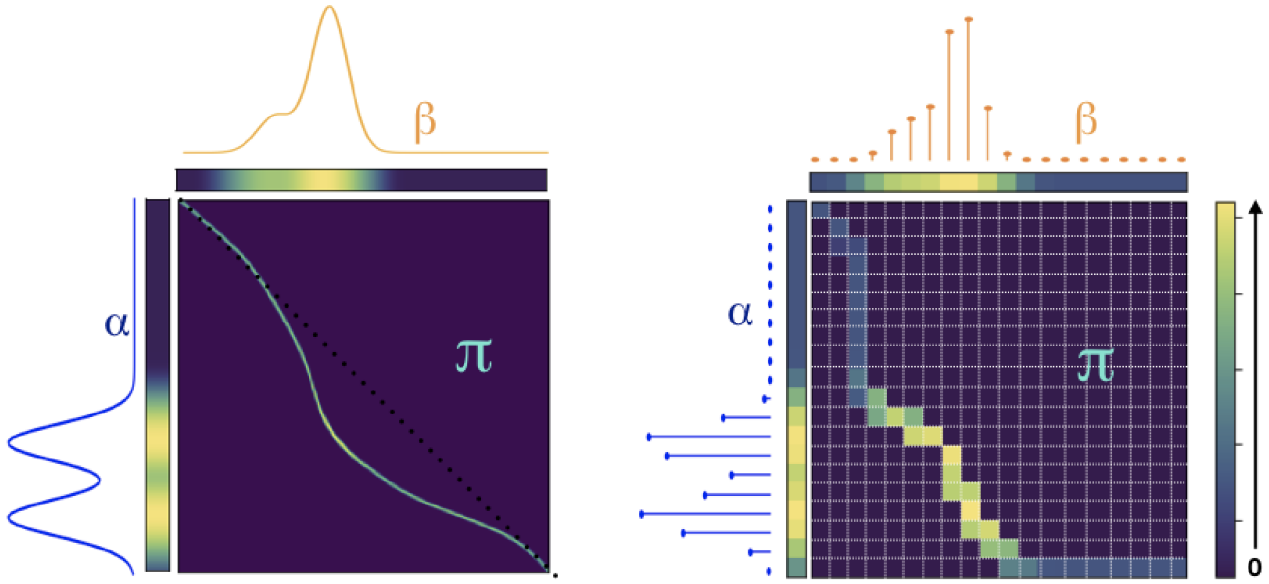


Fig. 1.5. Illustration of transportation plans for the continuous (left) and discrete (right) case. Taken from (Genevay, 2019).

Statistical complexity Consider now the general case of a pair of probability distributions $\alpha, \beta \in \mathcal{P}(\mathcal{X})$ with $\mathcal{X} \subset \mathbb{R}^d$. Comparing α and β can be done using empirical approximations $\alpha_n \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ and $\beta_n \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$ where $\mathbf{X}_1, \ldots, \mathbf{X}_n$ and $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ are i.i.d samples following α, β. A natural practical question is how many samples n are required to approximate a loss function $L(\alpha, \beta)$ using $L(\alpha_n, \beta_n)$?

On one hand, Sriperumbudur et al. (2012) showed that for MMDs, the rate of convergence is independent of the underlying dimension d:

$$\mathbb{E}\left|\,\mathrm{MMD}_k(\alpha_n, \beta_n) - \mathrm{MMD}_k(\alpha, \beta)\right| = O\left(n^{-\frac{1}{2}}\right) \quad . \tag{1.16}$$

On the other hand, OT has a catastrophic rate that decays exponentially slowly as the dimension grows. Consider OT with the cost function $C(x, y) = \|x - y\|^p$ and $d > 2$. Dudley (1969) showed that for $p = 1$:

$$\mathbb{E}\left|\,\mathrm{OT}(\alpha_n, \beta_n) - \mathrm{OT}(\alpha, \beta)\right| = O\left(n^{-\frac{1}{d}}\right) \quad , \tag{1.17}$$

which was later generalized by Fournier and Guillin (2015) for $p \geq 1$. Equation (1.17) seems to prohibit the use of OT in high dimensional settings as any empirical approximation would require exponentially many samples. But perhaps one can find a better estimator than the naive plug-in $\mathrm{OT}(\alpha_n, \beta_n)$? The good news is we have an answer. The bad news is the answer itself: Niles-Weed and Rigollet (2019) showed

that for **any estimator** $\widehat{\mathrm{OT}}(\alpha_n, \beta_n)$ of $\mathrm{OT}(\alpha, \beta)$, there exists a pair of measures $\alpha, \beta \in \mathcal{P}([0,1]^d)$ such that:

$$\mathbb{E}|\widehat{\mathrm{OT}}(\alpha_n, \beta_n) - \mathrm{OT}(\alpha, \beta)| \geq O\left((n\log(n))^{-\frac{1}{d}}\right) \quad . \tag{1.18}$$

As if the cubic numerical complexity was not enough, empirical OT is bound to fail in high dimensions.

But enough with the doom and gloom: what *can* we do ? In some aspects, MMD and OT are exact opposites: one is cheap and tractable in high dimensions but not suited for geometric applications, the other is computationally and statistically costly but performs well on such tasks. Could there be a middle ground ?

2 *How* optimal transport ?

The OT literature abounds with attempts of bringing down the complexities of OT. With no pretense of completeness, these attempts can be categorized in 3 different schools of thought:

1. Cherry-picking: restricting the analysis to a subset of measures that are regular enough such as Elliptical distributions or measures supported on low-dimensional manifolds.
2. Regularizing the measures: computing OT on projections of the data. The sliced Wasserstein approach (Bonneel et al., 2015) for instance, consists in aggregating the OT values computed on 1D projections of the data.
3. Regularizing the transport plan π by adding a Tikhonov penalty that makes the OT problem (1.13) strictly convex and thus easier to solve numerically.

All our contributions revolve around approach 3: the entropic formulation of optimal transport. As we will see in the following section, it defines the long-awaited bridge between OT and MMD norms. Moreover, it fits naturally with the unbalanced formulation of OT (1.2) given with the marginal KL discrepancies. First, we discuss a few examples of approaches 1 and 2.

2.1 Cherry-pick and regularize the measures

While computing OT is not an easy problem in high dimensions, it can actually be computed in closed form for *elliptically contoured* distributions (see remark below) with the quadratic cost $c(x,y) = \|x-y\|^2$. This closed form is thus specific for the 2-Wasserstein distance ($\mathcal{W}_2$) and is known as the Bures-Wasserstein metric.

2.1.1 The Bures-Wasserstein metric

Consider two multivariate Gaussians $\alpha = \mathcal{N}(\mathbf{a}, \mathbf{A})$ and $\beta = \mathcal{N}(\mathbf{b}, \mathbf{B})$ with $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$ and $\mathbf{A}, \mathbf{B} \in \mathcal{S}_+^d$. Olkin and Pukelsheim (1982) and (Dowson and Landau, 1982) independently showed that $\mathcal{W}_2^2$ is given by:

$$\mathcal{W}_2^2(\alpha, \beta) = \|\mathbf{a} - \mathbf{b}\|^2 + \mathcal{B}^2(\mathbf{A}, \mathbf{B}) \quad , \tag{1.19}$$

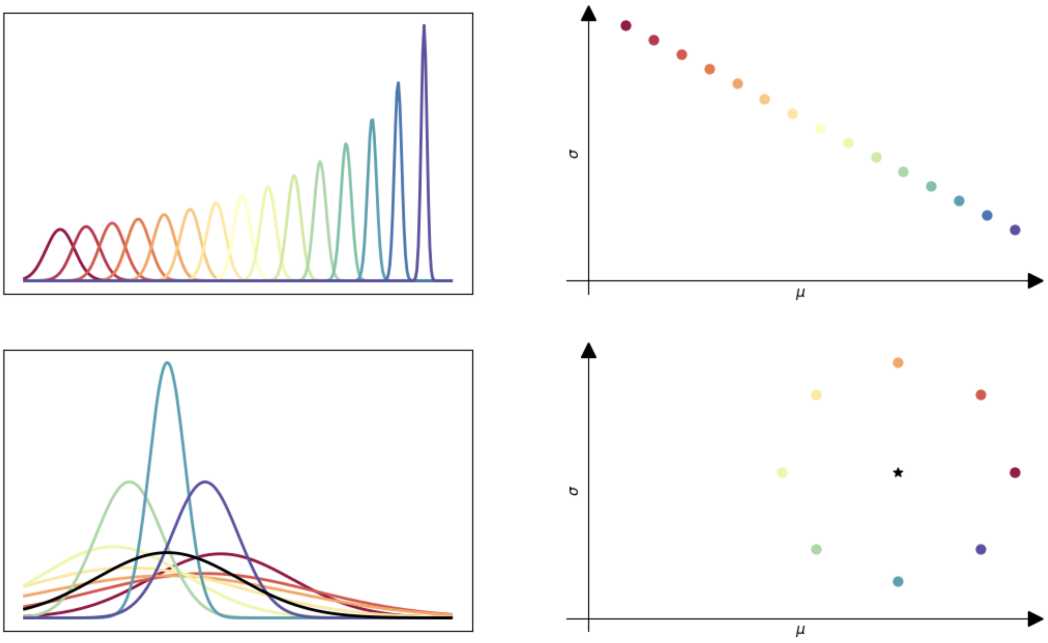


Fig. 1.6. Computing OT (W_2) between the univariate Gaussians (left) is equivalent to computing the Euclidean distance between their corresponding mappings on the (mean, standard-deviation) plane (right). The bottom row shows a set of Gaussians that are equidistant to the black Gaussian in the W_2 sense.

where:

$$\mathcal{B}^2(\mathbf{A}, \mathbf{B}) = \mathrm{Tr}(\mathbf{A}) + \mathrm{Tr}(\mathbf{B}) - 2\mathrm{Tr}((\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}})^{\frac{1}{2}}) \tag{1.20}$$

is the Bures metric on the cone of positive definite matrices (Bures, 1969). When $\mathbf{A}$ and $\mathbf{B}$ are diagonal, the Bures metric coincides with the Hellinger distance. Indeed, if $\mathbf{A} = \mathrm{diag}(\sigma_a)$ and $\mathbf{B} = \mathrm{diag}(\sigma_b)$, then $\mathcal{B}^2(\mathbf{A}, \mathbf{B}) = \|\sqrt{\sigma}_a - \sqrt{\sigma}_b\|_2^2$ where $\sqrt{.}$ on vectors is applied element-wise. Thus, for univariate Gaussians, the $\mathcal{W}_2$ corresponds to the Euclidean distance on the plane (mean, standard deviation), illustrated in Figure 1.6.

Remark 1 *The closed form* (1.19) *goes beyond Gaussian measures and can be extended to elliptical distributions (Gelbrich, 1990). Their name comes from the fact that they include distributions with a density function that has elliptical level sets. Formally, they can be characterized via a location and scale parameters* $\mathbf{m} \in \mathbb{R}^d$ *and* $\mathbf{S} \in \mathcal{S}_+^d$ *and can be transformed from one to the other via a linear transformation* $x \mapsto Ax + b$ *where* A *is positive definite.*

The Bures-Wasserstein not only provides a formula of OT for Elliptical distributions but it also gives a lower bound for all probability measures with a second order moment. Dowson and Landau (1982)

showed that for any $\alpha, \beta \in \mathcal{P}_2(\mathbb{R}^d)$ with respective mean and variance $\mathbf{a}, \mathbf{b}$ and $\mathbf{A}, \mathbf{B}$:

$$\|\mathbf{a} - \mathbf{b}\|^2 + \mathcal{B}^2(\mathbf{A}, \mathbf{B}) \leq \mathcal{W}_2^2(\alpha, \beta) \tag{1.21}$$

A simple upper bound can be derived by noticing that the independent coupling $\pi_0 = \alpha \otimes \beta$ has marginals α and β. Therefore, by definition of min, computing the OT loss with π_0 provides the upper bound:

$$\begin{aligned}
\mathcal{W}_2^2(\alpha, \beta) &\leq \int_{\mathbb{R}^p} \|x - y\|^2 \mathrm{d}\alpha(x)\mathrm{d}\beta(y) && (1.22)\\
&= \int_{\mathbb{R}^p} \|x\|^2 \mathrm{d}\alpha(x) + \int_{\mathbb{R}^p} \|y\|^2 \mathrm{d}\beta(y) - 2\int_{\mathbb{R}^p} xy \mathrm{d}\alpha(x)\mathrm{d}\beta(y) && (1.23)\\
&= \mathbf{E}_\alpha(X^2) + \mathbf{E}_\beta(X^2) - 2\mathbf{E}_\alpha(X)\mathbf{E}_\beta(X) && (1.24)\\
&= \mathbf{V}(\alpha) + \mathbf{V}(\beta) + \|\mathbf{E}_\alpha(X) - \mathbf{E}_\beta(X)\|^2 && (1.25)\\
&= \mathrm{Tr}(\mathbf{A}) + \mathrm{Tr}(\mathbf{B}) + \|\mathbf{a} - \mathbf{b}\|^2 && (1.26)
\end{aligned}$$

2.1.2 Low-dimensional projections

Another take at OT's curse of dimensionality is to consider projections on low-dimensional subspaces. While data in machine learning may be high dimensional, it has more often than not some – a priori unknown – low-dimensional structure. Instead of computing $\mathrm{OT}(\alpha, \beta)$ on the whole space $\mathbb{R}^d$, one could hope to find *the best* k-dimensional subspace on which the projections of α and β are most different. Formally, denoting the orthogonal projection of α on $E \subset \mathbb{R}^d$ by $P_{E\#}\alpha$, this quantity reads:

$$\mathrm{OT}_k(\alpha, \beta) = \sup_{\substack{E \subset \mathbb{R}^d \\ \dim(E) = k}} \mathrm{OT}(P_{E\#}\alpha, P_{E\#}\beta) \ , \tag{1.27}$$

which can be approximated by the empirical plug-in estimator $\mathrm{OT}_k(\alpha_n, \beta_n)$.

Numerical computation In practice, an exact computation of $\widehat{\mathrm{OT}_k}(\alpha_n, \beta_n)$ is potentially intractable. It can however be approximated using random projections or convex relaxation. The former led to the proposal of sliced Wasserstein distances (Rabin et al., 2011; Bonneel et al., 2015) that set $k = 1$ and average OT values on 1D lines, which amount to several sorting operations. Paty and Cuturi (2019) proposed a convex relaxation of (1.27) by making the key observation that the minimized quantity of $\mathcal{W}_2^2$ can be written:

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 \mathrm{d}\pi(x, y) = \mathrm{Tr}(\mathcal{V}_\pi) = \sum_{l=1}^{d} \lambda_l \ , \tag{1.28}$$

where $\mathcal{V}_\pi = \int_{\mathbb{R}^d \times \mathbb{R}^d} (x - y)(x - y)^\top \mathrm{d}\pi(x, y)$ is a second-order matrix with sorted eigenvalues $\lambda_1 \geq \cdots \geq \lambda_d$. Truncating (1.28) to the largest k eigenvalues leads to a tractable concave-convex max-min optimization problem that can be solved using saddle point algorithms.

Sample complexity Assuming that α and β are equal everywhere except on a k-dimensional subspace $\mathcal{U} \subset \mathbb{R}^d$ with $k \ll d$, Niles-Weed and Rigollet (2019) showed that, for this projection estimator, the sample complexity bound (1.17) can be improved. Formally, for the p-Wasserstein distance with $p \in [1, 2]$:

$$\mathbb{E}|\widehat{\mathrm{OT}_k}(\alpha_n, \beta_n) - \mathrm{OT}(\alpha, \beta)| = O\left(n^{-\frac{1}{k}} + \sqrt{\frac{d \log n}{n}}\right) \ , \tag{1.29}$$

where $n^{-\frac{1}{k}}$ is the cost of estimating OT on $\mathcal{U}$ and $\sqrt{\frac{d \log n}{n}}$ is the price to pay for not knowing $\mathcal{U}$ beforehand.

2.2 Regularize the transport plan: entropic OT

Except in dimension 1 where OT can be solved via sorting – as long as the ground cost function c can be written $c(x, y) = h(x - y)$ with a convex function h (Santambrogio, 2015) –, trading a bit of optimality for speed is becoming a necessity in machine learning applications. The "rebirth" of OT in machine learning research is mostly due to the computational edge entropic OT offers. Other regularizations based on ℓ_p norms were also investigated in the literature (Lorenz, Manns, and Meyer, 2019; Blondel, Seguy, and Rolet, 2018). Even though they come with nice sparsity-enhancing features, they do not "annihilate" the non-negativity constraint of the transport plan like entropy, which is crucial to obtain a fast and GPU-friendly dual ascent algorithm.

2.2.1 Sinkhorn's algorithm all the way: balanced, unbalanced and barycenters

Balanced OT Let $\alpha \overset{\text{def}}{=} \sum_{k=1}^{N} \mathbf{a}_i \delta_{x_i}$ and $\beta \overset{\text{def}}{=} \sum_{i=1}^{M} \mathbf{b}_i \delta_{y_i}$ be discrete measures in $\mathbb{R}^d$ with $\mathbf{a} \in \Delta_N$ and $\mathbf{b} \in \Delta_M$ where $\Delta_p \overset{\text{def}}{=} \{\mathbf{x} \in \mathbb{R}_+^p, \mathbf{x}^{\mathbb{1}} = 1\}$, known as the probability simplex. Let $\mathbf{C} \in \mathbb{R}^{p \times p}$ denote the ground cost matrix given by $\mathbf{C}_{ij} = c(x_i, y_j)$. On matrices, exp and log are applied element-wise and $\langle . \rangle$ denotes the Frobenius dot product. Cuturi (2013) proposed to add a strongly convex entropy penalty:

$$\mathrm{OT}_\varepsilon(\alpha, \beta) \overset{\text{def}}{=} \min_{\substack{\pi \in \mathbb{R}_+^{p \times p} \\ \pi \mathbb{1} = \mathbf{a}, \pi^\top \mathbb{1} = \mathbf{b}}} \langle \mathbf{C}, \pi \rangle + \varepsilon \langle \pi, \log(\pi) - 1 \rangle \ , \tag{1.30}$$

where $\varepsilon > 0$ is a fixed hyperparameter. With the linear map $\mathcal{A} : \pi \in \mathbb{R}_+^{p \times p} \mapsto (\pi \mathbb{1}, \pi^\top \mathbb{1}) \in \mathbb{R}_+^p \times \mathbb{R}_+^p$, the primal problem (1.30) can be written:

$$\mathrm{OT}_\varepsilon(\alpha, \beta) = \min_{\pi \in \mathbb{R}_+^{p \times p}} R(\pi) + \iota_{(\mathbf{a}, \mathbf{b})}(\mathcal{A}(\pi)) \ , \tag{1.31}$$

where $R(\pi) = \langle \mathbf{C}, \pi \rangle + \varepsilon \langle \pi, \log(\pi) - 1 \rangle$ and $\iota_a(x) = 0$ if $a = x$ and $+\infty$ otherwise.

The dual operator of $\mathcal{A}$ for the Frobenius dot product is given by: $\mathcal{A}^\star(\mathbf{f}, \mathbf{g}) \in \mathbb{R}_+^p \times \mathbb{R}_+^p \mapsto \mathbf{f} \oplus \mathbf{g} \in \mathbb{R}_+^{p \times p}$, where $\mathbf{f} \oplus \mathbf{g}$ denotes the matrix $(\mathbf{f}_i + \mathbf{g}_j)_{ij}$. Computing the Fenchel conjugates $\mathcal{R}^\star$ and $\iota^\star$, Fenchel

duality to (1.30) leads to the equivalent dual problem:

$$\begin{aligned} \mathrm{OT}_\varepsilon(\alpha,\beta) &= \max_{\mathbf{f},\mathbf{g}\in\mathbb{R}^p} -\iota^\star_{(\mathbf{a},\mathbf{b})}(-\mathbf{f},-\mathbf{g}) - \mathcal{R}^\star(\mathcal{A}^\star(\mathbf{f},\mathbf{g})) \\ &= \max_{\mathbf{f},\mathbf{g}\in\mathbb{R}^p} \langle \mathbf{f},\mathbf{a}\rangle + \langle \mathbf{g},\mathbf{b}\rangle - \varepsilon\langle e^{\frac{\mathbf{f}\oplus\mathbf{g}}{\varepsilon}} - 1, e^{-\frac{\mathbf{C}}{\varepsilon}}\rangle \ , \end{aligned} \tag{1.32}$$

Consider the change of variable $\mathbf{u} = e^{\frac{\mathbf{a}}{\varepsilon}}$ and $\mathbf{v} = e^{\frac{\mathbf{b}}{\varepsilon}}$ and $\mathbf{K} = e^{-\frac{\mathbf{C}}{\varepsilon}}$. The dual problem is a maximization of a concave function in $\mathbf{f}$ and $\mathbf{g}$. Performing block alternative gradient ascent on (1.32) with the aforementioned change of variable reads:

$$\mathbf{u} \leftarrow \frac{\mathbf{a}}{\mathbf{K}\mathbf{v}} \qquad \mathbf{v} \leftarrow \frac{\mathbf{b}}{\mathbf{K}^\top\mathbf{u}} \ , \tag{1.33}$$

and at optimality, the primal-dual relationship leads to the transport plan:

$$\pi = \mathrm{diag}(\mathbf{u})\mathbf{K}\,\mathrm{diag}(\mathbf{v}) \tag{1.34}$$

These iterations are guaranteed to converge at a linear rate as long as $\mathbf{K}$ has positive entries (Peyré and Cuturi, 2018). Strictly speaking, the convergence rate depends on the conditioning number of $\mathbf{K}$: the better the faster. Thus, in practice, taking low values of ε slows down the convergence.

Sinkhorn as a KL projection While the interest in entropic OT from the machine learning community is fairly recent, the formulation (1.30) dates back to the Schröedinger's bridge problem also known as *entropy maximization models* (Wilson, 1969). Nowadays, its appeal is mainly due to the simple, parallelizable and GPU friendly iterations (1.33). Better known under the name Sinkhorn's algorithm (Knopp and Sinkhorn, 1967), these iterations correspond to scaling operations that must be applied to a positive matrix (entry-wise) to make it doubly stochastic. From this perspective, it corresponds to a sequence of "projections" of the matrix $\mathrm{diag}(\mathbf{u})\mathbf{K}\,\mathrm{diag}(\mathbf{v})$ that make it "fit" the marginals $\mathbf{a}$ and $\mathbf{b}$ and was previously known as the "iterative projection fitting procedure" (IPFP). Benamou et al. (2015) formalized this idea by noticing that up to the additional constant $\langle \varepsilon\mathbf{K}, \mathbb{1}\rangle = \varepsilon\sum_{ij}\mathbf{K}_{ij}$, problem (1.30) is equivalent to a "Bregman" projection with the KL divergence:

$$\mathrm{OT}_\varepsilon(\alpha,\beta) = \min_{\substack{\pi\in\mathbb{R}_+^{p\times p} \\ \pi\mathbb{1}=\alpha, \pi^\top\mathbb{1}=\beta}} \varepsilon\mathrm{KL}(\pi|\mathbf{K}) \ , \tag{1.35}$$

where $\mathrm{KL}(\mathbf{A},\mathbf{B}) \overset{\text{def}}{=} \sum_{i,j}^n \mathbf{A}_{ij}\log\left(\frac{\mathbf{A}_{ij}}{\mathbf{B}_{ij}}\right) + \mathbf{B}_{ij} - \mathbf{A}_{ij}$ for $\mathbf{A},\mathbf{B} \in \mathbb{R}_+^{p\times p}$. At first sight, the formulation (1.35) seems to provide a second geometrical interpretation of entropic OT with an intuitive understanding of Sinkhorn's algorithm which is illustrated in Figure 1.7. A second glance shows that it actually has a major numerical contribution: the IPFP algorithm can also be used to solve the *fixed support* OT barycenter problem i.e when the support of the barycenter α is known a priori. This is encountered in computer graphics for instance where the support corresponds to pixel locations of an image. Formally, consider

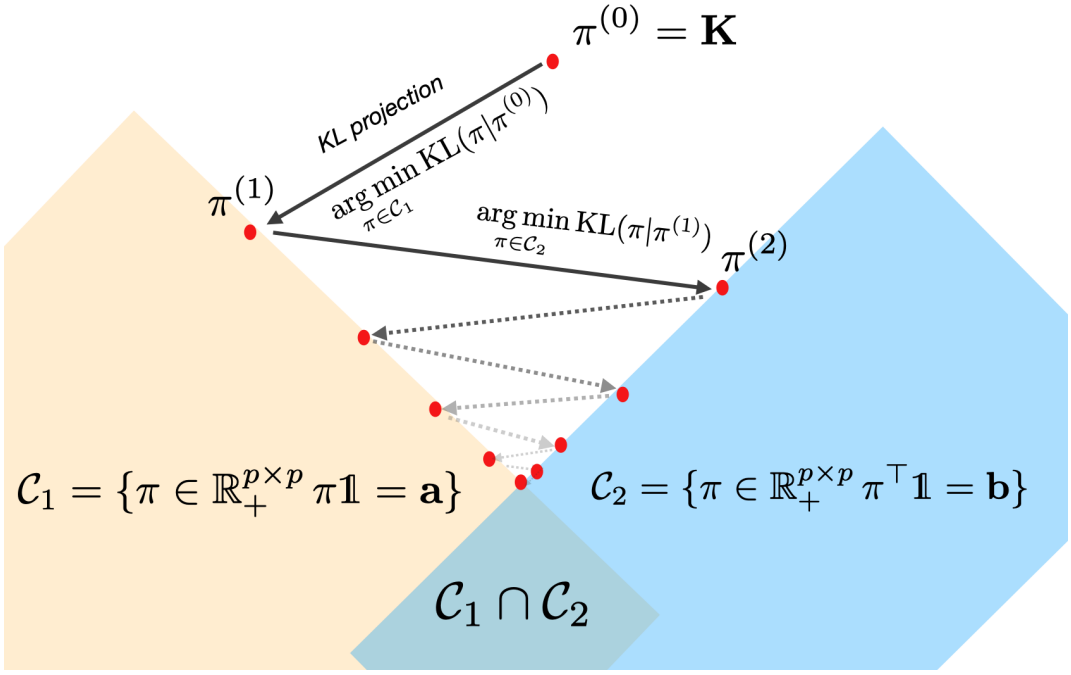


Fig. 1.7. Illustration of Sinkhorn's algorithm as an interative fitting procedure consisting in a sequence of KL projections that solve the equivalent formulation (1.35).

a sequence of discrete probability measures $\alpha_k \overset{\text{def}}{=} \sum_{i=1}^{p_k} \mathbf{a}_i^k \delta_{x_i^k}$ for $k = 1..K$ and $\alpha \overset{\text{def}}{=} \sum_{i=1}^{p} \mathbf{a}_i^k \delta_{x_i}$ with fixed and known support $x_1, \ldots, x_p$ but unknown weights $\mathbf{a}_1, \ldots, \mathbf{a}_p$. Let $\mathbf{C}_k$ denote the matrix with entries $\mathbf{C}_{kij} = c(x_i^k, x_j)$ and $\mathbf{K}_k = e^{-\frac{\mathbf{C}_k}{\varepsilon}}$. Optimization is performed with respect to the weights only and reads:

$$\min_{\mathbf{a} \in \Delta_p} \sum_{k=1}^{K} w_k \, \mathrm{OT}_\varepsilon(\alpha_k, \alpha) = \min_{\substack{\pi_1, \ldots, \pi_K \\ \pi_k \in \mathcal{C}_k \cap \mathcal{C}'}} \sum_{k=1}^{K} w_k \mathrm{KL}(\pi_k | \mathbf{K}_k) \;, \tag{1.36}$$

where $(w_k)_k \in \Delta_K$ is a fixed weight vector, $\mathcal{C}_k = \{\pi \in \mathbb{R}_+^{p_k \times p} | \pi \mathbb{1} = \mathbf{a}_k\}$ and $\mathcal{C}' = \{\pi \in \mathbb{R}_+^{p_k \times p} | \exists \mathbf{a} \in \Delta_p, \;\; \pi_k^\top \mathbb{1} = \mathbf{a}, \;\; \forall k = 1 \ldots K\}$. Solving (1.36) can be done via *Iterative Bregman projections* (IBP) which amounts to performing alternative minimization on one constraint set $\mathcal{C}$ at a time. Each step can be solved in closed form, leading to Sinkhorn-like iterations:

$$\mathbf{u}_k \leftarrow \frac{\mathbf{a}_k}{\mathbf{K}_k \mathbf{v}_k}, \qquad \mathbf{a} = \prod_{k=1}^{K} (\mathbf{K}_k^\top \mathbf{u}_k)^{w_k}, \qquad \mathbf{v}_k \leftarrow \frac{\mathbf{a}}{\mathbf{K}_k^\top \mathbf{u}_k} \;. \tag{1.37}$$

Unified entropic OT framework Perhaps all of the numerical elegance of entropic OT resides in the following unification proposed by Chizat et al. (2018b). Given a set of non-negative weights $(w_k)_k \in \Delta_K$

and a pair of convex separable scalar functions F_1 and F_2 operating on $\prod_{k=1}^K \mathbb{R}_+^{p_k}$ and $\mathbb{R}^{K\times p}$ respectively, it reads:

$$\min_{\pi\in\mathbb{R}_+^{p\times p^K}} \varepsilon\widehat{\mathrm{KL}}(\pi_1,\dots,\pi_K|\mathbf{K}_1,\dots,\mathbf{K}_K) + F_1(\pi_1\mathbb{1},\dots,\pi_K^\top\mathbb{1}) + F_2(\pi_1^\top\mathbb{1},\dots,\pi_K^\top\mathbb{1}). \tag{1.38}$$

with $\widehat{\mathrm{KL}}(\pi_1,\dots,\pi_K|\mathbf{K}^1,\dots,\mathbf{K}^K) \overset{\text{def}}{=} \sum_{k=1}^K w_k \mathrm{KL}(\pi_k|\mathbf{K}_k)$.

This extends entropic OT to the "unbalanced" setting where the transport plan π does not have the fit some input measures α, β *exactly*. Thus, α and β may be non-negative measures with different masses. While entropic OT can be recovered with $K = 1$ and $F_1(x) = \iota_{x=\alpha}$ and $F_2(x) = \iota_{x=\beta}$, the balanced barycenter problem (1.36) corresponds to the choice:

$$F_1(\pi_1\mathbb{1},\dots,\pi_K^\top\mathbb{1}) = \sum_{k=1} \iota_{\pi_k\mathbb{1}=\alpha_k} \tag{1.39}$$

$$F_2(\pi_1^\top\mathbb{1},\dots,\pi_K^\top\mathbb{1}) = \min_{\alpha\in\Delta_p} \sum_{k=1}^K \iota_{\pi_k^\top\mathbb{1}=\alpha} \tag{1.40}$$

Using Fenchel-Rockafellars's duality developments similar to (1.32), Chizat et al. (2018b) showed that performing dual ascent on the dual problem corresponds to the generic alternating iterations:

$$\begin{aligned} \mathbf{u}_1,\dots,\mathbf{u}_K &\leftarrow \mathrm{proxdiv}_{F_1}(\mathcal{K}(\mathbf{v}_1,\dots,\mathbf{v}_K)) \\ \mathbf{v}_1,\dots,\mathbf{v}_K &\leftarrow \mathrm{proxdiv}_{F_2}(\mathcal{K}^\top(\mathbf{u}_1,\dots,\mathbf{u}_K)) \end{aligned} \tag{1.41}$$

where the linear operator $\mathcal{K}$ and proxdiv are defined by:

$$\mathcal{K} : \mathbb{R}^{p^K} \to \prod_{k=1}^K \mathbb{R}_+^{p_k} \tag{1.42}$$

$$(\mathbf{x}_1,\dots,\mathbf{x}_K) \mapsto (\mathbf{K}_1\mathbf{x}_1,\dots,\mathbf{K}_K\mathbf{x}_K), \tag{1.43}$$

$$\mathrm{proxdiv}_F(\mathbf{z}) = \frac{1}{\mathbf{z}} \arg\min_{\mathbf{s}} F(s) + \varepsilon\widehat{\mathrm{KL}}(\mathbf{s}|\mathbf{z}) \tag{1.44}$$

Similarly, at optimality, each transport plan π_k is given by $\mathrm{diag}(u_k)\mathbf{K}_k\,\mathrm{diag}(v_k)$.

As long as the proxdiv operator can be computed in closed form, solving entropic OT covering balanced, unbalanced and barycenters problems can be done via very simple proxdiv operations (1.41). Table 1.2 provides the expression of the proxdiv operator of some divergences F_1 and F_2. For the sake of simplicity, we only cover unbalanced OT with the KL divergence. Examples of barycenters using $F = \gamma\mathrm{KL}$ are displayed in Figure 1.8. For low values of γ, the marginal constraints are not forced, thus very little transport occurs. We refer to (Chizat et al., 2018b) for other examples such as unbalanced OT with a Total Variation discrepancy or a range constraint.

Sinkhorn's algorithm is significantly faster on regular grids In general, as long as the proxdiv operator can be computed in closed form, each iteration of Sinkhorn has a complexity of $O(Kp^2)$ where K is the

OT setting	Divergence $F_2(\mathbf{x})$	$\text{proxdiv}_{F_2}(\mathbf{x})$
Balanced OT	$\iota_{x=\mathbf{a}}$	$\frac{\mathbf{a}}{\mathbf{x}}$
Unbalanced OT	$\gamma \text{KL}(\mathbf{x}\|\mathbf{a})$	$\left(\frac{\mathbf{a}}{\mathbf{x}}\right)^{\frac{\gamma}{\gamma+\varepsilon}}$
Balanced OT barycenter	$\min_{\mathbf{a}\in\Delta_p} \sum_{k=1}^K \iota_{\mathbf{x}_k=\mathbf{a}}$	$\frac{\mathbf{a}^\star}{\mathbf{x}}$ with $\mathbf{a}^\star = \prod_{k=1}^K (\mathbf{x}_k)^{w_k}$
Unbalanced OT barycenter	$\min_{\mathbf{a}\in\Delta_p} \sum_{k=1}^K \gamma \text{KL}(\mathbf{x}_k\|\mathbf{a})$	$\left(\frac{\mathbf{a}^\star}{\mathbf{x}}\right)^{\frac{\gamma}{\gamma+\varepsilon}}$ with $\mathbf{a}^\star = \left(\frac{\sum_{k=1}^K w_k \mathbf{x}_k^{\frac{\varepsilon}{\gamma+\varepsilon}}}{\sum_{k=1}^K w_k}\right)^{\frac{\gamma+\varepsilon}{\varepsilon}}$

Table 1.2: Examples of proxdiv operators from (Chizat et al., 2018b)

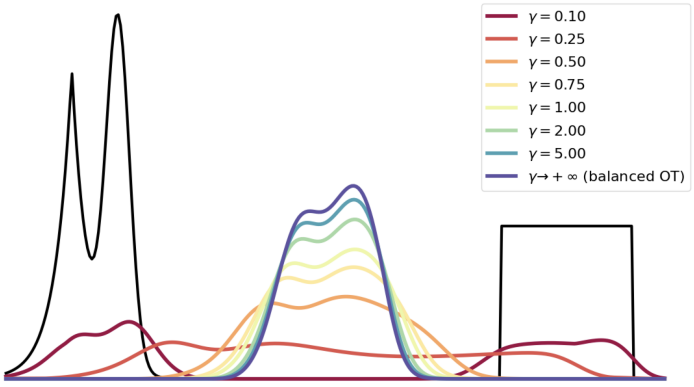


Fig. 1.8. Unbalanced barycenters of the two measures shown in black for various values of γ where F_1 and F_2 are defined as the two unbalanced KL divergences of table 1.2 respectively.

fixed number of the measures involved in the problem. This complexity can however be reduced to $O(Kp^{1+\frac{1}{d}})$ when working on regular grids of dimension d (Solomon et al., 2015) with the quadratic loss. Let's consider the simple example of images i.e $d = 2$. Assume for the sake of simplicity that the images are square with the same number of pixels equal to $p_1 = \cdots = p_K = p = m^2$. Let $\mathbf{z} \in \mathbb{R}_+^{m\times m}$ be an image with its vectorized format $\mathbf{z}' \in \mathbb{R}_+^{m^2}$. Let $1 \leq l \leq m^2$ denote a pixel with 2D coordinates $l = (l_x, l_y)$, $x, y \in [\![1, m]\!]$. Thus, the quadratic distance between two pixels l, k corresponds to: $\|l - k\|^2 = (l_x - k_x)^2 + (l_y - k_y)^2$

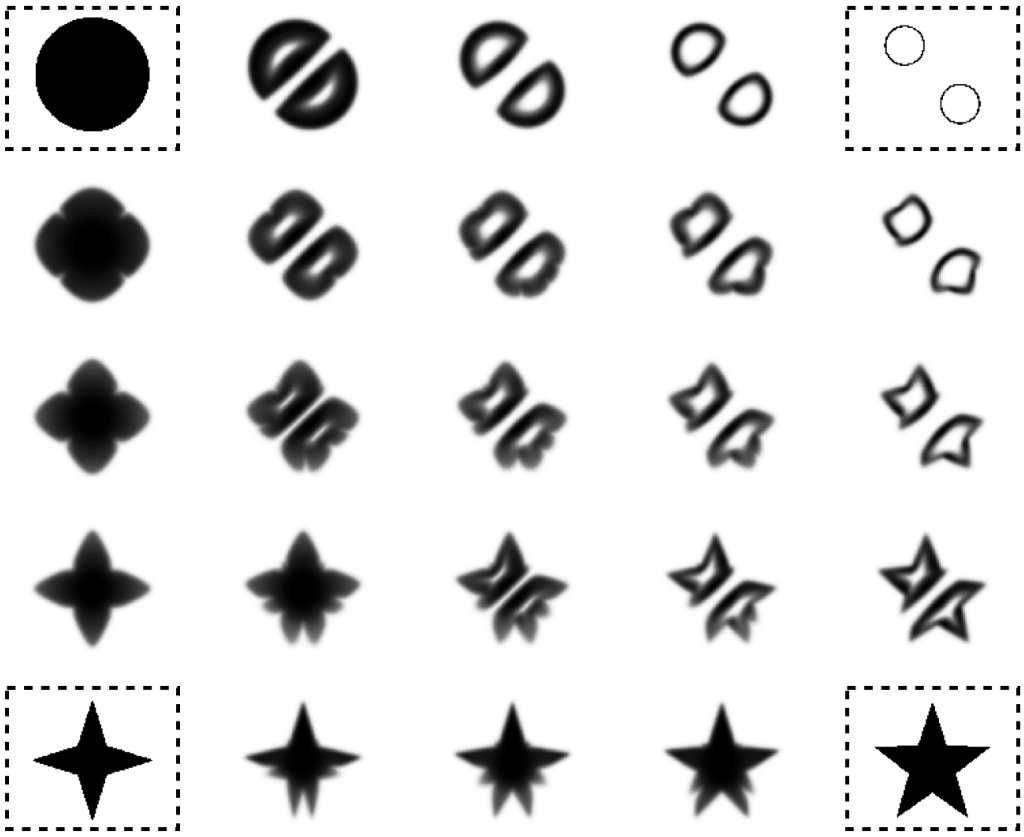


Fig. 1.9. Entropic OT interpolations (weighted balanced barycenters) of the four framed images for different sets of weights (w_k). Each image belongs to $\mathbb{R}^{p\times p}$ with $p = 400$. On a GPU, all 21 barycenters were computed in a few seconds.

and:

$$\begin{aligned}\mathcal{K}(\mathbf{z}')_k = \sum_{l=1}^{m^2} e^{-\frac{\|k-l\|^2}{\varepsilon}}\mathbf{z}'_l = \sum_{l=1}^{m^2} e^{-\frac{(l_x-k_x)^2+(l_y-k_y)^2}{\varepsilon}}\mathbf{z}'_l &= \sum_{l_x=1}^{m}\sum_{l_y=1}^{m} e^{-\frac{(l_x-k_x)^2+(l_y-k_y)^2}{\varepsilon}}\mathbf{z}_{l_x,l_y} \\ &= \sum_{l_y=1}^{m} e^{-\frac{(l_y-k_y)^2}{\varepsilon}} \sum_{l_x=1}^{m} e^{-\frac{(l_x-k_x)^2}{\varepsilon}}\mathbf{z}_{l_x,l_y} \\ &= \sum_{l_y=1}^{m} e^{-\frac{(l_y-k_y)^2}{\varepsilon}} [\mathbf{K}'\mathbf{z}]_{k_x,l_y} \\ &= [\mathbf{K}'\mathbf{z}\mathbf{K}']_{k_x,k_y} \ ,\end{aligned} \tag{1.45}$$

where $\mathbf{K}' \in \mathbb{R}_+^{m \times m}$ is the *smaller* kernel matrix with the entries $e^{-\frac{(i-j)^2}{\varepsilon}}$. Applying the operator $\mathcal{K}$ amounts to performing Gaussian convolutions along the rows and columns of $\mathbf{z}$ which has a complexity of $2m^3 = 2p^{\frac{3}{2}}$ instead of the p^2 operations of the usual matrix-vector product $\mathcal{K}(\mathbf{z}')$. The same Kernel separability trick applies to multi-dimensional data as long as the measures are defined on regular grids and the cost is quadratic. Figure 1.9 illustrates the barycenters of four images (at the corners) of size $p = (400 \times 400)$ for different interpolation weights. On a GPU, all barycenters were computed in a few seconds.

2.2.2 Entropic bias and the MMD-OT middle ground

Beyond discrete measures The definition of entropic OT given in (1.30) is specific to discrete measures as it defines the entropy function with respect to a uniform discrete measure over a finite set. Perhaps its most straightforward generalization would be that of the Lebesgue continuous case. Let $\mathcal{X} \subset \mathbb{R}^d$ be a compact space and $\alpha, \beta \in \mathcal{P}(\mathcal{X}, \mathcal{L})$ where $\mathcal{L}$ denotes the Lebesgue measure. Let c be a symmetric Lipschitz cost function over $\mathcal{X} \times \mathcal{X}$. Continuous entropic OT can be defined as:

$$\mathrm{OT}_\varepsilon^{\mathcal{L}}(\alpha, \beta) \stackrel{\text{def}}{=} \min_{\substack{\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{X}) \\ \pi_1 = \alpha, \pi_2 = \beta}} \int c \mathrm{d}\pi + \varepsilon \int \log\left(\frac{\mathrm{d}\pi}{\mathrm{d}\mathcal{L}}\right) \mathrm{d}\pi \ . \tag{1.46}$$

Identifying α, β and π with their Lebesgue densities leads to a problem that can be approximated via discrete OT computed on histograms *converging* towards those densities. Studying $\mathrm{OT}^{\mathcal{L}}$ can thus shed some light on the behavior of discrete entropic OT, as we will see in Chapter 2.

Both these formulations however do not cover instances where the measures are neither *both* discrete, nor *both* absolutely continuous. These limitations can be circumvented by noticing that as long as π has marginals α and β, its support will be included in the support of the product measure $\alpha \otimes \beta$. Formally, if $A \times B \subset \mathcal{X} \times \mathcal{X}$ is a Borel set such that $\alpha \otimes \beta(A \times B) = 0$ then $\alpha(A)\beta(B) = 0$ and thus either $\alpha(A) = 0$ or $\beta(B) = 0$. Since $A \times B \subset A \times \mathcal{X}$ and $A \times B \subset \mathcal{X} \times B$, it holds $\pi(A \times B) \leq \min(\pi(A \times \mathcal{X}), \pi(\mathcal{X} \times B)) = \min(\pi_1(A), \pi_2(B)) = \min(\alpha(A), \beta(B)) = 0$. Therefore, π is absolutely continuous with respect to $\alpha \otimes \beta$. Using the product measure as a reference, one can provide a generic definition of entropic OT:

$$\mathrm{OT}_\varepsilon^{\otimes}(\alpha, \beta) \stackrel{\text{def}}{=} \min_{\substack{\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{X}) \\ \pi_1 = \alpha, \pi_2 = \beta}} \int c \mathrm{d}\pi + \varepsilon \int \log\left(\frac{\mathrm{d}\pi}{\mathrm{d}\alpha \otimes \beta}\right) \mathrm{d}\pi \ . \tag{1.47}$$

MMD and OT interpolation The benefits of this formulation are numerous. For starters, regardless of the reference measure, when $\varepsilon \to +\infty$, OT_ε amounts to an entropy maximization leading to a $\lim_{\varepsilon \to +\infty} \pi_\varepsilon = \alpha \otimes \beta$. But when it comes to computing the limit of the OT value, $\lim_{\varepsilon \to +\infty} \mathrm{OT}_\varepsilon^{\otimes}(\alpha, \beta)$ is well defined and is given by $\int c \mathrm{d}\alpha \otimes \mathrm{d}\beta$, whereas $\lim_{\varepsilon \to +\infty} \mathrm{OT}_\varepsilon^{\mathcal{L}}(\alpha, \beta) = -\infty$. The former limit led several authors (Ramdas, Trillos, and Cuturi, 2017; Genevay, Peyre, and Cuturi, 2018; Feydy et al., 2019) to

propose the Sinkhorn divergence:

$$\mathrm{S}_\varepsilon(\alpha,\beta) = \mathrm{OT}_\varepsilon^{\otimes}(\alpha,\beta) - \frac{1}{2}(\mathrm{OT}_\varepsilon^{\otimes}(\alpha,\alpha) + \mathrm{OT}_\varepsilon^{\otimes}(\beta,\beta)) \ , \tag{1.48}$$

for which this limit becomes $\lim_{\varepsilon\to+\infty} \mathrm{S}_\varepsilon(\alpha,\beta) = \frac{1}{2}\int -c\mathrm{d}^2(\alpha-\beta)$. Thus, entropic OT interpolates between OT and an MMD distance if $-C$ is positive definite:

$$\mathrm{OT}(\alpha,\beta) \underset{\varepsilon\to 0}{\leftarrow} \mathrm{S}_\varepsilon(\alpha,\beta) \underset{\varepsilon\to+\infty}{\rightarrow} \frac{1}{2}\mathrm{MMD}_{-C}(\alpha,\beta) \tag{1.49}$$

In light of this result, could S_ε provide a middle ground in sample complexity ? Genevay et al. (2019) provides a positive answer with the complexity bound:

$$\mathbb{E}|\,\mathrm{S}_\varepsilon(\alpha_n,\beta_n) - \mathrm{S}_\varepsilon(\alpha,\beta)| = O\left(n^{-\frac{1}{2}}(\varepsilon^{-\frac{d}{2}}+1)e^{\frac{\kappa}{\varepsilon}}\right) \ , \tag{1.50}$$

where κ depends on the diameter of the compact set $\mathcal{X}$ and c. While the complexity in n is the same as the of MMDs, any practical use of (1.50) in high dimensions prohibits low values of ε. Thus, S_ε should not be seen or used as an approximation of OT, but as a well-established middle ground between OT and MMD metrics. But what properties make S_ε appropriate for machine learning or shape analysis applications ?

Properties of S_ε A well established result is the differentiability of entropic OT with gradients given by the optimal dual variables usually called in the OT theory *dual potentials*. Moreover, as long as $\mathcal{X}$ is a compact set and c induces a positive universal kernel $k(x,y) \overset{\text{def}}{=} e^{-\frac{c(x,y)}{\varepsilon}}$:

1. S_ε is non-negative: $\mathrm{S}_\varepsilon(\alpha,\beta) \geq 0, \qquad \mathrm{S}_\varepsilon(\alpha,\beta) = 0 \Leftrightarrow \alpha = \beta$.

2. S_ε is convex with respect to one of its arguments.

3. (1) leads to $\arg\min_\beta \mathrm{S}_\varepsilon(\alpha,\beta) = \alpha$. S_ε is said to be *debiased*.

4. S_ε measures the weak convergence in law: $\mathrm{S}_\varepsilon(\alpha_n,\alpha) \to 0 \Leftrightarrow \alpha_n \rightharpoonup \alpha$,

where the weak convergence is defined as:

$$\alpha_n \rightharpoonup \alpha \Leftrightarrow \int f\mathrm{d}\alpha_n \to \int f\mathrm{d}\alpha \quad \forall f \in \mathcal{C}(\mathcal{X}) \tag{1.51}$$

What about unbalanced OT? In the same fashion, entropic unbalanced OT can be defined for arbitrary non-negative measures $\mathcal{M}_+(\mathcal{X})$ as:

$$\underset{\varepsilon,\gamma}{\overset{\otimes}{\mathrm{UOT}}}(\alpha,\beta) \overset{\text{def}}{=} \inf_{\pi\in\mathcal{M}(\mathcal{X}\times\mathcal{X})} \int c\mathrm{d}\pi + \varepsilon\,\mathrm{KL}(\pi\|\alpha\otimes\beta) + \gamma\,\mathrm{KL}(\pi_1\|\alpha) + \gamma\,\mathrm{KL}(\pi_2\|\beta), \tag{1.52}$$

where $\gamma > 0$ and $\mathrm{KL}(\pi\|\alpha\otimes\beta) \overset{\text{def}}{=} \int_{\mathcal{X}} \log\left(\frac{\mathrm{d}\pi}{\mathrm{d}\alpha\mathrm{d}\beta}\right)\mathrm{d}\pi$.

To obtain similar properties for entropic unbalanced OT, a first attempt would be to consider a similar divergence:

$$(\alpha,\beta) \mapsto \underset{\varepsilon,\gamma}{\overset{\otimes}{\mathrm{UOT}}}(\alpha,\beta) - \frac{1}{2}(\underset{\varepsilon,\gamma}{\overset{\otimes}{\mathrm{UOT}}}(\alpha,\alpha) + \underset{\varepsilon,\gamma}{\overset{\otimes}{\mathrm{UOT}}}(\beta,\beta)) \ . \tag{1.53}$$

However, even with the positivity assumption of the kernel $k = e^{-c/\varepsilon}$, the divergence (1.53) does not verify non-negativity nor convexity which are violated when taking large mass discrepancies between the measures. To compensate them, one can add a quadratic penalty on this mass difference. The unbalanced Sinkhorn divergence proposed by Séjourné et al. (2019) reads:

$$\mathrm{S}^{\otimes}_{\varepsilon,\gamma}(\alpha,\beta) = \underset{\varepsilon,\gamma}{\overset{\otimes}{\mathrm{UOT}}}(\alpha,\beta) - \frac{1}{2}(\underset{\varepsilon,\gamma}{\overset{\otimes}{\mathrm{UOT}}}(\alpha,\alpha) + \underset{\varepsilon,\gamma}{\overset{\otimes}{\mathrm{UOT}}}(\beta,\beta)) + \frac{\varepsilon}{2}(\alpha(\mathcal{X}) - \beta(\mathcal{X}))^2 \ . \tag{1.54}$$

Similarly to the balanced case, $\mathrm{S}^{\otimes}_{\varepsilon,\gamma}$ is positive-definite and convex with respect to one if its argument. Moreover, it metrizes the convergence in law and has a sample complexity scaling with a similar dependency on n and ε to that of the bound (1.50).

2.2.3 The practitioner's dilemmas

The generic formulation $\mathrm{OT}^{\otimes}_{\varepsilon}$ is undoubtedly more principled from a theoretical point of view: it compares the entropic penalty of π relative to its maximum attained where $\pi = \alpha\otimes\beta$ and leads to the debiased divergence S_ε with all its virtuous properties. But in practice, when measures are discrete, are OT_ε and $\mathrm{OT}^{\otimes}_{\varepsilon}$ equivalent ? Does $\mathrm{OT}^{\otimes}_{\varepsilon}$ fit within the unified framework (1.38) of Chizat et al. (2018b) ?

The uniform and product measure and their Sinkhorn variations. For the sake of clarity, let us re-estate both formulations in the discrete case using a KL penalty. Let $\mathcal{X} = x_1,\dots,x_p$ be a finite set, $\alpha,\beta\in\mathcal{P}(\mathcal{X})$. On one hand, up to the additional constant $\varepsilon(\log(p)-1)$, the discrete OT discussed in (1.30) is equivalent to:

$$\mathrm{OT}^{\mathcal{U}}_{\varepsilon}(\alpha,\beta) \overset{\text{def}}{=} \min_{\substack{\pi\in\mathbb{R}_+^{p\times p}\\ \pi\mathbb{1}=\mathbf{a},\pi^\top\mathbb{1}=\mathbf{b}}} \langle\mathbf{C},\pi\rangle + \varepsilon\mathrm{KL}(\pi\|\mathcal{U}) \ , \tag{1.55}$$

where $\mathcal{U}_{\mathcal{X}^2}$ is the uniform measure over $\mathcal{X}^2$, weighting each x_i with $\frac{1}{p^2}$. Notice that in the discrete case one can always write for a feasible π:

$$\begin{aligned}\mathrm{KL}(\pi\|\alpha\otimes\beta) &= \sum_{i,j}^{p}\pi_{ij}\log\left(\frac{\pi_{ij}}{\mathbf{a}_i\mathbf{b}_j}\right)\\ &= \sum_{i,j}^{p}\pi_{ij}\log\left(\frac{\pi_{ij}}{1/p^2}\right) - \sum_{i,j}^{p}\pi_{ij}(2\log(p)+\log(\mathbf{a}_i)+\log(\mathbf{b}_j))\\ &= \mathrm{KL}(\pi\|\mathcal{U}) - 2\log(p) - \langle\log(\mathbf{a}),\mathbf{a}\rangle + \langle\log(\mathbf{b}),\mathbf{b}\rangle\\ &= \mathrm{KL}(\pi\|\mathcal{U}) - \mathrm{KL}(\mathbf{a}\|\mathcal{U}_{\mathcal{X}}) - \mathrm{KL}(\mathbf{b}\|\mathcal{U}_{\mathcal{X}})\end{aligned} \tag{1.56}$$

where we used the fact that $\mathbf{a}$ and $\mathbf{b}$ sum to 1. Thus, $\mathrm{OT}_\varepsilon^{\otimes}$ and $\mathrm{OT}_\varepsilon^{\mathcal{U}}$ are equivalent up to additive entropies of α and β:

$$\mathrm{OT}_\varepsilon^{\otimes}(\alpha,\beta) = \mathrm{OT}_\varepsilon^{\mathcal{U}}(\alpha,\beta) - \varepsilon\mathrm{KL}(\alpha\|\mathcal{U}_{\mathcal{X}}) - \varepsilon\mathrm{KL}(\beta\|\mathcal{U}_{\mathcal{X}}) \tag{1.57}$$

The dependency of this constant on α and β however induces some minor modifications to their dual problem and Sinkhorn's iterations. The equivalent dual problem of $\mathrm{OT}_\varepsilon^{\mathcal{U}}$ reads:

$$\max_{\mathbf{f},\mathbf{g}\in\mathbb{R}^p}\langle\mathbf{f},\mathbf{a}\rangle + \langle\mathbf{g},\mathbf{b}\rangle - \varepsilon\langle e^{\frac{\mathbf{f}\oplus\mathbf{g}-C}{\varepsilon}},1\rangle \ , \tag{1.58}$$

with optimality conditions given by:

$$e^{\frac{\mathbf{f}}{\varepsilon}} = \frac{\mathbf{a}}{\mathbf{K}e^{\frac{\mathbf{g}}{\varepsilon}}}, \quad e^{\frac{\mathbf{g}}{\varepsilon}} = \frac{\mathbf{b}}{\mathbf{K}^\top e^{\frac{\mathbf{f}}{\varepsilon}}}, \quad \pi = \mathrm{diag}(e^{\frac{\mathbf{f}}{\varepsilon}})\mathbf{K}\,\mathrm{diag}(e^{\frac{\mathbf{g}}{\varepsilon}}) \tag{1.59}$$

whereas the $\mathrm{OT}_\varepsilon^{\otimes}$ formulation has a slightly different dual problem:

$$\begin{aligned}\mathrm{OT}_\varepsilon^{\otimes}(\alpha,\beta) &\overset{\text{def}}{=} \min_{\substack{\pi\in\mathbb{R}_+^{p\times p}\\ \pi\mathbb{1}=\mathbf{a},\pi^\top\mathbb{1}=\mathbf{b}}} \langle\mathbf{C},\pi\rangle + \varepsilon\mathrm{KL}(\pi\|\mathbf{a}\otimes\mathbf{b})\\ &= \max_{\mathbf{f},\mathbf{g}\in\mathbb{R}^p}\langle\mathbf{f},\mathbf{a}\rangle + \langle\mathbf{g},\mathbf{b}\rangle - \varepsilon\langle e^{\frac{\mathbf{f}\oplus\mathbf{g}-C}{\varepsilon}},\mathbf{a}\otimes\mathbf{b}\rangle \ ,\end{aligned} \tag{1.60}$$

with optimality conditions given by:

$$e^{\frac{\mathbf{f}}{\varepsilon}} = \frac{1}{\mathbf{K}(\mathbf{b}\odot e^{\frac{\mathbf{g}}{\varepsilon}})}, \quad e^{\frac{\mathbf{g}}{\varepsilon}} = \frac{1}{\mathbf{K}^\top(\mathbf{a}\odot e^{\frac{\mathbf{f}}{\varepsilon}})}, \quad \pi = \mathrm{diag}(\mathbf{a})\,\mathrm{diag}(e^{\frac{\mathbf{f}}{\varepsilon}})\mathbf{K}\,\mathrm{diag}(e^{\frac{\mathbf{g}}{\varepsilon}})\,\mathrm{diag}(\mathbf{b}) \ . \tag{1.61}$$

While Sinkhorn's algorithm remains almost unchanged, the appearance of $\alpha\otimes\beta$ in the dual problem (1.60) reveals a key difference between $\mathrm{OT}_\varepsilon^{\mathcal{U}}$ and $\mathrm{OT}_\varepsilon^{\otimes}$. As a supremum of linear functions in $\mathbf{a}$ and $\mathbf{b}$, $\mathrm{OT}_\varepsilon^{\mathcal{U}}$ is **jointly** convex in $(\mathbf{a},\mathbf{b})$ whereas the product $\mathbf{a}\otimes\mathbf{b}$ in the dual $\mathrm{OT}_\varepsilon^{\otimes}$ prohibits joint convexity of $\mathrm{OT}_\varepsilon^{\otimes}$. In fact, Feydy et al., 2019 showed that $\mathrm{OT}_\varepsilon^{\otimes}$ is *concave* on the diagonal i.e $\alpha\to\mathrm{OT}^{\otimes}(\alpha,\alpha)$ is concave, which is

OT	Non-negative	Convex	Jointly convex	$\arg\min_\alpha \mathrm{OT}(\alpha,\beta)=\beta$	Sinkhorn for barycenters
$\mathrm{OT}_\varepsilon^{\mathcal{U}}$	X	✓	✓	X	✓
$\mathrm{OT}_\varepsilon^{\otimes}$	X	✓	X	X	X
S_ε	✓	✓	X	✓	✓ (chapter 2).
$\mathrm{UOT}_{\varepsilon,\gamma}^{\mathcal{U}}$	X	✓	✓	X	✓
$\mathrm{UOT}_{\varepsilon,\gamma}^{\otimes}$	X	✓	X	X	X
$\mathrm{S}_{\varepsilon,\gamma}^{\otimes}$	✓	✓	X	✓	X
$\mathrm{S}_{\varepsilon,\gamma}^{\mathcal{U}}$	✓	X	X	✓	✓ (chapter 2).

Table 1.3: Properties of different OT divergences **restricted on discrete measures with a symmetric positive semi-definite kernel matrix $\mathbf{K} \overset{\text{def}}{=} e^{-\frac{\mathbf{C}}{\varepsilon}}$.**

however useful to prove the convexity of S_ε. Moreover, barycenter problems with $\mathrm{OT}_\varepsilon^{\otimes}$ and S_ε cannot be written as a KL projection, thus the unified framework of (Chizat et al., 2018b) is lost.

Debiasing unbalanced OT Similar comparisons can be made for unbalanced OT. Debiasing UOT using the product measure ($\mathrm{S}_{\varepsilon,\gamma}^{\otimes}$) leads to – albeit interesting properties – loss functions for which barycenters cannot leverage fast GPU friendly algorithms offered by entropic regularization. For discrete measures on fixed supports, we can keep the appealing properties of Sinkhorn by defining UOT with respect to the uniform measure $\mathcal{U} \in \mathcal{P}(\mathcal{X}^2)$:

$$\underset{\varepsilon,\gamma}{\overset{\mathcal{U}}{\mathrm{UOT}}}(\alpha,\beta) \overset{\text{def}}{=} \inf_{\pi \in \mathcal{M}(\mathcal{X}\times\mathcal{X})} \int c\,\mathrm{d}\pi + \varepsilon\,\mathrm{KL}(\pi\|\mathcal{U}) + \gamma\,\mathrm{KL}(\pi_1\|\alpha) + \gamma\,\mathrm{KL}(\pi_2\|\beta), \tag{1.62}$$

and its debiased divergence:

$$\mathrm{S}_{\varepsilon,\gamma}^{\mathcal{U}}(\alpha,\beta) = \underset{\varepsilon,\gamma}{\overset{\mathcal{U}}{\mathrm{UOT}}}(\alpha,\beta) - \frac{1}{2}\left(\underset{\varepsilon,\gamma}{\overset{\mathcal{U}}{\mathrm{UOT}}}(\alpha,\alpha) + \underset{\varepsilon,\gamma}{\overset{\mathcal{U}}{\mathrm{UOT}}}(\beta,\beta)\right) \ . \tag{1.63}$$

The properties of these divergences for discrete measures are summarized in Table 1.3 and will be discussed in further detail in Chapter 2.

Numerical instability, scalability and Sinkhorn implementations Perhaps one the most notorious side effects of entropic regularization is the induced blurring of the optimal transportation plan. As ε increases, π_ε approaches the independent coupling $\alpha \otimes \beta$ which has maximum entropy and is illustrated in Figure 1.10. To tame this behavior and keep the appealing properties of OT, some applications may require small values of ε. However, when $\varepsilon \to 0$, most entries of the kernel $\mathbf{K} \overset{\text{def}}{=} e^{-\frac{\mathbf{c}}{\varepsilon}}$ vanish leading to numerical errors when dividing by $\mathbf{Ku}$ and $\mathbf{Kv}$. At the expense of losing parallelization, various Sinkhorn "stabilized" implementations that "absorb" large values of $\mathbf{u}$ and $\mathbf{v}$ in log-domain or that are computed

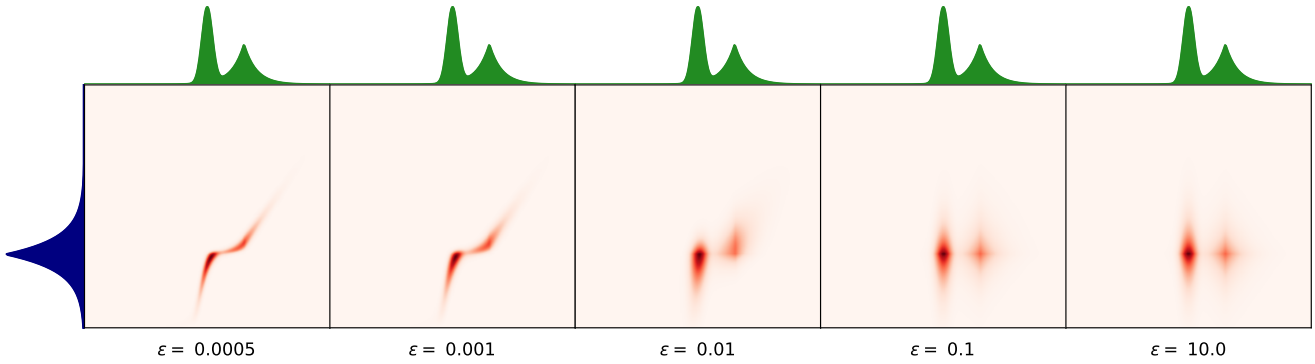


Fig. 1.10. Entropic blur of the transportation plan as ε increases.

entirely in log-domain using logsumexp routines are discussed in (Schmitzer, 2016) along with other multiscale procedures. Interested practitioners can find these Sinkhorn variants in the Python library POT (Flamary and Courty, 2017).

GPUs were the magic ingredient that brought back computational OT under the radar of applied mathematicians. While Sinkhorn's iterations may be simple and fast on GPUs, they require storing the ground cost matrix $\mathbf{C} \in \mathbb{R}_+^{p \times p}$ in memory which can be problematic as soon as p reaches a few thousands. This scalability limitation can be overcome by computing $c(x, y)$ *on the fly* when applying the logsumexp routines on non-tensorized data. This requires significant and non-trivial low level CUDA modifications, which, fortunately for everyone, is offered on a silver plater in the `KeOps` Python library (Charlier et al., 2020) with a subsequent package specific for geometric loss functions named `GeomLoss`[3](Feydy et al., 2019). With GeomLoss, computing entropic OT between millions of samples is no burden. For a comprehensive overview of shape analysis tools in geometry and all Sinkhorn's various implementations, we cannot recommend Jean Feydy's PhD manuscript highly enough (Feydy, 2020).

3 Outline and contributions

After having established all necessary background knowledge, we can now state our contributions which lay at the intersection of optimal transport, brain imaging and inverse problems. Our main purpose is to use OT to build a spatial prior P in a regularized setting of the form:

$$\min_{\mathbf{x}} L(\mathbf{x}) + \mu P(\mathbf{x}) \ , \tag{1.64}$$

where L is a data fidelity term and $\mu > 0$ a fixed hyperparameter.

Minimizing entropic OT losses however induce a bias in the minimizer called in the OT literature *entropic bias*. It can be defined as the simple case of the 1-measure barycenter: $\arg\min_\alpha \mathrm{OT}_\varepsilon(\alpha, \beta) \neq \beta$. One

[3] http://www.kernel-operations.io/geomloss/

of the virtues of S_ε is the lack of such bias at the expense of loosing Sinkhorn's algorithm for barycenters and joint convexity. From the practical perspective of problem (1.64), should we attempt to debias OT first or use the off-the-shelf unified framework of Chizat and counter entropy's blur with additional penalties ?

Chapter 2: Entropic optimal transport This chapter has two major contributions:

1. *Entropic OT for Gaussians.* Before providing a practical answer to the aforementioned question, it is crucial to understand what *exactly is* entropic bias. Doing so for arbitrary measures is not an easy task, so we turn our focus to multivariate Gaussians. This endeavor requires generalizing the convexity and differentiability results of entropic OT to measures with non-compact supports. We uncover a closed form of entropic OT similar to the Wasserstein-Bures metric. This closed form can be generalized to *Unbalanced* Gaussians i.e non-normalized Gaussians with an arbitrary mass. These closed forms provide the first test-case for theoretical conjectures of entropic OT and can serve as algorithmic benchmarks for stochastic Sinkhorn algorithms. To quantify the entropic bias for $\mathrm{OT}_\varepsilon^{\mathcal{L}}$, $\mathrm{OT}_\varepsilon^{\otimes}$ and S_ε, we characterize OT barycenters of multivariate Gaussians. We show that (1) $\mathrm{OT}_\varepsilon^{\mathcal{U}}$ / $\mathrm{OT}_\varepsilon^{\mathcal{L}}$ induces a blurring bias (increased variance), (2) $\mathrm{OT}_\varepsilon^{\otimes}$ produces a shrunk barycenter (decreased variance) and (3) S_ε has (almost) no bias.

2. *Algorithms for debiased balanced and unbalanced barycenters*. While it is straightforward to use the IBP algorithm to compute barycenters wih $\mathrm{OT}_\varepsilon^{\mathcal{U}}$, doing the same for the other divergences is not trivial. We propose a reweighted scheme to compute the barycenter of $\mathrm{OT}_\varepsilon^{\otimes}$ and a fast Sinkhorn-like algorithm to compute the debiased barycenter with S_ε. Finally, we discuss alternatives to the debiased unbalanced divergence $S_{\varepsilon,\gamma}$ to compute debiased unbalanced barycenters using Sinkhorn-like iterations.

Related publications:

- H. Janati et al, *Debiased Sinkhorn barycenters*, ICML'20.
- H. Janati et al, *Entropic OT between Gaussians has a closed form*, NeurIPS'20.

Chapter 3: Multi-task regression with an OT prior Armed with the necessary entropic OT knowledge, we can now have a take at (1.64) in the context of inverse brain imaging. This problem corresponds to locating neural sources given electro-magnetic measurements outside the head. Formally, it is tantamount to an ill-conditioned linear inverse problem. Our goal is to inform the model with spatial information by solving it jointly for multiple healthy individuals – referred to as *subjects*. The prior P acts a binder across the subjects leading the solution towards more spatially coherent neural patterns. Starting with the celebrated Group Lasso, several models based on block-sparsity norms are discussed and compared with

our OT based model. Our proposal is *aware* of the geometry of the cortex making it less prone to produce outliers. In practice, we show how this problem can be solved using proximal coordinate descent along with Sinkhorn's algorithm to reflect both the sources' sparsity and their spatial proximity. Experiments were conducted on both synthetic and real data and confronted to other brain imaging techniques.

Related publications:

- H. Janati et al, *Wasserstein regularization for sparse multi-task regression*, AISTATS'19.
- H. Janati et al, *Minimum Wasserstein Estimates: group level EEG-MEG source imaging via optimal transport*, IPMI'19.
- H. Janati et al, *Multi-subject source imaging with sparse multi-task regression*, Neuroimage 2020.

Chapter 4: Spatio-temporal optimal transport Analyzing EEG and MEG data with no regard to the temporal information is like cracking an egg with a hammer: however successful it may be, it is not why you bought the hammer in the first place. Unlike other brain imaging technologies, EEG and MEG measure brain activity up to milliseconds. Perhaps the most straightforward extension of OT to spatio-temporal data is to consider time as an additional feature. However, this approach would neglect its chronological order. Dynamic time warping (DTW) offers a principled way to compare time series based on some pre-defined cost function while being respectful of the chronology of the data. Setting this cost function to an OT loss would theoretically align time series by matching individual time frames that are spatially similar. However, DTW has two major limitations: it is not differentiable and is blind to temporal shifts. We show that its smooth variant, soft-DTW, is in fact not only differentiable but also increasing quadratically with time shifts. Combining soft-DTW and an entropy-bias-free formulation of UOT, we define a loss for spatio-temporal data and propose an off-the-shelf method to compute spatio-temporal barycenters.

Related publications:

- H. Janati et al, *Spatio-temporal alignments: optimal transport in space and time*, AISTATS'20.
- H. Janati et al, *Optimal transport barycenters for spatio-temporal data*, Submitted.

Chapter 2

Entropic Optimal transport

Let $\mathcal{X}$ be a Polish space and C a non-negative cost function on $\mathcal{X} \times \mathcal{X}$ such that $\mathrm{C}(x, y) = 0 \Leftrightarrow x = y$. Let m_1, m_2 be non-negative measures in $\mathcal{M}_+(\mathcal{X})$ such that $\alpha \ll m_1$ and $\beta \ll m_2$. Including both balanced and unbalanced settings, entropy regularized OT between $\alpha, \beta \in \mathcal{M}_+(\mathcal{X})$ with the reference measures $m_1, m_2 \in \mathcal{M}_+(\mathcal{X})$ and some marginal penalty function F can defined as:

$$\mathrm{OT}_\varepsilon^{F,m_1,m_2}(\alpha, \beta) \stackrel{\text{def}}{=} \min_{\pi \in \mathcal{M}_+(\mathcal{X}\times\mathcal{X})} \int_{\mathbb{R}^{d\times d}} \mathrm{C}\,\mathrm{d}\pi + \varepsilon \mathrm{KL}(\pi | m_1 \otimes m_2) + F(\pi_1|\alpha) + F(\pi_2|\beta) \ , \tag{2.1}$$

where $\varepsilon > 0$; π_1, π_2 denote the left and right marginals of π respectively; $m_1 \otimes m_2$ is the product measure of m_1 and m_2 and the relative entropy is defined as:

$$\mathrm{KL}(\pi | m_1 \otimes m_2) \stackrel{\text{def}}{=} \int_{\mathcal{X}\times\mathcal{X}} \log\left(\frac{\mathrm{d}\pi}{\mathrm{d}(m_1 \otimes m_2)}\right) \mathrm{d}m_1 \mathrm{d}m_2 + m_1(\mathcal{X})m_2(\mathcal{X}) - \pi(\mathcal{X}\times\mathcal{X}) \ . \tag{2.2}$$

Throughout this chapter, we will focus on two instances of F:

1. Balanced OT with $F(.|\alpha) = \iota_{.=\alpha}$. In this case, we use the notation $\mathrm{OT}_\varepsilon^{m_1,m_2}$.
2. Unbalanced OT with $F(.|\alpha) = \gamma \mathrm{KL}(.|\alpha)$, in which case we use the notation $\mathrm{UOT}_{\varepsilon,\gamma}^{m_1,m_2}$,

with the choices of references:

1. Counting measure: $m_1 = m_2 = \mathcal{U}$ for discrete measures.
2. Lebesgue measure: $m_1 = m_2 = \mathcal{L}$ for continuous measures.
3. Product measure: $m_1 = \alpha$, $m_2 = \beta$ for any measures; denoted by $\mathrm{OT}_\varepsilon^\otimes$ or $\mathrm{UOT}_{\varepsilon,\gamma}^\otimes$.

All the formulations above suffer from an *entropic bias* which refers to the fact that $\arg\min_\alpha \mathrm{OT}_\varepsilon^{F,m_1,m_2}(\alpha, \beta) \neq \beta$. Feydy et al. (2019) showed that in the balanced case this bias is fixed when considering the Sinkhorn divergence:

$$\mathrm{S}_\varepsilon(\alpha, \beta) = \mathrm{OT}_\varepsilon^\otimes(\alpha, \beta) - \frac{1}{2}(\mathrm{OT}_\varepsilon^\otimes(\alpha, \alpha) + \mathrm{OT}_\varepsilon^\otimes(\beta, \beta)) \ . \tag{2.3}$$

We have omitted the exponent $\otimes$ on S_ε on purpose: there is no need to define S_ε for each pair m_1, m_2 because they would be all be equal to each other. Indeed, Di Marino and Gerolin (2020) made the following key observation that characterizes the change of reference making all debiased divergences equal to each other:

$$\mathrm{OT}_\varepsilon^{m_1,m_2}(\alpha,\beta) = \mathrm{OT}_\varepsilon^{\otimes}(\alpha,\beta) + \varepsilon \mathrm{KL}(\alpha|m_1) + \varepsilon \mathrm{KL}(\beta|m_2) \ . \tag{2.4}$$

In practice, when computing OT barycenters, the obtained minimizer with OT_ε is either blurred or shrunk compared to what one would expect had ε been equal to 0. Would that be the case with S_ε ? Can we quantify this blurring / shrinking by studying a specific family of distributions ? In practice, can we compute debiased barycenters with S_ε but keep the computational advantage of Sinkhorn algorithms ? What would be the equivalent of S_ε in the unbalanced case ?

This chapter is dedicated to answering these questions as follows:

1. *OT in* $\mathbb{R}^d$: We start off this chapter by studying $\mathrm{OT}_\varepsilon^{\mathcal{L}}, \mathrm{OT}_\varepsilon^{\otimes}$ and S_ε on $\mathbb{R}^d$ with the quadratic cost and generalizing its differentiability and convexity properties previously known on measures with compact supports. Differentiability requires an additional assumption on the measures: we show that a sub-Gaussian tail is enough.

2. *OT for Gaussians*: We show that $\mathrm{OT}_\varepsilon^{F,\mathcal{L}}$ and $\mathrm{OT}_\varepsilon^{F,\otimes}$ have a closed form for Gaussian measures. These expressions provide new insights into the theoretical understanding of entropic OT and generalize the Bures-Wasserstein metric. For the balanced case, we characterize the barycenters of Gaussians for serveral OT formulations, showing that the debiased barycenter has an (almost) unaltered variance.

3. *Debiased algorithms for barycenters*. We provide fast simple Sinkhorn-like algorithms to compute debiased barycenters in both balanced and unbalanced cases for fixed support settings.

This chapter is based on:

- H. Janati et al, *Debiased Sinkhorn barycenters*, ICML'20.
- H. Janati et al, *Entropic OT between Gaussians has a closed form*, NeurIPS'20.

1 Entropic OT for measures with unbounded supports

The motivation behind this section is to use convexity and differentiability to study barycenters of Gaussians. These results were published in (Janati, Cuturi, and Gramfort, 2020a).

1.1 Convexity and differentiability of $\mathrm{OT}_\varepsilon^{\otimes}$ and S_ε

In this section, we set $\mathcal{X} = \mathbb{R}^d$ and $C(x,y) = \|x-y\|^2$. The set of continuous functions on $\mathbb{R}^d$ is denoted by $\mathcal{C}(\mathbb{R}^d)$. The set of probability measures with a second order moment is denoted by $\mathcal{P}_2(\mathbb{R}^d)$. For $\alpha \in \mathcal{P}(\mathbb{R}^d)$,

$\mathcal{L}_p(\mathbb{R}^d, \alpha)$ denotes the set of α-measurable functions $\mathbb{R}^d \to \mathbb{R}$ such that $\int |f|^p \, \mathrm{d}\alpha < +\infty$. Let $f \in \mathcal{L}_1(\mathbb{R}^d, \alpha)$, $g \in \mathcal{L}_1(\mathbb{R}^d, \beta)$ and denote $\langle \alpha, f \rangle = \int_{\mathbb{R}^d} f \, \mathrm{d}\alpha$. The tensor operators $\otimes$ and $\oplus$ denote respectively the mappings $f \otimes g : (x, y) \in \mathbb{R}^d \times \mathbb{R}^d \mapsto f(x).g(y)$ and $f \oplus g : (x, y) \in \mathbb{R}^d \times \mathbb{R}^d \mapsto f(x) + g(y)$. For the sake of convenience, we denote $\mathrm{OT}_\varepsilon^\otimes$ the generic OT formulation with the product measure as reference, formally given by:

$$\mathrm{OT}_\varepsilon^\otimes(\alpha, \beta) \stackrel{\text{def}}{=} \min_{\substack{\pi \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d) \\ \pi_{\#1} = \alpha, \pi_{\#2} = \beta}} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 \, \mathrm{d}\pi + \varepsilon \mathrm{KL}(\pi | \alpha \otimes \beta) \ , \tag{2.5}$$

We show that $\mathrm{OT}_\varepsilon^\otimes$ and S_ε are convex (w.r.t. one variable) and differentiable. Our differentiability proof is inspired from that of Feydy et al. (2019) where the compactness assumption of the space $\mathcal{X}$ is replaced with a sub-Gaussian tails assumption on the measures that allows one to apply Lebesgue's dominated convergence theorem on $\mathbb{R}^d$. The convexity proof is however novel and is solely based on the dual problem of $\mathrm{OT}_\varepsilon^\otimes$.

Dual problem Consider the Gaussian kernel $\mathrm{K}(x, y) \stackrel{\text{def}}{=} e^{-\frac{\|x-y\|^2}{\varepsilon}}$. Let $\alpha, \beta \in \mathcal{P}(\mathbb{R}^d)$. We define the linear operators on $\mathcal{K}$ and $\mathcal{K}^\top$ such that $\mathcal{K}(\mu) = \int_{\mathbb{R}^d} \mathrm{K}(x, y) \, \mathrm{d}\mu(y)$ and $\mathcal{K}^\top(\mu) = \int_{\mathbb{R}^d} K^\top(x, y) \, \mathrm{d}\mu(x)$ for any non-negative measure $\mu \in \mathcal{M}_+(\mathbb{R}^d)$. Problem (2.5) has a dual formulation given by:

$$\mathrm{OT}_\varepsilon^\otimes(\alpha, \beta) = \sup_{\substack{f \in \mathcal{L}_1(\mathbb{R}^d, \alpha) \\ g \in \mathcal{L}_1(\mathbb{R}^d, \beta)}} \int_{\mathbb{R}^d} f \, \mathrm{d}\alpha + \int_{\mathbb{R}^d} g \, \mathrm{d}\beta - \varepsilon \int_{\mathbb{R}^d \times \mathbb{R}^d} \exp\left(\frac{f \oplus g - C}{\varepsilon}\right) \mathrm{d}\alpha \, \mathrm{d}\beta + \varepsilon \ . \tag{2.6}$$

If α and β have finite second moments, (2.6) is well defined and a couple of dual potentials (f, g) are optimal if and only if they are solutions of Sinkhorn's equations (Mena and Niles-Weed, 2019):

$$\begin{aligned} e^{\frac{f}{\varepsilon}}.\mathcal{K}(e^{\frac{g}{\varepsilon}}.\beta) &= 1, \quad \alpha - a.e \ , \\ e^{\frac{g}{\varepsilon}}.\mathcal{K}^\top(e^{\frac{f}{\varepsilon}}.\alpha) &= 1, \quad \beta - a.e \ . \end{aligned} \tag{2.7}$$

and the optimal transport plan π is given by: $\pi = \exp\left(\frac{f \oplus g - C}{\varepsilon}\right).(\alpha \otimes \beta)$

Thus, at optimality the integral over $\mathbb{R}^d \times \mathbb{R}^d$ sums to 1 and:

$$\mathrm{OT}_\varepsilon^\otimes(\alpha, \beta) = \int_{\mathbb{R}^d} f \, \mathrm{d}\alpha + \int_{\mathbb{R}^d} g \, \mathrm{d}\beta \tag{2.8}$$

Symmetric terms $\mathrm{OT}_\varepsilon^\otimes(\alpha, \alpha)$ When $\alpha = \beta$, the symmetry of the problem leads to the existence of a symmetric pair of potentials (h, h). Indeed, if (f, g) is optimal (g, f) is also optimal. Moreover, since C is symmetric, the optimal transport plan π is also symmetric which leads to $f = g$. Thus the following proposition holds.

Proposition 1 *Let $\alpha \in \mathcal{P}_2(\mathbb{R}^d)$, it holds:*

$$\mathrm{OT}_\varepsilon^{\otimes}(\alpha,\alpha) = \sup_{h\in\mathcal{L}_1(\mathbb{R}^d,\alpha)} 2\int_{\mathbb{R}^d} h \,\mathrm{d}\alpha - \varepsilon \int_{\mathbb{R}^d\times\mathbb{R}^d} \exp\left(\frac{h\oplus h - C}{\varepsilon}\right) \mathrm{d}^2\alpha + \varepsilon \ , \tag{2.9}$$

Moreover, the supremum is attained at the unique (by strong concavity) autocorrelation potential $h \in \mathcal{L}_1(\mathbb{R}^d,\alpha)$ if and only if h is a solution of $e^{\frac{fh}{\varepsilon}}.\mathcal{K}(e^{\frac{h}{\varepsilon}}.\alpha) = 1, \ \ \alpha - a.e \ $, and at optimality it holds: $\frac{1}{2}\mathrm{OT}_\varepsilon^{\otimes}(\alpha,\alpha) = \int_{\mathbb{R}^d} h \,\mathrm{d}\alpha$.

Restriction on sub-Gaussians Feydy et al. (2019) showed the differentiability and convexity of $\mathrm{OT}_\varepsilon^{\otimes}$ and S_ε on measures with compact supports. On $\mathbb{R}^d$, more assumptions on α and β are required. Throughout this section we restrict $\mathrm{OT}_\varepsilon^{\otimes}$ and S_ε to the convex set of sub-Gaussian probability measures:

Assumption 1 *We restrict $\mathrm{OT}_\varepsilon^{\otimes}$ and S_ε to the set of sub-Gaussian probability measures $\mathcal{G}(\mathbb{R}^d) \overset{\text{def}}{=} \{\mu | \exists q > 0, \ \mathbb{E}_\mu(e^{\frac{\|X\|^2}{2dq^2}}) \leq 2\}$.*

Mena and Niles-Weed (2019) showed that if $\alpha,\beta \in \mathcal{G}(\mathbb{R}^d)$, there exists a pair of potentials (f,g) verifying the fixed point equations (2.7) on the whole space $\mathbb{R}^d$ that are bounded by quadratic functions. This result is key to show the differentiability of $\mathrm{OT}_\varepsilon^{\otimes}$ on $\mathcal{G}(\mathbb{R}^d)$.

Proposition 2 (Mena and Niles-Weed (2019), Prop. 6) *Let $\alpha,\beta \in \mathcal{G}(\mathbb{R}^d)$. There exists a pair of smooth functions (f,g) such that (2.7) holds on $\mathbb{R}^d$ and $\forall x,y \in \mathbb{R}^d$:*

$$\begin{aligned} -dq^2(1+\frac{1}{2}(\|x\|+\sqrt{2d}q)^2) \leq \frac{f(x)}{\varepsilon} \leq \frac{1}{2}(\|x\|+\sqrt{2d}q)^2 \\ -dq^2(1+\frac{1}{2}(\|y\|+\sqrt{2d}q)^2) \leq \frac{g(y)}{\varepsilon} \leq \frac{1}{2}(\|y\|+\sqrt{2d}q)^2 \end{aligned} \tag{2.10}$$

In the rest of this section, (f,g) denotes a pair of potentials defined by Proposition 2.

Differentiability We say that a function $F : \mathcal{G}(\mathbb{R}^d) \to \mathbb{R}$ is differentiable at α if there exists $\nabla F(\alpha) \in \mathcal{C}(\mathbb{R}^d)$ such that for any displacement $t\delta\alpha$ with $t \in [0,1]$ and $\delta\alpha = \alpha_1 - \alpha$ with $\alpha_1,\alpha \in \mathcal{G}(\mathbb{R}^d)$, and:

$$F(\alpha + t\delta\alpha) = F(\alpha) + t\langle \delta\alpha, \nabla F(\alpha)\rangle + o(t) \ , \tag{2.11}$$

where $\langle \delta\alpha, \nabla F(\alpha)\rangle = \int_{\mathbb{R}^d} \nabla F(\alpha) \,\mathrm{d}\delta\alpha$.

Proposition 3 *Let $\alpha,\beta \in \mathcal{G}(\mathbb{R}^d)$, and (f,g) their associated pair of dual potentials given by proposition 2. $\mathrm{OT}_\varepsilon^{\otimes}(\alpha,.)$ is differentiable on sub-Gaussian measures with unbounded supports and its gradient is given by:*

$$\nabla_\beta \mathrm{OT}_\varepsilon^{\otimes}(\alpha,\beta) = g \ . \tag{2.12}$$

PROOF. The proof is inspired from Feydy et al. (2019) in the case of measures with compact supports. The difference arises when taking the limit of integrals of the potentials. Thanks to assumption 1, proposition 2 provides an upper bound that allows to conclude by dominated convergence. Consider $\alpha, \beta, \alpha_1, \alpha_2, \beta_1, \beta_2 \in \mathcal{G}(\mathbb{R}^d)$ and denote the displacements $\delta\alpha = \alpha_1 - \alpha_2$ and $\delta\beta = \beta_1 - \beta_2$. Let Δ_t denote the ratio of (2.11):

$$\Delta_t = \frac{\mathrm{OT}_\varepsilon^\otimes(\alpha_t, \beta_t) - \mathrm{OT}_\varepsilon^\otimes(\alpha, \beta)}{t} \quad , \tag{2.13}$$

where $\alpha_t = \alpha + t\delta\alpha$ and $\beta_t = \beta + t\delta\beta$. Similarly to the proof of Proposition 2 of Feydy et al. (2019), we derive a lower and upper bound of Δ_t using suboptimal potentials. On one hand, the pair (f, g) is suboptimal for the dual problem defining $\mathrm{OT}_\varepsilon^\otimes(\alpha_t, \beta_t)$. Therefore:

$$\mathrm{OT}_\varepsilon^\otimes(\alpha_t, \beta_t) \geq \langle \alpha_t, f \rangle + \langle \beta_t, g \rangle - \varepsilon \langle \alpha_t \otimes \beta_t, \exp\left(\frac{f \oplus g - C}{\varepsilon}\right) \rangle + \varepsilon$$

Therefore, (2.6) and (2.7) lead to the lower bound:

$$\Delta_t \geq \langle \delta\alpha, f - \varepsilon \rangle + \langle \delta\beta, g - \varepsilon \rangle + o(1)$$

And similarly we get the upper bound:

$$\Delta_t \leq \langle \delta\alpha, f_t - \varepsilon \rangle + \langle \delta\beta, g_t - \varepsilon \rangle + o(1)$$

As $t \to 0$, $(\alpha_t, \beta) \to (\alpha, \beta)$. On one hand, Proposition 4 of Mena and Niles-Weed (2019) leads to the pointwise convergence of the sequence of potentials (f_t, g_t) towards (f, g). On the other hand, Proposition 2 implies that there exists $M > 0$ such that $|f_t(x)| \leq M\|x\|^2$ for all $x \in \mathbb{R}^d$. Given that any $\mu \in \mathcal{G}\sigma(\mathbb{R}^d)$ has a second order moment, by Lebesgue's dominated convergence we have $\langle \mu, f_t \rangle \to \langle \mu, f \rangle$. Similarly, $\langle \mu, g_t \rangle \to \langle \mu, g \rangle$. Finally, since $\langle \delta\alpha, \varepsilon \rangle = \langle \delta\beta, \varepsilon \rangle = 0$, we get as $t \to 0$, $\Delta_t \to \langle \delta\alpha, f \rangle + \langle \delta\beta, g \rangle$. Since f and g are smooth (Prop 2) and integrable with respect to any $\mu \in \mathcal{G}(\mathbb{R}^d)$, (2.11) holds for $\nabla \mathrm{OT}_\varepsilon^\otimes(\alpha, \beta) = (f, g)$. ■

The differentiability of S_ε follows immediately. Using the chain rule, 2 cancels $1/2$ and it holds:

Corollary 1 *Let $\alpha, \beta \in \mathcal{G}(\mathbb{R}^d)$, and (f, g) their associated pair of dual potentials given by proposition 2 and h_β the autocorrelation potential associated with β. $\mathrm{S}_\varepsilon^\otimes(\alpha, .)$ is differentiable on sub-Gaussian measures with unbounded supports and its gradient is given by:*

$$\nabla_\beta \mathrm{S}_\varepsilon^\otimes(\alpha, \beta) = g - h_\beta \quad . \tag{2.14}$$

Remark 2 *It is important to keep in mind that the notion of differentiability (and gradient) of the functions $\mathrm{OT}_\varepsilon^\otimes$ and S_ε differ from the usual Fréchet differentiability. Indeed, the space of probability measures $\mathcal{P}(\mathbb{R}^d)$ has an empty interior in the space of signed Radon measures $\mathcal{M}(\mathbb{R}^d)$. The definition adopted here defines derivatives along feasible directions in $\mathcal{P}(\mathbb{R}^d)$. This is however sufficient to characterize the convexity of S_ε and its stationary points (see appendix 5.2 for details).*

Convexity Now we turn to showing that S_ε is convex with respect to either one of its arguments separately. To do so, we prove the first order characterization of convexity of a differentiable function $F : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ given by:

$$F(\alpha) \geq F(\alpha') + \langle \alpha - \alpha', \nabla F(\alpha') \rangle \ , \tag{2.15}$$

As shown by the proof of the following Lemma, the positivity of K plays a key role in proving the convexity of S_ε.

Lemma 1 *Let $\alpha, \alpha' \in \mathcal{G}(\mathbb{R}^d)$ and let $h_\alpha, h_{\alpha'}$ denote their respective autocorrelation potentials given by proposition 1. Then if $\mathrm{K}(x,y) = e^{-\frac{\|x-y\|^2}{\varepsilon}}$:*

$$\int e^{\frac{h_\alpha(x)}{\varepsilon}} \mathrm{K}(x,y) e^{\frac{h_{\alpha'}(y)}{\varepsilon}} \, \mathrm{d}\alpha(x) \, \mathrm{d}\alpha'(y) \leq 1 \tag{2.16}$$

PROOF. The left side of (2.16) can be equivalently written using Fubini-Tonelli:

$$\begin{aligned} A &= \int e^{\frac{h_\alpha(x)}{\varepsilon}} \mathrm{K}(x,y) e^{\frac{h_{\alpha'}(y)}{\varepsilon}} \, \mathrm{d}\alpha(x) \, \mathrm{d}\alpha'(y) \\ &= \langle e^{\frac{h_\alpha}{\varepsilon}}.\alpha, \mathcal{K}(e^{\frac{h_{\alpha'}}{\varepsilon}}\alpha') \rangle \\ &= \langle e^{\frac{h_{\alpha'}}{\varepsilon}}.\alpha', \mathcal{K}^\top(e^{\frac{h_\alpha}{\varepsilon}}\alpha) \rangle \\ &= \langle e^{\frac{h_{\alpha'}}{\varepsilon}}.\alpha', \mathcal{K}(e^{\frac{h_\alpha}{\varepsilon}}\alpha) \rangle \ , \end{aligned}$$

where the last equality follows from the symmetry of K. Thus we have:

$$A = \frac{1}{2}\langle e^{\frac{h_{\alpha'}}{\varepsilon}}.\alpha', \mathcal{K}(e^{\frac{h_\alpha}{\varepsilon}}\alpha) \rangle + \frac{1}{2}\langle e^{\frac{h_\alpha}{\varepsilon}}.\alpha, \mathcal{K}(e^{\frac{h_{\alpha'}}{\varepsilon}}\alpha') \rangle \tag{2.17}$$

Since the optimal transport plans (primal solutions) associated with $\mathrm{OT}_\varepsilon^\otimes(\alpha,\alpha)$ and $\mathrm{OT}_\varepsilon^\otimes(\alpha',\alpha')$ integrate to 1, the right side of (2.16) can be written:

$$1 = \frac{1}{2}\langle e^{\frac{h_\alpha}{\varepsilon}}.\alpha, \mathcal{K}(e^{\frac{h_\alpha}{\varepsilon}}\alpha) \rangle + \frac{1}{2}\langle e^{\frac{h_{\alpha'}}{\varepsilon}}.\alpha', \mathcal{K}(e^{\frac{h_{\alpha'}}{\varepsilon}}\alpha') \rangle \tag{2.18}$$

Combining (2.17) with (2.18), it holds:

$$1 - A = \frac{1}{2}\langle r, \mathcal{K}(r) \rangle$$

where $r = e^{\frac{h_\alpha}{\varepsilon}}.\alpha - e^{\frac{h_{\alpha'}}{\varepsilon}}.\alpha'$. Since K is semi-definite positive, $1 - A \geq 0$. ■

Proposition 4 *Under assumption (1), S_ε is convex on sub-Gaussian measures with respect to either of its arguments.*

PROOF. Let $\beta \in \mathcal{G}(\mathbb{R}^d)$. Let $\alpha, \alpha' \in \mathcal{G}(\mathbb{R}^d)$. Let (f,g) and (f',g') denote the pair of potentials associated with $\mathrm{OT}_\varepsilon^\otimes(\alpha,\beta)$ and $\mathrm{OT}_\varepsilon^\otimes(\alpha',\beta)$ respectively and for any $\mu \in \mathcal{G}(\mathbb{R}^d)$, let h_μ denote the autocorrelation

potential associated with $\mathrm{OT}_\varepsilon^\otimes(\mu,\mu)$. The first order inequality (2.15) applied to $F = S_\varepsilon(.,\beta)$ is equivalent to:

$$\begin{aligned}
&(2.15) \Leftrightarrow \langle \alpha, f - h_\alpha\rangle + \langle \beta, g - h_\beta\rangle \geq \\
&\langle \alpha', f' - h_{\alpha'}\rangle + \langle \beta, g' - h_\beta\rangle + \langle \alpha - \alpha', f' - h_{\alpha'}\rangle \\
&\Leftrightarrow \langle \alpha, f - h_\alpha\rangle + \langle \beta, g\rangle \geq \langle \beta, g'\rangle + \langle \alpha, f' - h_{\alpha'}\rangle \\
&\Leftrightarrow \langle \alpha, f\rangle + \langle \beta, g\rangle \geq \langle \beta, g'\rangle + \langle \alpha, f' - h_{\alpha'} + h_\alpha\rangle \\
&\Leftrightarrow \mathrm{OT}_\varepsilon^\otimes(\alpha,\beta) \geq \langle \alpha, f' - h_{\alpha'} + h_\alpha\rangle + \langle \beta, g'\rangle
\end{aligned} \tag{2.19}$$

To show the last inequality we use the definition of the dual problem (2.6) and evaluate the dual function at the suboptimal potentials $(f' - h_{\alpha'} + h_\alpha, g')$. Doing so leads to:

$$\begin{aligned}
&\mathrm{OT}_\varepsilon^\otimes(\alpha,\beta) \geq \langle \alpha, f' - h_{\alpha'} + h_\alpha\rangle + \langle \beta, g'\rangle + \varepsilon \\
&- \varepsilon \int_{\mathbb{R}^d\times\mathbb{R}^d} \exp\left(\frac{(f' - h_{\alpha'} + h_\alpha) \oplus g' - C}{\varepsilon}\right) \mathrm{d}\alpha\, \mathrm{d}\beta \ .
\end{aligned}$$

To conclude, all we need to show is that,

$$\int_{\mathbb{R}^d\times\mathbb{R}^d} \exp\left(\frac{(f' - h_{\alpha'} + h_\alpha) \oplus g' - C}{\varepsilon}\right) \mathrm{d}\alpha\, \mathrm{d}\beta \leq 1 \tag{2.20}$$

By the Fubini-Tonelli theorem, the order of integration is irrelevant. First integrating with respect to β, we use the optimality conditions (2.7) on the pair (f', g') then on $h_{\alpha'}$:

$$\begin{aligned}
B &\stackrel{\text{def}}{=} \int_{\mathbb{R}^d\times\mathbb{R}^d} \exp\left(\frac{(f' - h_{\alpha'} + h_\alpha) \oplus g' - C}{\varepsilon}\right) \mathrm{d}\alpha\, \mathrm{d}\beta \\
&= \int_{\mathbb{R}^d} \exp\left(\frac{h_\alpha - h_{\alpha'}}{\varepsilon}\right) \mathrm{d}\alpha \\
&= \int_{\mathbb{R}^d\times\mathbb{R}^d} \exp\left(\frac{h_\alpha \oplus h_{\alpha'} - C}{\varepsilon}\right) \mathrm{d}\alpha\, \mathrm{d}\alpha'
\end{aligned}$$

Thus, Lemma 1 applies and we have $B \leq 1$. ■

1.2 Convexity and differentiability of $\mathrm{OT}_\varepsilon^\mathcal{L}$

We now turn to the continuous case with $m_1 = m_2 = \mathcal{L}$ the Lebesgue measure.

Dual problem Let α, β be continuous sub-Gaussian measures. Identifying α, β and π with their Lebesgue densities, The OT problem (2.1) has a dual problem given by:

$$\mathrm{OT}_\varepsilon^{\mathcal{L}}(\alpha,\beta) = \sup_{\substack{f\in\mathcal{L}_1(\alpha)\\ g\in\mathcal{L}_1(\beta)}} \langle f,\alpha\rangle + \langle g,\beta\rangle - \varepsilon \iint \exp\left(\frac{f(x)+g(y)-C(x,y)}{\varepsilon}\right) \mathrm{d}x\,\mathrm{d}y + \varepsilon \tag{2.21}$$

Notice that the convexity of $\mathrm{OT}_\varepsilon^{\mathcal{L}}$ follows immediately from (2.21) since it is a supremum of linear functions in α and β. The optimality conditions are equivalent to the marginal constraints of the primal problem (2.1). However, they are slightly different than those of $\mathrm{OT}_\varepsilon^{\otimes}$. Cancelling the gradient of the dual problem leads to the following system (Ivan Gentil, 2017):

$$\begin{aligned} e^{\frac{f}{\varepsilon}}\mathcal{K}(e^{\frac{g}{\varepsilon}}) &= \alpha \ , \\ e^{\frac{g}{\varepsilon}}\mathcal{K}^\top(e^{\frac{f}{\varepsilon}}) &= \beta \ . \end{aligned} \tag{2.22}$$

which in integral form can be written:

$$\begin{aligned} e^{\frac{f(x)}{\varepsilon}} \int e^{\frac{-C(x,y)+g(y)}{\varepsilon}} \mathrm{d}y &= \alpha(x) \ \ \forall x, \\ e^{\frac{g(x)}{\varepsilon}} \int e^{\frac{-C(y,x)+f(y)}{\varepsilon}} \mathrm{d}y &= \beta(x) \ \ \forall x, \end{aligned} \tag{2.23}$$

and the optimal transport plan's density π is given by: $\pi(x,y) = \exp\left(\frac{f(x)+g(y)-C(x,y)}{\varepsilon}\right)$

Thus, at optimality the integral over $\mathbb{R}^d \times \mathbb{R}^d$ sums to 1 and:

$$\mathrm{OT}_\varepsilon^{\mathcal{L}}(\alpha,\beta) = \langle f,\alpha\rangle + \langle g,\beta\rangle \tag{2.24}$$

Convexity and Differentiability Using absolute continuity continuity, one can rewrite the KL in the primal problem such that it holds (Di Marino and Gerolin, 2020):

$$\mathrm{OT}_\varepsilon^{\mathcal{L}}(\alpha,\beta) = \mathrm{OT}_\varepsilon^{\otimes}(\alpha,\beta) + \varepsilon\mathrm{KL}(\alpha|\mathcal{L}) + \varepsilon\mathrm{KL}(\beta|\mathcal{L}) \ . \tag{2.25}$$

We already showed that $\mathrm{OT}_\varepsilon^{\otimes}$ is convex (w.r.t. to one argument); KL is also convex (even jointly convex). Since the set of Lebesgue-continuous and sub-Gaussian measures is convex , $\mathrm{OT}_\varepsilon^{\mathcal{L}}$ is also convex with respect to one argument.

Identifying α with its density, it holds:

$$E(\alpha) \overset{\text{def}}{=} \mathrm{KL}(\alpha,\mathcal{L}) = \int \alpha(x)(\log(\alpha(x)) - 1)\,\mathrm{d}x \tag{2.26}$$

If $\alpha > 0$, then for any feasible displacement $h = h_1 - h_2$ with density functions h_1, h_2. The functional derivative of E in the direction h is given by: $\left[\frac{\mathrm{d}E(\alpha+th)}{\mathrm{d}t}\right]_{t=0} = \langle h, \log(\alpha)\rangle$. Thus, in the sense of the directional differentiation (2.11):

$$\nabla_\alpha \mathrm{KL}(\alpha, \mathcal{L}) = \log(\alpha) \tag{2.27}$$

Let (f, g) be a pair of optimal potentials for $\mathrm{OT}_\varepsilon^\otimes(\alpha, \beta)$. Following the differentiability of $\mathrm{OT}_\varepsilon^\otimes$ given by proposition 3 and the differentiability of KL, $\mathrm{OT}_\varepsilon^\mathcal{L}$ is differentiable on the set of sub-Gaussian measures with positive density functions and its gradient is given by: $\nabla_1 \mathrm{OT}_\varepsilon^\mathcal{L}(\alpha, \beta) = f - \varepsilon\log(\alpha)$. By a simple calculation, it is easy to show that $(f - \varepsilon\log(\alpha), g - \varepsilon\log(\beta))$ are actually solutions of the Sinkhorn equations (2.23). Similarly, given a solution (f_1, g_1) of (2.23), $(f_1 + \varepsilon\log(\alpha), g_1 + \varepsilon\log(\beta))$ are optimal potentials of $\mathrm{OT}_\varepsilon^\otimes$. Therefore, the following proposition holds:

Proposition 5 *Let $\alpha, \beta \in \mathcal{G}_\sigma(\mathbb{R}^d)$. If α and β are Lebesgue-continuous with positive density functions, then $\mathrm{OT}_\varepsilon^\mathcal{L}$ is differentiable and it holds:*

$$\nabla \mathrm{OT}_\varepsilon^\mathcal{L}(\alpha, \beta) = (f, g) \ , \tag{2.28}$$

where (f, g) is a pair of dual potentials verifying the fixed point equations (2.24).

2 Entropic OT for Gaussians

After having laid the ground for studying OT_ε on $\mathbb{R}^d$, we can now focus on Gaussian measures. Without invoking convexity and differentiability to study OT barycenters, one can first notice that Sinkhorn's optimality conditions could be stable for quadratic potentials if α and β are Gaussians. If successful, this attempt would provide a first closed form for entropic OT.

2.1 Closed form expressions

This section establishes closed form expressions for entropic between Gaussian measures. Except the last two propositions establishing the limits of UOT, all theoretical results were shown in (Janati et al., 2020a).

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, Jain, and Lim, 2018; Malagò, Montrucchio, and Pistone, 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, Georgiou, and Tannenbaum, 2018).

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, Chintala, and Bottou, 2017; Cuturi and Peyré, 2016). We focus in this section 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, Mielke, and Savaré, 2016; Liero, Mielke, and Savaré, 2018). The advantages of unbalanced entropic transport are numerous: it comes with favorable sample complexity regimes compared to unregularized OT (Genevay et al., 2019; Séjourné et al., 2019), can be cast as a loss with favorable properties (Genevay, Peyre, and Cuturi, 2018; Feydy et al., 2019), and can be evaluated using variations of the Sinkhorn algorithm (Genevay et al., 2016).

On the absence of closed forms of entropic 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, Trillos, and Cuturi, 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. One of our contributions is to fill this gap, and provide closed form expressions for balanced and unbalanced OT for Gaussian measures. We hope these formulas will prove 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.

2.1.1 Bures-Wasserstein and elliptical distributions

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:

$$W_2^2(\alpha, \beta) \stackrel{\text{def}}{=} \min_{\pi \in \Pi(\alpha,\beta)} \int_{\mathbb{R}^{d\times d}} \|x - y\|^2 \mathrm{d}\pi(x, y). \tag{2.29}$$

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 (2.29) can be equivalently rewritten using the *Monge* formulation, where $T_\sharp \mu = \nu$ *i.f.f.* for all Borel sets A, $\nu(T(A)) = \mu(A)$:

$$W_2^2(\alpha, \beta) = \min_{T : T_\sharp \alpha = \nu} \int_{\mathbb{R}^d} \|x - T(x)\|^2 \mathrm{d}\alpha(x). \tag{2.30}$$

The optimal map T^* in Equation (2.30) 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 (2.29) admits a closed form for Gaussian distributions, called the Wasserstein-Bures distance (a.k.a. the *Fréchet* distance):

$$W_2^2(\mathcal{N}(a,\mathbf{A}),\mathcal{N}(b,\mathbf{B})) = \|a-b\|^2 + \mathfrak{B}(\mathbf{A},\mathbf{B}), \tag{2.31}$$

where $\mathfrak{B}$ is the (squared) *Bures* distance (Bhatia, Jain, and Lim, 2018) between positive matrices:

$$\mathfrak{B}(\mathbf{A},\mathbf{B}) \stackrel{\text{def}}{=} \mathrm{Tr}\mathbf{A} + \mathrm{Tr}\mathbf{B} - 2\mathrm{Tr}(\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}})^{\frac{1}{2}}. \tag{2.32}$$

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

$$\mathbf{T}^{\mathbf{AB}} \stackrel{\text{def}}{=} \mathbf{A}^{-\frac{1}{2}}(\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}})^{\frac{1}{2}}\mathbf{A}^{-\frac{1}{2}} = \mathbf{B}^{\frac{1}{2}}(\mathbf{B}^{\frac{1}{2}}\mathbf{A}\mathbf{B}^{\frac{1}{2}})^{-\frac{1}{2}}\mathbf{B}^{\frac{1}{2}}, \tag{2.33}$$

which is related to the Bures gradient (w.r.t. the Frobenius inner product):

$$\nabla_{\mathbf{A}}\mathfrak{B}(\mathbf{A},\mathbf{B}) = \mathrm{Id} - \mathbf{T}^{\mathbf{AB}}. \tag{2.34}$$

$\mathfrak{B}(\mathbf{A},\mathbf{B})$ and its gradient can be computed efficiently on GPUs using Newton-Schulz iterations which are provided in Algorithm 1. The main bottleneck in computing $\mathbf{T}^{\mathbf{AB}}$ is that of computing matrix square

Algorithm 1 NS Monge Iterations

Require: 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**
    $\mathbf{T} \leftarrow (3\,\mathrm{Id} - \mathbf{Z}\mathbf{Y})/2$
    $\mathbf{Y} \leftarrow \mathbf{Y}\mathbf{T}$
    $\mathbf{Z} \leftarrow \mathbf{T}\mathbf{Z}$
  end while
  $\mathbf{Y} \leftarrow \sqrt{\frac{\|\mathbf{B}\|}{\|\mathbf{A}\|}}\mathbf{Y}$, $\mathbf{Z} \leftarrow \sqrt{\frac{\|\mathbf{A}\|}{\|\mathbf{B}\|}}\mathbf{Z}$
Ensure: $\mathbf{Y} = \mathbf{T}^{\mathbf{AB}}$, $\mathbf{Z} = \mathbf{T}^{\mathbf{BA}}$

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 $\mathbf{A}^{\frac{1}{2}}$ and $\mathbf{A}^{-\frac{1}{2}}$, and a second time to get $(\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}})^{\frac{1}{2}}$.

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 1. Using (2.34), 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.

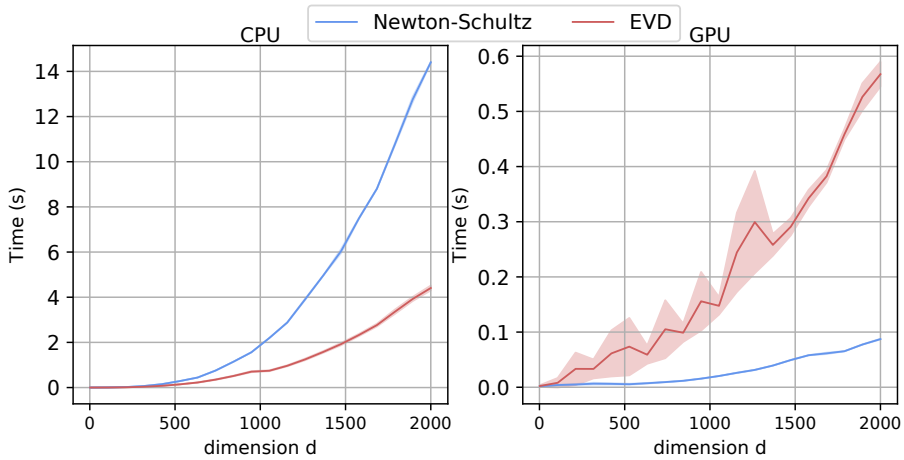


Fig. 2.1. Average run-time of Newton-Schulz and EVD to compute on CPUs and GPUs.

The main advantage of Newton-Schultz over SVD is that it its efficient scalability on GPUs, as illustrated in Figure 2.1.

Newton-Schulz iterations are quadratically convergent under the condition:

$$\| \mathrm{Id} - \left(\begin{smallmatrix} \mathbf{A} & 0 \\ 0 & \mathbf{B} \end{smallmatrix} \right)^2 \| < 1 \ ,$$

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 1 (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. (2.33) 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 2.1 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 $W(\mathrm{Id}_d, d)$ to which we add 0.1 Id 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 $\mathbf{A}_{evd}^{\frac{1}{2}}$ and $\mathbf{A}_{evd}^{-\frac{1}{2}}$ and stop the Newton-Schultz algorithm when its n-th running estimate $\mathbf{A}_n^{\frac{1}{2}}$ verifies: $\|\mathbf{A}_n^{\frac{1}{2}} - \mathbf{A}_{evd}^{\frac{1}{2}}\|_1 \leq 10^{-4}$. 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.

2.1.2 Balanced OT: entropic Bures-Wasserstein

Solving (2.29) can be quite challenging, even in a discrete setting (Peyré and Cuturi, 2018). Adding an entropic regularization term to (2.29) 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:

$$\mathrm{OT}^{\otimes}_{2\sigma^2}(\alpha,\beta) \stackrel{\text{def}}{=} \min_{\pi\in\Pi(\alpha,\beta)} \int_{\mathbb{R}^d\times\mathbb{R}^d} \|x-y\|^2 \mathrm{d}\pi(x,y) + 2\sigma^2\,\mathrm{KL}(\pi\|\alpha\otimes\beta), \tag{2.35}$$

where $\mathrm{KL}(\pi\|\alpha\otimes\beta) \stackrel{\text{def}}{=} \int_{\mathbb{R}^d} \log\left(\frac{\mathrm{d}\pi}{\mathrm{d}\alpha\mathrm{d}\beta}\right)\mathrm{d}\pi$ is the Kullback-Leibler divergence (or relative entropy). As in the original case (2.29), $\mathrm{OT}^{\otimes}_{2\sigma^2}$ can be studied with centered measures (i.e zero mean) with no loss of generality:

Lemma 2 *Let $\alpha,\beta\in\mathcal{P}$ and $\bar{\alpha},\bar{\beta}$ their respective centered transformations. It holds that*

$$\mathrm{OT}^{\otimes}_{2\sigma^2}(\alpha,\beta) = \mathrm{OT}^{\otimes}_{2\sigma^2}(\bar{\alpha},\bar{\beta}) + \|\mathbf{a}-\mathbf{b}\|^2. \tag{2.36}$$

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, $\forall\pi\in\Pi(\alpha,\beta)$,

1. $\bar{\pi}\in\Pi(\bar{\alpha},\bar{\beta})$,
2. $\mathrm{KL}(\pi\|\alpha\otimes\beta) = \mathrm{KL}(\bar{\pi}\|\bar{\alpha}\otimes\bar{\beta})$
3. $\int_{\mathbb{R}^d\times\mathbb{R}^d}\|x-y\|^2\mathrm{d}\bar{\pi}(x,y) = \int_{\mathbb{R}^d\times\mathbb{R}^d}\|(x-\mathbf{a})-(y-\mathbf{b})\|^2\mathrm{d}\pi(x,y) = \|\mathbf{a}-\mathbf{b}\|^2 + \int_{\mathbb{R}^d\times\mathbb{R}^d}\|x-y\|^2\mathrm{d}\pi(x,y)$

Plugging (i)-(iii) into (2.35), we get $\mathrm{OT}^{\otimes}_{2\sigma^2}(\alpha,\beta) = \mathrm{OT}^{\otimes}_{2\sigma^2}(\bar{\alpha},\bar{\beta}) + \|\mathbf{a}-\mathbf{b}\|^2$. ■

Dual problem and Sinkhorn's algorithm. Compared to (2.29), (2.35) enjoys additional properties, such as the uniqueness of the solution π^*. Moreover, problem (2.35) has the following dual formulation:

$$\mathrm{OT}^{\otimes}_{2\sigma^2}(\alpha,\beta) = \max_{\substack{f\in\mathcal{L}_1(\alpha),\\ g\in\mathcal{L}_1(\beta)}} \mathbb{E}_\alpha(f) + \mathbb{E}_\beta(g) - 2\sigma^2\left(\int_{\mathbb{R}^d\times\mathbb{R}^d} e^{\frac{f(x)+g(y)-\|x-y\|^2}{2\sigma^2}}\mathrm{d}\alpha(x)\mathrm{d}\beta(y) - 1\right). \tag{2.37}$$

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):

$$e^{\frac{f(x)}{2\sigma^2}}\left(\int_{\mathbb{R}^d} e^{\frac{-\|x-y\|^2+g(y)}{2\sigma^2}}\mathrm{d}\beta(y)\right) = 1, \quad e^{\frac{g(x)}{2\sigma^2}}\left(\int_{\mathbb{R}^d} e^{\frac{-\|x-y\|^2+f(y)}{2\sigma^2}}\mathrm{d}\alpha(y)\right) = 1. \tag{2.38}$$

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

$$\frac{\mathrm{d}\pi^\star}{\mathrm{d}\alpha\mathrm{d}\beta}(x,y) = e^{\frac{f(x)+g(y)-\|x-y\|^2}{2\sigma^2}}. \tag{2.39}$$

Starting from a pair of potentials (f_0, g_0), the optimality conditions (2.38) lead to an alternating dual ascent algorithm, which is equivalent to Sinkhorn's algorithm in log-domain:

$$\begin{aligned} g_{n+1} &= \left(y \in \mathbb{R}^d \to -2\sigma^2 \log \int_{\mathbb{R}^d} e^{\frac{-\|x-y\|^2 + f_n(x)}{2\sigma^2}} \mathrm{d}\alpha(x) \right), \\ f_{n+1} &= \left(x \in \mathbb{R}^d \to -2\sigma^2 \log \int_{\mathbb{R}^d} e^{\frac{-\|x-y\|^2 + g_{n+1}(y)}{2\sigma^2}} \mathrm{d}\beta(y) \right). \end{aligned} \tag{2.40}$$

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.

Closed form expression for Gaussian measures. When the measures are Gaussian, the algorithm above not only converges but its limit can be obtained analytically. The following theorem provides the closed form of entropic OT for Gaussian measures for both the product and the Lebesgue measure.

Theorem 1 *Let* $\mathbf{A}, \mathbf{B} \in \mathcal{S}^d_{++}$ *and* $\alpha = \mathcal{N}(\mathbf{a}, \mathbf{A})$ *and* $\beta = \mathcal{N}(\mathbf{b}, \mathbf{B})$.
Let $\mathbf{C}_\sigma = \mathbf{A}^{\frac{1}{2}}(\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}} + \frac{\sigma^4}{4}\mathrm{Id})^{\frac{1}{2}}\mathbf{A}^{-\frac{1}{2}} - \frac{\sigma^2}{2}\mathrm{Id}$, *then,*

$$\mathrm{OT}^{\otimes}_{2\sigma^2}(\alpha, \beta) = \|\mathbf{a} - \mathbf{b}\|^2 + \mathfrak{B}^{\otimes}_{2\sigma^2}(\mathbf{A}, \mathbf{B}) \tag{2.41}$$

$$\mathrm{OT}^{\mathcal{L}}_{2\sigma^2}(\alpha, \beta) = \|\mathbf{a} - \mathbf{b}\|^2 + \mathfrak{B}^{\mathcal{L}}_{2\sigma^2}(\mathbf{A}, \mathbf{B}) \tag{2.42}$$

where:

$$\begin{aligned} \mathcal{B}^{\otimes}_{2\sigma^2}(\mathbf{A}, \mathbf{B}) &= \mathrm{Tr}(\mathbf{A}) + \mathrm{Tr}(\mathbf{B}) - 2\mathrm{Tr}(\mathbf{C}_\sigma) + \sigma^2 \log\det\left(\frac{1}{\sigma^2}\mathbf{C}_\sigma + \mathrm{Id}\right) \\ \mathcal{B}^{\mathcal{L}}_{2\sigma^2}(\mathbf{A}, \mathbf{B}) &= \mathcal{B}^{\otimes}_{2\sigma^2}(\mathbf{A}, \mathbf{B}) - \sigma^2 \log\det((2\pi e)^2 \mathbf{A}\mathbf{B}) \ , \end{aligned} \tag{2.43}$$

Moreover, regardless of the reference measure, the Sinkhorn optimal transportation plan is also a Gaussian measure over $\mathbb{R}^d \times \mathbb{R}^d$ *given by*

$$\pi^\star = \mathcal{N}\left(\begin{pmatrix} \mathbf{a} \\ \mathbf{b} \end{pmatrix}, \begin{pmatrix} \mathbf{A} & \mathbf{C}_\sigma \\ \mathbf{C}_\sigma^\top & \mathbf{B} \end{pmatrix} \right). \tag{2.44}$$

Remark 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* $\mathrm{OT}^{\otimes}_{2\sigma^2}$ *given by Theorem 1 remains well-defined for positive semi-definite matrices. Moreover, unlike the Bures-Wasserstein metric and* $\mathrm{OT}^{\mathcal{L}}_{2\sigma^2}$, $\mathrm{OT}^{\otimes}_{2\sigma^2}$ *is differentiable even when* $\mathbf{A}$ *or* $\mathbf{B}$ *are singular.*

The proof of theorem 1 is broken down into smaller results, Propositions 6 to 8 and lemma 3. Using Lemma 2, we can focus in the rest of this section on centered Gaussians without loss of generality. Moreover, the formula for the Lebesgue measure $\mathrm{OT}^{\mathcal{L}}_\varepsilon$ can derived from that of $\mathrm{OT}^{\otimes}_\varepsilon$ using equation (2.25).

Sinkhorn's algorithm and quadratic potentials. We obtain a closed form solution of $\mathrm{OT}^{\otimes}_{2\sigma^2}$ by considering quadratic solutions of (2.38). The following key proposition characterizes the obtained potential after a pair of Sinkhorn iterations with quadratic forms.

Proposition 6 *Let $\alpha = \mathcal{N}(0, \mathbf{A})$ and $\beta = \mathcal{N}(0, \mathbf{B})$ and the Sinkhorn transform $T_\alpha : \mathbb{R}^{\mathbb{R}^d} \to \mathbb{R}^{\mathbb{R}^d}$:*

$$T_\alpha(h)(x) \stackrel{\text{def}}{=} -\log \int_{\mathbb{R}^d} e^{\frac{-\|x-y\|^2}{2\sigma^2}+h(y)} \mathrm{d}\alpha(y). \tag{2.45}$$

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 $\mathbf{X}' \stackrel{\text{def}}{=} \sigma^2\mathbf{X} + \sigma^2\mathbf{A}^{-1} + \mathrm{Id} \succ 0$. In that case,

1. *$T_\alpha(h) = \mathcal{Q}(\mathbf{Y}) + m'$ where $\mathbf{Y} = \frac{1}{\sigma^2}(\mathbf{X}'^{-1} - \mathrm{Id})$ and $m' \in \mathbb{R}$ is an additive constant*
2. *$T_\beta(T_\alpha(h))$ is well-defined and is also a quadratic form up to an additive constant, since $\mathbf{Y}' \stackrel{\text{def}}{=} \sigma^2\mathbf{Y} + \sigma^2\mathbf{B}^{-1} + \mathrm{Id} = \mathbf{X}'^{-1} + \sigma^2\mathbf{B}^{-1} \succ 0$ and (i) applies.*

PROOF. The exponent inside the integral can be written as:

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

which is integrable if and only if $\mathbf{X} + \mathbf{A}^{-1} + \frac{1}{\sigma^2}\mathrm{Id} \succ 0$. Moreover, up to a multiplicative factor, the exponentiated Sinkhorn transform is equivalent to a Gaussian convolution of an exponentiated quadratic form. Lemma 7 (see appendix) – which characterizes the Gaussian convolution of quadratic forms – applies:

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

Therefore $T_\alpha(h)$ is up to an additive constant given by $\mathcal{Q}(\frac{1}{\sigma^2}(\mathbf{X}'^{-1} - \mathrm{Id}))$. 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 $\sigma^2\mathbf{A}^{-1} + \mathrm{Id} \succ 0$, Proposition 6 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 7 *Starting with null potentials, Sinkhorn's algorithm is equivalent to the iterations:*

$$\mathbf{F}_{n+1} = \sigma^2\mathbf{A}^{-1} + \mathbf{G}_n^{-1}, \qquad \mathbf{G}_{n+1} = \sigma^2\mathbf{B}^{-1} + \mathbf{F}_{n+1}^{-1}, \tag{2.46}$$

with $\mathbf{F}_0 = \sigma^2\mathbf{A}^{-1} + \mathrm{Id}$ *and* $\mathbf{G}_0 = \sigma^2\mathbf{B}^{-1} + \mathrm{Id}$.

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

$$\mathbf{F} = \sigma^2\mathbf{U} + \sigma^2\mathbf{A}^{-1} + \mathrm{Id}, \qquad \mathbf{G} = \sigma^2\mathbf{V} + \sigma^2\mathbf{B}^{-1} + \mathrm{Id}\,. \tag{2.47}$$

Closed form solution. Taking the limit of Sinkhorn's equations (2.46) along with the change of variable (2.47), there exists a pair of optimal potentials determined up to an additive constant:

$$\frac{f}{2\sigma^2} = \mathcal{Q}(\mathbf{U}) = \mathcal{Q}\left(\frac{1}{\sigma^2}(\mathbf{G}^{-1} - \mathrm{Id})\right), \qquad \frac{g}{2\sigma^2} = \mathcal{Q}(\mathbf{V}) = \mathcal{Q}\left(\frac{1}{\sigma^2}(\mathbf{F}^{-1} - \mathrm{Id})\right), \tag{2.48}$$

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

$$\mathbf{F} = \sigma^2\mathbf{A}^{-1} + \mathbf{G}^{-1}, \qquad \mathbf{G} = \sigma^2\mathbf{B}^{-1} + \mathbf{F}^{-1}. \tag{2.49}$$

Let $\mathbf{C} \overset{\text{def}}{=} \mathbf{A}\mathbf{G}^{-1}$. Combining both equations of (2.49) in one leads to $\mathbf{G} = \sigma^2\mathbf{B}^{-1} + (\mathbf{G}^{-1} + \sigma^2\mathbf{A}^{-1})^{-1}$, which can be shown (see proof below) to be equivalent to

$$\mathbf{C}^2 + \sigma^2\mathbf{C} - \mathbf{A}\mathbf{B} = 0. \tag{2.50}$$

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 it has positive eigenvalues. Proposition 8 provides the only feasible solution of (2.50).

Proposition 8 *Let* $\sigma^2 \geq 0$ *and* $\mathbf{C}$ *satisfying Equation* (2.50). *Then,*

$$\mathbf{C} = \left(\mathbf{A}\mathbf{B} + \frac{\sigma^4}{4}\mathrm{Id}\right)^{\frac{1}{2}} - \frac{\sigma^2}{2}\mathrm{Id} = \mathbf{A}^{\frac{1}{2}}(\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}} + \tfrac{\sigma^4}{4}\mathrm{Id})^{\frac{1}{2}}\mathbf{A}^{-\frac{1}{2}} - \tfrac{\sigma^2}{2}\mathrm{Id}\,. \tag{2.51}$$

PROOF.

Combining the two equations in (2.49) yields

$$\begin{aligned}
&\mathbf{G} = \sigma^2\mathbf{B}^{-1} + (\mathbf{G}^{-1} + \sigma^2\mathbf{A}^{-1})^{-1} \\
&\Leftrightarrow \mathbf{G}\mathbf{A}^{-1} = \sigma^2\mathbf{B}^{-1}\mathbf{A}^{-1} + (\mathbf{A}\mathbf{G}^{-1} + \sigma^2\,\mathrm{Id})^{-1} \\
&\Leftrightarrow \mathbf{C}^{-1} = \sigma^2(\mathbf{A}\mathbf{B})^{-1} + (\mathbf{C} + \sigma^2\,\mathrm{Id})^{-1} \\
&\Leftrightarrow \mathbf{C}^{-1}(\mathbf{C} + \sigma^2\,\mathrm{Id}) = \sigma^2(\mathbf{A}\mathbf{B})^{-1}(\mathbf{C} + \sigma^2\,\mathrm{Id}) + \mathrm{Id} \\
&\Leftrightarrow \mathrm{Id} + \sigma^2\mathbf{C}^{-1} = \sigma^2(\mathbf{A}\mathbf{B})^{-1}(\mathbf{C} + \sigma^2\,\mathrm{Id}) + \mathrm{Id} \\
&\Leftrightarrow \mathbf{C} + \sigma^2\,\mathrm{Id} = \sigma^2(\mathbf{A}\mathbf{B})^{-1}(\mathbf{C} + \sigma^2\,\mathrm{Id})\mathbf{C} + \mathbf{C} \\
&\Leftrightarrow \mathbf{C}^2 + \sigma^2\mathbf{C} - \mathbf{A}\mathbf{B} = 0. \qquad (2.52)
\end{aligned}$$

Given that $\mathbf{A}$ and $\mathbf{G}^{-1}$ are positive, their product $\mathbf{C} = \mathbf{A}\mathbf{G}^{-1}$ can be written: $\mathbf{A}\mathbf{G}^{-1} = \mathbf{A}^{\frac{1}{2}}(\mathbf{A}^{\frac{1}{2}}\mathbf{G}^{-1}\mathbf{A}^{\frac{1}{2}})\mathbf{A}^{-\frac{1}{2}}$, thus $\mathbf{A}\mathbf{G}^{-1}$ is similar to the positive matrix $\mathbf{A}^{\frac{1}{2}}\mathbf{G}^{-1}\mathbf{A}^{\frac{1}{2}}$. Therefore, one can write an eigenvalue decomposition of $\mathbf{C} = \mathbf{P}\Sigma\mathbf{P}^{-1}$ with a positive diagonal matrix Σ. Substituting in (2.50), it follows that $\mathbf{C}$ and $\mathbf{A}\mathbf{B}$ share the same eigenvectors with modified eigenvalues. Thus, it is sufficient to find the real roots of the polynomial $x \mapsto x^2 + \sigma^2 x - ab$ with $a, b \in \mathbb{R}_{++}$ which are given by: $x_1 = -\frac{\sigma^2}{2} - \sqrt{ab + \frac{\sigma^4}{4}}$ and $x_2 = -\frac{\sigma^2}{2} + \sqrt{ab + \frac{\sigma^4}{4}}$. Since $\mathbf{C}$ is the product of the positive definite matrices $\mathbf{G}^{-1}$ and $\mathbf{A}$, its eigenvalues are all positive. Discarding the negative root, the closed form follows immediately. Indeed, by direct calculation, computing the square of the solution $\mathbf{C}$ leads to the equation (2.50):

$$\begin{aligned}
\mathbf{C}^2 &= \mathbf{A}\mathbf{B} + \frac{\sigma^4}{2}\,\mathrm{Id} - \sigma^2\left(\mathbf{A}\mathbf{B} + \frac{\sigma^4}{4}\,\mathrm{Id}\right)^{\frac{1}{2}} \\
&= \mathbf{A}\mathbf{B} - \sigma^2\mathbf{C}.
\end{aligned}$$

The second equality is obtained by observing that

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

i.e. that

$$\left(\mathbf{A}\mathbf{B} + \tfrac{\sigma^4}{4}\,\mathrm{Id}\right)^{\frac{1}{2}} = \mathbf{A}^{\frac{1}{2}}(\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}} + \tfrac{\sigma^4}{4}\,\mathrm{Id})^{\frac{1}{2}}\mathbf{A}^{-\frac{1}{2}}.$$

■

Using the transformation (2.48), we obtain the following corollary:

Corollary 2 *The optimal dual potentials of* (2.48) *can be given in closed form by:*

$$\mathbf{U} = \frac{\mathbf{B}}{\sigma^2}(\mathbf{C} + \sigma^2\,\mathrm{Id})^{-1} - \frac{\mathrm{Id}}{\sigma^2}, \qquad \mathbf{V} = (\mathbf{C} + \sigma^2\,\mathrm{Id})^{-1}\frac{\mathbf{A}}{\sigma^2} - \frac{\mathrm{Id}}{\sigma^2}. \qquad (2.53)$$

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

Optimal transportation plan and $\mathrm{OT}^{\otimes}_{2\sigma^2}$**.** Using Corollary 2 and (2.48), Equation (2.39) leads to a closed form expression of π. To conclude the proof of Theorem 1, we introduce lemma 3 that computes the $\mathrm{OT}^{\otimes}_{2\sigma^2}$ loss at optimality. Detailed technical proofs are provided in the appendix.

Lemma 3 *Let* $\mathbf{A}, \mathbf{B}, \mathbf{C}$ *be invertible matrices such that* $\mathbf{H} = \begin{pmatrix} \mathbf{A} & \mathbf{C} \\ \mathbf{C}^\top & \mathbf{B} \end{pmatrix} \succ 0$. *Let* $\alpha = \mathcal{N}(0, \mathbf{A}), \beta = \mathcal{N}(0, \mathbf{B})$, *and* $\pi = \mathcal{N}(0, \mathbf{H})$. *Then,*

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 \mathrm{d}\pi(x, y) = \mathrm{Tr}(\mathbf{A}) + \mathrm{Tr}(\mathbf{B}) - 2\mathrm{Tr}(\mathbf{C}), \tag{2.54}$$

$$\mathrm{KL}\left(\pi \| \alpha \otimes \beta\right) = \tfrac{1}{2}\left(\log\det \mathbf{A} + \log\det \mathbf{B} - \log\det \left(\begin{smallmatrix} \mathbf{A} & \mathbf{C} \\ \mathbf{C}^T & \mathbf{B} \end{smallmatrix}\right)\right). \tag{2.55}$$

Properties of $\mathrm{OT}^{\otimes}_{2\sigma^2}$**.** Theorem 1 shows that π has a Gaussian density. Proposition 9 allows to reformulate this optimization problem over couplings in $\mathbb{R}^{d\times d}$ with a positivity constraint.

Proposition 9 *Let* $\alpha = \mathcal{N}(0, \mathbf{A}), \beta = \mathcal{N}(0, \mathbf{B})$, *and* $\sigma^2 > 0$. *Then,*

$$\mathrm{OT}^{\otimes}_{2\sigma^2}(\alpha, \beta) = \min_{\mathbf{C}:\left(\begin{smallmatrix} \mathbf{A} & \mathbf{C} \\ \mathbf{C}^T & \mathbf{B} \end{smallmatrix}\right) \geq 0} \left\{ \mathrm{Tr}(\mathbf{A}) + \mathrm{Tr}(\mathbf{B}) - 2\mathrm{Tr}(\mathbf{C}) + \sigma^2 (\log\det \mathbf{AB} - \log\det \left(\begin{smallmatrix} \mathbf{A} & \mathbf{C} \\ \mathbf{C}^T & \mathbf{B} \end{smallmatrix}\right)) \right\} \tag{2.56}$$

$$= \min_{\mathbf{K} \in \mathbb{R}^{d\times d} : \|\mathbf{K}\|_{op} \leq 1} \mathrm{Tr}\mathbf{A} + \mathrm{Tr}\mathbf{B} - 2\mathrm{Tr}\mathbf{A}^{\frac{1}{2}}\mathbf{K}\mathbf{B}^{\frac{1}{2}} - \sigma^2 \ln\det(\mathrm{Id} - \mathbf{K}\mathbf{K}^\top). \tag{2.57}$$

Moreover, both (2.56) *and* (2.57) *are convex problems.*

We now study the convexity and differentiability of $\mathrm{OT}^{\otimes}_{2\sigma^2}$, which are more conveniently derived from the dual problem of (2.56) given as a positive definite program:

Proposition 10 *The dual problem of* (2.56) *can be written with no duality gap as*

$$\max_{\mathbf{F}, \mathbf{G} \succ 0} \left\{ \langle \mathrm{Id} - \mathbf{F}, \mathbf{A} \rangle + \langle \mathrm{Id} - \mathbf{G}, \mathbf{B} \rangle + \sigma^2 \log\det\left(\frac{\mathbf{FG} - \mathrm{Id}}{\sigma^4}\right) + \sigma^2 \log\det \mathbf{AB} + 2d\sigma^2 \right\}. \tag{2.58}$$

SKETCH OF PROOF. The dual problem follows from a simple application of Fenchel duality. The technical computation of the conjugate functions is provided in the appendix.

In the previous section, we have established that $\mathrm{OT}^{\otimes}_{\varepsilon}$ is differentiable on the set of sub-Gaussian measures with a gradient given by the optimal dual potentials. The following proposition re-establishes this statement for Gaussians and shows that minimizing the obtained loss on positive definite matrices $\mathfrak{B}^{\otimes}_{2\sigma^2}$ leads to a shrinking bias.

Proposition 11 *Assume* $\sigma > 0$ *and consider the pair* $\mathbf{U}, \mathbf{V}$ *of Corollary 2. Then*

(i) *The optimal pair* $(\mathbf{F}^*, \mathbf{G}^*)$ *of* (2.58) *is a solution to the fixed point problem* (2.49),

(ii) $\mathfrak{B}_{2\sigma^2}$ *is differentiable and:* $\nabla \mathfrak{B}^{\otimes}_{2\sigma^2}(\mathbf{A}, \mathbf{B}) = -(\sigma^2 \mathbf{U}, \sigma^2 \mathbf{V})$. *Thus:* $\nabla_{\mathbf{A}} \mathfrak{B}^{\otimes}_{2\sigma^2}(\mathbf{A}, \mathbf{B}) = \mathrm{Id} - \mathbf{B}^{\frac{1}{2}} \left((\mathbf{B}^{\frac{1}{2}} \mathbf{A} \mathbf{B}^{\frac{1}{2}} + \frac{\sigma^4}{4} \mathrm{Id})^{\frac{1}{2}} + \frac{\sigma^2}{2}$

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

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

SKETCH OF PROOF. We present the idea of the proof of each statement. The full technical details are provided in the appendix.

(i) The dual problem is concave in $\mathbf{F}$ and $\mathbf{G}$. Cancelling its gradient with respect to both leads to the same optimality conditions of the fixed point problem (2.49).

(ii) Applying Danskin's theorem leads to a gradient given by $(\mathrm{Id} - \mathbf{F} + \sigma^2 \mathbf{A}^{-1}, \mathrm{Id} - \mathbf{G} + \sigma^2 \mathbf{B}^{-1})$, which is equal to $-\sigma^2(\mathbf{U}, \mathbf{V})$ by virtue of the change of variable (2.47).

(iii) We compute the Hessian of $\mathbf{A} \mapsto \mathfrak{B}^{\otimes}_{2\sigma^2}$ and show its positivity. A simple counter-argument for joint convexity is provided in dimension 1.

(iv) $\mathfrak{B}^{\otimes}_{2\sigma^2}$ is convex and differentiable, we show that $\mathbf{A}_0$ verifies its first order optimality condition.

When $\mathbf{A}$ and $\mathbf{B}$ are not singular, by letting $\sigma \to 0$ in $\nabla_{\mathbf{A}} \mathfrak{B}_{2\sigma^2}(\mathbf{A}, \mathbf{B})$, we recover the gradient of the Bures metric given in (2.34).

Contrast with $\mathfrak{B}^{\mathcal{L}}_{2\sigma^2}$ Adding the individual entropies of α and β to the closed form of $\mathfrak{B}^{\otimes}_{2\sigma^2}$ leads to its closed form given in Theorem 1. Since these entropies are merely constants summing to $-\sigma^2 \log\det((2\pi e)^2 \mathbf{A}\mathbf{B})$, adding them to the dual problem (2.58):

$$\mathfrak{B}^{\mathcal{L}}_{2\sigma^2}(\mathbf{A}, \mathbf{B}) = \max_{\mathbf{F}, \mathbf{G} \succ 0} \left\{ \langle \mathrm{Id} - \mathbf{F}, \mathbf{A} \rangle + \langle \mathrm{Id} - \mathbf{G}, \mathbf{B} \rangle + \sigma^2 \log\det\left(\frac{\mathbf{F}\mathbf{G} - \mathrm{Id}}{\sigma^4} \right) - d\sigma^2 \log(2\pi) \right\} \tag{2.59}$$

As a supremum of linear functions in $(\mathbf{A}, \mathbf{B})$, $\mathfrak{B}^{\mathcal{L}}_{2\sigma^2}$ is jointly convex in $(\mathbf{A}, \mathbf{B})$, unlike $\mathfrak{B}^{\otimes}_{2\sigma^2}$. However, as clearly established by the presence of logdets of $\mathbf{A}$ and $\mathbf{B}$ in its closed form, its definition cannot be extended to the boundary of the cone of positive definite matrices. This is however not surprising, since for $\mathrm{OT}^{\mathcal{L}}_{\varepsilon}$ to be defined, the Gaussians must be absolutely continuous with respect to the Lebesgue measure in the first place. Finally, minimizing $\mathfrak{B}^{\mathcal{L}}_{2\sigma^2}$ leads to a blurring bias, which is analogous with the "usual" behavior of Sinkhorn's barycentric algorithm with the uniform reference measure in the discrete case. These properties – which can be deduced from those of $\mathfrak{B}^{\otimes}_{2\sigma^2}$ – are summarized in the following proposition.

Proposition 12 *Assume $\sigma > 0$, $\mathbf{A}, \mathbf{B}$ definite positive matrices and consider their associated pair $\mathbf{U}, \mathbf{V}$ of Corollary 2. Then:*

(i) $\mathfrak{B}^{\mathcal{L}}_{2\sigma^2}$ *is differentiable and:* $\nabla \mathfrak{B}^{\mathcal{L}}_{2\sigma^2}(\mathbf{A}, \mathbf{B}) = -\sigma^2(\mathbf{U} + \mathbf{A}^{-1}, \mathbf{V} + \mathbf{B}^{-1})$

(ii) $(\mathbf{A}, \mathbf{B}) \mapsto \mathfrak{B}^{\mathcal{L}}_{2\sigma^2}(\mathbf{A}, \mathbf{B})$ *is jointly convex.*

(iii) For a fixed $\mathbf{B}$*, the function* $\phi_{\mathbf{B}} : \mathbf{A} \mapsto \mathfrak{B}^{\mathcal{L}}_{2\sigma^2}(\mathbf{A}, \mathbf{B})$ *is minimized at* $\mathbf{A}_0 = \mathbf{B} + \sigma^2 \,\mathrm{Id}$.

2.1.3 Unbalanced OT: Entropic Gaussian-Hellinger-Kantorovich

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. We remind the reader of entropy-regularized unbalanced OT:

$$\mathrm{UOT}^{\otimes}_{2\sigma^2,\gamma}(\alpha, \beta) \stackrel{\text{def}}{=} \inf_{\pi \in \mathcal{M}_1^+} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 \mathrm{d}\pi(x, y) + 2\sigma^2 \,\mathrm{KL}(\pi \| \alpha \otimes \beta) + \gamma \,\mathrm{KL}(\pi_1 \| \alpha) + \gamma \,\mathrm{KL}(\pi_2 \| \beta), \quad (2.60)$$

where $\gamma > 0$ and π_1, π_2 are the marginal distributions of the coupling π.

Duality and optimality conditions. By definition of the KL divergence, the term $\mathrm{KL}(\pi \| \alpha \otimes \beta)$ in (2.60) is finite if and only if π admits a density with respect to $\alpha \otimes \beta$. Therefore (2.60) can be formulated as a variational problem:

$$\begin{aligned} \mathrm{UOT}_\sigma(\alpha, \beta) = \inf_{r \in \mathcal{L}_1(\alpha \otimes \beta)} \Big\{ & \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 r(x, y) \mathrm{d}\alpha(x) \mathrm{d}\beta(y) \\ & + 2\sigma^2 \,\mathrm{KL}(r \| \alpha \otimes \beta) + \gamma \,\mathrm{KL}(r_1 \| \alpha) + \gamma \,\mathrm{KL}(r_2 \| \beta) \Big\}, \end{aligned} \quad (2.61)$$

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 as: $\mathrm{KL}(f \| \mu) = \int_{\mathbb{R}^d} (f \log(f) + 1 - f) \mathrm{d}\mu$. As in (Chizat et al., 2018b), Fenchel-Rockafellar duality provides the following dual problem:

$$\begin{aligned} \mathrm{UOT}_\sigma(\alpha, \beta) = \sup_{\substack{f \in \mathcal{L}_\infty(\alpha) \\ g \in \mathcal{L}_\infty(\beta)}} \Big\{ & \gamma \int_{\mathbb{R}^d} (1 - e^{-\frac{f}{\gamma}}) \mathrm{d}\alpha + \gamma \int_{\mathbb{R}^d} (1 - e^{-\frac{g}{\gamma}}) \mathrm{d}\beta \\ & - 2\sigma^2 \int_{\mathbb{R}^d \times \mathbb{R}^d} (e^{\frac{-\|x-y\|^2 + f(x) + g(y)}{2\sigma^2}} - 1) \mathrm{d}\alpha(x) \mathrm{d}\beta(y) \Big\}. \end{aligned} \quad (2.62)$$

For which strong duality holds. Moreover, a maximizing sequence of potentials (f_n, g_n) weakly converges towards a pair of measurable functions (f, g) if they verify the optimality conditions (Rockafellar, 1970):

$$\frac{f(x)}{2\sigma^2} \stackrel{a.s}{=} -\tau \log \int_{\mathbb{R}^d} e^{\frac{g(y) - \|x-y\|^2}{2\sigma^2}} \mathrm{d}\beta(y), \quad \frac{g(x)}{2\sigma^2} \stackrel{a.s}{=} -\tau \log \int_{\mathbb{R}^d} e^{\frac{f(y) - \|x-y\|^2}{2\sigma^2}} \mathrm{d}\alpha(y), \quad (2.63)$$

where $\tau \overset{\text{def}}{=} \frac{\gamma}{\gamma+2\sigma^2}$.

In which case the (unique) optimal transportation plan is given by:

$$\frac{\mathrm{d}\pi}{\mathrm{d}\alpha \otimes \mathrm{d}\beta}(x,y) = e^{\frac{f(x)+g(y)-\|x-y\|^2}{2\sigma^2}}. \tag{2.64}$$

The following proposition provides a simple formula to compute $\mathrm{UOT}^{\otimes}_{2\sigma^2}$ at optimality. It shows that it is sufficient to know the total transported mass $\pi(\mathbb{R}^d \times \mathbb{R}^d)$.

Proposition 13 *Assume there exists an optimal transportation plan π^*, solution of (2.60), then:*

$$\mathrm{UOT}^{\otimes}_{2\sigma^2,\gamma}(\alpha,\beta) = \gamma(m_\alpha + m_\beta) + 2\sigma^2 m_\alpha m_\beta - 2(\sigma^2+\gamma)\pi^*(\mathbb{R}^d \times \mathbb{R}^d). \tag{2.65}$$

Unbalanced Entropic OT for scaled Gaussians. Let α and β be unbalanced 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 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:

$$\frac{f(\mathbf{x})}{2\sigma^2} = -\frac{1}{2}(x^\top \mathbf{U}\mathbf{x} - 2x^\top \mathbf{u}) + \log(m_u), \quad \frac{g(x)}{2\sigma^2} = -\frac{1}{2}(x^\top \mathbf{V}\mathbf{x} - 2x^\top \mathbf{v}) + \log(m_v). \tag{2.66}$$

Let σ and γ be the regularization parameters as in Equation (2.61), 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:

$$\mu = \begin{pmatrix} \mathbf{a} + \mathbf{A}\mathbf{X}^{-1}(\mathbf{b}-\mathbf{a}) \\ \mathbf{b} + \mathbf{B}\mathbf{X}^{-1}(\mathbf{a}-\mathbf{b}) \end{pmatrix} \tag{2.67}$$

$$\mathbf{H} = \begin{pmatrix} (\mathrm{Id} + \frac{1}{\lambda}\mathbf{C})(\mathbf{A} - \mathbf{A}\mathbf{X}^{-1}\mathbf{A}) & \mathbf{C} + (\mathrm{Id} + \frac{1}{\lambda}\mathbf{C})\mathbf{A}\mathbf{X}^{-1}\mathbf{B} \\ \mathbf{C}^\top + (\mathrm{Id} + \frac{1}{\lambda}\mathbf{C}^\top)\mathbf{B}\mathbf{X}^{-1}\mathbf{A} & (\mathrm{Id} + \frac{1}{\lambda}\mathbf{C}^\top)(\mathbf{B} - \mathbf{B}\mathbf{X}^{-1}\mathbf{B}) \end{pmatrix} \tag{2.68}$$

$$m_\pi = \sigma^{\frac{d\sigma^2}{\gamma+\sigma^2}} \left(m_\alpha m_\beta \det(\mathbf{C}) \sqrt{\frac{\det(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}})^\tau}{\det(\mathbf{A}\mathbf{B})}} \right)^{\frac{1}{\tau+1}} \frac{e^{-\frac{\|\mathbf{a}-\mathbf{b}\|^2_{\mathbf{X}^{-1}}}{2(\tau+1)}}}{\sqrt{\det(\mathbf{C} - \frac{2}{\gamma}\widetilde{\mathbf{A}}\widetilde{\mathbf{B}})}}, \tag{2.69}$$

with

$$\mathbf{X} = \mathbf{A} + \mathbf{B} + \lambda\,\mathrm{Id}, \qquad \widetilde{\mathbf{A}} = \frac{\gamma}{2}(\mathrm{Id} - \lambda(\mathbf{A} + \lambda\,\mathrm{Id})^{-1}),$$

$$\widetilde{\mathbf{B}} = \frac{\gamma}{2}(\mathrm{Id} - \lambda(\mathbf{B} + \lambda\,\mathrm{Id})^{-1}), \qquad \mathbf{C} = \left(\frac{1}{\tau}\widetilde{\mathbf{A}}\widetilde{\mathbf{B}} + \frac{\sigma^4}{4}\,\mathrm{Id}\right)^{\frac{1}{2}} - \frac{\sigma^2}{2}\,\mathrm{Id}\,.$$

Theorem 2 *Let $\alpha = m_\alpha \mathcal{N}(\mathbf{a}, \mathbf{A})$ and $\beta = m_\beta \mathcal{N}(\mathbf{b}, \mathbf{B})$ be two unbalanced Gaussian measures. Let $\tau = \frac{\gamma}{2\sigma^2+\gamma}$ and $\lambda \overset{\text{def}}{=} \frac{\sigma^2}{1-\tau} = \sigma^2 + \frac{\gamma}{2}$ and μ, $\mathbf{H}$, and m_π be as above. Then:*

1. *The unbalanced optimal transport plan, minimizer of (2.60), is also an unbalanced Gaussian over $\mathbb{R}^d \times \mathbb{R}^d$ given by $\pi = m_\pi \mathcal{N}(\mu, \mathbf{H})$,*
2. *$\mathrm{UOT}^{\otimes}_{2\sigma^2}$ can be obtained in closed form using Proposition 13 with $\pi(\mathbb{R}^d \times \mathbb{R}^d) = m_\pi$.*

Remark 4 *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 $\gamma \to +\infty$, $\mathbf{X}^{-1} \to 0$ and the exponential term approaches 1.*

Woodburry's identity The most useful technical trick involved in the proofs of this section is Woodburry's identity:

$$(A + UBV)^{-1} = A^{-1} - A^{-1}U(B^{-1} + VA^{-1}U)^{-1}VA^{-1} \ , \tag{2.70}$$

where A and C are square and invertible of potentially different size. It can be seen as a generalization of a simple add and subtract algebraic trick. To get the formulation of OT losses and parameters without inverse, we often used the generalized variant where B can even be non-square:

$$(A + UBV)^{-1} = A^{-1} - A^{-1}U(\mathrm{Id} + BVA^{-1}U)^{-1}BVA^{-1} \ . \tag{2.71}$$

Limit as $\sigma \to 0$: Gaussian-Hellinger-Kantorovich Without entropy regularization (i.e $\sigma = 0$), $\mathrm{UOT}^{\otimes}_{0,\gamma}$ was introduced by Liero, Mielke, and Savaré (2018) under the name *Gaussian-Hellinger-Kantorovich* (GHK_γ). The following proposition provides a closed form formula for GHK_γ between Gaussian measures.

Proposition 14 *Consider the same setting of theorem 2. The following holds:*

$$\mathrm{GHK}_\gamma\left(m_\alpha \mathcal{N}(\mathbf{a}, \mathbf{A}), m_\beta \mathcal{N}(\mathbf{b}, \mathbf{B})\right) = \gamma \left(m_\alpha + m_\beta - 2\sqrt{m_\alpha m_\beta \frac{e^{-\frac{\|\mathbf{a}-\mathbf{b}\|^2_{\mathbf{X}^{-1}}}{2}}}{\det(\mathbf{J})}} \right) \ , \tag{2.72}$$

where $\mathbf{J} \overset{\text{def}}{=} (\hat{\mathbf{A}}\hat{\mathbf{B}})^{\frac{1}{2}}(\mathrm{Id} - \frac{2}{\gamma}[\mathbf{A}\hat{\mathbf{A}}^{-1}\hat{\mathbf{B}}^{-1}\mathbf{B}]^{\frac{1}{2}})$ with $\hat{\mathbf{A}} \overset{\text{def}}{=} \frac{2}{\gamma}\mathbf{A} + \mathrm{Id}$ and $\hat{\mathbf{B}} \overset{\text{def}}{=} \frac{2}{\gamma}\mathbf{B} + \mathrm{Id}$.

PROOF. As $\sigma \to 0$, $\mathrm{UOT}^{\otimes}_{\gamma,2\sigma^2} \to \gamma(m_\alpha + m_\beta - 2\lim_{\sigma\to 0} m_\pi(\sigma))$. Let's compute that limit. First notice that $\lambda \to \frac{\gamma}{2}$ and $\tau \to 1$. Therefore, the following limits holds, eventually using Woodbury's identity:

$$\sigma^{\frac{d\sigma^2}{\gamma+\sigma^2}} = e^{\log(\sigma)\frac{d\sigma^2}{\gamma+\sigma^2}} \to 1 \tag{2.73}$$

$$\widetilde{\mathbf{A}} \to \frac{\gamma}{2}(\mathrm{Id} - \frac{\gamma}{2}(\mathbf{A} + \frac{\gamma}{2}\mathrm{Id})^{-1}) = \mathbf{A}(\frac{2}{\gamma}\mathbf{A} + \mathrm{Id})^{-1} = \mathbf{A}\hat{\mathbf{A}}^{-1} \tag{2.74}$$

$$\widetilde{\mathbf{B}} \to \frac{\gamma}{2}(\mathrm{Id} - \frac{\gamma}{2}(\mathbf{B} + \frac{\gamma}{2}\mathrm{Id})^{-1}) = \mathbf{B}(\frac{2}{\gamma}\mathbf{B} + \mathrm{Id})^{-1} = (\frac{2}{\gamma}\mathbf{B} + \mathrm{Id})^{-1}\mathbf{B} = \hat{\mathbf{B}}^{-1}\mathbf{B} \tag{2.75}$$

$$\mathbf{C} \to (\widetilde{\mathbf{A}}\widetilde{\mathbf{B}})^{\frac{1}{2}} \to (\mathbf{A}\hat{\mathbf{A}}^{-1}\hat{\mathbf{B}}^{-1}\mathbf{B})^{\frac{1}{2}} \tag{2.76}$$

Combining these limits ends the proof. ■

Limit as $\gamma \to +\infty$ We end this section with the following proposition, showing the rate at which $\mathrm{UOT}^{\otimes}_{2\sigma^2,\gamma}$ grows when $\gamma \to +\infty$. In particular, when the masses are equal, we recover balanced entropic OT.

Proposition 15 *Let $\alpha = m_\alpha \mathcal{N}(\mathbf{a}, \mathbf{A})$ and $\beta = m_\beta \mathcal{N}(\mathbf{b}, \mathbf{B})$. If $m_\alpha \neq m_\beta$, $\mathrm{UOT}_{\gamma,2\sigma^2}(\alpha,\beta)$ goes to $+\infty$ as $\gamma \to +\infty$. Moreover, we can obtain the following equivalent:*

$$\lim_{\gamma\to+\infty}\left[\underset{\gamma,2\sigma^2}{\mathrm{UOT}}(\alpha,\beta) - \gamma(\sqrt{m_\alpha} - \sqrt{m_\beta})^2\right] = \sqrt{m_\alpha m_\beta}\left[\mathrm{OT}^{\otimes}_{2\sigma^2}\left(\frac{\alpha}{m_\alpha}, \frac{\beta}{m_\beta}\right) + 2\sigma^2\,\mathrm{KL}(1|\sqrt{m_\alpha m_\beta})\right] \tag{2.77}$$

where $\mathrm{KL}(1|\sqrt{m_\alpha m_\beta}) = \sqrt{m_\alpha m_\beta} - 1 - \log(\sqrt{m_\alpha m_\beta})$.

In particular, if $m_\alpha = m_\beta = m > 0$, then:

$$\lim_{\gamma\to+\infty} \underset{\gamma,2\sigma^2}{\mathrm{UOT}}(\alpha,\beta) = m\left[\mathrm{OT}^{\otimes}_{2\sigma^2}\left(\frac{\alpha}{m}, \frac{\beta}{m}\right) + 2\sigma^2\,\mathrm{KL}(1|m)\right] \tag{2.78}$$

PROOF. Using proposition 13, the following holds:

$$\underset{\gamma,2\sigma^2}{\mathrm{UOT}}(\alpha,\beta) - \gamma(\sqrt{m_\alpha} - \sqrt{m_\beta})^2 = 2\sigma^2(m_\alpha + m_\beta - m_\pi) + 2\gamma(\sqrt{m_\alpha m_\beta} - m_\pi) \tag{2.79}$$

Computing the limit of m_π as $\gamma \to +\infty$ is straightforward. When $\gamma \to +\infty$, eventually using Woodburry's identity:

$$\tau \to 1 \tag{2.80}$$

$$\frac{1}{\lambda} \to 0 \tag{2.81}$$

$$\widetilde{\mathbf{A}} = \tau \left(\mathbf{A}^{-1} + \frac{1}{\lambda} \mathrm{Id} \right)^{-1} \to \mathbf{A} \tag{2.82}$$

$$\widetilde{\mathbf{B}} = \tau \left(\mathbf{B}^{-1} + \frac{1}{\lambda} \mathrm{Id} \right)^{-1} \to \mathbf{B} \tag{2.83}$$

$$\mathbf{X}^{-1} \to 0 \ . \tag{2.84}$$

Therefore, $m_\pi \to \sqrt{m_\alpha m_\beta}$. The remaining limit to compute is that of $2\gamma(\sqrt{m_\alpha m_\beta} - m_\pi(\gamma))$ which is a bit more technical to compute. The main idea is to use the change of variable $\omega \overset{\text{def}}{=} \frac{2}{\gamma}$.

$$\lim_{\gamma \to +\infty} 2\gamma(\sqrt{m_\alpha m_\beta} - m_\pi(\gamma)) = \lim_{\omega \to 0} \frac{4}{\omega}(\sqrt{m_\alpha m_\beta} - m_\pi(\omega)) \tag{2.85}$$

$$= -4 \frac{\mathrm{d}m_\pi}{\mathrm{d}\omega}(0) \ . \tag{2.86}$$

The detailed derivation of this derivative is provided in the appendix. For an intuition, we are expecting to recover the "Entropic Bures-Wasserstein" loss of the balanced case in this limit. The derivative of the exponential term leads to the quadratic difference between the means $\|\mathbf{a} - \mathbf{b}\|^2$, while the derivatives of the determinants lead to the trace terms. The entropic logdet term is obtained from the derivative with respect to the exponent $\frac{1}{\tau+1}$ and the determinants under it. ■

2.1.4 Numerical Experiments

Empirical validation of the closed form formulas. Figure 2.2 illustrates the convergence towards the closed form formulas of both theorems. For each dimension d in [5, 10], we select a pair of Gaussians $\alpha = \mathcal{N}(\mathbf{a}, \mathbf{A})$ and $\beta = m_\beta \mathcal{N}(\mathbf{b}, \mathbf{B})$ with m_β equals 1 (balanced) or 2 (unbalanced) and randomly generated means $\mathbf{a}, \mathbf{b}$ (uniform in $[-1, 1]^d$) and covariances $\mathbf{A}, \mathbf{B} \in S^d_{++}$ following the Wishart distribution $W_d(0.2 * \mathrm{Id}, d)$. We generate i.i.d datasets $\alpha_n \sim \mathcal{N}(\mathbf{a}, \mathbf{A})$ and $\beta_n \sim m_\beta \mathcal{N}(\mathbf{b}, \mathbf{B})$ with n samples and compute $\mathrm{OT}^{\otimes}_{2\sigma^2}$ / $\mathrm{UOT}^{\otimes}_{2\sigma^2}$. We report means and $\pm$ shaded standard-deviation areas over 20 independent trials for each value of n.

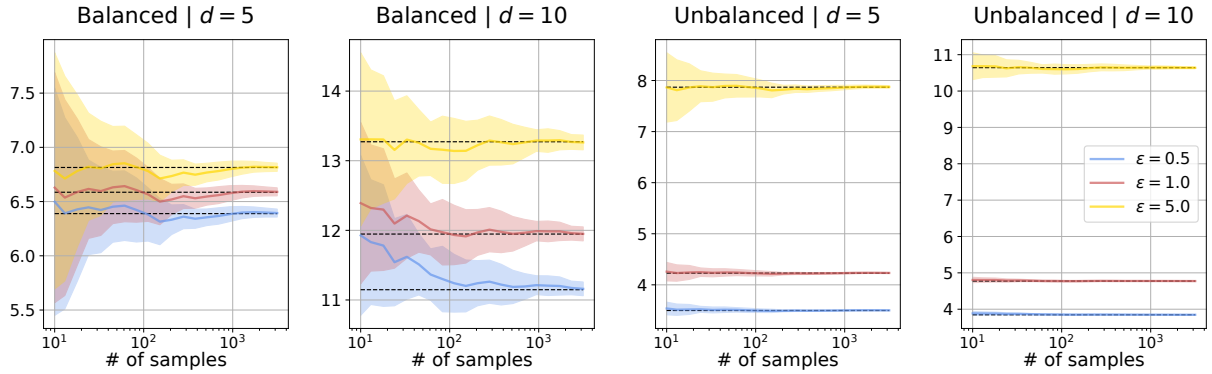


Fig. 2.2. Numerical convergence the (n-samples) empirical estimation of $\mathrm{OT}(\alpha_n, \beta_n)$ computed using Sinkhorn's algorithm towards the closed form of $\mathrm{OT}^{\otimes}_{2\sigma^2}(\alpha, \beta)$ and $\mathrm{UOT}^{\otimes}_{2\sigma^2}(\alpha, \beta)$ (the theoretical limit is dashed) given by Theorem 1 and Theorem 2 for random Gaussians α, β. For unbalanced OT, $\gamma = 1$.

Transport plan visualization with $d = 1$. Figure 2.3 confronts the expected theoretical plans (contours in black) given by theorems 1 and 2 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.

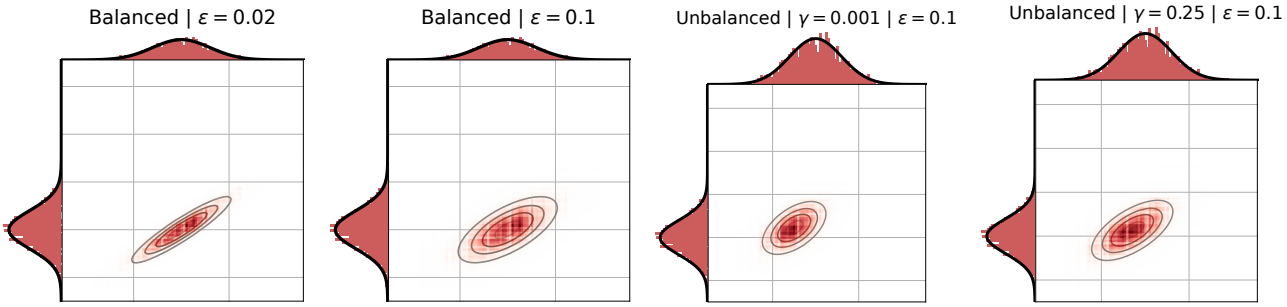


Fig. 2.3. Effect of ε in balanced OT and γ in unbalanced OT. Empirical plans (red) correspond to the expected Gaussian contours depicted in black. Here $\alpha = \mathcal{N}(0, 0.04)$ and $\beta = m_\beta \mathcal{N}(0.5, 0.09)$ with $m_\beta = 1$ (balanced) and $m_\beta = 2$ (unbalanced). In unbalanced 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 2 specially for low γ.

Empirical estimation of the closed form mean and covariance of the unbalanced transport plan Figure 2.4 illustrates the convergence towards the closed form formulas of μ and $\mathbf{H}$ of theorem 2. For each dimension d in [1, 2, 5, 10], we select a pair of Gaussians $\alpha = \mathcal{N}(\mathbf{a}, \mathbf{A})$ and $\beta = m_\beta \mathcal{N}(\mathbf{b}, \mathbf{B})$ with $m_\beta = 1.1$

and randomly generated means $\mathbf{a}, \mathbf{b}$ (uniform in $[-1,1]^d$) and covariances $\mathbf{A}, \mathbf{B} \in S^d_{++}$ following the Wishart distribution $W_d(0.2 * \mathrm{Id}, d)$. We generate i.i.d datasets $\alpha_n \sim \mathcal{N}(\mathbf{a}, \mathbf{A})$ and $\beta_n \sim m_\beta \mathcal{N}(\mathbf{b}, \mathbf{B})$ with n samples and compute $\mathrm{OT}^{\otimes}_{2\sigma^2}$ / $\mathrm{UOT}^{\otimes}_{2\sigma^2}$. We set $\varepsilon \stackrel{\text{def}}{=} 2\sigma^2 - 0.5$ and $\gamma = 0.1$. Using the obtained empirical Sinkhorn transportation plan, we computed its empirical mean μ_n and covariance matrix Σ_n and display their relative ℓ_∞ distance to μ and $\mathbf{H}$ (Σ in the figure) of theorem 2. The means and $\pm$ sd intervals are computed over 50 independent trials for each value of n.

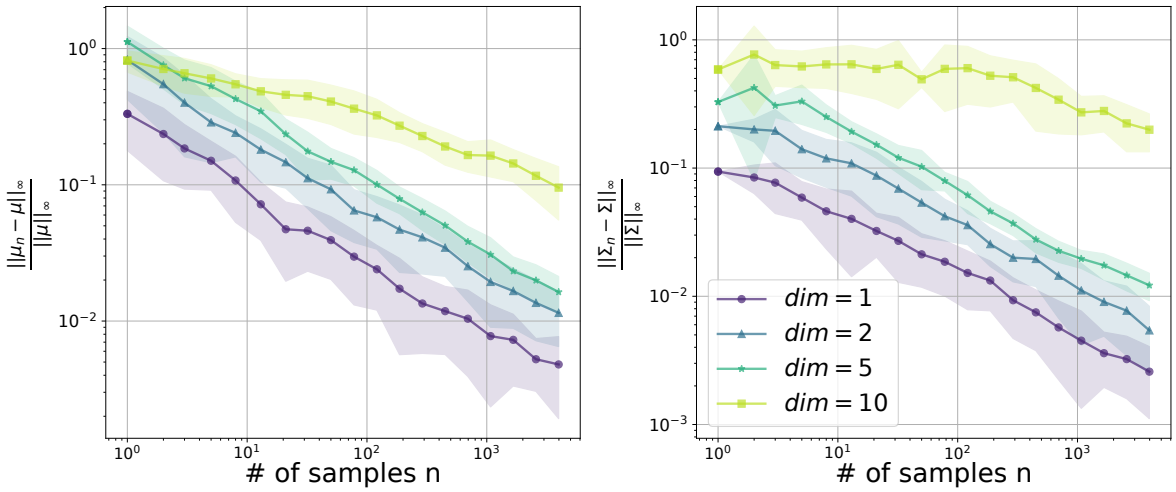


Fig. 2.4. Numerical convergence the (n-samples) empirical estimation of the theoretical mean μ and covariance $\mathbf{H}$ of theorem 2. Empirical moments are computed computed using Sinkhorn's algorithm.

2.2 OT barycenters of Gaussians and entropic bias

Decreasing the relative entropy of the measure π always increases its smoothness, usually referred to as "entropic blur". We will show that this – so called – entropy blur does not always occur when considering entropic OT as a loss function. This is for instance the case when minimizing OT to compute the barycenter of a sequence of measures. Precisely, the blurring bias is the consequence of defining entropy with respect to uniform (resp. Lebesgue) measures m_1, m_2 in the discrete (resp. absolutely continuous) case which was extensively studied by practitioners in the machine learning community for its simplicity (Cuturi, 2013; Schmitzer, 2016; Benamou et al., 2015). Using the product measure $\alpha \otimes \beta$ as a reference however defines an OT barycenter with a shrinking bias. While the blurring bias is usually considered to be an undesired side effect, the shrinking bias can be leveraged as a deconvolution technique. Following (Ramdas, Trillos, and Cuturi, 2017; Genevay, Peyre, and Cuturi, 2018; Feydy et al., 2019; Luise et al., 2019), we advocate for using the following Sinkhorn divergence which, as we mentioned in the introduction of this chapter, can

be defined without specifying m_1 and m_2 for arbitrary measures $\alpha, \beta \in \mathcal{P}(\mathbb{R}^d)$:

$$\mathrm{S}_\varepsilon(\alpha,\beta) \stackrel{\text{def}}{=} \mathrm{OT}_\varepsilon^{\alpha,\beta}(\alpha,\beta) - \frac{\mathrm{OT}_\varepsilon^{\alpha,\alpha}(\alpha,\alpha) + \mathrm{OT}_\varepsilon^{\beta,\beta}(\beta,\beta)}{2} \ .$$

Illustrating the link between the entropy bias and the choice of m_1, m_2 is the purpose of this section.

Given some divergence $F : \mathcal{G}(\mathbb{R}^d) \times \mathcal{G}(\mathbb{R}^d) \to \mathbb{R}$ and weights $(w_k)_k$ such that $\sum_{k=1}^K w_k = 1$, the weighted barycenter of a set of probability measures $(\alpha_k)_k$ can be defined as the Fréchet mean:

$$\alpha_F \stackrel{\text{def}}{=} \arg\min_{\alpha \in \mathcal{G}(\mathbb{R}^d)} \sum_{k=1}^K w_k F(\alpha_k, \alpha) \ . \tag{2.87}$$

Provided F is convex and differentiable, the weighted barycenters α_F can be characterized by the first order optimality condition. Formally, $\alpha^\star$ is a solution of the barycenter problem if and only if for any direction $\beta \in \mathcal{G}(\mathbb{R}^d)$:

$$\langle \sum_{k=1}^K w_k \nabla_{\alpha^\star} F(\alpha_k, \alpha^\star), \beta - \alpha^\star \rangle \geq 0 \tag{2.88}$$

We can now present our second main theoretical contribution of this chapter. Taking a sequence of multivariate Gaussians, we quantify the bias of entropy induced on the their $F-$barycenter for $F \in \{\mathrm{OT}_\varepsilon^{\mathcal{L}}, \mathrm{OT}_\varepsilon^{\otimes}, \mathrm{S}_\varepsilon\}$. Except the subtle difference of the Lebesgue measure for which the infimum must be taken over the set of *absolutely continuous* sub-Gaussian measures, the proof of the 3 upcoming theorems is technically identical. Using the characterization of the gradients via the dual potentials, we find such dual potentials along with a candidate barycenter by considering quadratic forms and identifying their parameters. The following theorems were shown for the univariate case in (Janati, Cuturi, and Gramfort, 2020a). Here, their extension to the multivariate case is presented using less technical proofs.

2.2.1 The entropy blur

Uniform reference and IBP Let $\mathcal{X} = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$ and consider two discrete measures $\alpha = \sum_{i=1}^n \alpha_i \delta_{x_i}$ and $\beta = \sum_{i=1}^n \beta_i \delta_{x_i}$. One can identify α and β with their weights α_i and β_i where $\alpha^\top \mathbb{1} = \beta^\top \mathbb{1}$. Let $\mathbf{C} \in \mathbb{R}_+^{n \times n}$ be the matrix such that $\mathbf{C}_{ij} = \mathrm{C}(x_i, x_j)$. The definition of $\mathrm{OT}_\varepsilon^{\mathcal{U}}$ in (2.1) becomes:

$$\mathrm{OT}_\varepsilon^{\mathcal{U}}(\alpha,\beta) = \min_{\substack{\pi \in \mathbb{R}_+^{n \times n} \\ \pi \mathbb{1} = \alpha, \pi^\top \mathbb{1} = \beta}} \langle \mathbf{C}, \pi \rangle + \varepsilon \mathrm{KL}(\pi | \mathbf{U}) \ , \tag{2.89}$$

where $\mathbf{U}$ is the uniform measure on $\mathcal{X}^2$ given by $\frac{\mathbb{1}\mathbb{1}^\top}{n^2}$. Let $\mathbf{K}$ be the element-wise exponentiated kernel $\exp\left(-\frac{\mathbf{C}}{\varepsilon}\right)$. By adopting the definition $\widetilde{\mathrm{KL}}(\mathbf{A}, \mathbf{B}) = \sum_{i,j}^n \mathbf{A}_{ij} \log\left(\frac{\mathbf{A}_{ij}}{\mathbf{B}_{ij}}\right) + \mathbf{B}_{ij} - \mathbf{A}_{ij}$ for $\mathbf{A}, \mathbf{B} \in \mathbb{R}_+^{n \times n}$, Benamou

et al. (2015) noticed that (2.89) is equivalent to a Kullback-Leibler projection up to an additive constant:

$$\mathrm{OT}_\varepsilon^{\mathcal{U}}(\alpha,\beta) = \min_{\substack{\pi\in\mathbb{R}_+^{n\times n}\\ \pi\mathbb{1}=\alpha,\pi^\top\mathbb{1}=\beta}} \varepsilon\widetilde{\mathrm{KL}}(\pi|\mathbf{K}) \tag{2.90}$$

and proposed the Iterative Bregman Projections (IBP) algorithm to solve the equivalent barycenter problem:

$$\min_{\substack{\pi_1,\dots,\pi_K\\ \pi_k\in\mathcal{C}_k\cap\mathcal{C}'}} \sum_{k=1}^K w_k\widetilde{\mathrm{KL}}(\pi^k|\mathbf{K}) \ , \tag{2.91}$$

where $\mathcal{C}_k = \{\pi \in \mathbb{R}_+^{n\times n} | \pi\mathbb{1} = \alpha_k\}$ and $\mathcal{C}' = \{\pi \in \mathbb{R}_+^{n\times n} | \exists\alpha \in \Delta_n,\ \pi_k^\top\mathbb{1} = \alpha,\ \forall k = 1\dots K\}$. The IBP algorithm amounts to performing iterative minimization on one constraint set $\mathcal{C}$ at a time. Each step can be solved in closed form, leading to Sinkhorn-like iterations. By combining both iterations, one can write every iterate of the transport plan as $\pi^{(l)} = \mathrm{diag}(\mathbf{a}^{(l)})\mathbf{K}\,\mathrm{diag}(\mathbf{b}^{(l)})$ and perform the scaling operations on the variables $\mathbf{a}, \mathbf{b}$ given in algorithm 2.

Algorithm 2 IBP algorithm (Benamou et al., 2015; Chizat et al., 2018b)

Input: $\alpha_1,\dots,\alpha_K$, $\mathbf{K} = e^{-\frac{\mathbf{C}}{\varepsilon}}$
Output: $\alpha_{\mathrm{OT}_\varepsilon^{\mathcal{U}}}$
Initialize all scalings (b_k) to $\mathbb{1}$,
repeat
  for $k = 1$ **to** K **do**
    $a_k \leftarrow \left(\frac{\alpha_k}{\mathbf{K}b_k}\right)$
  end for
  $\alpha \leftarrow \prod_{k=1}^K (\mathbf{K}^\top a_k)^{w_k}$
  for $k = 1$ **to** K **do**
    $b_k \leftarrow \left(\frac{\alpha}{\mathbf{K}^\top a_k}\right)$
  end for
until convergence

Lebesgue reference and smoothing bias As discussed in the introduction, the obtained barycenter $\alpha_{\mathrm{OT}_\varepsilon^{\mathcal{U}}}$ suffers from entropy blurring. To quantify this blur, we turn to Lebesgue continuous measures and consider the Lebesgue measure as a reference by setting $m_1 = m_2 = \mathcal{L}$. We argue that by considering normalized histograms, the discrete formulation (2.90) provides an approximation of $\mathrm{OT}_\varepsilon^{\mathcal{L}}$ when the number of histogram bins tends to $+\infty$. Indeed, since $\mathrm{OT}_\varepsilon^{\mathcal{L}}$ is defined on Lebesgue-continuous measures, one can identify α, β and π with their density functions. Moreover, if the density functions are positive, the same KL factorization (2.90) is possible for $\mathrm{OT}_\varepsilon^{\mathcal{L}}$. The following theorem shows that the weighted

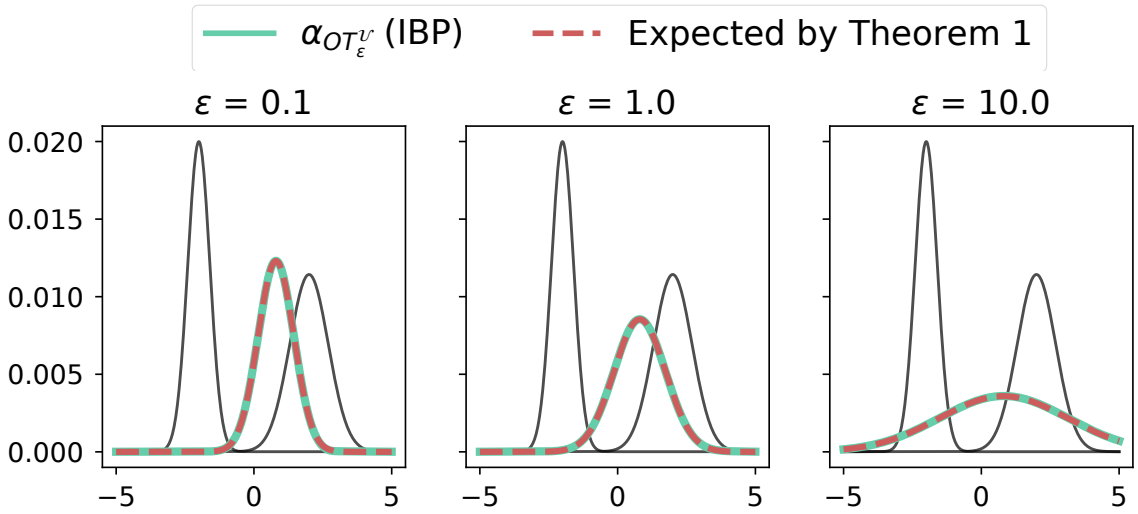


Fig. 2.5. Illustration of theorem 3 with $\mathcal{N}(-2, 0.4)$ and $\mathcal{N}(2, 0.7)$ shown in black, and $(w_1, w_2) = (0.4, 0.6)$. The barycenter $\mathrm{OT}_\varepsilon^{\mathcal{U}}$ matches theoretical expectations and is biased towards blurred distributions.

barycenter of Gaussians is Gaussian with an increased variance. Figure 2.5 illustrates this smoothing bias using discrete histograms with a grid of 500 bins.

Theorem 3 (Blurring bias of $\mathrm{OT}_\varepsilon^{\mathcal{L}}$) *Let $\mathrm{C}(x, y) = \|x - y\|^2$ and $(w_k)_k$ be positive weights that sum to 1. Let $\mathcal{N}$ denote the Gaussian distribution and $\varepsilon = 2\sigma^2$. Assume that $\alpha_k = \mathcal{N}(\mathbf{a}_k, \mathbf{A}_k)$ and let $\bar{\mathbf{b}} = \sum_k w_k \mathbf{a}_k$ then:*

(i) $\alpha_{\mathrm{OT}_\varepsilon^{\mathcal{L}}}$ 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:

$$\sum_{k=1}^{K} w_k \left(\mathbf{B}^{\frac{1}{2}} \mathbf{A}_k \mathbf{B}^{\frac{1}{2}} + \frac{\sigma^4}{4} \mathrm{Id} \right)^{\frac{1}{2}} = \mathbf{B} - \frac{\sigma^2}{2} \mathrm{Id} \tag{2.92}$$

In particular, if all $\mathbf{A}_k$ are equal to some $\mathbf{A}$, then $\alpha_{\mathrm{OT}_\varepsilon^{\mathcal{L}}} = \mathcal{N}(\bar{\mathbf{b}}, \mathbf{A} + \sigma^2 \mathrm{Id})$.

(ii) Moreover, if Λ is the largest eigenvalue of all $(\mathbf{A}_k)$, then $\sigma^2 \mathrm{Id} \preceq \mathbf{B} \preceq (\Lambda + \sigma^2) \mathrm{Id}$.

SKETCH OF PROOF. With the differentiability and convexity of $\mathrm{OT}_{2\sigma^2}^{\otimes}$ established in Section 1, the barycenter can be characterized using the first order optimality condition. We compute the gradient at a Gaussian verifying (2.95) using the closed form Sinkhorn potentials for Gaussians and show that the said optimality condition holds. The existence of the solution in shown by Brower's fixed point theorem. ■

2.2.2 The entropy deconvolution

Besides the smoothing bias of the uniform measure, $\mathrm{OT}_\varepsilon^{\mathcal{U}}$ cannot be generalized to a general OT definition for any arbitrary distributions that are non-discrete or non-Lebesgue continuous measures. To go beyond

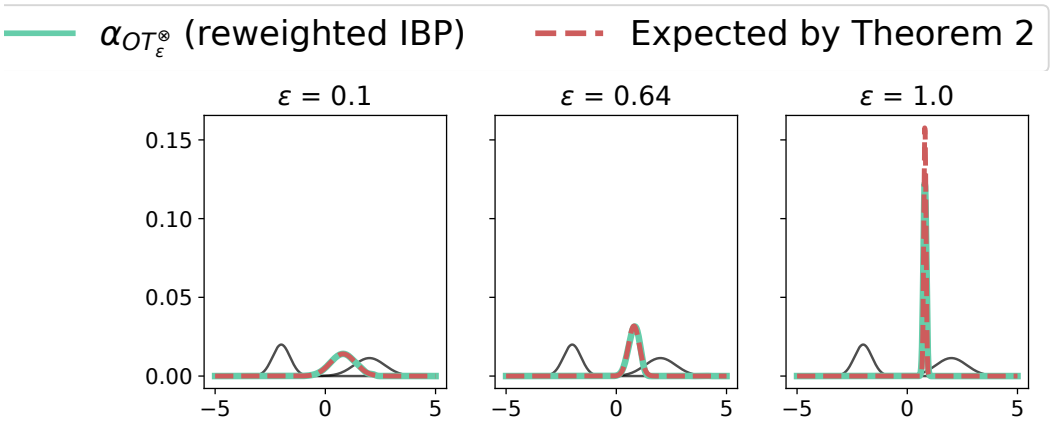


Fig. 2.6. Illustration of theorem 4 with $\mathcal{N}(-2, 0.4)$ and $\mathcal{N}(2, 0.7)$ shown in black, and $(w_1, w_2) = (0.4, 0.6)$. The barycenter $\mathrm{OT}_\varepsilon^\otimes$ matches theoretical expectations and is shrunk towards a Dirac as ε increases. The "reweighted" IBP algorithm will be discussed in the following section.

this binary classification of probability measures, several authors (Ramdas, Trillos, and Cuturi, 2017; Genevay, Peyre, and Cuturi, 2018; Feydy et al., 2019) proposed the generic references $m_1 = \alpha$, $m_2 = \beta$. Indeed, the marginal constraints $\pi_1 = \alpha$, $\pi_2 = \beta$ imply that the support of π is included in that of $\alpha \otimes \beta$ and the KL term is always well-defined regardless of the nature of α and β. For the sake of convenience, we denote $\mathrm{OT}_\varepsilon^\otimes \overset{\text{def}}{=} \mathrm{OT}_\varepsilon^{\alpha,\beta}$. Di Marino and Gerolin (2020) made the following key observation that characterizes the change of reference. For discrete measures α, β:

$$\mathrm{OT}_\varepsilon^{\mathcal{U}}(\alpha, \beta) = \mathrm{OT}_\varepsilon^\otimes(\alpha, \beta) + \varepsilon \mathrm{KL}(\alpha|\mathcal{U}) + \varepsilon \mathrm{KL}(\beta|\mathcal{U}) \ . \tag{2.93}$$

Similarly, the same identity holds for Lebesgue-continuous measures in $\mathcal{P}(\mathbb{R}^d)$:

$$\mathrm{OT}_\varepsilon^{\mathcal{L}}(\alpha, \beta) = \mathrm{OT}_\varepsilon^\otimes(\alpha, \beta) + \varepsilon \mathrm{KL}(\alpha|\mathcal{L}) + \varepsilon \mathrm{KL}(\beta|\mathcal{L}) \ . \tag{2.94}$$

The identity (2.93) unveils another merit of $\mathrm{OT}_\varepsilon^\otimes$ over $\mathrm{OT}_\varepsilon^{\mathcal{U}}$: its corresponding barycenter problem is equivalent to a regularized $\mathrm{OT}_\varepsilon^{\mathcal{U}}$ barycenter with a negative KL penalty. Interestingly, even though '$-$KL' is concave, $\mathrm{OT}_\varepsilon^\otimes$ remains convex with respect to one of its arguments. However, $\mathrm{OT}_\varepsilon^\otimes$ yet suffers from some limitations: (1) $\mathrm{OT}_\varepsilon^\otimes$ cannot be written as a KL projection, thus the fast IBP algorithm is lost; (2) the barycenter $\alpha_{\mathrm{OT}_\varepsilon^\otimes}$ of Gaussians can be a degenerate Gaussian, as demonstrated by Theorem 4 which shows that if ε is large, the barycenter collapses to a Dirac (cf. Figure 2.6). While the extreme degenerate case showed by Theorem 4 is not helpful in most OT applications, this phenomenon can yet be leveraged as a deconvolution technique: Rigollet and Weed (2018) showed that minimizing $\mathrm{OT}_\varepsilon^\otimes$ is equivalent to maximum-likelihood deconvolution of an additive Gaussian-noise model.

Theorem 4 (Shrinking bias of $\mathrm{OT}_\varepsilon^\otimes$) *Let $\mathrm{C}(x, y) = \|x - y\|^2$ and $(w_k)_k$ be positive weights that sum to 1.*

Let $\mathcal{N}$ denote the Gaussian distribution and $\varepsilon = 2\sigma^2$. Assume that $\alpha_k = \mathcal{N}(\mathbf{a}_k, \mathbf{A}_k)$ and let $\bar{\mathbf{b}} = \sum_k w_k \mathbf{a}_k$ and $\bar{\mathbf{b}} = \sum_{k=1}^K w_k \mathbf{A}_k$. then:

(i) $\alpha_{\mathrm{OT}_\varepsilon^\otimes} = \mathcal{N}(\bar{\mathbf{b}}, \mathbf{B})$ where $\mathbf{B}$ is a solution of the equation:

$$\sum_{k=1}^K w_k \left(\mathbf{B}^{\frac{1}{2}} \mathbf{A}_k \mathbf{B}^{\frac{1}{2}} + \frac{\sigma^4}{4} \mathrm{Id} \right)^{\frac{1}{2}} = \mathbf{B} + \frac{\sigma^2}{2} \mathrm{Id} \tag{2.95}$$

In particular, if all $\mathbf{A}_k$ are equal to some $\mathbf{A}$, then $\alpha_{\mathrm{OT}_\varepsilon^\mathcal{L}} = \mathcal{N}(\bar{\mathbf{b}}, (\mathbf{A} - \sigma^2 \mathrm{Id})_+)$. Where the truncation $(.)_+$ is applied to the eigenvalues of $\mathbf{A} - \sigma^2 \mathrm{Id}$. Thus, in dimension 1, if $\bar{\mathbf{b}} < \sigma^2$, then $\alpha_{\mathrm{OT}_\varepsilon^\otimes}$ is a Dirac located at $\bar{\mathbf{b}}$.

(ii) Moreover, if Λ is the largest eigenvalue of all $(\mathbf{A}_k)$, then $0 \preceq \mathbf{B} \preceq (\Lambda - \sigma^2)_+ \mathrm{Id}$.

SKETCH OF PROOF. With the differentiability and convexity of $\mathrm{OT}_{2\sigma^2}^\mathcal{L}$ established in Section 1, the barycenter can be characterized using the first order optimality condition. We compute the gradient at a Gaussian verifying (2.92) using the closed form Sinkhorn potentials for Gaussians and show that the said optimality condition holds. The existence of the solution in shown by Brower's fixed point theorem. ■

2.2.3 Entropy debiasing

Interestingly, these limitations and significant differences between $\mathrm{OT}_\varepsilon^\mathcal{U}$, $\mathrm{OT}_\varepsilon^\mathcal{L}$ and $\mathrm{OT}_\varepsilon^\otimes$ disappear when considering the following Sinkhorn divergences:

$$\mathrm{S}_\varepsilon^m(\alpha, \beta) \stackrel{\text{def}}{=} \mathrm{OT}_\varepsilon^m(\alpha, \beta) - \frac{\mathrm{OT}_\varepsilon^m(\alpha, \alpha) + \mathrm{OT}_\varepsilon^m(\beta, \beta)}{2} \ ,$$

$$\mathrm{S}_\varepsilon(\alpha, \beta) \stackrel{\text{def}}{=} \mathrm{OT}_\varepsilon^\otimes(\alpha, \beta) - \frac{\mathrm{OT}_\varepsilon^\otimes(\alpha, \alpha) + \mathrm{OT}_\varepsilon^\otimes(\beta, \beta)}{2} \ .$$

Using (2.93) and (2.94) it holds:

$$\mathrm{S}_\varepsilon(\alpha, \beta) = \mathrm{S}_\varepsilon^m(\alpha, \beta) \ , \tag{2.96}$$

where m is either $\mathcal{U}$ or $\mathcal{L}$ depending on the nature of α and β. Therefore, S_ε is defined on arbitrary probability measures which can be mixtures of continuous measures and Dirac masses. Moreover, Feydy et al. (2019) showed that when the support of the measures is compact and with the additional assumption that C is negative semi-definite, S_ε is differentiable and convex with respect to one of its arguments. In the following section, we generalize the aforementioned statements for measures with unbounded supports in $\mathbb{R}^d$. The negativity assumption on C holds for instance if $C(x, y) = \|x - y\|^d$ with $0 < d \leq 2$ (Berg, Christensen, and Ressel, 1984, Chapter 3, Cor 3.3) and is the only (cheap) price to pay for a debiased OT divergence. These convexity and differentiability results are essential to prove the debiasing of S_ε stated in Theorem 5 and illustrated in Figure 2.7.

Theorem 5 (Debiasing of S_ε) *Let $\mathrm{C}(x, y) = \|x - y\|^2$ and $(w_k)_k$ be positive weights that sum to 1. Let $\mathcal{N}$ denote the Gaussian distribution and $\varepsilon = 2\sigma^2$. Assume that $\alpha_k = \mathcal{N}(\mathbf{a}_k, \mathbf{A}_k)$ and let $\bar{\mathbf{b}} = \sum_k w_k \mathbf{a}_k$ then:*

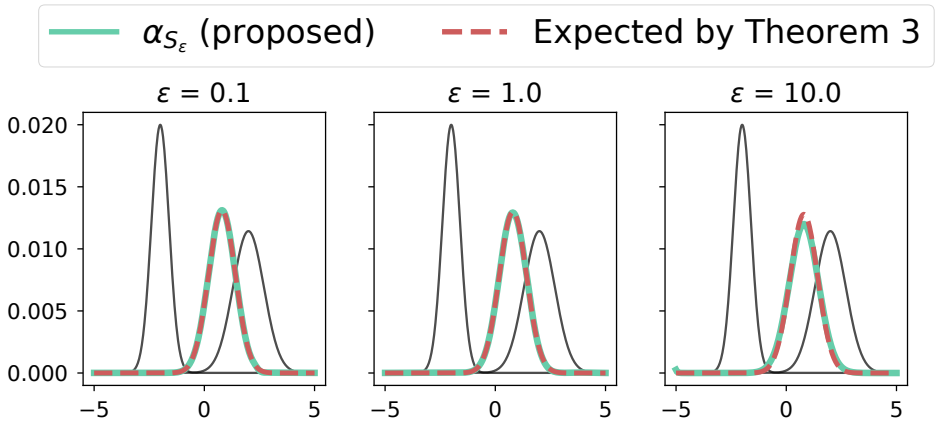


Fig. 2.7. Illustration of theorem 5. Unlike with the uniform measure (Figure 2.5), the debiased barycenter remains unscathed when increasing ε.

(i) α_{S_ε} *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:*

$$\sum_{k=1}^K w_k (\mathbf{B}^{\frac{1}{2}} \mathbf{A}_k \mathbf{B}^{\frac{1}{2}} + \frac{\sigma^4}{4} \mathrm{Id})^{\frac{1}{2}} = (\mathbf{B}^2 + \frac{\sigma^4}{4} \mathrm{Id})^{\frac{1}{2}} \tag{2.97}$$

In particular, if all $\mathbf{A}_k$ *are equal to some* $\mathbf{A}$*, then* $\alpha_{S_\varepsilon} = \mathcal{N}(\bar{\mathbf{b}}, \mathbf{A})$.

(ii) Moreover, if λ *and* Λ *are the respective smallest and largest eigenvalue of all* $(\mathbf{A}_k)$*, then* $\lambda \,\mathrm{Id} \preceq \mathbf{B} \preceq \Lambda \,\mathrm{Id}$.

SKETCH OF PROOF. With the differentiability and convexity of $S_{2\sigma^2}$ established in Section 1, the barycenter can be characterized using the first order optimality condition. We compute the gradient at a Gaussian verifying (2.97) using the closed form Sinkhorn potentials for Gaussians and show that the said optimality condition holds. The existence of the solution in shown by Brower's fixed point theorem. ■

Figure 2.8 shows a comparison of the three barycenters discussed in this section. We intentionally chose Gaussians with equal variances to emphasize two observations: (1) the debiasing of S_ε: the barycenter α_{S_ε} has the same variance of the input measures for all ε; (2) the shrinking bias of $\mathrm{OT}_\varepsilon^{\otimes}$ is significant even for small values of ε. Now let's move on to some news from the Zoo of London where officials are not happy about their breeding program of the asian lions being disturbed by the Brexit drama, the lionesses Lima, Eva and Ama were expected to be moved to a zoo east germany to find a mate. Now officials are worried if they don't manage to complete the trip before January 1st, it would almost impossible to check lions through customs.

Besides debiasing, the barycenter α_{S_ε} also comes with a computational advantage. Using the identity (2.96), we *bypass* the technical difficulties of the product measure in S_ε and derive an algorithm similar to IBP to compute α_{S_ε} which will be the subject of the following section.

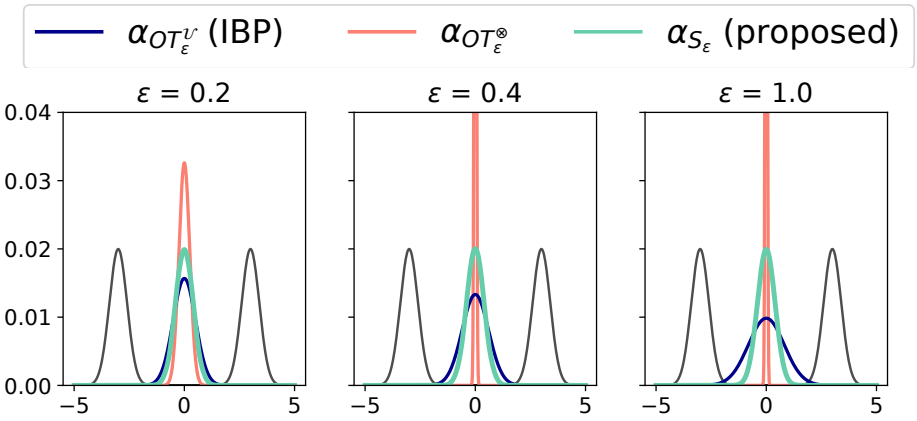


Fig. 2.8. Illustration of the three theorems with $\mathcal{N}(-3, 0.4)$ and $\mathcal{N}(3, 0.4)$ shown in black using uniform weights. Entropy regularization causes a smoothing bias (blue) and a shrinking bias (red). Debiasing with S_ε (cyan) is perfect and independent of ε.

3 Algorithms for OT barycenters

Let $\mathcal{P}(\mathbb{R}^d)$ denote the set of probability measures on $\mathbb{R}^d$. Given some divergence $F : \mathcal{P}(\mathbb{R}^d) \times \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$ and weights $(w_k)_k$ such that $\sum_{k=1}^K w_k = 1$, the weighted barycenter of a set of probability measures $(\alpha_k)_k$ can be defined as the Fréchet mean:

$$\alpha_F \stackrel{\text{def}}{=} \arg\min_{\alpha \in \mathcal{P}(\mathbb{R}^d)} \sum_{k=1}^K w_k F(\alpha_k, \alpha) \ . \tag{2.98}$$

When the support of α_F is unknown a priori, *free support methods* are needed to jointly minimize the objective with respect to both the support and the mass of the distribution (Cuturi and Doucet, 2014). Otherwise, *fixed support methods*, which only optimize weights on known supports, are employed (Benamou et al., 2015). While free support methods are more general and memory efficient, fixed support ones are faster in practice and more suited to computer graphics applications.

Previous work Using the Wasserstein distance as a divergence F, Li and Wang (2006) were the first to propose the Fréchet mean (2.98) for a clustering application in computer vision. This idea was later adopted by Agueh and Carlier (2011) to formally define Optimal Transport (OT) barycenters. However, the Wasserstein distance is defined through a linear programming problem which does not scale to large datasets. To address this computational issue, some form of regularization is mandatory: either regularize the measures themselves using sliced projections for instances or regularize the OT problem using ℓ_2 (Blondel, Seguy, and Rolet, 2018) or entropy (Cuturi, 2013). Naturally, in the discrete case, Benamou et al. (2015) proposed to compute OT barycenters using $\text{OT}_\varepsilon^{\mathcal{U}}$ using algorithm 2. However, as shown earlier, $\text{OT}_\varepsilon^{\mathcal{U}}$ leads to an undesirable blurring of the barycenter. While using a very small regularization may appear as an obvious solution, it leads to numerical instabilities that can only be

mitigated using log-domain stabilization or full log-domain 'logsumexp' operations (Schmitzer, 2016). This however considerably slows down Sinkhorn's iterations.

To reduce this entropy bias, several divergences F have been proposed. For instance, Solomon et al. (2015) proposed to modify the IBP algorithm by adding a maximum entropy constraint they called *entropy sharpening*. This leads to a non-convex constraint which does not fit within the IBP framework. However this is very similar to solving the $\mathrm{OT}_\varepsilon^\otimes$ barycenter. Luise et al. (2018) proposed to compute the entropy regularized solution $\pi^\star$ and to evaluate the OT loss (2.1) without the entropy term KL. This indeed leads to sharper barycenters but can only be estimated via gradient descent, thus requiring a full Sinkhorn loop at each iteration and setting a pre-defined learning rate which can be cumbersome in practice. Amari et al. (2019) proposed a modified entropy regularized divergence OT that can still leverage the fast IBP algorithm of Benamou et al. (2015) but requires a final deconvolution step with the kernel $\exp(-\frac{\mathbf{C}}{\varepsilon})$, which is only feasible when ε is small. With this same objective of non-blurred solutions, Dongdong et al. (2019) even called for a return to the original non-regularized Wasserstein barycenter and proposed an accelerated interior point methods algorithm.

Discrete measures on a finite space The purpose of this section is to derive a fast Sinkhorn-like algorithm to compute $\alpha_{\mathrm{S}_\varepsilon}$ on a fixed support. Let $\mathcal{X} = \{x_1, \dots, x_n\}$ be a finite grid of size n. With images for instance, each x_i would correspond to a pixel. We identify a probability measure $\alpha = \sum_{i=1}^n \alpha_i \delta_{x_i} \in \mathcal{P}(\mathcal{X})$ with its weights vector $(\alpha_i) \in \mathbb{R}_{++}^n$ such that $\sum_{i=1} \alpha_i = 1$. In the rest of this section, OT_ε and S_ε can be seen as functions operating on the interior of the probability simplex of $\mathbb{R}^n$ denoted by $\Delta_n = \{x \in \mathbb{R}_{++}^n | \sum_{i=1} x_i = 1\}$. We assume that the cost matrix $\mathbf{C} \in \mathbb{R}_+^{n\times n}$ is symmetric negative semi-definite (or equivalently, its associated kernel $\mathbf{K} = e^{-\frac{\mathbf{C}}{\varepsilon}}$ is positive semi-definite). This assumption holds for instance if $\mathbf{C}_{ij} = \|x_i - x_j\|^p$ with $p \in]0, 2]$ (see (Berg, Christensen, and Ressel, 1984, §3, Thm 2.2, Cor 3.3) for both claims).

3.1 Reweighted IBP for a deconvoluted barycenter

In practice (with discrete measures in the probability simplex Δ_p), even though the IBP formulation is lost with $\mathrm{OT}_\varepsilon^\otimes$, one can still compute the $\mathrm{OT}_\varepsilon^\otimes$ barycenter using a reweighted version of IBP. Indeed, using the identity (2.93), the $\mathrm{OT}_\varepsilon^\otimes$ barycenter problem is equivalent to the IBP barycenter problem with a non-convex penalty $-\mathrm{KL}$:

$$\arg\min_{\alpha\in\Delta_p} \sum_{k=1}^K w_k \,\mathrm{OT}_\varepsilon^{\mathcal{U}}(\alpha_k, \alpha) - \varepsilon \mathrm{KL}(\alpha, \mathcal{U}) \tag{2.99}$$

$$= \arg\min_{\alpha\in\Delta_p} \sum_{k=1}^K w_k \,\mathrm{OT}_\varepsilon^{\mathcal{U}}(\alpha_k, \alpha) - \langle \alpha, \log(\alpha) - 1\rangle \ . \tag{2.100}$$

The applied penalty is separable and can be written as the sum of non-convex functions $\sum_{i=1}^p -\alpha_i(\log(\alpha_i) - 1) \stackrel{\text{def}}{=} \sum_{i=1}^p g(\alpha_i)$. Gasso, Rakotomamonjy, and Canu (2009) showed that such problems can be solved using reweighted algorithms where the non-convex penalty is written as a difference of two convex

functions. The obtained reweighted algorithm is equivalent to minimization-majorization where g is replaced with a linear upper bound surrogate function. Using the gradient of g (given by $-\log(\alpha)$) to construct such a surrogate leads to a sequence of OT barycenter problems shown in Algorithm 3.

Algorithm 3 Reweighted IBP algorithm to solve (2.99)

Input: $\alpha_1, \dots, \alpha_K$, $\mathbf{K} = e^{-\frac{\mathbf{C}}{\varepsilon}}$
Output: $\alpha_{\mathrm{OT}_\varepsilon^\otimes}$
Initialize $\mathbf{x} = 0$.
repeat
  $\alpha \leftarrow \arg\min_{\alpha \in \Delta_p} \sum_{k=1}^K w_k \, \mathrm{OT}_\varepsilon^{\mathcal{U}}(\alpha_k, \alpha) + \langle \alpha, \mathbf{x} \rangle$
  $\mathbf{x} \leftarrow -\log(\alpha)$
until convergence

Now let's show how can the inner problem be solved using a modified IBP algorithm. Assume $\mathbf{x} \in \mathbb{R}^p$ is a fixed vector:

$$\arg\min_{\alpha \in \Delta_p} \sum_{k=1}^K w_k \, \mathrm{OT}_\varepsilon^{\mathcal{U}}(\alpha_k, \alpha) + \langle \alpha, \mathbf{x} \rangle \quad . \tag{2.101}$$

Introducing the concatenation $\pi = (\pi_1, \dots, \pi_K)$, we can use the theoretical framework of Chizat et al. (2018b) to write (2.101) as:

$$\min_{\pi \in \mathbb{R}_+^{p\,K}} \widehat{\mathrm{KL}}(\pi | \mathbf{K}) + F_1(\pi_1 \mathbb{1}, \dots, \pi_K^\top \mathbb{1}) + F_2(\pi_1^\top \mathbb{1}, \dots, \pi_K^\top \mathbb{1}) \tag{2.102}$$

with:

$$\widehat{\mathrm{KL}}(\pi | \mathbf{K}) = \sum_{k=1}^K w_k \widetilde{\mathrm{KL}}(\pi_k | \mathbf{K}) \tag{2.103}$$

$$F_1(\pi_1 \mathbb{1}, \dots, \pi_K^\top \mathbb{1}) = \sum_{k=1} \iota_{\pi_k \mathbb{1} = \alpha_k} \tag{2.104}$$

$$F_2(\pi_1^\top \mathbb{1}, \dots, \pi_K^\top \mathbb{1}) = \min_{\alpha \in \Delta_p} \sum_{k=1}^K \iota_{\pi_k^\top \mathbb{1} = \alpha} + \langle \alpha, \mathbf{x} \rangle \tag{2.105}$$

Chizat et al. (2018b) showed that the IBP algorithm 2 is an example of the general alternating iterations, starting from some initialized matrices $\mathbf{a}, \mathbf{b} \in \mathbb{R}^{p,K}$:

$$a \leftarrow \mathrm{proxdiv}_{F_1}(\mathbf{K}\mathbf{b}) \tag{2.106}$$

$$b \leftarrow \mathrm{proxdiv}_{F_2}(\mathbf{K}^\top \mathbf{a}) \tag{2.107}$$

where the proxdiv operator is defined as :

$$\text{proxdiv}_F(\mathbf{z}) = \frac{1}{\mathbf{z}} \arg\min_{\mathbf{s}} F(s) + \varepsilon \,\text{KL}(\mathbf{s}|\mathbf{z}) \tag{2.108}$$

Adapting the proof of Chizat et al. (2018b), we compute $\text{proxdiv}_{F_2}(\mathbf{z})$ by solving:

Proposition 16 *The proxdiv operator of F_2 can be given by:*

$$\text{proxdiv}_{F_2}(\mathbf{z}) = \frac{\alpha}{\mathbf{z}} \tag{2.109}$$

where:

$$\alpha = e^{-\mathbf{x}} \prod_{k=1}^{K} \mathbf{z}_k{}^{w_k} \tag{2.110}$$

PROOF. Computing the proxdiv operator of F_2 is equivalent to the problem:

$$\min_{\mathbf{s},\alpha \in \mathbb{R}_+^p} \varepsilon \sum_{k=1}^{K} w_k \,\text{KL}(\mathbf{s}_k, \mathbf{z}_k) + \iota_{s_k=\alpha} + \langle \alpha, \mathbf{x} \rangle \tag{2.111}$$

$$= \min_{\alpha \in \mathbb{R}_+^p} \varepsilon \sum_{k=1}^{K} w_k \,\text{KL}(\alpha, \mathbf{z}_k) + \langle \alpha, \mathbf{x} \rangle \tag{2.112}$$

Given that the problem above is convex, cancelling the gradient leads to its minimizer given by:

$$\sum_{k=1}^{K} w_k \log\left(\frac{\alpha}{\mathbf{z}_k}\right) + \mathbf{x} = 0 \tag{2.113}$$

$$\Leftrightarrow \alpha = e^{-\mathbf{x}} \odot \prod_{k=1}^{K} \mathbf{z}_k{}^{w_k} \tag{2.114}$$

The marginal constraints $\iota_{s_k=\alpha}$ impose that all s_k are equal to α at optimality, therefore diving by $\mathbf{z}$ leads to the proxdiv evaluation at $\mathbf{z}$. ■

Extending the proxdiv iterations leads to the detailed algorithm:

Computing the $\text{OT}_\varepsilon^{\otimes}$ barycenter thus requires several IBP loops. In section 2.2, we illustrated its shrinking bias on Gaussians. Now we turn to the debiased barycenter.

3.2 IBP for debiased barycenters

To obtain a fast iterative algorithm for the debiased barycenters α_{S_ε}, we are going to leverage the IBP algorithm through the uniform measure on $\mathcal{X}$ as follows. First, the identity (2.96) ensures that S_ε is

Algorithm 4 IBP algorithm with a linear penalty to solve (2.101)

Input: $\alpha_1, \dots, \alpha_K$, $\mathbf{K} = e^{-\frac{\mathbf{C}}{\varepsilon}}$, $\mathbf{x} \in \mathbb{R}^p$.
Output: $\alpha_{\mathrm{OT}_\varepsilon^{\mathcal{U}}}$
Initialize all scalings (b_k) to $\mathbb{1}$,
repeat
  for $k = 1$ **to** K **do**
    $a_k \leftarrow \left(\frac{\alpha_k}{\mathbf{K} b_k}\right)$
  end for
  $\alpha \leftarrow e^{-\mathbf{x}} \odot \prod_{k=1}^K (\mathbf{K}^\top a_k)^{w_k}$
  for $k = 1$ **to** K **do**
    $b_k \leftarrow \left(\frac{\alpha}{\mathbf{K}^\top a_k}\right)$
  end for
until convergence

independent of the reference measures. Thus, one can write:

$$\mathrm{S}_\varepsilon(\alpha, \beta) = \mathrm{OT}_\varepsilon^{\mathcal{U}}(\alpha, \beta) - \frac{\mathrm{OT}_\varepsilon^{\mathcal{U}}(\alpha, \alpha) + \mathrm{OT}_\varepsilon^{\mathcal{U}}(\beta, \beta)}{2} \quad .$$

Using (2.90), one can write $\mathrm{OT}_\varepsilon^{\mathcal{U}}(\alpha, \beta)$ as a KL projection. The remaining autocorrelation terms can be replaced by their dual problems to obtain the following proposition.

Algorithm 5 Debiased Sinkhorn Barycenter (Janati, Cuturi, and Gramfort, 2020a)

Input: $\alpha_1, \dots, \alpha_K$, $\mathbf{K} = e^{-\frac{\mathbf{C}}{\varepsilon}}$
Output: $\alpha_{\mathrm{S}_\varepsilon}$
Initialize all scalings $(b_k), d$ to $\mathbb{1}$,
repeat
  for $k = 1$ **to** K **do**
    $a_k \leftarrow \left(\frac{\alpha_k}{\mathbf{K} b_k}\right)$
  end for
  $\alpha \leftarrow d \odot \prod_{k=1}^K (\mathbf{K}^\top a_k)^{w_k}$
  for $k = 1$ **to** K **do**
    $b_k \leftarrow \left(\frac{\alpha}{\mathbf{K}^\top a_k}\right)$
  end for
  $d \leftarrow \sqrt{d \odot \left(\frac{\alpha}{\mathbf{K} d}\right)}$
until convergence

Proposition 17 *Let $\alpha_1, \dots, \alpha_K \in \Delta_n$ and $\mathbf{K} = e^{-\frac{\mathbf{C}}{\varepsilon}}$. Let π denote a sequence $\pi_1, \dots, \pi_K$ of transport plans in $\mathbb{R}_+^{n \times n}$ and the constraint sets $\mathcal{H}_1 = \{\pi | \forall k,\ \pi_k \mathbb{1} = \alpha_k\}$, and $\mathcal{H}_2 = \{\pi | \forall k\, \forall k',\ \pi_k^\top \mathbb{1} = \pi_{k'} \mathbb{1}\}$. The barycenter problem $\min_{\alpha \in \Delta_n} \sum_{k=1}^K w_k \mathrm{S}_\varepsilon(\alpha_k, \alpha)$ is equivalent to:*

$$\min_{\substack{\pi \in \mathcal{H}_1 \cap \mathcal{H}_2 \\ d \in \mathbb{R}_+^n}} \left[\varepsilon \sum_{k=1}^K w_k \widetilde{KL}(\pi_k | \mathbf{K}\, \mathrm{diag}(d)) + \frac{\varepsilon}{2} \langle d - \mathbb{1}, \mathbf{K}(d - \mathbb{1}) \rangle \right] . \tag{2.115}$$

where $\widetilde{KL}(\mathbf{A}, \mathbf{B}) = \sum_{i,j}^n \mathbf{A}_{ij} \log\left(\frac{\mathbf{A}_{ij}}{\mathbf{B}_{ij}}\right) + \mathbf{B}_{ij} - \mathbf{A}_{ij}$.

PROOF. The barycenter problem of S_ε only depends on $\mathrm{OT}_\varepsilon^{\mathcal{U}}(\alpha, \beta) - \frac{1}{2}(\mathrm{OT}_\varepsilon(\alpha, \alpha)$. Let's rewrite this expression using the IBP formulation and duality. the IBP formulation (2.90) is explicitly given by:

$$\mathrm{OT}_\varepsilon^{\mathcal{U}}(\alpha, \beta) = \min_{\substack{\pi \in \mathbb{R}_+^{n \times n} \\ \pi \mathbb{1} = \alpha, \pi^\top \mathbb{1} = \beta}} \varepsilon \widetilde{\mathrm{KL}}(\pi | \mathbf{K}) - \varepsilon \sum_{i,j} \mathbf{K}_{ij} \tag{2.116}$$

And the autocorrelation term can be expressed via its dual problem:

$$\begin{aligned} \mathrm{OT}_\varepsilon^{\mathcal{U}}(\alpha, \alpha) &= \max_{h \in \mathbb{R}^n} 2\langle h, \alpha \rangle - \varepsilon \langle e^{\frac{h}{\varepsilon}}, \mathbf{K} e^{\frac{h}{\varepsilon}} \rangle - \varepsilon \sum_{i,j} \mathbf{K}_{ij} && (2.117) \\ &= \max_{d \in \mathbb{R}_+^n} 2\langle \varepsilon \log(d), \alpha \rangle - \varepsilon \langle d, \mathbf{K} d \rangle - \varepsilon \sum_{i,j} \mathbf{K}_{ij} && (2.118) \\ &= - \min_{d \in \mathbb{R}_+^n} -2\langle \varepsilon \log(d), \alpha \rangle + \varepsilon \langle d, \mathbf{K} d \rangle + \varepsilon \sum_{i,j} \mathbf{K}_{ij} && (2.119) \end{aligned}$$

Moreover, on the constraint set $\mathcal{H}_1 \cap \mathcal{H}_2$, it holds $\alpha = \pi_k^\top$ for all k. Thus, denoting $\mathcal{H}_2(\alpha) = \{\pi | \forall k\, \forall k',\ \pi_k^\top \mathbb{1} = \alpha\}$ the following can be written:

$$
\begin{aligned}
&\arg\min_{\alpha\in\Delta_n} \sum_{k=1}^K w_k \, \mathrm{S}_\varepsilon(\alpha_k, \alpha) \\
&= \arg\min_{\alpha\in\Delta_n} \min_{\pi\in\mathcal{H}_1\cap\mathcal{H}_2(\alpha)} \sum_{k=1}^K w_k \varepsilon \widetilde{\mathrm{KL}}(\pi_k|\mathbf{K}) + \min_{d\in\mathbb{R}_+^n} -\langle \varepsilon \log(d), \alpha\rangle + \frac{1}{2}\varepsilon\langle d, \mathbf{K}d\rangle - \frac{1}{2}\varepsilon\sum_{i,j}\mathbf{K}_{ij} \\
&= \arg\min_{\alpha\in\Delta_n} \min_{\substack{\pi\in\mathcal{H}_1\cap\mathcal{H}_2(\alpha)\\ d\in\mathbb{R}_+^n}} \sum_{k=1}^K w_k \left(\varepsilon \widetilde{\mathrm{KL}}(\pi_k|\mathbf{K}) - \langle \varepsilon \log(d), \alpha\rangle\right) + \frac{1}{2}\varepsilon\langle d, \mathbf{K}d\rangle - \frac{1}{2}\varepsilon\sum_{i,j}\mathbf{K}_{ij} \\
&= \arg\min_{\alpha\in\Delta_n} \min_{\substack{\pi\in\mathcal{H}_1\cap\mathcal{H}_2(\alpha)\\ d\in\mathbb{R}_+^n}} \sum_{k=1}^K w_k \left(\varepsilon \widetilde{\mathrm{KL}}(\pi_k|\mathbf{K}) - \langle \varepsilon \log(d), \pi_k^\top \mathbb{1}\rangle\right) + \frac{1}{2}\varepsilon\langle d, \mathbf{K}d\rangle - \frac{1}{2}\varepsilon\sum_{i,j}\mathbf{K}_{ij} \\
&= \min_{\substack{\pi\in\mathcal{H}_1\cap\mathcal{H}_2\\ d\in\mathbb{R}_+^n}} \sum_{k=1}^K w_k \left(\varepsilon \widetilde{\mathrm{KL}}(\pi_k|\mathbf{K}) - \langle \varepsilon \log(d), \pi_k^\top \mathbb{1}\rangle\right) + \frac{1}{2}\varepsilon\langle d, \mathbf{K}d\rangle - \frac{1}{2}\varepsilon\sum_{i,j}\mathbf{K}_{ij} \\
&= \min_{\substack{\pi\in\mathcal{H}_1\cap\mathcal{H}_2\\ d\in\mathbb{R}_+^n}} \sum_{k=1}^K w_k \left(\varepsilon \widetilde{\mathrm{KL}}(\pi_k|\mathbf{K}\,\mathrm{diag}(d)) - \varepsilon\langle \mathbf{K}d, \mathbb{1}\rangle + \varepsilon\sum_{ij}\mathbf{K}_{ij}\right) + \frac{1}{2}\varepsilon\langle d, \mathbf{K}d\rangle - \frac{1}{2}\varepsilon\sum_{i,j}\mathbf{K}_{ij} \\
&= \min_{\substack{\pi\in\mathcal{H}_1\cap\mathcal{H}_2\\ d\in\mathbb{R}_+^n}} \sum_{k=1}^K \varepsilon w_k \widetilde{\mathrm{KL}}(\pi_k|\mathbf{K}\,\mathrm{diag}(d)) - \varepsilon\langle \mathbf{K}d, \mathbb{1}\rangle + \frac{1}{2}\varepsilon\langle d, \mathbf{K}d\rangle + \frac{1}{2}\varepsilon\sum_{i,j}\mathbf{K}_{ij} \\
&= \min_{\substack{\pi\in\mathcal{H}_1\cap\mathcal{H}_2\\ d\in\mathbb{R}_+^n}} \sum_{k=1}^K \varepsilon w_k \mathrm{KL}(\pi_k|\mathbf{K}\,\mathrm{diag}(d)) + \frac{\varepsilon}{2}\langle d - \mathbb{1}, \mathbf{K}(d-\mathbb{1})\rangle \quad .
\end{aligned}
$$

■

Since $\widetilde{\mathrm{KL}}$ is jointly convex and $\mathbf{K}$ is assumed positive-definite, the objective (2.115) is convex. Minimizing (2.115) with respect to π leads to the barycenter problem $\alpha_{\mathrm{OT}_\varepsilon^{\mathcal{U}}}$ (2.91) with the modified kernel $\mathbf{K}\,\mathrm{diag}(d)$. This problem can be solved via the fast IBP algorithm. Minimizing with respect to d leads to the Sinkhorn fixed point equation $d = \frac{\sum_{w_k}\pi_k^\top \mathbb{1}}{\mathbf{K}d}$ for which there exists a converging sequence (Knight, Ruiz, and Uçar, 2014):

$$
d_{n+1} \leftarrow \sqrt{\frac{d_n \odot \sum_{w_k}\pi_k^\top \mathbb{1}}{\mathbf{K}d}} \, (\star) \tag{2.120}
$$

Given that (2.115) is smooth and convex, alternate minimization – which amounts to perform IBP and (*) iterations – converges towards its minimum. However, we notice that in practice, either taking one iteration or fully optimizing the subproblems produces the same minimizer. We thus propose to combine one IBP iteration with the update ($\star$), which leads to Algorithm 5.

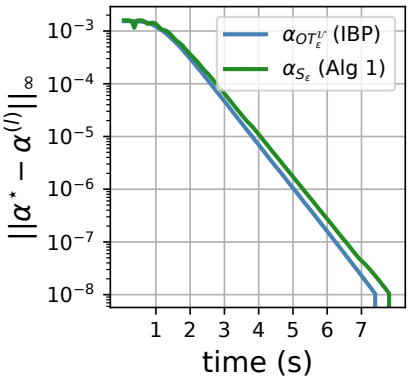


Fig. 2.9. Convergence to the true barycenters of univariate Gaussians $\mathcal{N}(-0.5, 0.1)$ and $\mathcal{N}(0.5, 0.1)$. Algorithm 5 is as fast as IBP with a linear convergence rate.

Convergence of Algorithm 5 A convergence proof of IBP can obtained using alternating Bregman projections (See (Benamou et al., 2015) and the references theirein). For Algorithm 5 however, similar techniques are not successful. Using the theoretical barycenters of Gaussians given by theorems 3 and 5, we can monitor the convergence to the ground truth (Figure 2.9). Theoretically, both IBP and algorithm 5 have a $\mathcal{O}(Kn^2)$ complexity per iteration, convergence guarantees will be the subject of future work.

3.3 Debiasing unbalanced OT

In the balanced case, changing the entropic reference measure from the uniform to the product measure is tantamount to an additional entropy of the input measures:

$$\mathrm{OT}_\varepsilon^{\mathcal{U}}(\alpha, \beta) = \mathrm{OT}_\varepsilon^{\otimes}(\alpha, \beta) + \varepsilon \mathrm{KL}(\alpha|\mathcal{U}) + \varepsilon \mathrm{KL}(\beta|\mathcal{U}) \ . \tag{2.121}$$

Equalities like (2.121) are possible in balanced OT because of the marginal constraints of the transport plan. In unbalanced OT, such identities do not hold. Thus, the reference measure has a much more important role in unbalanced OT.

3.3.1 Reference measure and bias

Let's re-visit the quantities of interest of UOT in the discrete case. Let α, β be two non-negative measures with a fixed support given by $\mathcal{X} = \{x_1, \dots, x_p\} \subset \mathbb{R}^d$. They can be identified with vectors of non-negative weights i.e $\mathbf{a}, \mathbf{b} \in \mathbb{R}_+^p$. Let $\mathbf{C}$ be the cost matrix filled with entries $\mathbf{C}_{ij} = c(x_i, x_j)$ for some non-negative symmetric cost function $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$. Denoting $\mathcal{U}$ the uniform non-negative measure in $\mathcal{X}^2$ assigning the weight 1 to each (x_i, x_j), we define:

$$\underset{\varepsilon,\gamma}{\overset{\mathcal{U}}{\mathrm{UOT}}}(\alpha,\beta) = \min_{\pi\in\mathbb{R}_+{}^{p\times p}} \langle\pi,\mathbf{C}\rangle + \varepsilon\mathrm{KL}(\pi|\mathcal{U}) + \gamma\mathrm{KL}(\pi\mathbb{1}|\mathbf{x}) + \gamma\mathrm{KL}(\pi^\top\mathbb{1}|\mathbf{y}) \ , \tag{2.122}$$

$$\underset{\varepsilon,\gamma}{\overset{\otimes}{\mathrm{UOT}}}(\alpha,\beta) = \min_{\pi\in\mathbb{R}_+{}^{p\times p}} \langle\pi,\mathbf{C}\rangle + \varepsilon\mathrm{KL}(\pi|\alpha\otimes\beta) + \gamma\mathrm{KL}(\pi\mathbb{1}|\mathbf{x}) + \gamma\mathrm{KL}(\pi^\top\mathbb{1}|\mathbf{y}) \ . \tag{2.123}$$

Both formulations fall within the framework of (Chizat et al., 2018b) and have equivalent dual problems up to additional constants which depend only on $\mathbf{C}$ and $\mathcal{X}$:

$$\underset{\varepsilon,\gamma}{\overset{\mathcal{U}}{\mathrm{UOT}}}(\alpha,\beta) = \max_{f,g\in\mathbb{R}^p} -\gamma\langle\mathbf{a},e^{-\frac{f}{\gamma}}-1\rangle - \gamma\langle\mathbf{b},e^{-\frac{g}{\gamma}}-1\rangle - \varepsilon\langle e^{\frac{g\oplus g-\mathbf{C}}{\varepsilon}}-1,\mathbf{U}\rangle \tag{2.124}$$

$$\underset{\varepsilon,\gamma}{\overset{\otimes}{\mathrm{UOT}}}(\alpha,\beta) = \max_{f,g\in\mathbb{R}^p} -\gamma\langle\mathbf{a},e^{-\frac{f}{\gamma}}-1\rangle - \gamma\langle\mathbf{b},e^{-\frac{g}{\gamma}}-1\rangle - \varepsilon\langle e^{\frac{f\oplus g-\mathbf{C}}{\varepsilon}}-1,\mathbf{a}\otimes\mathbf{b}\rangle \tag{2.125}$$

Moreover, with the change of variables: $\omega = \frac{\gamma}{\gamma+\varepsilon}$, $\mathbf{K} = e^{-\frac{\mathbf{C}}{\varepsilon}}$, $\mathbf{u} = e^{\frac{f}{\varepsilon}}$, $\mathbf{v} = e^{\frac{g}{\varepsilon}}$, the optimal dual points are the respective solutions of the fixed point problems and can be used to compute the gradients (Feydy et al., 2017; Séjourné et al., 2019):

$$\text{For } \underset{\varepsilon,\gamma}{\overset{\mathcal{U}}{\mathrm{UOT}}}: \quad \mathbf{u} = \left(\frac{\mathbf{a}}{\mathbf{K}\mathbf{v}}\right)^\omega \quad , \quad \mathbf{v} = \left(\frac{\mathbf{b}}{\mathbf{K}^\top\mathbf{u}}\right)^\omega \tag{2.126}$$

$$\nabla \underset{\varepsilon,\gamma}{\overset{\mathcal{U}}{\mathrm{UOT}}}(\alpha,\beta) = \gamma(\mathbb{1}-\mathbf{u}^{-\frac{\varepsilon}{\gamma}},\mathbb{1}-\mathbf{v}^{-\frac{\varepsilon}{\gamma}}) \tag{2.127}$$

$$\text{For } \underset{\varepsilon,\gamma}{\overset{\otimes}{\mathrm{UOT}}}: \quad \mathbf{u} = \left(\frac{\mathbb{1}}{\mathbf{K}(\mathbf{b}\odot\mathbf{v})}\right)^\omega \quad , \quad \mathbf{v} = \left(\frac{\mathbb{1}}{\mathbf{K}^\top(\mathbf{a}\odot\mathbf{u})}\right)^\omega \tag{2.128}$$

$$\nabla \underset{\varepsilon,\gamma}{\overset{\otimes}{\mathrm{UOT}}}(\alpha,\beta) = (\gamma+\varepsilon\mathbf{b}^\top\mathbb{1}-(\gamma+\varepsilon)\mathbf{u}^{-\frac{\varepsilon}{\gamma}},\gamma+\varepsilon\mathbf{a}^\top\mathbb{1}-(\gamma+\varepsilon)\mathbf{v}^{-\frac{\varepsilon}{\gamma}}) \tag{2.129}$$

and the optimal transport plans are given by:

$$\pi^{\mathcal{U}} = \mathrm{diag}(\mathbf{u})\mathbf{K}\,\mathrm{diag}(\mathbf{v}), \quad \pi^{\otimes} = \mathrm{diag}(\mathbf{a}\odot\mathbf{u})\mathbf{K}\,\mathrm{diag}(\mathbf{v}\odot\mathbf{b}), \tag{2.130}$$

These iterations are a generalization of the Sinkhorn algorithm which corresponds to $\omega = 1$ i.e $\gamma = +\infty$. When the measures α,β are equal to each other, the symmetries of $\mathbf{C}$, π and that of the dual problem lead to a dual solution on the diagonal:

Corollary 3 *Let $\mathbf{a}\in\mathbb{R}_+^p$. The associated optimal dual (identical) scalings $\mathbf{u}$, $\mathbf{v}$ to computing $\mathrm{UOT}^{\mathcal{U}}_{\varepsilon,\gamma}(\mathbf{a},\mathbf{b})$ are given by the solution of the fixed point problem: $\mathbf{u} = \left(\frac{\mathbf{a}}{K\mathbf{u}}\right)^\phi$*

The following proposition shows that similar to the balanced case, UOT divergences induce blurring and shrinking biases depending on the reference measure.

Proposition 18 *Let* $\mathbf{b} \in \mathbb{R}^p_{++}$ *and assume that* $\varepsilon > 0$ *is small enough for* $\mathbf{K} \overset{\text{def}}{=} e^{-\frac{\mathbf{c}}{\varepsilon}}$ *to be invertible. Then:*

$$\operatorname*{arg\,min}_{\mathbf{a} \in \mathbb{R}^p_{++}} \overset{\mathcal{U}}{\underset{\varepsilon,\gamma}{\mathrm{UOT}}}(\mathbf{a}, \mathbf{b}) = \mathbf{K}\left(\frac{\mathbf{b}}{\mathbf{K}^\top \mathbb{1}}\right)^\omega \tag{2.131}$$

$$\operatorname*{arg\,min}_{\mathbf{a} \in \mathbb{R}^p_{++}} \overset{\otimes}{\underset{\varepsilon,\gamma}{\mathrm{UOT}}}(\mathbf{a}, \mathbf{b}) = \kappa^\omega \mathbf{K}^{\top -1}\left(\left(\frac{\mathbf{b}}{\mathbf{K}^{-1}(\kappa \mathbb{1})}\right)^\omega\right) \quad , \tag{2.132}$$

where:

$$\kappa \overset{\text{def}}{=} \left(\frac{\gamma + \varepsilon \mathbf{b}^\top \mathbb{1}}{\gamma + \varepsilon}\right)^{\frac{\varepsilon+\gamma}{\varepsilon}} \quad . \tag{2.133}$$

PROOF. The dual problem (2.124) shows that $\mathrm{UOT}^{\mathcal{U}}_{\varepsilon,\gamma}$ is a supremum of linear functions in $(\mathbf{a}, \mathbf{b})$. Thus, $\mathrm{UOT}^{\mathcal{U}}_{\varepsilon,\gamma}$ is jointly convex in $(\mathbf{a}, \mathbf{b})$. Moreover, Séjourné et al. (2019) showed that $\mathrm{UOT}^{\otimes}_{\varepsilon,\gamma}$ is convex in α and in β but not jointly. Since they are differentiable, cancelling their gradients lead to the desired formulas. ■

3.3.2 Debiased unbalanced divergences

Unlike balanced OT, debiasing unbalanced OT cannot be tackled regardless of the reference measure. Motivated by computational aspects, we propose a debiased UOT loss based on the uniform measure. But first, we acknowledge the more generic formulation using the product measure.

Debiasing $\mathrm{UOT}^{\otimes}_{\varepsilon,\gamma}$ As discussed in the introduction, a natural idea would be to consider a debiasing similar to the balanced case by proposing:

$$(\alpha, \beta) \mapsto \overset{\otimes}{\underset{\varepsilon,\gamma}{\mathrm{UOT}}}(\alpha, \beta) - \frac{1}{2}\left(\overset{\otimes}{\underset{\varepsilon,\gamma}{\mathrm{UOT}}}(\alpha, \alpha) + \overset{\otimes}{\underset{\varepsilon,\gamma}{\mathrm{UOT}}}(\beta, \beta)\right) \quad . \tag{2.134}$$

However, (2.134) does not verify non-negativity nor convexity which are violated when taking large mass discrepancies between the measures. Séjourné et al. (2019) proposed to redress it by adding a quadratic penalty on this mass difference:

$$\mathrm{S}^{\otimes}_{\varepsilon,\gamma}(\alpha, \beta) \overset{\text{def}}{=} \overset{\otimes}{\underset{\varepsilon,\gamma}{\mathrm{UOT}}}(\alpha, \beta) - \frac{1}{2}\left(\overset{\otimes}{\underset{\varepsilon,\gamma}{\mathrm{UOT}}}(\alpha, \alpha) + \overset{\otimes}{\underset{\varepsilon,\gamma}{\mathrm{UOT}}}(\beta, \beta)\right) + \frac{\varepsilon}{2}(\alpha(\mathcal{X}) - \beta(\mathcal{X}))^2 \quad . \tag{2.135}$$

Similarly to the balanced case, as long as $\mathbf{K}$ is positive definite, $\mathrm{S}^{\otimes}_{\varepsilon,\gamma}$ is non-negative and convex with respect to one if its argument. Moreover, since it is defined through the product measure, it is not restricted to discrete measures and was originally studied by Séjourné et al. (2019) for generic measures supported

on compact sets. When it comes to computing barycenters however, no formulation close to Sinkhorn's algorithm can be obtained. One must therefore revert back to first order descent methods where each descent iteration requires a full Sinkhorn loop to compute the gradient.

Debiasing $\mathrm{UOT}^{\mathcal{U}}_{\varepsilon,\gamma}$ When restricted to discrete measures supported on the same finite and fixed set $\mathcal{X}$, we can show that the natural debiased divergence is enough for debiased barycenters to be defined and computed with fast Sinkhorn-like iterations. We propose the following divergence:

$$\mathrm{S}^{\mathcal{U}}_{\varepsilon,\gamma}(\alpha,\beta) \stackrel{\text{def}}{=} \underset{\varepsilon,\gamma}{\overset{\mathcal{U}}{\mathrm{UOT}}}(\alpha,\beta) - \frac{1}{2}(\underset{\varepsilon,\gamma}{\overset{\mathcal{U}}{\mathrm{UOT}}}(\alpha,\alpha) + \underset{\varepsilon,\gamma}{\overset{\mathcal{U}}{\mathrm{UOT}}}(\beta,\beta)) \ . \tag{2.136}$$

Since the support $\mathcal{X}$ and the kernel $\mathbf{K}$ are fixed, we can identify α, β with their weight vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^p_+$. In the rest of this section, $\mathrm{UOT}^{\otimes}_{\varepsilon,\gamma}$ and $\mathrm{UOT}^{\mathcal{U}}_{\varepsilon,\gamma}$ are considered functions over $\mathbb{R}^p_+ \times \mathbb{R}^p_+$.

Non-negativity To show that $\mathrm{S}^{\mathcal{U}}_{\varepsilon,\gamma}$ is non-negative, we assume that the kernel $\mathbf{K} = e^{-\frac{\mathbf{C}}{\varepsilon}}$ is positive semi-definite. This is the case for example with $\mathbf{C}_{ij} = \|x_i - x_j\|^l$ with $0 < l \leq 2$ (Berg, Christensen, and Ressel, 1984) if the support of the measures is given by $\{x_1, \dots, x_p\} \subset \mathbb{R}^d$.

Proposition 19 *Let* $\mathbf{a}, \mathbf{b} \in \mathbb{R}^p_+$. *If* $\mathbf{K} = e^{-\frac{\mathbf{C}}{\varepsilon}}$ *is positive semi-definite:*

$$\mathrm{S}^{\mathcal{U}}_{\varepsilon,\gamma}(\mathbf{a},\mathbf{b}) \geq 0$$

Moreover, if $\mathbf{K}$ *is positive definite,* $\mathrm{S}^{\mathcal{U}}_{\varepsilon,\gamma}(\mathbf{a},\mathbf{b}) = 0 \Leftrightarrow \mathbf{a} = \mathbf{b}$.

PROOF. Let $\mathbf{c}$ and $\mathbf{d}$ denote the solutions of the fixed point problems: $\mathbf{c} = \left(\frac{\mathbf{a}}{\mathbf{K}\mathbf{c}}\right)^{\omega}$ and $\mathbf{d} = \left(\frac{\mathbf{b}}{\mathbf{K}\mathbf{d}}\right)^{\omega}$. With the change of variable $\mathbf{u} = e^{\frac{f}{\varepsilon}}$ and $\mathbf{v} = e^{\frac{g}{\varepsilon}}$, let $(\mathbf{u},\mathbf{v}) \to \mathcal{D}(\mathbf{u},\mathbf{v})$ denote the dual function of (2.124). On one hand, by Corollary 3, $\mathrm{UOT}^{\mathcal{U}}_{\varepsilon,\gamma}(\mathbf{a},\mathbf{a}) = \max_{\mathbf{u},\mathbf{v}\in\mathbb{R}^p_+} \mathcal{D}(\mathbf{u},\mathbf{v}) = \mathcal{D}(\mathbf{c},\mathbf{c})$. Similarly, $\mathrm{UOT}^{\mathcal{U}}_{\varepsilon,\gamma}(\mathbf{b},\mathbf{b}) = D(\mathbf{d},\mathbf{d})$. On the other hand, by definition of the max $\mathrm{UOT}^{\mathcal{U}}_{\varepsilon,\gamma}(\mathbf{a},\mathbf{b}) \geq \mathcal{D}(\mathbf{c},\mathbf{d})$. Therefore:

$$\begin{aligned}
\mathrm{S}^{\mathcal{U}}_{\varepsilon,\gamma}(\mathbf{a},\mathbf{b}) &\geq D(\mathbf{c},\mathbf{d}) - \frac{1}{2}(D(\mathbf{c},\mathbf{c}) + D(\mathbf{d},\mathbf{d})) \\
&= \varepsilon\left[-\langle \mathbf{c}\otimes\mathbf{d}, \mathbf{K}\rangle + \frac{1}{2}\langle \mathbf{c}\otimes\mathbf{c}, \mathbf{K}\rangle + \frac{1}{2}\langle \mathbf{d}\otimes\mathbf{d}, \mathbf{K}\rangle\right] \\
&= \varepsilon\left[-\langle \mathbf{c}, \mathbf{K}\mathbf{d}\rangle + \frac{1}{2}\langle \mathbf{c}, \mathbf{K}\mathbf{c}\rangle + \frac{1}{2}\langle \mathbf{d}, \mathbf{K}\mathbf{d}\rangle\right] \\
&= \frac{\varepsilon}{2}\langle \mathbf{c}-\mathbf{d}, \mathbf{K}(\mathbf{c}-\mathbf{d})\rangle \geq 0
\end{aligned}$$

Where the last inequality follows from the positivity of $\mathbf{K}$. If $\mathbf{K}$ is positive definite, the last inequality is strict unless $\mathbf{c} = \mathbf{d}$, in which case the fixed point equations lead to $\mathbf{a} = \mathbf{b}$. ■

Coercivity Regardless of the nature of $\mathbf{K}$, we will now show that $S(., \mathbf{b})$ is coercive for any fixed $\mathbf{b}$. To do so, we first show that $\mathrm{UOT}^{\mathcal{U}}_{\varepsilon,\gamma}$ only depends on the sums of transported mass:

Proposition 20 *Let $\mathbf{a}, \mathbf{b} \in \mathbb{R}^p_+$ and $\pi_{\mathbf{a},\mathbf{b}} \in \mathbb{R}^{p\times p}_+$ their associated transport plan. Then:*

$$\mathrm{S}^{\mathcal{U}}_{\varepsilon,\gamma}(\mathbf{a}, \mathbf{b}) = (\varepsilon + 2\gamma)(\frac{1}{2}\|\pi_{\mathbf{a},\mathbf{a}}\|_1 + \frac{1}{2}\|\pi_{\mathbf{b},\mathbf{b}}\|_1 - \|\pi_{\mathbf{a},\mathbf{b}}\|_1) \tag{2.137}$$

SKETCH OF PROOF. Let $\mathbf{a}, \mathbf{b} \in \mathbb{R}^p_+$. And let $\mathbf{u}, \mathbf{v}$ the dual scalings associated with the dual problem of $\mathrm{UOT}^{\mathcal{U}}_{\varepsilon,\gamma}(\mathbf{a}, \mathbf{b})$. The corresponding primal solution is given by $\pi_{ij} = \mathbf{u}_i \mathbf{K}_{ij} \mathbf{v}_j$. Therefore, using the fixed point equations (2.126), we have: $\|\pi_{\mathbf{a},\mathbf{b}}\|_1 = \langle \mathbf{u}, \mathbf{K}\mathbf{v}\rangle = \langle \mathbf{a}, \mathbf{u}^{-\frac{\varepsilon}{\gamma}}\rangle = \langle \mathbf{v}, \mathbf{K}^\top \mathbf{u}\rangle = \langle \mathbf{b}, \mathbf{v}^{-\frac{\varepsilon}{\gamma}}\rangle$. Therefore, at optimality, the dual function (2.124) is equal to:

$$\underset{\varepsilon,\gamma}{\overset{\mathcal{U}}{\mathrm{UOT}}}(\mathbf{a}, \mathbf{b}) = -(\varepsilon + 2\gamma)\|\pi_{\mathbf{a},\mathbf{b}}\|_1 + \gamma(\|\mathbf{a}\|_1 + \|\mathbf{b}\|_1) \quad + \varepsilon\|\mathbf{K}\|_1,$$

Writing $\mathrm{UOT}^{\mathcal{U}}_{\varepsilon,\gamma}(\mathbf{a}, \mathbf{a})$ and $\mathrm{UOT}^{\mathcal{U}}_{\varepsilon,\gamma}(\mathbf{b}, \mathbf{b})$ in the same way ends the proof. ■

To prove that $\mathrm{UOT}^{\mathcal{U}}_{\varepsilon,\gamma}$ is coercive, we bound $\|\pi_{\mathbf{a},\mathbf{b}}\|_1$ with the ℓ_1 norms of $\mathbf{a}$ and $\mathbf{b}$:

Lemma 4 *Let $\mathbf{a}, \mathbf{b} \in \mathbb{R}^p_+$ and $\pi_{\mathbf{a},\mathbf{b}} \in \mathbb{R}^{p\times p}_+$ their associated transport plan. Let $\kappa = \min_{i,j} e^{-\frac{\mathrm{C}_{ij}}{\gamma}}$. We have the following bounds on the total transported mass:*

$$\kappa\|\mathbf{a}\|_1\|\mathbf{b}\|_1 \leq \|\pi_{\mathbf{a},\mathbf{b}}\|_1^{2+\frac{\varepsilon}{\gamma}} \leq p^{\frac{3}{2}}\|\mathbf{a}\|_1\|\mathbf{b}\|_1 \tag{2.138}$$

PROOF. The first order optimality condition of the primal UOT problem (2.122) reads for all $i, j \in [\![1, p]\!]$:

$$\varepsilon \log(\pi_{ij}) - \varepsilon \log(\mathbf{K}_{ij}) + \gamma \log\left(\frac{\pi_{i.}^\top \mathbb{1} \pi_{.j}^\top \mathbb{1}}{\mathbf{x}_i \mathbf{y}_j}\right) = 0 \tag{2.139}$$

$$\Leftrightarrow \pi_{ij}^{\frac{\varepsilon}{\gamma}} \pi_{i.}^\top \mathbb{1} \pi_{.j}^\top \mathbb{1} = \mathbf{a}_i \mathbf{b}_j e^{-\frac{\mathrm{C}_{ij}}{\gamma}} \tag{2.140}$$

On one hand we have:

$$\begin{aligned}
\sum_{i,j}^{p} \pi_{ij}^{\frac{\varepsilon}{\gamma}} \pi_{i.}^\top \mathbb{1} \pi_{.j}^\top \mathbb{1} &\leq \|\pi\|_\infty^{\frac{\varepsilon}{\gamma}} \sum_{i,j}^{p} \pi_{i.}^\top \mathbb{1} \pi_{.j}^\top \mathbb{1} \\
&= \|\pi\|_\infty^{\frac{\varepsilon}{\gamma}} \|\pi\|_1^2 \\
&\leq \|\pi\|_1^{2+\frac{\varepsilon}{\gamma}}
\end{aligned}$$

On the other hand, using Jensen's inequality in the second step:

$$\begin{aligned}\sum_{i,j}^{p} \pi_{ij}^{\frac{\varepsilon}{\gamma}} \pi_{i.}^{\top} \mathbb{1} \pi_{.j}^{\top} \mathbb{1} &\geq \sum_{i,j}^{p} \pi_{ij}^{\frac{\varepsilon}{\gamma}+2} \\ &\geq p^2 \left(\frac{\sum_{i,j} \pi_{ij}}{p^2} \right)^{\frac{\varepsilon}{\gamma}+2} \\ &\geq p^{-2-2\frac{\varepsilon}{\gamma}} \|\pi\|_1^{2+\frac{\varepsilon}{\gamma}}\end{aligned}$$

Finally, since $\kappa \leq \min_{ij} e^{-\frac{\mathbf{c}_{ij}}{\gamma}} \leq 1$ we get the desired inequalities. ■

Proposition 21 *For $\mathbf{b} \in \mathbb{R}_+^p$, the function $\mathbf{a} \mapsto \mathrm{S}_{\varepsilon,\gamma}^{\mathcal{U}}(\mathbf{a}, \mathbf{b})$ is coercive.*

PROOF. Lemma 4 and proposition 20, we get, with $\zeta = \frac{1}{2+\frac{\varepsilon}{\gamma}}$:

$$\underset{\varepsilon,\gamma}{\overset{\mathcal{U}}{\mathrm{UOT}}}(\mathbf{a}, \mathbf{b}) \geq \kappa(\|\mathbf{a}\|_1^{2\zeta} + \|\mathbf{b}\|_1^{2\zeta}) - p^{\frac{3}{2}} \|\mathbf{a}\|_1^{\zeta} \|\mathbf{b}\|_1^{\zeta} \tag{2.141}$$

Therefore: $\|\mathbf{a}\|_1 \to +\infty \Rightarrow \mathrm{S}_{\varepsilon,\gamma}^{\mathcal{U}}(\mathbf{a}, \mathbf{b}) \to +\infty$ ■

Differentiability $\mathrm{UOT}_{\varepsilon,\gamma}^{\mathcal{U}}(., \mathbf{b})$ is differentiable, and its gradient is given by $\gamma(1 - \mathbf{a}^{-\frac{\varepsilon}{\gamma}})$ where $\mathbf{u}$ is the solution of the fixed equation (2.126) (Feydy et al., 2017). Thus, $\mathrm{S}_{\varepsilon,\gamma}^{\mathcal{U}}$ is also differentiable. If $\mathbf{K}$ is positive semi-definite then $\mathrm{S}_{\varepsilon,\gamma}^{\mathcal{U}} \geq 0$ and thus, from the following proposition we conclude that all its stationary points are minimizers:

Proposition 22 *Let $\mathbf{a}, \mathbf{b} \in \mathbb{R}_{++}^p$ be a stationary point of $\mathrm{S}_{\varepsilon,\gamma}^{\mathcal{U}}$ i.e $\nabla \mathrm{S}_{\varepsilon,\gamma}^{\mathcal{U}}(\mathbf{a}, \mathbf{b}) = (\mathbf{0}, \mathbf{0})$. Then, $\mathrm{S}_{\varepsilon,\gamma}^{\mathcal{U}}(\mathbf{a}, \mathbf{b}) = 0$. Moreover, if $\mathbf{K}$ is positive definite, then $\mathbf{a} = \mathbf{b}$.*

PROOF. Let $\mathbf{u}, \mathbf{v}, \mathbf{c}, \mathbf{d}$ the solutions of the fixed problems:

$$\mathbf{u} = \left(\frac{\mathbf{a}}{K\mathbf{v}} \right)^{\omega} \quad , \quad \mathbf{v} = \left(\frac{\mathbf{b}}{K\mathbf{u}} \right)^{\omega}, \quad \mathbf{c} = \left(\frac{\mathbf{a}}{K\mathbf{c}} \right)^{\omega}, \quad \mathbf{d} = \left(\frac{\mathbf{b}}{K\mathbf{d}} \right)^{\omega} \tag{2.142}$$

Applying the chain rule, $\frac{1}{2}$ disappears and we get: $\nabla_a \mathrm{UOT}_{\varepsilon,\gamma}^{\mathcal{U}}(\mathbf{a}, \mathbf{b}) = \gamma(\mathbf{c}^{-\frac{\varepsilon}{\gamma}} - \mathbf{u}^{-\frac{\varepsilon}{\gamma}})$ and $\nabla_b \mathrm{UOT}_{\varepsilon,\gamma}^{\mathcal{U}}(\mathbf{a}, \mathbf{b}) = \gamma(\mathbf{d}^{-\frac{\varepsilon}{\gamma}} - \mathbf{v}^{-\frac{\varepsilon}{\gamma}})$. If $(\mathbf{a}, \mathbf{b})$ is a stationary point of $\mathrm{UOT}_{\varepsilon,\gamma}^{\mathcal{U}}$, then we immediately have $\mathbf{u} = \mathbf{c}$ and $\mathbf{v} = \mathbf{d}$. The fixed point equations lead to $\mathbf{Kv} = \mathbf{Ku} = \mathbf{Kc} = \mathbf{Kd}$. The transported mass between $\mathbf{a}$ and $\mathbf{b}$ is given by:

$\|\pi_{\mathbf{a},\mathbf{b}}\|_1 = \langle \mathbf{u}, \mathbf{K}\mathbf{v}\rangle = \langle \mathbf{v}, \mathbf{K}\mathbf{u}\rangle$. Therefore, using Proposition 20, $\mathrm{UOT}^{\mathcal{U}}_{\varepsilon,\gamma}$ can be written:

$$\begin{aligned}
\underset{\varepsilon,\gamma}{\overset{\mathcal{U}}{\mathrm{UOT}}}(\mathbf{a},\mathbf{b}) &= \frac{\varepsilon+2\gamma}{2}(\langle \mathbf{c}, \mathbf{K}\mathbf{c}\rangle + \langle \mathbf{d}, \mathbf{K}\mathbf{d}\rangle - 2\langle \mathbf{u}, \mathbf{K}\mathbf{v}\rangle) \\
&= \frac{\varepsilon+2\gamma}{2}(\langle \mathbf{c}, \mathbf{K}\mathbf{c}\rangle + \langle \mathbf{d}, \mathbf{K}\mathbf{d}\rangle - \langle \mathbf{u}, \mathbf{K}\mathbf{v}\rangle - \langle \mathbf{v}, \mathbf{K}\mathbf{u}\rangle) \\
&= \frac{\varepsilon+2\gamma}{2}(\langle \mathbf{c}+\mathbf{d}-\mathbf{u}-\mathbf{v}, \mathbf{K}\mathbf{u}\rangle) \\
&= 0
\end{aligned}$$

Moreover, if $\mathbf{K}$ is positive definite, $\mathbf{K}\mathbf{u} = \mathbf{K}\mathbf{v}$ leads to $\mathbf{u} = \mathbf{v}$ and thus $\mathbf{a} = \mathbf{b}$. ■

Remark 5 *It is important to keep in mind that all the properties shown for* $\mathrm{UOT}^{\mathcal{U}}_{\varepsilon,\gamma}$ *(coercivity, non-negativity) hold only if the measures are defined over the same support* $\mathcal{X}$ *for all measures. This is crucial for* $\mathbf{K}$ *to be symmetric and defined regardless of the measures themselves.*

3.3.3 Debiased unbalanced barycenters

This section extends the debiased barycenter of (Janati, Cuturi, and Gramfort, 2020a) to the unbalanced case. As of the time of writing, its contributions are not published yet.

Let $\mathbf{a}_1, \dots, \mathbf{a}_K \in \mathbb{R}_+^p$ and $w_1, \dots, w_K$ a sequence of positive weights adding to 1. $\mathrm{UOT}^{\mathcal{U}}_{\varepsilon,\gamma}$ is non-negative and coercive, thus its barycenter problem is well defined:

$$\min_{\mathbf{b}\in\mathbb{R}_+^p} \mathcal{J}(\mathbf{b}) \overset{\text{def}}{=} \min_{\mathbf{a}\in\mathbb{R}_+^p} \sum_{k=1}^{K} w_k \, \mathrm{S}^{\mathcal{U}}_{\varepsilon,\gamma}(\mathbf{a}_k, \mathbf{b}) \tag{2.143}$$

UOT is differentiable, and its gradient is given by $\gamma(1-\mathbf{u}^{-\frac{\varepsilon}{\gamma}}, 1-\mathbf{v}^{-\frac{\varepsilon}{\gamma}})$ where $(\mathbf{u},\mathbf{v})$ is the solution of the fixed equation (2.126) (Feydy et al., 2017). Thus, using the chain rule, $\mathcal{J}$ is also differentiable and its gradient is given by:

$$\nabla\mathcal{J}(\mathbf{b}) = \gamma(\mathbf{c}^{-\frac{\varepsilon}{\gamma}} - \sum_{k=1}^{K} w_k \mathbf{v}_k^{-\frac{\varepsilon}{\gamma}}) \tag{2.144}$$

where, using the usual exponential change of variables, $\mathbf{c}, \mathbf{v}_1, \dots, \mathbf{v}_K, \mathbf{a}_1, \dots, \mathbf{a}_K \in \mathbb{R}_+^p$ verify the fixed point equations:

$$\mathbf{u}_k = \left(\frac{\mathbf{a}_k}{\mathbf{K}\mathbf{v}_k}\right)^{\omega} \quad , \quad \mathbf{v}_k = \left(\frac{\mathbf{b}}{\mathbf{K}^\top \mathbf{u}_k}\right)^{\omega} \quad , \quad \mathbf{c} = \left(\frac{\mathbf{b}}{\mathbf{K}\mathbf{c}}\right)^{\omega} \tag{2.145}$$

Without studying the convexity of $\mathcal{J}$, we can show that any stationary point of $\mathcal{J}$ is actually a global minimum. Thus, it is sufficient to solve $\nabla\mathcal{J}(\mathbf{b}) = 0$ to compute the $\mathrm{UOT}^{\mathcal{U}}_{\varepsilon,\gamma}$ barycenter. The following lemma plays a major role in proving this statement.

Lemma 5 (Suboptimality) *Let* $\mathbf{a}, \mathbf{b} \in \mathbb{R}_+^p$. *Let* $\mathbf{u}, \mathbf{v} \in \mathbb{R}_+^p$ *be the optimal dual variables associated with* $\mathrm{UOT}^{\mathcal{U}}_{\varepsilon,\gamma}(\mathbf{a}, \mathbf{b})$ *i.e the solutions of the optimality conditions* $\mathbf{u} = \left(\frac{\mathbf{a}}{\mathbf{K}\mathbf{v}}\right)^{\omega}$ *and* $\mathbf{v} = \left(\frac{\mathbf{b}}{\mathbf{K}^\top \mathbf{u}}\right)^{\omega}$. *Then for any* $\mathbf{x}, \mathbf{y} \in \mathbb{R}_+^p$*:*

$$\gamma\langle \mathbf{a}, \mathbf{x}^{-\frac{\varepsilon}{\gamma}}\rangle + \gamma\langle \mathbf{b}, \mathbf{y}^{-\frac{\varepsilon}{\gamma}}\rangle + \varepsilon\langle \mathbf{x}, \mathbf{K}\mathbf{y}\rangle \geq (\varepsilon + 2\gamma)\langle \mathbf{u}, \mathbf{K}\mathbf{v}\rangle \tag{2.146}$$

PROOF. Using the same change of variable $\mathbf{x} = e^{\frac{f}{\varepsilon}}, \mathbf{y} = e^{\frac{g}{\varepsilon}}$, the dual problem (2.124) can be written:

$$\begin{aligned}
\underset{\varepsilon,\gamma}{\overset{\mathcal{U}}{\mathrm{UOT}}}(\mathbf{a}, \mathbf{b}) &= \max_{\mathbf{x},\mathbf{y}\in\mathbb{R}_+^p} -\gamma\langle \mathbf{a}, \mathbf{x}^{-\frac{\varepsilon}{\gamma}} - 1\rangle - \gamma\langle \mathbf{b}, \mathbf{y}^{-\frac{\varepsilon}{\gamma}} - 1\rangle - \varepsilon\langle \mathbf{x}\otimes\mathbf{y} - 1, \mathbf{K}\rangle \\
&= \max_{\mathbf{x},\mathbf{y}\in\mathbb{R}_+^p} -\gamma\langle \mathbf{a}, \mathbf{x}^{-\frac{\varepsilon}{\gamma}} - 1\rangle - \gamma\langle \mathbf{b}, \mathbf{y}^{-\frac{\varepsilon}{\gamma}} - 1\rangle - \varepsilon\langle \mathbf{x}\otimes\mathbf{y} - 1, \mathbf{K}\rangle \\
&= \max_{\mathbf{x},\mathbf{y}\in\mathbb{R}_+^p} -\gamma\langle \mathbf{a}, \mathbf{x}^{-\frac{\varepsilon}{\gamma}}\rangle - \gamma\langle \mathbf{b}, \mathbf{y}^{-\frac{\varepsilon}{\gamma}}\rangle - \varepsilon\langle \mathbf{x}, \mathbf{K}\mathbf{y}\rangle - \varepsilon\|\mathbf{K}\|_1 + \gamma(\|\mathbf{a}\|_1 + \|\mathbf{b}\|_1)
\end{aligned}$$

Since $\mathbf{u}, \mathbf{v}$ are the solution of the dual problem above, at optimality it holds:

$$\langle \mathbf{a}, \mathbf{u}^{-\frac{\varepsilon}{\gamma}}\rangle = \langle \mathbf{u}^{\frac{1}{\omega}}\mathbf{K}\mathbf{v}, \mathbf{u}^{-\frac{\varepsilon}{\gamma}}\rangle = \langle \mathbf{u}, \mathbf{K}\mathbf{v}\rangle$$

Similarly:

$$\langle \mathbf{b}, \mathbf{v}^{-\frac{\varepsilon}{\gamma}}\rangle = \langle \mathbf{v}^{\frac{1}{\omega}}\mathbf{K}^\top\mathbf{u}, \mathbf{v}^{-\frac{\varepsilon}{\gamma}}\rangle = \langle \mathbf{v}, \mathbf{K}^\top\mathbf{u}\rangle = \langle \mathbf{u}, \mathbf{K}\mathbf{v}\rangle$$

Thus:

$$\underset{\varepsilon,\gamma}{\overset{\mathcal{U}}{\mathrm{UOT}}}(\mathbf{a}, \mathbf{b}) = -(\varepsilon + 2\gamma)\langle \mathbf{u}, \mathbf{K}\mathbf{v}\rangle - \varepsilon\|\mathbf{K}\|_1 + \gamma(\|\mathbf{a}\|_1 + \|\mathbf{b}\|_1)$$

By the definition of the max operator, it holds for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}_+^p$:

$$\gamma\langle \mathbf{a}, \mathbf{x}^{-\frac{\varepsilon}{\gamma}}\rangle + \gamma\langle \mathbf{b}, \mathbf{y}^{-\frac{\varepsilon}{\gamma}}\rangle + \varepsilon\langle \mathbf{x}, \mathbf{K}\mathbf{y}\rangle \geq (\varepsilon + 2\gamma)\langle \mathbf{u}, \mathbf{K}\mathbf{v}\rangle$$

□

Since $\mathcal{J}$ is coercive, it has at least one global minimum. The following proposition shows that this minimum is unique, potentially attained at multiple minimizers.

Proposition 23 *Let* $\bar{\mathbf{b}} \in \mathbb{R}_+^p$ *such that* $\nabla\mathcal{J}(\bar{\mathbf{b}}) = \mathbf{0}$. *Then for any* $\mathbf{z} \in \mathbb{R}_+^p$ *it holds:*

$$\mathcal{J}(\mathbf{z}) \geq \mathcal{J}(\bar{\mathbf{b}})$$

.

PROOF. Let $\mathbf{d}_1, \dots, \mathbf{d}_K$ the symmetric dual variables used to compute $\mathrm{UOT}^{\mathcal{U}}_{\varepsilon,\gamma}(\mathbf{a}_k, \mathbf{a}_k)$ for $k = 1..K$ i.e the solutions of $\mathbf{d}_k = \left(\frac{\mathbf{a}_k}{\mathbf{K}\mathbf{d}_k}\right)^{\omega}$. Let $\mathbf{z} \in \mathbb{R}^p_+$ and its associated dual variables $\mathbf{c}', \mathbf{u}'_1, \dots, \mathbf{u}'_K, \mathbf{v}'_1, \dots, \mathbf{v}'_K$ used to compute $\mathrm{UOT}^{\mathcal{U}}_{\varepsilon,\gamma}(\mathbf{z}, \mathbf{z}), \mathrm{UOT}^{\mathcal{U}}_{\varepsilon,\gamma}(\mathbf{a}_1, \mathbf{z}), \dots, \mathrm{UOT}^{\mathcal{U}}_{\varepsilon,\gamma}(\mathbf{a}_K, \mathbf{z})$. Therefore, it holds:

$$\mathcal{J}(\mathbf{z}) = (\varepsilon + 2\gamma) \sum_{k=1}^{K} w_k \left(\frac{1}{2}(\langle \mathbf{c}', \mathbf{K}\mathbf{c}' \rangle + \langle \mathbf{d}_k, \mathbf{K}\mathbf{d}_k \rangle) - \langle \mathbf{u}'_k, \mathbf{K}\mathbf{v}'_k \rangle \right)$$

Let $\mathbf{c}, \mathbf{u}_1, \dots, \mathbf{u}_K, \mathbf{v}_1, \dots, \mathbf{v}_K$ denote the dual variables verifying (2.145) associated with $\bar{\mathbf{b}}$ and the $(\mathbf{a}_k)_k$. Since $\nabla \mathcal{J}(\bar{\mathbf{b}}) = \mathbf{0}$, it holds: $\mathbf{c}^{-\frac{\varepsilon}{\gamma}} = \sum_{k=1}^{K} w_k \mathbf{v}_k^{-\frac{\varepsilon}{\gamma}}$, therefore:

$$\sum_{k=1}^{K} w_k \langle \mathbf{a}_k, \mathbf{K}\mathbf{v}_k \rangle = \sum_{k=1}^{K} w_k \langle \bar{\mathbf{b}}, \mathbf{v}_k^{-\frac{\varepsilon}{\gamma}} \rangle = \langle \bar{\mathbf{b}}, \mathbf{c}^{-\frac{\varepsilon}{\gamma}} \rangle = \langle \mathbf{c}, \mathbf{K}\mathbf{c} \rangle \tag{2.147}$$

Thus, evaluate $\mathcal{J}$ at $\bar{\mathbf{b}}$ leads to:

$$\begin{aligned} \mathcal{J}(\bar{\mathbf{b}}) &= (\varepsilon + 2\gamma) \sum_{k=1}^{K} w_k \left(\frac{1}{2}(\langle \mathbf{c}, \mathbf{K}\mathbf{c} \rangle + \langle \mathbf{d}_k, \mathbf{K}\mathbf{d}_k \rangle) - \langle \mathbf{u}_k, \mathbf{K}\mathbf{v}_k \rangle \right) \\ &= \frac{1}{2}(\varepsilon + 2\gamma) \left(\sum_{k=1}^{K} w_k \langle \mathbf{d}_k, \mathbf{K}\mathbf{d}_k \rangle - \langle \mathbf{c}, \mathbf{K}\mathbf{c} \rangle \right) \end{aligned}$$

Thus, the statement we wish to prove is equivalent to:

$$\mathcal{J}(\mathbf{z}) \geq \mathcal{J}(\bar{\mathbf{b}}) \Leftrightarrow \frac{1}{2}(\varepsilon + 2\gamma) \left(\langle \mathbf{c}', \mathbf{K}\mathbf{c}' \rangle + \langle \mathbf{c}, \mathbf{K}\mathbf{c} \rangle \right) \geq (\varepsilon + 2\gamma) \sum_{k=1}^{K} w_k \langle \mathbf{u}'_k, \mathbf{K}\mathbf{v}'_k \rangle \tag{2.148}$$

For each element of the sum in the right side above, let's derive un upper bound using Lemma 5. Consider the sub-optimal dual variables $(\mathbf{x}_k, \mathbf{z}_k) = (\mathbf{u}_k, \mathbf{v}_k \odot \frac{\mathbf{c}'}{\mathbf{c}})$. It holds:

$$\gamma \langle \mathbf{a}_k, \mathbf{u}_k^{-\frac{\varepsilon}{\gamma}} \rangle + \gamma \langle \mathbf{z}, (\mathbf{v}_k \odot \frac{\mathbf{c}'}{\mathbf{c}})^{-\frac{\varepsilon}{\gamma}} \rangle + \varepsilon \langle \mathbf{v}_k \odot \frac{\mathbf{c}'}{\mathbf{c}}, \mathbf{K}^\top \mathbf{u}_k \rangle \geq (\varepsilon + 2\gamma) \langle \mathbf{u}'_k, \mathbf{K}\mathbf{v}'_k \rangle \tag{2.149}$$

Applying the weighted sum and using the optimality conditions along with $\mathcal{J}(\bar{\mathbf{b}}) = 0$, the elements in the left side can be further simplified as:

$$\sum_{k=1}^{K} w_k \langle \mathbf{a}_k, \mathbf{u}_k{}^{-\frac{\varepsilon}{\gamma}} \rangle = \sum_{k=1}^{K} w_k \langle \bar{\mathbf{b}}, \mathbf{v}_k{}^{-\frac{\varepsilon}{\gamma}} \rangle = \sum_{k=1}^{K} w_k \langle \bar{\mathbf{b}}, \mathbf{c}^{-\frac{\varepsilon}{\gamma}} \rangle = \langle \mathbf{c}, \mathbf{K}\mathbf{c} \rangle$$

$$\sum_{k=1}^{K} w_k \langle \mathbf{z}, (\mathbf{v}_k \odot \frac{\mathbf{c}'}{\mathbf{c}})^{-\frac{\varepsilon}{\gamma}} \rangle = \langle \mathbf{z} \odot (\frac{\mathbf{c}'}{\mathbf{c}})^{-\frac{\varepsilon}{\gamma}}, \sum_{k=1}^{K} w_k \mathbf{v}_k^{-\frac{\varepsilon}{\gamma}} \rangle = \langle \mathbf{z} \odot (\frac{\mathbf{c}'}{\mathbf{c}})^{-\frac{\varepsilon}{\gamma}}, \mathbf{c}^{-\frac{\varepsilon}{\gamma}} \rangle = \langle \mathbf{z}, \mathbf{c}'^{-\frac{\varepsilon}{\gamma}} \rangle = \langle \mathbf{c}', \mathbf{K}\mathbf{c}' \rangle$$

$$\sum_{k=1}^{K} w_k \langle \mathbf{v}_k \odot \frac{\mathbf{c}'}{\mathbf{c}}, \mathbf{K}^\top \mathbf{u}_k \rangle = \langle \frac{\mathbf{c}'}{\mathbf{c}} \odot \bar{\mathbf{b}}, \sum_{k=1}^{K} w_k \mathbf{v}_k{}^{-\frac{\varepsilon}{\gamma}} \rangle = \langle \mathbf{c}' \odot \bar{\mathbf{b}}, \mathbf{c}^{-\frac{\gamma+\varepsilon}{\gamma}} \rangle = \langle \mathbf{c}', \mathbf{K}\mathbf{c} \rangle$$

Therefore, summing over equation (2.149):

$$\gamma \langle \mathbf{c}, \mathbf{K}\mathbf{c} \rangle + \gamma \langle \mathbf{c}', \mathbf{K}\mathbf{c}' \rangle + \varepsilon \langle \mathbf{c}', \mathbf{K}\mathbf{c} \rangle \geq (\varepsilon + 2\gamma) \sum_{k=1}^{K} w_k \langle \mathbf{u}'_k, \mathbf{K}\mathbf{v}'_k \rangle$$

On another side, since $\mathbf{K}$ is positive semi-definite, it holds:

$$\langle \mathbf{c} - \mathbf{c}', \mathbf{K}(\mathbf{c} - \mathbf{c}') \rangle \geq 0 \Rightarrow \frac{1}{2}(\langle \mathbf{c}, \mathbf{K}\mathbf{c} \rangle + \langle \mathbf{c}', \mathbf{K}\mathbf{c}' \rangle) \geq \langle \mathbf{c}', \mathbf{K}\mathbf{c} \rangle$$

Combining the last two inequalities leads to (2.148) ending the proof. □.

To solve the barycenter problem (2.143), it is sufficient to solve the fixed point system:

$$\mathbf{u}_k = \left(\frac{\mathbf{a}_k}{\mathbf{K}\mathbf{v}_k} \right)^{\omega} \quad , \quad \mathbf{v}_k = \left(\frac{\bar{\mathbf{b}}}{\mathbf{K}^\top \mathbf{u}_k} \right)^{\omega} \quad , \quad \mathbf{c} = \left(\frac{\bar{\mathbf{b}}}{\mathbf{K}\mathbf{c}} \right)^{\omega} \quad , \quad \sum_{k=1}^{K} w_k \mathbf{v}_k{}^{-\frac{\varepsilon}{\gamma}} = \mathbf{c}^{-\frac{\varepsilon}{\gamma}} \tag{2.150}$$

which – combining the last 3 equations – is equivalent to:

$$\mathbf{u}_k = \left(\frac{\mathbf{a}_k}{\mathbf{K}\mathbf{v}_k} \right)^{\omega} , \quad \mathbf{v}_k = \left(\frac{\bar{\mathbf{b}}}{\mathbf{K}^\top \mathbf{u}_k} \right)^{\omega} , \quad \mathbf{c} = \left(\frac{\bar{\mathbf{b}}}{\mathbf{K}\mathbf{c}} \right)^{\omega} , \quad \bar{\mathbf{b}} = \mathbf{c}^{\frac{1}{\omega}} \left(\sum_{k=1}^{K} w_k (\mathbf{K}^\top \mathbf{u}_k)^{1-\omega} \right)^{\frac{1}{1-\omega}} \tag{2.151}$$

These equations are very similar to the barycentric Sinkhorn algorithm of Chizat et al. (2018b). Indeed, disregarding the symmetric equation in $\mathbf{c}$ and setting $\mathbf{c} = \mathbb{1}_p$ in the update of $\bar{\mathbf{b}}$, we recover Sinkhorn's iterations for the $\text{UOT}^{\mathcal{U}}_{\varepsilon,\gamma}$ barycenter. These updates lead to Algorithm 6. While the theoretical analysis of its convergence is left for future work, we empirically observe that it converges regardless of the initialization of the dual variables. More importantly, it leads to sharper barycenters than the (biased) UOT barycenters for almost no additional computational cost.

Remark 6 *The proposition 23 may seem to indicate that $\mathcal{J}$ has a positive curvature. However, it is easy to show that $S^{\mathcal{U}}_{\varepsilon,\gamma}$ is not convex in dimension 1. Indeed, taking a dimension $p = 1$ leads to $\mathbf{K} = 1$ and the Sinkhorn equations*

Algorithm 6 Debiased unbalanced $\mathrm{S}^{\mathcal{U}}_{\varepsilon,\gamma}$ barycenter.

Input: $\mathbf{a}_1, \dots, \mathbf{a}_K \in \mathbb{R}^p_+$, **parameters** $\varepsilon, \gamma > 0$, $\mathbf{K} \stackrel{\text{def}}{=} e^{-\frac{\mathbf{C}}{\varepsilon}}$
Output: $\bar{\mathbf{b}}$, **the** $\mathrm{UOT}^{\mathcal{U}}_{\varepsilon,\gamma}$ **barycenter of** $(\mathbf{a}_1, \dots, \mathbf{a}_K)$
Initialize $\mathbf{c} = \mathbf{v}_1 = \cdots = \mathbf{v}_K = \mathbb{1}_p$, **set** $\omega = \frac{\gamma}{\gamma+\varepsilon}$
while Not converged **do**
  for $k = 1$ **to** K **do**
    $\mathbf{u}_k = \left(\frac{\mathbf{a}_k}{\mathbf{K}\mathbf{v}_k}\right)^{\omega}$
  end for
  $\bar{\mathbf{b}} = \mathbf{c}^{\frac{1}{\omega}} \left(\sum_{k=1}^K w_k (\mathbf{K}^\top \mathbf{u}_k)^{1-\omega}\right)^{\frac{1}{1-\omega}}$
  for $k = 1$ **to** K **do**
    $\mathbf{v}_k = \left(\frac{\bar{\mathbf{b}}}{\mathbf{K}^\top \mathbf{u}_k}\right)^{\omega}$
  end for
  $\mathbf{c} = \left(\frac{\bar{\mathbf{b}}}{\mathbf{K}\mathbf{c}}\right)^{\omega}$
end while

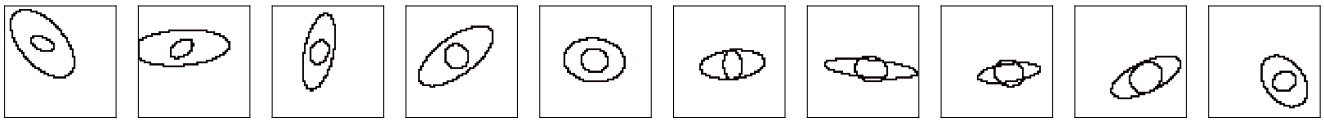


Fig. 2.10. All 10 nested ellipses images used to compute the barycenters of Figure 2.11.

can be solved in closed form. We obtain for $a, b \in \mathbb{R}_+$*:*

$$\mathrm{S}^{\mathcal{U}}_{\varepsilon,\gamma}(a,b) = (\varepsilon + 2\gamma)\left(\frac{a^{\frac{2\omega}{\omega+1}} + b^{\frac{2\omega}{\omega+1}}}{2} - (ab)^{\frac{\omega}{\omega+1}}\right)$$

Since $\omega < 1$*, in dimension 1,* $a \mapsto \mathrm{S}^{\mathcal{U}}_{\varepsilon,\gamma}(a,b)$ *is strictly concave provided* a *is large enough.*

3.4 Experiments

Now we turn to showing the practical benefits of debiased barycenters in terms of accuracy, speed and performance. We compare the following barycenters.

3.4.1 Balanced OT

We compare several formulations of OT barycenters:

1. $\mathrm{OT}^{\mathcal{U}}_{\varepsilon}$: OT with the uniform measure; computed using the IBP algorithm (Algorithm 2).

2. S_ε: Proposed debiased divergence; computed using the proposed algorithm (Algorithm 5).

3. $\mathrm{OT}_\varepsilon^{\otimes}$: Computed using iterative IBP within the minimization-majorization framework (Algorithm 3).

4. A_ε: Sharp barycenters introduced by Luise et al. (2018). Given the optimal transport plan π_ε computed by solving $\mathrm{OT}_\varepsilon^{\mathcal{U}}$, A_ε is defined by evaluating the OT loss without entropy:

$$A_\varepsilon(\alpha, \beta) = \langle \mathbf{C}, \pi_\varepsilon \rangle \ . \tag{2.152}$$

   Its barycenter can be computed using accelerated gradient descent with automatic differentiation through Sinkhorn's algorithm. This method required considerable manual effort to tune the learning rate in order to get an acceptable barycenter and was more prone to numerical instabilities.

5. Free support barycenters with S_ε: introduced by Luise et al. (2019), we used the online Python code provided by the authors which amounts to add or remove a dirac particle at each iteration and update their weights using Frank-Wolf's algorithm. The algorithm is stopped when no particules are created / removed.

6. W: non regularized Wasserstein distance. We used the accelerated interior point methods introduced by Dongdong et al. (2019) with the online matlab implementation provided by the authors.

Debiased barycenters of ellipses To demonstrate how debiased barycenters α_{S_ε} reduce smoothing and are computationally competitive with $\alpha_{\mathrm{OT}_\varepsilon^{\mathcal{U}}}$, we compare the barycenters of 10 randomly generated nested ellipses displayed in Figure 2.10. We simulate each ellipse by generating random major and minor radii with a moving a center from the top left quarter corner to the bottom right quarter corner. The box constraints of the random generators of the radii are manually picked so that ellipses are more likely to be nested with an assymetric surrounding ellipse (see supplementary code). The full list of 10 images used to compute the barycenters is displayed in Figure 2.10. Each image has 60×60 pixels. The ground OT cost function is the squared Euclidean cost over the unit square $[0, 1]^2$. For entropy regularized distances (All except W), we set ε to the lowest value guaranteeing no numerical instabilities in Sinkhorn's algorithm (this was particularly an issue for *Sharp barycenters* α_{A_ε} of Luise et al. (2018)). Now we detail the algorithm used for each divergence F defining each barycenter α_F of the experiment in Figure 2.11. We use the same termination criterion for all methods based on a maximum relative change of the barycenters set to 10^{-5}. For $\alpha_{S_\varepsilon}, \alpha_{\mathrm{OT}_\varepsilon^{\mathcal{U}}}, \alpha_{\mathrm{OT}_\varepsilon^{\otimes}}, \alpha_{A_\varepsilon}$, we use the convolution trick introduced by Solomon et al. (2015) which amounts to computing the kernel operation $\mathbf{K}a$ on a vectorized image a by applying a Gaussian convolution on the rows and the columns of a, thereby reducing the complexity of one Debiased / IBP iteration from $O(n^2)$ to $O(n^{\frac{3}{2}})$. Figure 2.11 shows that even though α_{A_ε} and α_W are not blurred compared to $\alpha_{\mathrm{OT}_\varepsilon^{\mathcal{U}}}$, they cannot compete computationally with Sinkhorn-like algorithms. The debiased barycenter is sharp and runs in about the same time as $\alpha_{\mathrm{OT}_\varepsilon^{\mathcal{U}}}$. Besides, the shrinking bias of $\mathrm{OT}_\varepsilon^{\otimes}$ unfolded by theorem 4 is illustrated in the degeneracy of the ellipse $\alpha_{\mathrm{OT}_\varepsilon^{\otimes}}$.

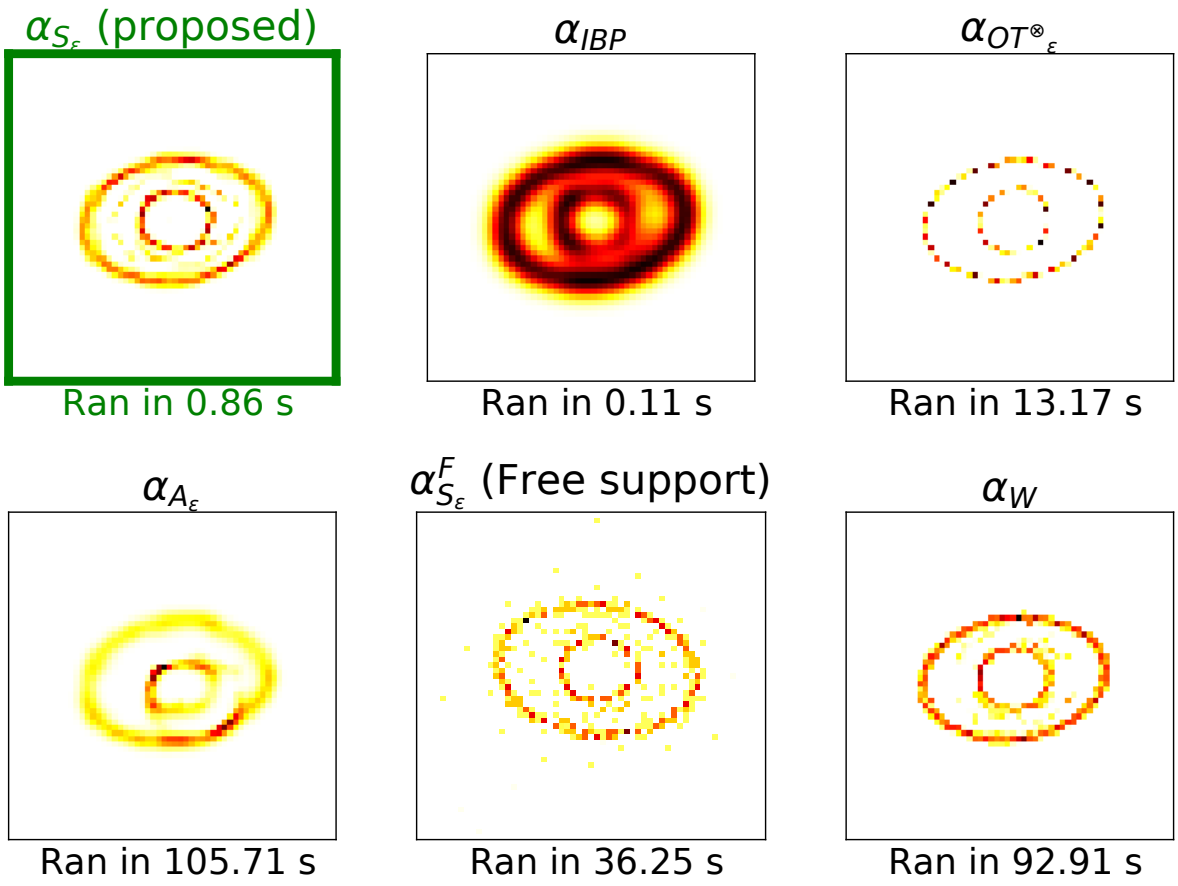


Fig. 2.11. Barycenters of the 10 nested ellipses shown in Figure 2.10. Results illustrate the reduced blurring of the proposed approach and running times presented below each image demonstrate the computational efficiency. All 6 barycenters were computed on a laptop with an Intel Core i5 3.1 GHz Processor.

Barycenters of 3D shapes The original 3D shapes (tore and rabbit) are taken from the `PyVista` (Sullivan and Kaszynski, 2019) Python library and displayed in Figure 2.12. We preprocess the original meshes as follows. Each mesh is smoothed by 100 iterations of a Laplacian operator then the coordinates are centered and rescaled to fit within 95% of the cube $(-1,1)^3$. We sample 3D histograms of both meshes on a uniform 3D grid of size 200^3. Both histograms are normalized and regularized by adding a 10^{-10} weight to avoid numerical errors. We set the lowest stable regularization $\varepsilon = 0.01$ for the ground cost defined as the squared Euclidean distance over the $(-1,1)^3$ cube. We compute weighted barycenters with the IBP algorithm 2 and the proposed debiased Sinkhorn barycenter algorithm 5. The different interpolations correspond to weights $(w, 1-w)$ where $w \in [0, 0.25, 0.5, 0.75, 1]$. We set the cost matric $\mathbf{C}$ to the squared Euclidean distance on the unit cube and set $\varepsilon = 0.01$. Results presented in Figures 2.13 and 2.14 using

Fig. 2.12. Input meshes used to compute the barycenters of 3D meshes.

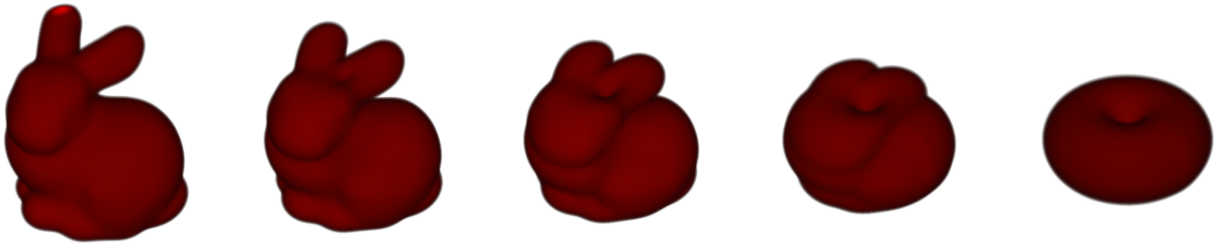


Fig. 2.13. Interpolation of two 3D shapes a $(200)^3$ uniform grid with IBP illustrating a clear blurring bias of $\mathrm{OT}_\varepsilon^{\mathcal{U}}$.

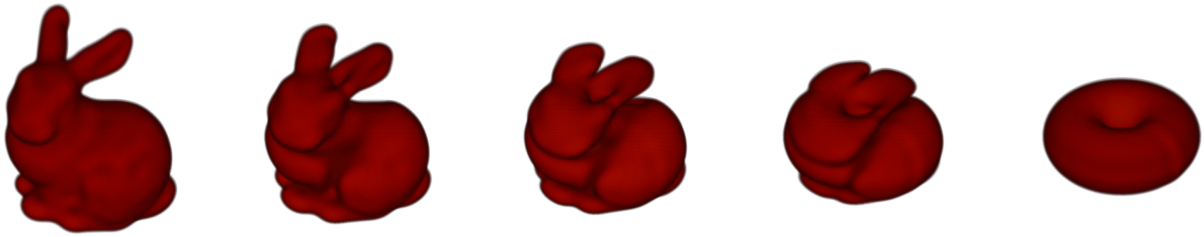


Fig. 2.14. Interpolation of two 3D shapes on a $(200)^3$ uniform grid with the proposed Debiased Sinkhorn (Alg 5). The interpolation is sharper and completes in about the same time as figure 2.13 (5 seconds on a GPU).

$\mathrm{OT}_\varepsilon^{\mathcal{U}}$ and S_ε qualitatively demonstrate that S_ε leads to sharper edges, while in both cases it takes a few seconds to compute on a GPU. Again, the kernel operation $\mathbf{K}a$ on a vectorized 3D grid a can be computed via a sequence of 3 Gaussian convolutions on each axis (x, y, z) which reduces the complexity of one Debiased / IBP iteration from $O(n^2)$ to $O(n^{\frac{4}{3}})$.

Optimal transport barycentric embeddings One of the many machine learning applications of OT barycenters is to compute low-dimensional barycentric embeddings. Introduced by Bonneel, Peyré, and

Cuturi (2016), OT barycentric coordinates are defined as follows. Given a dictionary $\mathcal{A}$ of distributions $\alpha_1, \dots, \alpha_K$ and $w \in \Delta_K$, let $\alpha_F(w) = \arg\min_\alpha \sum_{k=1}^K w_k F(\alpha_k, \alpha)$ for some OT divergence F. The OT coordinates $\hat{w}$ of a distribution β are defined as the weights of the barycenter $\alpha_F(w)$ best approximating β for a given divergence. Using a quadratic divergence, it reads:

$$\hat{w} = \arg\min_{w \in \Delta_K} \|\alpha_F(w) - \beta\|^2 \ . \tag{2.153}$$

To leverage the differentiability of the IBP iterations, Bonneel, Peyré, and Cuturi (2016) used the divergence $\mathrm{OT}_\varepsilon^{\mathcal{U}}$ and proposed to substitute the minimizer $\alpha_F(w)$ with the l-th IBP iterate $\alpha_F^{(l)}(w)$. Differentiating the barycenter nets $\alpha_F^{(l)}(w)$ with respect to w can be done via automatic differentiation, while the full minimization can be done using accelerated gradient descent using a soft-max reparametrization. Here we use the ADAM optimizer of the pyTorch library (Paszke et al., 2017). To evaluate the benefits of debiasing, we take 500 samples of the MNIST dataset (LeCun and Cortes, 2010) with 100 instances of each digit (0-1-2-3-4). We select 10% of the dataset (a subset of 50 images; ergo K=50) at random as our learning dictionary $\mathcal{A}$ and compute the barycentric coordinates of the remaining 90% subset denoted as $\mathcal{D}$. Thus, for each image among the 450 samples of $\mathcal{D}$, we compute the closest (in squared ℓ_2) weighted barycenter of the elements of $\mathcal{A}$ by optimizing over the weights. Thus, each image is represented by a vector of weights $w \in \Delta_K$. Our new embedded dataset is now a table of shape (450×50). We then use this embedding to train a Random Forest Classifier with 100 estimators using scikit-learn's (Pedregosa et al., 2011) default parameters (version 0.21.3) and compute a 10-fold cross-validation. Figure 2.15 displays the accuracy scores for $F = \mathrm{OT}_\varepsilon^{\mathcal{U}}$ and $F = \mathrm{S}_\varepsilon$ for 20 different randomized selections of the dictionary $\mathcal{A}$. The debiased S_ε improves accuracy and is less sensitive to the setting of ε.

3.4.2 Unbalanced barycenters

A proper application of debiased unbalanced barycenters will be the subject of Chapter 4. We provide nonetheless a toy example with barycenters of scaled Gaussians in Figure 2.16. Both Gaussians have the same variance but a 3 to 1 mass ratio.

4 Limitations and future perspectives

Entropic OT is now considered to be a well-established loss function for comparing and averaging probability / positive measures. The purpose of this chapter was to study the various entropic OT formulations both from a theoretical perspective as well as a practical one. This is yet far from being a closed book. Several questions remain open. We conclude this chapter with a brief discussion over potential future directions.

Entropic OT over non-compact sets Without compactness of the underlying space, an integrable upper bound on the optimal dual potentials is required to obtain the differentiability of entropic OT. In our case, we assumed that measures have sub-Gaussian tails which leads to quadratic bounds provided by Mena

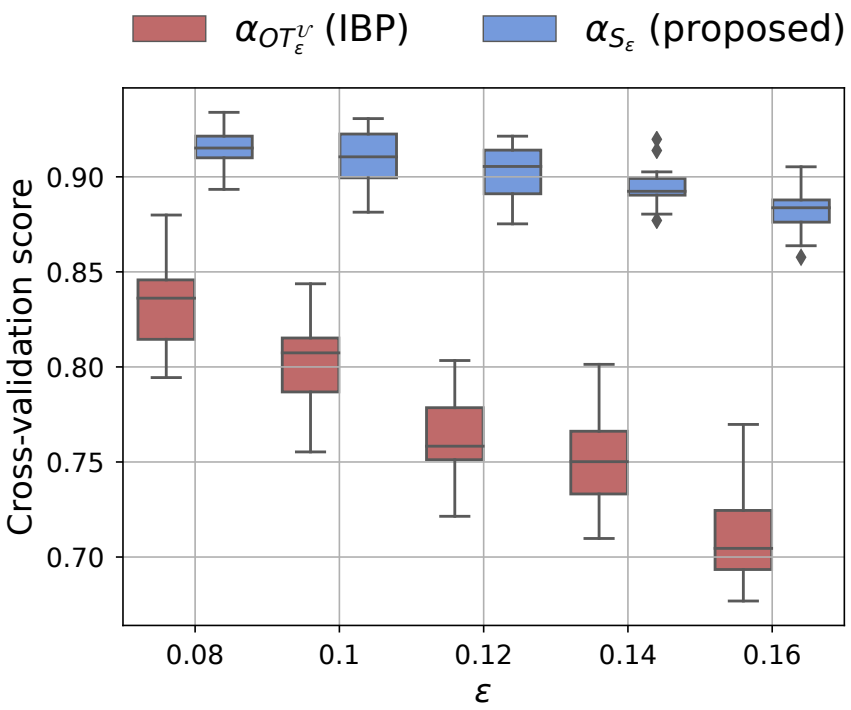


Fig. 2.15. Cross-validation accuracy with 95% confidence intervals obtained on 500 MNIST images using barycentric embedding with S_ε or $OT_\varepsilon^{\mathcal{U}}$. Debiasing of S_ε improves performance. S_ε is less sensitive to ε.

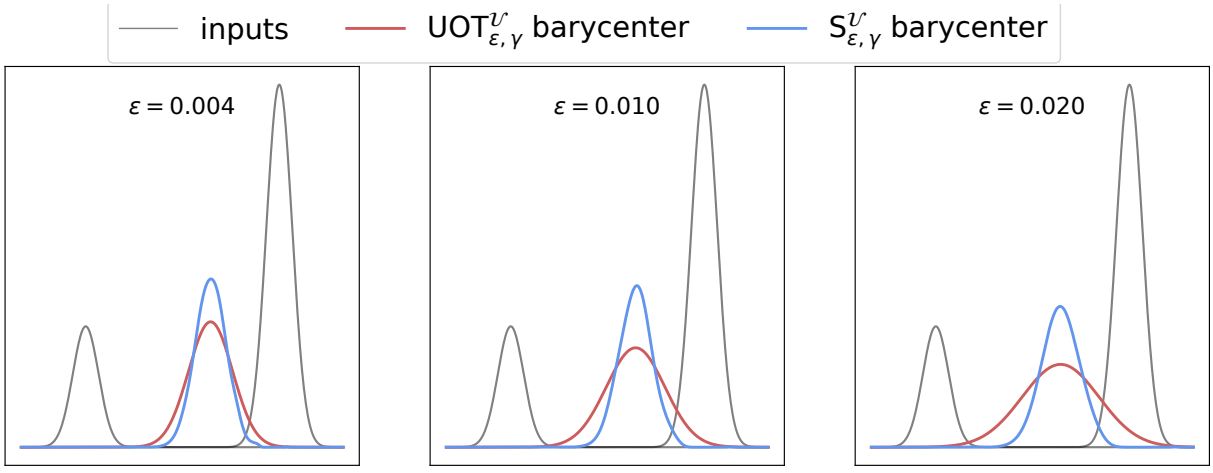


Fig. 2.16. Barycenters of 2 unbalanced Gaussian distributions. The debiased barycenter $S_{\varepsilon,\gamma}^{\mathcal{U}}$ is less sensitive to entropy regularization (ε) than $UOT_{\varepsilon,\gamma}^{\mathcal{U}}$.

and Niles-Weed (2019). Enlarging this set of measures can be done by establishing these bounds directly. Moreover, it is important to keep in mind that the sub-Gaussian assumptions is intimately linked with the use of the quadratic cost.

Entropic OT closed forms A long-standing missing piece of the entropic OT puzzle was the existence of closed form expressions for non-trivial parameters. The closed form we obtained is similar to that of the Bures-Wasserstein metric. Since the latter can be generalized to Elliptical distributions, It is thus natural to wonder whether such an extension is possible for entropic OT as well. Our proofs are however based on the stability of Sinkhorn's equations for quadratic potentials. At the heart of this stability property lies the simplicity of integrating exponentials of quadratic forms. If the same reasoning is followed, exploring other distributions would probably require changing the cost function.

Entropic OT barycenters for Gaussians The purpose of our barycenter theorems was to highlight how the variance of the barycenter is effected for different OT formulations. Yet we cannot help but ask how can such barycenters be computed ? All equations that define the variance of the barycenter can be written as fixed point equations. Numerically, the natural fixed point algorithm of these equations converge for virtually any initialization. Whether these iterations are contractant for some metric is still an open question even for the non-entropic case studied by Agueh and Carlier (2011).

Debiased barycentric algorithms We provided two GPU friendly algorithms to compute debiased barycenters for both balanced and unbalanced measures. While these algorithms converge well in practice, numerical evidence suggests that they are not contractant for the Thompson metric (for which Sinkhorn is). Theoretical understanding of these algorithms thus requires going beyond usual entropic OT convergence techniques.

5 Appendix

We provide in this appendix the technical proofs of the closed forms and the Gaussian barycenters theorems.

5.1 Proofs of the closed forms

Proof of Proposition 7. PROOF. Let $\mathbf{U}_0 = \mathbf{V}_0 = 0$. Applying Proposition 6 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 $\frac{g_n}{2\sigma^2} = \mathcal{Q}(\mathbf{V}_n)$ where:

$$\mathbf{V}_{n+1} = \frac{1}{\sigma^2}((\sigma^2\mathbf{U}_n + \sigma^2\mathbf{A}^{-1} + \mathrm{Id})^{-1} - \mathrm{Id})$$
$$\mathbf{U}_{n+1} = \frac{1}{\sigma^2}((\sigma^2\mathbf{V}_{n+1} + \sigma^2\mathbf{B}^{-1} + \mathrm{Id})^{-1} - \mathrm{Id}).$$

The change of variable:

$$\mathbf{F}_n = \sigma^2\mathbf{U}_n + \sigma^2\mathbf{A}^{-1} + \mathrm{Id}$$
$$\mathbf{G}_n = \sigma^2\mathbf{V}_n + \sigma^2\mathbf{B}^{-1} + \mathrm{Id}$$

leads to (2.46).

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

$$\mathbf{F} \leftarrow \sigma^2\mathbf{A}^{-1} + (\sigma^2\mathbf{B}^{-1} + \mathbf{F}^{-1})^{-1} \tag{2.154}$$

Let $\forall \mathbf{X} \in \mathcal{S}^d_{++}, \phi(\mathbf{X}) \stackrel{\text{def}}{=} \sigma^2\mathbf{A}^{-1} + (\sigma^2\mathbf{B}^{-1} + \mathbf{X}^{-1})^{-1} \in \mathcal{S}^d_{++}$. The first differential of ϕ admits the following expression:

$$\forall \mathbf{X} \in \mathcal{S}^d_{++}, \forall \mathbf{H} \in \mathbb{R}^{d\times d}, D\phi(\mathbf{X})[\mathbf{H}] = (\mathrm{Id} + \sigma^2\mathbf{X}\mathbf{B}^{-1})^{-1}\mathbf{H}(\sigma^2\mathbf{B}^{-1}\mathbf{X} + \mathrm{Id})^{-1}. \tag{2.155}$$

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

$$\|(\mathrm{Id} + \sigma^2\mathbf{X}\mathbf{B}^{-1})^{-1}\|_{\mathrm{op}} = \|(\mathrm{Id} + \sigma^2\mathbf{B}^{-\frac{1}{2}}\mathbf{X}\mathbf{B}^{-\frac{1}{2}})^{-1}\|_{\mathrm{op}} < 1 \ ,$$

which implies that $\|D\phi(\mathbf{X})\|_{\mathrm{op}} < 1$ for $\mathbf{X} \in \mathcal{S}^d_{++}$ and $\sigma^2 > 0$. The same arguments hold for the iterates $(\mathbf{G}_n)_{n\geq 0}$.

From (2.154) and using Weyl's inequality, we can bound the smallest eigenvalue of $\mathbf{F}_n$ from under: $\forall n, \lambda_d(\mathbf{F}_n) \geq \frac{\sigma^2}{\lambda_1(\mathbf{A})}$ (where $\lambda_d(\mathbf{F})$ is the smallest eigenvalue of $\mathbf{F}$ and $\lambda_1(\mathbf{A})$ is the biggest eigenvalue of $\mathbf{A}$).

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

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

Which proves the uniform bound ■

Proof of Lemma 3 PROOF. It follows from elementary properties of Gaussian measures that the first and second marginals of π are respectively α and β. Hence,

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 \mathrm{d}\pi(x, y) = \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x\|^2 \mathrm{d}\pi(x, y) + \int_{\mathbb{R}^d \times \mathbb{R}^d} \|y\|^2 \mathrm{d}\pi(x, y) - 2 \int_{\mathbb{R}^d \times \mathbb{R}^d} \langle x, y \rangle \mathrm{d}\pi(x, y) \tag{2.156}$$

$$= \int_{\mathbb{R}^d} \|x\|^2 \mathrm{d}\alpha(x) + \int_{\mathbb{R}^d} \|y\|^2 \mathrm{d}\beta(y) - 2 \int_{\mathbb{R}^d \times \mathbb{R}^d} \langle x, y \rangle \mathrm{d}\pi(x, y) \tag{2.157}$$

$$= \mathrm{Tr}(\mathbf{A}) + \mathrm{Tr}(\mathbf{B}) - 2\mathrm{Tr}(\mathbf{C}). \tag{2.158}$$

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

$$\mathrm{KL}\left(\pi \| \alpha \otimes \beta\right) = \tfrac{1}{2} \left(\mathrm{Tr}\left[\begin{pmatrix} \mathbf{A} & 0 \\ 0 & \mathbf{B} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{A} & \mathbf{C} \\ \mathbf{C}^T & \mathbf{B} \end{pmatrix} \right] - 2n + \log\det \begin{pmatrix} \mathbf{A} & 0 \\ 0 & \mathbf{B} \end{pmatrix} - \log\det \begin{pmatrix} \mathbf{A} & \mathbf{C} \\ \mathbf{C}^T & \mathbf{B} \end{pmatrix} \right) \tag{2.159}$$

$$= \tfrac{1}{2} \left(\log\det \mathbf{A} + \log\det \mathbf{B} - \log\det \begin{pmatrix} \mathbf{A} & \mathbf{C} \\ \mathbf{C}^T & \mathbf{B} \end{pmatrix} \right). \tag{2.160}$$

Optimal transport plan and $\mathrm{OT}^{\otimes}_{2\sigma^2}$

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

with $\Gamma \overset{\text{def}}{=} \begin{pmatrix}\frac{\mathbf{F}}{\sigma^2} & -\frac{\mathrm{Id}}{\sigma^2}\\ -\frac{\mathrm{Id}}{\sigma^2} & \frac{\mathbf{G}}{\sigma^2}\end{pmatrix}$. Moreover, since $\frac{\mathbf{G}}{2\sigma^2} \succ 0$, and its Schur complement satisfies $\frac{\mathbf{F}}{\sigma^2} - \frac{1}{\sigma^2}\mathbf{G}^{-1} = \mathbf{A}^{-1} \succ 0$, we have that $\Gamma \succ 0$. Therefore π is a Gaussian $\mathcal{N}(\mathbf{H})$ with the covariance matrix given by the block inverse formula:

$$
\begin{aligned}
\mathbf{H} &= \Gamma^{-1} && (2.161)\\
&= \sigma^2\begin{pmatrix}(\mathbf{F}-\mathbf{G}^{-1})^{-1} & (\mathbf{G}\mathbf{F}-\mathrm{Id})^{-1}\\ (\mathbf{F}\mathbf{G}-\mathrm{Id})^{-1} & (\mathbf{G}-\mathbf{F}^{-1})^{-1}\end{pmatrix} && (2.162)\\
&= \begin{pmatrix}\mathbf{A} & \mathbf{C}\\ \mathbf{C}^\top & \mathbf{B}\end{pmatrix}, && (2.163)
\end{aligned}
$$

where we used the optimality equations (2.49) and the definition of $\mathbf{C} = \mathbf{A}\mathbf{G}^{-1}$.

We can now conclude the proof of Theorem 1 by computing $\mathrm{OT}^{\otimes}_{2\sigma^2}(\alpha,\beta)$ using Lemma 3. Let $\mathbf{R} = \mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}}$. Using the closed form expression of $\mathbf{C}$ in (2.51), it first holds that

$$
\mathbf{Z} \overset{\text{def}}{=} \mathbf{A}^{-\frac{1}{2}}\mathbf{C}\mathbf{A}^{\frac{1}{2}} = (\mathbf{R}+\tfrac{\sigma^4}{4}\,\mathrm{Id})^{\frac{1}{2}} - \tfrac{\sigma^2}{2}\,\mathrm{Id}\,. \tag{2.164}
$$

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

$$\begin{aligned}
\det \left(\begin{smallmatrix} \mathbf{A} & \mathbf{C} \\ \mathbf{C}^T & \mathbf{B} \end{smallmatrix} \right) &= \det(\mathbf{A}) \det(\mathbf{B} - \mathbf{C}^\top \mathbf{A}^{-1} \mathbf{C}) \\
&= \det(\mathbf{A}^{\frac{1}{2}} \mathbf{B} \mathbf{A}^{\frac{1}{2}} - \mathbf{A}^{\frac{1}{2}} \mathbf{C}^\top \mathbf{A}^{-1} \mathbf{C} \mathbf{A}^{\frac{1}{2}}) \\
&= \det(\mathbf{R} - \mathbf{Z}^\top \mathbf{Z}) \\
&= \det(\mathbf{R} - \mathbf{Z}^2) \qquad (2.165) \\
&= \det(\sigma^2 (\mathbf{R} + \frac{\sigma^4}{4} \operatorname{Id})^{\frac{1}{2}} - \frac{\sigma^4}{2} \operatorname{Id}) \\
&= (\frac{\sigma^2}{2})^d \det((4\mathbf{R} + \sigma^4 \operatorname{Id})^{\frac{1}{2}} - \sigma^2 \operatorname{Id}).
\end{aligned}$$

Since the matrices inside the determinant commute, we can use the identity $\pi - \mathbf{Q} = (\pi^2 - \mathbf{Q}^2)(\pi + \mathbf{Q})^{-1}$ to get rid of the negative sign. Equation (2.165) then becomes:

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

Plugging this expression in (2.55), the determinant of $\mathbf{A}$ and $\mathbf{B}$ cancel out and we finally get:

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

Proof of Proposition 9 PROOF. Using Lemma 3, eq. (2.35) becomes

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

which gives eq. (2.56). Let us now prove eq. (2.57). 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. $\mathbf{K} \in \mathbb{R}^d : \|\mathbf{K}\|_{\text{op}} \leq 1$) such that $\mathbf{C} = \mathbf{A}^{\frac{1}{2}} \mathbf{K} \mathbf{B}^{\frac{1}{2}}$ (Bhatia, 2007, Ch. 1).[1] With this parameterization, we have (using Schur complements) that

$$\begin{aligned}
\det \left(\begin{smallmatrix} \mathbf{A} & \mathbf{C} \\ \mathbf{C}^T & \mathbf{B} \end{smallmatrix} \right) &= \det \mathbf{B} \det(\mathbf{A} - \mathbf{C}\mathbf{B}^{-1}\mathbf{C}^\top) \\
&= \det \mathbf{B} \det \mathbf{A} \det(\operatorname{Id} - \mathbf{K}\mathbf{K}^\top)
\end{aligned}$$

[1] Another immediate NSC is $\mathbf{A} \geq \mathbf{C}\mathbf{B}^{-1}\mathbf{C}^T$

Hence, injecting this in Equation (2.56), we have the following equivalent problem:

$$\mathfrak{B}^{\otimes}_{\sigma^2}(\mathbf{A},\mathbf{B}) = \min_{\mathbf{K}\in\mathbb{R}^{d\times d}:\|\mathbf{K}\|_{\mathrm{op}}\leq 1} \mathrm{Tr}\mathbf{A} + \mathrm{Tr}\mathbf{B} - 2\mathrm{Tr}\mathbf{A}^{\frac{1}{2}}\mathbf{K}\mathbf{B}^{\frac{1}{2}} - \sigma^2 \ln\det(\mathrm{Id} - \mathbf{K}\mathbf{K}^\top) \tag{2.166}$$

Let's prove that both problems are convex.

- (2.56): The set $\{\mathbf{C} : \left(\begin{smallmatrix}\mathbf{A} & \mathbf{C}\\ \mathbf{C}^T & \mathbf{B}\end{smallmatrix}\right) \succeq 0\}$ is convex, since $\left(\begin{smallmatrix}\mathbf{A} & \mathbf{C}_1\\ \mathbf{C}_1^T & \mathbf{B}\end{smallmatrix}\right) \succeq 0$ and $\left(\begin{smallmatrix}\mathbf{A} & \mathbf{C}_2\\ \mathbf{C}_2^T & \mathbf{B}\end{smallmatrix}\right) \succeq 0$ implies that $\left(\begin{smallmatrix}\mathbf{A} & (1-\theta)\mathbf{C}_1+\theta\mathbf{C}_2\\ (1-\theta)\mathbf{C}_1^T+\theta\mathbf{C}_2^T & \mathbf{B}\end{smallmatrix}\right) = (1-\theta)\left(\begin{smallmatrix}\mathbf{A} & \mathbf{C}_1\\ \mathbf{C}_1^T & \mathbf{B}\end{smallmatrix}\right) + \theta\left(\begin{smallmatrix}\mathbf{A} & \mathbf{C}_2\\ \mathbf{C}_2^T & \mathbf{B}\end{smallmatrix}\right) \succeq 0$. 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 (2.56) is convex.

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

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

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

Proof of Proposition 10 PROOF. By Proposition 9, (2.56) is convex, hence strong duality holds. Ignoring the terms not depending on $\mathbf{C}$, problem (2.56) can be written using the redundant parameterization $\mathbf{X} = \left(\begin{smallmatrix}\mathbf{X}_1 & \mathbf{X}_2\\ \mathbf{X}_3 & \mathbf{X}_4\end{smallmatrix}\right)$:

$$\mathfrak{D}(\mathbf{A},\mathbf{B}) \stackrel{\mathrm{def}}{=} \min_{\substack{\mathbf{X}\succ 0\\ \mathbf{X}_1=\mathbf{A},\mathbf{X}_4=\mathbf{B}}} -\mathrm{Tr}(\mathbf{X}_2) - \mathrm{Tr}(\mathbf{X}_3) - \sigma^2\log\det(\mathbf{X}) \tag{2.167}$$

$$= \min_{\substack{\mathbf{X}\succ 0\\ \mathbf{X}_1=\mathbf{A},\mathbf{X}_4=\mathbf{B}}} -\langle \mathbf{X}, \left(\begin{smallmatrix}0 & \mathrm{Id}\\ \mathrm{Id} & 0\end{smallmatrix}\right)\rangle - \sigma^2\log\det(\mathbf{X}) \tag{2.168}$$

$$= \min_{\substack{\mathbf{X}\succ 0\\ \mathbf{X}_1=\mathbf{A},\mathbf{X}_4=\mathbf{B}}} \mathcal{F}(\mathbf{X}), \tag{2.169}$$

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

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

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:

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

Where the last equality follows from the fact that Id and $\mathbf{G}$ commute. Therefore, reinserting the discarded trace terms, the dual problem of (2.56) can be written as

$$\begin{aligned}
\max_{\mathbf{F},\mathbf{G}\succ 0} \Big\{ &-\langle \mathbf{F},\, \mathbf{A}\rangle - \langle \mathbf{G},\, \mathbf{B}\rangle + \sigma^2 \log\det(\mathbf{F}\mathbf{G} - \mathrm{Id}) \\
&+ \mathrm{Tr}(\mathbf{A}) + \mathrm{Tr}(\mathbf{B}) + \sigma^2 \log\det \mathbf{A}\mathbf{B} + 2d\sigma^2(1 - \log\sigma^2)) \Big\}.
\end{aligned} \tag{2.170}$$

Proof of Proposition 11 PROOF. (i) *Optimality:* Canceling out the gradients in eq. (2.58) leads to the following optimality conditions:

$$\begin{aligned}
-A + \sigma^2 \mathbf{G}(\mathbf{F}\mathbf{G} - \mathrm{Id})^{-1} &= 0 \\
-B + \sigma^2 (\mathbf{F}\mathbf{G} - \mathrm{Id})^{-1}\mathbf{F} &= 0,
\end{aligned} \tag{2.171}$$

i.e.

$$\begin{aligned}
\mathbf{F} &= \sigma^2 \mathbf{A}^{-1} + \mathbf{G}^{-1} \\
\mathbf{G} &= \sigma^2 \mathbf{B}^{-1} + \mathbf{F}^{-1}
\end{aligned} \tag{2.172}$$

Thus $(\mathbf{F}, \mathbf{G})$ is a solution of the Sinkhorn fixed point equation (2.49).

(ii) *Differentiabilty:* Using Danskin's theorem on problem (2.58) 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 7, we recover the dual potentials of Corollary 2:

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

Using Corollary 2, it holds that

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

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

(iii) *Convexity:* Assume without loss of generality that $\mathbf{B}$ is fixed and let $G : \mathbf{B} \mapsto \nabla_{\mathbf{A}} \mathfrak{B}^{\otimes}_{\sigma^2}(\mathbf{A}, \mathbf{B})$. As long as $\sigma > 0$, G is differentiable as a composition of differentiable functions. Let's show that the Hessian of $\psi : \mathbf{A} \mapsto \mathfrak{B}^{\otimes}_{\sigma^2}(\mathbf{A}, \mathbf{B})$ is a positive quadratic form. Take a direction $\mathbf{H} \in \mathcal{S}^d_+$. It holds:

$$\begin{aligned}\nabla^2_{\mathbf{A}} \mathfrak{B}^{\otimes}_{\sigma^2}(\mathbf{A}, \mathbf{B})(\mathbf{H}, \mathbf{H}) &= \langle \mathbf{H}, \mathrm{Jac}_G(\mathbf{A})(\mathbf{H})\rangle \\ &= \mathrm{Tr}(\mathbf{H}\,\mathrm{Jac}_G(\mathbf{A})(\mathbf{H})).\end{aligned}$$

For the sake of clarity, let's write $G(\mathbf{A}) = \mathrm{Id} - L(W(\phi(\mathbf{A})))$ with the following intermediary functions:

$$\begin{aligned}L &: \mathbf{A} \mapsto \mathbf{B}^{\frac{1}{2}}\mathbf{A}\mathbf{B}^{\frac{1}{2}} \\ Q &: \mathbf{A} \mapsto \mathbf{A}^{\frac{1}{2}} \\ \phi &: \mathbf{A} \mapsto Q(L(\mathbf{A}) + \frac{\sigma^4}{4}\mathrm{Id}) \\ W &: \mathbf{A} \mapsto (\mathbf{A} + \frac{\sigma^2}{2}\mathrm{Id})^{-1}.\end{aligned}$$

Moreover, their derivatives are given by:

$$\begin{aligned}
\mathrm{Jac}_L(\mathbf{A})(\mathbf{H}) &= \mathbf{B}^{\frac{1}{2}}\mathbf{H}\mathbf{B}^{\frac{1}{2}} \\
\mathrm{Jac}_W(\mathbf{A})(\mathbf{H}) &= -(\mathbf{A}+\frac{\sigma^2}{2}\,\mathrm{Id})^{-1}\mathbf{H}(\mathbf{A}+\frac{\sigma^2}{2}\,\mathrm{Id})^{-1} \\
\mathrm{Jac}_Q(\mathbf{A})(\mathbf{H}) &= \mathbf{Z},
\end{aligned}$$

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:

$$\begin{aligned}
\mathrm{Jac}_G(\mathbf{A})(\mathbf{H}) &= -\,\mathrm{Jac}_L(W(\phi(\mathbf{A})))(\mathrm{Jac}_W(\phi(\mathbf{A}))(\mathrm{Jac}_\phi(\mathbf{A})(\mathbf{H}))) \\
&= -\mathbf{B}^{\frac{1}{2}}\,\mathrm{Jac}_W(\phi(\mathbf{A}))(\mathrm{Jac}_\phi(\mathbf{A})(\mathbf{H}))\mathbf{B}^{\frac{1}{2}} \\
&= \mathbf{B}^{\frac{1}{2}}\left(\phi(\mathbf{A})+\frac{\sigma^2}{2}\,\mathrm{Id}\right)^{-1}\mathrm{Jac}_\phi(\mathbf{A})(\mathbf{H})\left(\phi(\mathbf{A})+\frac{\sigma^2}{2}\,\mathrm{Id}\right)^{-1}\mathbf{B}^{\frac{1}{2}} \\
&= \mathbf{B}^{\frac{1}{2}}\left(\mathbf{D}^{\frac{1}{2}}+\frac{\sigma^2}{2}\,\mathrm{Id}\right)^{-1}\mathrm{Jac}_\phi(\mathbf{A})(\mathbf{H})\left(\mathbf{D}^{\frac{1}{2}}+\frac{\sigma^2}{2}\,\mathrm{Id}\right)^{-1}\mathbf{B}^{\frac{1}{2}}.
\end{aligned}$$

Again using the chain rule:

$$\begin{aligned}
\mathbf{Y} \overset{\text{def}}{=} \mathrm{Jac}_\phi(\mathbf{A})(\mathbf{H}) &= \mathrm{Jac}_Q(L(\mathbf{A})+\frac{\sigma^4}{4}\,\mathrm{Id})((\mathrm{Jac}_L(\mathbf{A}))(\mathbf{H})) \\
&= \mathrm{Jac}_Q(L(\mathbf{A})+\frac{\sigma^4}{4}\,\mathrm{Id})(\mathbf{B}^{\frac{1}{2}}\mathbf{H}\mathbf{B}^{\frac{1}{2}}) \\
&= \mathrm{Jac}_Q(\mathbf{D})(\mathbf{B}^{\frac{1}{2}}\mathbf{H}\mathbf{B}^{\frac{1}{2}}).
\end{aligned}$$

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:

$$\begin{aligned}\nabla^2_{\mathbf{A}}\mathfrak{B}^{\otimes}_{\sigma^2}(\mathbf{A},\mathbf{B})(\mathbf{H},\mathbf{H}) &= \langle \mathbf{H}, \mathrm{Jac}_G(\mathbf{A})(\mathbf{H})\rangle \\ &= \mathrm{Tr}\left(\mathbf{H}\,\mathrm{Jac}_G(\mathbf{A})(\mathbf{H})\right) \\ &= \mathrm{Tr}\left(\mathbf{H}\mathbf{B}^{\frac{1}{2}}\left(\mathbf{D}^{\frac{1}{2}}+\frac{\sigma^2}{2}\,\mathrm{Id}\right)^{-1}\mathbf{Y}\left(\mathbf{D}^{\frac{1}{2}}+\frac{\sigma^2}{2}\,\mathrm{Id}\right)^{-1}\mathbf{B}^{\frac{1}{2}}\right) \\ &= \mathrm{Tr}\left(\mathbf{B}^{\frac{1}{2}}\mathbf{H}\mathbf{B}^{\frac{1}{2}}\left(\mathbf{D}^{\frac{1}{2}}+\frac{\sigma^2}{2}\,\mathrm{Id}\right)^{-1}\mathbf{Y}\left(\mathbf{D}^{\frac{1}{2}}+\frac{\sigma^2}{2}\,\mathrm{Id}\right)^{-1}\right).\end{aligned}$$

Since $\mathbf{H}$ and $\mathbf{Y}$ are positive, the matrices $\mathbf{B}^{\frac{1}{2}}\mathbf{H}\mathbf{B}^{\frac{1}{2}}$ and $\left(\mathbf{D}^{\frac{1}{2}}+\frac{\sigma^2}{2}\,\mathrm{Id}\right)^{-1}\mathbf{Y}\left(\mathbf{D}^{\frac{1}{2}}+\frac{\sigma^2}{2}\,\mathrm{Id}\right)^{-1}$ are positive semi-definite as well. Their product is similar to a positive semi-definite matrix, therefore the trace above is non-negative.

Given that $\mathbf{A}$ and $\mathbf{H}$ are arbitrary positive semi-definite matrices, it holds that

$$\nabla^2_{\mathbf{A}}\mathfrak{B}^{\otimes}_{\sigma^2}(\mathbf{A},\mathbf{B})(\mathbf{H},\mathbf{H}) \geq 0$$

Therefore, $\mathbf{A} \mapsto \mathfrak{B}^{\otimes}_{\sigma^2}(\mathbf{A},\mathbf{B})$ is convex.

Counter-example of joint convexity: If $\mathfrak{B}^{\otimes}_{\sigma^2}$ were jointly convex , then $\delta \overset{\text{def}}{=}: \mathbf{A} \to \mathfrak{B}^{\otimes}_{\sigma^2}(\mathbf{A},\mathbf{A})$ would be a convex function.

In the 1-dimensional case with $\sigma = 1$, one can see that this would be equivalent to $x \to \ln((x^2+1)^{\frac{1}{2}} + 1) - (x^2+1)^{\frac{1}{2}}$ being convex, whereas it is in fact strictly concave.

(iv) Minimizer of $\phi_{\mathbf{B}}$ With fixed $\mathbf{B}$, cancelling the gradient of $\phi_{\mathbf{B}} \overset{\text{def}}{=}: \mathbf{A} \mapsto \mathfrak{B}^{\otimes}_{\sigma^2}(\mathbf{A},\mathbf{B})$ leads to $\mathbf{A} = \mathbf{B} - \sigma^2\,\mathrm{Id}$ which is well defined if and only if $\mathbf{B} \succeq \sigma^2\,\mathrm{Id}$. However, if $\mathbf{B} - \sigma^2\,\mathrm{Id}$ is not positive semi-definite, write the eigenvalue decomposition: $\mathbf{B} = \pi\Sigma\pi^\top$ and define $\mathbf{A}_0 \overset{\text{def}}{=} \pi(\Sigma - \sigma^2\,\mathrm{Id})_+\pi^\top$ where the operator $x_+ = \max(x,0)$ is applied element-wise. Then:

$$\begin{aligned}\nabla_{\mathbf{A}}\phi_{\mathbf{B}}(\mathbf{A}_0) &= \mathrm{Id} - \pi\Sigma^{\frac{1}{2}}\pi^\top\left((\pi(\Sigma^2-\sigma^2\Sigma)_+\pi^\top + \frac{\sigma^4}{4}\,\mathrm{Id})^{\frac{1}{2}} + \frac{\sigma^2}{2}\,\mathrm{Id}\right)^{-1}\pi\Sigma^{\frac{1}{2}}\pi^\top \\ &= \mathrm{Id} - \pi\Sigma^{\frac{1}{2}}\left(((\Sigma^2-\sigma^2\Sigma)_+ + \frac{\sigma^4}{4}\,\mathrm{Id})^{\frac{1}{2}} + \frac{\sigma^2}{2}\,\mathrm{Id}\right)^{-1}\Sigma^{\frac{1}{2}}\pi^\top \\ &= \mathrm{Id} - \pi\Sigma^{\frac{1}{2}}\left((\Sigma-\sigma^2\,\mathrm{Id})_+ + \sigma^2\,\mathrm{Id}\right)^{-1}\Sigma^{\frac{1}{2}}\pi^\top \\ &= \pi(\mathrm{Id} - \Sigma^{\frac{1}{2}}\left((\Sigma-\sigma^2\,\mathrm{Id})_+ + \sigma^2\,\mathrm{Id}\right)^{-1}\Sigma^{\frac{1}{2}})\pi^\top \\ &= \frac{1}{\sigma^2}\pi(\sigma^2\,\mathrm{Id} - \Sigma)_+\pi^\top\end{aligned}$$

Thus, given that $(\Sigma - \sigma^2 \operatorname{Id})_+ (\sigma^2 \operatorname{Id} - \Sigma)_+ = 0$, it holds, for any $\mathbf{H} \in \mathcal{S}_+^d$:

$$\begin{aligned}\langle \mathbf{H} - \mathbf{A}_0, \nabla_{\mathbf{A}} \phi_{\mathbf{B}}(\mathbf{A}_0) \rangle &= \langle \pi^\top \mathbf{H} \pi - (\Sigma - \sigma^2 \operatorname{Id})_+, (\sigma^2 \operatorname{Id} - \Sigma)_+ \rangle \\ &= \langle \pi^\top \mathbf{H} \pi, (\sigma^2 \operatorname{Id} - \Sigma)_+ \rangle \\ &= \operatorname{Tr}(\pi^\top \mathbf{H} \pi (\sigma^2 \operatorname{Id} - \Sigma)_+) \geq 0\end{aligned}$$

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 $\mathbf{A}_0$.

Proof of Proposition 13 PROOF. Using Fubini-Tonelli along with the optimality conditions (2.63), the double integral can be written:

$$\begin{aligned}\pi(\mathbb{R}^d \times \mathbb{R}^d) &= \int_{\mathbb{R}^d \times \mathbb{R}^d} e^{\frac{-\|x-y\|^2 + f(x) + g(y)}{2\sigma^2}} \mathrm{d}\alpha(x) \mathrm{d}\beta(y) \\ &= \int_{\mathbb{R}^d} \left(\int_{\mathbb{R}^d} e^{\frac{-\|x-y\|^2 + f(x)}{2\sigma^2}} \mathrm{d}\alpha(x) \right) e^{\frac{g(y)}{2\sigma^2}} \mathrm{d}\beta(y) \\ &= \int_{\mathbb{R}^d} e^{\frac{g(y)}{2\sigma^2}(1 - \frac{1}{\tau})} \mathrm{d}\beta(y) \\ &= \int_{\mathbb{R}^d} e^{-\frac{g(y)}{\gamma}} \mathrm{d}\beta(y)\end{aligned}$$

And similarly: $\pi(\mathbb{R}^d \times \mathbb{R}^d) = \int_{\mathbb{R}^d} e^{-\frac{f(x)}{\gamma}} \mathrm{d}\alpha(x)$. Therefore, the three integrals in the dual objective (2.62) are equal to $\pi(\mathbb{R}^d \times \mathbb{R}^d)$ which ends the proof.

Lemma 6 *[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:*

$$P_\alpha(x) + P_\beta(x) = -\frac{1}{2}\left((\mathbf{x} - \mathbf{c})^\top \mathbf{C}(\mathbf{x} - \mathbf{c}) + q_{\alpha,\beta} \right) \tag{2.173}$$

where:

$$\begin{cases} \mathbf{C} & = \mathbf{A} + \mathbf{B} \\ (\mathbf{A} + \mathbf{B})\mathbf{c} & = (\mathbf{A}\mathbf{a} + \mathbf{B}\mathbf{b}) \\ q_{\alpha,\beta} & = \mathbf{a}^\top \mathbf{A}\mathbf{a} + \mathbf{b}^\top \mathbf{B}\mathbf{b} - c^\top \mathbf{C}\mathbf{c} \end{cases} \tag{2.174}$$

In particular, if $\mathbf{C} = \mathbf{A} + \mathbf{B}$ is invertible, then:

$$\begin{cases} \mathbf{c} = \mathbf{C}^{-1}(\mathbf{A}\mathbf{a} + \mathbf{B}\mathbf{b}) \\ \mathbf{c}^\top \mathbf{C}\mathbf{c} = (\mathbf{A}\mathbf{a} + \mathbf{B}\mathbf{b})^\top \mathbf{C}^{-1}(\mathbf{A}\mathbf{a} + \mathbf{B}\mathbf{b}) \end{cases} \tag{2.175}$$

PROOF. On one hand,

$$\begin{aligned} P_\alpha(x) + P_\beta(x) &= -\frac{1}{2}\Big((\mathbf{x}-\mathbf{a})^\top \mathbf{A}(\mathbf{x}-\mathbf{a}) + (\mathbf{x}-\mathbf{b})^\top \mathbf{B}(\mathbf{x}-\mathbf{b})\Big) \\ &= -\frac{1}{2}\Big(\mathbf{x}^\top(\mathbf{A}+\mathbf{B})\mathbf{x} - 2\mathbf{x}^\top(\mathbf{A}\mathbf{a}+\mathbf{B}\mathbf{b}) + \mathbf{a}^\top\mathbf{A}\mathbf{a} + \mathbf{b}^\top\mathbf{B}\mathbf{b}\Big) \end{aligned}$$

On the other hand, for an arbitrary $\gamma = (\mathbf{c}, \mathbf{C})$ and $q \in \mathbb{R}$:

$$\begin{aligned} P_\gamma(x) - \frac{q}{2} &= -\frac{1}{2}\Big((\mathbf{x}-\mathbf{c})^\top \mathbf{C}(\mathbf{x}-\mathbf{C}) + q\Big) \\ &= -\frac{1}{2}\Big(x^\top \mathbf{C} x - 2x^\top \mathbf{C}\mathbf{c} + \mathbf{c}^\top\mathbf{C}\mathbf{c} + q\Big) \end{aligned}$$

If $\mathbf{A} \neq \mathbf{B}$, identification of the parameters of both quadratic forms leads to (2.174).

Lemma 7 *[Gaussian convolution of generic quadratic forms] Let $\mathbf{A} \in S_d$ and $\mathbf{a} \in \mathbb{R}^d$ and $\sigma > 0$ such that $\sigma^2\mathbf{A} + \mathrm{Id} \succ 0$. 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 $\mathcal{N}(0, \frac{\mathrm{Id}}{\sigma^2})$ is given by:*

$$\mathcal{N}(0, \frac{\mathrm{Id}}{\sigma^2}) \star \exp\left(\mathcal{Q}_\alpha\right) \stackrel{\text{def}}{=} \int_{\mathbb{R}^d} \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left(-\frac{1}{2\sigma^2}\|. - y\|^2 + \mathcal{Q}_\alpha(y)\right) \mathrm{d}y = c_\alpha \exp(\mathcal{Q}(\mathbf{G}\mathbf{a}, \mathbf{G}\mathbf{A})) \quad (2.176)$$

where:

$$\begin{aligned} \mathbf{G} &= (\sigma^2\mathbf{A} + \mathrm{Id})^{-1} \\ c_\alpha &= \frac{e^{\frac{\sigma^2\mathbf{a}^\top\mathbf{G}\mathbf{a}}{2}}}{\sqrt{\det(\sigma^2\mathbf{A} + \mathrm{Id})}} \end{aligned}$$

PROOF. Using Lemma 6 one can write for any $x \in \mathbb{R}^d$ considered fixed:

$$\begin{aligned} -\frac{1}{2\sigma^2}\|x-y\|^2 + \mathcal{Q}_\alpha(y) &= \mathcal{Q}(x, \frac{\mathrm{Id}}{\sigma^2})(y) + \mathcal{Q}(\mathbf{a}, \mathbf{A})(y) \\ &= \mathcal{Q}(\mathbf{a} + \frac{x}{\sigma^2}, \mathbf{A} + \frac{\mathrm{Id}}{\sigma^2})(y) - \frac{1}{2\sigma^2}\|x\|^2 \\ &= P(\sigma\mathbf{a} + \frac{x}{\sigma^2}, \mathbf{A} + \frac{\mathrm{Id}}{\sigma^2})(y) + h(x) \end{aligned}$$

with $h(x) = -\frac{1}{2}\left(\frac{1}{\sigma^2}\|x\|^2 - \frac{1}{\sigma^2}(\sigma^2\mathbf{a}+x)^\top(\sigma^2\mathbf{A}+\mathrm{Id})^{-1}(\sigma^2\mathbf{a}+x)\right)$. Therefore, the convolution integral is finite if and only if $\mathbf{A} + \frac{\mathrm{Id}}{\sigma^2} \succ 0$ in which case we get the integral of a Gaussian density:

$$\begin{aligned}\frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}}\int_{\mathbb{R}^d}\exp\left(\mathcal{Q}f(\mathbf{a}+\frac{x}{\sigma^2},\mathbf{A}+\frac{\mathrm{Id}}{\sigma^2})(y)+h(x)\right)\mathrm{d}(y) &= \sqrt{\frac{\det(2\pi(\mathbf{A}+\frac{\mathrm{Id}}{\sigma^2})^{-1})}{(2\pi\sigma^2)^n}}e^{h(x)}\\ &= \frac{e^{h(x)}}{\sqrt{\det(\sigma^2\mathbf{A}+\mathrm{Id})}}\end{aligned}$$

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:

$$\begin{aligned}h_2(x) &= \frac{1}{\sigma^2}(\|x\|^2 - x^\top(\sigma^2\mathbf{A}+\mathrm{Id})^{-1}x\\ h_1(x) &= -2x^\top(\sigma^2\mathbf{A}+\mathrm{Id})^{-1}\mathbf{a}\\ h_0 &= -\sigma^2\mathbf{a}^\top(\sigma^2\mathbf{A}+\mathrm{Id})^{-1}\mathbf{a}\end{aligned}$$

Finally, we can factorize h_2 and h_0 using Woodbury's matrix identity which holds even for a singular matrix $\mathbf{A}$:

$$(\sigma^2\mathbf{A}+\mathrm{Id})^{-1} = \mathrm{Id} - \sigma^2(\sigma^2\mathbf{A}+\mathrm{Id})^{-1}\mathbf{A} \qquad \text{(Woodbury's identity)}$$

Let $\mathbf{G} = (\sigma^2\mathbf{A}+\mathrm{Id})^{-1}$.

$$\begin{aligned}h_2(x) &= \frac{1}{\sigma^2}(\|x\|^2 - x^\top(\mathrm{Id} - \sigma^2(\sigma^2\mathbf{A}+\mathrm{Id})^{-1}\mathbf{A})x\\ &= x^\top(\sigma^2\mathbf{A}+\mathrm{Id})^{-1}\mathbf{A}x\\ &= x^\top\mathbf{G}\mathbf{A}x\\ h_1(x) &= -2x^\top\mathbf{G}\mathbf{a}\\ h_0 &= -\sigma^2\mathbf{a}^\top(\sigma^2\mathbf{A}+\mathrm{Id})^{-1}\mathbf{a}\\ &= -\sigma^2\mathbf{a}^\top\mathbf{G}\mathbf{a}\end{aligned}$$

Therefore, $h(x) = -\frac{1}{2}\left(x^\top\mathbf{G}\mathbf{A}x - 2x^\top\mathbf{G}\mathbf{a} - \sigma^2\mathbf{a}^\top\mathbf{G}\mathbf{a}\right) = \mathcal{Q}(\mathbf{G}\mathbf{a},\mathbf{G}\mathbf{A})(x) + \frac{\sigma^2\mathbf{a}^\top\mathbf{G}\mathbf{a}}{2}$.

Proof of theorem 2 – closed form of unbalanced Gaussians 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. M

Proposition 24 *Let α be an unbalanced Gaussians given by$m_\alpha \mathcal{N}(\mathbf{a}, \mathbf{A})$. Let $\tau = \frac{\gamma}{2\sigma^2+\gamma}$. Define the unbalanced Sinkhorn transform $T : \mathbb{R}^{\mathbb{R}^d} \to \mathbb{R}^{\mathbb{R}^d}$:*

$$T_\alpha(h)(x) \stackrel{\text{def}}{=} -\tau \log \int_{\mathbb{R}^d} e^{\frac{-\|x-y\|^2}{2\sigma^2}+h(y)} \mathrm{d}\alpha(y) \tag{2.177}$$

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 $\mathbf{F} \stackrel{\text{def}}{=} \sigma^2\mathbf{U} + \sigma^2\mathbf{A}^{-1} + \mathrm{Id} \succ 0$, in which case $T_\alpha(h) = \mathcal{Q}(\mathbf{v}, \mathbf{V}) + \log(m_v)$ with the identified parameters:

$$\mathbf{V} = \tau \frac{1}{\sigma^2}(\mathbf{F}^{-1} - \mathrm{Id}) \tag{2.178}$$

$$\mathbf{v} = -\tau \mathbf{F}^{-1}(\mathbf{A}^{-1}\mathbf{a} + \mathbf{u}) \tag{2.179}$$

$$m_v = \left(\frac{\sqrt{\det(\mathbf{A})\det(\mathbf{F})}}{m_u m_\alpha e^{\frac{q_{u,\alpha}}{2}} \sigma^{2d}} \right)^\tau \tag{2.180}$$

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:

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

which is integrable if and only if $\mathbf{U} + \mathbf{A}^{-1} + \frac{1}{\sigma^2}\mathrm{Id} \succ 0 \Leftrightarrow \mathbf{F} \succ 0$. Moreover, up to a multiplicative factor, the exponentiated Sinkhorn transform is equivalent to a Gaussian convolution of an exponentiated

quadratic form. Lemma 7 applies:

$$
\begin{aligned}
e^{-T_\alpha(h)} &= \int_{\mathbb{R}^d} e^{\frac{-\|x-y\|^2}{2\sigma^2}+f(y)}\mathrm{d}\alpha(y) \\
&= m_u m_\alpha \frac{\exp(-\frac{1}{2}\mathbf{a}^\top\mathbf{A}^{-1}\mathbf{a})}{\sqrt{\det(2\pi\mathbf{A})}} \int_{\mathbb{R}^d} e^{\frac{-\|x-y\|^2}{2\sigma^2}+\mathcal{Q}(\mathbf{u},\mathbf{U})(y)+\mathcal{Q}(\mathbf{A}^{-1}\mathbf{a},\mathbf{A}^{-1})(y)}\mathrm{d}y \\
&= m_u m_\alpha \frac{\exp(-\frac{1}{2}\mathbf{a}^\top\mathbf{A}^{-1}\mathbf{a})}{\sqrt{\det(2\pi\mathbf{A})}} \sqrt{(2\pi\sigma^2)^{2d}} \exp\left(\mathcal{N}(\sigma^2\,\mathrm{Id})\right) \star \exp\left(\mathcal{Q}(\mathbf{u}+\mathbf{A}^{-1}\mathbf{a}, \mathbf{U}+\mathbf{A}^{-1})\right) \\
&= m_u m_\alpha \frac{\sigma^{2d}\exp(-\frac{1}{2}\mathbf{a}^\top\mathbf{A}^{-1}\mathbf{a})}{\sqrt{\det(\mathbf{A})}} \exp\left(\mathcal{N}(\sigma^2\,\mathrm{Id})\right) \star \exp\left(\mathcal{Q}(\mathbf{u}+\mathbf{A}^{-1}\mathbf{a}, \mathbf{U}+\mathbf{A}^{-1})\right) \\
&= m_u m_\alpha \frac{\sigma^{2d}\exp(-\frac{1}{2}\mathbf{a}^\top\mathbf{A}^{-1}\mathbf{a})}{\sqrt{\det(\mathbf{A})}} c_\alpha \exp\left(\mathcal{Q}(\mathbf{F}^{-1}(\mathbf{u}+\mathbf{A}^1\mathbf{a}), \mathbf{F}^{-1}(\mathbf{U}+\mathbf{A}^{-1})\right). \\
&= m_u m_\alpha \frac{\sigma^{2d}\exp(-\frac{1}{2}\mathbf{a}^\top\mathbf{A}^{-1}\mathbf{a})}{\sqrt{\det(\mathbf{A})}} c_\alpha \exp\left(\mathcal{Q}(\mathbf{F}^{-1}(\mathbf{u}+\mathbf{A}^1\mathbf{a}), \frac{1}{\sigma^2}\mathbf{F}^{-1}(\mathbf{F}-\mathrm{Id})\right). \\
&= m_u m_\alpha \frac{\sigma^{2d}\exp(-\frac{1}{2}\mathbf{a}^\top\mathbf{A}^{-1}\mathbf{a})}{\sqrt{\det(\mathbf{A})}} c_\alpha \exp\left(\mathcal{Q}(\mathbf{F}^{-1}(\mathbf{u}+\mathbf{A}^1\mathbf{a}), \frac{1}{\sigma^2}(\mathrm{Id}-\mathbf{F}^{-1})\right).
\end{aligned}
$$

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 equation of m_v. Unlike 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 (2.63).

Proposition 25 *The pair of quadratic forms (f, g) of (2.66) verifies the optimality conditions (2.63) if and only if:*

$$
\begin{aligned}
\mathbf{F} &\stackrel{\text{def}}{=} \sigma^2\mathbf{A}^{-1} + \sigma^2\mathbf{U} + \mathrm{Id} \succ 0 \\
\mathbf{G} &\stackrel{\text{def}}{=} \sigma^2\mathbf{B}^{-1} + \sigma^2\mathbf{V} + \mathrm{Id} \succ 0,
\end{aligned}
\tag{2.181}
$$

$$
\begin{aligned}
m_v\left(\frac{m_u m_\alpha e^{\frac{q_{u,\alpha}}{2}}\sigma^d}{\sqrt{\det(\mathbf{A})\det(\mathbf{F})}}\right)^\tau &= 1 \\
\mathbf{v} &= -\tau\mathbf{F}^{-1}(\mathbf{A}^{-1}\mathbf{a}+\mathbf{u}) \\
\mathbf{G} &= \tau\mathbf{F}^{-1} + \sigma^2\mathbf{B}^{-1} + (1-\tau)\,\mathrm{Id} \\
q_{u,\alpha} &= \frac{\sigma^2}{\tau^2}\mathbf{v}^\top\mathbf{F}\mathbf{v} - \mathbf{a}^\top\mathbf{A}^{-1}\mathbf{a}
\end{aligned}
\qquad
\begin{aligned}
m_u\left(\frac{m_v m_\beta e^{\frac{q_{v,\beta}}{2}}\sigma^d}{\sqrt{\det(\mathbf{B})\det(\mathbf{G})}}\right)^\tau &= 1 \\
\mathbf{u} &= -\tau\mathbf{G}^{-1}(\mathbf{B}^{-1}\mathbf{b}+\mathbf{v}) \\
\mathbf{F} &= \tau\mathbf{G}^{-1} + \sigma^2\mathbf{A}^{-1} + (1-\tau)\,\mathrm{Id} \\
q_{v,\beta} &= \frac{\sigma^2}{\tau^2}\mathbf{u}^\top\mathbf{G}\mathbf{u} - \mathbf{b}^\top\mathbf{B}^{-1}\mathbf{b}
\end{aligned}
\tag{2.182}
$$

PROOF. The equations on $m_u, m_v, \mathbf{u}, \mathbf{v}$ follow immediately from Proposition 24. Using the definition of $\mathbf{F}$ and $\mathbf{G}$, substituting $\mathbf{U}$ and $\mathbf{F}$ leads to the equations in $\mathbf{F}$ and $\mathbf{G}$

We now turn to solve the system (2.182). 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 does 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 some change of variables. Let $\lambda \overset{\text{def}}{=} \frac{\sigma^2}{1-\tau} = \sigma^2 + \frac{\gamma}{2}$.

$$
\begin{aligned}
&\begin{cases} \mathbf{F} &= \tau \mathbf{G}^{-1} + \sigma^2 \mathbf{A}^{-1} + (1-\tau)\,\mathrm{Id} \\ \mathbf{G} &= \tau \mathbf{F}^{-1} + \sigma^2 \mathbf{B}^{-1} + (1-\tau)\,\mathrm{Id} \end{cases} \\
\Leftrightarrow &\begin{cases} \mathbf{F} &= \left(\frac{\mathbf{G}}{\tau}\right)^{-1} + \frac{\sigma^2}{\tau}\tau(\mathbf{A}^{-1} + \frac{1}{\lambda}\,\mathrm{Id}) \\ \frac{\mathbf{G}}{\tau} &= \mathbf{F}^{-1} + \frac{\sigma^2}{\tau}(\mathbf{B}^{-1} + \frac{1}{\lambda}\,\mathrm{Id}) \end{cases} \\
\Leftrightarrow &\begin{cases} \mathbf{F} &= \widetilde{\mathbf{G}}^{-1} + \sigma^2(\frac{\widetilde{\mathbf{A}}}{\tau})^{-1} \\ \widetilde{\mathbf{G}} &= \mathbf{F}^{-1} + \sigma^2 \widetilde{\mathbf{B}}^{-1} \end{cases}
\end{aligned}
$$

which correspond to the balanced OT fixed point equations (2.49) associated with the pair $(\frac{\widetilde{\mathbf{A}}}{\tau}, \widetilde{\mathbf{B}})$ with the change of variables:

$$\widetilde{\mathbf{G}} \overset{\text{def}}{=} \frac{\mathbf{G}}{\tau} \tag{2.183}$$

$$\widetilde{\mathbf{A}} \overset{\text{def}}{=} \tau(\mathbf{A}^{-1} + \frac{1}{\lambda}\,\mathrm{Id})^{-1} \tag{2.184}$$

$$\widetilde{\mathbf{B}} \overset{\text{def}}{=} \tau(\mathbf{B}^{-1} + \frac{1}{\lambda}\,\mathrm{Id})^{-1} \tag{2.185}$$

Notice that since $0 < \tau < 1$, $\widetilde{\mathbf{A}}$ and $\widetilde{\mathbf{B}}$ are well-defined and positive definite. Therefore, Proposition 8 applies and we can write in closed form:

$$
\begin{aligned}
\mathbf{C} \overset{\text{def}}{=} \widetilde{\mathbf{A}}\widetilde{\mathbf{G}}^{-1} &= \left(\frac{1}{\tau}\widetilde{\mathbf{A}}\widetilde{\mathbf{B}} + \frac{\sigma^4}{4}\,\mathrm{Id}\right)^{\frac{1}{2}} - \frac{\sigma^2}{2}\,\mathrm{Id} \\
&= \widetilde{\mathbf{A}}^{\frac{1}{2}}\left(\frac{1}{\tau}\widetilde{\mathbf{A}}^{\frac{1}{2}}\widetilde{\mathbf{B}}\widetilde{\mathbf{A}}^{\frac{1}{2}} + \frac{\sigma^4}{4}\,\mathrm{Id}\right)^{\frac{1}{2}}\widetilde{\mathbf{A}}^{-\frac{1}{2}} - \frac{\sigma^2}{2}\,\mathrm{Id}
\end{aligned}
\tag{2.186}
$$

And similarly by symmetry:

$$\widetilde{\mathbf{B}}\mathbf{F}^{-1} = \left(\frac{1}{\tau}\widetilde{\mathbf{B}}\widetilde{\mathbf{A}} + \frac{\sigma^4}{4}\,\mathrm{Id}\right)^{\frac{1}{2}} - \frac{\sigma^2}{2}\,\mathrm{Id} = \mathbf{C}^\top \tag{2.187}$$

Therefore we obtain $\mathbf{F}$ and $\mathbf{G}$ in closed form:

$$\mathbf{F} = \widetilde{\mathbf{B}}\mathbf{C}^{-1} \tag{2.188}$$

$$\mathbf{G} = \mathbf{C}^{-1}\widetilde{\mathbf{A}} \tag{2.189}$$

Finally, to obtain the formulas of $\widetilde{\mathbf{A}}$ and $\widetilde{\mathbf{B}}$ of Theorem 2, use Woodburry's identity to write:

$$\begin{aligned}\widetilde{\mathbf{B}} &= \tau\lambda(\mathrm{Id} - \lambda(\mathbf{B} + \lambda\,\mathrm{Id})^{-1}) \\ &= \frac{\gamma}{\gamma + 2\sigma^2}\frac{2\sigma^2 + \gamma}{2}(\mathrm{Id} - \lambda(\mathbf{B} + \lambda\,\mathrm{Id})^{-1}) \\ &= \frac{\gamma}{2}(\mathrm{Id} - \lambda(\mathbf{B} + \lambda\,\mathrm{Id})^{-1})\end{aligned}$$

the same applies for $\widetilde{\mathbf{A}}$.

Identifying u and v. Combining the equations in $\mathbf{u}$ and $\mathbf{v}$ leads to:

$$\begin{aligned}&\mathbf{v} = -\tau\mathbf{F}^{-1}(\mathbf{A}^{-1}\mathbf{a} + \tau\mathbf{u}) \\ &\Leftrightarrow \mathbf{F}\mathbf{v} = -\tau\mathbf{A}^{-1}\mathbf{a} - \tau\mathbf{u} \\ &\Leftrightarrow \mathbf{F}\mathbf{v} = -\tau\mathbf{A}^{-1}\mathbf{a} + \tau^2\mathbf{G}^{-1}(\mathbf{B}^{-1}\mathbf{b} + \mathbf{v}) \\ &\Leftrightarrow \mathbf{G}\mathbf{F}\mathbf{v} = -\tau\mathbf{G}\mathbf{A}^{-1}\mathbf{a} + \tau^2(\mathbf{B}^{-1}\mathbf{b} + \mathbf{v}) \\ &\Leftrightarrow (\mathbf{G}\mathbf{F} - \tau^2\,\mathrm{Id})\mathbf{v} = -\tau\mathbf{G}\mathbf{A}^{-1}\mathbf{a} + \tau^2\mathbf{B}^{-1}\mathbf{b}\end{aligned}$$

Similarly, $(\mathbf{F}\mathbf{G} - \tau^2\,\mathrm{Id})\mathbf{u} = -\tau\mathbf{F}\mathbf{B}^{-1}\mathbf{b} + \tau^2\mathbf{A}^{-1}\mathbf{a}$. Moreover, since $0 < \tau < 1$, it holds$(\mathbf{F} - \tau^2\mathbf{G}^{-1}) \succ (\mathbf{F} - \tau\mathbf{G}^{-1}) = \sigma^2\widetilde{\mathbf{A}}^{-1} \succ 0$. Therefore, $(\mathbf{F}\mathbf{G} - \tau^2\,\mathrm{Id}) = (\mathbf{F} - \tau^2\mathbf{G}^{-1}\,\mathrm{Id})\mathbf{G}$ is invertible. The same applies for $(\mathbf{G}\mathbf{F} - \tau^2\,\mathrm{Id})$.

Finally, both equations can be vectorized:

$$\begin{pmatrix}\mathbf{G}\mathbf{F} - \tau^2\,\mathrm{Id} & 0 \\ 0 & \mathbf{F}\mathbf{G} - \tau^2\,\mathrm{Id}\end{pmatrix}\begin{pmatrix}\mathbf{v} \\ \mathbf{u}\end{pmatrix} = \begin{pmatrix}-\tau\mathbf{G} & \tau^2\,\mathrm{Id} \\ \tau^2\,\mathrm{Id} & -\tau\mathbf{F}\end{pmatrix}\begin{pmatrix}\mathbf{A}^{-1} & 0 \\ 0 & \mathbf{B}^{-1}\end{pmatrix}\begin{pmatrix}\mathbf{a} \\ \mathbf{b}\end{pmatrix} \tag{2.190}$$

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:

$$(m_u m_v)^{\tau+1} = \left(\frac{\sqrt{\det(\mathbf{A}\mathbf{B})\det(\mathbf{F}\mathbf{G})}}{\sigma^{2d} m_\alpha m_\beta}\right)^{\tau}\exp(-\frac{\tau}{2}(q_{u,\alpha} + q_{v,\beta})) \tag{2.191}$$

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:

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

with $\mu \stackrel{\text{def}}{=} \begin{pmatrix}\mathbf{A}^{-1}\mathbf{a}+\mathbf{u} \\ \mathbf{B}^{-1}\mathbf{b}+\mathbf{v}\end{pmatrix}$ and $\Gamma \stackrel{\text{def}}{=} \begin{pmatrix}\frac{\mathbf{F}}{\sigma^2} & -\frac{\mathrm{Id}}{\sigma^2} \\ -\frac{\mathrm{Id}}{\sigma^2} & \frac{\mathbf{G}}{\sigma^2}\end{pmatrix}$. Let's show that $\Gamma \succ 0$. Since $\frac{\mathbf{G}}{2\sigma^2} \succ 0$, it is sufficient to show that Schur complement $\frac{\mathbf{F}}{\sigma^2} - \frac{1}{\sigma^2}\mathbf{G}^{-1} \succ 0$. On one hand, with

$$\frac{\mathbf{F}-\mathbf{G}^{-1}}{\sigma^2} = \tau\widetilde{\mathbf{A}}^{-1} - \frac{1}{\lambda}\mathbf{G}^{-1}$$

On the other hand, almost by definition $\widetilde{\mathbf{A}} \prec \tau\lambda\,\mathrm{Id}$ and $\widetilde{\mathbf{B}} \prec \tau\lambda\,\mathrm{Id}$. Thus for any $x \in \mathbb{R}^d$:

$$x^\top \frac{\widetilde{\mathbf{A}}^{\frac{1}{2}}\widetilde{\mathbf{B}}\widetilde{\mathbf{A}}^{\frac{1}{2}}}{\tau} x \leq \lambda\|\widetilde{\mathbf{A}}^{\frac{1}{2}}x\|^2 = \lambda x^\top \widetilde{\mathbf{A}} x \leq \tau\lambda^2\|x\|^2,$$

which implies

$$\left(\frac{\widetilde{\mathbf{A}}^{\frac{1}{2}}\widetilde{\mathbf{B}}\widetilde{\mathbf{A}}^{\frac{1}{2}}}{\tau} + \frac{\sigma^4}{4}\,\mathrm{Id}\right)^{\frac{1}{2}} \prec \sqrt{\tau\lambda^2 + \frac{\sigma^4}{4}}\,\mathrm{Id} = \frac{\lambda}{2}(\sqrt{4\tau + (1-\tau)^2})\,\mathrm{Id} = \frac{\lambda(1+\tau)}{2}\,\mathrm{Id}\,.$$

Therefore, using the second equality of (2.186) and inverting (2.188) to obtain $\mathbf{G}^{-1}$:

$$\begin{aligned}
x^\top \mathbf{G}^{-1} x &= x^\top \widetilde{\mathbf{A}}^{-\frac{1}{2}} \left(\left(\frac{\widetilde{\mathbf{A}}^{\frac{1}{2}} \widetilde{\mathbf{B}} \widetilde{\mathbf{A}}^{\frac{1}{2}}}{\tau} + \frac{\sigma^4}{4} \mathrm{Id} \right)^{\frac{1}{2}} - \frac{\sigma^2}{2} \mathrm{Id}) \right) \widetilde{\mathbf{A}}^{-\frac{1}{2}} x \\
&= (\widetilde{\mathbf{A}}^{-\frac{1}{2}} x)^\top \left(\left(\frac{\widetilde{\mathbf{A}}^{\frac{1}{2}} \widetilde{\mathbf{B}} \widetilde{\mathbf{A}}^{\frac{1}{2}}}{\tau} + \frac{\sigma^4}{4} \mathrm{Id} \right)^{\frac{1}{2}} - \frac{\lambda(1-\tau)}{2} \mathrm{Id}) \right) (\widetilde{\mathbf{A}}^{-\frac{1}{2}} x) \\
&\leq (\widetilde{\mathbf{A}}^{-\frac{1}{2}} x)^\top \left(\frac{\lambda(1+\tau)}{2} \mathrm{Id} - \frac{\lambda(1-\tau)}{2} \mathrm{Id}) \right) (\widetilde{\mathbf{A}}^{-\frac{1}{2}} x) \\
&= \tau\lambda x^\top \widetilde{\mathbf{A}}^{-1} x.
\end{aligned}$$

Thus $\mathbf{G}^{-1} \prec \tau\lambda \widetilde{\mathbf{A}}^{-1}$. 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 $z \stackrel{\text{def}}{=} \binom{x}{y}$:

$$\begin{aligned}
\frac{\mathrm{d}\pi}{\mathrm{d}x\mathrm{d}y}(x,y) &= \omega \exp\left(\mathcal{Q}(\mu,\Gamma)(x,y)\right) \\
&= \omega \exp\left(-\frac{1}{2}(z^\top \Gamma z - 2 z^\top \mu) \right) \\
&= \omega \exp\left(\frac{1}{2}\mu^\top \Gamma^{-1} \mu - \frac{1}{2}(z - \Gamma^{-1}\mu)^\top \Gamma (z - \Gamma^{-1}\mu)) \right) \\
&= \omega e^{\frac{1}{2}\mu^\top \Gamma^{-1}\mu} \mathcal{N}(\mathbf{H}\mu, \mathbf{H})(z),
\end{aligned}$$

where $\mathbf{H} = \Gamma^{-1}$.

Detailed expressions. To conclude the proof of Theorem 2, we need to simplify the formulas of m, $\mathbf{H}\mu$ and $\mathbf{H}$. First, we will start with the mean $\mathbf{H}\mu$.

Hμ Using the optimality conditions of Proposition 25 and the closed form formula of $\mathbf{v}$ and $\mathbf{u}$:

$$\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 \operatorname{Id} & 0 \\ 0 & \mathbf{F}\mathbf{G} - \tau^2 \operatorname{Id} \end{pmatrix}^{-1}\begin{pmatrix} -\tau\mathbf{G} & \tau^2 \operatorname{Id} \\ \tau^2 \operatorname{Id} & -\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 \operatorname{Id} & 0 \\ 0 & \mathbf{F}\mathbf{G} - \tau^2 \operatorname{Id} \end{pmatrix}^{-1}\begin{pmatrix} \mathbf{G} & -\tau \operatorname{Id} \\ -\tau \operatorname{Id} & \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 \operatorname{Id})^{-1} \\ -\tau(\mathbf{F}\mathbf{G} - \tau^2 \operatorname{Id})^{-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 \operatorname{Id} \\ \tau \operatorname{Id} & \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} \operatorname{Id} & \tau\mathbf{G}^{-1} \\ \tau\mathbf{F}^{-1} & \operatorname{Id} \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{2.192}$$

Therefore:

$$\begin{aligned}
\mathbf{H}\mu &= \sigma^2\begin{pmatrix} \mathbf{F} & -\operatorname{Id} \\ -\operatorname{Id} & \mathbf{G} \end{pmatrix}^{-1}\begin{pmatrix} \operatorname{Id} & \tau\mathbf{G}^{-1} \operatorname{Id} \\ \tau\mathbf{F}^{-1} \operatorname{Id} & \operatorname{Id} \end{pmatrix}^{-1}\begin{pmatrix} \mathbf{A}^{-1} & 0 \\ 0 & \mathbf{B}^{-1} \end{pmatrix}\begin{pmatrix} \mathbf{a} \\ \mathbf{b} \end{pmatrix} \\
&= \sigma^2\left(\begin{pmatrix} \operatorname{Id} & \tau\mathbf{G}^{-1} \operatorname{Id} \\ \tau\mathbf{F}^{-1} \operatorname{Id} & \operatorname{Id} \end{pmatrix}\begin{pmatrix} \mathbf{F} & -\operatorname{Id} \\ -\operatorname{Id} & \mathbf{G} \end{pmatrix}\right)^{-1}\begin{pmatrix} \mathbf{A}^{-1} & 0 \\ 0 & \mathbf{B}^{-1} \end{pmatrix}\begin{pmatrix} \mathbf{a} \\ \mathbf{b} \end{pmatrix} \\
&= \sigma^2\begin{pmatrix} \mathbf{F} - \tau\mathbf{G}^{-1} & -(1-\tau) \operatorname{Id} \\ -(1-\tau) \operatorname{Id} & \mathbf{G} - \tau\mathbf{F}^{-1} \end{pmatrix}^{-1}\begin{pmatrix} \mathbf{A}^{-1} & 0 \\ 0 & \mathbf{B}^{-1} \end{pmatrix}\begin{pmatrix} \mathbf{a} \\ \mathbf{b} \end{pmatrix} \\
&= \sigma^2\begin{pmatrix} \sigma^2\mathbf{A}^{-1} + (1-\tau) \operatorname{Id} & -(1-\tau) \operatorname{Id} \\ -(1-\tau) \operatorname{Id} & \sigma^2\mathbf{B}^{-1} + (1-\tau) \operatorname{Id} \end{pmatrix}^{-1}\begin{pmatrix} \mathbf{A}^{-1} & 0 \\ 0 & \mathbf{B}^{-1} \end{pmatrix}\begin{pmatrix} \mathbf{a} \\ \mathbf{b} \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{A}^{-1} + \operatorname{Id} & -\lambda \operatorname{Id} \\ -\lambda \operatorname{Id} & \mathbf{B}^{-1} + \lambda \operatorname{Id} \end{pmatrix}^{-1}\begin{pmatrix} \mathbf{A}^{-1} & 0 \\ 0 & \mathbf{B}^{-1} \end{pmatrix}\begin{pmatrix} \mathbf{a} \\ \mathbf{b} \end{pmatrix}
\end{aligned} \tag{2.193}$$

Let's compute the inverse of:

$$\mathbf{Z} \overset{\text{def}}{=} \begin{pmatrix} \mathbf{A}^{-1} + \frac{1}{\lambda} \operatorname{Id} & -\frac{1}{\lambda} \operatorname{Id} \\ -\frac{1}{\lambda} \operatorname{Id} & \mathbf{B}^{-1} + \frac{1}{\lambda} \operatorname{Id} \end{pmatrix}. \tag{2.194}$$

Let $\mathbf{S}$ and $\mathbf{S}'$ be the respective Schur complements of $\mathbf{A}^{-1} + \frac{1}{\lambda}\operatorname{Id}$ and $\mathbf{B}^{-1} + \frac{1}{\lambda}\operatorname{Id}$ in $\mathbf{Z}$. The block inverse formula writes:

$$\mathbf{Z}^{-1} = \begin{pmatrix} \mathbf{S} & \frac{1}{\lambda}\mathbf{S}(\mathbf{B}^{-1} + \frac{1}{\lambda}\operatorname{Id})^{-1} \\ \frac{1}{\lambda}(\mathbf{A}^{-1} + \frac{1}{\lambda}\operatorname{Id})^{-1}\mathbf{S} & \mathbf{S}' \end{pmatrix}.$$

Using Woodbury's identity twice and denoting $\mathbf{X} \stackrel{\text{def}}{=} \mathbf{A} + \mathbf{B} + \lambda \operatorname{Id}$:

$$\begin{aligned} \mathbf{S} &= (\mathbf{A}^{-1} + \frac{1}{\lambda}\operatorname{Id} - \frac{1}{\lambda^2}(\mathbf{B}^{-1} + \frac{1}{\lambda}\operatorname{Id})^{-1})^{-1} \\ &= (\mathbf{A}^{-1} + (\mathbf{B} + \lambda \operatorname{Id})^{-1})^{-1} \\ &= (\mathbf{A} - \mathbf{A}(\mathbf{A} + \mathbf{B} + \lambda \operatorname{Id})^{-1}\mathbf{A}) \\ &= \mathbf{A} - \mathbf{A}\mathbf{X}^{-1}\mathbf{A}. \end{aligned}$$

And similarly: $\mathbf{S}' = \mathbf{B} - \mathbf{B}\mathbf{X}^{-1}\mathbf{B}$. The off-diagonal blocks can be simplified as well:

$$\begin{aligned} \frac{1}{\lambda}\mathbf{S}(\mathbf{B}^{-1} + \frac{1}{\lambda}\operatorname{Id})^{-1} &= \frac{1}{\lambda}(\mathbf{A}^{-1} + (\mathbf{B} + \lambda \operatorname{Id})^{-1})^{-1}(\mathbf{B}^{-1} + \frac{1}{\lambda}\operatorname{Id})^{-1} \\ &= (\mathbf{A}^{-1} + (\mathbf{B} + \lambda \operatorname{Id})^{-1})^{-1}(\lambda \operatorname{Id} + \mathbf{B}\operatorname{Id})^{-1}\mathbf{B} \\ &= \Big((\mathbf{B} + \lambda \operatorname{Id}) - (\mathbf{B} + \lambda \operatorname{Id})(\mathbf{A} + \mathbf{B} + \lambda \operatorname{Id})^{-1}(\mathbf{B} + \lambda \operatorname{Id})\Big)(\lambda \operatorname{Id} + \mathbf{B}\operatorname{Id})^{-1}\mathbf{B} \\ &= \mathbf{B} - (\mathbf{B} + \lambda \operatorname{Id})\mathbf{X}^{-1}\mathbf{B} \\ &= \mathbf{B} - (\mathbf{X} - \mathbf{A})\mathbf{X}^{-1}\mathbf{B} \\ &= \mathbf{A}\mathbf{X}^{-1}\mathbf{B}. \end{aligned}$$

Similarly, $\frac{1}{\lambda}(\mathbf{A}^{-1} + \frac{1}{\lambda}\operatorname{Id})^{-1}\mathbf{S} = \mathbf{B}\mathbf{X}^{-1}\mathbf{A}$. Thus, the inverse of $\mathbf{Z}$ is given by:

$$\mathbf{Z}^{-1} = \begin{pmatrix} \mathbf{A} - \mathbf{A}\mathbf{X}^{-1}\mathbf{A} & \mathbf{A}\mathbf{X}^{-1}\mathbf{B} \\ \mathbf{B}\mathbf{X}^{-1}\mathbf{A} & \mathbf{B} - \mathbf{B}\mathbf{X}^{-1}\mathbf{B} \end{pmatrix}. \tag{2.195}$$

and finally:

$$\begin{aligned} \mathbf{H}\mu = \mathbf{Z}^{-1}\begin{pmatrix} \mathbf{A}^{-1} & 0 \\ 0 & \mathbf{B}^{-1} \end{pmatrix}\begin{pmatrix} \mathbf{a} \\ \mathbf{b} \end{pmatrix} &= \begin{pmatrix} \operatorname{Id} - \mathbf{A}\mathbf{X}^{-1} & \mathbf{A}\mathbf{X}^{-1} \\ \mathbf{B}\mathbf{X}^{-1} & \operatorname{Id} - \mathbf{B}\mathbf{X}^{-1} \end{pmatrix}\begin{pmatrix} \mathbf{a} \\ \mathbf{b} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{a} + \mathbf{A}\mathbf{X}^{-1}(\mathbf{b} - \mathbf{a}) \\ \mathbf{b} + \mathbf{B}\mathbf{X}^{-1}(\mathbf{a} - \mathbf{b}) \end{pmatrix} \end{aligned}$$

Finding the covariance matrix H. To compute $\mathbf{H} = \left(\frac{1}{\sigma^2}\begin{pmatrix} \mathbf{F} & -\operatorname{Id} \\ -\operatorname{Id} & \mathbf{G} \end{pmatrix}\right)^{-1}$ 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:

$$\begin{aligned}\frac{1}{\sigma^2}\begin{pmatrix}\mathbf{F} & -\mathrm{Id}\\ -\mathrm{Id} & \mathbf{G}\end{pmatrix}\begin{pmatrix}\mathrm{Id} & \tau\mathbf{F}^{-1}\\ \tau\mathbf{G}^{-1} & \mathrm{Id}\end{pmatrix} &= \frac{1}{\sigma^2}\begin{pmatrix}\mathbf{F}-\tau\mathbf{G}^{-1} & -(1-\tau)\,\mathrm{Id}\\ -(1-\tau)\,\mathrm{Id} & \mathbf{G}-\tau\mathbf{F}^{-1}\end{pmatrix}\\ &= \begin{pmatrix}\mathbf{A}^{-1}+\frac{1}{\lambda}\mathrm{Id} & -\frac{1}{\lambda}\mathrm{Id}\\ -\frac{1}{\lambda}\mathrm{Id} & \mathbf{B}^{-1}+\frac{1}{\lambda}\mathrm{Id}\end{pmatrix},\end{aligned}$$

where the last equality follows from the optimality conditions (2.182). Therefore:

$$\mathbf{H} = \begin{pmatrix}\mathrm{Id} & \tau\mathbf{F}^{-1}\\ \tau\mathbf{G}^{-1} & \mathrm{Id}\end{pmatrix}\begin{pmatrix}\mathbf{A}^{-1}+\frac{1}{\lambda}\mathrm{Id} & -\frac{1}{\lambda}\mathrm{Id}\\ -\frac{1}{\lambda}\mathrm{Id} & \mathbf{B}^{-1}+\frac{1}{\lambda}\mathrm{Id}\end{pmatrix}^{-1}.$$

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}\mathrm{Id} & \tau\mathbf{F}^{-1}\\ \tau\mathbf{G}^{-1} & \mathrm{Id}\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}\mathrm{Id} & \tau\mathbf{C}\widetilde{\mathbf{B}}^{-1}\\ \mathbf{C}^\top\widetilde{\mathbf{A}}^{-1} & \mathrm{Id}\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}\mathrm{Id} & \mathbf{C}(\mathbf{B}^{-1}+\frac{1}{\lambda}\mathrm{Id})\\ \mathbf{C}^\top(\mathbf{A}^{-1}+\frac{1}{\lambda}\mathrm{Id}) & \mathrm{Id}\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}\mathrm{Id} & \mathbf{C}(\mathbf{B}^{-1}+\frac{1}{\lambda}\mathrm{Id})\\ \mathbf{C}^\top(\mathbf{A}^{-1}+\frac{1}{\lambda}\mathrm{Id}) & \mathrm{Id}\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}\mathrm{Id} & \frac{1}{\lambda}\mathbf{C}(\lambda\,\mathrm{Id}+\mathbf{B})\mathbf{B}^{-1}\\ \frac{1}{\lambda}\mathbf{C}^\top\mathbf{C}(\lambda\,\mathrm{Id}+\mathbf{A})\mathbf{A}^{-1} & \mathrm{Id}\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}\mathrm{Id} & \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} & \mathrm{Id}\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})(\mathrm{Id}-\mathbf{X}^{-1}\mathbf{B})\\ \frac{1}{\lambda}\mathbf{C}^\top(\mathbf{X}-\mathbf{B})(\mathrm{Id}-\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}(\mathrm{Id}+\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\,\mathrm{Id}+\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}(\mathrm{Id}+\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\,\mathrm{Id}+\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} & (\mathrm{Id}+\frac{1}{\lambda}\mathbf{C}^\top)(\mathbf{B}-\mathbf{B}\mathbf{X}^{-1}\mathbf{B})\end{pmatrix}\\
&= \begin{pmatrix}(\mathrm{Id}+\frac{1}{\lambda}\mathbf{C})(\mathbf{A}-\mathbf{A}\mathbf{X}^{-1}\mathbf{A}) & \mathbf{C}+(\mathrm{Id}+\frac{1}{\lambda}\mathbf{C})\mathbf{A}\mathbf{X}^{-1}\mathbf{B}\\ \mathbf{C}^\top+(\mathrm{Id}+\frac{1}{\lambda}\mathbf{C}^\top)\mathbf{B}\mathbf{X}^{-1}\mathbf{A} & (\mathrm{Id}+\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:

$$\frac{\mathrm{d}\pi}{\mathrm{d}x\mathrm{d}y}(x,y) = \omega e^{\frac{1}{2}\mu^\top \Gamma^{-1}\mu}\sqrt{\det(2\pi\mathbf{H})}\mathcal{N}(\mathbf{H}\mu,\mathbf{H})(z), \tag{2.196}$$

where

$$\begin{aligned}
\omega &= \frac{m_\alpha m_\beta}{\sqrt{\det(4\pi^2\mathbf{AB})}} m_u m_v e^{-\frac{1}{2}(\mathbf{a}^\top\mathbf{A}^{-1}\mathbf{a}+\mathbf{b}^\top\mathbf{B}^{-1}\mathbf{b})} \\
&= \frac{m_\alpha m_\beta}{\sqrt{\det(4\pi^2\mathbf{AB})}} \left(\frac{\sqrt{\det(\mathbf{AB})\det(\mathbf{FG})}}{\sigma^{2d} m_\alpha m_\beta}\right)^{\frac{\tau}{\tau+1}} e^{-\frac{\tau}{2(\tau+1)}(q_{u,\alpha}+q_{v,\beta})} e^{-\frac{1}{2}(\mathbf{a}^\top\mathbf{A}^{-1}\mathbf{a}+\mathbf{b}^\top\mathbf{B}^{-1}\mathbf{b})} \\
&= \frac{1}{(2\pi)^d}\left(\frac{m_\alpha m_\beta}{\sqrt{\det(\mathbf{AB})}}\right)^{\frac{1}{\tau+1}} \left(\frac{\sqrt{\det(\mathbf{FG})}}{\sigma^{2d}}\right)^{\frac{\tau}{\tau+1}} e^{-\frac{\tau}{2(\tau+1)}(q_{u,\alpha}+q_{v,\beta})} e^{-\frac{1}{2}(\mathbf{a}^\top\mathbf{A}^{-1}\mathbf{a}+\mathbf{b}^\top\mathbf{B}^{-1}\mathbf{b})}.
\end{aligned}$$

First, let's simplify the argument of the exponential terms. Isolating the terms that depend only on the input means $\mathbf{a}, \mathbf{b}$ it holds: $q_{u,\alpha} + q_{v,\beta} = \frac{\sigma^2}{\tau^2}(\mathbf{v}^\top\mathbf{F}\mathbf{v} + \mathbf{u}^\top\mathbf{G}\mathbf{u}) + \mathbf{a}^\top\mathbf{A}^{-1}\mathbf{a} + \mathbf{b}^\top\mathbf{B}^{-1}\mathbf{b}$. Therefore, the full exponential argument is given by:

$$\phi \overset{\text{def}}{=} \mu^\top\Gamma^{-1}\mu - \frac{\tau}{\tau+1}\frac{\sigma^2}{\tau^2}(\mathbf{v}^\top\mathbf{F}\mathbf{v} + \mathbf{u}^\top\mathbf{G}\mathbf{u}) - \frac{1}{\tau+1}(\mathbf{a}^\top\mathbf{A}^{-1}\mathbf{a} + \mathbf{b}^\top\mathbf{B}^{-1}\mathbf{b}) \tag{2.197}$$

On one hand, using Equation (2.193) we replace μ:

$$\begin{aligned}
\mu^\top\Gamma^{-1}\mu &= \mu^\top\mathbf{H}\mu \\
&= \sigma^2 \begin{pmatrix}\mathbf{A}^{-1}\mathbf{a}\\ \mathbf{B}^{-1}\mathbf{b}\end{pmatrix}^\top \begin{pmatrix}\mathrm{Id} & \tau\mathbf{F}^{-1}\\ \tau\mathbf{G}^{-1} & \mathrm{Id}\end{pmatrix}^{-1} \begin{pmatrix}\mathbf{F} & -\mathrm{Id}\\ -\mathrm{Id} & \mathbf{G}\end{pmatrix}^{-1} \begin{pmatrix}\mathrm{Id} & \tau\mathbf{G}^{-1}\\ \tau\mathbf{F}^{-1} & \mathrm{Id}\end{pmatrix}^{-1} \begin{pmatrix}\mathbf{A}^{-1}\mathbf{a}\\ \mathbf{B}^{-1}\mathbf{b}\end{pmatrix}
\end{aligned}$$

On the other hand:

$$\begin{aligned}
\frac{\sigma^2}{\tau^2}(\mathbf{v}^\top\mathbf{F}\mathbf{v} + \mathbf{u}^\top\mathbf{G}\mathbf{u}) &= \sigma^2((\mathbf{A}^{-1}\mathbf{a}+\mathbf{u})^\top\mathbf{F}^{-1}(\mathbf{A}^{-1}\mathbf{a}+\mathbf{u}) + (\mathbf{B}^{-1}\mathbf{b}+\mathbf{v})^\top\mathbf{G}^{-1}(\mathbf{B}^{-1}\mathbf{b}+\mathbf{v})) \\
&= \sigma^2\mu^\top \begin{pmatrix}\mathbf{F}^{-1} & 0\\ 0 & \mathbf{G}^{-1}\end{pmatrix}\mu \\
&= \sigma^2 \begin{pmatrix}\mathbf{A}^{-1}\mathbf{a}\\ \mathbf{B}^{-1}\mathbf{b}\end{pmatrix}^\top \begin{pmatrix}\mathrm{Id} & \tau\mathbf{F}^{-1}\\ \tau\mathbf{G}^{-1} & \mathrm{Id}\end{pmatrix}^{-1} \begin{pmatrix}\mathbf{F}^{-1} & 0\\ 0 & \mathbf{G}^{-1}\end{pmatrix} \begin{pmatrix}\mathrm{Id} & \tau\mathbf{G}^{-1}\\ \tau\mathbf{F}^{-1} & \mathrm{Id}\end{pmatrix}^{-1} \begin{pmatrix}\mathbf{A}^{-1}\mathbf{a}\\ \mathbf{B}^{-1}\mathbf{b}\end{pmatrix}
\end{aligned}$$

Let $\mathbf{J} = \begin{pmatrix} \mathrm{Id} & \tau\mathbf{G}^{-1} \\ \tau\mathbf{F}^{-1} & \mathrm{Id} \end{pmatrix}$ and $\mathbf{K} = \begin{pmatrix} \mathbf{F} & 0 \\ 0 & \mathbf{G} \end{pmatrix}$. It holds:

$$\mu^\top \Gamma^{-1}\mu - \frac{\tau}{\tau+1}\frac{\sigma^2}{\tau^2}(\mathbf{v}^\top\mathbf{F}\mathbf{v} + \mathbf{u}^\top\mathbf{G}\mathbf{u}) = \begin{pmatrix} \mathbf{A}^{-1}\mathbf{a} \\ \mathbf{B}^{-1}\mathbf{b} \end{pmatrix}^\top \mathbf{J}^{\top -1}(\mathbf{H} - \frac{\sigma^2\tau}{\tau+1}\mathbf{K}^{-1})\mathbf{J}^{-1}\begin{pmatrix} \mathbf{A}^{-1}\mathbf{a} \\ \mathbf{B}^{-1}\mathbf{b} \end{pmatrix}$$

Let's 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 $\mathbf{J}\mathbf{K} = \begin{pmatrix} \mathbf{F} & \tau\,\mathrm{Id} \\ \tau\,\mathrm{Id} & \mathbf{G} \end{pmatrix}$. 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}\mathrm{Id} \\ -\frac{1}{\sigma^2}\mathrm{Id} & -\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\,\mathrm{Id} \\ \tau\,\mathrm{Id} & \mathbf{G} \end{pmatrix}^{-1}(\mathbf{J}\mathbf{K}^\top)^\top\right) \\
&= \frac{\tau+1}{\tau\sigma^2}\left(-\begin{pmatrix} \mathbf{F} & \tau\,\mathrm{Id} \\ \tau\,\mathrm{Id} & \mathbf{G} \end{pmatrix}\begin{pmatrix} \mathrm{Id} & \tau\mathbf{F}^{-1} \\ \tau\mathbf{G}^{-1} & \mathrm{Id} \end{pmatrix} + (\tau+1)\begin{pmatrix} \mathbf{F} & \tau\,\mathrm{Id} \\ \tau\,\mathrm{Id} & \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))\,\mathrm{Id} \\ (-2\tau + \tau(\tau+1))\,\mathrm{Id} & -\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)\,\mathrm{Id} \\ -(1-\tau)\,\mathrm{Id} & \mathbf{G} - \tau\mathbf{F}^{-1} \end{pmatrix} \\
&= (\tau+1)\begin{pmatrix} \mathbf{A}^{-1} + \frac{1}{\lambda}\mathrm{Id} & -\frac{1}{\lambda}\mathrm{Id} \\ -\frac{1}{\lambda}\mathrm{Id} & \mathbf{B}^{-1} + \frac{1}{\lambda}\mathrm{Id} \end{pmatrix} \\
&= (\tau+1)\mathbf{Z}
\end{aligned}$$

Therefore:

$$\mu^\top \Gamma^{-1}\mu - \frac{\tau}{\tau+1}\frac{\sigma^2}{\tau^2}(\mathbf{v}^\top\mathbf{F}\mathbf{v} + \mathbf{u}^\top\mathbf{G}\mathbf{u}) = \frac{1}{\tau+1}\begin{pmatrix} \mathbf{A}^{-1}\mathbf{a} \\ \mathbf{B}^{-1}\mathbf{b} \end{pmatrix}^\top \mathbf{Z}^{-1}\begin{pmatrix} \mathbf{A}^{-1}\mathbf{a} \\ \mathbf{B}^{-1}\mathbf{b} \end{pmatrix} \tag{2.198}$$

The full exponential argument ϕ defined in Equation (2.197) 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 (2.196) leads to:

$$\begin{aligned}
m_\pi &\overset{\text{def}}{=} \pi(\mathbb{R}^d\times\mathbb{R}^d)\\
&= \sqrt{\det(\mathbf{H})}\left(\frac{m_\alpha m_\beta}{\sqrt{\det(\mathbf{A}\mathbf{B})}}\right)^{\frac{1}{\tau+1}}\left(\frac{\sqrt{\det(\mathbf{F}\mathbf{G})}}{\sigma^{2d}}\right)^{\frac{\tau}{\tau+1}} e^{-\frac{1}{2(\tau+1)}\left(\|\mathbf{a}-\mathbf{b}\|^2_{\mathbf{X}^{-1}}\right)}.
\end{aligned}$$

The determinants can be easily expressed as functions of $\mathbf{C}$. First notice that:

$$\det(\mathbf{H}) = \frac{1}{\det(\Gamma)} = \frac{\sigma^{4d}}{\det(\mathbf{F}\mathbf{G}-\mathrm{Id})},$$

and using the definition of $\mathbf{C}$, it holds that

$$\mathbf{F}\mathbf{G} = \widetilde{\mathbf{B}}\mathbf{C}^{-2}\widetilde{\mathbf{A}}.$$

Therefore, $\det(\mathbf{FG}) = \frac{\det(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}})}{\det(\mathbf{C})^2}$. Keeping in mind that the closed form expression of $\mathbf{C}$ given in (2.188) is applied to the pair $(\frac{1}{\tau}\widetilde{\mathbf{A}}, \widetilde{\mathbf{B}})$ in the unbalanced case, it holds: $\mathbf{C}^2 + \sigma^2\mathbf{C} = \frac{1}{\tau}\widetilde{\mathbf{A}}\widetilde{\mathbf{B}}$. Thus:

$$\begin{aligned}
\mathbf{FG} - \mathrm{Id} &= \widetilde{\mathbf{B}}\mathbf{C}^{-2}\widetilde{\mathbf{A}}(\mathrm{Id} - \widetilde{\mathbf{A}}^{-1}\mathbf{C}^2\widetilde{\mathbf{B}}^{-1}) \\
&= \widetilde{\mathbf{B}}\mathbf{C}^{-2}\widetilde{\mathbf{A}}(\mathrm{Id} - \widetilde{\mathbf{A}}^{-1}(\frac{1}{\tau}\widetilde{\mathbf{A}}\widetilde{\mathbf{B}} - \sigma^2\mathbf{C})\widetilde{\mathbf{B}}^{-1}) \\
&= \widetilde{\mathbf{B}}\mathbf{C}^{-2}\widetilde{\mathbf{A}}(\frac{(1-\tau)}{\tau}\,\mathrm{Id} + \sigma^2\widetilde{\mathbf{A}}^{-1}\mathbf{C}\widetilde{\mathbf{B}}^{-1}) \\
&= \sigma^2\widetilde{\mathbf{B}}\mathbf{C}^{-2}\widetilde{\mathbf{A}}(-\frac{2}{\gamma}\,\mathrm{Id} + \widetilde{\mathbf{A}}^{-1}\mathbf{C}\widetilde{\mathbf{B}}^{-1}) \\
&= \sigma^2\widetilde{\mathbf{B}}\mathbf{C}^{-2}(-\frac{2}{\gamma}\widetilde{\mathbf{A}}\widetilde{\mathbf{B}} + \mathbf{C})\widetilde{\mathbf{B}}^{-1},
\end{aligned}$$

therefore

$$\det(\mathbf{FG} - \mathrm{Id}) = \sigma^{2d}\frac{\det((-\frac{2}{\gamma}\widetilde{\mathbf{A}}\widetilde{\mathbf{B}} + \mathbf{C})}{\det(\mathbf{C})^2}.$$

Replacing the determinant formulas of $\mathbf{FG}$ and $\mathbf{FG} - \mathrm{Id}$ and re-arranging the common terms $\det(\mathbf{C})$ and σ leads to:

$$\begin{aligned}
\pi(\mathbb{R}^d \times \mathbb{R}^d) &= \frac{\left(m_\alpha m_\beta \sigma^{2d}\det(\mathbf{C})\sqrt{\frac{\det(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}})^\tau}{\det(\mathbf{AB})}}\right)^{\frac{1}{\tau+1}}}{\sqrt{\frac{\det(\mathbf{C}-\frac{2}{\gamma}\widetilde{\mathbf{A}}\widetilde{\mathbf{B}})}{\sigma^{2d}}}} e^{-\frac{1}{2(\tau+1)}(\|\mathbf{a}-\mathbf{b}\|^2_{\mathbf{X}^{-1}})} \\
&= \sigma^{d(\frac{2}{\tau+1}-1)}\frac{\left(m_\alpha m_\beta \det(\mathbf{C})\sqrt{\frac{\det(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}})^\tau}{\det(\mathbf{AB})}}\right)^{\frac{1}{\tau+1}}}{\sqrt{\det(\mathbf{C}-\frac{2}{\gamma}\widetilde{\mathbf{A}}\widetilde{\mathbf{B}})}} e^{-\frac{1}{2(\tau+1)}(\|\mathbf{a}-\mathbf{b}\|^2_{\mathbf{X}^{-1}})} \\
&= \sigma^{d\frac{1-\tau}{\tau+1}}\frac{\left(m_\alpha m_\beta \det(\mathbf{C})\sqrt{\frac{\det(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}})^\tau}{\det(\mathbf{AB})}}\right)^{\frac{1}{\tau+1}}}{\sqrt{\det(\mathbf{C}-\frac{2}{\gamma}\widetilde{\mathbf{A}}\widetilde{\mathbf{B}})}} e^{-\frac{1}{2(\tau+1)}(\|\mathbf{a}-\mathbf{b}\|^2_{\mathbf{X}^{-1}})} \\
&= \sigma^{\frac{d\sigma^2}{\sigma^2+\gamma}}\frac{\left(m_\alpha m_\beta \det(\mathbf{C})\sqrt{\frac{\det(\widetilde{\mathbf{A}}\widetilde{\mathbf{B}})^\tau}{\det(\mathbf{AB})}}\right)^{\frac{1}{\tau+1}}}{\sqrt{\det(\mathbf{C}-\frac{2}{\gamma}\widetilde{\mathbf{A}}\widetilde{\mathbf{B}})}} e^{-\frac{1}{2(\tau+1)}(\|\mathbf{a}-\mathbf{b}\|^2_{\mathbf{X}^{-1}})}
\end{aligned} \tag{2.199}$$

Deriving a closed form for $\mathrm{UOT}^{\otimes}_{2\sigma^2}$. Using Equation (2.199), a direct application of Proposition 13 yields

$$\mathrm{UOT}^{\otimes}_{2\sigma^2}(\alpha,\beta) = \gamma(m_\alpha + m_\beta) + 2\sigma^2(m_\alpha m_\beta) - 2(\sigma^2+2\gamma)m_{\pi^\star}. \tag{2.200}$$

This ends the proof of Theorem 2.

We end this first section with the technical details of the proof of Proposition 15 restated below.

Proposition 26 *Let $\alpha = m_\alpha \mathcal{N}(\mathbf{a},\mathbf{A})$ and $\beta = m_\beta \mathcal{N}(\mathbf{b},\mathbf{B})$. If $m_\alpha \neq m_\beta$, $\mathrm{UOT}_{\gamma,2\sigma^2}(\alpha,\beta)$ goes to $+\infty$ as $\gamma \to +\infty$. Moreover, we can obtain the following equivalent:*

$$\lim_{\gamma\to+\infty}\left[\underset{\gamma,2\sigma^2}{\mathrm{UOT}}(\alpha,\beta) - \gamma(\sqrt{m_\alpha}-\sqrt{m_\beta})^2\right] = \sqrt{m_\alpha m_\beta}\left[\mathrm{OT}^{\otimes}_{2\sigma^2}\left(\frac{\alpha}{m_\alpha},\frac{\beta}{m_\beta}\right) + 2\sigma^2\,\mathrm{KL}(1|\sqrt{m_\alpha m_\beta})\right] \tag{2.201}$$

where $\mathrm{KL}(1|\sqrt{m_\alpha m_\beta}) = \sqrt{m_\alpha m_\beta} - 1 - \log(\sqrt{m_\alpha m_\beta})$.

In particular, if $m_\alpha = m_\beta = m > 0$, then:

$$\lim_{\gamma\to+\infty}\underset{\gamma,2\sigma^2}{\mathrm{UOT}}(\alpha,\beta) = m\left[\mathrm{OT}^{\otimes}_{2\sigma^2}\left(\frac{\alpha}{m},\frac{\beta}{m}\right) + 2\sigma^2\,\mathrm{KL}(1|m)\right] \tag{2.202}$$

PROOF. Using proposition 13, the following holds:

$$\underset{\gamma,2\sigma^2}{\mathrm{UOT}}(\alpha,\beta) - \gamma(\sqrt{m_\alpha}-\sqrt{m_\beta})^2 = 2\sigma^2(m_\alpha m_\beta - m_\pi) + 2\gamma(\sqrt{m_\alpha m_\beta} - m_\pi) \tag{2.203}$$

Computing the limit of m_π as $\gamma \to +\infty$ is straightforward. When $\gamma \to +\infty$, eventually using Woodburry's identity:

$$\tau \to 1 \tag{2.204}$$

$$\frac{1}{\lambda} \to 0 \tag{2.205}$$

$$\widetilde{\mathbf{A}} = \tau\left(\mathbf{A}^{-1} + \frac{1}{\lambda}\mathrm{Id}\right)^{-1} \to \mathbf{A} \tag{2.206}$$

$$\widetilde{\mathbf{B}} = \tau\left(\mathbf{B}^{-1} + \frac{1}{\lambda}\mathrm{Id}\right)^{-1} \to \mathbf{B} \tag{2.207}$$

$$\mathbf{X}^{-1} \to 0\ . \tag{2.208}$$

Therefore, $m_\pi \to \sqrt{m_\alpha m_\beta}$. And it holds:

$$\lim_{\gamma\to+\infty}\left[\underset{\gamma,2\sigma^2}{\mathrm{UOT}}(\alpha,\beta) - \gamma(\sqrt{m_\alpha}-\sqrt{m_\beta})^2\right] = 2\sigma^2\sqrt{m_\alpha m_\beta}(\sqrt{m_\alpha m_\beta}-1) \tag{2.209}$$

$$+ \lim_{\gamma\to+\infty} 2\gamma(\sqrt{m_\alpha m_\beta} - m_\pi(\gamma)) \tag{2.210}$$

The remaining limit to compute is that of $2\gamma(\sqrt{m_\alpha m_\beta} - m_\pi(\gamma))$ which is a bit more technical to compute. The main idea is to use the change of variable $\omega \overset{\text{def}}{=} \frac{2}{\gamma}$.

$$\lim_{\gamma \to +\infty} 2\gamma(\sqrt{m_\alpha m_\beta} - m_\pi(\gamma)) = \lim_{\omega \to 0} \frac{4}{\omega}(\sqrt{m_\alpha m_\beta} - m_\pi(\omega)) \tag{2.211}$$

$$= -4\frac{\mathrm{d}m_\pi}{\mathrm{d}\omega}(0) \ . \tag{2.212}$$

First, let's write all elements of m_π as a function of ω:

$$\tau = \frac{\gamma}{\gamma + 2\sigma^2} = \frac{1}{1 + \omega\sigma^2} \tag{2.213}$$

$$\frac{1}{\tau + 1} = \frac{\gamma + 2\sigma^2}{2\gamma + 2\sigma^2} = 1 - \frac{1}{2 + \omega\sigma^2} \tag{2.214}$$

$$\frac{1}{\lambda} = \frac{2}{\gamma + 2\sigma^2} = \frac{\omega}{1 + \omega\sigma^2} \tag{2.215}$$

$$\widetilde{\mathbf{A}}(\omega) = \tau\left(\mathbf{A}^{-1} + \frac{1}{\lambda}\,\mathrm{Id}\right)^{-1} = \left((1 + \omega\sigma^2)\mathbf{A}^{-1} + \omega\,\mathrm{Id}\right)^{-1} \tag{2.216}$$

$$\widetilde{\mathbf{B}}(\omega) = \tau\left(\mathbf{B}^{-1} + \frac{1}{\lambda}\,\mathrm{Id}\right)^{-1} = \left((1 + \omega\sigma^2)\mathbf{B}^{-1} + \omega\,\mathrm{Id}\right)^{-1} \tag{2.217}$$

$$\frac{\mathbf{X}^{-1}}{\tau + 1}(\omega) = (1 - \frac{1}{2 + \sigma^2\omega})(\mathbf{A} + \mathbf{B} + \frac{1}{\omega}\,\mathrm{Id})^{-1} = \omega(1 - \frac{1}{2 + \sigma^2\omega})(\omega\mathbf{A} + \omega\mathbf{B} + (\omega\sigma^2 + 1)\,\mathrm{Id})^{-1} \tag{2.218}$$

$$\mathbf{C}(\omega) = \left((1 + \omega\sigma^2)\widetilde{\mathbf{A}}\widetilde{\mathbf{B}} + \frac{\sigma^4}{4}\,\mathrm{Id}\right)^{\frac{1}{2}} - \frac{\sigma^2}{2}\,\mathrm{Id} \tag{2.219}$$

To differentiate $\omega \mapsto m_\pi(\omega)$ at 0, we re-write it as follows:

$$m_\pi(\omega) = \exp\left[f(\omega) + g(\omega) + h(\omega)\right] \ , \tag{2.220}$$

where

$$f(\omega) = \frac{d\sigma^2\omega}{2 + \sigma^2\omega}\log(\sigma) \tag{2.221}$$

$$g(\omega) = g_0(\omega)\left[G + g_1(\omega) + g_2(\omega)\right] + g_3(\omega) \tag{2.222}$$

$$h(\omega) = -\frac{1}{2}(\mathbf{a} - \mathbf{b})^\top(\frac{\mathbf{X}^{-1}}{\tau + 1}(\omega))(\mathbf{a} - \mathbf{b}) \ , \tag{2.223}$$

where

$$g_0(\omega) = 1 - \frac{1}{2 + \omega\sigma^2} \tag{2.224}$$

$$G = \log(m_\alpha m_\beta) - \frac{1}{2}\log\det(\mathbf{AB}) \tag{2.225}$$

$$g_1(\omega) = \log\det(\mathbf{C}(\omega)) \tag{2.226}$$

$$g_2(\omega) = \frac{1}{2(1+\sigma^2\omega)}\log\det(\widetilde{\mathbf{A}}(\omega)\widetilde{\mathbf{B}}(\omega)) \tag{2.227}$$

$$g_3(\omega) = -\frac{1}{2}\log\det(\mathbf{C}(\omega) - \omega\widetilde{\mathbf{A}}(\omega)\widetilde{\mathbf{B}}(\omega)) \ . \tag{2.228}$$

With these notations, it holds:

$$\frac{\mathrm{d}m_\pi}{\mathrm{d}\omega}(0) = (\frac{\mathrm{d}f}{\mathrm{d}\omega}(0) + \frac{\mathrm{d}g}{\mathrm{d}\omega}(0) + \frac{\mathrm{d}h}{\mathrm{d}\omega}(0))\exp(f(0)+g(0)+h(0)) \tag{2.229}$$

$$= (\frac{\mathrm{d}f}{\mathrm{d}\omega}(0) + \frac{\mathrm{d}g}{\mathrm{d}\omega}(0) + \frac{\mathrm{d}h}{\mathrm{d}\omega}(0))m_\pi(0) \tag{2.230}$$

$$= (\frac{\mathrm{d}f}{\mathrm{d}\omega}(0) + \frac{\mathrm{d}g}{\mathrm{d}\omega}(0) + \frac{\mathrm{d}h}{\mathrm{d}\omega}(0))\sqrt{m_\alpha m_\beta} \ , \tag{2.231}$$

where the derivative of g can be detailed further:

$$\frac{\mathrm{d}g}{\mathrm{d}\omega}(0) = \frac{\mathrm{d}g_0}{\mathrm{d}\omega}(G + g_1(0) + g_2(0)) + g_0(0)\left(\frac{\mathrm{d}g_1}{\mathrm{d}\omega} + \frac{\mathrm{d}g_2}{\mathrm{d}\omega}\right)(0) + \frac{\mathrm{d}g_3}{\mathrm{d}\omega}(0) \tag{2.232}$$

$$= \frac{\sigma^2}{4}\left(\log(m_\alpha m_\beta) + \log\det(\mathbf{C}(0))\right) + \frac{1}{2}\left(\frac{\mathrm{d}g_1}{\mathrm{d}\omega} + \frac{\mathrm{d}g_2}{\mathrm{d}\omega}\right)(0) + \frac{\mathrm{d}g_3}{\mathrm{d}\omega}(0) \tag{2.233}$$

We evaluate the derivative of each component.

Computing $\frac{\mathrm{d}f}{\mathrm{d}\omega}(0)$ The function f is defined as:

$$f(\omega) = \frac{d\sigma^2\omega}{2+\sigma^2\omega}\log(\sigma) \tag{2.234}$$

$$= d\log(\sigma)\left(1 - \frac{2}{2+\sigma^2\omega}\right) \tag{2.235}$$

Its derivative is given by:

$$\frac{\mathrm{d}f}{\mathrm{d}\omega}(\omega) = d\log(\sigma)\frac{2\sigma^2}{(2+\sigma^2\omega)^2} \tag{2.236}$$

thus $\frac{\mathrm{d}f}{\mathrm{d}\omega}(0) = \frac{1}{2}d\sigma^2\log(\sigma)$.

Computing $\frac{\mathrm{d}h}{\mathrm{d}\omega}(0)$ The function h is defined as:

$$h(\omega) = -\frac{1}{2}(\mathbf{a}-\mathbf{b})^\top(\frac{\mathbf{X}^{-1}}{\tau+1}(\omega))(\mathbf{a}-\mathbf{b}) \tag{2.237}$$

$$= -\frac{1}{2}\omega\psi(\omega) \tag{2.238}$$

where

$$\psi(\omega) \stackrel{\text{def}}{=} (1-\frac{1}{2+\sigma^2\omega})(\mathbf{a}-\mathbf{b})^\top(\omega\mathbf{A}+\omega\mathbf{B}+(\omega\sigma^2+1)\,\mathrm{Id})^{-1}(\mathbf{a}-\mathbf{b}) \ . \tag{2.239}$$

As a derivative of a product, it holds:

$$\frac{\mathrm{d}h}{\mathrm{d}\omega}(0) = -\frac{1}{2}\psi(0) = -\frac{1}{4}\|\mathbf{a}-\mathbf{b}\|^2 \ . \tag{2.240}$$

Derivatives of $\widetilde{\mathbf{A}}(\omega)\widetilde{\mathbf{B}}(\omega)$ and $\mathbf{C}(\omega)$ Differentiation through the elements of g requires computing derivatives of $\omega \mapsto \widetilde{\mathbf{A}}(\omega)\widetilde{\mathbf{B}}(\omega)$ and $\omega \mapsto \mathbf{C}(\omega)$. For the sake of clarity, we introduce the following notations.

The matrix inverse and square root operations are denoted respectively by $\mathrm{inv} : \mathbf{A} \mapsto \mathbf{A}^{-1}$ and $\mathcal{R} : \mathbf{A} \mapsto \mathbf{A}^{\frac{1}{2}}$. First, we remind the reader of the differentials of these applications along with det and logdet. The differential operator of a function F at $\mathbf{A}$ is a linear operator denoted by $\mathcal{J}_F(\mathbf{A})$.

$$\mathcal{J}_{\mathrm{inv}}(\mathbf{A})(\mathbf{H}) = -\mathbf{A}^{-1}\mathbf{H}\mathbf{A}^{-1} \tag{2.241}$$

$$\mathcal{J}_{\det}(\mathbf{A})(\mathbf{H}) = \mathrm{Tr}(\mathbf{A},\mathbf{H}) \tag{2.242}$$

$$\mathcal{J}_{\log\det}(\mathbf{A})(\mathbf{H}) = \mathrm{Tr}(\mathbf{A}^{-1},\mathbf{H}) \tag{2.243}$$

$$\mathcal{J}_{\mathcal{R}}(\mathbf{A})(\mathbf{H}) \text{ is the only positive definite solution } \mathbf{Z} \text{ of } \Leftrightarrow \mathbf{A}^{\frac{1}{2}}\mathbf{Z}+\mathbf{Z}\mathbf{A}^{\frac{1}{2}} = \mathbf{H} \tag{2.244}$$

Let's compute the derivative of $\mathcal{V} : \omega \mapsto \widetilde{\mathbf{A}}(\omega)\widetilde{\mathbf{B}}(\omega)$ at 0. First, we can simplify that expression by writing:

$$\widetilde{\mathbf{A}}(\omega)\widetilde{\mathbf{B}}(\omega) = \left[\left((1+\omega\sigma^2)\mathbf{B}^{-1}+\omega\,\mathrm{Id}\right)\left((1+\omega\sigma^2)\mathbf{A}^{-1}+\omega\,\mathrm{Id}\right)\right]^{-1} \tag{2.245}$$

$$= \left[(1+\omega\sigma^2)^2\mathbf{B}^{-1}\mathbf{A}^{-1}+\omega(1+\omega\sigma^2)(\mathbf{A}^{-1}+\mathbf{B}^{-1})+\omega^2\,\mathrm{Id}\right]^{-1} \tag{2.246}$$

$$\stackrel{\text{def}}{=} \mathbf{M}(\omega)^{-1} \tag{2.247}$$

Applying the chain rule, the derivative of $\mathcal{V}$ for some direction $h > 0$:

$$\mathcal{J}_{\mathcal{V}}(\omega)(h) = \mathcal{J}_{\text{inv}}(\mathbf{M}(\omega))\left(\mathcal{J}_{\mathbf{M}}(\omega)(h)\right) \tag{2.248}$$
$$= -\mathbf{M}(\omega)^{-1}\mathcal{J}_{\mathbf{M}}(\omega)(h)\mathbf{M}(\omega)^{-1} \tag{2.249}$$
$$\tag{2.250}$$

where

$$\mathcal{J}_{\mathbf{M}}(\omega)(h) = h\left(2\sigma^2(1+\sigma^2\omega\,\mathrm{Id})\mathbf{B}^{-1}\mathbf{A}^{-1} + (1+2\sigma^2\omega)(\mathbf{A}^{-1}+\mathbf{B}^{-1}) + 2\omega\,\mathrm{Id}\right) \tag{2.251}$$

Evaluating at 0 leads to:

$$\mathcal{J}_{\mathcal{V}}(0)(h) = -h\mathbf{A}\left(2\sigma^2\,\mathrm{Id} + \mathbf{A} + \mathbf{B}\right)\mathbf{B}\ . \tag{2.252}$$

We can now establish the derivative of $\omega \mapsto \mathbf{C}(\omega)$ at 0 as well. First re-rewrite the definition of $\mathbf{C}$:

$$\mathbf{C}(\omega) = \left((1+\omega\sigma^2)\widetilde{\mathbf{A}}(\omega)\widetilde{\mathbf{B}}(\omega) + \frac{\sigma^4}{4}\,\mathrm{Id}\right)^{\frac{1}{2}} - \frac{\sigma^2}{2}\,\mathrm{Id} \tag{2.253}$$
$$= \left((1+\omega\sigma^2)\mathcal{V}(\omega) + \frac{\sigma^4}{4}\,\mathrm{Id}\right)^{\frac{1}{2}} - \frac{\sigma^2}{2}\,\mathrm{Id} \tag{2.254}$$
$$\stackrel{\text{def}}{=} \mathbf{S}(\omega)^{\frac{1}{2}} - \frac{\sigma^2}{2}\,\mathrm{Id} \tag{2.255}$$

Thus, by the chain rule:

$$\mathcal{J}_{\mathbf{C}}(\omega)(h) = \mathcal{J}_{\mathcal{R}}(\mathbf{S}(\omega))(\mathcal{J}_{\mathbf{S}}(\omega)(h)) \tag{2.256}$$
$$= \mathcal{J}_{\mathcal{R}}(\mathbf{S}(\omega))\left(h\sigma^2\mathcal{V}(\omega) + (1+\sigma^2\omega)\mathcal{J}_{\mathcal{V}}(\omega)(h)\right) \tag{2.257}$$
$$\tag{2.258}$$

Substituting at 0:

$$\mathcal{J}_{\mathbf{C}}(0)(h) = \mathcal{J}_{\mathcal{R}}(\mathbf{S}(0))\left(h\sigma^2\mathbf{AB} - h\mathbf{A}\left(2\sigma^2\,\mathrm{Id} + \mathbf{A} + \mathbf{B}\right)\mathbf{B}\right) \tag{2.259}$$
$$= \mathcal{J}_{\mathcal{R}}(\mathbf{S}(0))\left(-h\sigma^2\mathbf{AB} - h\mathbf{A}\left(\mathbf{A}+\mathbf{B}\right)\mathbf{B}\right) \tag{2.260}$$

Thus $\mathcal{J}_{\mathbf{C}}(0)(h)$ is the only positive definite solution $\mathbf{Z}$ of:

$$\mathbf{S}(0)^{\frac{1}{2}}\mathbf{Z} + \mathbf{Z}\mathbf{S}(0)^{\frac{1}{2}} = -h\left(\sigma^2\mathbf{AB} + \mathbf{A}(\mathbf{A}+\mathbf{B})\mathbf{B}\right) \tag{2.261}$$
$$\Leftrightarrow (\mathbf{AB} + \frac{\sigma^4}{4}\,\mathrm{Id})^{\frac{1}{2}}\mathbf{Z} + \mathbf{Z}(\mathbf{AB} + \frac{\sigma^4}{4}\,\mathrm{Id})^{\frac{1}{2}} = -h\left(\sigma^2\mathbf{AB} + \mathbf{A}(\mathbf{A}+\mathbf{B})\mathbf{B}\right) \tag{2.262}$$

Computing $\frac{\mathrm{d}g_1}{\mathrm{d}\omega}(0)$ The function g_1 is defined as: $g_1 : \omega \mapsto \log\det(\mathbf{C}(\omega))$. The chain rule dictates:

$$\frac{\mathrm{d}g_1}{\mathrm{d}\omega}(0) = \mathcal{J}_{\log\det}(\mathbf{C}(0))(\mathcal{J}_{\mathbf{C}}(0)(h)) \tag{2.263}$$
$$= \mathrm{Tr}(\mathbf{C}(0)^{-1}\mathcal{J}_{\mathbf{C}}(0)(h)) \tag{2.264}$$
$$= \mathrm{Tr}(\mathbf{C}(0)^{-1}\mathbf{Z}) \tag{2.265}$$

Computing $\frac{\mathrm{d}g_2}{\mathrm{d}\omega}(0)$ The function g_2 can be written:

$$g_2(\omega) = \frac{1}{2(1+\sigma^2\omega)}\log\det(\widetilde{\mathbf{A}}(\omega)\widetilde{\mathbf{B}}(\omega)) \tag{2.266}$$
$$= \frac{1}{2(1+\sigma^2\omega)}\log\det(\mathcal{V}(\omega)) \tag{2.267}$$

Thus:

$$\frac{\mathrm{d}g_2}{\mathrm{d}\omega}(0)(h) = -\frac{\sigma^2 h}{2}\log\det(\mathcal{V}(0)) + \frac{1}{2}\mathcal{J}_{\log\det}(\mathbf{V}(0))\left(\mathcal{J}_{\mathcal{V}}(0)(h)\right) \tag{2.268}$$
$$= -\frac{\sigma^2 h}{2}\log\det(\mathcal{V}(0)) + \frac{1}{2}\mathrm{Tr}\left(\mathbf{V}(0)^{-1}\left(\mathcal{J}_{\mathcal{V}}(0)(h)\right)\right) \tag{2.269}$$
$$= -\frac{\sigma^2 h}{2}\log\det(\mathbf{AB}) + \frac{1}{2}\mathrm{Tr}\left(\mathbf{B}^{-1}\mathbf{A}^{-1}\left(-h\mathbf{A}\left(2\sigma^2\,\mathrm{Id} + \mathbf{A} + \mathbf{B}\right)\mathbf{B}\right)\right) \tag{2.270}$$
$$= -\frac{\sigma^2 h}{2}\log\det(\mathbf{AB}) - \frac{h}{2}\mathrm{Tr}\left(\mathbf{A}^{-1}\left(\mathbf{A}\left(2\sigma^2\,\mathrm{Id} + \mathbf{A} + \mathbf{B}\right)\mathbf{B}\right)\mathbf{B}^{-1}\right) \tag{2.271}$$
$$= -\frac{\sigma^2 h}{2}\log\det(\mathbf{AB}) - \frac{h}{2}\mathrm{Tr}\left(2\sigma^2\,\mathrm{Id} + \mathbf{A} + \mathbf{B}\right) \tag{2.272}$$

Computing $\frac{\mathrm{d}g_3}{\mathrm{d}\omega}(0)$ The function g_3 can be written:

$$g_3(\omega) = -\frac{1}{2}\log\det(\mathbf{C}(\omega) - \omega\widetilde{\mathbf{A}}(\omega)\widetilde{\mathbf{B}}(\omega)) \tag{2.273}$$
$$= -\frac{1}{2}\log\det(\mathbf{C}(\omega) - \omega\mathcal{V}(\omega)) \ . \tag{2.274}$$

Since the derivative of $\omega\mathcal{V}(\omega)$ at 0 is given by $\mathcal{V}(0)$, the chain rule dictates:

$$\frac{\mathrm{d}g_3}{\mathrm{d}\omega}(0)(h) = -\frac{1}{2}\mathcal{J}_{\log\det}(\mathbf{C}(0))\left(\mathcal{J}_{\mathbf{C}}(0)(h) - h\mathcal{V}(0)\right) \tag{2.275}$$

$$= -\frac{1}{2}\mathrm{Tr}\left(\mathbf{C}(0)^{-1}(\mathcal{J}_{\mathbf{C})}(0)(h) - h\mathcal{V}(0))\right) \tag{2.276}$$

$$= -\frac{1}{2}\mathrm{Tr}\left(\mathbf{C}(0)^{-1}(\mathbf{Z} - h\mathbf{AB})\right) \tag{2.277}$$

Computing $\frac{\mathrm{d}g}{\mathrm{d}\omega}(0)$ We showed earlier that:

$$\frac{\mathrm{d}g}{\mathrm{d}\omega}(0) = \frac{\sigma^2}{4}\left(\log(m_\alpha m_\beta) + \log\det(\mathbf{C}(0))\right) + \frac{1}{2}\left(\frac{\mathrm{d}g_1}{\mathrm{d}\omega} + \frac{\mathrm{d}g_2}{\mathrm{d}\omega}\right)(0) + \frac{\mathrm{d}g_3}{\mathrm{d}\omega}(0) \ . \tag{2.278}$$

We can detail the derivatives above:

$$\frac{1}{2}\left(\frac{\mathrm{d}g_1}{\mathrm{d}\omega} + \frac{\mathrm{d}g_2}{\mathrm{d}\omega}\right)(0)(h) + \frac{\mathrm{d}g_3}{\mathrm{d}\omega}(0)(h) = \frac{1}{2}\left(\mathrm{Tr}(\mathbf{C}(0)^{-1}\mathbf{Z})\right. \tag{2.279}$$

$$\left. -\frac{h\sigma^2}{2}\log\det(\mathbf{AB}) - \frac{h}{2}\mathrm{Tr}\left(2\sigma^2\,\mathrm{Id} + \mathbf{A} + \mathbf{B}\right)\right) - \frac{1}{2}\mathrm{Tr}\left(\mathbf{C}(0)^{-1}(\mathbf{Z} - h\mathbf{AB})\right) \ . \tag{2.280}$$

The terms depending on $\mathbf{Z}$ cancel out and it holds:

$$\frac{1}{2}\left(\frac{\mathrm{d}g_1}{\mathrm{d}\omega} + \frac{\mathrm{d}g_2}{\mathrm{d}\omega}\right)(0) + \frac{\mathrm{d}g_3}{\mathrm{d}\omega}(0) = \frac{1}{4}\left(\sigma^2\log\det(\mathbf{AB}) - \mathrm{Tr}\left(2\sigma^2\,\mathrm{Id} + \mathbf{A} + \mathbf{B} - 2\mathbf{C}(0)^{-1}\mathbf{AB}\right)\right) \ , \tag{2.281}$$

where, $\mathbf{C}(0) = \left(\mathbf{AB} + \frac{\sigma^4}{4}\,\mathrm{Id}\right)^{\frac{1}{2}} - \frac{\sigma^2}{2}\,\mathrm{Id}$ corresponds to the solution $\mathbf{C}_0$ of the fixed point equation shown in balanced OT:

$$\mathbf{C}_0^2 + \sigma^2\mathbf{C}_0 = \mathbf{AB} \Rightarrow \mathbf{C}_0^{-1}(\mathbf{AB}) = \mathbf{C}_0 + \sigma^2\,\mathrm{Id} \ . \tag{2.282}$$

Thus:

$$\frac{1}{2}\left(\frac{\mathrm{d}g_1}{\mathrm{d}\omega} + \frac{\mathrm{d}g_2}{\mathrm{d}\omega}\right)(0) + \frac{\mathrm{d}g_3}{\mathrm{d}\omega}(0) = -\frac{1}{4}\left(\sigma^2\log\det(\mathbf{AB}) + \mathrm{Tr}\left(\mathbf{A} + \mathbf{B} - 2\mathbf{C}_0\right)\right) \ . \tag{2.283}$$

Finally, summing up the constituents of g:

$$\frac{\mathrm{d}g}{\mathrm{d}\omega}(0) = \frac{\sigma^2}{4}\left(\log(m_\alpha m_\beta) + \log\det(\mathbf{C}_0)\right) - \frac{1}{4}\left(\sigma^2\log\det(\mathbf{AB}) + \mathrm{Tr}\left(\mathbf{A}+\mathbf{B}-2\mathbf{C}_0\right)\right) \tag{2.284}$$

$$= \frac{\sigma^2}{4}\left(\log(m_\alpha m_\beta) + \log\det(\mathbf{C}_0(\mathbf{AB})^{-1})\right) - \frac{1}{4}\mathrm{Tr}\left(\mathbf{A}+\mathbf{B}-2\mathbf{C}_0\right) \tag{2.285}$$

$$= \frac{\sigma^2}{4}\left(\log(m_\alpha m_\beta) + \log\det((\mathbf{C}_0+\sigma^2\,\mathrm{Id})^{-1})\right) - \frac{1}{4}\mathrm{Tr}\left(\mathbf{A}+\mathbf{B}-2\mathbf{C}_0\right) \tag{2.286}$$

$$= \frac{\sigma^2}{4}\left(\log(m_\alpha m_\beta) - \log\det(\mathbf{C}_0+\sigma^2\,\mathrm{Id})\right) - \frac{1}{4}\mathrm{Tr}\left(\mathbf{A}+\mathbf{B}-2\mathbf{C}_0\right) \tag{2.287}$$

$$= \frac{\sigma^2}{4}\left(\log(m_\alpha m_\beta) - 2d\log(\sigma) - \log\det(\frac{\mathbf{C}_0}{\sigma^2}+\mathrm{Id})\right) - \frac{1}{4}\mathrm{Tr}\left(\mathbf{A}+\mathbf{B}-2\mathbf{C}_0\right) \tag{2.288}$$

$$= -\frac{1}{4}\mathfrak{B}^{\otimes}_{\sigma^2}(\mathbf{A},\mathbf{B}) + \frac{\sigma^2}{4}\log(m_\alpha m_\beta) - \frac{1}{2}d\sigma^2\log(\sigma)\ . \tag{2.289}$$

Summing up everything

$$\frac{\mathrm{d}m_\pi}{\mathrm{d}\omega}(0) = \sqrt{m_\alpha m_\beta}(\frac{\mathrm{d}f}{\mathrm{d}\omega}(0) + \frac{\mathrm{d}g}{\mathrm{d}\omega}(0) + \frac{\mathrm{d}h}{\mathrm{d}\omega}(0)) \tag{2.290}$$

$$= \sqrt{m_\alpha m_\beta}\left(\frac{1}{2}d\sigma^2\log(\sigma) - \frac{1}{4}\|\mathbf{a}-\mathbf{b}\|^2 + \frac{\mathrm{d}g}{\mathrm{d}\omega}(0)\right) \tag{2.291}$$

$$= \sqrt{m_\alpha m_\beta}\left(-\frac{1}{4}\|\mathbf{a}-\mathbf{b}\|^2 - \frac{1}{4}\mathfrak{B}^{\otimes}_{\sigma^2}(\mathbf{A},\mathbf{B}) + \frac{\sigma^2}{4}\log(m_\alpha m_\beta)\right) \tag{2.292}$$

$$= \sqrt{m_\alpha m_\beta}\left(-\frac{1}{4}\,\mathrm{OT}^{\otimes}_{2\sigma^2}(\frac{\alpha}{m_\alpha},\frac{\beta}{m_\beta}) + \frac{\sigma^2}{4}\log(m_\alpha m_\beta)\right) \tag{2.293}$$

Thus:

$$\lim_{\gamma\to+\infty} 2\gamma(\sqrt{m_\alpha m_\beta} - m_\pi(\gamma)) = \lim_{\omega\to 0}\frac{4}{\omega}(\sqrt{m_\alpha m_\beta} - m_\pi(\omega)) \tag{2.294}$$

$$= -4\frac{\mathrm{d}m_\pi}{\mathrm{d}\omega}(0) \tag{2.295}$$

$$= \sqrt{m_\alpha m_\beta}\left(\mathrm{OT}^{\otimes}_{2\sigma^2}(\frac{\alpha}{m_\alpha},\frac{\beta}{m_\beta}) - \sigma^2\log(m_\alpha m_\beta)\right)\ . \tag{2.296}$$

Adding $2\sigma^2\sqrt{m_\alpha m_\beta}(\sqrt{m_\alpha m_\beta}-1)$ from (2.209) ends the proof. ■

5.2 Proofs of the Gaussian barycenter theorems

Convexity and Optimality condition In this section we show how the notion of differentiability along feasible directions in $\mathcal{P}(\mathbb{R}^d)$ is enough to characterize convexity and first order optimality conditions. Consider an arbitrary function F on the space of probability measures.

Definition 3 *F is said to be differentiable at $\alpha \in \mathcal{P}(\mathbb{R}^d)$, if and only if there exists $\nabla F(\alpha) \in \mathcal{C}(\mathbb{R}^d)$ such that for any displacement $\delta\alpha = \alpha_1 - \alpha_2$ with $\alpha_1, \alpha_2 \in \mathcal{P}(\mathbb{R}^d)$:*

$$F(\alpha + t\delta\alpha) = F(\alpha) + t\langle \delta\alpha, \nabla F(\alpha)\rangle + o(t) \ , \tag{2.297}$$

where $\langle \eta, \nabla F(\alpha)\rangle = \int_{\mathbb{R}^d} \nabla F(\alpha)\, \mathrm{d}\eta$.

Proposition 27 (convexity) *Assume F is differentiable on $\mathcal{P}(\mathbb{R}^d)$. F is convex If and only if for all $\alpha, \alpha' \in \mathcal{P}(\mathbb{R}^d)$:*

$$F(\alpha) \geq F(\alpha') + \langle \alpha - \alpha', \nabla F(\alpha')\rangle \ , \tag{2.298}$$

PROOF. ($\Rightarrow$). Assume (2.298) holds. Let $\lambda \in [0,1]$ and $\alpha_\lambda = \lambda\alpha + (1 - \lambda\alpha')$ with arbitrary probability measures α, α'. Applying (2.298) twice with $\alpha' = \alpha_\lambda$ leads to:

$$\begin{aligned} F(\alpha) &\geq F(\alpha_\lambda) + \langle \alpha - \alpha_\lambda, \nabla F(\alpha_\lambda)\rangle \\ F(\alpha') &\geq F(\alpha_\lambda) + \langle \alpha' - \alpha_\lambda, \nabla F(\alpha_\lambda)\rangle \end{aligned}$$

Multiplying the first equation by λ and the second one by $1 - \lambda$ before summing leads to:

$$\lambda F(\alpha) + (1 - \lambda)F(\alpha') \geq F(\alpha_\lambda).$$

Thus F is convex.

($\Leftarrow$). Assume F is convex. Let $\lambda \in (0,1)$. Convexity implies that:

$$\begin{aligned} & F(\lambda\alpha + (1-\lambda)\alpha') \leq \lambda F(\alpha) + (1-\lambda)F(\alpha') \\ & \Rightarrow F(\alpha' + \lambda(\alpha - \alpha')) \leq \lambda F(\alpha) + (1-\lambda)F(\alpha') \\ & \Rightarrow F(\alpha') + \lambda\langle \alpha - \alpha', \nabla F(\alpha')\rangle + o(\lambda) \leq \lambda F(\alpha) + (1-\lambda)F(\alpha') \\ & \Rightarrow \lambda\langle \alpha - \alpha', \nabla F(\alpha')\rangle + o(\lambda) \leq \lambda F(\alpha) - \lambda F(\alpha') \\ & \Rightarrow \langle \alpha - \alpha', \nabla F(\alpha')\rangle + \frac{o(\lambda)}{\lambda} \leq F(\alpha) - F(\alpha') \end{aligned}$$

Letting $\lambda \to 0$ leads to (2.298). ■

Proposition 28 (Optimality condition) *Assume F is differentiable and convex on $\mathcal{P}(\mathbb{R}^d)$ then $\alpha^\star$ minimizes F if and only if $\langle \nabla F(\alpha^\star), \alpha - \alpha^\star\rangle \geq 0$.*

PROOF. ($\Rightarrow$) Assume $\alpha^\star$ is a minimizer of F. Let $t >$. Since $\mathcal{P}(\mathbb{R}^d)$ is convex, we can write for any $\alpha \in \mathcal{P}(\mathbb{R}^d)$:

$$F(\alpha^\star) \leq F(\alpha^\star + t(\alpha - \alpha^\star))$$

For t small enough, we can use (2.297) on the right-hand side:

$$F(\alpha^\star) \leq F(\alpha^\star) + t\langle \alpha - \alpha^\star, \nabla F(\alpha^\star)\rangle + o(t)$$

Dividing by t and letting $t \to 0$ leads to $\langle \alpha - \alpha^\star, \nabla F(\alpha^\star)\rangle \geq 0$ for all α.

($\Leftarrow$) Assume $\langle \nabla F(\alpha^\star), \alpha - \alpha^\star\rangle \geq 0$. Proposition 27 applies and (2.298) allows to conclude that $\alpha^\star$ is a minimizer of F. ■

Proofs of the barycenter theorems For any probability measure α, let $\bar{\alpha}$ denote its centered transformation $\alpha - \mathbb{E}_\alpha(X)$. Let F be any of the divergences $\mathrm{OT}^{\mathcal{L}}_{2\sigma^2}$, $\mathrm{OT}^{\otimes}_{2\sigma^2}$ or $S_{2\sigma^2}$. Thanks to Lemma 2, the barycenter problem can be restricted to centered distributions. Indeed, the following holds:

$$\mathrm{OT}^{\otimes}_{2\sigma^2}(\alpha, \beta) = \|\mathbb{E}_\alpha(X) - \mathbb{E}_\beta(X)\|^2 + \mathrm{OT}^{\otimes}_{2\sigma^2}(\bar{\alpha}, \bar{\beta}) \ . \tag{2.299}$$

Similarly, since for Lebesgue continuous measures the entropy is irrelevant to the centering of the distribution $\mathrm{KL}(\alpha||\mathcal{L}) = \mathrm{KL}(\bar{\alpha}||\mathcal{L})$, it holds thanks to (2.94):

$$\mathrm{OT}^{\mathcal{L}}_{2\sigma^2}(\alpha, \beta) = \|\mathbb{E}_\alpha(X) - \mathbb{E}_\beta(X)\|^2 + \mathrm{OT}^{\mathcal{L}}_{2\sigma^2}(\bar{\alpha}, \bar{\beta}) \ . \tag{2.300}$$

And finally:

$$S_{2\sigma^2}(\alpha, \beta) = \|\mathbb{E}_\alpha(X) - \mathbb{E}_\beta(X)\|^2 + S_{2\sigma^2}(\bar{\alpha}, \bar{\beta}) \ . \tag{2.301}$$

Thus, each barycenter problem is equivalent to:

$$\begin{aligned}
&\min_{\beta \in \mathcal{G}} \sum_{k=1}^K w_k F(\alpha_k, \beta) \\
&= \min_{\beta \in \mathcal{G}} \sum_{k=1}^K w_k \|\mathbf{a}_k - \mathbb{E}_\beta(X)\|^2 + w_k F(\bar{\alpha_k}, \bar{\beta}) \\
&= \min_{\substack{\mathbf{b} \in \mathbb{R}^d \\ \beta \in \mathcal{G}, \mathbb{E}_\beta(\mathbf{X}) = 0}} \sum_{k=1}^K w_k \|\mathbf{a}_k - \mathbf{b}\|^2 + w_k F(\bar{\alpha_k}, \beta)
\end{aligned} \tag{2.302}$$

Therefore, since both arguments are independent, we can first minimize over $\mathbf{b}$ to obtain $\mathbb{E}_\beta(\mathbf{X}) = \mathbf{b} = \sum_{k=1}^K w_k \mathbf{a}_k$, which provides the mean of the barycenter that is identical in all 3 theorems. Without loss of generality, we assume from now on that $\mathbf{a}_k = 0$ for all k.

We have showed that $F \in \{\mathrm{OT}^{\mathcal{L}}_{2\sigma^2}, \mathrm{OT}^{\otimes}_{2\sigma^2}, S_{2\sigma^2}\}$ is differentiable and convex (w.r.t. one measure at a time) on sub-Gaussian measures. Characterizing the barycenter can thus be done using the first order optimality condition:

$$\beta = \operatorname*{arg\,min}_{\alpha \in \mathcal{G}(\mathbb{R}^d)} \sum_{k=1}^K w_k F(\alpha_k, \beta) \Leftrightarrow \text{for any } \mu \in \mathcal{G} \quad \langle \sum_{k=1}^K w_k \nabla_\beta F(\alpha_k, \beta), \mu - \beta\rangle \geq 0 \tag{2.303}$$

We remind the reader that the notion of differentiability is different from the usual Fréchet differentiability: a function $F : \mathcal{G} \to \mathbb{R}$ is differentiable at α if there exists $\nabla F(\alpha) \in \mathcal{C}(\mathbb{R}^d)$ such that for any displacement $t\delta\alpha$ with $t > 0$ and $\delta\alpha = \alpha_1 - \alpha_2$ with $\alpha_1, \alpha_2 \in \mathcal{G}$, and:

$$F(\alpha + t\delta\alpha) = F(\alpha) + t\langle \delta\alpha, \nabla F(\alpha)\rangle + o(t) \ , \tag{2.304}$$

where $\langle \delta\alpha, \nabla F(\alpha)\rangle = \int_{\mathbb{R}^d} \nabla F(\alpha) \mathrm{d}\delta\alpha$.

Variance of the $S_{2\sigma^2}$ barycenter: theorem 5 Let (f_k, g_k) denote the potentials associated with $\mathrm{OT}^{\otimes}_{2\sigma^2}(\alpha_k, \beta)$ and h_β the autocorrelation potential associated with $\mathrm{OT}^{\otimes}_{2\sigma^2}(\beta, \beta)$. If β is sub-Gaussian, it holds: $\nabla_\beta S_{2\sigma^2}(\alpha_k, \beta) = g_k - h$. Therefore, from (2.298) a probability measure β is the debiased barycenter if and only if for any direction $\mu \in \mathcal{G}$, the optimality condition holds:

$$\begin{aligned} &\langle \sum_{k=1}^{K} w_k \nabla_\beta S_{2\sigma^2}(\alpha_k, \beta), \mu - \beta\rangle \geq 0 \\ &\Leftrightarrow \sum_{k=1}^{K} w_k \langle g_k - h_\beta, \mu - \beta\rangle \geq 0 \end{aligned} \tag{2.305}$$

Moreover, the potentials $(f_k), (g_k)$ and h must verify the Sinkhorn optimality conditions (2.38) for all k and for all x β-a.s and y α-a.s:

$$\begin{cases} e^{\frac{f_k(x)}{2\sigma^2}} \left(\int_{\mathbb{R}^d} e^{\frac{-\|x-y\|^2 + g_k(y)}{2\sigma^2}} \mathrm{d}\beta(y) \right) = 1, \quad e^{\frac{g_k(x)}{2\sigma^2}} \left(\int_{\mathbb{R}^d} e^{\frac{-\|x-y\|^2 + f_k(y)}{2\sigma^2}} \mathrm{d}\alpha_k(y) \right) = 1. \\ e^{\frac{h(x)}{2\sigma^2}} \left(\int_{\mathbb{R}^d} e^{\frac{-\|x-y\|^2 + h_\beta(y)}{2\sigma^2}} \mathrm{d}\beta(y) \right) = 1. \end{cases} \tag{2.306}$$

We are going to show that for the Gaussian measure β given in the statement of the theorem is well-defined and verifies all optimality conditions (2.306). 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 (2.302)). The Sinkhorn equations can therefore be written as a system on positive definite matrices:

$$\mathbf{F}_k = \sigma^2 \mathbf{A}_k^{-1} + \mathbf{G}_k^{-1}, \ \ \mathbf{G}_k = \sigma^2 \mathbf{B} + \mathbf{F}_k^{-1}, \ \ \mathbf{H} = \sigma^2 \mathbf{B} + \mathbf{H}^{-1}$$

where for all k:

$$\begin{aligned} \frac{f_k}{2\sigma^2} &= \mathcal{Q}(\frac{1}{\sigma^2}(\mathbf{G}_k^{-1} - \mathrm{Id})) + \frac{f_k(0)}{2\sigma^2} \\ \frac{g_k}{2\sigma^2} &= \mathcal{Q}(\frac{1}{\sigma^2}(\mathbf{F}_k^{-1} - \mathrm{Id})) + \frac{g_k(0)}{2\sigma^2} \\ \frac{h_\beta}{2\sigma^2} &= \mathcal{Q}(\frac{1}{\sigma^2}(\mathbf{H}^{-1} - \mathrm{Id})) + \frac{h_\beta(0)}{2\sigma^2} \end{aligned} \tag{2.307}$$

Moreover, provided $\mathbf{B}$ exists and is positive definite, the system (2.307) has a unique set of solutions $(\mathbf{F}_k)_k, (\mathbf{G}_k)_k, \mathbf{H}$ given by:

$$\mathbf{F}_k = \mathbf{B}\mathbf{C}_k^{-1}, \;\; \mathbf{G}_k = \mathbf{C}_k^{-1}\mathbf{A}_k, \;\; \mathbf{H} = \mathbf{B}^{-1}\mathbf{J} \tag{2.308}$$

where $\mathbf{C}_k = (\mathbf{A}_k\mathbf{B} + \frac{\sigma^4}{4}\,\mathrm{Id})^{\frac{1}{2}} - \frac{\sigma^2}{2}\,\mathrm{Id}$ and $\mathbf{J} = (\mathbf{B}^2 + \frac{\sigma^4}{4}\,\mathrm{Id})^{\frac{1}{2}} + \frac{\sigma^2}{2}\,\mathrm{Id}$. Therefore, the gradient in (2.305) can be written:

$$\begin{aligned}\sum_{k=1}^{K} w_k(g_k - h_\beta) &= \mathcal{Q}(2(\sum_{k=1}^{K} w_k\mathbf{F}_k^{-1} - \mathbf{H}^{-1})) + \sum_{w=1}^{K} w_k g_k(0) - h_\beta(0) \\ &= \mathcal{Q}(2(\sum_{k=1}^{K} w_k\mathbf{F}_k^{-1} - \mathbf{H}^{-1})) + m\end{aligned} \tag{2.309}$$

for some constant $m \in \mathbb{R}$. Let's compute the matrix defining the quadratic form:

$$\begin{aligned}&\sum_{k=1}^{K} w_k\mathbf{C}_k\mathbf{B}^{-1} - \mathbf{J}^{-1}\mathbf{B} \\ &= \sum_{k=1}^{K} w_k\mathbf{B}^{-\frac{1}{2}}(\mathbf{B}^{\frac{1}{2}}\mathbf{A}_k\mathbf{B}^{\frac{1}{2}} + \frac{\sigma^4}{4}\,\mathrm{Id})^{\frac{1}{2}}\mathbf{B}^{-\frac{1}{2}} - \mathbf{B}^{-1}(\mathbf{B}^2 + \frac{\sigma^4}{4}\,\mathrm{Id})^{\frac{1}{2}} \\ &= \sum_{k=1}^{K} w_k\mathbf{B}^{-\frac{1}{2}}(\mathbf{B}^{\frac{1}{2}}\mathbf{A}_k\mathbf{B}^{\frac{1}{2}} + \frac{\sigma^4}{4}\,\mathrm{Id})^{\frac{1}{2}}\mathbf{B}^{-\frac{1}{2}} - \mathbf{B}^{-\frac{1}{2}}(\mathbf{B}^2 + \frac{\sigma^4}{4}\,\mathrm{Id})^{\frac{1}{2}}\mathbf{B}^{-\frac{1}{2}} \\ &= \mathbf{B}^{-\frac{1}{2}}\left(\sum_{k=1}^{K} w_k(\mathbf{B}^{\frac{1}{2}}\mathbf{A}_k\mathbf{B}^{\frac{1}{2}} + \frac{\sigma^4}{4}\,\mathrm{Id})^{\frac{1}{2}} - (\mathbf{B}^2 + \frac{\sigma^4}{4}\,\mathrm{Id})^{\frac{1}{2}}\right)\mathbf{B}^{-\frac{1}{2}}\end{aligned} \tag{2.310}$$

which is null if $\mathbf{B}$ is a solution of the equation:

$$\sum_{k=1}^{K} w_k(\mathbf{B}^{\frac{1}{2}}\mathbf{A}_k\mathbf{B}^{\frac{1}{2}} + \frac{\sigma^4}{4}\,\mathrm{Id})^{\frac{1}{2}} = (\mathbf{B}^2 + \frac{\sigma^4}{4}\,\mathrm{Id})^{\frac{1}{2}}. \tag{2.311}$$

Therefore, the gradient is constant and equal to m. For any probability measure $\mu \in \mathcal{G}$:

$$\begin{aligned}\langle\sum_{k=1}^{K} w_k\nabla_\beta S_{2\sigma^2}(\alpha_k, \beta), \mu - \beta\rangle &= \langle\sum_{k=1}^{K} w_k g_k - h_\beta, \mu - \beta\rangle \\ &= \langle m, \mu - \beta\rangle \\ &= 0\end{aligned} \tag{2.312}$$

since both measures integrate to 1. Therefore, the optimality condition holds.

To end the proof, all we need to show is that (2.311) admits a positive definite solution. To show the

existence of a solution, the same proof of Agueh and Carlier (2011) applies. Indeed, let λ_k and Λ_k denote respectively the smallest and largest eigenvalue of $\mathbf{A}_k$. Let $\lambda = \min_k \lambda_k$ and $\Lambda = \max_k \Lambda_k$. Let $K_{\lambda,\Lambda}$ be the convex compact subset of positive definite matrices $\mathbf{B}$ such that $\Lambda\,\mathrm{Id} \succeq \mathbf{B} \succeq \lambda\,\mathrm{Id}$. Define the map:

$$\begin{aligned} T : & K_{\lambda,\Lambda} \to \mathcal{S}_+^d \\ & \mathbf{B} \mapsto \left(\left(\sum_{k=1}^K w_k (\mathbf{B}^{\frac{1}{2}} \mathbf{A}_k \mathbf{B}^{\frac{1}{2}} + \frac{\sigma^4}{4}\,\mathrm{Id})^{\frac{1}{2}} \right)^2 - \frac{\sigma^4}{4}\,\mathrm{Id} \right)^{\frac{1}{2}} \end{aligned}$$

Now for any $\mathbf{B} \in K_{\lambda,\Lambda}$, it holds:

$$\lambda\,\mathrm{Id} \preceq T(\mathbf{B}) \preceq \Lambda\,\mathrm{Id}\,. \tag{2.313}$$

T is therefore a continuous function that maps $K_{\lambda,\Lambda}$ to itself, thus Brouwer's fixed-point theorem guarantees the existence of a solution. ■

Variance of the $\mathrm{OT}^{\otimes}_{2\sigma^2}$ barycenter: theorem 4 Let (f_k, g_k) denote the potentials associated with $\mathrm{OT}^{\otimes}_{2\sigma^2}(\alpha_k, \beta)$. If β is sub-Gaussian, it holds: $\nabla_\beta \mathrm{OT}_{2\sigma^2}(\alpha_k, \beta) = g_k$. Therefore, from (2.298), a probability measure β is the $\mathrm{OT}^{\otimes}_{2\sigma^2}$ barycenter if and only if for any direction $\mu \in \mathcal{G}$, the optimality condition holds:

$$\begin{aligned} \langle \sum_{k=1}^K w_k \nabla_\beta \mathrm{OT}_{2\sigma^2}(\alpha_k, \beta), \mu - \beta \rangle \geq 0 \\ \Leftrightarrow \sum_{k=1}^K w_k \langle g_k, \mu - \beta \rangle \geq 0 \end{aligned} \tag{2.314}$$

Moreover, the potentials $(f_k), (g_k)$ must verify the Sinkhorn optimality conditions (2.38) for all k and for all x β-a.s and y α-a.s:

$$\left\{ e^{\frac{f_k(x)}{2\sigma^2}} \left(\int_{\mathbb{R}^d} e^{\frac{-\|x-y\|^2 + g_k(y)}{2\sigma^2}} \mathrm{d}\beta(y) \right) = 1, \quad e^{\frac{g_k(x)}{2\sigma^2}} \left(\int_{\mathbb{R}^d} e^{\frac{-\|x-y\|^2 + f_k(y)}{2\sigma^2}} \mathrm{d}\alpha_k(y) \right) = 1. \right. \tag{2.315}$$

We are going to show that for the Gaussian measure β given in the statement of theorem 4 is well-defined and verifies all optimality conditions (2.315). 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 (2.302)). The Sinkhorn equations can therefore be written as a system on positive definite matrices:

$$\mathbf{F}_k = \sigma^2 \mathbf{A}_k^{-1} + \mathbf{G}_k^{-1}, \;\; \mathbf{G}_k = \sigma^2 \mathbf{B} + \mathbf{F}_k^{-1}$$

where for all k:

$$\begin{aligned}\frac{f_k}{2\sigma^2} &= \mathcal{Q}(\frac{1}{\sigma^2}(\mathbf{G}_k^{-1} - \mathrm{Id})) + \frac{f_k(0)}{2\sigma^2} \\ \frac{g_k}{2\sigma^2} &= \mathcal{Q}(\frac{1}{\sigma^2}(\mathbf{F}_k^{-1} - \mathrm{Id})) + \frac{g_k(0)}{2\sigma^2}\end{aligned} \tag{2.316}$$

Moreover, provided $\mathbf{B}$ exists and is positive definite, the system (2.307) has a unique set of solutions $(\mathbf{F}_k)_k, (\mathbf{G}_k)_k$ given by:

$$\mathbf{F}_k = \mathbf{B}\mathbf{C}_k^{-1}, \;\; \mathbf{G}_k = \mathbf{C}_k^{-1}\mathbf{A}_k, \tag{2.317}$$

where $\mathbf{C}_k = (\mathbf{A}_k\mathbf{B} + \frac{\sigma^4}{4}\mathrm{Id})^{\frac{1}{2}} - \frac{\sigma^2}{2}\mathrm{Id}$. Therefore, the gradient in (2.314) can be written:

$$\begin{aligned}\sum_{k=1}^K w_k g_k &= \mathcal{Q}(2(\sum_{k=1}^K w_k \mathbf{F}_k^{-1} - \mathrm{Id})) + \sum_{w=1}^K w_k g_k(0) \\ &= \mathcal{Q}\left(2(\sum_{k=1}^K w_k \mathbf{F}_k^{-1} - \mathrm{Id})\right) + m\end{aligned} \tag{2.318}$$

for some $m \in \mathbb{R}$. Let's compute the matrix of the quadratic form:

$$\begin{aligned}\sum_{k=1}^K w_k \mathbf{C}_k\mathbf{B}^{-1} - \mathrm{Id} &= \sum_{k=1}^K w_k \mathbf{B}^{-\frac{1}{2}}(\mathbf{B}^{\frac{1}{2}}\mathbf{A}_k\mathbf{B}^{\frac{1}{2}} + \frac{\sigma^4}{4}\mathrm{Id})^{\frac{1}{2}}\mathbf{B}^{-\frac{1}{2}} - \frac{\sigma^2}{2}\mathbf{B}^{-1} - \mathrm{Id} \\ &= \mathbf{B}^{-\frac{1}{2}}\left(\sum_{k=1}^K w_k(\mathbf{B}^{\frac{1}{2}}\mathbf{A}_k\mathbf{B}^{\frac{1}{2}} + \frac{\sigma^4}{4}\mathrm{Id})^{\frac{1}{2}} - \frac{\sigma^2}{2}\mathrm{Id} + \mathbf{B}\right)\mathbf{B}^{-\frac{1}{2}}\end{aligned} \tag{2.319}$$

which is null if $\mathbf{B}$ is a solution of the equation:

$$\sum_{k=1}^K w_k(\mathbf{B}^{\frac{1}{2}}\mathbf{A}_k\mathbf{B}^{\frac{1}{2}} + \frac{\sigma^4}{4}\mathrm{Id})^{\frac{1}{2}} = \mathbf{B} + \frac{\sigma^2}{2}\mathrm{Id}\,. \tag{2.320}$$

Therefore, for any probability measure $\mu \in \mathcal{G}$:

$$\begin{aligned}\langle\sum_{k=1}^K w_k \nabla_\beta \mathrm{OT}^{\otimes}_{2\sigma^2}(\alpha_k, \beta), \mu - \beta\rangle &= \langle\sum_{k=1}^K w_k g_k, \mu - \beta\rangle \\ &= \langle m, \mu - \beta\rangle \\ &= m\int(\mathrm{d}\mu - \mathrm{d}\beta) \\ &= 0\end{aligned} \tag{2.321}$$

since both measures integrate to 1. Therefore, the optimality condition holds.

To end the proof, all we need to show is that (2.99) admits a positive definite solution. To show the existence of a solution, the same proof of Agueh and Carlier (2011) applies. Indeed, let Λ_k denote the largest eigenvalue of $\mathbf{A}_k$ and $\Lambda = \max_k \Lambda_k$. Let K_Λ be the convex compact subset of positive definite matrices $\mathbf{B}$ such that $(\Lambda - \sigma^2)_+ \operatorname{Id} \succeq \mathbf{B} \succeq 0$. Define the map:

$$\begin{aligned} T : & K_\Lambda \to \mathcal{S}_+^d \\ & \mathbf{B} \mapsto \sum_{k=1}^K w_k (\mathbf{B}^{\frac{1}{2}} \mathbf{A}_k \mathbf{B}^{\frac{1}{2}} + \frac{\sigma^4}{4} \operatorname{Id})^{\frac{1}{2}} - \frac{\sigma^2}{2} \operatorname{Id} \end{aligned}$$

Now for any $\mathbf{B} \in K_\Lambda$, it holds:

$$0 \preceq T(\mathbf{B}) \preceq (\Lambda - \sigma^2)_+ \operatorname{Id}. \tag{2.322}$$

T is therefore a continuous function that maps K_Λ to itself, thus Brouwer's fixed-point theorem guarantees the existence of a solution. ■

Variance of the $\mathrm{OT}_{2\sigma^2}^{\mathcal{L}}$ barycenter: theorem 3 Let (f_k, g_k) denote the potentials associated with $\mathrm{OT}_{2\sigma^2}^{\mathcal{L}}(\alpha_k, \beta)$. If β is sub-Gaussian, it holds: $\nabla_\beta \mathrm{OT}_{2\sigma^2}^{\mathcal{L}}(\alpha_k, \beta) = g_k + 2\sigma^2 \log\left(\frac{\mathrm{d}\beta}{\mathrm{d}\mathcal{L}}\right)$. Therefore, from (2.15), a probability measure β is the $\mathrm{OT}_{2\sigma^2}^{\mathcal{L}}$ barycenter if and only if for any direction $\mu \in \mathcal{G}$, the optimality condition holds:

$$\begin{aligned} & \langle \sum_{k=1}^K w_k \nabla_\beta \mathrm{OT}_{2\sigma^2}^{\mathcal{L}}(\alpha_k, \beta), \mu - \beta \rangle \geq 0 \\ \Leftrightarrow & \sum_{k=1}^K w_k \langle g_k + 2\sigma^2 \log\left(\frac{\mathrm{d}\beta}{\mathrm{d}\mathcal{L}}\right), \mu - \beta \rangle \geq 0 \end{aligned} \tag{2.323}$$

Moreover, the potentials $(f_k), (g_k)$ must verify the Sinkhorn optimality conditions (2.38) for all k and for all x β-a.s and y α-a.s:

$$\left\{ e^{\frac{f_k(x)}{2\sigma^2}} \left(\int_{\mathbb{R}^d} e^{\frac{-\|x-y\|^2 + g_k(y)}{2\sigma^2}} \mathrm{d}\beta(y) \right) = 1, \quad e^{\frac{g_k(x)}{2\sigma^2}} \left(\int_{\mathbb{R}^d} e^{\frac{-\|x-y\|^2 + f_k(y)}{2\sigma^2}} \mathrm{d}\alpha_k(y) \right) = 1. \right. \tag{2.324}$$

We are going to show that for the Gaussian measure β given in the statement of theorem 3 is well-defined and verifies all optimality conditions (2.324). 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 (2.302)).

On one hand, its density with respect to the Lebesgue measure verifies:

$$\log\left(\frac{\mathrm{d}\beta}{\mathrm{d}\mathcal{L}}\right)(x) = -\frac{1}{2}x^\top \mathbf{B}^{-1}x + \frac{1}{2}\log\det(2\pi\mathbf{B}) \tag{2.325}$$

$$= \mathcal{Q}(\mathbf{B}^{-1})(x) + m \tag{2.326}$$

for some constant $m \in \mathbb{R}$.

On the other hand, the Sinkhorn equations corresponding can therefore be written as a system on positive definite matrices:

$$\mathbf{F}_k = \sigma^2\mathbf{A}_k^{-1} + \mathbf{G}_k^{-1}, \;\; \mathbf{G}_k = \sigma^2\mathbf{B} + \mathbf{F}_k^{-1}$$

where for all k:

$$\begin{aligned} \frac{f_k}{2\sigma^2} &= \mathcal{Q}(\frac{1}{\sigma^2}(\mathbf{G}_k^{-1} - \mathrm{Id})) + \frac{f_k(0)}{2\sigma^2} \\ \frac{g_k}{2\sigma^2} &= \mathcal{Q}(\frac{1}{\sigma^2}(\mathbf{F}_k^{-1} - \mathrm{Id})) + \frac{g_k(0)}{2\sigma^2} \end{aligned} \tag{2.327}$$

Moreover, provided $\mathbf{B}$ exists and is positive definite, the system (2.307) has a unique set of solutions $(\mathbf{F}_k)_k, (\mathbf{G}_k)_k$ given by:

$$\mathbf{F}_k = \mathbf{B}\mathbf{C}_k^{-1}, \;\; \mathbf{G}_k = \mathbf{C}_k^{-1}\mathbf{A}_k, \tag{2.328}$$

where $\mathbf{C}_k = (\mathbf{A}_k\mathbf{B} + \frac{\sigma^4}{4}\mathrm{Id})^{\frac{1}{2}} - \frac{\sigma^2}{2}\mathrm{Id}$. Therefore, the gradient in (2.314) can be written:

$$\begin{aligned} &\sum_{k=1}^K w_k g_k + 2\sigma^2\log\left(\frac{\mathrm{d}\beta}{\mathrm{d}\mathcal{L}}\right) = \mathcal{Q}(2(\sum_{k=1}^K w_k\mathbf{F}_k^{-1} - \mathrm{Id} + \sigma^2\mathbf{B}^{-1})) + \sum_{w=1}^K w_k g_k(0) + 2\sigma^2 m \\ &= \mathcal{Q}\left(2(\sum_{k=1}^K w_k\mathbf{F}_k^{-1} - \mathrm{Id} + \sigma^2\mathbf{B}^{-1})\right) + m' \end{aligned} \tag{2.329}$$

for some constant $m' \in \mathbb{R}$. Let's compute the matrix of the quadratic form:

$$\begin{aligned} \sum_{k=1}^K w_k\mathbf{C}_k\mathbf{B}^{-1} - \mathrm{Id} + \sigma^2\mathbf{B}^{-1} &= \sum_{k=1}^K w_k\mathbf{B}^{-\frac{1}{2}}(\mathbf{B}^{\frac{1}{2}}\mathbf{A}_k\mathbf{B}^{\frac{1}{2}} + \frac{\sigma^4}{4}\mathrm{Id})^{\frac{1}{2}}\mathbf{B}^{-\frac{1}{2}} + \frac{\sigma^2}{2}\mathbf{B}^{-1} - \mathrm{Id} \\ &= \mathbf{B}^{-\frac{1}{2}}\left(\sum_{k=1}^K w_k(\mathbf{B}^{\frac{1}{2}}\mathbf{A}_k\mathbf{B}^{\frac{1}{2}} + \frac{\sigma^4}{4}\mathrm{Id})^{\frac{1}{2}} + \frac{\sigma^2}{2}\mathrm{Id} - \mathbf{B}\right)\mathbf{B}^{-\frac{1}{2}} \end{aligned} \tag{2.330}$$

which is null if $\mathbf{B}$ is a solution of the equation:

$$\sum_{k=1}^{K} w_k (\mathbf{B}^{\frac{1}{2}} \mathbf{A}_k \mathbf{B}^{\frac{1}{2}} + \frac{\sigma^4}{4} \mathrm{Id})^{\frac{1}{2}} = \mathbf{B} - \frac{\sigma^2}{2} \mathrm{Id}\,. \tag{2.331}$$

Therefore, for any probability measure $\mu \in \mathcal{G}$:

$$\begin{aligned} \langle \sum_{k=1}^{K} w_k \nabla_\beta \mathrm{OT}^{\mathcal{L}}_{2\sigma^2}(\alpha_k, \beta), \mu - \beta \rangle &= \langle \sum_{k=1}^{K} w_k g_k + 2\sigma^2 \log\left(\frac{\mathrm{d}\beta}{\mathrm{d}\mathcal{L}}\right), \mu - \beta \rangle \\ &= \langle m + m', \mu - \beta \rangle \\ &= m + m' \int (\mathrm{d}\mu - \mathrm{d}\beta) \\ &= 0 \end{aligned} \tag{2.332}$$

since both measures integrate to 1. Therefore, the optimality condition holds.

To end the proof, all we need to show is that (2.92) admits a positive definite solution. To show the existence of a solution, the same proof of Agueh and Carlier (2011) applies. Indeed, let Λ_k denote the largest eigenvalue of $\mathbf{A}_k$ and $\Lambda = \max_k \Lambda_k$. Let $K_{\sigma^2,\Lambda}$ be the convex compact subset of positive definite matrices $\mathbf{B}$ such that $(\Lambda + \sigma^2)\,\mathrm{Id} \succeq \mathbf{B} \succeq \sigma^2\,\mathrm{Id}$. Define the map:

$$\begin{aligned} T : K_{\sigma^2,\Lambda} &\to \mathcal{S}^d_+ \\ \mathbf{B} &\mapsto \sum_{k=1}^{K} w_k (\mathbf{B}^{\frac{1}{2}} \mathbf{A}_k \mathbf{B}^{\frac{1}{2}} + \frac{\sigma^4}{4} \mathrm{Id})^{\frac{1}{2}} + \frac{\sigma^2}{2} \mathrm{Id} \end{aligned}$$

Now for any $\mathbf{B} \in K_{\sigma^2,\Lambda}$, it holds:

$$\sigma^2\,\mathrm{Id} \preceq T(\mathbf{B}) \preceq (\Lambda + \sigma^2)\,\mathrm{Id}\,. \tag{2.333}$$

T is therefore a continuous function that maps $K_{\sigma^2,\Lambda}$ to itself, thus Brouwer's fixed-point theorem guarantees the existence of a solution in $K_{\sigma^2,\Lambda}$. ■

Chapter 3

Optimal transport as a multi-task prior

A learner is facing several learning tasks. Intuitively, one may think that isolated and focused learning on one single task at a time leads to greater mastery of the given assignments. Unless the tasks have nothing in common, this intuition is more likely to be wrong. Take the example of an archer practicing his aiming skills with different ranges, wind conditions and moving targets. Two methods are offered before them. Method 1: practice one scheme at a time until reaching mastery before moving on to the other. Method 2: practice throws in a mixed-up scheme. Even though method 1 is common and more appealing, chances are that in test time, when evaluated on a random setting in the Olympics, method 2 proves to be the strategy to follow. This argument is based on behavioral experiments in different types of exercises in sports, arts and mnemonics (Brown, Roediger, and McDaniel, 2014). It can be explained by two factors. By following method 1, repeated trials of the same task may lead to great performances in training sessions that are not necessarily reproducible at evaluation (*overfitting*). However, by following method 2 the subject not only learns each task but also learns to adapt from one to another (*generalization*).

These psychology and cognitive science observations relate to similar ones in machine learning. By acquiring knowledge from slightly different sources, multi-task learning models can generalize better and avoid overfitting. Moreover, in high dimensional settings (when the dimension of the problem p is larger than the number of samples n), solving the problems jointly can be seen as a way of *pooling* together more data if the learned entities share some underlying structure (Caruana, 1993). This chapter is devoted to modeling this underlying structure in the context of brain imaging across different individuals. Taking into account the spatial geometry of the cortex is a crucial ingredient for an accurate portrayal of the similarity across tasks. Enters optimal transport.

This chapter is based on:

- H. Janati et al, *Wasserstein regularization for sparse multi-task regression*, AISTATS'19.
- H. Janati et al, *Minimum Wasserstein Estimates: group level EEG-MEG source imaging via optimal transport*, IPMI'19.
- H. Janati et al, *Multi-subject source imaging with sparse multi-task regression*, Neuroimage 2020.

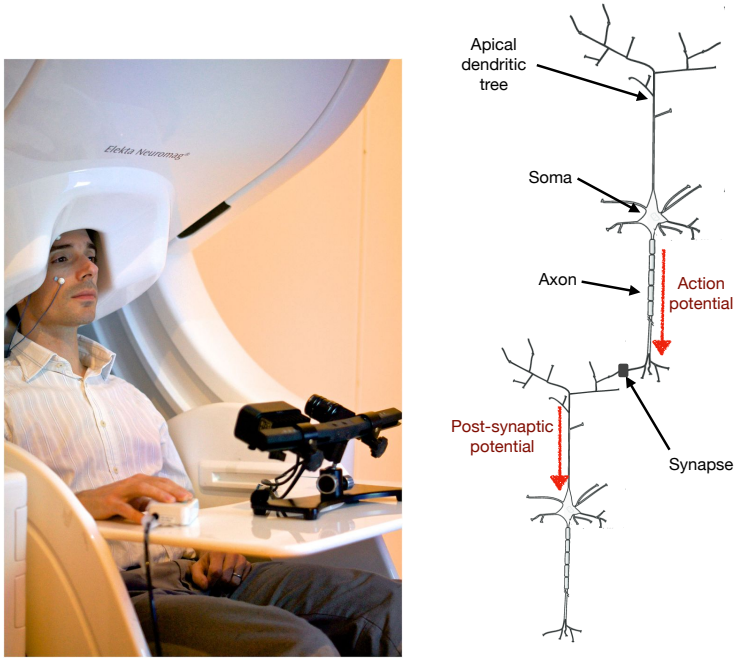


Fig. 3.1. Left: Subject wearing a MEG scanner and using a controller for a cognitive experiment (Proudfoot et al., 2014). **Right:** schematic of two pyramidal neurons.

1 Brain imaging

An area of the brain is active when neurons in that area emit electrical currents. If strong enough, these electrical currents can be sensed by measuring the electrical potentials at the surface of the head (electroencephalography or EEG). The magnetic field created by those neural currents (magneto-encaphalography or MEG) can also be measured outside the head using magnetic sensors. EEG and MEG are direct brain imaging techniques in the sense that they measure the electrical brain activity with no intermediary middleman. On the contrary, indirect modalities rely on some physiological phenomenon correlated with brain activity. Functional magnetic resonance imaging (fMRI) for instance relies on measuring the rapid delivery of blood to neural cells known as the *haemodynamic response*. These inherent differences lead to different characteristics: because fMRI relies on magnetic resonance, it provides 3D activation maps of the entire brain with an accuracy reaching less than 1mm (Duyn, 2012). However, blood delivery is slow and delayed in time thereby limiting the temporal resolution of fMRI to the order of 1 second. EEG and MEG provide the complimentary picture: high temporal resolution up to a millisecond, but low spatial resolution since measurements are taken from a distance. Recovering the brain sources from these measurements is known as source localization (Baillet, 2017; Baillet, Mosher, and Leahy, 2001a).

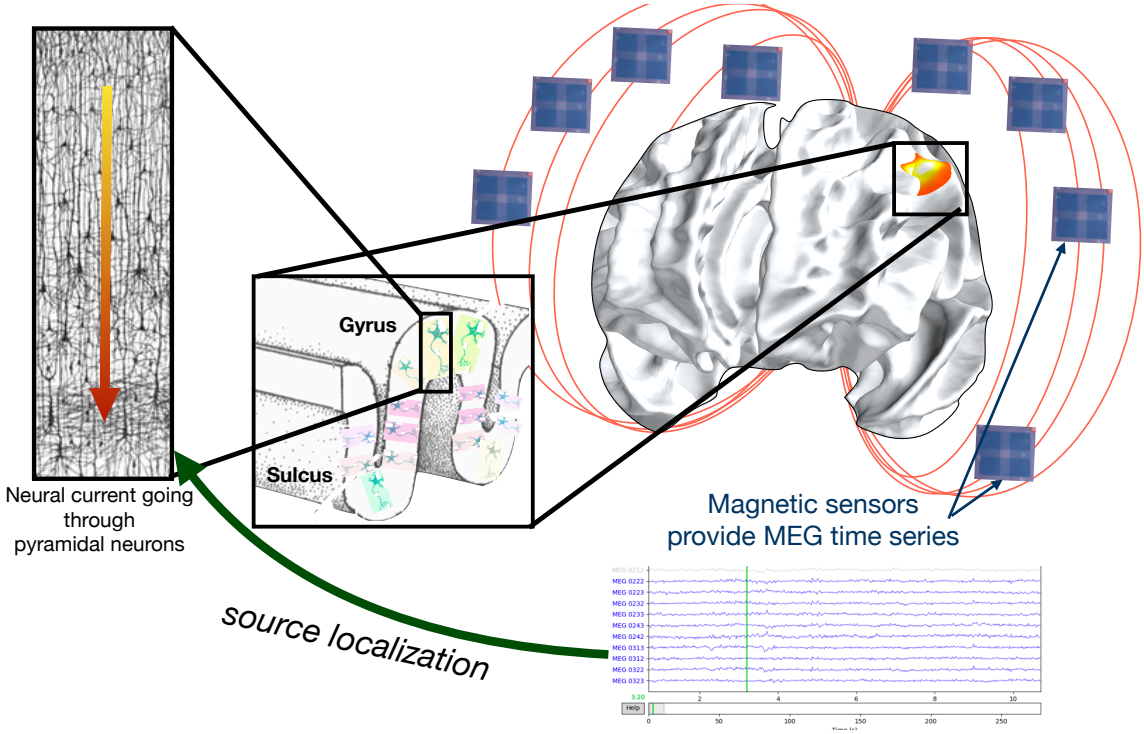


Fig. 3.2. Schematic illustrating neural activity from brain sources to MEG data. Source localization consists in inferring the location and amplitudes of neural currents from electromagnetic field measurements. The drawing of the pyramidal neurons – which contribute most to EEG and MEG signals – are taken from (Ramón y Cajal, 1899).

1.1 From brain recordings to brain activity: source localization

Sudden changes in the concentration of the ions in contact with a pyramidal neuron's membrane create un *action potential* (Figure 3.1). Action potentials are electrical impulses that travel along the neuron's "long cable" called *axon* before being transmitted to neighboring neurons at their "junctions" (*synapses*). At the receiving neuron, *post-synaptic potentials* are generated in its apical dendritic tree. These electrical impulses are however very weak: around a few µA. To create a magnetic field strong enough to be captured by MEG sensors, these impulses must be coordinated in time and orientation. This is the main reason why MEG measurements are most likely due to post-synaptic potentials in the columnar organization of the cortex where large pyramidal neurons have parallel apical dendrites (Nunez and Srinivasan, 2006) as shown in Figure 3.2. Localizing the underlying neural activity at the origin of the signals is a linear inverse problem known as *source localization* (Baillet, Mosher, and Leahy, 2001b; Becker et al., 2015; Michel et al., 2004; Wipf and Nagarajan, 2009). To solve this inverse problem and localize the underlying brain activity, several linear and non-linear methods have been proposed and reviewed in the

literature (Fuchs et al., 1999; Becker et al., 2015; Vega-Hernández et al., 2008). One commonly employed strategy to tackle the inverse problem is beamforming (Van Veen et al., 1997; Gross et al., 2001) or more generally scanning techniques (Mosher and Leahy, 1999; Mäkelä et al., 2018). A popular alternative casts the problem as a *linear regression problem* in high dimension; an approach commonly referred to as the *imaging approach* (Baillet, Mosher, and Leahy, 2001b). However, and despite the linearity of the forward model, this inverse problem is inherently difficult as it is "ill-posed". Indeed, the number of potential sources is much larger than the number of MEG and EEG sensors, which implies that, even in the absence of noise, different neural activity patterns could result in the same electromagnetic field measurements. Moreover, each source is modeled as a dipole for which both an amplitude and an orientation must be inferred. This fact makes M/EEG source localization particularly challenging in the presence of multiple simultaneous active regions in the brain. Incorporating additional prior information is mandatory. Such information can be formulated for instance using Bayesian framework (Haufe et al., 2009; Cai et al., 2020) or by exploiting the temporal axis of the data (Castaño-Candamil et al., 2015).

Cortically constrained source spaces In the imaging approach that leads to regression models, the position of the potential current sources in the brain need to be defined. Given a segmentation of the MRI scan of each subject, the sources, which are modeled as electric current dipoles, can be either placed on a regular grid spanning the entire brain volume, or positioned along the cortical mantle (Dale et al., 2000). When working with such a cortically constrained model, and since synchronized currents flowing along the apical dendrites of cortical pyramidal neurons are thought to be mostly responsible for M/EEG signals (Okada, 1993), it is then possible to constrain the dipole orientations to be normal to the cortical surface. Doing so, solving the inverse problem amounts to estimating the amplitudes of current dipoles whose positions and orientations are fixed a priori. The ensemble of possible candidate dipoles forms what is generally referred to as the *source space*. In this work, we will consider cortically constrained source spaces.

Forward modeling Let n denote the number of sensors (EEG and/or MEG) and p the number of dipoles in the source space. Following Maxwell's equations, at each time instant, the electromagnetic field measurements $\mathbf{b} \in \mathbb{R}^n$ are a linear combination of the fields produced by all sources $\mathbf{x} \in \mathbb{R}^p : \mathbf{b} = \mathbf{L}\mathbf{x}$. The linear forward operator $\mathbf{L} \in \mathbb{R}^{n \times p}$ is called the *leadfield* or *gain matrix*. Factoring noise in the measurements $\mathbf{y} \in \mathbb{R}^n$ leads to:

$$\mathbf{y} = \mathbf{b} + \varepsilon = \mathbf{L}\mathbf{x} + \eta \ , \tag{3.1}$$

where η is some noise vector that is generally assumed Gaussian distributed $\mathcal{N}(0, \mathbf{\Sigma})$. In practice, $\mathbf{L}$ is computed by solving Maxwell's equations using for example a boundary element method (BEM) (Hämäläinen and Sarvas, 1987; Mosher, Leahy, and Lewis, 1999; Kybic et al., 2005).

Whitening Since M/EEG signals are correlated by design, the noise covariance matrix $\mathbf{\Sigma}$ is not diagonal. For the inverse problem to be cast as a least squares regression problem, one needs to apply a whitening transformation to the data. Given an estimate $\hat{\mathbf{\Sigma}}$, that can be obtained from pre-stiumus baseline periods,

the whitening step amounts to computing the transformed data $\hat{\mathbf{\Sigma}}^{-\frac{1}{2}}\mathbf{y}$ and $\hat{\mathbf{\Sigma}}^{-\frac{1}{2}}\mathbf{L}$ (Engemann and Gramfort, 2015). In the rest of this chapter, we will assume that the data are whitened, meaning that the noise has unit variance at each sensor, and that it is uncorrelated across sensors.

1.2 Ill-conditioning and prior biases

When using the imaging approach, one limits the set of possible solutions by making prior hypotheses on the nature of the source distributions. Formally, this can be achieved by optimizing a regularized data fitting loss:

$$\mathbf{x}^\star = \arg\min_{\mathbf{x}\in\mathbb{R}^p} \frac{1}{2n}\|\mathbf{y} - \mathbf{L}\mathbf{x}\|_2^2 + \omega(\mathbf{x}) \ , \tag{3.2}$$

where ω is a prior function.

Perhaps one of the most simple regularizers is the *Ridge* penalty, the solutions of which are known as *minimum-norm estimates* (MNE) (Hämäläinen and Ilmoniemi, 1994):

$$\mathbf{x}^{\text{MNE}} = \arg\min_{\mathbf{x}\in\mathbb{R}^p} \frac{1}{2n}\|\mathbf{y} - \mathbf{L}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_2^2, \tag{3.3}$$

The inverse solutions discussed above typically employ penalties ω that are increasing functions of the source amplitudes. This inherently induces a bias towards sources in the superficial layers of the cortex (Kohler et al., 1996; Lin et al., 2006). Indeed, deep sources require larger amplitude values than superficial ones to produce electromagnetic fields with similar strength. To circumvent this problem, one can normalize the columns of the leadfield $\mathbf{L}$ by a fraction of their norms (Lin et al., 2006; Gramfort et al., 2013b). In all our experiments we use a depth weighting of 0.9. Formally, this means that every column $\mathbf{L}_{.j}$ is normalized by $\|\mathbf{L}_{.j}\|^{0.9}$.

The proposal of minimum-norm estimates (3.3) lead to several variants variants such as dSPM (Dale et al., 2000) which relies on noise normalization or sLORETA (Pascual-Marqui, 2002), proposed to correct for the depth bias induced by the ℓ_2 norm (Ahlfors, Ilmoniemi, and Hämäläinen, 1992). These methods have linear solutions and are very cheap to compute but promote weak and distributed neural patterns that inevitably lead to low spatial accuracy. When studying the brain response of specific and simple cognitive tasks, it is more principled to favor strong and sparse sources as long as they explain the data. Sparsity can be promoted using an ℓ_1 norm penalty. The resulting problem, known as *minimum current estimates* (MCE) (Uutela, Hämäläinen, and Somersalo, 1999) in the field of M/EEG, and Lasso (Tibshirani, 1996) in the machine learning community, reads:

$$\mathbf{x}^\star = \arg\min_{\mathbf{x}\in\mathbb{R}^p} \frac{1}{2n}\|\mathbf{y} - \mathbf{L}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1 \ . \tag{3.4}$$

The possible choices of ω are virtually limitless: block-sparse norms can be used to defined *mixed-norms estimates* (MxNE) (Strohmeier et al., 2016) and their time-frequency variant (TF-MxNE) (Gramfort et al., 2013b) to leverage the spatio-temporal dynamics of M/EEG signals. If other imaging data are available

such as fMRI (Zhongming Liu, Lei Ding, and Bin He, 2006; Ou et al., 2010) or diffusion MRI (Deslauriers-Gauthier et al., 2017), it is also possible to use them as prior information for example in hierarchical Bayesian models (Sato et al., 2018). While such techniques have had some success, source estimation in the presence of complex multi-dipole configurations remains a challenge. To address it, one idea is to leverage the anatomical and functional diversity of multi-subject datasets to improve localization results.

1.3 Multi-subject source localization

The idea of using multi-subject information to improve the spatial accuracy of M/EEG source imaging has been proposed before in the neuroimaging community. Before presenting our contributions, let's discuss some of the related proposals in the literature.

Related work Larson, Maddox, and Lee (2014) hypothesized that different anatomies across subjects allow for different point spread functions that only overlap on one source location. Averaging across subjects thereby increases the accuracy of source localization. On fMRI data, Varoquaux et al. (2011) proposed a probabilistic dictionary learning model to infer activation maps jointly across a cohort of subjects. A similar idea lead Litvak and Friston (2008) to propose a Bayesian hierarchical model to cope with inter-subject functional variability. Their model uses a multiple sparse prior defined using 256 bilateral patches, making the assumption that the same few patches are active with different amplitudes for all subjects. Each patch is created using a Laplacian diffusion of a source on the cortical mesh. This Laplacian operator is defined with a smoothness coefficient that can vary between zero and one, and defaults to 0.6 in the reference implementation. However, when the number of subjects increases, it requires more patches to cope with the diversity of cortical orientation patterns. This becomes problematic for a number of subjects as small as 10. The assumption of a common set of active patches was then relaxed by Kozunov and Ossadtchi (2015) who proposed a Bayesian model inspired by Litvak and Friston (2008). Instead of defining spatial patches as priors, GALA models spatial coherence across subjects through a number of fixed covariance matrices that embody both group similarities and individual signatures. The weights of these covariance matrices are then learned adaptively from the data. The novelty of the GALA model lies in the design of some fixed covariance matrices to model functional similarity across subjects using the geometry of the cortex. This is achieved using a Gaussian Kernel defined on the adjacency matrix of the cortical mesh. The correlation between two different sources is thereby inversely proportional to the geodesic distance that separates them.

Multi-task regression As source imaging can be cast as estimating a regression model, source imaging for a set of subjects can be formulated as solving a set of coupled regression problems. In the statistical machine learning literature, such supervised learning problems are commonly referred to as *multi-task* prediction problems (Caruana, 1993). Notice that in this context, a *task* must be perceived as a *machine learning task*, not as a *cognitive* one. *Multi-task regression* is thus equivalent to *multi-subject regression*. In the M/EEG source imaging literature, to our knowledge, the only contribution formulating the problem as a multi-task regression model employs a Group Lasso with an ℓ_{21} block sparse norm (Lim et al., 2017).

However, the Group Lasso promotes neural sources that are either active for all subjects, or for none of them, similarly to (Litvak and Friston, 2008).

Although, the assumption of perfectly overlapping functional activity across subjects can be justified when aiming for coarse localization results, it gets more unrealistic as we aim for fine spatial resolution in the order of millimeters. In the following section, we investigate several multi-task regression models that relax the aforementioned overlapping sources assumption and propose a new flexible regularized model defined through a combination of Wasserstein metrics and non-convex ℓ_q $(0 < q \leq 1)$ pseudo-norms. These priors formalize the two main assumptions of our model: (1) focal and strong activation foci are favored over weak distributed ones; (2) responses across subjects are spatially close if presented with the same cognitive stimulus. To promote spatial proximity, we minimize an optimal transport (Wasserstein) distance across subjects, hence the name of our method: Minimum Wasserstein Estimates (MWE$_q$). Our experiments show that the choice of q defining the ℓ_q is very important: lower values tend to promote sparser solutions and yield models easier to tune in practice (Strohmeier, Gramfort, and Haueisen, 2015).

Anatomical alignment and common source space In order to estimate the source amplitudes jointly for S different subjects, it is necessary to have a correspondence between their cortical source spaces. To do so, one needs some anatomical alignment procedure between the cortical surfaces of the different subjects. In this work we follow the methods implemented by FreeSurfer. The *morphing* procedure uses the sulci and gyri patterns which are matched in an auxiliary spherical inflated cortical surface (Fischl, Sereno, and Dale, 1999; Gramfort et al., 2013a). Defining a source space in a template anatomy – here fsaverage – it is possible to morph it to each individual subject. Doing so, the source spaces for all subjects have the same number of dipoles and have some spatial correspondence. The resulting leadfields $\mathbf{L}^{(1)}, \ldots, \mathbf{L}^{(S)}$ have therefore the same dimensions $(n \times p)$ with aligned columns; a given column maps to the same brain region across all subjects – note that each leadfield is computed using the anatomical head model of its corresponding individual subject. While this procedure constructs aligned leadfield operators, assuming that the obtained correspondence across subjects is also functional, does not necessarily hold (Robinson et al., 2014). Using optimal transport, our proposed model goes beyond this rigid one-to-one alignment and allows for some spatial flexibility of the activation foci. Before introducing our model, we remind the reader of the general framework of multi-task regression.

2 Joint multi-task regression

Jointly estimating the current density $\mathbf{x}^{(s)}$ of each subject s can be expressed as a multi-task regression problem where some coupling prior is assumed on $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(S)}$ through a penalty Ω:

$$\min_{\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(S)} \in \mathbb{R}^p} \frac{1}{2n} \sum_{s=1}^{S} \|\mathbf{y}^{(s)} - \mathbf{L}^{(s)}\mathbf{x}^{(s)}\|_2^2 + \Omega(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(S)}) \ . \tag{3.5}$$

2.1 Block-sparse models

Let $\mathbf{X} \in \mathbf{R}^{p \times S}$ denote the horizontal stacking of the individual source vectors $\mathbf{X} = [\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(S)}]$. Mixed norms are defined on matrices as combinations of ℓ_p norms applied on their columns. For $a, b \geq 1$, the ℓ_{ab} norm is defined as:

$$\ell_{ab}(\mathbf{X}) = \left(\sum_{j=1}^{p} \left(\sum_{s=1}^{S} \mathbf{X}_{js}^{a} \right)^{\frac{1}{a}} \right)^{\frac{1}{b}} \quad . \tag{3.6}$$

When setting $b = 1$, the resulting mixed norm applies an ℓ_1 norm to the p ℓ_a norms of the columns of the matrix. Thus, minimizing such a penalty would promote sparsity of the *entire* columns of that matrix. This is particularly interesting when the desired prior should reflect some *structured* sparsity. In the context of source imaging, this corresponds to a strict consensus across subjects to decide whether a source is active or not. One of the most simple examples of such norms is the Group-Lasso penalty.

The Group-Lasso The Group Lasso (Yuan and Lin, 2006; Lim et al., 2017) is defined using the ℓ_{21} norm. In practice, using some hyper-parameter $\mu > 0$, it reads:

$$\min_{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(S)} \in \mathbb{R}^p} \frac{1}{2n} \sum_{s=1}^{S} \|\mathbf{y}^{(s)} - \mathbf{L}^{(s)} \mathbf{x}^{(s)}\|_2^2 + \mu \ell_{21}(\mathbf{X}) \quad . \tag{3.7}$$

The double sum penalty can be seen as an ℓ_1 penalty applied to a vector of ℓ_2 norms taken across subjects: only some ℓ_2 norms are non-zero. Therefore, a source is canceled out for all subjects or for none of them. In general multi-task learning settings, this is relevant for settings where all tasks are explained by the exact same relevant features. In practice however, the degree to which the active features overlap across tasks is not known a priori. Negahban and Wainwright (2008) showed that in the simple case of two tasks, if the fraction of overlapping sources is less than $2/3$ then Group Lasso has a lower probability of correct support identification than a Lasso estimator solved independently for each task. Such a high bar is not likely to be met when data do not follow such a rigid structured sparsity simply because in real life, data are *dirty*. For Dirty data, we need "dirty models".

Dirty models The assumption of identical sources for all subjects is clearly not realistic, Dirty models (Jalali et al., 2010) relax this assumption by decomposing the source vector of each subject s into two parts: $\mathbf{x}^{(s)} = \overset{c}{\mathbf{X}}^{(s)} + \overset{d}{\mathbf{X}}^{(s)}$, where the support of $\overset{c}{\mathbf{X}}^{(s)}$ is common to all subjects and $\overset{d}{\mathbf{X}}^{(s)}$ is different for each one. Originally, Jalali et al. (2010) introduced Dirty models with an $\ell_{1\infty}$ block norm on $\overset{c}{\mathbf{X}}$. For the sake of convenience and comparison with the Group Lasso, we use the (practically equivalent) ℓ_{21} norm. This variant of Dirty models reads:

$$\min_{\overset{c}{\mathbf{X}}, \overset{d}{\mathbf{X}}} \in \mathbb{R}^{pS} \frac{1}{2n} \sum_{s=1}^{S} \|\mathbf{y}^{(s)} - \mathbf{L}^{(s)} \mathbf{x}^{(s)}\|_2^2 + \mu \|\overset{c}{\mathbf{X}}\|_{21} + \lambda \|\overset{d}{\mathbf{X}}\|_{11} \quad , \tag{3.8}$$

with tuning hyper-parameters $\mu, \lambda \geq 0$.

Since the ℓ_{11} norm induces un-structured sparsity on all sources, Dirty models can be seen as a middle ground between an independent Lasso and a Group Lasso. The advantage of Dirty models over the Group Lasso is that it is agnostic with respect to the degree of similarities across subjects. Moreover, when tuning the hyper-parameters μ and λ, a comparison with Group-Lasso is not needed since it is included in the "tuning path" of Dirty models. The following proposition provides sufficient conditions for falling back to Group Lasso or independant Lasso which guides the selection of hyper-parameter candidates. Figure 3.3 visualizes this proposition by showing the nature of the obtained solutions on a grid of hyper-parameters.

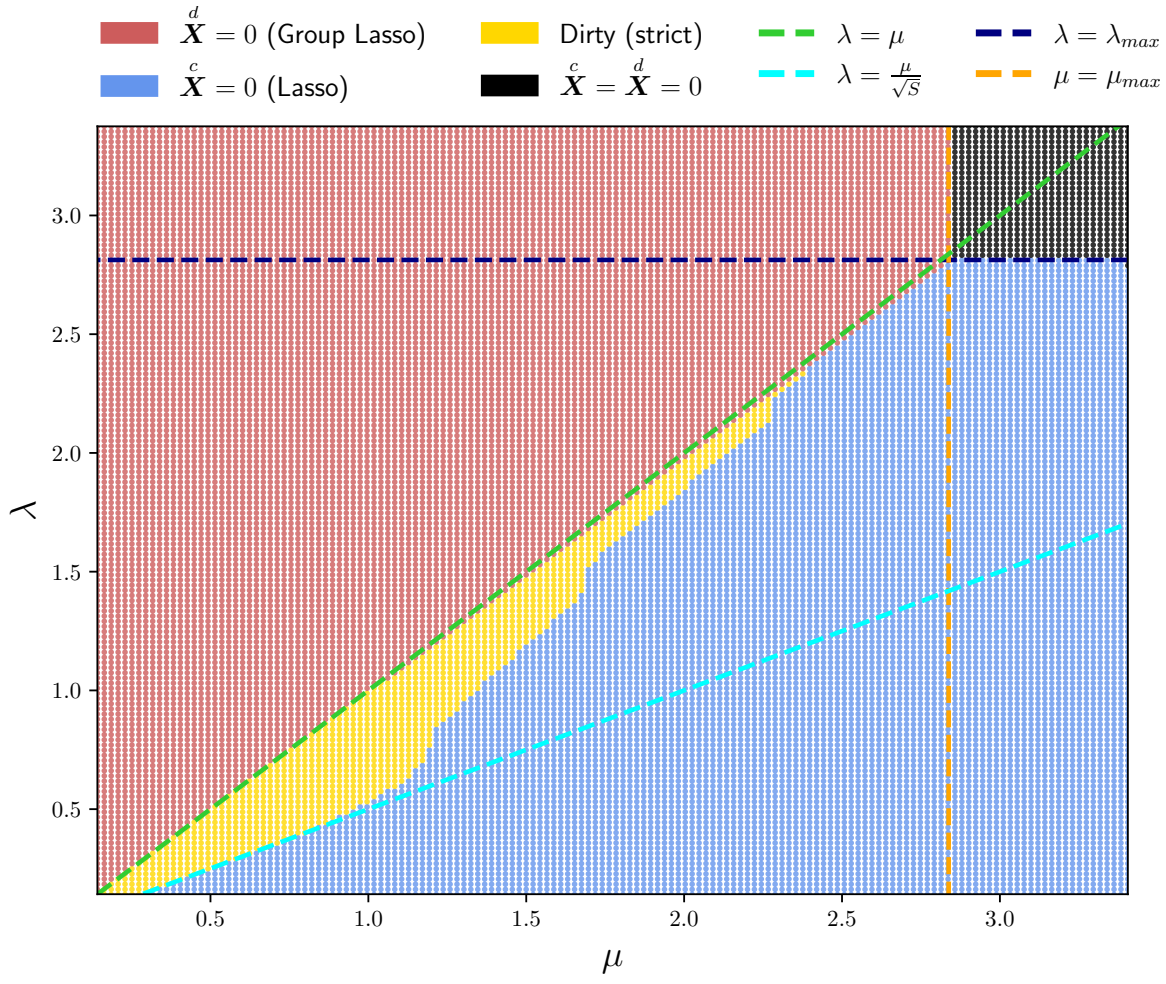


Fig. 3.3. Illustration of proposition 29. Dirty models do not need to be tuned over a full square grid: outside the highlighted slopes, it is equivalent to a Group Lasso or an independent Lasso. Experiment run with random Gaussian data with 4 tasks.

Proposition 29 *Let* $\mathbf{R} \overset{\text{def}}{=} [{\mathbf{L}^{(1)}}^\top \mathbf{b}^{(1)}, \dots, {\mathbf{L}^{(S)}}^\top \mathbf{b}^{(S)}] \in \mathbb{R}^{n,S}$ *and define the maximum hyper-parameter values* $\mu_{\max} \overset{\text{def}}{=} \frac{\|\mathbf{R}\|_{2\infty}}{n}$ *and* $\lambda_{\max} \overset{\text{def}}{=} \frac{\|\mathbf{R}\|_\infty}{n}$. *Let* $\mathbf{X}^\star = \overset{c}{\mathbf{X}}{}^\star + \overset{d}{\mathbf{X}}{}^\star$ *be a solution of Dirty models (3.8). Then the following holds:*

$$\mu \geq \sqrt{S}\lambda \Rightarrow \overset{c}{\mathbf{X}}{}^\star = 0 \tag{3.9}$$

$$\lambda \geq \mu \Rightarrow \overset{d}{\mathbf{X}}{}^\star = 0 \tag{3.10}$$

$$\lambda \geq \lambda_{\max} \text{ and } \mu \geq \mu_{\max} \Leftrightarrow \overset{d}{\mathbf{X}}{}^\star = \overset{c}{\mathbf{X}}{}^\star = 0 \ . \tag{3.11}$$

SKETCH OF PROOF. The proof of the three statements is derived using first order optimality conditions. The sub-differential sets of $\mu\ell_{21}$ and $\lambda\ell_{11}$ at 0 are given by the balls $\mu\mathcal{B}_{2\infty}$ and $\lambda\mathcal{B}_\infty$ respectively, leading to the quantities $\mu_{\max}$ and $\lambda_{\max}$. Moreover, any optimal dual variable must be included in the intersection of the aforementioned balls. Loosely speaking, when $\mu \geq \sqrt{S}\lambda$, it holds $\lambda\mathcal{B}_\infty \subset \mu\mathcal{B}_{2\infty}$, thus the $\mu\ell_{21}$ penalty becomes irrelevant for optimality: Dirty models are equivalent to a Lasso. Inversely, when $\mu \leq \lambda$, $\mu\mathcal{B}_{2\infty} \subset \lambda\mathcal{B}_\infty$, thus the $\lambda\ell_{11}$ penalty becomes irrelevant for optimality and Dirty models are then equivalent to a Group Lasso. Detailed technical derivations are provided in the appendix. ■

Multi-level Lasso It is noteworthy to keep in mind that structured sparsity can be obtained without necessarily invoking mixed norms. Perhaps a more simple and intuitive solution is to consider a multiplicative model for inverse solutions by writing them as a product of an *equal* vector $\mathbf{E}$ and a *different* matrix $\mathbf{D}$:

$$\mathbf{x}_j^{(s)} = \mathbf{e}_j \mathbf{D}_j^{(s)} \tag{3.12}$$

Lozano and Swirszcz (2012) proposed this model and applied an ℓ_1 penalty on both components $\mathbf{e} \in \mathbb{R}^p$ and $\mathbf{D} \in \mathbb{R}^{p,S}$. For the model to be identifiable, $\mathbf{e}$ is constrained to be non-negative. The *Multi-level Lasso* reads:

$$\min_{\mathbf{e}\in\mathbb{R}_+^p, \mathbf{D}\in\mathbb{R}^{p\times S}} \frac{1}{2n} \sum_{s=1}^{S} \|\mathbf{y}^{(s)} - \mathbf{L}^{(s)}(\mathbf{e} \odot \mathbf{E}^{(s)})\|^2 + \mu\|\mathbf{e}\|_1 + \lambda\|\mathbf{D}\|_{11} \ . \tag{3.13}$$

Due to the multiplicative formulation, the multi-level Lasso is not a convex problem. Moreover, it can be related to group-norm models highlighting a form a structured sparsity prior. As shown by Lozano and Swirszcz (2012), (3.13) is equivalent to a standard multi-task regression problem with the non-convex group pseudo norm $\ell_{1\frac{1}{2}}$ regularization:

$$\min_{\mathbf{X}\in\mathbb{R}^{p\times S}} \frac{1}{2n} \sum_{s=1}^{S} \|\mathbf{y}^{(s)} - \mathbf{L}^{(s)}\mathbf{X}^{(s)}\|^2 + 2\sqrt{\mu\lambda} \sum_{j=1}^{p} \sqrt{\|\mathbf{X}_j\|_1} \ . \tag{3.14}$$

Both Dirty models and Multi-level Lasso can be solved via alternating optimization where each update is carried out via proximal coordinate descent.

2.2 MWE: minimum Wasserstein estimates

Block-sparse models can be used to promote consistency across tasks. This consistency is however limited by a form of *pointwise* comparison of features. For source localization, these features are not arbitrarily ordered: their spatial structure matters. Our aim is to use OT metrics to compare source estimates by taking into account the geodesic distances between their locations. The application of OT in brain imaging is however not novel. Haufe et al. (2008) used the Wasserstein distance (a.k.a Earth Mover's Distance) as an evaluation metric to compare M/EEG source imaging methods on simulated experiments. Leveraging its fast entropic variant, Gramfort, Peyré, and Cuturi (2015) proposed to compute average brain patterns for fMRI and M/EEG group studies. Given these successes, one cannot help but wonder about the benefits of including OT in the model design not only as a pre- or post- processing step.

2.2.1 MWE$_1$: sparse entropic OT regularization

Consider the finite metric space $(\mathcal{X}, \|.\|_2)$ where each element of $X = \{1, \dots, p\}$ corresponds to a vertex of the source space. Let $\mathbf{C}$ be the matrix where $\mathbf{C}_{ij}$ corresponds to the geodesic distance $\|i - j\|$ between vertices i and j. The unbalanced entropic OT loss $\mathrm{UOT}^{\mathcal{U}}_{\varepsilon,\gamma}$ introduced in chapter 2 can be used to compare non-negative activation maps $\mathbf{a}, \mathbf{b} \in \mathbb{R}^p_+$:

$$\underset{\varepsilon,\gamma}{\overset{\mathcal{U}}{\mathrm{UOT}}}(\mathbf{a}, \mathbf{b}) \overset{\text{def}}{=} \min_{\pi \in \mathbb{R}_+{}^{p \times p}} \underbrace{\varepsilon \mathrm{KL}(\pi | e^{-\frac{\mathbf{C}}{\varepsilon}})}_{\text{transport cost + entropy}} + \underbrace{\gamma \mathrm{KL}(\pi \mathbb{1} | \mathbf{a}) + \gamma \mathrm{KL}(\pi^\top \mathbb{1} | \mathbf{b})}_{\text{relaxed marginal constraints}} \ . \tag{3.15}$$

A straightforward extension to signed measures proposed by several authors (Mainini, 2012; Profeta and Sturm, 2018) is to compare positive and negative parts separately:

$$\widetilde{\mathrm{UOT}}(\mathbf{a}, \mathbf{b}) \overset{\text{def}}{=} \underset{\varepsilon,\gamma}{\overset{\mathcal{U}}{\mathrm{UOT}}}(\mathbf{a}^+, \mathbf{b}^+) + \underset{\varepsilon,\gamma}{\overset{\mathcal{U}}{\mathrm{UOT}}}(\mathbf{a}^-, \mathbf{b}^-) \ , \tag{3.16}$$

where $\mathbf{x}^{(s)+} = \max(\mathbf{x}^{(s)+}, 0)$ and $\mathbf{x}^{(s)-} = \max(-\mathbf{x}^{(s)+}, 0)$ for any $\mathbf{x} \in \mathbb{R}^p$.

Minimizing $\widetilde{\mathrm{UOT}}$ across subjects would lead to both spatial and sign consistency across subjects. In the context of MEG, the sign of the source estimates indicates the polarity of the dipoles which is defined using the convention that positive currents are flowing out of the cortex (from deep cortical layers to superficial ones), while negative currents are flowing into the cortex (Gramfort et al., 2013c; Tadel et al., 2011). Along with a sparsity constraint, we propose to use as a coupling prior in (3.25):

$$\Omega_{\mathrm{MWE}_1}(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(S)}) \overset{\text{def}}{=} \underbrace{\lambda \sum_{s=1}^{S} \|\mathbf{x}^{(s)}\|_1}_{\text{Sparsity}} + \underbrace{\mu \min_{\bar{\boldsymbol{x}} \in \mathbb{R}^p} \frac{1}{S} \sum_{s=1}^{S} \widetilde{\mathrm{UOT}}(\mathbf{x}^{(s)}, \bar{\boldsymbol{x}})}_{\text{Spatial variance}} \ , \tag{3.17}$$

where $\mu, \lambda \geq 0$ are tuning hyperparameters. The minimized OT sum in (3.17) measures the average geodesic distance between all the $\mathbf{x}^{(s)}$ and their barycenter $\bar{\boldsymbol{x}}$. It can thus be seen as quantification of

the spatial variability of the source estimates. This spatial averaging is illustrated in Figure 3.4 where we considered two averaging methods of simulated brain patterns of 5 different subjects. When the brain patterns are close but non-overlapping, the usual (Euclidean) average leads to blurring and spatial distortions while the OT barycenter conserves the shape of its input.

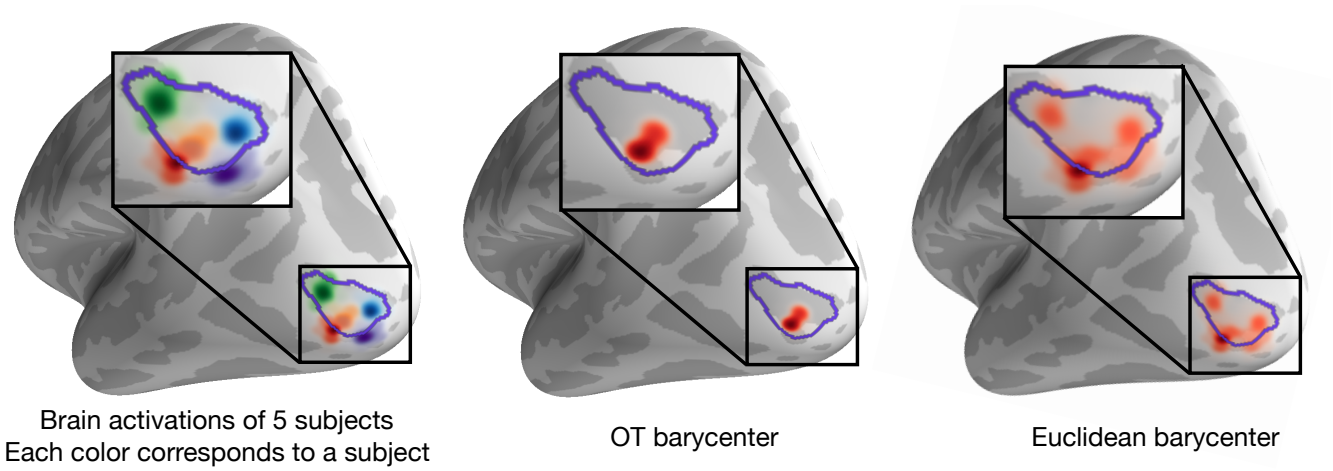


Fig. 3.4. Illustration of the UOT barycenter $\bar{\mathbf{x}}$ (middle) of 5 activations inputs $\mathbf{x}^{(s)}$ (left) with random amplitudes between 20 and 30 nAm in the *middle and occipital lunatus sulcus* defined by the *aparc.a2009s* segmentation. $\bar{\mathbf{x}}$ is located at the average location of the inputs with an average amplitude levels. The Euclidean barycenter (right) is the usual mean: it creates undesirable blurring. Promoting subject-level source estimates that are close to a population average for UOT promotes spatial proximity between all activation foci.

Solving MWE$_1$ Even though $\mathrm{UOT}^{\mathcal{U}}_{\varepsilon,\gamma}$ is differentiable and jointly convex, computing its gradient requires running a full Sinkhorn loop with a learning-rate that must be tuned. Instead, we would like to take advantage of both Sinkhorn's algorithm and the separability of the ℓ_1 norm for which proximal coordinate descent – at least in practice – outperforms proximal gradient descent methods. By combining (3.15), (3.16) and (3.17), we obtain an objective function taking as arguments:

$$\left((\mathbf{x}^{(s)+})_s, (\mathbf{x}^{(s)-})_s, (\pi^{(s)+})_s, (\pi^{(s)-})_s, \bar{\mathbf{x}}^+, \bar{\mathbf{x}}^-\right).$$

The only coupled terms in this loss function are the KL terms which is jointly convex. We propose to minimize it by alternating.

When minimizing with respect to one $\mathbf{x}^{(s)+}$ (or $\mathbf{x}^{(s)-}$), the resulting problem can be written (dropping the exponents for simplicity):

$$\min_{\mathbf{x}\in\mathbb{R}^p_+} \frac{1}{2n}\|\mathbf{y}-\mathbf{L}\mathbf{x}\|_2^2 + \frac{\mu\gamma}{S}\mathrm{KL}(\mathbf{m}|\mathbf{x}) + \lambda\|\mathbf{x}\|_1 \ , \tag{3.18}$$

Algorithm 7 MWE_1 algorithm (Janati et al., 2019)

Input: $\mu, \varepsilon, \gamma, \lambda$ and cost matrix $\mathbf{C}$. data $(\mathbf{y}^{(s)})_s (\mathbf{L}^{(s)})_s$.
Output: Solutions $(\mathbf{x}^{(s)})$ of MWE_1
repeat
  for $s = 1$ **to** S **do**
    Update $\mathbf{x}^{(s)+}$ with proximal coordinate descent to solve (3.18).
    Update $\mathbf{x}^{(s)-}$ with proximal coordinate descent to solve (3.18).
  end for
  Update left marginals $\mathbf{m}^{(1)+}, \dots, \mathbf{m}^{(S)+}$ and the barycenter $\bar{\boldsymbol{x}}^{+}$ with Sinkhorn's algorithm.
  Update left marginals $\mathbf{m}^{(1)-} \dots, \mathbf{m}^{(S)-}$ and the barycenter $\bar{\boldsymbol{x}}^{-}$ with Sinkhorn's algorithm.
until convergence

where $\mathbf{m} \overset{\text{def}}{=} \pi \mathbb{1}$. In addition to the data fidelity and sparsity terms, the group prior is individually applied to each subject through the KL term. The individual source estimates are pushed towards the marginals of the transport plan linking them to the barycenter $\bar{\mathbf{x}}$. The following proposition shows that problem (3.18) can be solved using proximal coordinate descent (Richtárik and Takáč, 2014).

Proposition 30 *Problem (3.18) is equivalent to the non-negative regularized regression problem with a separable penalty g:*

$$\min_{\mathbf{x} \in \mathbb{R}_+^p} \frac{1}{2n} \|\mathbf{y} - \mathbf{L}\mathbf{x}\|_2^2 + \sum_{j=1}^{p} g_j(\mathbf{x}_j) \ , \tag{3.19}$$

where $g_j : x \in \mathbb{R}_+ \mapsto -\mathbf{a}_j \log(x) + bx \in \mathbb{R}$ with $\mathbf{a} \overset{\text{def}}{=} \frac{\mu\gamma}{S}\mathbf{m}$ and $b = \lambda + \frac{\mu\gamma}{S}$. Moreover, g_j is convex and its proximal operator is given by:

$$\text{prox}_{g_j}(y) = \frac{1}{2}\left[-b + y + \sqrt{(b-y)^2 + 4\mathbf{a}_j}\right] \ . \tag{3.20}$$

PROOF. Proof from (Janati, Cuturi, and Gramfort, 2019). The definition of $\text{KL}(\mathbf{m}, \mathbf{x}) = \sum_{j=1}^{p} \mathbf{m}_j(\log(\mathbf{m}_j) - \log(\mathbf{x}_j)) + \mathbf{x}_j - \mathbf{m}_j$ along with the positivity constraint on $\mathbf{x}$ leads to the equivalent formulation (3.19). KL and the ℓ_1 norm are convex, thus g is convex. Its proximal operator is defined as:

$$\text{prox}_{g_j}(y) = \arg\min_{z>0} \frac{1}{2}\|z - y\|^2 + g_j(z) \tag{3.21}$$

$$= \arg\min_{z>0} \frac{1}{2}\|z - y\|^2 - \mathbf{a}_j \log(z) + bz \tag{3.22}$$

The first order optimality condition leads to a second order equation in z. Keeping the non-negative solution ends the proof. ■

Remark 7 *The* $-\log$ *penalty seems to act as a barrier and forbid the sources to be sparse. This is true when* $\mathbf{a}_j > 0$*, which is desired to promote similarity towards the barycenter. However, the individual entries of* $\mathbf{a}$ *are not necessarily positive. In the event of* $\mathbf{a}_j = 0$*,* g *is equivalent to a non-negative* ℓ_1 *norm. Moreover, even when* $\mathbf{a}_j = 0$*, the proximal operator* (3.20) *is still valid as it coincides with that of a non-negative lasso:*

$$\mathbf{a}_j = 0 \Rightarrow \operatorname{prox}_{g_j}(y) = \frac{1}{2}\left[-b + y + |b - y|\right] \tag{3.23}$$

$$= \frac{1}{2}\max(y - b, 0) \tag{3.24}$$

The second update with respect to $((\mathbf{P}^{(s)+})_s, (\mathbf{P}^{(s)-})_s, \bar{\boldsymbol{x}}^+, \bar{\boldsymbol{x}}^-)$ can be cast as two UOT barycenter problems, carried out using generalized Sinkhorn iterations (Chizat et al., 2018b) (a.k.a IBP algorithm). Note that one does not need to compute the transport plans $\pi^{(s)}$ since inferring every source estimate $\mathbf{x}$ only requires the knowledge of the left marginal $\mathbf{m} = \mathbf{P}\mathbb{1}$ which does not require storing $\mathbf{P}$ in memory. Moreover, while coordinate descent iterations are linear in the number of sources p, optimal transport iterations are linear in the number of subjects S and quadratic in the number of sources p. However, the algorithm can be significantly sped up using (1) GPUs for optimal transport iterations; (2) removing sparse sources from the computing of the OT barycenter; (3) warm-start within the inner alternating operations of the convex subproblem MWE_1. In all our experiments, we set the initial source estimates to $\mathbf{x}_+ = \mathbf{x}_- = \mathbb{1}/p$.

2.2.2 Concomitant MWE_1: adaptive noise level normalization

One of the drawbacks of MWE_1 is that the ℓ_1 hyper-parameter λ is common to all subjects. This implicitly assumes that the level of noise is the same across subjects. Following the work of Ndiaye et al. (2017) and Massias et al. (2018) on the smoothed concomitant Lasso, we propose to extend MWE by inferring the specific noise standard deviation $\sigma^{(s)}$ along with the regression coefficient $\mathbf{x}^{(s)}$ of each subject.

Concomitant estimation Inferring both the sources and their standard deviations in linear regression models can be via maximization of a penalized maximum likelihood of a joint distribution. Even though such a problem can be made convex through a change of variable (Städler, Bühlmann, and Geer, 2010), it does not fit within the usual "smooth + proximable" framework. Following the seminal work of Huber (1981) on robust estimation, Owen (2007) proposed a *concomitant* problem for the Lasso problem with the constraint $\sigma > 0$, also known as Scaled Lasso for which proximal coordinate descent methods are still applicable. The positivity constraint is however problematic for primal dual methods that require dividing by σ. Ndiaye et al. (2017) proposed to overcome this issue by adding a "smoothing" constraint so that minimization is performed over a closed set. Similarly, we define concomitant MWE_1 as:

$$\min_{\substack{\mathbf{x}^{(1)},\dots,\mathbf{x}^{(S)} \in \mathbb{R}^p \\ \sigma^{(1)},\dots,\sigma^{(S)} \in [\sigma_0, +\infty]}} \sum_{s=1}^{S} \frac{1}{2n\sigma^{(s)}} \|\mathbf{y}^{(s)} - \mathbf{L}^{(s)}\mathbf{x}^{(s)}\|_2^2 + \frac{\sigma^{(s)}}{2} + \Omega_{\text{MWE}_1}(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(S)}) \ , \tag{3.25}$$

where σ_0 is a pre-defined constant. In practice σ_0 can be set as a small fraction of the initial estimate of the standard deviation $\sigma_0 = \alpha \min_s \frac{\|\mathbf{y}^{(s)}\|}{\sqrt{n}}$. In our experiments we set $\alpha = 0.01$, making sure that it does not affect the solutions by checking that the estimated $\hat{\sigma}^{(s)}$ are largely superior to σ_0.

Solving (3.25) Adding the estimation of the standard deviations $\sigma^{(s)}$ does not change the convexity of the problem. Adding an update with respect to σ leads to the iteration:

$$\sigma^{(s)} \leftarrow \frac{\|\mathbf{Y}^{(s)} - \mathbf{L}^{(s)}\mathbf{x}^{(s)}\|_2}{\sqrt{n}} \wedge \sigma_0 \ . \tag{3.26}$$

The only modification to the update of the sources $\mathbf{x}^{(s)}$ is the change of the hyper-parameters μ and λ which finds themselves being multiplied by $\sigma^{(s)}$.

2.2.3 Concomitant MWE$_{0.5}$: fighting entropic blur

Given our lengthy discussions over debiasing entropic OT in chapter 2, it would make sense to replace $\mathrm{UOT}^{\mathcal{U}}_{\varepsilon,\gamma}$ with its debiased loss function $\mathrm{S}^{\mathcal{U}}_{\varepsilon,\gamma}$ for which we can compute barycenters via Sinkhorn's algorithm. This alternative would conceptually lead to less blurry barycenters. In practice however, minimizing with respect to the sources $\mathbf{x}^{(s)}$ could not be done via proximal coordinate descent anymore. Instead, we opt for substituting the ℓ_1 penalty with the more sparsity enhancing penalty $\ell_{0.5}$.

Non-convex separable penalties and re-weighted algorithms In the machine learning community, using non-convex penalties as better proxy of the exact sparsity norm ℓ_0 is not new. The particular case of $\ell_{0.5}$ is known as adaptive Lasso or re-weighted Lasso (Candès, Wakin, and Boyd, 2008). As long as the non-convex penalty is separable, Gasso, Rakotomamonjy, and Canu (2009) showed that the resulting problem can be solved via multiple convex sub-problems. Let $L : \mathbb{R}^p \to \mathbb{R}$ and $g : \mathbb{R}_+ \to \mathbb{R}$ be a convex and non-convex functions respectively. Assume for the sake of simplicity that g is differentiable. Then the problem

$$\min_{x \in \mathbb{R}^p} L(x) + \sum_{j=1}^{p} g(|\mathbf{x}_j|) \ , \tag{3.27}$$

can be solved via the sequence of weighted Lasso problems starting with $\mathbf{w} = \mathbb{1}$:

$$\mathbf{x}^{(k+1)} \leftarrow \operatorname*{arg\,min}_{x \in \mathbb{R}^p} L(x) + \|\mathbf{w}^{(k)} \odot \mathbf{x}\|_1 \tag{3.28}$$

$$\mathbf{w}_j^{(k+1)} \leftarrow g'(\mathbf{x}_j^{(k+1)}) \quad \text{for all } j \tag{3.29}$$

This iterative reweighting scheme can be be seen as a majorization-minimization procedure where g is majorized by its local linear approximation which is then minimized as a convex surrogate function.

Algorithm 8 Reweighted $\text{MWE}_{0.5}$ (Janati et al., 2020b)

Initialize weights $\mathbf{w}^{(s)} = \mathbb{1}$ for $s = 1 \dots S$
repeat
  Solve MWE_1 (Algorithm 7) with the weighted ℓ_1 norms $\|\mathbf{w}^{(s)} \odot \mathbf{x}^{(s)}\|_1$
  $\mathbf{w}_j^{(s)} = \frac{1}{2\sqrt{|\mathbf{x}_j^{(s)}|}}$ for all s, j
until convergence

Reweighted concomitant MWE Replacing ℓ_1 with $\ell_{0.5}$ in our MWE estimator is motivated by several factors:

1. Promote sparser estimates to fight back against the entropic blur induce by UOT
2. Its established de-biasing of the amplitudes of the sources and improved support identification in the context of source localization (Strohmeier et al., 2016)
3. It can be solved as a sequence of MWE_1 sub-problems.

Formally, the concomitant $\text{MWE}_{0.5}$ problem reads:

$$\min_{\substack{\mathbf{x}^{(1)},\dots,\mathbf{x}^{(S)} \in \mathbb{R}^p \\ \sigma^{(1)},\dots,\sigma^{(S)} \in [\sigma_0, +\infty]}} \sum_{s=1}^{S} \frac{1}{2n\sigma^{(s)}} \|\mathbf{y}^{(s)} - \mathbf{L}^{(s)}\mathbf{x}^{(s)}\|_2^2 + \frac{\sigma^{(s)}}{2} + \Omega_{\text{MWE}_{0.5}}(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(S)}) \ , \tag{3.30}$$

where:

$$\Omega_{\text{MWE}_{0.5}}(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(S)}) \stackrel{\text{def}}{=} \lambda \underbrace{\|\mathbf{x}^{(s)}\|_{0.5}}_{\text{Sparsity}} + \underbrace{\mu \min_{\bar{\mathbf{x}} \in \mathbb{R}^p} \frac{1}{S} \sum_{s=1}^{S} \widetilde{\text{UOT}}(\mathbf{x}^{(s)}, \bar{\mathbf{x}})}_{\text{Spatial variance}} \ . \tag{3.31}$$

Computing the update rule for $g : x \mapsto \sqrt{x}$ leads to Algorithm 8. In practice however, when some $\mathbf{x}_j^{(s)} = 0$, the majorization step will cause an overflow error in the weights $\mathbf{w}$. In practice, one can simply filter out the corresponding null features or set $\mathbf{w}_j^{(s)} = \frac{1}{2\sqrt{|\mathbf{x}_j^{(s)}| + \eta}}$ where η is a small value as proposed by Gasso, Rakotomamonjy, and Canu (2009). We adopt this strategy and set $\eta = 10^{-6}$ in all our experiments.

Hyperparameters of UOT The parameters defining the Wasserstein distance $\widetilde{\text{UOT}}$ are ε (entropy regularization) and γ (marginal relaxation). Large values of ε accelerate the convergence of the Sinkhorn algorithm but induce an undesired blurring of the source estimates. Very Low values however lead to numerical instability. We set ε to 0.002 divided by the median of the ground metric $\mathbf{C}$ which provides a good trade-off between computation speed and sharpness of the barycenter. With the same reasoning, low values of γ allow for a "free" transport, thus the barycenter converges towards a blurred uniform

distribution. In chapter 2, the closed form for unbalanced OT between Gaussians indicates that γ should scale as the distance between their means $\|\mathbf{a} - \mathbf{b}\|^2$ for the transported mass not to vanish. Thus, γ should be set proportionally to the values of $\mathbf{C}$. In all our experiments, we set $\gamma = \|\mathbf{C}\|_\infty$.

3 Experiments

We compared all the multi-task regression models discussed so far on MEG source localization problems for both simulations and real data. We believe the results we obtained are as important as how we obtained them. Well documented software matters.

3.1 Software

Benchmarking sparse multi-task regression problems required implementing all the aforementioned methods. For the sake of reproducibility, we decided to share our Python implementation of several multi-task regression models for both general machine learning applications as well as EEG/MEG source localization.

`mutar`: Multi-task regression in Python The `MuTaR` library follows the `scikit-learn` API and provides solvers for the following models:

1. Independent Lasso
2. Independent Reweighted Lasso
3. Group Lasso
4. Dirty Models
5. Multi-level Lasso
6. (Concomitant) MWE$_q$ for $q \in \{0.5, 1\}$.

The `MuTaR` webpage provides several illustrating examples (Fig 3.5) and can be accessed via the link:

hichamjanati.github.io/mutar/

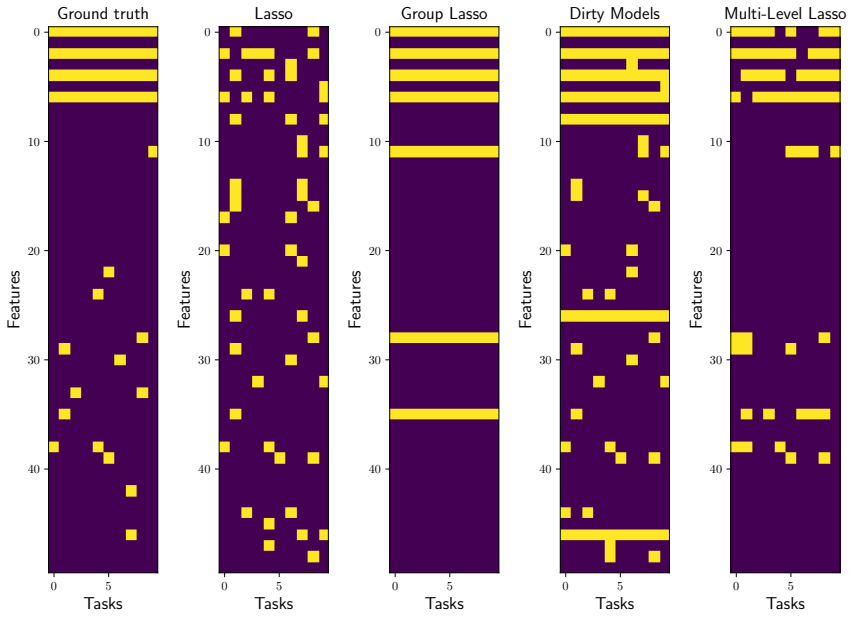


Fig. 3.5. Example taken from MuTaR's documentation displaying the supports of inverse solutions with random Gaussian data.

`groupmne`: multi-subject EEG / MEG source localization `MNE-Python` is undoubtedly the reference for EEG and MEG analysis in Python. Starting from raw MEG and EEG data, it provides users with simple functions from computing the source space and the leadfield operator to solving the inverse problem with ℓ_2, ℓ_1, $\ell_{0.5}$ and several other priors. `GroupMNE` is a MNE-python derivative project that provides functions to prepare the leadfields for muti-subject source localization and calls `MuTaR` to solve the inverse problem. The `groupMNE` webpage is available at:

hichamjanati.github.io/groupmne/

3.2 Results

To the best of our knowledge, only the Group Lasso model was previously used for multi-subject source localization with EEG/MEG (Lim et al., 2017). The extent to which multi-task regression improves source localization is thus still unknown. Is there a limiting number of subjects for which performance stops improving ? Could subject variability degrade that performance ? Using simulated MEG data, we attempt

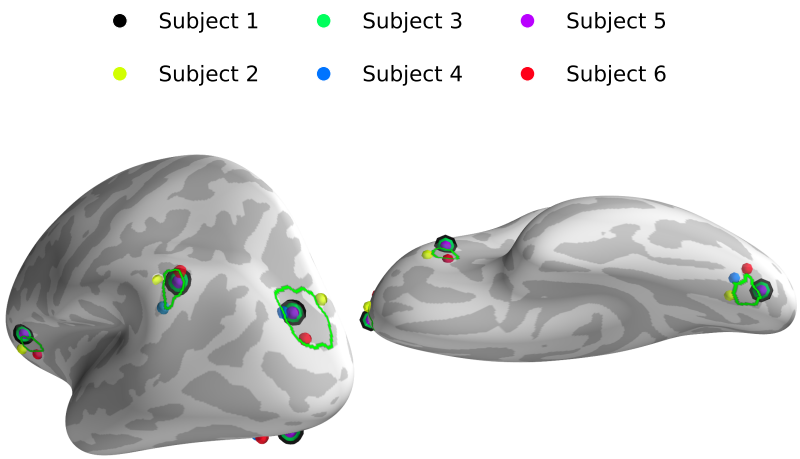


Fig. 3.6. Example of a simulated source configuration with 5 activations for $S = 6$ subjects - one activation per label. The 5 labels – highlighted within green borders – are taken from the aparc.a2009s FreeSurfer Destrieux parcellation (Destrieux et al., 2010). Different radii are used to distinguish overlapping sources. Here, subjects 1, 3 and 5 share the exact same source locations.

to provide answers to these questions by first exploring "the best case scenario" for each model. Is multi-task regression worth it ?

3.2.1 Simulations with semi-real data

By semi-real data we mean that we simulate MEG data $\mathbf{y}$ with real leadfield matrices $\mathbf{L}$ extracted from the public Cam-CAN dataset (Taylor et al., 2017). We use the MRI scan of each subject to compute a source space and its associated leadfield comprising 2562 sources per hemisphere (Gramfort et al., 2013a). Keeping only MEG gradiometer channels, we have $n = 204$ observations per subject. To keep the simulation settings simple, we restrict all leadfields to the left hemisphere. We thus have $S = 32$ leadfields with $p = 2562$. We simulate an inverse solution $\mathbf{x}^s$ with 5 sources (5-sparse vector) by randomly selecting one source per label (a.k.a. region of interest) among 5 pre-defined labels using the *aparc.a2009s* parcellation of the Destrieux atlas (Destrieux et al., 2010). To model functional consistency, 50% of the subjects share sources at the same locations, the remaining 50% have sources randomly generated in the same labels (see Figure 3.6 for an example). Their amplitudes are taken uniformly between 20 and 30 nAm. Their sign is taken at random with a Bernoulli distribution (0.5) for each label (all subjects share the same polarity of currents in a given label). We simulate $\mathbf{y}$ using the real forward model of each subject

with a covariance matrix σI_n. We set σ so as to have an average signal-to-noise ratio across subjects equal to 4 ($\text{SNR} \overset{\text{def}}{=} \sum_{s=1}^{S} \frac{\|\mathbf{L}^{(s)}\mathbf{x}^{(s)}\|}{S\sigma}$).

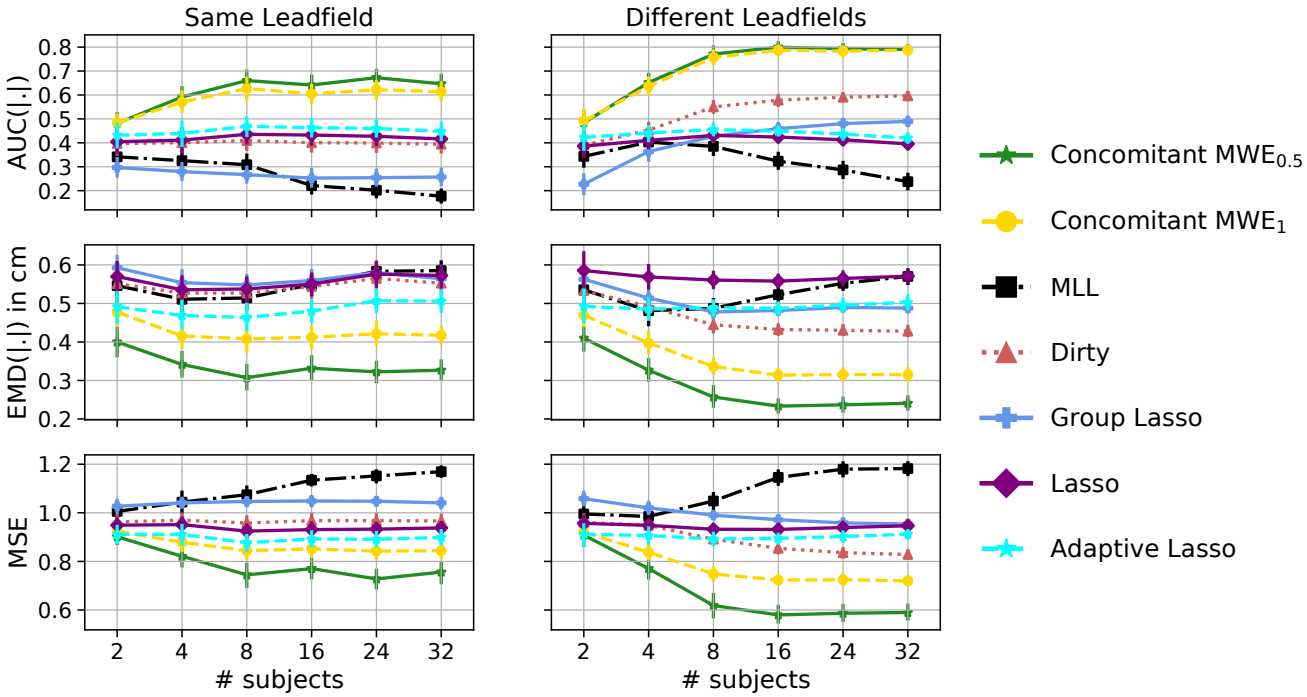


Fig. 3.7. Performance of different models over 30 trials in terms of PR-AUC, EMD and MSE. Each simulation trial uses a different source configuration (5 sources) and noise. The leadfields were derived from the MRI scans of the Cam-CAN dataset. For the same leadfield column, the shared leadfield across subjects is randomly picked for each run.

Performance evaluation We evaluate the performance of all models knowing the ground truth by comparing the best estimates on a grid of hyperparameters in terms of three metrics: the mean squared error (MSE) to quantify accuracy in amplitude estimation, area under the curve (AUC) of the precision-recall curve (PR), and a generalized Earth mover distance (EMD) to assess supports estimation, as done by Haufe et al. (2008). We use the PR-AUC computed between the absolute values of the coefficients and the true supports. Similarly, the EMD is computed between normalized absolute values of sources. Since $\mathbf{C}$ is expressed in centimeters, EMD can be seen as an expectation of the geodesic distance in millimeters between the truth and the source estimates. For a better intuitive interpretation of the EMD, we compute the EMD per source, i.e we divide it by 5. When increasing the number of subjects, the sets of used leadfields are increasing (the list of leadfields is ordered). For the "same leadfields" condition, we randomly pick a leadfield for all subjects for each run. We perform 30 different trials (with different true activations and noise) and report the mean within a 95% confidence interval in Figure 3.7.

Simulation results Various observations can be made. The Group Lasso (Lim et al., 2017) performs poorly – even compared to independent Lasso – which is expected since simulated sources are not perfectly overlapping for all subjects. This result is supported by theoretical evidence. As mentioned in the introduction, for the simple case of 2 subjects, one can show that if the fraction of overlapping sources is less than 2/3, Group Lasso performs worse than independent Lasso (Negahban and Wainwright, 2008). MWE$_q$ however benefits from the presence of more subjects by leveraging spatial proximity. The mean AUC increases from 0.4 (Lasso) to 0.8. The average error EMD distance is reduced from 6 mm (Lasso) to nearly 2 mm. Finally, even if both MWE$_q$ models show a similar AUC score, the proposed reweighting allows MWE$_{0.5}$ to outperform MWE$_1$ by a significant margin in terms of amplitude estimation as quantified by MSE. Finally, by inducing more sparsity, the $\ell_{0.5}$ norm of MWE$_{0.5}$ reduces the number of false positives which are located far from the true sources, thereby reducing the EMD distance by 1 mm compared to MWE$_1$. The case of Multi-level Lasso is more difficult to assess. When examining the obtained solutions, we find that it most often than not has several false negatives i.e it promote solutions that are *too sparse*, specially as the number of subjects increase. This is perhaps due to its non-convex formulation: once a feature is 0, it is removed for all subjects. Further analysis of its initialization and hyper-parameter tuning should be pursued. Finally, we can appreciate the improvement of multi-task models when increasing the number subjects, especially when using different leadfield matrices. We argue that this improvement is the consequence of the different folding patterns of the cortex across subject. Indeed, these folding differences lead to different dipole orientations of the same source across subjects, thereby increasing the chances of an accurate localization.

3.2.2 Experiments on MEG data

Datasets description We use two publicly available MEG datasets: DS117 (Wakeman and Henson, 2015) and Cam-CAN (Taylor et al., 2017). DS117 provides MRI, MEG, EEG and fMRI data of 16 healthy subjects to whom were presented images of famous, unfamiliar and scrambled faces. The fusiform face area (FFA) which specializes in facial recognition activates around 170 ms after stimulus (Henson et al., 2011; Kanwisher, McDermott, and Chun, 1997a). We pick the time point in the contrast response *famous* vs *scrambled* with the peak response for each subject within the interval 150-200 ms after stimulus. Similarly, Cam-CAN provides MEG, EEG and MRI data of around 650 healthy subjects with several types of tasks. We select the youngest 32 subjects (aged between 18 years and 29 years) and use their MEG recordings to study the auditory N100 response. We average the responses of 3 stimuli: 300Hz, 600Hz and 1200Hz with a total of 60 trials. We pick the time point with the peak response within 80-120 ms after stimulus. For both datasets, the leadfield operator of each subject was obtained from their T1 MRI scan using a cortically constrained source space formed by about 2500 candidate dipoles per hemisphere.

Model selection For all lasso-type models, there exists $\lambda_{\max}$ such that for $\lambda \geq \lambda_{\max}$ the inverse solution is 0 everywhere. For instance, with ℓ_1 and $\ell_{0.5}$ we have $\lambda_{\max} = \frac{\|\mathbf{L}^\top \mathbf{y}\|_\infty}{n}$ (Rakotomamonjy, Gasso, and Salmon, 2019). This allows to set λ in a relative scale between 0 and 1, making this choice less sensitive to the data. In practice, one can pick a certain value in [0, 1] based on the number of active sources, which is the heuristic used in the following experiments with real data. Even though the choice of $\lambda_{\max}$ does not

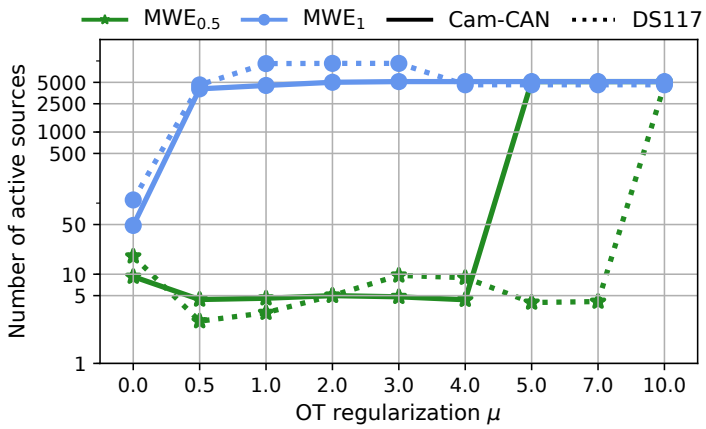


Fig. 3.8. Number of active sources for MWE models with $\lambda = 30\%$. Color encodes the different models; line-style encodes the different datasets. The mean is reported across all subjects. With MWE$_{0.5}$, a similar phase transition occurs for both datasets after a certain $\mu_{\max}$.

theoretically guarantee null source estimates with MWE$_{0.5}$, we observe experimentally that reweighting and the OT regularizer promote even more sparsity with a lower λ compared to Lasso models. We use the same relative scaling to set λ for MWE$_{0.5}$. The OT regularization parameter μ controls the level of consistency across subjects. Figure 3.8 shows that for the reweighted MWE$_{0.5}$, there exists a phase transition at a certain value $\mu_{\max}$, after which the source estimates lose all sparsity and cover the entire cortical mantle uniformly. MWE$_1$ however shrinks the source estimates towards 0 but fails to produce sparse solutions. In practice, based on the complexity of the topographic maps of the MEG data, we select λ and μ that lead to – on average – a 2-sparse solution with Cam-CAN ($\lambda = 30\%, \mu = 3$) and a 6-sparse solution with DS117 ($\lambda = 20\%, \mu = 0.5$).

MWE for population imaging The standard approach to obtain the source estimates from a group of subjects is to average the estimates obtained independently for each subject. Euclidean averaging however induces undesired blurring and sparsity is lost even when the individual solutions are sparse. Figure 3.10 shows that MWE$_{0.5}$ prevents that from happening. Moreover, the latent variable $\bar{x}$ of MWE$_{0.5}$ is sharper and more informative at a population level. To compare with single-subject solvers, we compute MCE and reweighted MCE solutions by selecting independently for each subject a λ such that the solution is 2-sparse (resp. 6-sparse) for Cam-CAN (resp. DS117). For dSPM, we use the default hyperparameter value $1/\text{SNR}^2$ with $\text{SNR} = 3$. For multi-subject models, we compare with the hierarchical Bayesian model GALA (Kozunov and Ossadtchi, 2015). GALA is a Bayesian model with a multivariate Gaussian prior with covariance matrices defined such that neighboring vertices are highly correlated. GALA has however a certain degree of flexibility as it models both similar and specific types of activation across subjects. The hyperparameters setting this similarity-specificity trade-off are inferred from the data, which is one of the

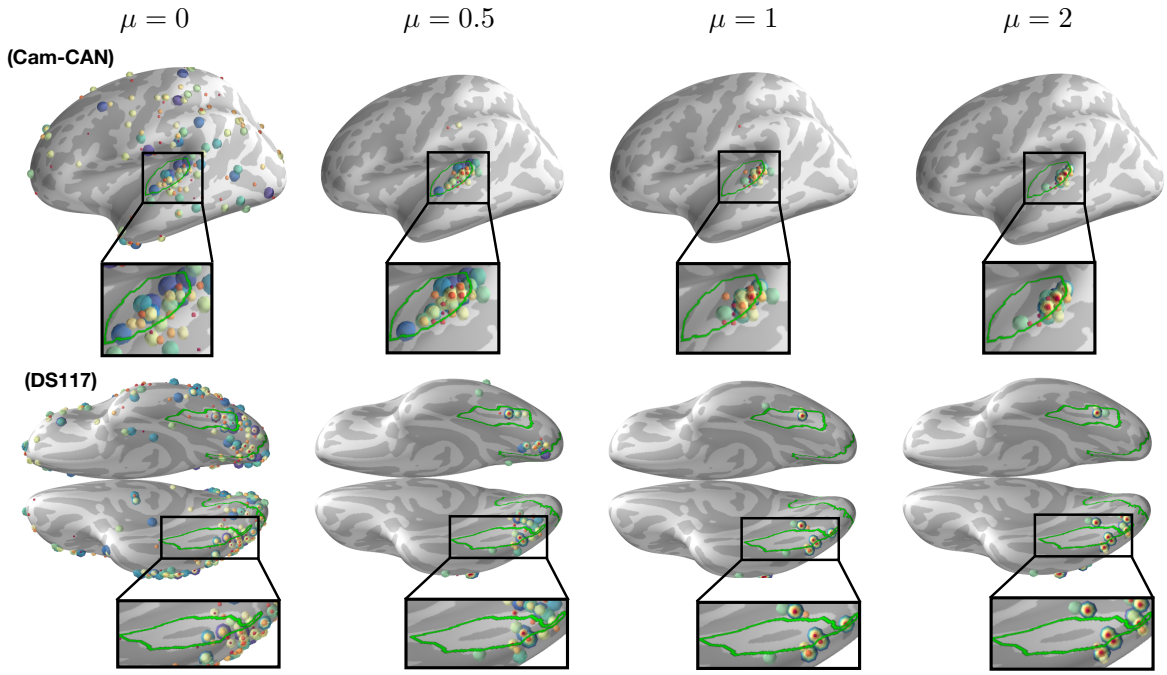


Fig. 3.9. Support of source estimates of $\text{MWE}_{0.5}$ recovered in the auditory task of Cam-CAN with 32 subjects (top) and the visual task of DS117 with 16 subjects (bottom). Each color corresponds to a subject. Different radii are displayed for a better distinction of sources. Increasing μ with $\mu < \mu_{\text{max}}$ promotes functional consistency across subjects. Top: Cam-CAN dataset ($\lambda = 30\%$). Bottom: DS117 dataset ($\lambda = 20\%$).

main features of GALA. In our experiments, we used the default fixed covariance matrices in the code kindly provided by the authors of GALA. The green borders highlight regions of interest. For Cam-CAN, we use the *neurosynth* (Yarkoni, 2014) label corresponding to the *auditory cortex* thresholded at 15 and projected on the surface of the temporal lobe. For DS117, we rely on the *aparc a2009s* segmentation to show both the fusiform gyrus and the primary visual cortex V1. With Cam-CAN, the Euclidean average of the obtained minimum Wasserstein estimates is focal, located right in the auditory cortex. However, The average Lasso and dSPM estimates are dispersed around the auditory cortex with a substantial blurring due to averaging. The visual task of DS117 appears to be the most challenging for several reasons which explain the low amplitude sources. These reasons are discussed in detail at the end of this chapter.

Comparison with fMRI The EEG/MEG inverse problem has an infinite number of solutions. We proposed to regularize it in two ways: (1) at a subject level by favoring focal sources; (2) at a population level by promoting spatial proximity between activation foci. However, one could argue that $\text{MWE}_{0.5}$ promotes consistency at the expense of proper fitting of individual data. To address this concern we compute the standardized fMRI Z-score of the conditions *famous vs scrambled faces*. We compare minimum

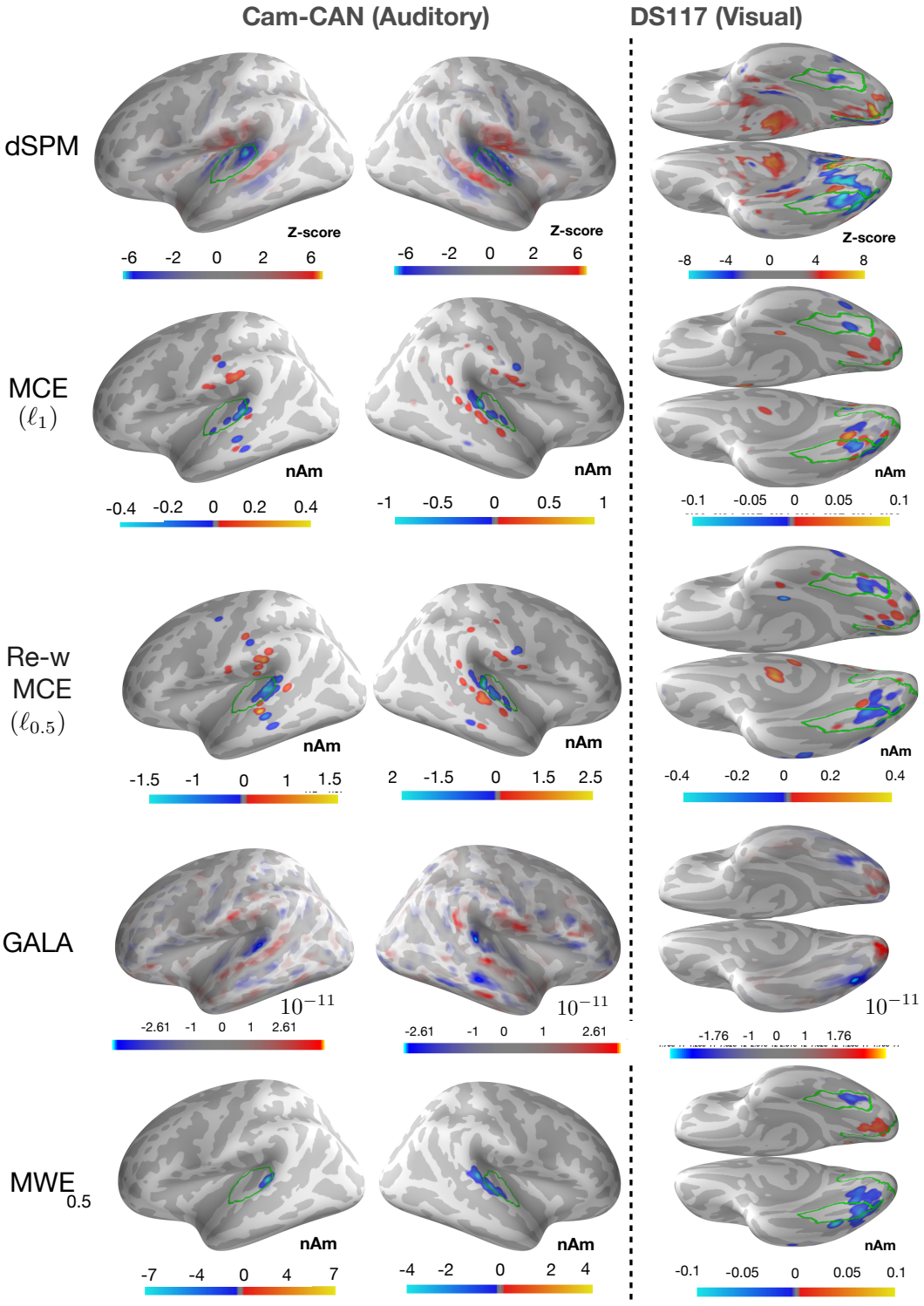


Fig. 3.10. Average source estimates of different solvers. **Left:** Cam-CAN dataset. **Right:** DS117 dataset. MWE reduces blurring by promoting functional consistency. No thresholding was applied on the source estimates except for the dSPM Z-scores.

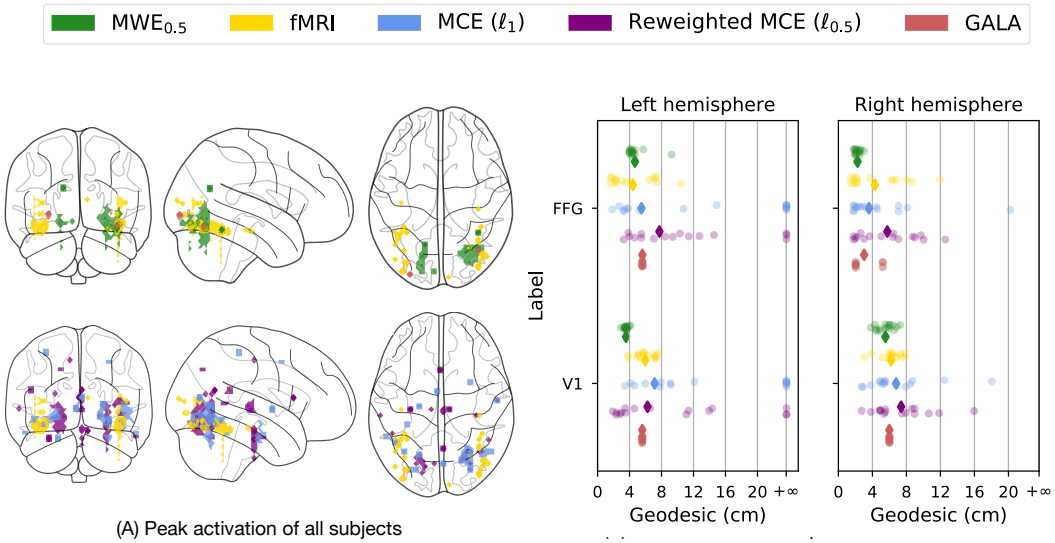


Fig. 3.11. (A) Peak activation in each hemisphere of all subjects of the DS117 dataset (each dot represents the peak of each subject). (B) Mean geodesic distance between the peak foci and the vertices of the labels FFA and V1. For some subjects, MCE / reweighted MCE produce 6-sparse solutions entirely in the right hemisphere, to which the goedesic $+\infty$ is assigned. Notice that MWE tends to be closer to the FFG than MCE. The glass brains of the remaining subjects are displayed in the appendix.

current estimates (MCE or Lasso) (Uutela, Hämäläinen, and Somersalo, 1999), reweighted MCE, $\text{MWE}_{0.5}$ and fMRI by computing for each subject the mean geodesic distance between the mode of the neural activation map of each subject and the vertices of the Fusiform-gyrus (FFG) as well as the primary visual cortex (V1). Figure 3.11 (B) shows that the distribution of MWE geodesics is closer to that of fMRI z-maps. By promoting functional similarity, MWE disregards the spurious activation that are far from the regions of interest. Moreover, one can notice that some 6-sparse MCE models cancel out all sources in the left hemisphere (subjects with a geodesic equal to $+\infty$). Figure 3.11 (A) shows using glass-brains (Abraham et al., 2014) the distribution of the peak activation of all subjects (each dot corresponds to a subject). Multi-subject models (GALA and MWE) do not display the spurious activation in the temporal lobes and the medial wall recovered by MCE and Reweighted MCE. While both GALA and MWE are more consistent with fMRI, GALA promotes an almost identical solution for all subjects: functional consistency is favored at the expense of individual signatures.

4 Discussion

The M/EEG source imaging problem is a notoriously hard inverse problem, in particular when the underlying neural activity is distributed over different coactive brain regions. To tackle this problem, this work proposes to jointly localize sources for a population of subjects by casting the estimation as a multi-task regression problem.

Embracing this formulation of multi-task regression, this work develops three key ideas. First it proposes to use non-linear registration to obtain subject specific leadfield matrices that are spatially aligned. Second it copes with the issue of inter-subject spatial variability of functional activations using optimal transport. Finally, it makes use of non-convex sparsity priors ($\ell_{0.5}$) and joint inference of source estimates and noise variances σ_k^2 to obtain accurate source amplitudes.

The classic pipeline of a M/EEG group source imaging study is to perform source localization independently across subjects using inverse solvers such as MNE, MCE, sLORETA, dSPM or MxNE. The group-level analysis is then carried out as a post-processing step by averaging the source estimates of each subject or by aggregating Z-scores in a multiple tests comparison (Takeda et al., 2019). This is usually done thanks to a non-linear registration and by averaging of the estimates after mapping them to the same brain template. In this work, a different approach based on multi-task regression is proposed. The non-linear registration is used to compute leadfield matrices that are spatially aligned. A source space formed by candidate dipoles are defined on the average brain geometry and this source space is warped to individual anatomies for which Maxwell equations are solved numerically. By doing so, we demonstrate improvements in terms of source localization accuracy. This is significant evidence that anatomical variability can be more a blessing than a curse for group level M/EEG source imaging.

This statement is actually inline with the work of Larson, Maddox, and Lee (2014), who suggested that anatomical differences between subjects can improve the accuracy of the averaged source estimates by emphasizing common sources across subjects. Our simulations confirm this hypothesis not only for averaged estimates but also for individual ones. Indeed, all the multi-task models studied in our simulations improve with more subjects. One possible explanation of why anatomical differences help is that anatomical variability combined with functional similarities lead to non-redundant information across subjects. Take the example of a shared source across subjects. Different folding patterns of the cortical mantle would lead to different (normal) orientations of the current dipole. Since the relative position of the sensors is not changed, the leadfields – having different sensitivity maps – generate measurements with more information, i.e higher rank. Quantitatively, our simulations with semi-real data show that multi-subject inverse solvers improve the localization error by almost 4 mm per source.

By pooling together data from multiple subjects one can increase the number of measurements, hence make the problem less ill-posed. Yet, this cannot be done without taking into consideration differences between subjects, especially the spatial variability in activation patterns. To cope with this issue when averaging brain patterns both in M/EEG and fMRI, Wasserstein distances have proven efficient (Gramfort, Peyré, and Cuturi, 2015). Through this work, we explained how they could be included directly in the inverse solver. Thanks to their ability to model spatial proximity between source estimates, the MWE model allows to promote functional similarities across subjects using the geometry of the cortical mantle. Fortunately, the computation of the Wasserstein barycenter does not lead to a computational bottleneck.

In our experiments, 40% to 60% of time is spent on optimal transport versus proximal coordinate descent. Thanks to careful optimization procedures based on Sinkhorn iterations and block coordinate descent algorithms, the model proposed here runs in a few minutes on empirical M/EEG datasets.

Beyond the use of Wasserstein metrics to cope with spatial misalignments, the proposed MWE_q model brings in two important ingredients from the statistics literature employing sparsity promoting regularizations: concomitant estimation and convex reweighted schemes. By using concomitant estimation, the MWE_q model can cope with the different noise levels and signal-to-noise ratios for the different subjects. This is particularly critical to have the number of hyperparameters of the model that is fixed and does not scale with the number of subjects. In theory, for source imaging with a solver such as dSPM or sLORETA, that is applied independently for all subjects, the regularization parameters could be tuned for each dataset. The MWE_q model has a list of regularization parameters that does not depend on the number of subjects. Besides, results from Figures 4 and 5 demonstrate the benefit of $\text{MWE}_{0.5}$ vs. MWE_1. Employing a more aggressive sparsity promoting regularization improves in particular the source amplitude estimation as shown by the MSE metric. Also, as demonstrated empirically in Figure 3.8, it greatly simplifies the setting of the regularization parameter μ as solutions become suddenly much less sensitive to the choice of this parameter. Indeed, in practice one can set the sparsity hyperparameter λ based on the number of active sources. Meanwhile, the OT hyperparameter can be set as $\mu = \frac{1}{2}\mu_{\max}$ where $\mu_{\max}$ is the smallest μ for which the solutions are suddenly dense. Finally, note that other priors could be used along with the optimal transport regularizer $\tilde{W}$. The same optimization strategy would apply as long as the penalty g is separable across sources and subjects, i.e it can be written as $g(\mathbf{x}) = \sum_{s,j} g_0(\mathbf{x}_j^{(s)})$. For instance, one can define $g_0(x) = x^2$ to favor distributed sources over focal ones similarly to dSPM or sLORETA.

From a more neuroscientific perspective, the model presented here has potentially interesting consequences. Results on Cam-CAN demonstrate that the sources obtained with $\text{MWE}_{0.5}$ have a higher spatial specificity. As seen in Figure 7, the inferred activation foci are well limited to primary auditory cortices while solvers that are not based on a group-level multi-task regression model lead to spurious activations next to secondary somatosensory cortices and on middle temporal gyrus. On DS117 dataset, the cognitive task performed by the subjects is more advanced, complicating the discussion of the results in terms of localization. Yet, the availability of the fMRI data allows for a quantification of the activation foci between MEG and fMRI. While it is often repeated that fMRI and M/EEG sources are different, and thus brain activation maps obtained by these different modalities should not necessarily match, Kujala et al. (2014) provide evidence that fMRI correlates with source-localized MEG activity in many regions of the brain specially with the high frequency components, suggesting that similarities between the two methods should not be overlooked. Our results point in the same direction, demonstrating that the proposed method reduces the gap between MEG source imaging and fMRI.

5 Appendix

Proof of proposition 29.

Proposition 31 *Let* $\mathbf{R} \overset{\text{def}}{=} [{\mathbf{L}^{(1)}}^\top \mathbf{b}^{(1)}, \dots, {\mathbf{L}^{(S)}}^\top \mathbf{b}^{(S)}] \in \mathbb{R}^{n,S}$ *and define the maximum hyper-parameter values* $\mu_{\max} \overset{\text{def}}{=} \frac{\|\mathbf{R}\|_{2\infty}}{n}$ *and* $\lambda_{\max} \overset{\text{def}}{=} \frac{\|\mathbf{R}\|_\infty}{n}$. *Let* $\mathbf{X}^\star = \overset{c}{\mathbf{X}}^\star + \overset{d}{\mathbf{X}}^\star$ *be a solution of Dirty models (3.8). Then the following holds:*

$$\mu \geq \sqrt{S}\lambda \Rightarrow \overset{c}{\mathbf{X}}^\star = 0 \tag{3.32}$$

$$\lambda \geq \mu \Rightarrow \overset{d}{\mathbf{X}}^\star = 0 \tag{3.33}$$

$$\lambda \geq \lambda_{\max} \text{ and } \mu \geq \mu_{\max} \Leftrightarrow \overset{d}{\mathbf{X}}^\star = \overset{c}{\mathbf{X}}^\star = 0 \ . \tag{3.34}$$

PROOF. Let's denote the block diagonal matrix $\mathbf{G} \overset{\text{def}}{=} \text{diag}(\mathbf{L}^{(1)}, \dots, \mathbf{L}^{(S)}) \in \mathbb{R}^{Sn \times Sp}$ and the vertical stackings $\overset{c}{\mathbf{X}}' \overset{\text{def}}{=} [\overset{c}{\mathbf{x}}^{(1)}, \dots, \overset{c}{\mathbf{x}}^{(S)}] \in \mathbb{R}^{Sp}$ and $\overset{d}{\mathbf{X}}' \overset{\text{def}}{=} [\overset{d}{\mathbf{x}}^{(1)}, \dots, \overset{d}{\mathbf{x}}^{(S)}] \in \mathbb{R}^{Sp}$. The reverse shaping operation $\mathbf{X}' \in \mathbb{R}^{pS} \mapsto \mathbf{X} \in \mathbb{R}^{Sp}$ is denoted by φ. The resulting observed data can be reshaped as $\mathbf{B}' \overset{\text{def}}{=} \mathbf{L}'\mathbf{X}' = \mathbf{L}'(\overset{c}{\mathbf{X}}' + \overset{d}{\mathbf{X}}') \in \mathbb{R}^{Sn}$. With these changes, problem (3.8) can be written:

$$\min_{\overset{c}{\mathbf{X}}, \overset{d}{\mathbf{X}}} \in \mathbb{R}^{Sp} \frac{1}{2n}\|\mathbf{L}'(\overset{c}{\mathbf{X}}' + \overset{d}{\mathbf{X}}') - \mathbf{B}'\|^2 + \mu\|\overset{c}{\mathbf{X}}\|_{21} + \lambda\|\overset{d}{\mathbf{X}}\|_{11} \ , \tag{3.35}$$

The optimality condition for problem (3.8) reads:

$$0 \in \frac{1}{n}\phi({\mathbf{L}'}^\top(\mathbf{L}'\overset{c}{\mathbf{X}}^\star + \mathbf{L}'\overset{d}{\mathbf{X}}^\star - \mathbf{B}')) + \mu\partial_{\ell_{21}}(\overset{c}{\mathbf{X}}^\star) \ , \tag{3.36}$$

$$0 \in \frac{1}{n}\phi({\mathbf{L}'}^\top(\mathbf{L}'\overset{c}{\mathbf{X}}^\star + \mathbf{L}'\overset{d}{\mathbf{X}}^\star - \mathbf{B}')) + \lambda\partial_{\ell_{11}}(\overset{d}{\mathbf{X}}^\star) \ . \tag{3.37}$$

Therefore, there exist $\overset{c}{\mathbf{Z}} \in \partial_{\ell_{21}}(\overset{c}{\mathbf{X}}^\star)$ and $\overset{d}{\mathbf{Z}} \in \partial_{\ell_{11}}(\overset{d}{\mathbf{X}}^\star)$ such that:

$$0 = \frac{1}{n}\phi({\mathbf{L}'}^\top(\mathbf{L}'\overset{c}{\mathbf{X}}^\star + \mathbf{L}'\overset{d}{\mathbf{X}}^\star - \mathbf{B}')) + \mu\overset{c}{\mathbf{Z}} \ , \tag{3.38}$$

$$0 = \frac{1}{n}\phi({\mathbf{L}'}^\top(\mathbf{L}'\overset{c}{\mathbf{X}}^\star + \mathbf{L}'\overset{d}{\mathbf{X}}^\star - \mathbf{B}')) + \lambda\overset{d}{\mathbf{Z}} \ . \tag{3.39}$$

Thus:

$$\mu\overset{c}{\mathbf{Z}} = \lambda\overset{d}{\mathbf{Z}} \ . \tag{3.40}$$

The sub-differentials of ℓ_{21} and ℓ_{11} are given by:

$$\partial_{\ell_{21}}(\mathbf{X})_j = \begin{cases} \frac{\mathbf{X}_j}{\|\mathbf{X}_j\|_2} & \text{if} \quad \mathbf{X}_j \neq 0 \\ \{\mathbf{z} \in \mathbb{R}^S, \|\mathbf{z}\|_2 \leq 1\} & \text{if} \quad \mathbf{X}_j = 0 \end{cases} \tag{3.41}$$

$$\partial_{\ell_{11}}(\mathbf{X})_j^{(s)} = \begin{cases} \text{sign}(\mathbf{X}_j^{(k)}) & \text{if} \quad \mathbf{X}_j^{(s)} \neq 0 \\ \{z \in \mathbb{R}, |\mathbf{z}| \leq 1\} & \text{if} \quad \mathbf{X}_j^{(s)} = 0 \end{cases} \tag{3.42}$$

Therefore, for any j and s:

$$\|\overset{c}{\mathbf{Z}}_j\|_2 \leq 1 \ , \tag{3.43}$$

$$|\overset{d}{\mathbf{Z}^{(s)}}_j| \leq 1 \ . \tag{3.44}$$

Thus, equation (3.40) leads to:

$$\begin{aligned} \overset{c}{\mathbf{X}}_j \neq 0 \Rightarrow \mu \frac{\overset{c}{\mathbf{X}_j}}{\|\overset{c}{\mathbf{X}}_j\|_2} &= \lambda \overset{d}{\mathbf{Z}}_j \\ \Rightarrow \mu = \lambda \|\overset{d}{\mathbf{Z}}_j\|_2 &\leq \lambda S \ . \end{aligned}$$

Hence:

$$\mu > \lambda S \Rightarrow \overset{c}{\mathbf{X}} = 0 \ . \tag{3.45}$$

Similarly:

$$\begin{aligned} \overset{d}{\mathbf{X}^{(s)}}_j \neq 0 \Rightarrow \mu \overset{c}{\mathbf{Z}^{(s)}}_j &= \lambda \text{sign}(\overset{d}{\mathbf{X}^{(s)}}_j) \\ \Rightarrow \mu |\overset{c}{\mathbf{Z}^{(s)}}_j| &= \lambda \\ \Rightarrow \mu &\geq \lambda \end{aligned}$$

Hence:

$$\mu < \lambda \Rightarrow \overset{d}{\mathbf{X}} = 0 \ . \tag{3.46}$$

Finally:

$$(\overset{c}{\mathbf{X}}^\star, \overset{d}{\mathbf{X}}^\star) = (0,0) \Leftrightarrow \exists \mathbf{Z} \in \mu \partial_{\ell_{21}}(0) \cap \partial_{\ell_{11}}(0) \, \frac{1}{n} \mathbf{L}'^\top \mathbf{B}' + \mathbf{Z} = 0 \tag{3.47}$$

$$\Leftrightarrow \frac{1}{n} \|\mathbf{L}'^\top \mathbf{B}'\|_{2\infty} \leq \mu \text{ and } \frac{1}{n} \|\mathbf{L}'^\top \mathbf{B}'\|_{\infty} \leq \lambda \tag{3.48}$$

$$\Leftrightarrow \mu_{\max} \leq \mu \text{ and } \lambda_{\max} \leq \lambda \ . \tag{3.49}$$

■

Glass brains for all subjects

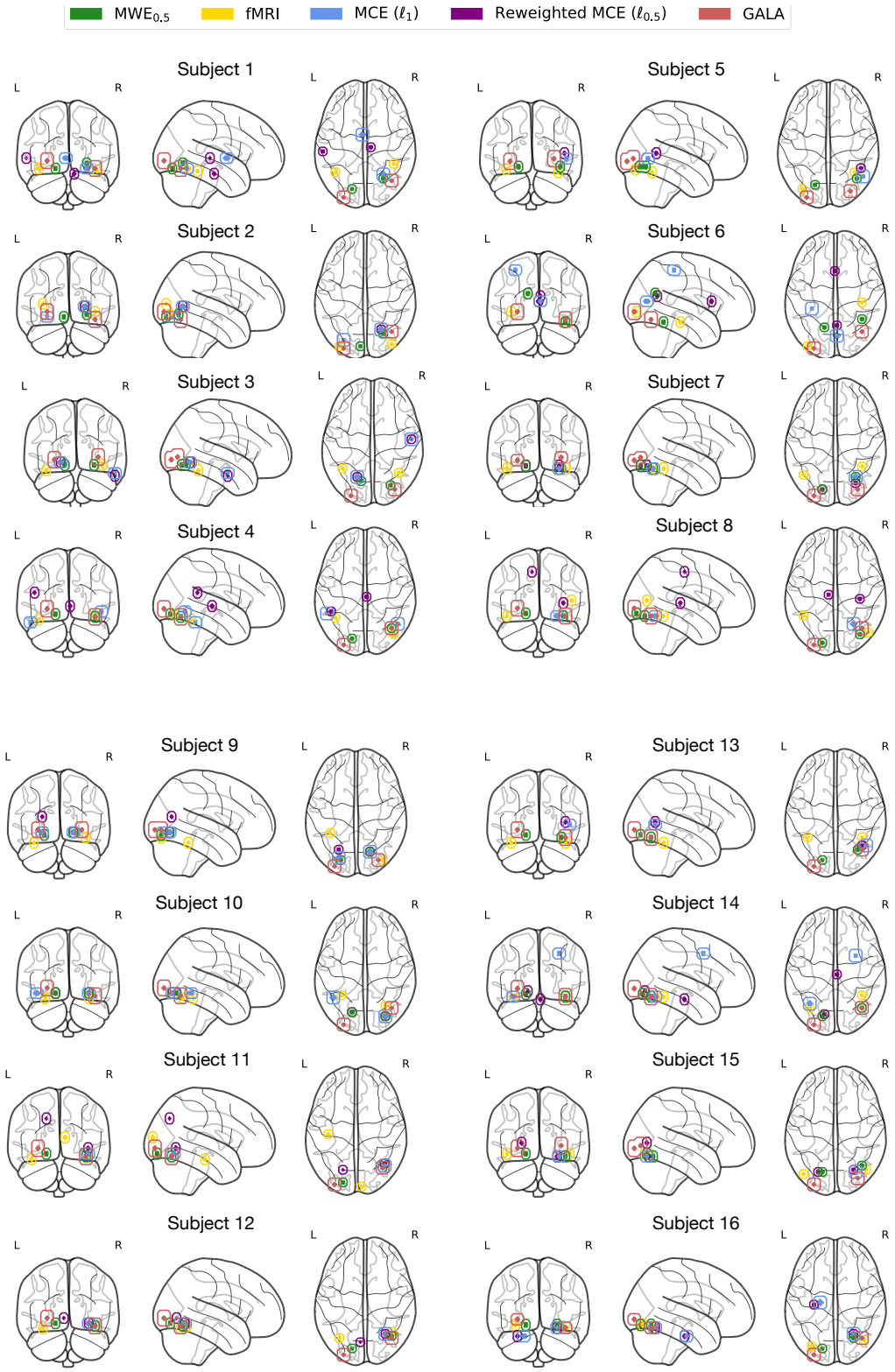


Fig. 3.12. Peak activation foci on each hemisphere of all subjects of the DS117 dataset.

Chapter 4

Spatio-temporal Optimal transport

Spatio-temporal data consist of time series in which each time sample is multivariate and lives in a certain coordinate system equipped with a natural distance. Such a coordinate system can correspond to 2D or 3D positions in space, pixel positions etc. This setting is encountered in several machine learning problems. Multi-target tracking for example, involves the prediction of the time indexed positions of several objects or particles (Doucet et al., 2002). In brain imaging, magnetoencephalography (MEG) and functional magnetic resonance imaging (fMRI) yield measurements of neural activity in multiple positions and at multiple time points (Gramfort et al., 2011). Quantifying spatio-temporal variability in brain activity can allow to compare different clinical populations which is the main applied motivation of this chapter.

As discussed in Chapter 2, OT metrics such as the Wasserstein distance can capture spatial variations between probability distributions. Given a transport cost function – commonly referred to as ground metric – the Wasserstein distance computes the optimal transportation plan between two measures. Its heavy computational cost can be significantly reduced by using entropy regularization (Cuturi, 2013). Additionally, it is also possible to extend its definition to handle measures with different total mass using the unbalanced OT formulation of Chizat et al. (2018b), which also relies on entropic regularization, pending some minor modifications to Sinkhorn's algorithm (Frogner et al., 2015; Chizat et al., 2018b). To take into account the temporal dimension, one could define the ground metric as a combination of spatial and temporal shifts similarly to the definition of TL^p distances (Thorpe et al., 2017). This method however ignores the chronological order of the data and requires a tuning parameter to settle the tradeoff between spatial and temporal transport cost. This is one of the main features of Dynamic Time Warping (DTW).

Given a pairwise distance matrix between all time points of two time series of respective lengths m, n, DTW computes the minimum-cost alignment between the time series (Sakoe and Chiba, 1978) while preserving the chronological order of the data. Indeed, the DTW optimization problem is constrained on alignments where no temporal back steps are allowed. It can be seen as an OT-like problem where the transport plan must not respect the marginal constraints but instead is a binary matrix with at least one non-zero entry per line and per column, and where the cumulated non-zero path is formed by $\rightarrow, \downarrow, \searrow$ steps exclusively. However, the binary nature of this set makes the DTW loss non-differentiable which is a major limitation when DTW is used as a loss function. To circumvent this issue, several authors introduced smoothed versions of DTW (Saigo et al., 2004; Cuturi, 2011; Cuturi and Blondel, 2017). Instead of selecting *the* minimum cost alignment, Global Alignment Kernels (GAK) (Saigo et al., 2004; Cuturi, 2011) compute a weighted cost on the whole set of possible alignments. Similarly, the soft-minimum generalization

approach of Cuturi and Blondel (2017) – called soft-DTW – provides a similar framework to that of GAK where gradients can easily be computed used a backpropagation of Bellman's equation (Bellman, 1952).

We show that this soft version of DTW has another property: it increases quadratically with temporal shifts. It can thus be considered a "transportation" loss for time series. For spatio-temporal data, it is only natural to combine DTW with Optimal transport. To do so however, we need to make use of unbalanced optimal transport to allow different time samples to have different total masses. As with balanced OT, we propose a debiased version of Unbalanced OT and study its properties. We conclude this chapter with a barycenter algorithm for spatio-temporal data.

This chapter is based on:

- H. Janati et al, *Spatio-temporal alignments: optimal transport in space and time*, AISTATS'20.
- H. Janati et al, *Optimal transport barycenters for spatio-temporal data*, Submitted.

1 OT in time

1.1 Soft dynamic time warping

Consider two multivariate time series $\mathbf{x} \in \mathbb{R}^{p,T_1}$ and $\mathbf{y} \in \mathbb{R}^{p,T_2}$ with respective lengths T_1, T_2 and having observations in $\mathbb{R}^p$. DTW is defined through some pairwise distance matrix $\Delta(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{T_1,T_2}$ between all their time points such that the cost of a given alignment function $\sigma : [\![1, T_1]\!] \to [\![1, T_2]\!]$ is equal to $\sum_{i=1}^{T_1} \Delta(\mathbf{x}_i, \mathbf{y}_{\sigma(i)})$. To guarantee the preservation of the chronology of the data, σ must be increasing and verify $\sigma(1) = 1$ and $\sigma(T_1) = T_2$. The resulting optimization problem is however better posed as a minimization of $\sum_{i=1}^{T_1} \sum_{j=1}^{T_2} A_{ij}\Delta(\mathbf{x}_i, \mathbf{y}_j)$ over the set of binary alignments A on the rectangular lattice $[\![1, T_1]\!] \times [\![1, T_2]\!]$ where no temporal back steps are allowed. This amounts to considering binary matrices with a non-zero path linking the corners of the lattice $(1, 1)$ (upper left) and (T_1, T_2) (bottom right) using $\rightarrow, \downarrow, \searrow$ steps exclusively (Sakoe and Chiba, 1978). Figure 4.1 displays a toy example of such an alignment. Formally, DTW is defined as:

$$\mathbf{dtw}(\mathbf{x}, \mathbf{y}; \Delta) = \min\{\langle \mathbf{A}, \Delta(\mathbf{x}, \mathbf{y})\rangle, \mathbf{A} \in \mathcal{A}_{T_1,T_2}\} \ , \tag{4.1}$$

where $\langle ., . \rangle$ denotes the Frobenius dot product. The binary nature of the constraint set in (4.1) makes the DTW loss non-differentiable which is a major limitation when DTW is used as a loss function. To circumvent this issue, several authors introduced regularized variants of DTW (Saigo et al., 2004; Cuturi, 2011; Cuturi and Blondel, 2017). Instead of selecting *the* minimum cost alignment, Global Alignment Kernels (GAK) for instance (Saigo et al., 2004; Cuturi, 2011) compute a weighted cost of all possible alignments with a certain smoothing hyperparameter. Similarly, the soft-minimum generalization approach of (Cuturi and Blondel, 2017) – called soft-DTW – provides a similar framework to that of GAK

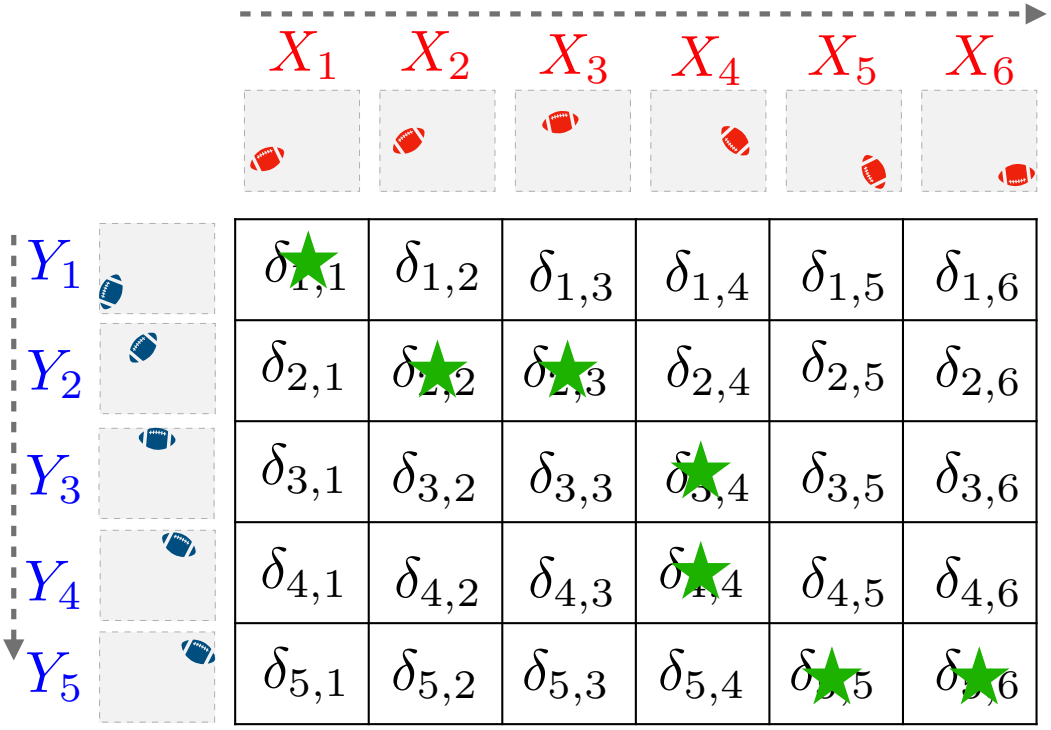


Fig. 4.1. Example of Dynamic Time Warping alignment between two time series of images given a pairwise distance matrix.

that includes DTW as a sub-case:

$$\mathbf{dtw}_\beta(\mathbf{x}, \mathbf{y}; \Delta) = \mathrm{softmin}_\beta\{\langle \mathbf{A}, \Delta(\mathbf{x}, \mathbf{y})\rangle, \mathbf{A} \in \mathcal{A}_{T_1,T_2}\} \ , \tag{4.2}$$

where the soft-minimum operator of a set $\mathcal{A}$ with parameter $\beta \geq 0$ is defined as:

$$\mathrm{softmin}_\beta(\mathcal{A}) = \begin{cases} -\beta \log\left(\sum_{a\in\mathcal{A}} e^{-a/\beta}\right) & \text{if } \beta > 0 \\ \min \mathcal{A} & \text{if } \beta = 0 \end{cases} \tag{4.3}$$

In particular, softmin is continuous at 0 so that when $\beta \to 0$, $\mathbf{dtw}_\beta$ approaches **dtw**.

Forward recursion Figure 4.1 illustrates two time series of images and their cost matrix Δ. The path from (1, 1) to (5, 6) is an example of a feasible alignement in $\mathcal{A}_{5,6}$. When $\beta = 0$, the soft-minimum is a minimum and $\mathbf{dtw}_\beta$ falls back to the classical DTW metric. Nevertheless, it can still be computed using the dynamic program of Algorithm 9 with a soft-min instead of min operator.

Algorithm 9 BP recursion to compute $\mathbf{dtw}_\beta$ (Cuturi and Blondel, 2017)

Input: data $\mathbf{x}, \mathbf{y}$ soft-min parameter β and distance function δ
Output: $\mathbf{dtw}_\beta(\mathbf{x}, \mathbf{y}) = r_{T_1,T_2}$
$r_{0,0} = 0; r_{0,j} = r_{i,0} = \infty$ for $i \in [\![T_1]\!], j \in [\![T_2]\!]$
for $i = 1$ **to** T_1 **do**
  for $j = 1$ **to** T_2 **do**
    $r_{i,j} = \delta(\mathbf{x}_i, \mathbf{y}_j) + \text{softmin}_\beta(r_{i-1,j-1}, r_{i-1,j}, r_{i,j-1})$
  end for
end for

Algorithmic differentiation When $\beta > 0$, differentiating (4.2) with respect to $\mathbf{x}$ yields:

$$\nabla_{\mathbf{x}} \mathbf{dtw}_\beta(\mathbf{x}, \mathbf{y}, \Delta) = \left(\frac{\partial \Delta(\mathbf{x}, \mathbf{y})}{\partial \mathbf{x}}\right)^\top E_\beta(\mathbf{x}, \mathbf{y}) \ , \tag{4.4}$$

where $E_\beta(\mathbf{x}, \mathbf{y}) \overset{\text{def}}{=} \frac{\partial \mathbf{dtw}_\beta}{\partial \Delta}(\mathbf{x}, \mathbf{y}) = \frac{\sum_{A \in \mathcal{A}_{T_1,T_2}} e^{-\frac{\langle \mathbf{A}, \Delta(\mathbf{x},\mathbf{y}) \rangle}{\beta}} A}{\sum_{A \in \mathcal{A}_{T_1,T_2}} e^{-\frac{\langle \mathbf{A}, \Delta(\mathbf{x},\mathbf{y}) \rangle}{\beta}}}$ can be interpreted as a weighted average alignment. To compute $E_\beta(\mathbf{x}, \mathbf{y})$, (Cuturi and Blondel, 2017) proposed to back propagate the forward recursion of Algorithm 9, starting from E_{T_1,T_2} down to $E_{0,0}$. Indeed, the value of $\mathbf{dtw}_\beta$ is stored in the last alignment cost r_{T_1,T_2}. Thus differentiating $\mathbf{dtw}_\beta$ with respect to any $r_{i,j}$ only involves the terms of $r_{i-1,j}, r_{i,j-1}$ and $r_{i-1,j-1}$. Differentiating the softmin operation of the forward pass yields the backward recursion of Algorithm 10.

Algorithm 10 Backward recursion to differentiate sdtw (Cuturi and Blondel, 2017).

Input: $\mathbf{x}, \mathbf{y}$, parameter β, distance δ and intermediary alignment matrix R
Output: $E = E_\beta(\mathbf{x}, \mathbf{y})$
$r_{i,m+1} = r_{n+1,j} = -\infty, i \in [\![n]\!], j \in [\![m]\!]$
$e_{i,m+1} = r_{n+1,j} = 0, i \in [\![n]\!], j \in [\![m]\!]$
$\delta_{i,m+1} = \delta_{n+1,j} = 0, i \in [\![n]\!], j \in [\![m]\!]$
$\delta_{n+1,m+1} = 0, e_{n+1,m+1} = 1, r_{n+1,m+1} = r_{n,m}$
for $i = 1$ **to** n **do**
  for $j = 1$ **to** m **do**
    $a = \exp \frac{1}{\beta}(r_{i+1,j} - r_{i,j} - \delta_{i+1,j})$
    $b = \exp \frac{1}{\beta}(r_{i,j+1} - r_{i,j} - \delta_{i,j+1})$
    $c = \exp \frac{1}{\beta}(r_{i+1,j+1} - r_{i,j} - \delta_{i+1,j+1})$
    $e_{i,j} = a e_{i+1,j} + b e_{i,j+1} + c e_{i+1,j+1}$
  end for
end for

The cardinality of the set of feasible alignments $\mathcal{A}_{T_1,T_2}$ plays a crucial role in our proofs. It is known as the Delannoy number $D(T_1-1, T_2-1)$ (Cuturi, 2011). For the sake of convenience, we consider the shifted Delannoy sequence starting at $n = m = 1$ so that: $\text{card}(\mathcal{A}_{m,n}) = D_{m,n}$ for all integers $m, n \geq 1$. Formally, the Delannoy sequence can be defined recursively by:

Definition 4 (Delannoy sequence) *The Delannoy number $D_{m,n}$ corresponds to the number of paths from $(1,1)$ to (m,n) in a $(m \times n)$ lattice where only $\rightarrow, \downarrow, \searrow$ movements are allowed. It can also be defined with the recursion $\forall m, n \in \mathbb{N}^\star$:*

$$D_{1,n} = D_{m,1} = 1 \tag{4.5}$$
$$D_{m+1,n+1} = D_{m,n+1} + D_{m+1,n} + D_{m,n} \ . \tag{4.6}$$

1.2 New bounds of Delannoy numbers

In this section we provide several new inequalities bounding the growth of Delannoy numbers. These inequalities are crucial in both proving that $\mathbf{dtw}_\beta$ is sensitive to temporal shifts and in providing a practical heuristic to set β depending on temporal sensitivity. First we start by studying the central Delannoy numbers $D_{n,n}$.

Proposition 32 *Let $c = 1 + \sqrt{2}$ and $\sigma = \frac{21}{22}c^2 - 5$. The central (diagonal) Delannoy sequence $D_m \overset{\text{def}}{=} D_{m,m}$ verifies:*

$$\frac{D_{m+1}}{D_m} \leq c^2 \frac{m}{m+\frac{1}{2}} \quad \forall m \geq 1 \tag{4.7}$$
$$\frac{D_{m+1}}{D_m} \geq c^2 \frac{m}{m+\sigma} \quad \forall m \geq 5 \tag{4.8}$$

PROOF. In (Janati, Cuturi, and Gramfort, 2020b), we showed the weaker result $\frac{D_{m+1}}{D_m} \leq c^2$. The same proof can be adapted to obtain tighter bounds depending on m. The central (or diagonal) Delannoy numbers D_m verify the 2-stages recursion equation for any $m \geq 2$ (Stanley, 2011):

$$mD_{m+1} = (6m-3)D_m - (m-1)D_{m-1} \tag{4.9}$$

We are going to prove both inequalities by induction.

Inequality (4.7) For $m = 1$, we have $D_2 = 3 \leq 2 + \frac{4}{3}\sqrt{2} = \frac{2}{3}c^2 = \frac{2}{3}c^2 D_1$. Assume that (4.8) holds for some $m \geq 2$. From (4.9) and the induction assumption:

$$(m+1)D_{m+2} = (6m+3)D_{m+1} - mD_m \tag{4.10}$$

$$\leq (6m+3)D_{m+1} - m\frac{m+\frac{1}{2}}{c^2 m}D_{m+1} \tag{4.11}$$

$$\leq (6m+3-\frac{m+\frac{1}{2}}{c^2})D_{m+1} \tag{4.12}$$

$$= \frac{(6c^2-1)m + \frac{6c^2-1}{2}}{c^2}D_{m+1} \tag{4.13}$$

$$= c^2(m+\frac{1}{2})D_{m+1} \tag{4.14}$$

where we used the fact that $1/c^2 = \frac{1}{3+2\sqrt{2}} = 3 - 2\sqrt{2} = 6 - c^2$, hence $6c^2 - 1 = c^4$. Therefore:

$$\frac{D_{m+2}}{D_{m+1}} \leq c^2 \frac{m+\frac{1}{2}}{m+1} \quad .$$

To conclude, it suffices to show that for all $m \geq 5$:

$$\frac{m+\frac{1}{2}}{m+1} \leq \frac{m+1}{m+\frac{3}{2}} \tag{4.15}$$

which is equivalent to:

$$(m+\frac{1}{2})(m+\frac{3}{2}) \leq (m+1)^2 \tag{4.16}$$

$$\Leftrightarrow m^2 + 2m + \frac{3}{4} \leq m^2 + 2m + 1 \tag{4.17}$$

$$\Leftrightarrow \frac{3}{4} \leq 1 \tag{4.18}$$

Inequality (4.8) For $m = 5$, we have with numerical evaluation $\frac{D_6}{D_5} - c^2\frac{5}{5+\sigma} \geq 0$. Assume that (4.8) holds for some $m \geq 6$. From (4.9) and the induction assumption:

$$(m+1)D_{m+2} = (6m+3)D_{m+1} - mD_m \tag{4.19}$$
$$\geq (6m+3)D_{m+1} - m\frac{m+\sigma}{c^2 m}D_{m+1} \tag{4.20}$$
$$= (6m+3-\frac{m+\sigma}{c^2})D_{m+1} \tag{4.21}$$
$$= \frac{(6c^2-1)m + 3c^2 - \sigma}{c^2}D_{m+1} \tag{4.22}$$
$$= c^2(m + \frac{3c^2-\sigma}{c^4})D_{m+1} \tag{4.23}$$

where we used the fact that $1/c^2 = \frac{1}{3+2\sqrt{2}} = 3 - 2\sqrt{2} = 6 - c^2$, hence $6c^2 - 1 = c^4$. Therefore:

$$\frac{D_{m+2}}{D_{m+1}} \geq c^2\frac{m + \frac{3c^2-\sigma}{c^4}}{m+1} \quad .$$

To conclude, it suffices to show that for all $m \geq 2$:

$$\frac{m + \frac{3c^2-\sigma}{c^4}}{m+1} \geq \frac{m+1}{m+\sigma+1} \tag{4.24}$$

which is equivalent to:

$$(m + \frac{3c^2-\sigma}{c^4})(m+\sigma+1) \geq (m+1)^2$$
$$\Leftrightarrow (\frac{3c^2-\sigma}{c^4} + \sigma - 1)m + \frac{3c^2-\sigma}{c^4}(\sigma+1) - 1 \geq 0 \tag{4.25}$$

Numerical evaluation shows that $\frac{3c^2-\sigma}{c^4} + \sigma - 1 \geq 0.06$ and that $\frac{3c^2-\sigma}{c^4}(\sigma+1) - 1 \geq -0.24$. Thus (4.25) is verified for $m \geq 5$. ■

The main purpose of proposition 32 is the following corollary which is crucial to derive a simple heuristic to set the hyperparameter β when using $\mathbf{dtw}_\beta$ in practice.

Corollary 4 *Let $T > m \geq 1$, $c = 1 + \sqrt{2}$ and $\sigma = \frac{21c^2}{22} - 5$ ($\approxeq 0.56$). The central Delannoy numbers verify:*

$$c^{2(T-m)}\frac{D_m}{D_T} \geq \left(\frac{T}{me}\right)^{\frac{1}{2}} \quad \textit{for } m \geq 1$$
$$c^{2(T-m)}\frac{D_m}{D_T} \leq \left(\frac{T-1}{m-1}\right)^{\sigma} \quad \textit{for } m \geq 5 \tag{4.26}$$

PROOF. Combining both inequalities of proposition 32 leads to:

$$\begin{aligned}
&\prod_{k=m}^{T-1} c^2 \frac{k}{k+\sigma} \leq \frac{D_T}{D_m} \leq \prod_{k=m}^{T-1} c^2 \frac{k}{k+\frac{1}{2}} \\
\Leftrightarrow &\prod_{k=m}^{T-1} \frac{k+\frac{1}{2}}{k} \leq c^{2(T-m)} \frac{D_m}{D_T} \leq \prod_{k=m}^{T-1} \frac{k+\sigma}{k} \\
\Leftrightarrow &\prod_{k=m}^{T-1} \frac{k+\frac{1}{2}}{k} \leq c^{2(T-m)} \frac{D_m}{D_T} \leq \prod_{k=m}^{T-1} \frac{k+\sigma}{k} \\
\Leftrightarrow &\exp\left[\sum_{k=m}^{T-1} \log\left(1+\frac{1}{2k}\right)\right] \leq c^{2(T-m)} \frac{D_m}{D_T} \leq \exp\left[\sum_{k=m}^{T-1} \log\left(1+\frac{\sigma}{k}\right)\right]
\end{aligned}$$

Let $z \in [\frac{1}{2}, 1[$. Using the inequalities $\frac{x}{1+x} \leq \log(1+x) \leq x$ which holds for $x \geq -1$:

$$\begin{aligned}
&\sum_{k=m}^{T-1} \frac{z}{z+k} \leq \sum_{k=m}^{T-1} \log\left(1+\frac{z}{k}\right) \leq \sum_{k=m}^{T-1} \frac{z}{k} \\
\Rightarrow &z \sum_{k=m}^{T-1} \frac{1}{1+k} \leq \sum_{k=m}^{T-1} \log\left(1+\frac{z}{k}\right) \leq z \sum_{k=m}^{T-1} \frac{1}{k} \\
\Leftrightarrow &z\left(\sum_{k=0}^{T-1} \frac{1}{1+k} - \sum_{k=0}^{m-1} \frac{1}{1+k}\right) \leq \sum_{k=m}^{T-1} \log\left(1+\frac{z}{k}\right) \leq z\left(\sum_{k=1}^{T-1} \frac{1}{k} - \sum_{k=1}^{m-1} \frac{1}{k}\right) \\
\Leftrightarrow &z\left(\sum_{k=1}^{T} \frac{1}{k} - \sum_{k=1}^{m} \frac{1}{k}\right) \leq \sum_{k=m}^{T-1} \log\left(1+\frac{z}{k}\right) \leq z\left(\sum_{k=1}^{T-1} \frac{1}{k} - \sum_{k=1}^{m-1} \frac{1}{k}\right)
\end{aligned}$$

Finally, using the classical bounds of the Harmonic series (Chen and Qi, 2003):

$$\log(n) + \gamma + \frac{1}{2n+1} \leq \sum_{i=1}^{n} \frac{1}{i} \leq \log(n) + \gamma + \frac{1}{2n-1} \ , \tag{4.27}$$

it holds:

$$\begin{aligned}\frac{1}{z}\sum_{k=m}^{T-1}\log\left(1+\frac{z}{k}\right) &\leq \log(T-1)+\frac{1}{2T-2}-\log(m-1)-\frac{1}{2m-1}\\ &\leq \log(T)+\frac{1}{2T-2}-\log(m-1)-\frac{1}{2m-1}\\ &\leq \log\left(\frac{T-1}{m-1}\right)+\underbrace{\frac{1}{2}\frac{2(m-T)+1}{(T-1)(2m-1)}}_{<0}\\ &\leq \log\left(\frac{T-1}{m-1}\right)\end{aligned}$$

and similarly:

$$\begin{aligned}\frac{1}{z}\sum_{k=m}^{T-1}\log\left(1+\frac{z}{k}\right) &\geq \log(T)+\frac{1}{2T+1}-\log(m)-\frac{1}{2m+2}\\ &\geq \log\left(\frac{T}{m}\right)-1\end{aligned}$$

Taking the exponential after substituting z by σ (resp. $\frac{1}{2}$) provides the upper (resp. lower) bound. ■

We now turn to provide bounds on the off-diagonal Delannoy numbers which lead to the quadratic lower bound of $\mathbf{dtw}_\beta$. The following proposition was established in (Janati, Cuturi, and Gramfort, 2020b).

Proposition 33 *Let $c = 1+\sqrt{2}$. $\forall m, i \in \mathbb{N}^\star$:*

$$D_{m,m+i} \leq c\Phi_{m,i}D_{m,m+i-1} \tag{4.28}$$

$$c\Psi_{m,i}D_{m,m+i} \leq D_{m+1,m+i} \tag{4.29}$$

Where

$$\begin{cases}\Phi_{m,i} = 1-\frac{(1-\frac{1}{c})(i-1)+\frac{1}{c}}{m+i-1}\\ \Psi_{m,i} = 1+\frac{(1-\frac{1}{c})(i-1)}{m}\end{cases}$$

SKETCH OF PROOF. We prove both statements jointly with a double recurrence reasoning. The initializing for $i = 1$ is immediately obtained using the bounded growth proposition 32. To show the induction step, we rely on the recursion equation (4.6). For the sake of clarity, the full proof is provided in the appendix.

By applying proposition 33 to all $i \in [\![1, k]\!]$, the product of all the obtained inequalities leads to:

Proposition 34 *Let $k \in [\![1, T-1]\!]$, for any $m, m', k \in \mathbb{N}^\star$ such that $m + m' \leq T - 1$ and $k \leq \min(T - m, m' - 1)$. Using the notations of proposition 33 for Φ and Ψ:*

$$\log\left(\frac{D_{m,m}D_{m',m'}}{D_{m+k,m}D_{m'-k,m'}}\right) \geq \log\left(\prod_{i=1}^{k}\frac{\Psi_{m'-i,i}}{\Phi_{m,i}}\right) \geq P(k) \tag{4.30}$$

where $P(k) = \alpha k(k-1) + \rho k + \frac{1}{3T}$ with $\alpha = \frac{2-\sqrt{2}}{T} > 0$ and $\rho = \frac{3\sqrt{2}-4}{3T} > 0$ Where $\alpha = \frac{2-\sqrt{2}}{2}(\frac{1}{m'} + \frac{1}{m+m'}) > 0$ and $\rho = \frac{3\sqrt{2}-4}{3T} > 0$.

PROOF. Iterating the inequalities of proposition 33, we have on one hand with the first inequality:

$$\frac{D_{m,m}}{D_{m,m+k}} \geq \frac{1}{c^k \prod_{i=1}^{k}\Phi_{m,i}} \ , \tag{4.31}$$

and on the other hand with the second inequality:

$$\frac{D_{m+k,m+k}}{D_{m,m+k}} \geq c^k \prod_{i=0}^{k-1}\Psi_{m+i,k-i} = c^k \prod_{i=1}^{k}\Psi_{m+k-i,i} \ .$$

With the change of variable $m' = m + k$ and the symmetry of Delannoy numbers, we have:

$$\frac{D_{m',m'}}{D_{m',m'-k}} \geq c^k \prod_{i=1}^{k}\Psi_{m'-i,i} \ . \tag{4.32}$$

Taking the product of (4.31) and (4.32) leads to the first lower bound. Let $a = 1 - \frac{1}{c}$, it also holds:

$$\log\left(\prod_{i=1}^{k}\frac{\Psi_{m'-i,i}}{\Phi_{m,i}}\right) = \sum_{i=1}^{k}\log(\Psi_{m'-i,i}) - \log(\Phi_{m,i}) \tag{4.33}$$

Using the inequality $\frac{x}{1+x} \leq \log(1+x) \leq x$ for $x > -1$ on both logarithms we have, on one hand:

$$\begin{aligned}\log(\Psi_{m'-i,i}) &= \log\left(1 + \frac{a(i-1)}{m'-i}\right) \\ &\geq \frac{a(i-1)}{m'-i+a(i-1)} = \frac{a(i-1)}{m'-\frac{i}{c}-a} \\ &\geq \frac{a(i-1)}{m'}\end{aligned} \tag{4.34}$$

and on the other hand:

$$\begin{aligned} -\log(\Phi_{m,i}) &= -\log\left(1 - \frac{a(i-1)+\frac{1}{c}}{m+i-1}\right) \\ &\geq \frac{a(i-1)+\frac{1}{c}}{m+i-1} \geq \frac{a(i-1)+\frac{1}{c}}{m+m'} \\ &\geq \frac{a(i-1)}{m+m'} + \frac{1}{cT} \ . \end{aligned} \tag{4.35}$$

Finally, combining equations (4.34) and (4.35), the formula $\sum_{i=1}^{k}(i-1) = \frac{k(k-1)}{2}$ leads the quadratic function P. ■

1.3 $\mathbf{dtw}_\beta$ increases quadratically with temporal shifts

Temporal shifts Let $\mathbf{x}$ and $\mathbf{y}$ be two time series. When studying the properties of $\mathbf{dtw}_\beta$, the dimensionality of the time series is irrelevant since it is compressed when computing the cost matrix Δ. Thus, to study temporal shifts, we assume in this section that $\mathbf{x}$ and $\mathbf{y}$ are univariate and belong to $\mathbb{R}^T$. To properly define temporal shifts, we introduce a few preliminary notions. We name the first (respectively, last) time index where $\mathbf{x}$ fluctuates the *onset* (respectively, the *offset*) of $\mathbf{x}$ and denote it by $\text{on}(\mathbf{x})$ (respectively, $\text{off}(\mathbf{x})$). The *fluctuation set* of $\mathbf{x}$ is denoted by $\text{fluc}(\mathbf{x})$ and corresponds to all time indices between the onset and the offset. Formally:

$$\text{on}(\mathbf{x}) = \underset{i\in[\![1,T-1]\!]}{\arg\min}\{\mathbf{x}_{i+1} \neq \mathbf{x}_i\} \tag{4.36}$$

$$\text{off}(\mathbf{x}) = \underset{i\in[\![1,T-1]\!]}{\arg\max}\{\mathbf{x}_{i+1} \neq \mathbf{x}_i\} \tag{4.37}$$

$$\text{fluc}(\mathbf{x}) = \{i \in [\![1,T]\!], \text{on}(\mathbf{x}) \leq i \leq \text{off}(\mathbf{x})\} \tag{4.38}$$

For $\mathbf{x}$ and $\mathbf{y}$ to be temporally shifted with respect to each other, their values must agree both within and outside their (different) fluctuation sets.

Definition 5 (Temporal k-shift) *Let $\mathbf{x}$ and $\mathbf{y}$ be two time series in $\mathbb{R}^T$ and $k \in [\![1, T-1]\!]$. We say that $\mathbf{y}$ is temporally k-shifted with respect to $\mathbf{x}$ and write $\mathbf{y} = \mathbf{x}_{+_k}$ if and only if:*

$$\begin{aligned} &\text{on}(\mathbf{y}) = \text{on}(\mathbf{x}) + k \\ &\text{off}(\mathbf{y}) = \text{off}(\mathbf{x}) + k \\ &i \leq \text{on}(\mathbf{x}), j \leq \text{on}(\mathbf{y}) \Rightarrow \mathbf{x}_i = \mathbf{y}_j \\ &i \geq \text{off}(\mathbf{x}), j \geq \text{off}(\mathbf{y}) \Rightarrow \mathbf{x}_i = \mathbf{y}_j \\ &i \in \text{fluc}(\mathbf{x}), j \in \text{fluc}(\mathbf{y}), j - i = k \Rightarrow \mathbf{x}_i = \mathbf{y}_j \ . \end{aligned} \tag{4.39}$$

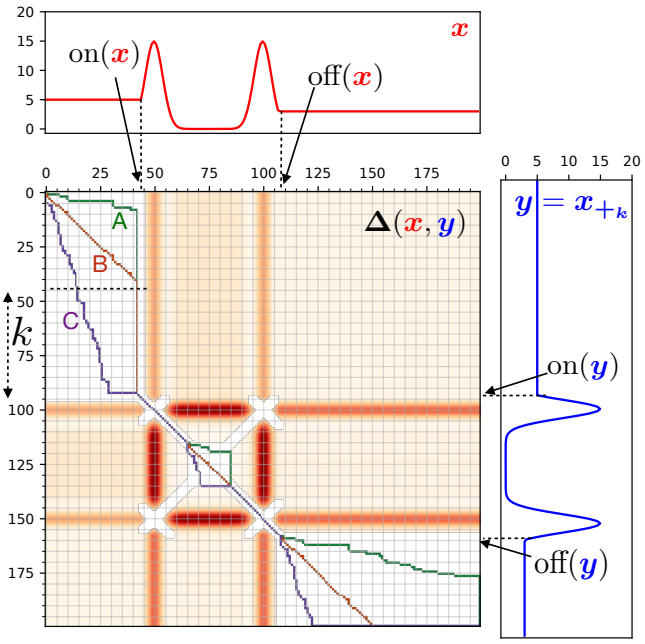


Fig. 4.2. Example of 3 DTW alignment paths (A, B and C) between $\mathbf{x}$ and $\mathbf{y} = \mathbf{x}_{+_k}$ with a temporal 50-shift. The heatmap of the distance matrix Δ shows (white) rectangles where all paths A, B, C have an equal DTW cost of 0. These areas correspond to time durations where $\mathbf{x}$ and $\mathbf{x}_{+_k}$ are constant. It is noteworthy that when shifting one time series, among the areas crossed by the alignments A, B, C, only the two white rectangles outside the fluctuation set change in size.

An example of a temporal 50-shift is illustrated in Figure 4.2. The heatmap of the squared Euclidean cost matrix Δ shows three rectangular white areas where all alignments A, B and C have the same cost of 0. Since $\mathbf{dtw}_0$ is defined as the minimum of all alignment costs, all these paths are equivalent. Temporal k-shifts change the set of alignments with cost 0 but do not change the $\mathbf{dtw}_0$ value. However, when $\beta > 0$, $\mathbf{dtw}_\beta$ computes a weighted sum of all possible paths, which is affected by temporal shifts by including the number of equivalent paths. The cardinality of $\mathcal{A}_{m,n}$ is known as the Delannoy number $D(m-1, n-1)$(Sulanke, 2003), as reported in (Cuturi, 2011). For the sake of convenience, we consider the shifted Delannoy sequence starting at $n = m = 1$ so that: $\text{card}(\mathcal{A}_{m,n}) = D_{m,n}$. If β is positive but small enough, the alignements with 0 cost dominate the $\mathbf{dtw}_\beta$ logsumexp. This leads to proposition 35.

Proposition 35 *Let $k \in [\![1, T-1]\!]$, let $m = \text{on}(\mathbf{x})$ and $m' = T - \text{off}(\mathbf{x})$. Let $\mu = \min_{i,j}\{\Delta(\mathbf{x},\mathbf{x})_{ij} | \Delta(\mathbf{x},\mathbf{x})_{ij} > 0\}$. If $0 < \beta \leq \frac{\mu}{\log(3TD_{T,T})}$:*

$$\mathbf{dtw}_\beta(\mathbf{x}, \mathbf{x}_{+_k}) - \mathbf{dtw}_\beta(\mathbf{x}, \mathbf{x}) \geq \beta \log\left(\frac{D_{m,m} D_{m',m'}}{D_{m+k,m} D_{m'-k,m'}}\right) - \frac{\beta}{3T} \tag{4.40}$$

PROOF. Given a pairwise distance matrix $\Delta(\mathbf{x}, \mathbf{y})$, the soft-DTW dissimilarity is defined as $\mathbf{dtw}_\beta(\mathbf{x}, \mathbf{y}) = -\beta \log\left(\sum_{A \in \mathcal{A}_{T,T}} e^{-\frac{\langle A, \Delta(\mathbf{x},\mathbf{y})\rangle}{\beta}}\right)$. The set of all possible costs can be written: $C = \{\langle A, \Delta(\mathbf{x}, \mathbf{y})\rangle, A \in \mathcal{A}_{T,T}\}$. Dropping duplicates, let $d_0 < d_1, \ldots, < d_G$ denote all unique values in C. And finally let n_i be the number of alignments A such that $\langle A, \Delta(\mathbf{x}, \mathbf{y})\rangle = d_i$. We have:

$$\mathbf{dtw}_\beta(\mathbf{x}, \mathbf{y}) = -\beta \log\left(\sum_{A \in \mathcal{A}_{T,T}} e^{-\frac{\langle A, \Delta(\mathbf{x},\mathbf{y})\rangle}{\beta}}\right) = -\beta \log\left(\sum_{i=0}^{G} n_i e^{-\frac{d_i}{\beta}}\right) \quad . \tag{4.41}$$

When $\mathbf{y} = \mathbf{x}$, we have $d_0 = 0$. Isolating the first element of the sum we get:

$$\mathbf{dtw}_\beta(\mathbf{x}, \mathbf{x}) = -\beta \log(n_0) - \beta \log\left(1 + \sum_{i=1}^{G} \frac{n_i}{n_0} e^{-\frac{d_i}{\beta}}\right) \leq -\beta \log(n_0) \quad . \tag{4.42}$$

Similarly, when $\mathbf{y}$ is temporally k-shifted with respect to $\mathbf{x}$, we also have $d_0 = 0$. Adding an exponent $'$ on terms that depend on the time series $\mathbf{x}_{+k}$, we have:

$$\begin{aligned}\mathbf{dtw}_\beta(\mathbf{x}, \mathbf{x}_{+k}) = -\beta \log(n'_0) - \beta \log\left(1 + \sum_{i=1}^{G} \frac{n'_i}{n'_0} e^{-\frac{d'_i}{\beta}}\right) &\geq -\beta \log(n'_0) - \beta \sum_{i=1}^{G} \frac{n'_i}{n'_0} e^{-\frac{d'_i}{\beta}} \\ &\geq -\beta \log(n'_0) - \beta D_{T,T} e^{-\frac{d'_1}{\beta}}\end{aligned} \tag{4.43}$$

However, since the set of values taken by $\Delta(\mathbf{x}, \mathbf{x})$ and $\Delta(\mathbf{x}, \mathbf{x}_{+k})$ are the same, we have $d_i = d'_i$ (but $n_i \neq n'_i$ apriori) and the assumption on β provides:

$$\begin{aligned}&\beta \leq \frac{\mu}{\log(3TD_{T,T})} \\ &\Rightarrow \beta \leq \frac{d_1}{\log(3TD_{T,T})} \\ &\Rightarrow e^{\frac{-d'_1}{\beta}} \leq \frac{1}{3TD_{T,T}} \\ &\Rightarrow -\beta D_{T,T} e^{\frac{-d'_1}{\beta}} \geq -\frac{\beta}{3T}\end{aligned} \tag{4.44}$$

Combining (4.42), (4.43) and (4.44) leads to:

$$\mathbf{dtw}_\beta(\mathbf{x}, \mathbf{x}_{+k}) - \mathbf{dtw}_\beta(\mathbf{x}, \mathbf{x}) \geq \beta \log\left(\frac{n_0}{n'_0}\right) - \frac{\beta}{3T} \tag{4.45}$$

Now let's develop the term $\frac{n_0}{n'_0}$. n'_0 corresponds to the number of equivalent alignments with 0 cost which can be given by $D_{\text{on}(\mathbf{x}),\text{on}(\mathbf{y})} \Omega D_{T-\text{off}(\mathbf{x}),T-\text{off}(\mathbf{y})}$, where Ω is the number of 0 cost alignments within the cross

product of the fluctuation sets. However, temporal shifts do not change Ω but only change the outermost sets. For instance, considering the example of Figure 4.2 one can see that only rectangles outside the fluctuation set are affected. Therefore, Ω cancels out in $\frac{n_0}{n_0'}$ and we get the desired bound. ■

Combining proposition 35 with the lower bound in 34 leads to the main theoretical result of this chapter:

Theorem 6 *Let $k \in [\![1, T-1]\!]$, let $m = \text{on}(\mathbf{x})$ and $m' = T - \text{off}(\mathbf{x})$.*
Let $\mu = \min_{i,j}\{\Delta(\mathbf{x},\mathbf{x})_{ij} | \Delta(\mathbf{x},\mathbf{x})_{ij} > 0\}$. If $0 < \beta \leq \frac{\mu}{\log(3TD_{T,T})}$:

$$\mathbf{dtw}_\beta(\mathbf{x}, \mathbf{x}_{+_k}) - \mathbf{dtw}_\beta(\mathbf{x}, \mathbf{x}) \geq P(k) \stackrel{\text{def}}{=} \beta\alpha k(k-1) + \beta\rho k \tag{4.46}$$

Where $\alpha = \frac{2-\sqrt{2}}{2}(\frac{1}{m'} + \frac{1}{m+m'}) > 0$ and $\rho = \frac{3\sqrt{2}-4}{3T} > 0$.

1.4 Setting β to control temporal sensitivity

While the previously considered time series covered a wide range of scenarios, the obtained result requires β to be too small, thereby not providing any insight on how $\mathbf{dtw}_\beta$ behaves when β increases. In the following paragraph, we relax this assumption on β in order to find a tighter lower bound than the one given in theorem 6. We consider the simplified setting of Dirac univariate time series $\mathbf{x}, \mathbf{y}$ such that $\mathbf{y}$ is ahead of $\mathbf{x}$ by k time steps (see Figure 4.1). Formally, let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^T$ such that for some $t^\star \in [\![1, T]\!]$ and $1 \leq k \leq T - t^\star$:

$$\begin{aligned} t \neq t^\star &\Rightarrow \mathbf{x}_t = 0 \\ t \neq t^\star + k &\Rightarrow \mathbf{y}_t = 0 \\ \mathbf{x}_{t^\star} &= \mathbf{y}_{t^\star+k} = c \in \mathbb{R} \end{aligned} \tag{4.47}$$

This simplified setting allows for tighter bounds of Soft-DTW.

Proposition 36 *Consider $\mathbf{x}$ and $\mathbf{y}$ as defined in (4.47). Let $r = \Delta(c, 0)$, $T \geq 6$, $c = 1 + \sqrt{2}$ and $\sigma = \frac{21c^2}{22} - 5$ ($\approxeq$ 0.56).Then:*

$$\mathbf{dtw}_\beta(\mathbf{x}, \mathbf{y}) - \mathbf{dtw}_\beta(\mathbf{x}, \mathbf{x}) \geq -\beta \log\left(e^{-P(k)}(1 - \lambda_\beta) + \lambda_\beta H\right) \tag{4.48}$$

where: $\lambda_\beta = e^{-\frac{r}{\beta}}$, $H = 92T^\sigma$ and P is the quadratic bound defined in Proposition 34.

PROOF. First let's upper bound $\mathbf{dtw}_\beta(\mathbf{x}, \mathbf{y})$. Notice that since $\mathbf{x}, \mathbf{y}$ are Dirac time series, the elements of the distance matrix $\Delta(\mathbf{x}, \mathbf{y})$ are either equal to 0 or r. Therefore, the cost of any path A given by $\langle A, \Delta(\mathbf{x}, \mathbf{y})\rangle$ can be written as qr for some $q \in \mathbb{N}$. More specifically, q corresponds to the number of times the path A meets the non-zero elements of $\Delta(\mathbf{x}, \mathbf{y})$. Therefore, denoting the number of feasible alignments

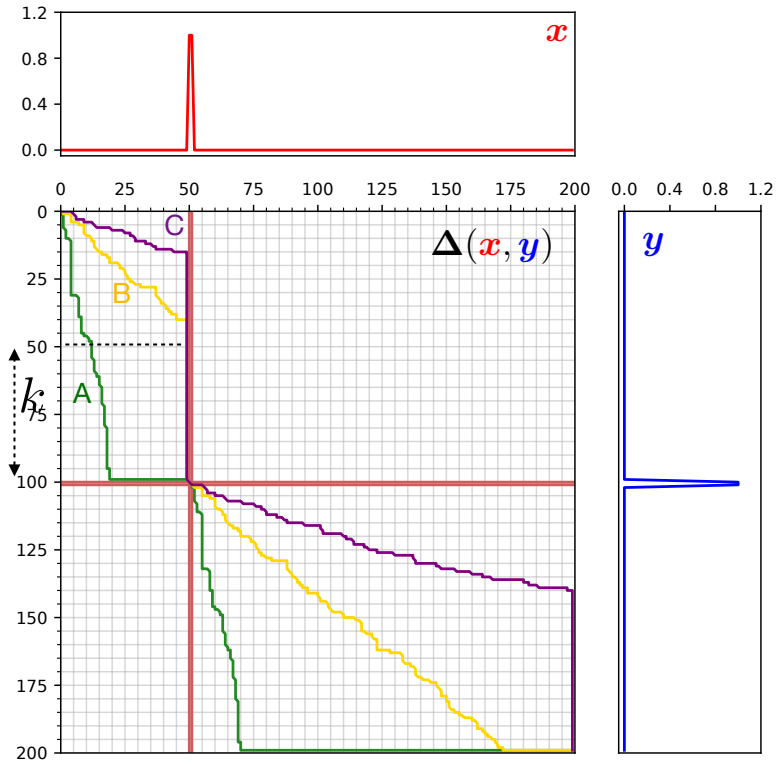


Fig. 4.3. Example of 3 DTW alignment paths (A, B and C) between **x** and **y** with a temporal 50-shift. The heatmap of the distance matrix Δ (here squared Euclidean) shows 2 red bars where the distance is not equal to 0 except at their intersection. An alignment path has 0 cost if and only if it does not cross the red lines.

corresponding to each q by $M_q(\mathbf{x}, \mathbf{y})$ it holds:

$$\sum_{A \in \mathcal{A}_{T,T}} e^{-\frac{\langle A, \Delta(\mathbf{x},\mathbf{y}) \rangle}{\beta}} = \sum_{q=0}^{T+k} M_q(\mathbf{x}, \mathbf{y}) e^{-\frac{qr}{\beta}} = M_0(\mathbf{x}, \mathbf{y}) + e^{-\frac{r}{\beta}} \sum_{q=1}^{T+k} M_k(\mathbf{x}, \mathbf{y}) e^{-\frac{r(q-1)}{\beta}} \tag{4.49}$$

Therefore:

$$\begin{aligned} \sum_{A \in \mathcal{A}_{T,T}} e^{-\frac{\langle A, \Delta(\mathbf{x},\mathbf{y}) \rangle}{\beta}} &\leq M_0(\mathbf{x}, \mathbf{y}) + e^{-\frac{r}{\beta}} \sum_{q=1}^{T+k} M_k(\mathbf{x}, \mathbf{y}) \\ &= M_0(\mathbf{x}, \mathbf{y}) + e^{-\frac{r}{\beta}} (D_T - M_0(\mathbf{x}, \mathbf{y})) \\ &= (1 - \lambda_\beta) M_0(\mathbf{x}, \mathbf{y}) + \lambda_\beta D_T \end{aligned}$$

and similarly:

$$\sum_{A \in \mathcal{A}_{T,T}} e^{-\frac{\langle A, \Delta(\mathbf{x},\mathbf{x}) \rangle}{\beta}} \geq M_0(\mathbf{x},\mathbf{x})$$

where $M_0(\mathbf{x},\mathbf{y}) = D_{t^\star-1,t^\star-1+k} D_{T-t^\star,T-t^\star-k}$ and $M_0(\mathbf{x},\mathbf{x}) = D_{t^\star-1} D_{T-t^\star}$. Therefore, combining both inequalities after introducing $-\log$ leads to:

$$\begin{aligned}
\mathbf{dtw}_\beta(\mathbf{x},\mathbf{y}) - \mathbf{dtw}_\beta(\mathbf{x},\mathbf{x}) &\geq -\beta \log\left(\frac{M_0(\mathbf{x},\mathbf{y})(1-\lambda_\beta) + \lambda_\beta D_T}{M_0(\mathbf{x},\mathbf{x})}\right) \\
&= -\beta \log\left(\frac{D_{t^\star-1,t^\star-1+k} D_{T-t^\star,T-t^\star-k}}{D_{t^\star-1} D_{T-t^\star}}(1-\lambda_\beta) + \lambda_\beta \frac{D_T}{D_{t^\star-1} D_{T-t^\star}}\right)
\end{aligned}$$

On one hand applying Proposition 34 with $m = t^\star - 1$ and $m' = T - m - 1$ provides:

$$\frac{D_{t^\star-1,t^\star-1+k} D_{T-t^\star,T-t^\star-k}}{D_{t^\star-1} D_{T-t^\star}} \leq e^{-P(k)}$$

And on the other, using Corollary 4 we can get the H upper bound. For $t^\star \geq 1$:

$$\frac{D_T}{D_{T-t^\star}} \leq \sqrt{\frac{(t^\star-1)e}{T}} \frac{1}{c^{2(T-t^\star+1)}} \leq \sqrt{\frac{t^\star e}{T}} c^{2t^\star}$$

and if $t^\star - 1 \geq 5$:

$$\frac{D_5}{D_{t^\star-1}} \leq \left(\frac{T-t^\star-1}{4}\right)^\sigma c^{2(T-t^\star-5)} \leq \left(\frac{T-t^\star}{4}\right)^\sigma \frac{1}{c^{2(t^\star-6)}}$$

Combining the two leads to:

$$\frac{D_T}{D_{t^\star-1} D_{T-t^\star}} \leq \frac{c^{12}}{D_5} \left(\frac{T-t^\star}{4}\right)^\sigma \sqrt{\frac{t^\star e}{T}}$$

Maximizing the upper bound with respect to $t^\star$ leads to the maximizer $t^\star = \frac{T}{2\sigma+1}$. Substituting shows that:

$$(T-t^\star)^\sigma \sqrt{\frac{t^\star}{T}} \leq \left(T \frac{2\sigma}{1+2\sigma}\right)^\sigma \sqrt{\frac{1}{1+2\sigma}}$$

Finally, numerical evaluation gets rid of the constants:

$$\frac{D_T}{D_{t^\star-1} D_{T-t^\star}} \leq H$$

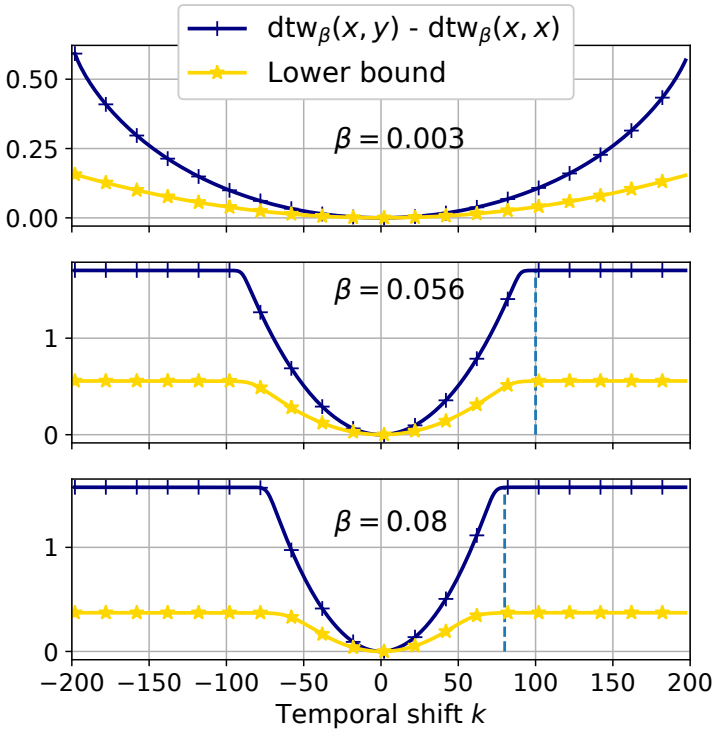


Fig. 4.4. Lower bound P of $\mathbf{dtw}_\beta$ along with the $k_{\max}$ value given by proposition 37.

If however $t^\star \leq 5$:

$$\frac{D_T}{D_{t^\star-1}D_{T-t^\star}} \leq \sqrt{\frac{t^\star e}{T}}\left(\frac{c^2}{3}\right)^{t^\star} \leq \sqrt{e}\left(\frac{c^2}{3}\right)^{5} \approx 45.63 < 92 < H$$

Multiplying by $1 - \lambda_\beta \geq 0$, adding $\lambda_\beta H$ and applying $-\log$ ends the proof ■.

Effect of β When β is small enough, λ_β goes to 0, thus the lower bound of Proposition 36 can be very well approximated by the quadratic bound $P(k)$. However when β increases, $\lambda_\beta H$ increases which will dominate the log argument for a sufficiently large k. Using the example of Figure 4.1, we compute both sides of Equation (4.48) for 3 values of β. Figure 4.4 shows that $\mathbf{dtw}_\beta$ saturates for a certain temporal shift $k_{\max}$ beyond which it is no longer sensitive to temporal lags. This phase transition is also observed by the lower bound (4.48). This provides a heuristic to set β based on a predefined $k_{\max}$ corresponding to the largest temporal shift the user is willing to capture. Notice that such a point does not always exist (when β is too small) as it may be larger than the time series length T (see top example, Figure 4.4).

Proposition 37 *Let $T \geq 6$, $1 \leq k_{\max}$ and $0 < \eta < 1$. Using the same notations of Proposition 36, define the lower bound function:*

$$\mathrm{LB}_\beta : k \mapsto -\beta \log\left(e^{-P(k)}(1-\lambda_\beta) + \lambda_\beta H\right)$$

If $\beta \geq \frac{r}{P(k_{\max}) + \log((e^\eta - 1)H)}$ then:

$$0 \leq \frac{\lim\limits_{k \to +\infty} \mathrm{LB}_\beta(k) - \mathrm{LB}_\beta(k_{\max})}{\beta} \leq \eta \tag{4.50}$$

PROOF. It is straightforward to see that $\lim\limits_{k \to +\infty} \mathrm{LB}_\beta(k) = -\beta \log(\lambda_\beta H)$. Therefore on one hand:

$$\begin{aligned}
&\frac{\lim\limits_{k \to +\infty} \mathrm{LB}_\beta(k) - \mathrm{LB}_\beta(k_{\max})}{\beta} \leq \eta \\
\Leftrightarrow &\log\left(e^{-P(k)}(1-\lambda_\beta) + \lambda_\beta H\right) - \log(\lambda_\beta H) \leq \eta \\
\Leftrightarrow &e^{-P(k_{\max})}(1-\lambda_\beta) + \lambda_\beta H \leq e^\eta \lambda_\beta H \\
\Leftrightarrow &\lambda_\beta \geq \frac{e^{-P(k_{\max})}}{(e^\eta - 1)H + e^{-P(k_{\max})}}
\end{aligned}$$

On the other hand:

$$\begin{aligned}
&\beta \geq \frac{r}{P(k_{\max}) + \log\left((e^\eta - 1)H\right)} \\
\Rightarrow &-\frac{r}{\beta} \geq -P(k_{\max}) - \log\left((e^\eta - 1)H\right) \\
\Rightarrow &\lambda_\beta \geq \frac{e^{-P(k_{\max})}}{(e^\eta - 1)H} \geq \frac{e^{-P(k_{\max})}}{(e^\eta - 1)H + e^{-P(k_{\max})}}
\end{aligned}$$

Thus the upper bound in (4.50) holds. The lower bound follows from the positivity of $e^{-P(k)}(1-\lambda_\beta)$ ■.

Proposition 37 provides a sufficient condition to set β such that the lower bound LB saturates for a certain $k_{\max}$. In the examples shown in Figure 4.4, β was set using this heuristic with $\eta = 0.01$ and $k_{\max} \in \{500, 100, 80\}$ respectively top to bottom. The dotted vertical lines highlight the choice of $k_{\max}$ which is very close to the saturation point of $\mathbf{dtw}_\beta$. In practice, we use the same heuristic by setting $r = \max_{i,j} \Delta(\mathbf{x}_i, \mathbf{y}_j)$.

2 OT in space

To perform temporal matching across different time points of spatio-temporal data, we take advantage of the *debiased unbalanced entropic OT* divergence $\mathrm{S}^{\mathcal{U}}_{\varepsilon,\gamma}$ introduced in Chapter 2. Provided some minor assumptions, this divergence is non-negative and its Fréchet means $\arg\min \sum_{k=1}^K w_k \mathrm{S}^{\mathcal{U}}_{\varepsilon,\gamma}(\mathbf{a}_k, \mathbf{b})$ can be

computed in parallel on GPUs. For the sake of clarity, we provide a brief reminder of the its definition and main properties.

2.1 Unbalanced entropic OT

When data are endowed with geometrical properties, OT metrics such as the Wasserstein distance can capture spatial variations between probability distributions. Given a transport cost function – commonly referred to as ground metric – the Wasserstein distance computes the optimal transportation plan between two measures. Its heavy computational cost can be significantly reduced by using entropy regularization (Cuturi, 2013). Additionally, it is also possible to extend its definition to handle measures with different total mass using the unbalanced OT formulation of Chizat et al. (2018b), which also relies on entropic regularization, pending some minor modifications to Sinkhorn's algorithm (Frogner et al., 2015). Formally, Let α, β be two non-negative measures with a fixed support given by $\mathcal{X} = \{x_1, \dots, x_p\} \subset \mathbb{R}^d$. They can be identified with vectors of non-negative weights $\mathbf{a}, \mathbf{b} \in \mathbb{R}_+^p$. Let $\mathbf{C}$ be the cost matrix filled with entries $\mathbf{C}_{ij} = c(x_i, x_j)$ for some non-negative symmetric cost function $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ such that $\mathbf{K} \stackrel{\text{def}}{=} e^{-\frac{\mathbf{C}}{\varepsilon}}$. Denoting $\mathcal{U}$ the uniform non-negative measure in $\mathcal{X}^2$ assigning the weight 1 to each (x_i, x_j), we define:

$$\underset{\varepsilon,\gamma}{\overset{\mathcal{U}}{\text{UOT}}}(\alpha, \beta) = \min_{\pi \in {\mathbb{R}_+}^{p \times p}} \langle \pi, \mathbf{C} \rangle + \varepsilon \text{KL}(\pi | \mathcal{U}) + \gamma \text{KL}(\pi \mathbb{1} | \mathbf{x}) + \gamma \text{KL}(\pi^\top \mathbb{1} | \mathbf{y}) \tag{4.51}$$

$\text{UOT}^{\mathcal{U}}_{\varepsilon,\gamma}$ is however not positive. Moreover, its barycenters are biased towards smooth and blurred measures which inevitably neglect the fine-grained aspects of the data. To reduce this effect, we propose the following divergence:

$$\text{S}^{\mathcal{U}}_{\varepsilon,\gamma}(\alpha, \beta) \stackrel{\text{def}}{=} \underset{\varepsilon,\gamma}{\overset{\mathcal{U}}{\text{UOT}}}(\alpha, \beta) - \frac{1}{2}\left(\underset{\varepsilon,\gamma}{\overset{\mathcal{U}}{\text{UOT}}}(\alpha, \alpha) + \underset{\varepsilon,\gamma}{\overset{\mathcal{U}}{\text{UOT}}}(\beta, \beta)\right) \ . \tag{4.52}$$

2.2 Debiased spatial barycenters

The following proposition regroups all relevant properties of $\text{S}^{\mathcal{U}}_{\varepsilon,\gamma}$ shown in Chapter 2.

Proposition 38 *Let $\mathbf{a}_1, \dots, \mathbf{a}_K, \mathbf{a}, \mathbf{b} \in \mathbb{R}_+^p$ and $w_1, \dots, w_K$ a sequence of positive weights adding to 1. Assume that $\mathbf{K}$ is positive semi-definite. Let $\mathcal{J}\mathbf{b} \mapsto \sum_{k=1}^K w_k \text{S}^{\mathcal{U}}_{\varepsilon,\gamma}(\mathbf{a}_k, \mathbf{b})$.*

1. *$\text{S}^{\mathcal{U}}_{\varepsilon,\gamma}$ is differentiable, non-negative and coercive but not convex.*
2. *If $\mathbf{K}$ is positive definite then $\text{S}^{\mathcal{U}}_{\varepsilon,\gamma}(\mathbf{a}, \mathbf{b}) = 0 \Leftrightarrow \mathbf{a} = \mathbf{b}$.*
3. *The Fréchet mean $\bar{\mathbf{b}} = \arg\min_{\mathbf{b}} \mathcal{J}(\mathbf{b})$ is well-defined and is equivalent to $\nabla \mathcal{J}(\bar{\mathbf{b}}) = \mathbf{0}$*

Cancelling the gradient of $\mathcal{J}$ is equivalent to solving the fixed point system in $(\mathbf{u}_k)_k, (\mathbf{v}_k)_k, \mathbf{c}, \bar{\mathbf{b}}$:

$$\mathbf{u}_k = \left(\frac{\mathbf{a}_k}{\mathbf{K}\mathbf{v}_k}\right)^\omega, \quad \mathbf{v}_k = \left(\frac{\bar{\mathbf{b}}}{\mathbf{K}^\top \mathbf{u}_k}\right)^\omega, \quad \mathbf{c} = \left(\frac{\bar{\mathbf{b}}}{\mathbf{K}\mathbf{c}}\right)^\omega, \quad \bar{\mathbf{b}} = \mathbf{c}^{\frac{1}{\omega}} \left(\sum_{k=1}^K w_k (\mathbf{K}^\top \mathbf{u}_k)^{1-\omega}\right)^{\frac{1}{1-\omega}} \tag{4.53}$$

An iterating algorithm to solve the system (4.53) is provided below.

Algorithm 11 Debiased unbalanced $\mathrm{S}^{\mathcal{U}}_{\varepsilon,\gamma}$ barycenter.

Input: $\mathbf{a}_1, \ldots, \mathbf{a}_K \in \mathbb{R}_+^p$, **parameters** $\varepsilon, \gamma > 0$, $\mathbf{K} \overset{\text{def}}{=} e^{-\frac{\mathbf{C}}{\varepsilon}}$
Output: $\bar{\mathbf{b}}$, **the** $\mathrm{UOT}^{\mathcal{U}}_{\varepsilon,\gamma}$ **barycenter of** $(\mathbf{a}_1, \ldots, \mathbf{a}_K)$
Initialize $\mathbf{c} = \mathbf{v}_1 = \cdots = \mathbf{v}_K = \mathbb{1}_p$, **set** $\omega = \frac{\gamma}{\gamma+\varepsilon}$
while Not converged **do**
  for $k = 1$ **to** K **do**
    $\mathbf{u}_k = \left(\frac{\mathbf{a}_k}{\mathbf{K}\mathbf{v}_k}\right)^\omega$
  end for
  $\bar{\mathbf{b}} = \mathbf{c}^{\frac{1}{\omega}} \left(\sum_{k=1}^K w_k (\mathbf{K}^\top \mathbf{u}_k)^{1-\omega}\right)^{\frac{1}{1-\omega}}$
  for $k = 1$ **to** K **do**
    $\mathbf{v}_k = \left(\frac{\bar{\mathbf{b}}}{\mathbf{K}^\top \mathbf{u}_k}\right)^\omega$
  end for
  $\mathbf{c} = \left(\frac{\bar{\mathbf{b}}}{\mathbf{K}\mathbf{c}}\right)^\omega$
end while

3 OT in space and time

3.1 The spatio-temporal loss and barycenters

$\mathrm{S}^{\mathcal{U}}_{\varepsilon,\gamma}$ is coercive with respect to each of its arguments, moreover, if $\mathbf{K} \overset{\text{def}}{=} e^{-\frac{\mathbf{C}}{\varepsilon}}$ is positive semi-definite, then $\mathrm{S}^{\mathcal{U}}_{\varepsilon,\gamma}$ is non-negative. Therefore, $\mathbf{dtw}_\beta$ is well defined with $\mathrm{S}^{\mathcal{U}}_{\varepsilon,\gamma}$ as a cost function, a loss function we named STA: *Spatio-Temporal Alignment*.

Definition 6 (STA) *We define the STA loss as:*

$$\mathbf{sta}_\beta(\mathbf{x}, \mathbf{y}) = \mathbf{dtw}_\beta(\mathbf{x}, \mathbf{y}; \mathrm{S}^{\mathcal{U}}_{\varepsilon,\gamma}) \tag{4.54}$$

With $\mathrm{S}^{\mathcal{U}}_{\varepsilon,\gamma}$ as a cost function, two time points $\mathbf{x}_i, \mathbf{y}_j$ would be temporally aligned if they are close spatially, formally, if $\mathrm{S}^{\mathcal{U}}_{\varepsilon,\gamma}(\mathbf{x}_i, \mathbf{y}_j)$ is small enough.

Consider a dataset with N multivariate time series $\mathbf{x}_1, \dots, \mathbf{x}_N$ assumed to have the same dimension p and respective time lengths $T_1, \dots, T_N$. Let $w_1, \dots, w_n$ be a set of positive weights summing to one. The Soft-DTW barycenter with cost $\mathrm{S}^{\mathcal{U}}_{\varepsilon,\gamma}$ and fixed length T is defined as:

$$\bar{\mathbf{x}} = \underset{\mathbf{x}\in\mathbb{R}^{p,T}}{\arg\min} \sum_{i=1}^{N} w_i \mathbf{sta}_\beta(\mathbf{x}_i, \mathbf{x}) \tag{4.55}$$

$$= \underset{\mathbf{x}\in\mathbb{R}^{p,T}}{\arg\min} \sum_{i=1}^{N} w_i \mathbf{dtw}_\beta(\mathbf{x}_i, \mathbf{x}; \mathrm{S}^{\mathcal{U}}_{\varepsilon,\gamma}) \tag{4.56}$$

$$= \underset{\mathbf{x}\in\mathbb{R}^{p,T}}{\arg\min} - \sum_{i=1}^{N} w_i \beta \log \left(\sum_{A\in\mathcal{A}_{T,T_i}} e^{-\frac{\langle A, \mathrm{S}^{\mathcal{U}}_{\varepsilon,\gamma}(\mathbf{x},\mathbf{x}_i)\rangle}{\beta}} \right) \tag{4.57}$$

Alternating optimization Since $\mathrm{S}^{\mathcal{U}}_{\varepsilon,\gamma}$ is differentiable, the most straightforward solution to (4.55) would probably be to use a Quasi-Newton method. However, computing each gradient step would require $T\sum_{i=1}^{N} T_i$ Sinkhorn runs. Instead, we use Fenchel duality to obtain an alternating optimization problem that non only avoids the computation of the gradients of $\mathrm{S}^{\mathcal{U}}_{\varepsilon,\gamma}$ but also spares us any form of step-size backtracking. This is given by proposition 39.

Proposition 39 *Denote the sets of binary matrices $\mathcal{A}_{T_i,T}$ with some arbitrary indexation $\mathcal{A}_{T_i,T} = \{A_i^1, \dots, A_i^{D_{T_i,T}}\}$ and let S_K denote the probability simplex of $\mathbb{R}^K$. For any coercive cost function Δ, the Soft-DTW problem (4.55) is equivalent to the alternating optimization problem:*

$$\min_{\mathbf{x}\in\mathbb{R}^{p,T}} \min_{\substack{\theta_1\in S_{D_{T_1,T}} \\ \dots \\ \theta_N\in S_{D_{T_N,T}}}} \sum_{i=1}^{N} \langle \sum_{k=1}^{D_{T_k,T}} w_i \theta_i^k A_i^k, \Delta(\mathbf{x}_i, \mathbf{x})\rangle + \beta H(\theta_i) \tag{4.58}$$

PROOF. A standard result in convex optimization theory states that the Fenchel conjugate of entropy is logsumexp. Formally, for $\mathbf{x} \in \mathbb{R}^K$:

$$(\beta H)^\star(\mathbf{x}) \overset{\text{def}}{=} \max_{\theta\in S_K} \langle \mathbf{x}, \theta\rangle - \beta H(\theta) = \beta \log \left(\sum_{k=1}^{K} e^{-\frac{\mathbf{x}_k}{\beta}} \right) \tag{4.59}$$

Thus:

$$-\beta \log \left(\sum_{k=1}^{K} e^{-\frac{\mathbf{x}_k}{\beta}} \right) = \min_{\theta\in S_K} \langle -\mathbf{x}, \theta\rangle + \beta H(\theta). \tag{4.60}$$

Therefore, the barycenter loss (4.55) can be written:

$$\min_{\mathbf{x}\in\mathbb{R}^{p,T}} \sum_{i=1}^{N} w_i \mathbf{dtw}_\beta(\mathbf{x}_i, \mathbf{x}) = \min_{\mathbf{x}\in\mathbb{R}^{p,T}} \sum_{i=1}^{N} -w_i\beta \log\left(\sum_{A\in\mathcal{A}_{T,T_i}} e^{-\frac{\langle A,\Delta(\mathbf{x},\mathbf{x}_i)\rangle}{\beta}}\right) \tag{4.61}$$

$$= \min_{\mathbf{x}\in\mathbb{R}^{p,T}} \sum_{i=1}^{N} w_i \min_{\theta_i\in S_{D_{T_i,T}}} \langle \theta_i, \Big(\langle A_i^1, \Delta(\mathbf{x}_i,\mathbf{x})\rangle, \dots, \langle A^{D_{T_i,T}}{}_i, \Delta(\mathbf{x}_i,\mathbf{x})\rangle\Big)\rangle + \beta H(\theta_i) \tag{4.62}$$

$$= \min_{\mathbf{x}\in\mathbb{R}^{p,T}} \sum_{i=1}^{N} w_i \min_{\theta_i\in S_{D_{T_i,T}}} \sum_{k=1}^{D_{T_k,T}} w_i \langle \theta_i^k A_i^k, \Delta(\mathbf{x}_i,\mathbf{x})\rangle + \beta H(\theta_i) \tag{4.63}$$

$$= \min_{\mathbf{x}\in\mathbb{R}^{p,T}} \min_{\substack{\theta_1\in S_{D_{T_1,T}} \\ \theta_N\in \dot{\dots} S_{D_{T_N,T}}}} \sum_{i=1}^{N} \langle \sum_{k=1}^{D_{T_k,T}} w_i\theta_i^k A_i^k, \Delta(\mathbf{x}_i,\mathbf{x})\rangle + \beta H(\theta_i) \ , \tag{4.64}$$

where the last equality follows from the separability of the sum with respective to the θ_i ■

The major benefit of the dual formulation of Proposition (39) is the ability to compute Fréchet means of Δ directly. This will be in particular crucial when we define Δ as $S^{\mathcal{U}}_{\varepsilon,\gamma}$ divergence, briefly revisited in the following section. $S^{\mathcal{U}}_{\varepsilon,\gamma}$ Fréchet means can be computed using the variant of Sinkhorn's algorithm studied in chapter 2. These Sinkhorn-type algorithms are known to be orders of magnitude faster than gradient based methods (Cuturi and Peyré, 2018). While minimizing with respect to the θ_i seems computationally unfeasible due their large dimension, their update is actually not required to compute the new $\mathbf{x}$. Instead, one needs to update the matrices $\mathbf{Z}_i \stackrel{\text{def}}{=} \sum_{k=1}^{D_{T_k,T}} w_i\theta_i^k A_i^k$ which are *exactly* given by the gradients $\frac{\partial \mathbf{dtw}_\beta(\mathbf{x}_i,\mathbf{x})}{\partial\Delta}(\mathbf{x}_i,\mathbf{x})$. Indeed, given that the loss is convex in θ_i, for a fixed $\mathbf{x}$, the optimal θ_i verifies the KKT conditions for some Lagrange multiplier λ_i:

$$\begin{cases} \langle A_i^k, \Delta(\mathbf{x}_i,\mathbf{x})\rangle + \beta\log(\theta_i^k) - \lambda_i = 0 \\ \sum_{k=1}^{D_{T_i,T}} \theta_i^k = 1 \end{cases}$$

which leads to:

$$\theta_i^k = \frac{e^{-\frac{\langle A_i^k,\Delta(\mathbf{x}_i,\mathbf{x})\rangle}{\beta}}}{\sum_{k=1}^{D_{T_k,T}} e^{-\frac{\langle A_i^k,\Delta(\mathbf{x}_i,\mathbf{x})\rangle}{\beta}}} \tag{4.65}$$

Thus:

$$\mathbf{Z}_i \stackrel{\text{def}}{=} \sum_{k=1}^{D_{T_k,T}} w_i\theta_i^k A_i^k = \frac{\sum_{k=1} e^{-\frac{\langle A_i^k,\Delta(\mathbf{x}_i,\mathbf{x})\rangle}{\beta}} A_i^k}{\sum_{k=1}^{D_{T_k,T}} e^{-\frac{\langle A_i^k,\Delta(\mathbf{x}_i,\mathbf{x})\rangle}{\beta}}} = \frac{\partial \mathbf{dtw}_\beta(\mathbf{x}_i,\mathbf{x})}{\partial\Delta}(\mathbf{x}_i,\mathbf{x}) \ , \tag{4.66}$$

which can be computed using Algorithm 10. Notice that to update $\mathbf{x}$ computing the θ_i is not necessary, it is sufficient to update the $\mathbf{Z}_i$. This leads to algorithm 12.

Initialization It is important to keep in mind the loss (4.58) is not jointly convex in $\mathbf{x}$ and θ. Thus, algorithm 12 is not guaranteed to converge to a global minimum. Nevertheless, in our experiments, initializing $\mathbf{x}$ with a uniform distribution leads to meaningful barycenters with the desired spatio-temporal properties.

Algorithm 12 Soft-DTW barycenter.

Input: $\mathbf{x}_1, \dots, \mathbf{x}_N$, $\mathbf{x}_0$, weights $w_1, \dots, w_N$, parameter β
Output: solution of (4.58)
Initialize $\mathbf{x} = \mathbf{x}_0 \in \mathbb{R}^{p,T}$, **compute** $\Delta(\mathbf{x}_i, \mathbf{x})$ **for all** $i = 1..N$
while not converged **do**
  for $i = 1$ **to** N **do**
    Update $\mathbf{Z}_i$ with Algorithm 10
  end for
  for $t = 1$ **to** T **do**
    $\mathbf{x}^t = \arg\min_{\mathbf{a} \in \mathbb{R}^p} \sum_{i=1}^{N} \sum_{t=1}^{T_i} w_i \mathbf{Z}_i^t \Delta(\mathbf{x}_i^t, \mathbf{a})$
  end for
end while

3.2 Experiments

We perform several experiments on both synthetic and real data to answer two main questions:

1. Can $\mathbf{sta}_\beta$ discriminate between several classes with spatio-temporal differences ?
2. Is the $\mathbf{sta}_\beta$ barycenter robust to the spatio-temporal variability of the data ?

Discriminative power of $\mathbf{sta}_\beta$ Brain imaging data recordings report the brain activity both in space and time. Thanks to their high temporal resolution, Electroencephalography and magnetoencephalography can capture response latencies in the order of a millisecond. Abnormal differences in latency, amplitude and topography of brain signals are important biomarkers of several conditions of the central nervous system such as multiple sclerosis (Whelan et al., 2010) or amblyopia (Sokol, 1983). We argue here that $\mathbf{sta}_\beta$ can aggregate all these differences in a meaningful dissimilarity score. To illustrate this, we use the average brain surface derived from MRI scans and provided by the FreeSurfer software (Fischl, 2012). We compute a triangulated mesh of 642 vertices on the left hemisphere and simulate 4 types of signals as follows. We set $T = 20$ and select 2 activation time points $t_1 = 5$ and $t_2 = 15$. We also select two brain regions in the visual cortex given by FreeSurfer's segmentation known as *V1* (primary visual cortex) and *MT* (middle temporal visual area) which are defined on 17 and 8 vertices respectively. Each generated time series peaks at t_1 or t_2, in a random vertex in V1 or MT with a random amplitude between 1 and 3. For the signals to be more realistic, we apply a Gaussian filter along the temporal and the spatial axes. Examples of the generated data are displayed in Figure 4.5. We generate $N = 200$ samples (50 per time point / brain region) and compute the pairwise dissimilarity matrices $\mathbf{dtw}_\beta$ and

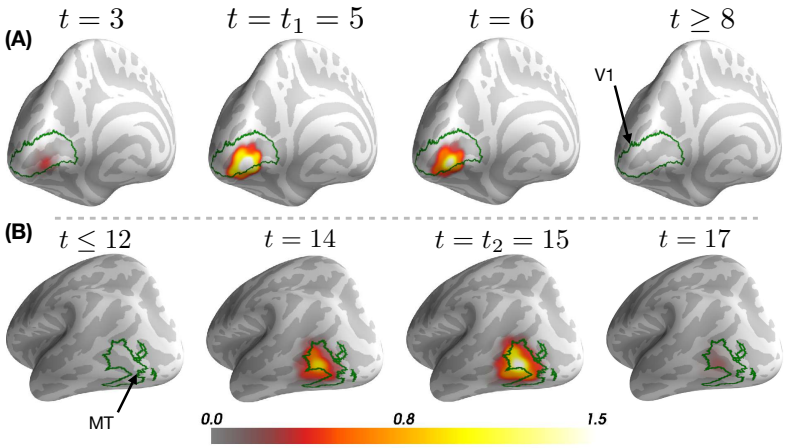


Fig. 4.5. Two examples of the simulated time series. **(A)** brain signal in V1 with a peak at $t = t_1$. **(B)** brain signal in MT with a peak at $t = t_2$. The borders of the brain regions V1 and MT are highlighted in green.

$\mathbf{sta}_\beta$ with $\beta = 0$ and $\beta = 0.1$. Figure 4.6 shows the t-distributed Stochastic Neighbor Embedding (t-SNE) (Maaten and Hinton, 2008; Pedregosa et al., 2011) of the data. As expected, $\mathbf{dtw}_\beta$ cannot capture spatial variability regardless of β. With $\beta = 0$, $\mathbf{sta}_\beta$ separates the data according to the brain region only. Only with positive β can $\mathbf{sta}_\beta$ identify all four groups. Computing the full $\mathbf{sta}_\beta$ dissimilarity matrix performed $\frac{1}{2}N(N+1) \times T^2 = 8040000$ Sinkhorn loops between 642 dimensional inputs. The whole experiment completed in 10 minutes on our DGX-1 station. Python code and data can be found in https://github.com/hichamjanati/spatio-temporal-alignements.

Averaging of real brain imaging data Studying the function of the various regions of the Human brain is one of the primary goals of neuroimaging research. These studies usually involve a group of healthy individuals (subjects) or patients who perform a series of tasks while having their neural activity recorded from which active regions of the brain are localized. However, drawing conclusions at a population level requires an aggregation function that combines the individual active sources of each subject. While averaging may seem like a straightforward and simple solution, it does not take into account the anatomical differences across subjects which lead to spatially blurred means. Moreover, the brain responses of the different subjects are never synced in time, specially when working with Electro-encephalogrphy (EEG) or Magneto-encephalography (MEG) data which have a high temporal resolution of the order of 1 millisecond. We use public the EEG/MEG dataset DS117 (Wakeman and Henson, 2015) and compute the spatio-temporal source configuration of 6 subjects who were shown images of Human faces using MNE-Python (Gramfort et al., 2013c). Here the support of our measures $\mathcal{A}$ is taken to be the set of 642 vertices that define the cortical mesh of the brain. The OT ground metric $\mathbf{C}$ is defined as the quadratic length of the shortest path on the triangulated mesh. We compute 3 different averages: a Euclidean mean, a $\mathrm{S}^{\mathcal{U}}_{\varepsilon,\gamma}$ barycenter (independently across time) and a spatio-temporal STA

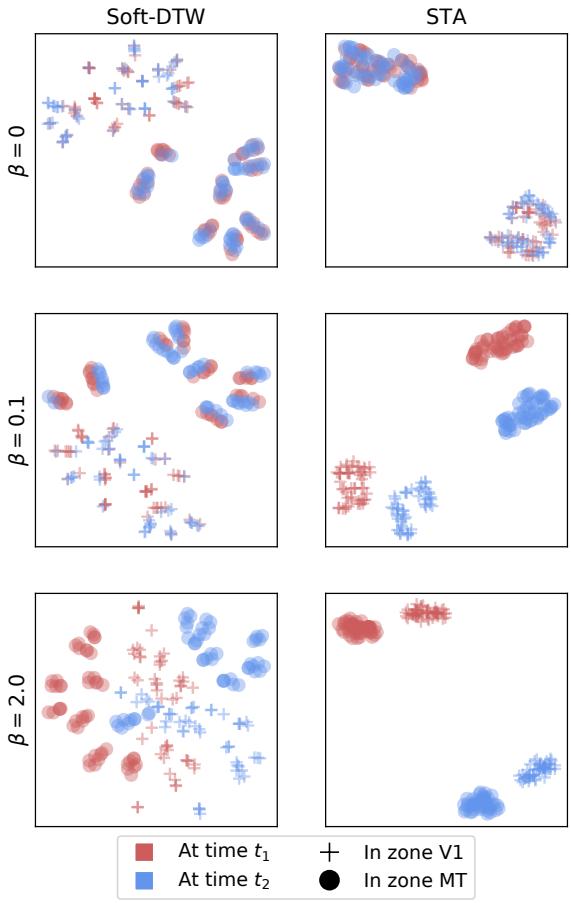


Fig. 4.6. t-SNE visualization of the simulated brain signals in two different regions, at two different time instants. With $\beta > 0$, $\mathbf{sta}_\beta$ can discriminate between all four groups. Increasing β leads to a higher temporal sensitivity.

barycenter with $k_{\max} = 20$. As shown in Figure 4.7, the first burst in the neural response is a visually evoked potential (known as P1) that arises around 100ms after the stimulus (Slotnick et al., 1999) in the primary visual cortex (blue). Then, at around 170 ms, an evoked response that is specific to the display of faces occurs in a small region known as the Fusiform Face Area (Green) (FFA) (Bentin et al., 1996; Kanwisher, McDermott, and Chun, 1997b). The delimited regions of interest were selected using the meta-analysis tool Neurosynth (Yarkoni, 2014).

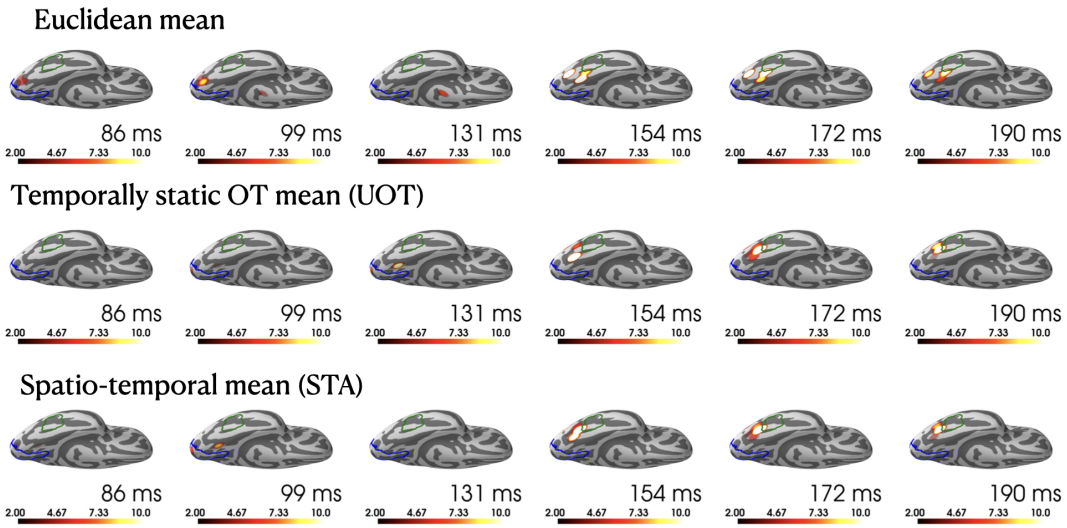


Fig. 4.7. Barycenters of the spatio-temporal neural activity of 6 subjects taken from the DS117 dataset. The STA barycenter shows a more focused activation around the Fusiform Form Face Area (green) than the other methods. Unlike the OT barycenter, STA shows a more plausible time occurrence of the first evoked response around 100ms.

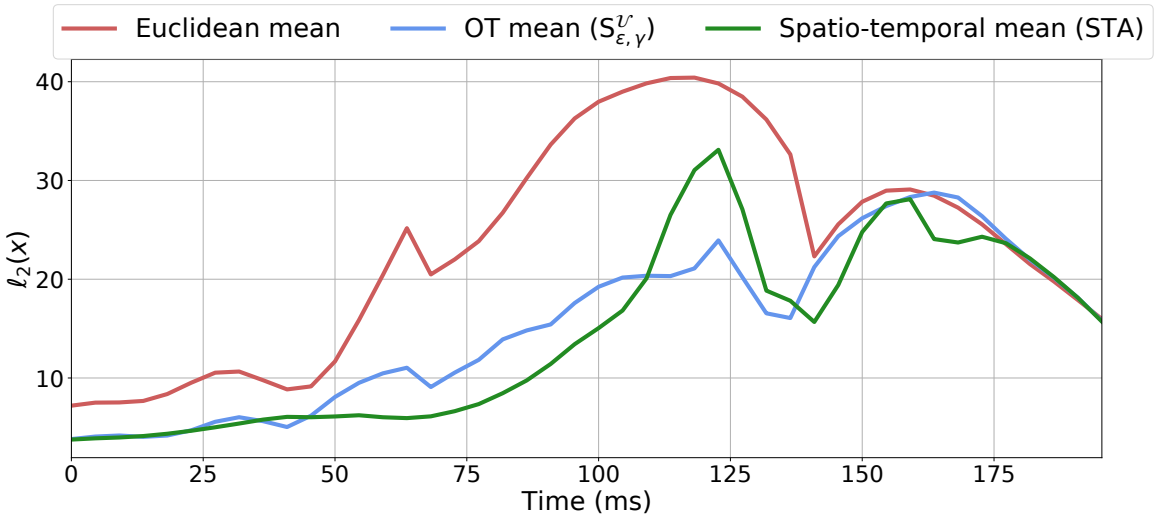


Fig. 4.8. The ℓ_2 norm (across space) of the barycenters displayed Figure 4.7 shows a clear temporal sensitivity of STA as it identifies the two expected evoked responses to visual facial stimulus.

To further assess the temporal sensitivity of STA, we display in Figure 4.8 the ℓ_2 norm across space of the 3 barycenters of Figure 4.7. The two evoked responses are more pronounced when using the STA barycenter.

Forecasting the motion of handwritten letters

Dataset We evaluate the performance of STA in a prediction task using a publicly available dataset of handwritten letters where the position of a pen are tracked in time (Williams, M.Toussaint, and Storkey., 2006). We subsample the data both spatially and temporally so as to keep 13 time points of (30×30) images for each time series. Each image can thus be seen as a screenshot at a certain time during the writing motion. To make the task a bit more challenging, we randomly shift each time series spatially (resp. temporally) by 0 to 10 pixels in each direction (by keeping 5 to 13 time points evenly selected). The dataset is composed of 20 samples of each one of the letters ("a", "b", "c", "v"), thus the full shape of the dataset is (100, 13, 30, 30).

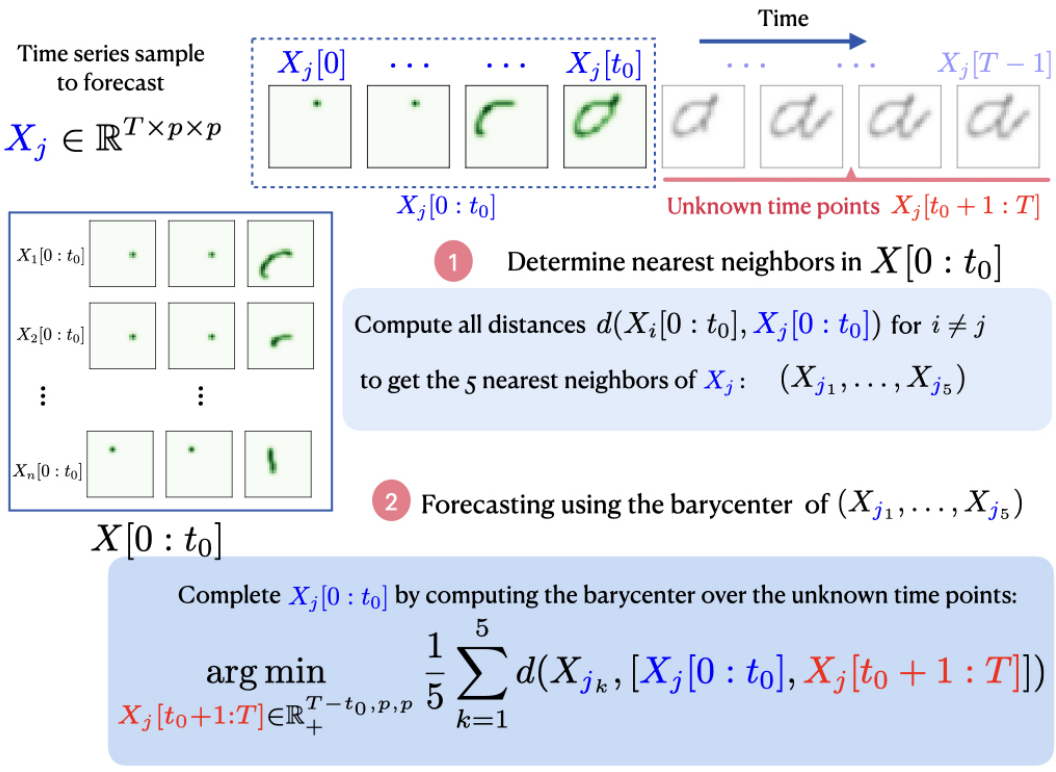


Fig. 4.9. Sketch explaining the forecasting pipeline used with the handwritten letters experiment.

Forecasting We propose to use barycenters as a forecasting method. For each time series $\mathbf{x}$ in the dataset, knowing only the first $t_0 < 13$ time points, we would like to predict the rest. First, based only on the observed $t_0 = 5$ time points, we select the closest 5 neighbors of $\mathbf{x}$ in the data based on some loss function d. We denote these nearest neighbors $\mathbf{x}'_1, \dots, \mathbf{x}'_5$. Next, we predict the future of $\mathbf{x}$ by computing the d-barycenter of $(\mathbf{x}'_k)_{k=1..5}$ while keeping the first t_0 observations of $\mathbf{x}$ fixed. The full pipeline is illustrated in Figure 4.9. The predictions obtained for the example shown in Figure 4.9 are illustrated in Figure 4.10. While ℓ_2 based method clearly fail to identify neighbors in the same class ("a"), OT based methods do not. Moreover, thanks to temporal variability, STA provides a more accurate prediction of the remaining time points than OT alone.

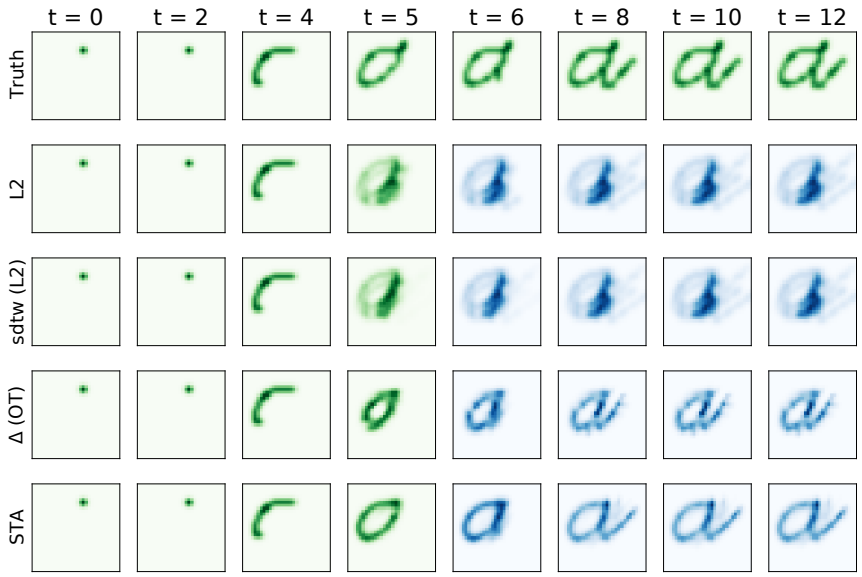


Fig. 4.10. Forecasts of a handwritten letter time series. The green time points are fixed and considered known for all models. Blue observations are predicted. As expected L2 based methods fail to identify neighbors in the same class.

Figure 4.11 shows a more quantitative comparison where we evaluate the accuracy of the predictions for all samples in the dataset with the ℓ_2 and the EMD (Earth mover distance) metrics averaged across the 8 predicted time points. To compute the EMD scores, we normalize all images so that their values add up to 1 and define EMD with the Euclidean quadratic cost between pixel coordinates. On both metrics and for all letters, *STA* outperforms the other loss functions.

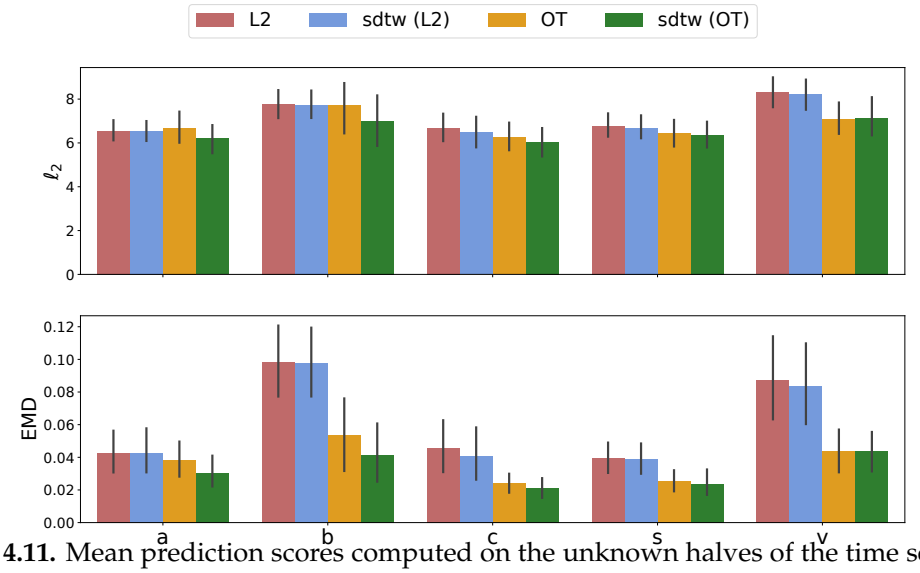


Fig. 4.11. Mean prediction scores computed on the unknown halves of the time series for each letter. The EMD score is computed between each true image and its prediction after normalization to the probability simplex. The ground metric is the Euclidean distance between pixel locations, normalized so that EMD is within [0-1].

4 Appendix

Proposition 33 is our most technical contribution, its demonstration – published in (Janati, Cuturi, and Gramfort, 2020b) – requires considerable care. Similarly to the bounded growth proposition 32, we would like to bound the off-diagonal Delannoy numbers with their closest diagonal numbers with a bound depending on k. We do so incrementally by comparing the off-diagonal number $D_{m,m+k}$ with $D_{m,m+k-1}$ and $D_{m+1,m+k}$. The proposition states:

Proposition 40 (Proposition 33) *Let $c = 1 + \sqrt{2}$. $\forall m, k \in \mathbb{N}^\star$:*

$$A(m,k): \quad D_{m,m+k} \leq c\Phi_{m,k} D_{m,m+k-1} \tag{4.67}$$

$$B(m,k): \quad c\Psi_{m,k} D_{m,m+k} \leq D_{m+1,m+k} \tag{4.68}$$

Where

$$\begin{cases} \Phi_{m,k} = 1 - \frac{(1-\frac{1}{c})(k-1)+\frac{1}{c}}{m+k-1} \\ \Psi_{m,k} = 1 + \frac{(1-\frac{1}{c})(k-1)}{m} \end{cases}$$

It is noteworthy that since $1 - 1/c = 2 - \sqrt{2} > 0$, we have for all m, k $\Phi_{m,k} \leq 1$ and $\Psi_{m,k} \geq 1$. When both Ψ and Φ are constant and equal to 1, we get two constant bounds equal to c. The role of Φ and Ψ is to have tighter bounds when k increases. The demonstration is based on an induction reasoning on m. That is, we would like to show for all m the statement: $P(m) : (\forall k \geq 1)\ A(m,k)$ and $B(m,k)$. To assist the

reader, we visualize the proof on Figure 4.12 which describes all the steps of the induction. For the sake of clarity, we isolate the following technical Lemma before proving the proposition.

Lemma 8 *Let $c = 1 + \sqrt{2}$ and $m, k \geq 1$. The sequences Φ and Ψ verify the inequalities:*

$$c\Psi_{m,k+1}\Phi_{m,k+1} \leq \left(\frac{1}{c} + \Psi_{m,k} + \Phi_{m,k+1}\right) \leq c\Phi_{m+1,k}\Psi_{m,k} \tag{4.69}$$

PROOF. First, a notation to make calculations easier, let $\alpha = 1 - \frac{1}{c}$. Then we have:

$$\begin{cases} \Phi_{m,k} = 1 - \frac{a(k-1)+\frac{1}{c}}{m+k-1} \\ \Psi_{m,k} = 1 + \frac{a(k-1)}{m} \end{cases}$$

The middle term can be written using $2 + \frac{1}{c} = c$,

$$\begin{aligned} \frac{1}{c} + \Psi_{m,k} + \Phi_{m,k+1} &= 2 + \frac{1}{c} + \frac{a(k-1)}{m} - \frac{ak + \frac{1}{c}}{m+k} \\ &= c + \frac{a(k-1)}{m} - \frac{ak + \frac{1}{c}}{m+k} \quad . \end{aligned}$$

Let's start by proving the right inequality.

1. Right inequality: The right side can be written:

$$c\Phi_{m+1,k}\Psi_{m,k} = c + c\left[\frac{ak}{m} - \frac{a(k-1)+\frac{1}{c}}{m+k} - \frac{a(k-1)}{m}\frac{\left(a(k-1)+\frac{1}{c}\right)}{m+k}\right] \tag{4.70}$$

The inequality we want to prove is equivalent to, dropping the first c: For all $m, k \geq 1$:

$$\begin{aligned} &\frac{a(k-1)}{m} - \frac{ak+\frac{1}{c}}{m+k} \leq c\left[\frac{a(k-1)}{m} - \frac{a(k-1)+\frac{1}{c}}{m+k} - \frac{a(k-1)}{m}\frac{\left(a(k-1)+\frac{1}{c}\right)}{m+k}\right] \\ &\Leftrightarrow a(k-1)(m+k) - m(ak+\frac{1}{c}) \leq c\left[a(k-1)(m+k) - m\left(a(k-1)+\frac{1}{c}\right)\right. \\ &\qquad\qquad \left. -a(k-1)\left(a(k-1)+\frac{1}{c}\right)\right] \\ &\Leftrightarrow akm + ak^2 - am - ak - mak - \frac{m}{c} \leq c\left[akm + ak^2 - ma - ak - akm + ma\right. \\ &\qquad\qquad \left. -\frac{m}{c} - a^2(k-1)^2 - \frac{a}{c}k + \frac{a}{c}\right] \\ &\Leftrightarrow a\left(c - ac - 1\right)k^2 + ac\left(2a-1\right)k + m\left(a + \frac{1}{c} - 1\right) + a - a^2c \geq 0 \end{aligned}$$

However, $c - ac - 1 = 0$ and $a + \frac{1}{c} - 1 = 0$. Thus, the left side above gives rise to an affine function f in k defined as: $f(k) = ac(2a-1)k + a - a^2c$ that verifies $f(1) = ac(2a-1) + a - a^2c = 0$, and since its slope $ac(2a-1) = a(2c-3) = a(\sqrt{2}-1) > 0$, we have $f(k) \geq 0, \quad \forall k \geq 1$. Therefore, since all inductions above are equivalent to each other, the right inequality is proven.

2. Left inequality: Similarly, the left side can be written:

$$c\Phi_{m,k+1}\Psi_{m,k+1} = c + c\left[\frac{ak}{m} - \frac{ak+\frac{1}{c}}{m+k} - \frac{ak}{m}\frac{\left(ak+\frac{1}{c}\right)}{m+k}\right] \tag{4.71}$$

Again c cancels out, and the inequality is equivalent to, for all $m, k \geq 1$:

$$\begin{aligned}
&\frac{a(k-1)}{m} - \frac{ak+\frac{1}{c}}{m+k} \geq c\left[\frac{ak}{m} - \frac{ak+\frac{1}{c}}{m+k} - \frac{ak}{m}\frac{\left(ak+\frac{1}{c}\right)}{m+k}\right] \\
&\Leftrightarrow akm + ak^2 - am - ak - mak - \frac{m}{c} \geq c\left[akm + ak^2 - akm - \frac{m}{c} - a^2k^2 - \frac{ak}{c}\right] \\
&\Leftrightarrow a\left(c - ac - 1\right)k^2 + m\left(a + \frac{1}{c} - 1\right) \geq 0
\end{aligned}$$

However, $c - ac - 1 = 0$ and $a + \frac{1}{c} - 1 = 0$. Thus, we indeed have the last inequality. Therefore, since all inductions above are equivalent to each other, the left inequality is proven. ■

Proof of proposition 33 We can now describe our induction proof. We would like to show for all m the statement: $P(m) : (\forall k \geq 1)\ A(m,k)$ and $B(m,k)$.

0. intialization step For $m = 1$, on one hand we have for all $k \geq 1 : D_{1,k} = 1$ and $c\Phi_{1,k} = 1 + \frac{c-2}{k} = 1 + \frac{\sqrt{2}-1}{k} \geq 1$, thus we have $A(1,k)\ \forall k$. On the other hand, one can easily show that $D_{2,1+k} = 2k+1$ and that $c\Psi_{1,k} = (c-1)k + 1 = \sqrt{2}k + 1 \leq 2k+1$, since $D_{1,k+1} = 1$, we have $B(1,k)\ \forall k$.

1. induction step (on m) . Let $m \geq 2$ and assume $A(m,k)$ and $B(m,k)$ are true for all $k \geq 1$. We first start by proving $A(m+1,k)$ for any $k \geq 1$.

1.1 $A(m,k)$ **and** $B(m,k)\ (\forall k) \Rightarrow A(m+1,k)\ (\forall k)$**:** We show this directly for any $k \geq 1$. Using the recursive definition of Delannoy numbers (4.9) applied to left side of $A(m+1,k)$ we have:

$$D_{m+1,m+k+1} = D_{m+1,m+k} + D_{m,m+k+1} + D_{m,m+k}\ . \tag{4.72}$$

Applying $A(m,k+1)$ to the second term of the right side we get: $D_{m,m+k+1} \leq c\Phi_{m,k+1}D_{m,m+k}$; and applying $B(m,k)$ to the third term, we get: $D_{m,m+k} \leq \frac{D_{m+1,m+k}}{c\Psi_{m,k}}$. Which sums up to:

$D_{m+1,m+k+1} \leq \left(1 + \frac{1}{c\Psi_{m,k}} + \frac{\Phi_{m,k+1}}{\Psi_{m,k}}\right) D_{m+1,m+k}$. To conclude $A(m+1,k)$, all we need is:

$$\left(1 + \frac{1}{c\Psi_{m,k}} + \frac{\Phi_{m,k+1}}{\Psi_{m,k}}\right) \leq c\Phi_{m+1,k} \ , \tag{4.73}$$

which follows directly from the right inequality of Lemma 8. We have thus proven $A(m+1,k)$ for any arbitrary $k \geq 1$.

1.2 $A(m,k)$**,** $A(m+1,k)$**,** $B(m,k)\ (\forall k) \Rightarrow B(m+1,k)\ (\forall k)$**:** We prove the statement $B(m+1,k)\ (\forall k)$ via an induction reasoning on k.

1.2.0 initialization step (k = 1) . For $k = 1$, we have to show that:

$$c\Psi_{m+1,1} D_{m+1,m+2} \leq D_{m+2}.$$

On one hand, we have $\Psi_{m+1,1} = 1$. On the other hand, using the recursion definition (4.6) we get: $D_{m+2} = D_{m+1,m+2} + D_{m+2,m+1} + D_{m+1}$. And by symmetry of Delannoy numbers: $D_{m+2} = 2\, D_{m+1,m+2} + D_{m+1}$. Now using the growth proposition 32 on D_{m+1}, we have: $D_{m+1,m+2} \leq \frac{c^2-1}{2c^2} D_{m+2}$. Since $c = 1+\sqrt{2}$, we have $\frac{c^2-1}{2c^2} = \frac{1}{c}$ which concludes $B(m+1,1)$.

1.2.1 induction step (on k) :. Let $k \geq 1$ and assume $B(m+1,k)$ is true, let's prove that $B(m+1,k+1)$ is true as well. $B(m+1,k+1)$ can be written: $c\Psi_{m+1,k+1} D_{m+1,m+k+2} \leq D_{m+2,m+k+2}$. Again, using the recursion definition, we have:

$$D_{m+2,m+k+2} = D_{m+1,m+k+2} + D_{m,m+k+1} + D_{m+1,m+k+1} + D_{m+2,m+k+1} \tag{4.74}$$

Applying the already proven $A(m+1,\ k')$ (for all k') to the second member of the right side, we have: $D_{m+1,m+k+1} \geq \frac{D_{m+1,m+k+2}}{c\Phi_{m+1,k+1}}$. Similarly, applying the induction (on k) assumption $B(m+1,\ k)$ to the third member, we get: $D_{m+2,m+k+1} \geq c\Psi_{m+1,k} D_{m+1,m+k+1}$. Which sums up to: $D_{m+2,m+k+2} \geq \left(1 + \frac{1}{c\Phi_{m+1,k+1}} + \frac{\Psi_{m+1,k}}{\Phi_{m+1,k+1}}\right) D_{m+1,m+k+2}$. To conclude $B(m+1,k+1)$, all we need is:

$$c\Psi_{m+1,k+1} \leq \left(1 + \frac{1}{c\Phi_{m+1,k+1}} + \frac{\Psi_{m+1,k}}{\Phi_{m+1,k+1}}\right) \tag{4.75}$$

Which follows directly from the left inequality of Lemma 8, where m is substituted with $m+1$. Therefore, $B(m+1,k+1)$ is true, ending the induction proof on k.

Hence, $B(m+1,k)$ holds for any $k \geq 1$, the induction on proof on m is complete. ■

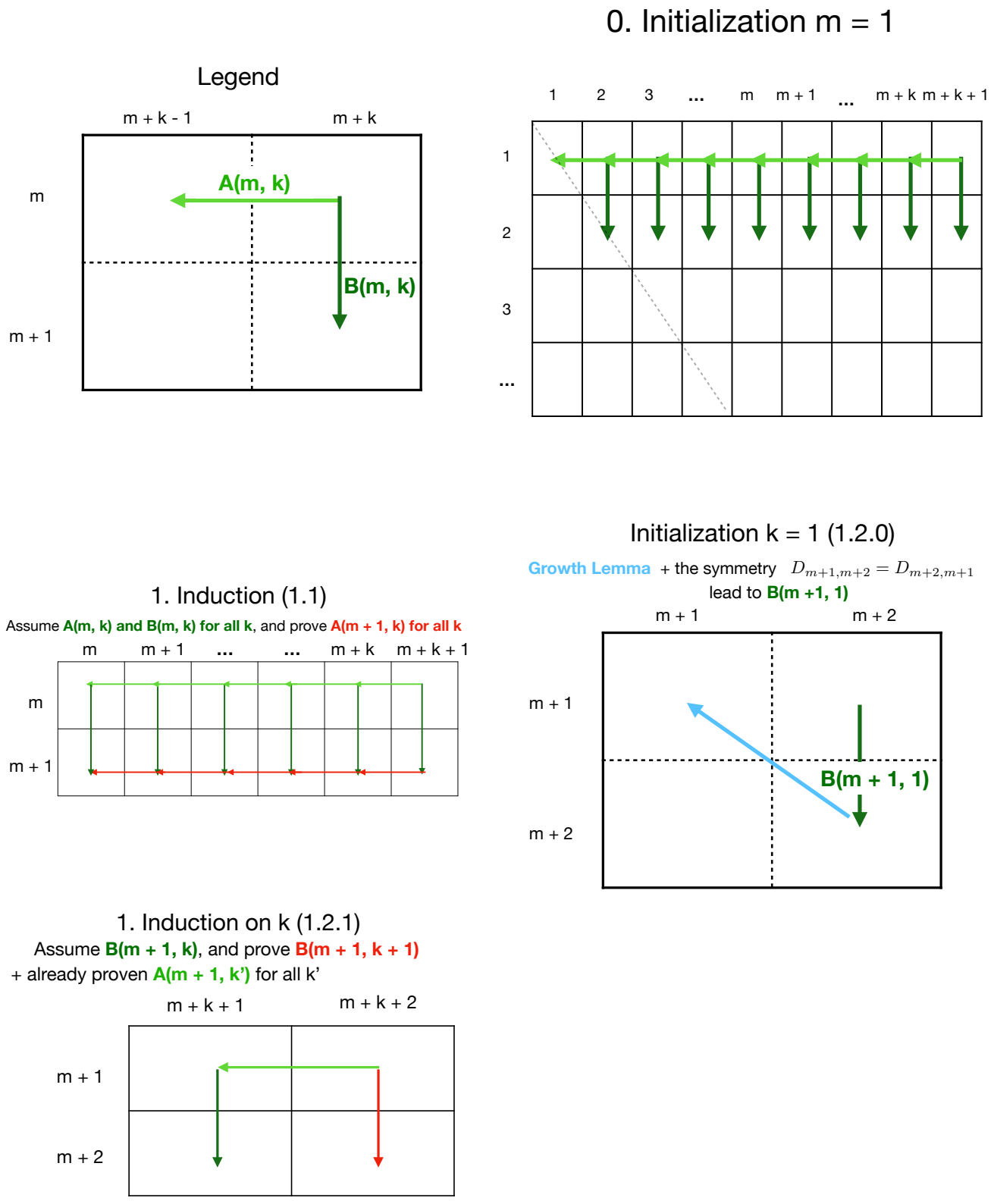


Fig. 4.12. Visualization of the proof of proposition 33. The key steps are 1.1 and 1.2.1, where given the top and left arrows, one must derive the right and bottom arrows.

Chapter 5

Conclusion

As highlighted by the last two chapters, brain imaging was the leading motivation of this PhD. To design the ultimate prior that combines the data of multiple subjects, we must exploit all facets of MEG data and inverse problems. To that end, we involved tools from optimal transport theory, time series and sparse optimization. As any good chef would recommend, inspecting the quality of our ingredients is a must.

Sparsity and MEG source localization Source localization is an inherently ill-posed inverse problem. Using a sparse prior in cognitive experiments where few regions of the brain are expected to be active is not only a mere "natural" idea. The adaptive (a.k.a re-weighted) Lasso can recover sparse sources with no amplitude bias while explaining the variance of the data (Strohmeier, Haueisen, and Gramfort, 2014). Moreover, blind source separation techniques such as ICA can recover components with single dipoles in a complete unsupervised way (i.e without knowledge of the leadfield matrix) (Makeig et al., 1997; Delorme et al., 2012). From a biological perspective, anatomical evidence shows that MEG and EEG are most sensitive to the pyramidal neurons which have a columnar organization within the superficial layers of the cortex (Nunez and Srinivasan, 2006). Following the argument of sufficient reason of Poincaré (1902), science should favor such simple sparse models as long as they fit the data. However, sparse sources are not necessarily accurate solutions. Our contribution was to bind these subject-specific sparse solutions together using optimal transport for a more spatially informed solver. As our experiments suggest, combining the data of different subjects is significantly more informative for various multi-task models and provides solutions that are similar to fMRI activation maps which do not require solving any inverse problem. These encouraging preliminary results for multi-subject inverse problems must be followed by work facilitating hyperparameter tuning for seeing these results being used in a realistic clinical setting.

Time series and spatio-temporal data Exploiting the temporal dimension of MEG data was conducted with the intention of defining a spatio-temporal MEG source localization prior. Dynamic time warping (DTW) has long been applied to clinical data for pattern detection (Karamzadeh et al., 2013) and temporal alignment (Talakoub et al., 2015). Our main contribution was to show that its smooth variant (soft-DTW) is sensitive to temporal shifts and can be virtually defined with any alignment cost function. To some extent, on one hand, we have illustrated the benefits of jointly solving the inverse problem for multiple subjects using OT and sparsity priors. On the other hand, combining OT with soft-DTW captures both time and

space and can be used to average spatio-temporal sources using Sinkhorn's algorithm. Conceptually, replacing OT with STA in the multi-task prior of chapter 3 would allow us to exploit the temporal, spatial and multi-subject axes at once. What could go wrong ? Well, as a house of cards, everything could crumble: the foundations must be reliable. First of all, as our proof demonstrates, the obtained temporal sensitivity is more due to the combinatorial nature of the Delannoy sets than it is an inherent property with which soft-DTW is designed. Thus, how temporal vs spatial differences are aggregated in its final value requires further exploration. Moreover, the order of magnitude of its value – which is not necessarily positive – depends a lot on the size of the time series and the time points where sources are active. Such variability would only make model selection harder. In the context of machine learning and signal processing however, we believe that STA can provide a useful tool to compare and average spatio-temporal data with fast GPU friendly algorithms.

Optimal transport Unlike the other contributions mentioned above, the results presented in chapter 2 are relatively more mature. From theory to practice, we have analyzed entropic OT in its multiple formulations. As a practitioner, working on fixed support brain anatomies guided our focus on fast Sinkhorn iterations for measures defined on grids including debiasing techniques. As a mathematician, our borderline obsession with a bias-free entropic OT and its barycenters led to the study of Gaussians and to the discovery their closed forms. We hope that these contributions will pave the way for a deeper understanding of entropic optimal transport and its algorithms both for neuroimaging and beyond.

Appendix A

Introduction en Français

Idéalement, la poursuite de toute activité scientifique commence par un sentiment d'émerveillement, qui, par le biais d'un raisonnement et d'une recherche plus poussés, se transforme en un graphe de connaissances composé de questions grossières et fines. Il peut sembler évident que la capacité d'une personne à fournir des réponses et à étendre le graphe dépend fortement du degré d'"intérêt" de la question. Mais ne serait-ce pas plutôt l'inverse ? Un sujet ne devient "intéressant" qu'après avoir maîtrisé son contexte, ce qui conduit à un sentiment - peut-être infondé - d'être capable de fournir des réponses à ses questions ouvertes. Absorber la quantité d'informations nécessaire pour atteindre cet état peut prendre des jours, des mois, voire des années. Ainsi, d'un point de vue optimiste, tout peut être intéressant si on l'observe suffisamment longtemps. Pour le sujet qui nous occupe, nous espérons qu'après avoir lu cette introduction, "assez longtemps" ne sera pas trop long.

1 Pourquoi le transport optimal ?

Nous commençons par motiver le transport optimal (OT) sous deux angles différents. Premièrement, en illustrant son utilisation pratique en neuro-imagerie – qui sera le sujet principal du chapitre 3. Ensuite, en montrant comment il s'inscrit dans le paysage statistique.

1.1 Point de vue pragmatique: données d'imagerie cérébrale

L'objectif de l'imagerie cérébrale fonctionnelle est d'étudier l'activité cérébrale. Considérons un modèle de la surface du cerveau donné par un maillage triangulé de p sommets. L'activité cérébrale peut être illustrée en pondérant chaque sommet par un nombre qui peut correspondre ou être proportionnel à l'intensité du courant électrique à l'emplacement de ce sommet.

1.1.1 Comparaison de schémas neuronaux

La comparaison de deux cartes d'activation différentes (ensembles de poids dans $\mathbb{R}_+^p$ [1]) peut être effectuée à l'aide de n'importe quelle fonction de distance dans $\mathbb{R}^p$. Cependant, une telle comparaison ne prendra

[1] Les cartes d'activation peuvent être des vecteurs signés, ce point sera abordé plus en détail au chapitre 3.

pas en compte les disparités spatiales entre les cartes d'activation. En effet, la réduction de ces cartes à des paires de vecteurs de poids ne tient pas compte de toutes les informations contenues dans la structure triangulée de leur maille sous-jacente : l'ordre des sommets importe. La figure A.1 en présente deux exemples. Sachant que l'objectif de l'imagerie cérébrale est de mettre en évidence la fonction des différentes régions du cerveau, la comparaison de la paire **(a)** doit tenir compte de la distance physique entre les régions actives. Sous réserve de telles mesures, une distance entre cette paire de cartes pourrait simplement correspondre à la géodésique entre leurs sommets avec une intensité maximale. Cette idée n'est cependant pas facile à généraliser à des schémas neuronaux complexes (Figure A.1 **(b)**). Lever cette géodésique pour comparer de telles cartes est précisément l'objectif du transport optimal.

Kantorovich OT Cette généralisation nécessite de voir la paire de cartes d'intensité comme des distributions de masse qui doivent être transportées de l'une à l'autre de manière à minimiser une fonction de coût, qui, dans notre cas, est donnée par la géodésique. Cela impose une première restriction importante : la paire de vecteurs de poids doit avoir des entrées non négatives et s'additionner à la même masse totale égale à 1, c'est-à-dire qu'ils appartiennent au simplexe de probabilité Δ_p. Formellement, si nous numérotons les sommets de 1 à p et désignons $\mathbf{x},\ par \in \Delta_p$, alors, la formulation de Kantorovich de l'OT pour la fonction de coût c est donnée par :

$$\mathrm{OT}(\mathbf{x},\mathbf{y}) \stackrel{\text{def}}{=} \min_{\pi \in \mathbb{R}_+^{p\times p} \pi\mathbb{1}=\mathbf{x}, \pi 4op \mathbb{1}=\mathbf{y}} \sum_{i,j}^{p} c(i,j)\pi_{ij} = \langle \mathbf{C}, \pi \rangle \ , \tag{A.1}$$

où $\mathbf{C} \in \mathbb{R}^{p\times p}$ est la matrice avec l'entrée générale $\mathbf{C}_{ij} = c(i,j)$. Le minimiseur π est un tableau conjoint discret avec des marginaux égaux à $\mathbf{x}$ et $\mathbf{y}$ qui minimise le coût de transport $\langle \mathbf{C}, \pi \rangle$. Ce coût a donc la même unité que $\mathbf{C}$ et peut être considéré comme le déplacement moyen optimal entre la paire de cartes d'activation.

Unbalanced OT La formulation (A.1) peut être utile comme métrique de validation dans les simulations où les cartes d'activation sont préalablement projetées sur le simplexe. Cependant, l'OT ne peut pas a priori être utilisée pour comparer les cartes d'activation de deux individus différents ou de deux points temporels différents : la différence dans les amplitudes globales des cartes d'activation importe. La comparaison de vecteurs de poids avec des masses *unbalanced* peut se faire en relaxant les contraintes marginales de (A.1) et en les remplaçant par des divergences lâches qui pénalisent leur violation. L'utilisation du Kullback-Leibler comme divergence conduit à *unbalanced* OT entre $\mathbf{x},\ par \in \mathbb{R}_+^p$ (Liero, Mielke, and Savaré, 2016) :

$$\mathrm{UOT}(\mathbf{x},\mathbf{y}) \stackrel{\text{def}}{=} \min_{\pi \in \mathbb{R}_+{}^{p\times p}} \langle \mathbf{C}, \pi \rangle + \gamma \mathrm{KL}(\pi\mathbb{1}|\mathbf{x}) + \gamma \mathrm{KL}(\pi^\top \mathbb{1}|\mathbf{y}) \ , \tag{A.2}$$

où $\gamma > 0$ est un hyperparamètre qui contrôle le déplacement de masse. Lorsque γ est petit, les marginales de π peuvent être très éloignées de $\mathbf{x}$ et de $\mathbf{y}$, donc très peu de masse est transportée. En pratique, il

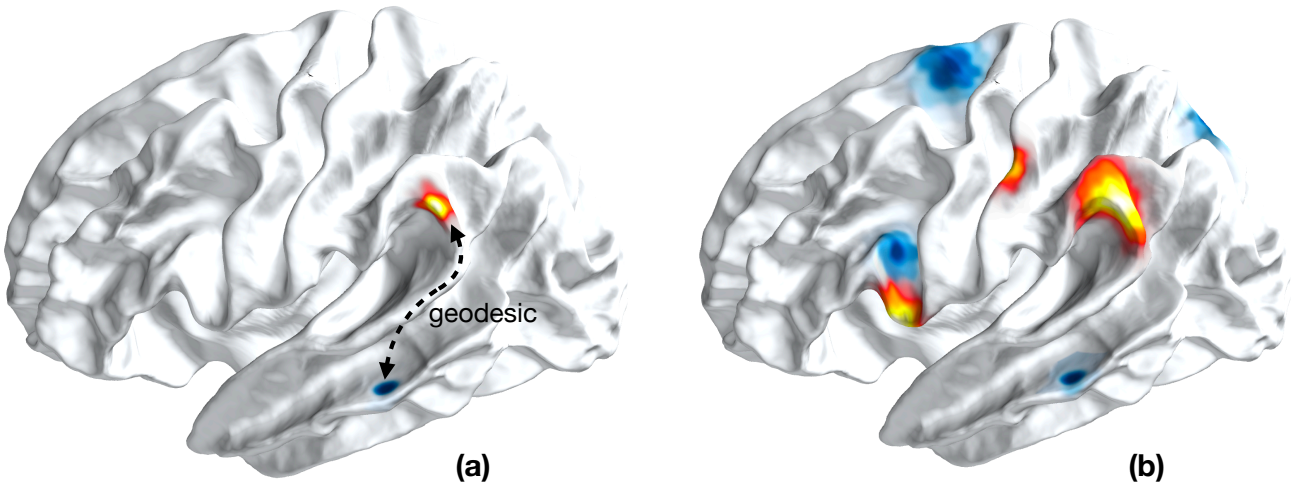


Fig. A.1. Exemples de paires de cartes d'activation cérébrale. S'il est facile et intuitif de comparer la paire de cartes mono-atomiques **(a)** en calculant la géodésique entre leurs emplacements, calculer une telle distance pour la paire **(b)** n'est pas aussi évident.

devrait être fixé relativement aux valeurs de $\mathbf{C}$. Aller au-delà de $\|\mathbf{C}\|_\infty$ conduit en pratique à des plans de transport π presque indiscernables les uns des autres.

1.1.2 Moyennes de mesures neurales

Pour comprendre le fonctionnement du cerveau humain sain, les études de neuro-imagerie sont généralement menées sur un grand groupe de sujets soumis au même protocole expérimental. La synthèse des résultats de ces études nécessite une méthode d'agrégation des multiples cartes cérébrales. Habituellement, les anatomies cérébrales individuelles sont cartographiées sur un "modèle cérébral" commun en faisant correspondre les modèles de convolution cérébrale similaires (note : gyri et sillons) les uns aux autres. Maintenant que les cartes résultantes sont définies sur la même anatomie, toute moyenne de Fréchet peut être utilisée pour définir le cerveau fonctionnel moyen (Gramfort, Peyré, and Cuturi, 2015).

Étant donné K cartes d'activation $\mathbf{x}_1, \ldots, \mathbf{x}_K$ et une fonction de perte F, leur moyenne F-Fréchet est définie par :

$$\arg\min_{\mathbf{x}} \frac{1}{K} \sum_{k=1}^{K} F(\mathbf{x}, \mathbf{x}_k) \tag{A.3}$$

La façon la plus directe de calculer la moyenne des cartes cérébrales est sans aucun doute la moyenne euclidienne, c'est-à-dire en prenant $F(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|^2$. Cependant, même lors de la réalisation d'une même tâche cognitive, la variabilité fonctionnelle entre les individus empêchera les différentes cartes

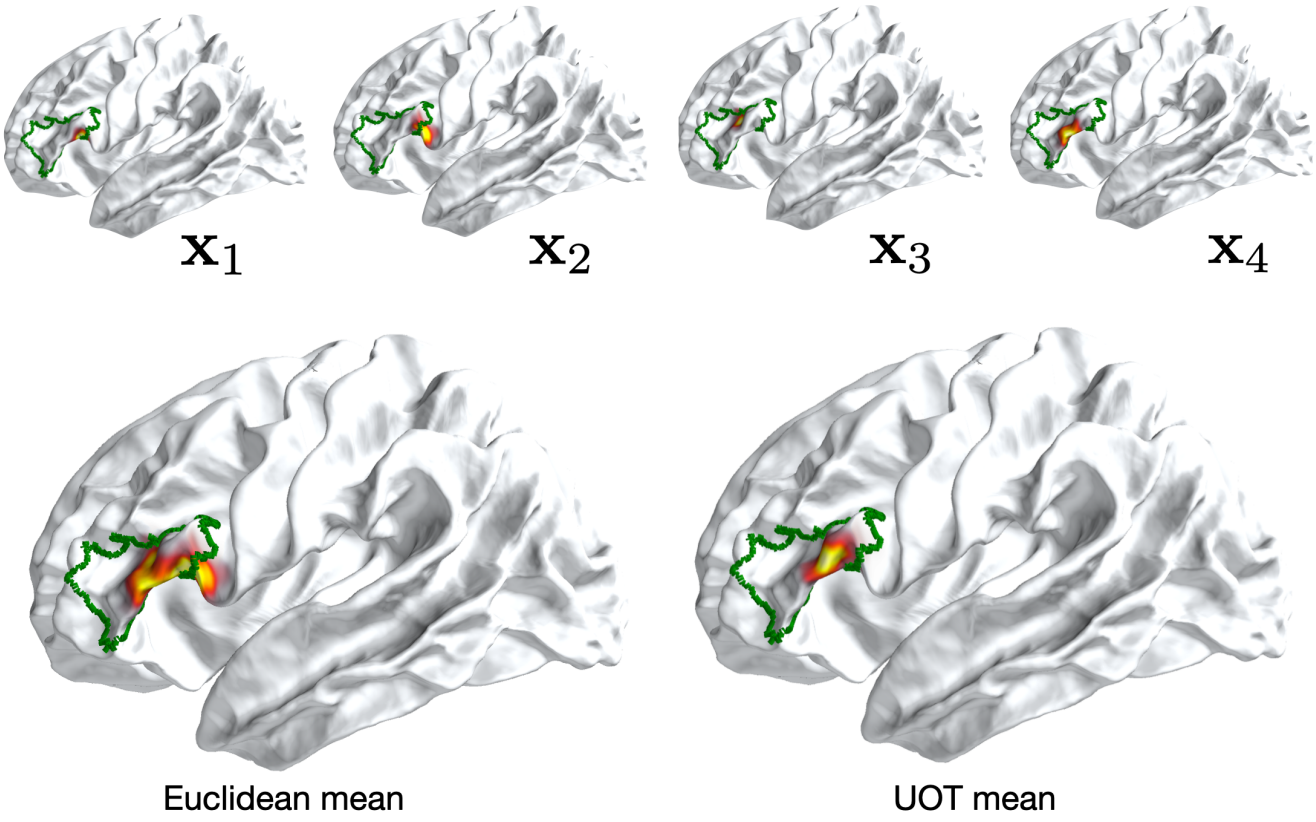


Fig. A.2. Barycentres euclidiens et UOT de 4 cartes d'activation simulées. L'UOT ne souffre pas de l'artefact de flou lié à la moyenne.

d'activation de se superposer parfaitement : Des régions *fonctionnellement* identiques ne sont pas nécessairement *spatialement* identiques (Poline et al., 2010; Allena et al., 2012). L'établissement de la moyenne de ces cartes conduit inévitablement à une moyenne floue. La figure A.2 compare les moyennes de Fréchet (ou barycentres) obtenues avec la perte quadratique et avec UOT : l'exploitation de la *métrie du sol* donnée par la géodésique est cruciale pour obtenir des moyennes significatives.

1.2 Point de vue du statisticien et du géomètre

La "conscience géométrique" des méthodes d'OT discutées ci-dessus est possible parce que nous considérons les cartes d'activation comme des distributions sur le maillage triangulé du cortex. Jusqu'à présent, nous avons supposé que les sommets de ce maillage sont fixes pour toutes les cartes d'activation, ce qui signifie qu'elles sont définies sur le même support fixe. Cette hypothèse permet d'utiliser des algorithmes plus simples et plus rapides qui n'opèrent que sur les poids de ces mesures. Cependant, l'étude théorique

de l'OT nous oblige à abandonner cette hypothèse et à étudier l'OT comme un moyen de comparer des mesures de probabilité avec des supports potentiellement différents.

1.2.1 f-Divergences

La comparaison de mesures de probabilité sur un espace $\mathcal{X}$ est un élément constitutif des statistiques et des modèles d'apprentissage automatique. Ce rôle est joué par plusieurs outils tels que la fonction de Kullback-Leibler ou la variation totale. Ces fonctions appartiennent à la grande famille des divergences de Csiszár introduites pour la première fois par Rényi (1961) et étudiées ensuite par Csiszár (1963). Elles peuvent être définies sur l'ensemble des mesures arbitraires non négatives $\mathcal{M}_+(\mathcal{X})$. Les divergences de Csiszár sont également connues dans la littérature comme des divergences f, car elles sont définies par une fonction *entropie* f.

Definition 7 (f-divergence) *Soit $f : \mathbb{R} \to \bar{\mathbb{R}}_+$ une fonction convexe et semi-continue inférieure, telle que $f(1) = 0$. et $f(\mathbb{R}_-^\star) = +\infty$. Définissons la constante $f_\infty \stackrel{\text{def}}{=} \lim_{p\to+\infty} \frac{f(p)}{p}$. En adoptant la convention $+\infty \times 0 = 0$, la divergence de Csiszár associée à f, communément appelée divergence f, est définie sur l'ensemble des mesures non négatives $\mathcal{M}_+(\mathcal{X})$ comme :*

$$D_f(\alpha, \beta) \stackrel{\text{def}}{=} \int_{\mathcal{X}} f\left(\frac{\mathrm{d}\alpha}{\mathrm{d}\beta}\right) \mathrm{d}\beta + f_\infty \int_{\mathcal{X}} \mathrm{d}\alpha^{perp} \ , \tag{A.4}$$

où $\alpha^\perp$ est la composante singulière de la décomposition de Lebesgue $\alpha = \frac{\mathrm{d}\alpha}{\mathrm{d}\beta}\beta + \alpha^\perp$.

Lorsque α admet une densité de Lebesgue par rapport à β, la composante singulière $\alpha^\perp$ est égale à 0. Ainsi, le deuxième terme de (A.4) disparaît. Le tableau A.1 présente quelques exemples de divergences de Csiszár avec leurs fonctions d'entropie associées f. L'une des caractéristiques les plus attrayantes de cette famille de divergences est leur formulation simple avec un coût de calcul linéaire. Cependant, elles sont limitées à une gamme très restreinte d'applications en raison de deux limitations majeures :

1. La formulation de la décomposition de Lebesgue rompt leur continuité par rapport à un déplacement de position d'un atome dans leur support.

Divergence	$f(p)$
Kullback-Leibler	$p\log(p) - p + 1$
Total variation	$\frac{1}{2}\lvert p-1\rvert$
Reverse Kullback-Leibler	$-\log(p)$
Pearson $\mathcal{X}^2$-divergence	$(p-1)^2$
Hellinger distance	$2p - 4\sqrt{p} + 2$

Table A.1: Exemples de divergences de Csiszár pour différentes fonctions d'entropie.

2. Même dans le cas de mesures absolument continues, les densités de leurs entrées sont comparées ponctuellement, négligeant ainsi toute géométrie sous-jacente de $\mathcal{X}$.

Vous trouverez d'autres exemples et propriétés des divergences de Csiszár dans (Liese and Vajda, 2006).

1.2.2 Normes MMD

Pour aller au-delà de cette comparaison "ponctuelle" des mesures, il faut prendre en compte une certaine interaction croisée entre les mesures. Cette intuition est particulièrement accessible lorsqu'on considère une paire de mesures discrètes $\alpha = \sum_{i=1}^{p} \alpha_i \delta_{x_i}$ et $\beta = \sum_{j=1}^{p} \beta_j \delta_{y_j}$. Si leurs supports se chevauchent – ce qui est nécessaire pour que les divergences f soient bien définies – $KL(\alpha, \beta)$ par exemple comparera les poids sur une base univoque avant d'appliquer une somme. Faire la somme de toutes les paires possibles (α_i, β_j) serait non seulement une comparaison plus complète mais permettrait également d'inclure une certaine notion de distance entre les positions (x_i, y_j). Cette inclusion est communément appelée "élévation de la géométrie" de $\mathcal{X}$. Par exemple, l'inclusion des positions (x_i, y_j) dans ce calcul par le biais d'un ensemble de poids $w_{ij} = K(x_i, y_j)$ pour une certaine fonction k conduit à la formule : $\sum_{i,j} w_{ij}(\alpha_i - \beta_i)(\alpha_j - \beta_j)$. Remarquez que cette formule n'impose aucune restriction sur $(x_i)_i$ et $(y_j)_j$, elle reste donc bien définie même si les supports de α et β sont disjoints. Cela conduit à la définition des normes *maximum mean discrepancy* (MMD) (Gretton et al., 2006) ou des normes de Kernel :

Definition 8 (Normes MMD) *Soit $\mathcal{X}$ un espace compact et K un noyau positif c'est-à-dire une fonction symétrique continue sur $\mathcal{X}^2$ telle que :*

- $K(x, y) = h(x - y)$ *pour une fonction h quelconque*
- $\||\alpha\|_K^2 \stackrel{\text{def}}{=} \int_{\mathcal{X}^2} K \mathrm{d}^2\alpha = \int_{\mathcal{X}^2} K(x, y)\mathrm{d}\alpha(x)\mathrm{d}\alpha(y) \geq 0$ *pour tout α in $\mathcal{M}_+(\mathcal{X})$.*

Pour tout $\alpha, \beta \in \mathcal{M}_+(\mathcal{X})$, la MMD distance *entre α et β peut être définie comme :*

$$\mathrm{MMD}_K(\alpha, \beta) \stackrel{\text{def}}{=} \|\alpha - \beta\|_K^2 \tag{A.5}$$

Contrairement aux divergences f qui nécessitent l'existence de la densité de Lebesgue $\frac{\mathrm{d}\alpha}{\mathrm{d}\beta}$, les normes MMD sont bien définies pour des mesures arbitraires dans $\mathcal{M}_+(\mathcal{X})$. Cependant, même si elles lèvent formellement toute géométrie définie par leur noyau, dans les applications géométriques, elles ne produisent pas de résultats satisfaisants. Par exemple, en reprenant l'exemple précédent de la moyenne de données de neuro-imagerie définies sur un support anatomique fixe $\mathbf{x}_1, \ldots, \mathbf{x}_N$, la perte de Fréchet MMD se lit pour des vecteurs de poids $\mathbf{a}, \mathbf{b}_1, \ldots, \mathbf{b}_N$ *dans* $\mathbb{R}_+^p$ et une matrice Kernel avec les entrées $\mathbf{K}_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$:

$$L(\mathbf{a}) \stackrel{\text{def}}{=} \frac{1}{N} \sum_{n=1}^{N} \|\mathbf{a} - \mathbf{b}_n\|_K^2 = \frac{1}{N} \sum_{n=1}^{N} \sum_{k=1}^{N} \left(\langle \mathbf{a}, \mathbf{K}\mathbf{a} \rangle + \langle \mathbf{b}_n, \mathbf{K}\mathbf{b}_n \rangle - 2\langle \mathbf{a}, \mathbf{K}\mathbf{b}_n \rangle \right) \quad . \tag{A.6}$$

Tant que $\mathbf{K}$ est une matrice définie positive, L est une fonction convexe et coercitive. L'annulation de son gradient conduit au barycentre $\mathbf{a} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{b}_n$, qui correspond à la moyenne euclidienne usuelle indépendamment du choix de K : la géométrie de l'espace sous-jacent est totalement ignorée. Mais avant de porter des jugements hâtifs et de condamner complètement les MMD, peut-être que le fait de prendre des supports libres conduirait à un barycentre plus "géométriquement" conscient ? Lorsqu'elle est restreinte aux mesures de Dirac, la MMD agit comme une perte sur l'espace sous-jacent tant que $h(0) = 0$:

$$\mathrm{MMD}_k(\delta_x, \delta_y) = K(x,x) + K(y,y) - 2K(x,y) = -2K(x,y) \ . \tag{A.7}$$

Cette perte peut même être une distance sur l'espace caractéristique $\mathcal{X}$. Par exemple, lorsque k est le noyau de la distance énergétique : $k(x,y) = -\|x-y\|$, la MMD correspond à la norme ℓ_2 entre les Diracs, pour lesquels le *average* dirac serait situé à leurs emplacements médians. Aussi encourageant que cela puisse être, la prise en compte de nuages de points avec de multiples atomes révèle une limitation majeure des MMD connue sous le nom de *écran de champ électrique*. De la même manière que l'effet sur une charge électrique est dominé par les interactions avec les particules voisines, le gradient MMD d'une seule particule - lors de l'ajustement de la densité - disparaît numériquement en dehors d'un rayon de courte portée. Formellement, étant donné une mesure cible $\beta \overset{\text{def}}{=} \frac{1}{N}\sum_{i=1}^{N} \delta_{y_i}$, ajuster β correspond à minimiser sur les positions d'une mesure $\alpha \overset{\text{def}}{=} \frac{1}{M}\sum_{i=1}^{M} \delta_{x_i}$ la quantité $\mathrm{MMD}_k(\alpha(x_1, \dots, x_M), \beta)$. Avec le noyau $k(x,y) = -2\|x-y\|$ par exemple, en supposant qu'aucune des particules ne se chevauche, la descente directe par rapport à une particule x_l est donnée par :

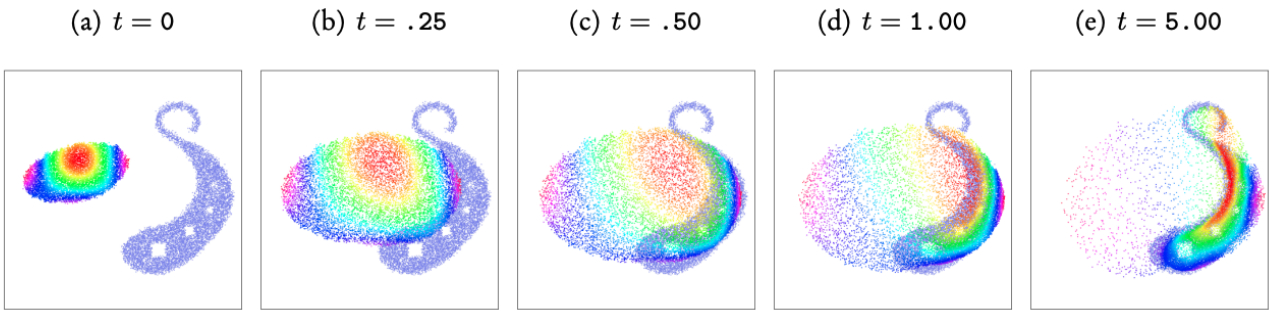
$$-\nabla_{x_l}\mathrm{MMD}_k(\alpha,\beta) = 2\sum_{i\neq l} \frac{x_l - x_i}{\|x_l - x_i\|} - \sum_{j=1}^{N} \frac{x_l - y_j}{|x_l - y_j\|} \tag{A.8}$$

Sous l'influence de la première somme, les particules x_l subissent une force *répulsive* qui s'oppose à la force *attractive* de β. La figure A.3 illustre cet effet d'amortissement : les particules à l'extrême gauche sont dispersées autour de leur emplacement d'origine.

Puisque nous sommes principalement intéressés par la comparaison de mesures basées sur leur forme globale, cette illustration montre que la géométrie de $\mathcal{X}$ intervient "trop tard" dans le calcul des MMD, agissant simplement comme une fonction de pondération. Au lieu de calculer les interactions tous azimuts, cette géométrie sous-jacente pourrait peut-être indiquer quelles particules interagissent avec quelles ?

1.2.3 Transport optimal

Si α et β ont le même nombre de particules de Dirac avec des poids uniformes, une "bonne" fonction de perte d'ajustement de la densité L devrait faire correspondre chaque particule δ_{x_i} à sa destination *final* $\delta_{y_{\sigma(i)}}$, pour une certaine carte d'affectation $\sigma : [\![1,N]\!] \to [\![1,N]\!]$. Idéalement, les étapes de descente de gradient effectuées par chaque particule devraient être proportionnelles à la distance qu'elles doivent parcourir. Par exemple, avec une taille d'étape fixe ω, des gradients de la forme : $x_i \mapsto \frac{1}{\omega}(x_i - y_{\sigma(i)})$ conduiraient à la convergence en une seule itération de descente pour toutes les particules de α. Ces



(graphiques)

Fig. A.3. Tiré de la documentation de KeOps (Charlier et al., 2020). Ajustement de la densité du nuage de points de gauche à la distribution de droite en utilisant un flux de gradient avec la distance énergétique $\text{MMD}_{-\|.\|}$. Les particules à l'extrême gauche sont dispersées loin de la distribution cible en raison de leurs interactions répulsives dominantes avec les particules voisines. Les différentes couleurs servent uniquement au suivi visuel des trajectoires des particules.

gradients "idéaux" peuvent être obtenus avec la fonction de perte :

$$\frac{\omega}{2}\sum_{i=1}^{N}\|x_i - y_{\sigma(i)}\|^2 \ . \tag{A.9}$$

Dans un souci de normalisation, prenons $\omega = \frac{1}{N}$ et définissons l'affectation σ comme la permutation optimale avide dans l'ensemble des permutations de $[\![1, N]\!]$ à $[\![1, N]\!]$ qui minimise (**??**). La fonction de perte obtenue correspond à la première distance de transport optimale proposée par Monge (1781) :

$$\text{OT}(\alpha, \beta) = \min_{\sigma \in G(N)} \frac{1}{2N}\sum_{i=1}^{N}\|x_i - y_{\sigma(i)}\|^2 \ . \tag{A.10}$$

Un exemple simple et intuitif de σ peut être retrouvé en dimension 1 : il correspond à une opération de tri sur la droite réelle du vecteur $y_1, \ldots, y_N$ qui est illustré dans la Figure **??**. La formulation de Monge de l'OT peut donc être vue comme une généralisation du tri aux espaces multidimensionnels.

En pratique toutefois, les mesures peuvent avoir des nombres d'atomes différents (statistiques non paramétriques), avec des poids potentiellement non uniformes (cartes cérébrales fonctionnelles) :

$$\alpha = \sum_{i=1}^{N} a_i \delta_{x_i} \qquad \beta = \sum_{j=1}^{M} b_j \delta_{y_j} \tag{A.11}$$

Dans de tels contextes, une fonction d'affectation peut ne pas exister. Une formulation plus inclusive de l'OT consiste à voir les mesures non pas comme des "particules" à assigner mais comme un "volume de

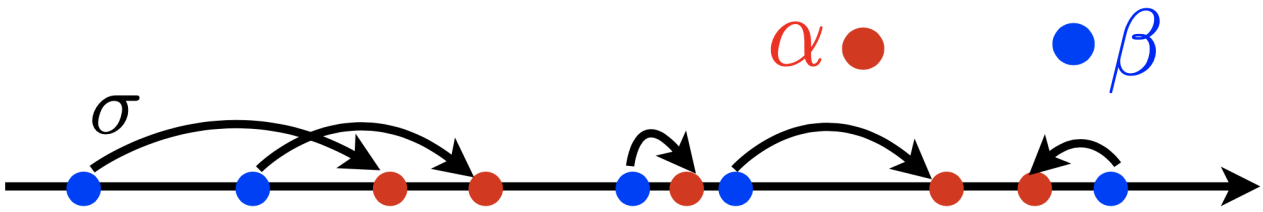


Fig. A.4. OT sur la ligne réelle correspond à une affectation de tri σ. Tiré de (Peyré and Cuturi, 2018).

fluide" à transporter : une masse individuelle α_i n'est pas simplement transférée à un endroit différent mais est *splittée* et déplacée à travers pour *remplir* plusieurs endroits cibles. Ce plan de transport "non déterministe" peut être donné par une matrice $\pi \in \mathbb{R}^{N\times M}$ telle que $\pi_{i,j}$ correspond à la fraction de masse transportée de $a_i\delta_{x_i}$ à $b_j\delta_{y_j}$. Ainsi, pour garantir un transport complet, π doit vérifier : $\pi\mathbb{1} = a$ et $\pi^\top\mathbb{1} = b$. Formellement, cette formulation généralisée d'OT correspond au problème, introduit par Kantorovich (1942) :

$$\mathrm{OT}(\alpha,\beta) = \min_{\substack{\pi \\ in\mathbb{R}_+^{N\times M} \\ pi \\ \mathbb{1}=a,\pi^\top\mathbb{1}=b}} \frac{1}{2}\sum_{i=1}^{N} \|x_i - y_j\|^2\pi_{ij} \ . \tag{A.12}$$

Puisque α et β sont des mesures de probabilité, l'ensemble de contraintes de (A.12) fait de π un tableau conjoint avec les marginaux α et β. Une généralisation directe aux mesures de probabilité génériques avec une fonction de coût symétrique arbitraire $C : \mathcal{X}^2 \to \mathbb{R}_+$ cherche un couplage $\pi \in \mathcal{P}(\mathcal{X}^2)$ avec des marginaux $\pi_1 = \alpha$ et $\pi_2 = \beta$:

$$\mathrm{OT}(\alpha,\beta) = \min_{\substack{\pi\in\mathcal{P}(\mathcal{X}^2) \\ \pi_1=\alpha,\pi_2=\beta}} \int_{\mathcal{X}^2} C\mathrm{d}\pi \ . \tag{A.13}$$

En particulier, la fonction de coût $c(x,y) = \|x-y\|^p$ définit la distance de Wasserstein d'ordre p :

$$\mathcal{W}_p^p(\alpha,\beta) = \min_{\substack{\pi\in\mathcal{P}(\mathcal{X}^2) \\ \pi_1=\alpha,\pi_2=\beta}} \int_{\mathcal{X}^2} \|x-y\|^p\mathrm{d}\pi(x,y) \ . \tag{A.14}$$

Des exemples de plans de transport discrets et continus π sont illustrés dans la figure A.5. Remarquez que les deux formulations des équations (A.10) et (A.12) coïncident avec (A.13) lorsqu'elles sont restreintes à leur domaine de définition.

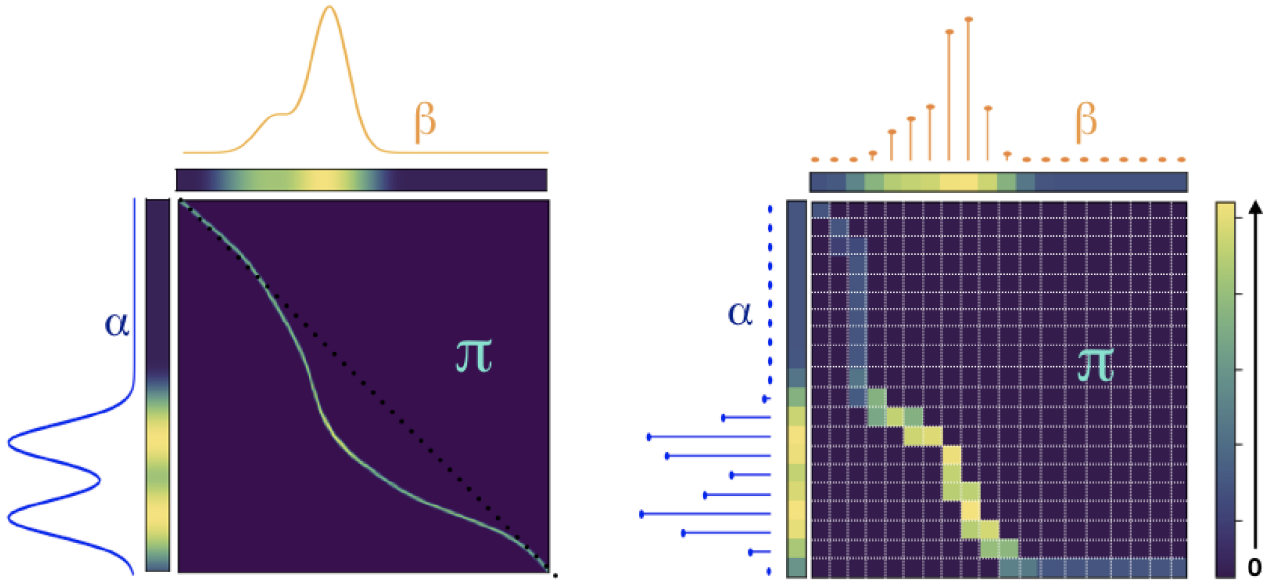


Fig. A.5. Illustration des plans de transport pour le cas continu (à gauche) et discret (à droite). Tiré de (Genevay, 2019).

1.2.4 Complexité statistique et informatique

Contrairement aux MMD, les gradients OT ne s'estompent pas pour les longues distances. De plus, ils "lèvent" la géométrie de $\mathcal{X}$ pour comparer les distributions en optimisant le "transport de masse" qui tient compte de la forme globale des mesures. Cependant, ces propriétés attrayantes ont un prix qui n'est pas abordable pour la plupart des statisticiens et des praticiens de l'apprentissage automatique.

Computational complexity (complexité informatique) Considérons les deux mesures discrètes α, β définies dans (A.11). Par souci de simplicité, supposons que $N = M$. En pratique, le nombre d'atomes N peut correspondre au nombre de bins d'un histogramme, au nombre de sommets d'un maillage ou au nombre de pixels d'une image. En ce qui concerne les applications d'apprentissage automatique, la complexité en N est la plus importante. La distance MMD $\|\alpha - \beta\|_k^2$ peut être donnée par la forme fermée :

$$\|\alpha - \beta\|_k^2 = \langle a, \mathbf{K}a \rangle + \langle b, \mathbf{K}b \rangle - 2\langle a, \mathbf{K}b \rangle \tag{A.15}$$

qui nécessite un nombre exact d'opérations donné par $2N^2 + 3N + 3 = O(N^2)$. Cependant, le calcul de l'OT nécessite la résolution du problème de programmation linéaire (A.12), ce qui peut être réalisé à l'aide de variantes de l'algorithme du réseau simplex et présente donc une complexité inquiétante de $O(N^3 \log(N))$. La réduction de cette complexité par la régularisation est cruciale pour la plupart des utilisations pratiques et fera l'objet de la section 2.

Complexité statistique Considérons maintenant le cas général d'une paire de distributions de probabilité α, β *dans*$\mathcal{P}(\mathcal{X})$ avec $\mathcal{X} \subset \mathbb{R}^d$. La comparaison de α et β peut être effectuée à l'aide d'approximations empiriques $\alpha_n \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ et $\beta_n \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$ où $\mathbf{X}_1, \dots, \mathbf{X}_n$ et $\mathbf{Y}_1, \dots, \mathbf{Y}_n$ sont des échantillons i. i.d. suivant α, β. Une question pratique naturelle est de savoir combien d'échantillons n sont nécessaires pour approximer une fonction de perte $L(\alpha, \beta)$ en utilisant $L(\alpha_n, \beta_n)$?

D'une part, Sriperumbudur et al. (2012) ont montré que pour les MMD, le taux de convergence est indépendant de la dimension sous-jacente d :

$$\mathbb{E}|\operatorname{MMD}_k(\alpha_n, \beta_n) - \operatorname{MMD}_k(\alpha, \beta)| = O\left(n^{-\frac{1}{2}}\right) \quad . \tag{A.16}$$

D'autre part, OT a un taux de catastrophe qui décroît exponentiellement lentement lorsque la dimension augmente. Considérons OT avec la fonction de coût $C(x, y) = \|x - y\|^p$ et $d > 2$. Dudley (1969) ont montré que pour $p = 1$:

$$\mathbb{E}|\operatorname{OT}(\alpha_n, \beta_n) - \operatorname{OT}(\alpha, \beta)| = O\left(n^{-\frac{1}{d}}\right) \quad , \tag{A.17}$$

qui a ensuite été généralisée par Fournier and Guillin (2015) pour $p \geq 1$. L'équation (A.17) semble interdire l'utilisation de l'OT dans des contextes à haute dimension, car toute approximation empirique nécessiterait un nombre exponentiel d'échantillons. Mais peut-être peut-on trouver un meilleur estimateur que le plug-in naïf $\operatorname{OT}(\alpha_n, \beta_n)$? La bonne nouvelle est que nous avons une réponse. La mauvaise nouvelle est la réponse elle-même : Niles-Weed and Rigollet (2019) a montré que pour **tout estimateur** $\widehat{\operatorname{OT}}(\alpha_n, \beta_n)$ de $\operatorname{OT}(\alpha, \beta)$, il existe une paire de mesures $\alpha, \beta \in \mathcal{P}([0,1]^d)$ telle que :

$$\mathbb{E}|\widehat{\operatorname{OT}}(\alpha_n, \beta_n) - \operatorname{OT}(\alpha, \beta)| \geq O\left((n \log(n))^{-\frac{1}{d}}\right) \quad . \tag{A.18}$$

Comme si la complexité numérique cubique ne suffisait pas, l'OT empirique est voué à l'échec en haute dimension.

Mais assez de pessimisme : que pouvons-nous faire ? À certains égards, MMD et OT sont exactement opposés : l'un est bon marché et exploitable en haute dimension mais ne convient pas aux applications géométriques, l'autre est coûteux en termes de calcul et de statistique mais donne de bons résultats pour ce type de tâches. Pourrait-il y avoir un juste milieu ?

2 *Comment* le transport optimal ?

La littérature sur l'OT regorge de tentatives de réduction de la complexité de l'OT. Sans prétendre à l'exhaustivité, ces tentatives peuvent être classées en 3 écoles de pensée différentes :

1. Cherry-picking : restreindre l'analyse à un sous-ensemble de mesures qui sont suffisamment régulières telles que les distributions elliptiques ou les mesures supportées sur des collecteurs de faible dimension.

2. Regularizing the measures : calcul des OT sur les projections des données. L'approche de Wasserstein en tranches (Bonneel et al., 2015) par exemple, consiste à agréger les valeurs d'OT calculées sur des projections 1D des données.

3. Régularisation du plan de transport π en ajoutant une pénalité de Tikhonov qui rend le problème OT (A.13) strictement convexe et donc plus facile à résoudre numériquement.

Toutes nos contributions tournent autour de l'approche 3 : la formulation entropique du transport optimal. Comme nous le verrons dans la section suivante, elle définit le pont tant attendu entre les normes OT et MMD. De plus, elle s'adapte naturellement à la formulation déséquilibrée d'OT (A.2) donnée avec les divergences marginales de KL. Tout d'abord, nous discutons de quelques exemples des approches 1 et 2.

2.1 Choix et régularisation des mesures

Bien que le calcul de l'OT ne soit pas un problème facile en haute dimension, il peut en fait être calculé en forme fermée pour les distributions *elliptiquement contournées* (voir remarque ci-dessous) avec le coût quadratique $c(x, y) = \|x - y\|^2$. Cette forme fermée est donc spécifique à la distance de 2-Wasserstein ($\mathcal{W}_2$) et est connue comme la métrique de Bures-Wasserstein.

2.1.1 La métrique de Bures-Wasserstein

Considérons deux gaussiennes multivariées $\alpha = \mathcal{N}(\mathbf{a}, \mathbf{A})$ et $\beta = \mathcal{N}(\mathbf{b}, \mathbf{B})$ avec $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$ et $\mathbf{A}, \mathbf{B} \in \mathcal{S}_+^d$. Olkin and Pukelsheim (1982) et (Dowson and Landau, 1982) ont montré indépendamment que $\mathcal{W}_2^2$ est donné par :

$$\mathcal{W}_2^2(\alpha, \beta) = \|ba - \mathbf{b}\|^2 + \mathcal{B}^2(\mathbf{A}, \mathbf{B}) \ , \tag{A.19}$$

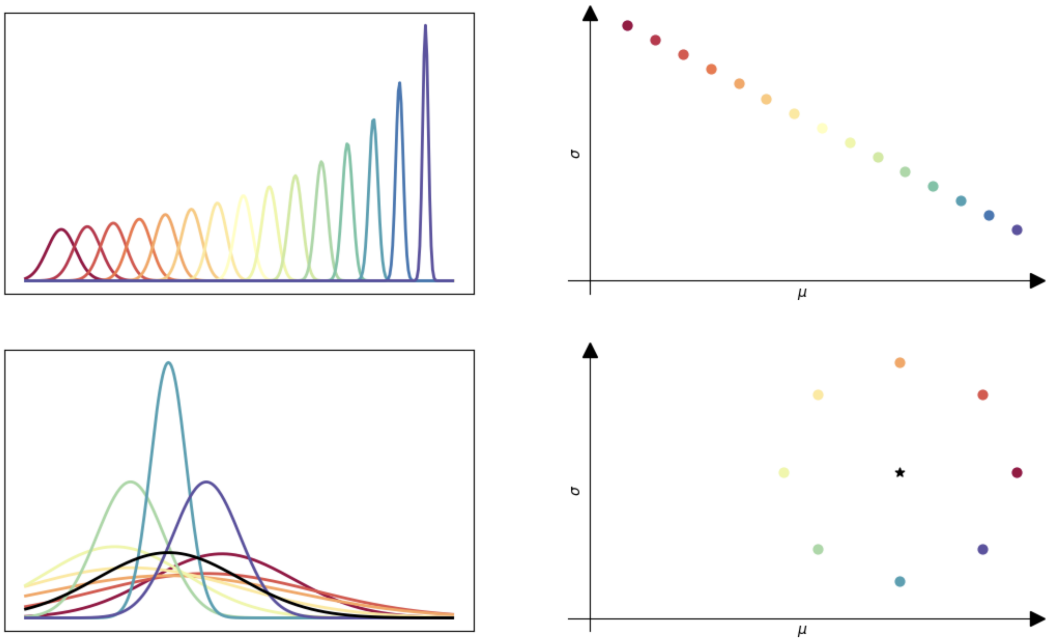
où :

$$\mathcal{B}^2(\mathbf{A}, \mathbf{B}) = \mathrm{Tr}(\mathbf{A}) + \mathrm{Tr}(\mathbf{B}) - 2\mathrm{Tr}((\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}})^{\frac{1}{2}}) \tag{A.20}$$

est la métrique de Bures sur le cône des matrices définies positives (Bures, 1969). Lorsque $\mathbf{A}$ et $\mathbf{B}$ sont diagonales, la métrique de Bures coïncide avec la distance de Hellinger. En effet, si $\mathbf{A} = \mathrm{diag}(\sigma_a)$ et $\mathbf{B} = \mathrm{diag}(\sigma_b)$, alors $\mathcal{B}^2(\mathbf{A}, \mathbf{B}) = \|\sqrt{\sigma}_a - \sqrt{\sigma}_b\|_2^2$ où $\sqrt{.}$ sur les vecteurs est appliqué élément par élément. Ainsi, pour les gaussiennes univariées, le $\mathcal{W}_2$ correspond à la distance euclidienne sur le plan (moyenne, écart-type), illustrée dans la figure A.6.

Remark 8 *La forme fermée* (**??**) *va au-delà des mesures gaussiennes et peut être étendue aux distributions elliptiques (Gelbrich, 1990). Leur nom vient du fait qu'elles incluent des distributions dont la fonction de densité a des ensembles de niveaux elliptiques. Formellement, elles peuvent être caractérisées par un emplacement et des paramètres d'échelle* $\mathbf{m} \in \mathbb{R}^d$ *et* $\mathbf{S} \in \mathcal{S}_+^d$ *et peuvent être transformées de l'une à l'autre par une transformation linéaire* $x \mapsto Ax + b$ *où* A *est défini positif.*

La formule de Bures-Wasserstein fournit non seulement une formule d'OT pour les distributions elliptiques mais elle donne également une borne inférieure pour toutes les mesures de probabilité avec un



(graphique)

Fig. A.6. Calculer l'OT (W_2) entre les gaussiennes univariées (à gauche) est équivalent à calculer la distance euclidienne entre leurs mappings correspondants sur le plan (moyenne, écart-type) (droite). La rangée du bas montre un ensemble de gaussiennes qui sont équidistantes de la gaussienne noire dans le sens W_2.

moment du second ordre. Dowson and Landau (1982) ont montré que pour toute $\alpha, \beta \in \mathcal{P}_2(\mathbb{R}^d)$ avec moyenne et variance respectives $\mathbf{a}, \mathbf{b}$ et $\mathbf{A}, \mathbf{B}$:

$$\mathbf{a} - \mathbf{b}\|^2 + \mathcal{B}^2(\mathbf{A}, \mathbf{B}) \leq \mathcal{W}_2^2(\alpha, \beta) \tag{A.21}$$

Une simple limite supérieure peut être dérivée en remarquant que le couplage indépendant $\pi_0 = \alpha \otimes \beta$ a des marginaux α et β. Par conséquent, par définition de min, le calcul de la perte OT avec π_0 fournit la limite supérieure :

$$\begin{aligned}
\mathcal{W}_2^2(\alpha, \beta) &\leq \int_{\mathbb{R}^p} \|x - y\|^2 \mathrm{d}\alpha(x)\mathrm{d}\beta(y) && \text{(A.22)}\\
&= \int_{\mathbb{R}^p} \|x\|^2 \mathrm{d}\alpha(x) + \int_{\mathbb{R}^p} \|y\|^2 \mathrm{d}\beta(y) - 2\int_{\mathbb{R}^p} xy \mathrm{d}\alpha(x)\mathrm{d}\beta(y) && \text{(A.23)}\\
&= \mathbf{E}_\alpha(X^2) + \mathbf{E}_\beta(X^2) - 2\mathbf{E}_\alpha(X)\mathbf{E}_\beta(X) && \text{(A.24)}\\
&= \mathbf{V}(\alpha) + \mathbf{V}(\beta) + \|\mathbf{E}_\alpha(X) - \mathbf{E}_\beta(X)\|^2 && \text{(A.25)}\\
&= \mathrm{Tr}(\mathbf{A}) + \mathrm{Tr}(\mathbf{B}) + \|\mathbf{a} - \mathbf{b}\|^2 && \text{(A.26)}
\end{aligned}$$

2.1.2 Projections en basse dimension

Une autre approche de la malédiction de la dimensionnalité de l'OT consiste à envisager des projections sur des sous-espaces de faible dimension. Bien que les données de l'apprentissage automatique puissent être de haute dimension, elles présentent le plus souvent une structure de basse dimension, inconnue a priori. Au lieu de calculer $\mathrm{OT}(\alpha, \beta)$ sur l'espace entier $\mathbb{R}^d$, on peut espérer trouver *le meilleur* sous-espace k-dimensionnel sur lequel les projections de α et β sont les plus différentes. Formellement, en désignant la projection orthogonale de α sur le sous-ensemble $E\mathbb{R}^d$ par $P_{E\#}\alpha$, cette quantité s'écrit:

$$\mathrm{OT}_k(\alpha, \beta) = \sup_{\substack{E \subset \mathbb{R}^d \\ \dim(E)=k}} \mathrm{OT}(P_{E\#}\alpha, P_{E\#}\beta) \ , \tag{A.27}$$

qui peut être approximé par l'estimateur empirique plug-in $\mathrm{OT}_k(\alpha_n, \beta_n)$.

Calcul numérique En pratique, un calcul exact de $\widehat{\mathrm{OT}_k}(\alpha_n, \beta_n)$ est potentiellement intraitable. Il peut cependant être approché en utilisant des projections aléatoires ou une relaxation convexe. La première a conduit à la proposition de distances de Wasserstein tranchées (Rabin et al., 2011; Bonneel et al., 2015) qui fixent $k = 1$ et font la moyenne des valeurs OT sur des lignes 1D, ce qui revient à plusieurs opérations de tri. Paty and Cuturi (2019) ont proposé une relaxation convexe de (A.27) en faisant l'observation clé que la quantité minimisée de $\mathcal{W}_2^2$ peut s'écrire :

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 \mathrm{d}\pi(x, y) = \mathrm{Tr}(\mathcal{V}_\pi) = \sum_{l=1}^{d} \lambda_l \ , \tag{A.28}$$

où $\mathcal{V}_\pi = \int_{\mathbb{R}^d \times \mathbb{R}^d} (x - y)(x - y)^\top \mathrm{d}\pi(x, y)$ est une matrice du second ordre avec des valeurs propres triées $\lambda_1 \geq \cdots \geq \lambda_d$. La troncature de (A.28) aux k plus grandes valeurs propres conduit à un problème d'optimisation max-min concave-convexe traçable qui peut être résolu à l'aide d'algorithmes de point de selle.

Complexité de l'échantillon En supposant que α et β sont égaux partout sauf dans un sous-espace à k dimensions $\mathcal{U} \subset \mathbb{R}^d$ avec $k \ll d$, Niles-Weed and Rigollet (2019) a montré que, pour cet estimateur par projection, la limite de complexité d'échantillon (A.17) peut être améliorée. Formellement, pour la distance p-Wasserstein avec $p \in [1, 2]$:

$$\mathbb{E}|\widehat{\mathrm{OT}_k}(\alpha_n, \beta_n) - \mathrm{OT}(\alpha, \beta)| = O\left(n^{-\frac{1}{k}} + \sqrt{\frac{d \log n}{n}}\right) \ , \tag{A.29}$$

où $n^{-\frac{1}{k}}$ est le coût de l'estimation de OT sur $\mathcal{U}$ et $\sqrt{\frac{d \log n}{n}}$ est le prix à payer pour ne pas connaître $\mathcal{U}$ à l'avance.

2.2 Régulariser le plan de transport : OT entropique

Sauf en dimension 1 où l'OT peut être résolu via un tri – tant que la fonction de coût au sol c peut être écrite $c(x, y) = h(x - y)$ avec une fonction convexe h (Santambrogio, 2015) –, troquer un peu d'optimalité pour la vitesse devient une nécessité dans les applications d'apprentissage automatique. La "renaissance" de l'OT dans la recherche sur l'apprentissage automatique est principalement due à l'avantage computationnel qu'offre l'OT entropique. D'autres régularisations basées sur les normes ℓ_p ont également été étudiées dans la littérature (Lorenz, Manns, and Meyer, 2019; Blondel, Seguy, and Rolet, 2018). Même si elles présentent de belles caractéristiques d'amélioration de la spartialité, elles n'annihilent pas la contrainte de non-négativité du plan de transport comme l'entropie, ce qui est crucial pour obtenir un algorithme d'ascension double rapide et convivial pour les GPU.

2.2.1 L'algorithme de Sinkhorn: équilibré, déséquilibré et barycentres

Balanced OT Supposons que $\alpha \overset{\text{def}}{=} \sum_{k=1}^{N} \mathbf{a}_i \delta_{x_i}$ et $\beta \overset{\text{def}}{=} \sum_{i=1}^{M} \mathbf{b}_i \delta_{y_i}$ soient des mesures discrètes dans $\mathbb{R}^d$ avec $\mathbf{a} \in \Delta_N$ et $\mathbf{b} \in \Delta_M$ où $\Delta_p \overset{\text{def}}{=} \{\mathbf{x} \in \mathbb{R}_+^p, \mathbf{x}^{\mathbb{1}} = 1\}$, connu sous le nom de simplex de probabilité. Soit $\mathbf{C} \in \mathbb{R}^{p \times p}$ la matrice de coût du terrain donnée par $\mathbf{C}_{ij} = c(x_i, y_j)$. Sur les matrices, exp et log sont appliqués élément par élément et $\langle . \rangle$ désigne le produit scalaire de Frobenius. Cuturi (2013) ont proposé d'ajouter une pénalité d'entropie fortement convexe :

$$\mathrm{OT}_\varepsilon(\alpha, \beta) \overset{\text{def}}{=} \min_{\substack{\pi \in \mathbb{R}_+^{p \times p} \\ \pi \mathbb{1} = \mathbf{a}, \pi^\top \mathbb{1} = \mathbf{b}}} \langle \mathbf{C}, \pi \rangle + \varepsilon \langle \pi, \log(\pi) - 1 \rangle \ , \tag{A.30}$$

où $\varepsilon > 0$ est un hyperparamètre fixe. Avec la carte linéaire $\mathcal{A} : \pi \in \mathbb{R}_+^{p \times p} \mapsto (\pi \mathbb{1}, \pi^\top \mathbb{1}) \in \mathbb{R}_+^p \times \mathbb{R}_+^p$, le problème primaire (A.30) peut être écrit :

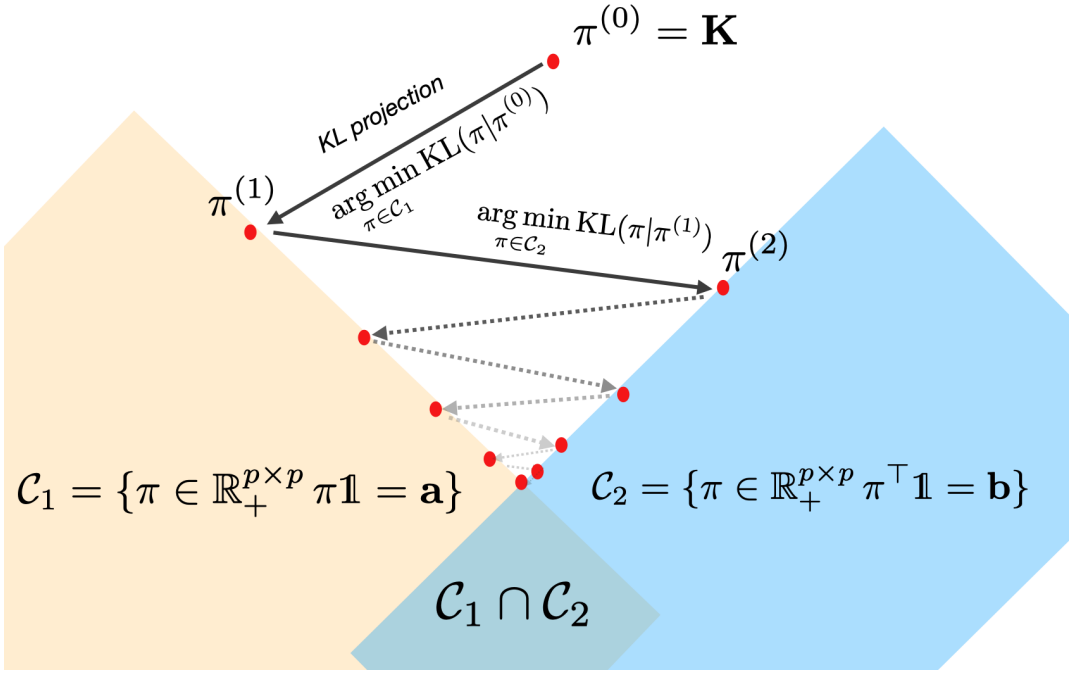
$$\mathrm{OT}_\varepsilon(\alpha, \beta) = \min_{\pi \in \mathbb{R}_+^{p \times p}} R(\pi) + \iota_{(\mathbf{a}, \mathbf{b})}(\mathcal{A}(\pi)) \ , \tag{A.31}$$

où $R(\pi) = \langle \mathbf{C}, \pi \rangle + \varepsilon \langle \pi, \log(\pi) - 1 \rangle$ et $\iota_a(x) = 0$ si $a = x$ et $+\infty$ sinon.

L'opérateur dual de $\mathcal{A}$ pour le produit scalaire de Frobenius est donné par : $\mathcal{A}^\star(\mathbf{f}, \mathbf{g}) \in \mathbb{R}_+^p \times \mathbb{R}_+^p \mapsto \mathbf{f} \oplus \mathbf{g} \in \mathbb{R}_+^{p \times p}$, où $\mathbf{f} \oplus \mathbf{g}$ désigne la matrice $(\mathbf{f}_i + \mathbf{g}_j)_{ij}$. En calculant les conjugués de Fenchel $\mathcal{R}^\star$ et $\iota^\star$, la dualité de Fenchel à (A.30) conduit au problème dual équivalent :

$$\begin{aligned} \mathrm{OT}_\varepsilon(\alpha, \beta) &= \max_{\mathbf{f}, \mathbf{g} \in \mathbb{R}^p} -\iota^\star_{(\mathbf{a}, \mathbf{b})}(-\mathbf{f}, -\mathbf{g}) - \mathcal{R}^\star(\mathcal{A}^\star(\mathbf{f}, \mathbf{g})) \\ &= \max_{\mathbf{f}, \mathbf{g} \in \mathbb{R}^p} \langle \mathbf{f}, \mathbf{a} \rangle + \langle \mathbf{g}, \mathbf{b} \rangle - \varepsilon \langle e^{\frac{\mathbf{f} \oplus \mathbf{g}}{\varepsilon}} - 1, e^{-\frac{\mathbf{C}}{\varepsilon}} \rangle \ , \end{aligned} \tag{A.32}$$

Considérons le changement de variable $\mathbf{u} = e^{\frac{\mathbf{a}}{\varepsilon}}$ et $\mathbf{v} = e^{\frac{\mathbf{b}}{\varepsilon}}$ et $\mathbf{K} = e^{-\frac{\mathbf{C}}{\varepsilon}}$. Le problème dual est une maximisation d'une fonction concave dans $\mathbf{f}$ et $\mathbf{g}$. L'exécution de la montée de gradient alternative par



(graphiques)

Fig. A.7. Illustration de l'algorithme de Sinkhorn comme procédure d'ajustement interactif consistant en une séquence de projections KL qui résolvent la formulation équivalente (A.35).

blocs sur (A.32) avec le changement de variable susmentionné donne les résultats suivants :

$$\mathbf{u} \leftarrow \frac{\mathbf{a}}{\mathbf{K}\mathbf{v}} \qquad \mathbf{v} \leftarrow \frac{\mathbf{b}}{\mathbf{K}^\top \mathbf{u}} \,, \tag{A.33}$$

et à l'optimalité, la relation primal-dual conduit au plan de transport :

$$\pi = \mathrm{diag}(\mathbf{u}) \mathbf{K} \, \mathrm{diag}(\mathbf{v}) \tag{A.34}$$

Ces itérations sont garanties converger à un taux linéaire tant que $\mathbf{K}$ a des entrées positives (Peyré and Cuturi, 2018). Strictement parlant, le taux de convergence dépend du nombre de conditionnement de $\mathbf{K}$: plus il est élevé, plus il est rapide. Ainsi, en pratique, prendre de faibles valeurs de ε ralentit la convergence.

Sinkhorn comme projection KL Bien que l'intérêt de la communauté de l'apprentissage automatique pour les OT entropiques soit assez récent, la formulation (A.30) remonte au problème du pont de Schr'oedinger également connu sous le nom de *modèles de maximisation de l'entropie* (Wilson, 1969). De nos jours, son attrait est principalement dû aux itérations simples, parallélisables et compatibles avec les GPU

(A.33). Plus connues sous le nom d'algorithme de Sinkhorn (Knopp and Sinkhorn, 1967), ces itérations correspondent aux opérations de mise à l'échelle qui doivent être appliquées à une matrice positive (au niveau des entrées) pour la rendre doublement stochastique. De ce point de vue, elle correspond à une séquence de "projections" de la matrice $\mathrm{diag}(\mathbf{u})\mathbf{K}\,\mathrm{diag}(\mathbf{v})$ qui lui font "épouser" les marginales $\mathbf{a}$ et $\mathbf{b}$ et était précédemment connue sous le nom de "procédure itérative d'ajustement par projection" (IPFP). Benamou et al. (2015) ont formalisé cette idée en remarquant que jusqu'à la constante supplémentaire $\langle \varepsilon\mathbf{K}, \mathbb{1}\rangle = \varepsilon \sum_{ij} \mathbf{K}_{ij}$, le problème (A.30) est équivalent à une projection "Bregman" avec la divergence KL :

$$\mathrm{OT}_\varepsilon(\alpha,\beta) = \min_{\substack{\pi\in\mathbb{R}_+^{p\times p}\\ pi\mathbb{1}=\alpha, \pi^\top\mathbb{1}=\beta}} \varepsilon \mathrm{KL}(\pi|\mathbf{K}) \ , \tag{A.35}$$

où $\mathrm{KL}(\mathbf{A},\mathbf{B}) \stackrel{\text{def}}{=} \sum_{i,j} \mathbf{A}_{ij} \log\left(\frac{\mathbf{A}_{ij}}{\mathbf{B}_{ij}}\right) + \mathbf{B}_{ij} - \mathbf{A}_{ij}$ pour $\mathbf{A},\mathbf{B} \in \mathbb{R}_+^{p\times p}$. À première vue, la formulation (A.35) semble fournir une deuxième interprétation géométrique de l'OT entropique avec une compréhension intuitive de l'algorithme de Sinkhorn qui est illustré dans la figure A.7. Un deuxième coup d'œil montre qu'il a en fait une contribution numérique majeure : l'algorithme IPFP peut également être utilisé pour résoudre le *fixed support* OT, c'est-à-dire lorsque le support du barycentre α est connu a priori. Cela se rencontre par exemple en infographie où le support correspond aux emplacements des pixels d'une image. Formellement, on considère une séquence de mesures de probabilité discrètes $\alpha_k \stackrel{\text{def}}{=} \sum_{i=1}^{p_k} \mathbf{a}_i^k \delta_{x_i^k}$ pour $k = 1..K$ et $\alpha \stackrel{\text{def}}{=} \sum_{i=1}^{p} \mathbf{a}_i^k \delta_{x_i}$ avec un support fixe et connu $x_1,\dots,x_p$ mais des poids inconnus $\mathbf{a}_1,\dots,\mathbf{a}_p$. Soit $\mathbf{C}_k$ la matrice avec des entrées $\mathbf{C}_{kij} = c(x_i^k, x_j)$ et $\mathbf{K}_k = e^{-\frac{\mathbf{C}_k}{\varepsilon}}$. L'optimisation est effectuée par rapport aux poids seulement et se lit comme suit :

$$\min_{\mathbf{a}\in\Delta_p} \sum_{k=1}^K w_k \,\mathrm{OT}_\varepsilon(\alpha_k,\alpha) = \min_{\substack{\pi_1,\dots,\pi_K\\ \pi_k\in\mathcal{C}_k\cap\mathcal{C}'}} \sum_{k=1}^K w_k \mathrm{KL}(\pi_k|\mathbf{K}_k) \ , \tag{A.36}$$

où $(w_k)_k \in \Delta_K$ est un vecteur de poids fixe, $\mathcal{C}_k = \{\pi \in \mathbb{R}_+^{p_k\times p} | \pi\mathbb{1} = \mathbf{a}_k\}$ et $\mathcal{C}' = \{\pi \in \mathbb{R}_+^{p_k\times p} | \exists \mathbf{a} \in \Delta_p,\ \pi_k^\top\mathbb{1} = \mathbf{a},\ \forall k = 1\dots K\}$. La résolution de (A.36) peut être effectuée via *Projections itératives de Bregman* (IBP), ce qui revient à effectuer une minimisation alternative sur un ensemble de contraintes $\mathcal{C}$ à la fois. Chaque étape peut être résolue en forme fermée, ce qui conduit à des itérations de type Sinkhorn :

$$\mathbf{u}_k \leftarrow \frac{\mathbf{a}_k}{\mathbf{K}_k\mathbf{v}_k}, \qquad \mathbf{a} = \prod_{k=1}^K (\mathbf{K}_k^\top \mathbf{u}_k)^{w_k}, \qquad \mathbf{v}_k \leftarrow \frac{\mathbf{a}}{\mathbf{K}_k^\top \mathbf{u}_k} \ . \tag{A.37}$$

Cadre unifié d'OT entropique Toute l'élégance numérique de l'OT entropique réside peut-être dans l'unification suivante proposée par Chizat et al. (2018b). Étant donné un ensemble de poids non négatifs $(w_k)_k \in \Delta_K$ et une paire de fonctions scalaires convexes séparables F_1 et F_2 opérant sur $\prod_{k=1}^K \mathbb{R}_+^{p_k}$ et $\mathbb{R}^{K\times p}$

respectivement, on obtient :

$$\min_{\pi \in \mathbb{R}_+^{p \times p^K}} \varepsilon \widehat{\mathrm{KL}}(\pi_1, \dots, \pi_K | \mathbf{K}_1, \dots, \mathbf{K}_K) + F_1(\pi_1 \mathbb{1}, \dots, \pi_K^\top \mathbb{1}) + F_2(\pi_1^\top \mathbb{1}, \dots, \pi_K^\top \mathbb{1}). \tag{A.38}$$

avec $\widehat{\mathrm{KL}}(\pi_1, \dots, \pi_K | \mathbf{K}^1, \dots, \mathbf{K}^K) \stackrel{\text{def}}{=} \sum_{k=1}^K w_k \mathrm{KL}(\pi_k | \mathbf{K}_k)$.

Cela étend la théorie de l'OT entropique au cadre "déséquilibré" où le plan de transport π ne doit pas s'adapter à certaines mesures d'entrée α, β *exact*. Ainsi, α et β peuvent être des mesures non négatives avec des masses différentes. Alors que l'OT entropique peut être retrouvé avec $K = 1$ et $F_1(x) = \iota_{x=\alpha}$ et $F_2(x) = \iota_{x=\beta}$, le problème du barycentre équilibré (A.36) correspond au choix :

$$F_1(\pi_1 \mathbb{1}, \dots, \pi_K^\top \mathbb{1}) = \sum_{k=1} \iota_{\pi_k \mathbb{1} = \alpha_k} \tag{A.39}$$

$$F_2(\pi_1^\top \mathbb{1}, \dots, \pi_K^\top \mathbb{1}) = \min_{\alpha \in \Delta_p} \sum_{k=1}^K \iota_{\pi_k^\top \mathbb{1} = \alpha} \tag{A.40}$$

En utilisant les développements de la dualité de Fenchel-Rockafellars similaires à ceux de (A.32), Chizat et al. (2018b) a montré qu'effectuer une ascension duale sur le problème dual correspond aux itérations alternées génériques :

$$\begin{aligned} \mathbf{u}_1, \dots, \mathbf{u}_K &\leftarrow \mathrm{proxdiv}_{F_1}(\mathcal{K}(\mathbf{v}_1, \dots, \mathbf{v}_K)) \\ \mathbf{v}_1, \dots, \mathbf{v}_K &\leftarrow \mathrm{proxdiv}_{F_2}(\mathcal{K}^\top(\mathbf{u}_1, \dots, \mathbf{u}_K)) \end{aligned} \tag{A.41}$$

où l'opérateur linéaire $\mathcal{K}$ et proxdiv sont définis par :

$$\mathcal{K} : \mathbb{R}^{p^K} \to \prod_{k=1}^K \mathbb{R}_+^{p_k} \tag{A.42}$$

$$(\mathbf{x}_1, \dots, \mathbf{x}_K) \mapsto (\mathbf{K}_1 \mathbf{x}_1, \dots, \mathbf{K}_K \mathbf{x}_K), \tag{A.43}$$

$$\mathrm{proxdiv}_F(\mathbf{z}) = \frac{1}{\mathbf{z}} \arg\min_{\mathbf{s}} F(s) + \varepsilon \widehat{\mathrm{KL}}(\mathbf{s} | \mathbf{z}) \tag{A.44}$$

De même, à l'optimalité, chaque plan de transport π_k est donné par $\mathrm{diag}(u_k) \mathbf{K}_k \mathrm{diag}(v_k)$.

Tant que l'opérateur proxdiv peut être calculé sous une forme fermée, la résolution des problèmes entropiques OT couvrant les centres équilibrés, déséquilibrés et barycentriques peut se faire via des opérations proxdiv très simples (A.41). Le tableau A.2 fournit l'expression de l'opérateur proxdiv de certaines divergences F_1 et F_2. Pour des raisons de simplicité, nous ne couvrons que les OT non équilibrés avec la divergence de KL. Des exemples de barycentres utilisant $F = \gamma \mathrm{KL}$ sont présentés dans la figure A.8. Pour les faibles valeurs de γ, les contraintes marginales ne sont pas forcées, et donc très peu de transport se produit. Nous nous référons à (Chizat et al., 2018b) pour d'autres exemples tels que les OT déséquilibrés avec un écart de variation totale ou une contrainte de plage.

OT	Divergence $F_2(\mathbf{x})$	$\text{proxdiv}_{F_2}(\mathbf{x})$
Balancé	$\iota_{x=\mathbf{a}}$	$\frac{\mathbf{a}}{\mathbf{x}}$
Non balancé	$\gamma \text{KL}(\mathbf{x}\|\mathbf{a})$	$\left(\frac{\mathbf{a}}{\mathbf{x}}\right)^{\frac{\gamma}{\gamma+\varepsilon}}$
Barycentre	$\min_{\mathbf{a}\in\Delta_p} \sum_{k=1}^K \iota_{\mathbf{x}_k=\mathbf{a}}$	$\frac{\mathbf{a}^\star}{\mathbf{x}}$ with $\mathbf{a}^\star = \prod_{k=1}^K (\mathbf{x}_k)^{w_k}$
Barycentre non balancé	$\min_{\mathbf{a}\in\Delta_p} \sum_{k=1}^K \gamma \text{KL}(\mathbf{x}_k\|\mathbf{a})$	$\left(\frac{\mathbf{a}^\star}{\mathbf{x}}\right)^{\frac{\gamma}{\gamma+\varepsilon}}$ with $\mathbf{a}^\star = \left(\frac{\sum_{k=1}^K w_k \mathbf{x}_k^{\frac{\varepsilon}{\gamma+\varepsilon}}}{\sum_{k=1}^K w_k}\right)^{\frac{\gamma+\varepsilon}{\varepsilon}}$

Table A.2: Exemples de proxdiv (Chizat et al., 2018b)

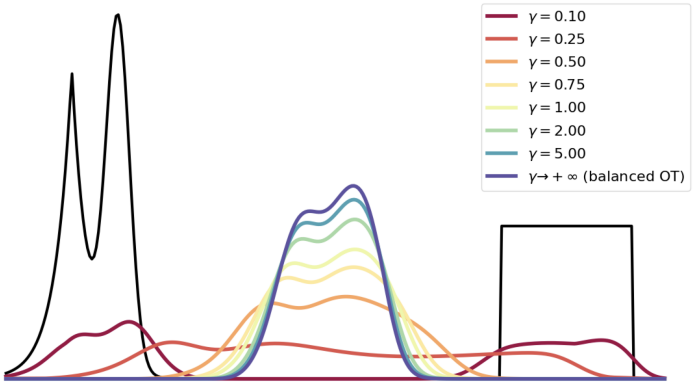
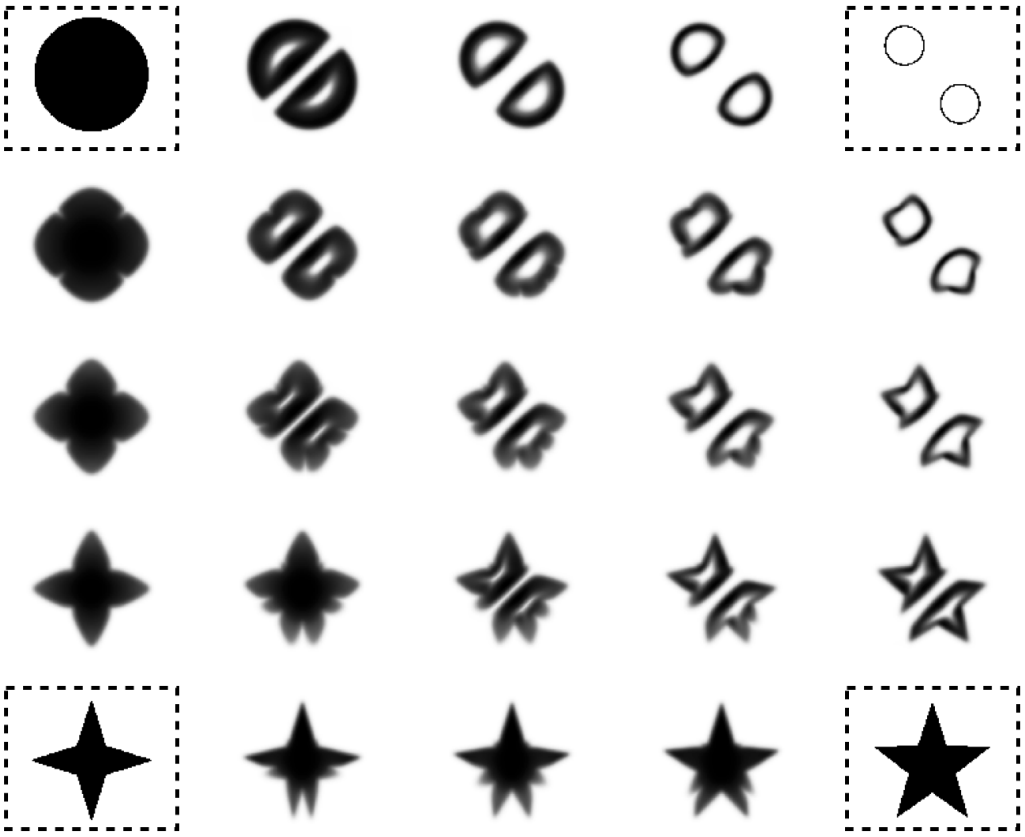


Fig. A.8. Barycentres non équilibrés des deux mesures indiquées en noir pour différentes valeurs de γ où F_1 et F_2 sont définis comme les deux divergences de KL non équilibrées du tableau A.2 respectivement.

L'algorithme de Sinkhorn est significativement plus rapide sur les grilles régulières En général, tant que l'opérateur de proxdivision peut être calculé sous une forme fermée, chaque itération de Sinkhorn a une complexité de $O(Kp^2)$ où K est le nombre fixe de mesures impliquées dans le problème. Cette complexité peut toutefois être réduite à $O(Kp^{1+\frac{1}{d}})$ lorsqu'on travaille sur des grilles régulières de dimension d (Solomon et al., 2015) avec la perte quadratique. Considérons l'exemple simple des images, c'est-à-dire $d = 2$. Supposons pour simplifier que les images sont carrées avec le même nombre de pixels égal à $p_1 = \cdots = p_K = p = m^2$. Soit $\mathbf{z} \in \mathbb{R}_+^{m\times m}$ une image avec son format vectoriel $\mathbf{z}' \in \mathbb{R}_+^{m^2}$. Soit $1 \leq l \leq m^2$ un pixel de coordonnées 2D $l = (l_x, l_y)$, $x, y \in [\![1, m]\!]$. Ainsi, la distance quadratique entre deux pixels l, k



(graphiques)

Fig. A.9. Interpolations OT entropiques (barycentres équilibrés pondérés) des quatre images encadrées pour différents ensembles de poids (w_k). Chaque image appartient à $\mathbb{R}^{p\times p}$ avec $p = 400$. Sur un GPU, les 21 barycentres ont été calculés en quelques secondes.

correspond à : $\|l-k\|^2 = (l_x-k_x)^2 + (l_y-k_y)^2$ et :

$$\begin{aligned}\mathcal{K}(\mathbf{z}')_k = \sum_{l=1}^{m^2} e^{-\frac{\|k-l\|^2}{\varepsilon}}\mathbf{z}'_l = \sum_{l=1}^{m^2} e^{-\frac{(l_x-k_x)^2+(l_y-k_y)^2}{\varepsilon}}\mathbf{z}'_l &= \sum_{l_x=1}^{m}\sum_{l_y=1}^{m} e^{-\frac{(l_x-k_x)^2+(l_y-k_y)^2}{\varepsilon}}\mathbf{z}_{l_x,l_y} \\ &= \sum_{l_y=1}^{m} e^{-\frac{(l_y-k_y)^2}{\varepsilon}} \sum_{l_x=1}^{m} e^{-\frac{(l_x-k_x)^2}{\varepsilon}}\mathbf{z}_{l_x,l_y} \\ &= \sum_{l_y=1}^{m} e^{-\frac{(l_y-k_y)^2}{\varepsilon}}[\mathbf{K}'\mathbf{z}]_{k_x,l_y} \\ &= [\mathbf{K}'\mathbf{z}\mathbf{K}']_{k_x,k_y}\ ,\end{aligned} \tag{A.45}$$

où $\mathbf{K}' \in \mathbb{R}_+^{m \times m}$ est la matrice du noyau *smaller* avec les entrées $e^{-\frac{(i-j)^2}{\varepsilon}}$. Appliquer l'opérateur $\mathcal{K}$ revient à effectuer des convolutions gaussiennes le long des lignes et des colonnes de $\mathbf{z}$, ce qui a une complexité de $2m^3 = 2p^{\frac{3}{2}}$ au lieu des p^2 opérations du produit matrice-vecteur habituel $\mathcal{K}(\mathbf{z}')$. La même astuce de séparabilité du noyau s'applique aux données multidimensionnelles pour autant que les mesures soient définies sur des grilles régulières et que le coût soit quadratique. La figure A.9 illustre les barycentres de quatre images (aux coins) de taille $p = (400 \times 400)$ pour différents poids d'interpolation. Sur un GPU, tous les barycentres ont été calculés en quelques secondes.

2.2.2 Biaissement entropique et compromis MMD-OT

Au-delà des mesures discrètes La définition de l'OT entropique donnée dans (A.30) est spécifique aux mesures discrètes puisqu'elle définit la fonction entropique par rapport à une mesure discrète uniforme sur un ensemble fini. Sa généralisation la plus simple serait peut-être celle du cas continu de Lebesgue. Soit $\mathcal{X} \subset \mathbb{R}^d$ un espace compact et $\alpha, \beta \in \mathcal{P}(\mathcal{X}, \mathcal{L})$ où $\mathcal{L}$ désigne la mesure de Lebesgue. Soit c une fonction de coût symétrique Lipschitz sur $\mathcal{X} \times \mathcal{X}$. L'OT entropique continu peut être défini comme :

$$\mathrm{OT}_\varepsilon^{\mathcal{L}}(\alpha, \beta) \stackrel{\text{def}}{=} \min_{\substack{\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{X}) \\ pi_1 = \alpha, \pi_2 = \beta}} \int c \mathrm{d}\pi + \varepsilon \int \log\left(\frac{\mathrm{d}\pi}{\mathrm{d}\mathcal{L}}\right) \mathrm{d}\pi \ . \tag{A.46}$$

Identifier α, β et π avec leurs densités de Lebesgue conduit à un problème qui peut être approximé via des OT discrets calculés sur des histogrammes *convergents* vers ces densités. L'étude de $\mathrm{OT}^{\mathcal{L}}$ peut donc éclairer le comportement de l'OT entropique discret, comme nous le verrons au chapitre **??**.

Ces deux formulations ne couvrent cependant pas les cas où les mesures ne sont ni *both* discrètes, ni *both* absolument continues. Ces limitations peuvent être contournées en remarquant que tant que π possède des marginales α et β, son support sera inclus dans le support de la mesure produit $\alpha \otimes \beta$. Formellement, si $A \times B \subset \mathcal{X} \times \mathcal{X}$ est un ensemble de Borel tel que $\alpha \otimes \beta(A \times B) = 0$ alors $\alpha(A)\beta(B) = 0$ et donc soit $\alpha(A) = 0$ soit $\beta(B) = 0$. Puisque $A \times B \subset A\mathcal{X}$ et $A \times B \subset \mathcal{X} \times B$, il est vrai que $\pi(A \times B) \leq \min(\pi(A\mathcal{X}), \pi(\mathcal{X} \times B)) = \min(\pi_1(A), \pi_2(B)) = \min(\alpha(A), \beta(B)) = 0$. Par conséquent, π est absolument continue par rapport à $\alpha \otimes \beta$. En utilisant la mesure du produit comme référence, on peut donner une définition générique de l'OT entropique :

$$\mathrm{OT}_\varepsilon^{\otimes}(\alpha, \beta) \stackrel{\text{def}}{=} \min_{\substack{\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{X}) \\ \pi_1 = \alpha, \pi_2 = \beta}} \int c \mathrm{d}\pi + \varepsilon \int \log\left(\frac{\mathrm{d}\pi}{\mathrm{d}\alpha \otimes \beta}\right) \mathrm{d}\pi \ . \tag{A.47}$$

MMD et interpolation OT Les avantages de cette formulation sont nombreux. Pour commencer, quelle que soit la mesure de référence, lorsque $\varepsilon \to +\infty$, OT_ε revient à une maximisation de l'entropie conduisant à un $lim_{\varepsilon \to +\infty} \pi_\varepsilon = \alpha \otimes \beta$. Mais lorsqu'il s'agit de calculer la limite de la valeur OT, $\lim_{\varepsilon \to +\infty} \mathrm{OT}_\varepsilon^{\otimes}(\alpha, \beta)$ est bien défini et est donné par $\int c \mathrm{d}\alpha \otimes \mathrm{d}\beta$, alors que $\lim_{\varepsilon \to +\infty} \mathrm{OT}_\varepsilon^{\mathcal{L}}(\alpha, \beta) = -\infty$. La première limite a conduit plusieurs auteurs (Ramdas, Trillos, and Cuturi, 2017; Genevay, Peyre, and Cuturi, 2018; Feydy

et al., 2019) à proposer la divergence de Sinkhorn :

$$\mathrm{S}_\varepsilon(\alpha,\beta) = \mathrm{OT}_\varepsilon^{\otimes}(\alpha,\beta) - \frac{1}{2}(\mathrm{OT}_\varepsilon^{\otimes}(\alpha,\alpha) + \mathrm{OT}_\varepsilon^{\otimes}(\beta,\beta)) \ , \tag{A.48}$$

pour laquelle cette limite devient $\lim_{\varepsilon\to+\infty} \mathrm{S}_\varepsilon(\alpha,\beta) = \frac{1}{2}\int -c\mathrm{d}^2(\alpha-\beta)$. Ainsi, l'OT entropique interpole entre l'OT et une distance MMD si $-C$ est défini positif :

$$\mathrm{OT}(\alpha,\beta) \underset{\varepsilon\to 0}{Leftarrow} \mathrm{S}_\varepsilon(\alpha,\beta) \underset{\varepsilon\to+\infty}{\to} \frac{1}{2}\mathrm{MMD}_{-C}(\alpha,\beta) \tag{A.49}$$

À la lumière de ce résultat, S_ε pourrait-il fournir un terrain d'entente en matière de complexité d'échantillonnage ? Genevay et al. (2019) fournit une réponse positive avec la limite de complexité :

$$\mathbb{E}|\,\mathrm{S}_\varepsilon(\alpha_n,\beta_n) - \mathrm{S}_\varepsilon(\alpha,\beta)| = O\left(n^{-\frac{1}{2}}(\varepsilon^{-\frac{d}{2}}+1)e^{\frac{\kappa}{\varepsilon}}\right) \ , \tag{A.50}$$

où κ dépend du diamètre de l'ensemble compact $\mathcal{X}$ et de c. Bien que la complexité en n soit la même que celle des MMD, toute utilisation pratique de (A.50) en haute dimension interdit les faibles valeurs de ε. Ainsi, S_ε ne doit pas être considéré ou utilisé comme une approximation de l'OT, mais comme un moyen terme bien établi entre les métriques OT et MMD. Mais quelles sont les propriétés qui rendent S_ε approprié pour les applications d'apprentissage automatique ou d'analyse de forme ?

Propriétés de S_ε Un résultat bien établi est la différentiabilité de l'OT entropique avec des gradients donnés par les variables duales optimales généralement appelées dans la théorie de l'OT *potentiels duaux*. De plus, tant que $\mathcal{X}$ est un ensemble compact et que c induit un noyau universel positif $k(x,y) \overset{\text{def}}{=} e^{-\frac{c(x,y)}{\varepsilon}}$:

1. S_ε is non-negative : $\mathrm{S}_\varepsilon(\alpha,\beta) \geq 0, \qquad \mathrm{S}_\varepsilon(\alpha,\beta) = 0 \Leftrightarrow \alpha = \beta$.
   L'élément S_ε est convexe par rapport à l'un de ses arguments.
   L'énoncé (1) conduit à $\arg\min_\beta \mathrm{S}_\varepsilon(\alpha,\beta) = \alpha$. On dit que S_ε est *debiased*.
   L'élément S_ε mesure la convergence faible en loi:$\mathrm{S}_\varepsilon(\alpha_n,\alpha) \to 0 \Leftrightarrow \alpha_n \rightharpoonup \alpha$,

où la convergence faible est définie comme :

$$\alpha_n \rightharpoonup \alpha \Leftarrow \int f\mathrm{d}\alpha_n \to \int f\mathrm{d}\alpha \quad \forall f \in \mathcal{C}(\mathcal{X}) \tag{A.51}$$

Qu'en est-il des OT déséquilibrés? De la même manière, l'OT entropique non équilibré peut être défini pour des mesures arbitraires non négatives $\mathcal{M}_+(\mathcal{X})$ comme :

$$\underset{\varepsilon,\gamma}{\overset{\otimes}{\mathrm{UOT}}}(\alpha,\beta) \overset{\text{def}}{=} \inf_{\pi\in\mathcal{M}(\mathcal{X}\times\mathcal{X})} \int c\mathrm{d}\pi + \varepsilon\,\mathrm{KL}(\pi\|\alpha\otimes\beta) + \gamma\,\mathrm{KL}(\pi_1\|\alpha) + \gamma\,\mathrm{KL}(\pi_2\|\beta), \tag{A.52}$$

où $\gamma > 0$ et $\mathrm{KL}(\pi\|\alpha\otimes\beta) \overset{\text{def}}{=} \int_{\mathcal{X}} \log\left(\frac{\mathrm{d}\pi}{\mathrm{d}\alpha\mathrm{d}\beta}\right)\mathrm{d}\pi$.

Pour obtenir des propriétés similaires pour les OT entropiques déséquilibrés, une première tentative serait de considérer une divergence similaire :

$$(\alpha, \beta) \mapsto \underset{\varepsilon,\gamma}{\overset{\otimes}{\mathrm{UOT}}}(\alpha, \beta) - \frac{1}{2}(\underset{\varepsilon,\gamma}{\overset{\otimes}{\mathrm{UOT}}}(\alpha, \alpha) + \underset{\varepsilon,\gamma}{\overset{\otimes}{\mathrm{UOT}}}(\beta, \beta)) \ . \tag{A.53}$$

Cependant, même avec l'hypothèse de positivité du noyau $k = e^{-c/\varepsilon}$, la divergence (A.53) ne vérifie pas la non-négativité ni la convexité qui sont violées lorsqu'on prend de grands écarts de masse entre les mesures. Pour les compenser, on peut ajouter une pénalité quadratique sur cette différence de masse. La divergence de Sinkhorn déséquilibrée proposée par Séjourné et al. (2019) se lit comme suit :

$$\mathrm{S}^{\otimes}_{\varepsilon,\gamma}(\alpha, \beta) = \underset{\varepsilon,\gamma}{\overset{\otimes}{\mathrm{UOT}}}(\alpha, \beta) - \frac{1}{2}(\underset{\varepsilon,\gamma}{\overset{\otimes}{\mathrm{UOT}}}(\alpha, \alpha) + \underset{\varepsilon,\gamma}{\overset{\otimes}{\mathrm{UOT}}}(\beta, \beta)) + \frac{\varepsilon}{2}(\alpha(\mathcal{X}) - \beta(\mathcal{X}))^2 \ . \tag{A.54}$$

Comme dans le cas équilibré, $\mathrm{S}^{\otimes}_{\varepsilon,\gamma}$ est définie positivement et convexe par rapport à un si son argument. De plus, elle métrique la convergence en loi et a une complexité d'échantillon qui s'échelonne avec une dépendance similaire sur n et ε à celle de la borne (A.50).

2.2.3 Les dilemmes du praticien

La formulation générique $\mathrm{OT}^{\otimes}_{\varepsilon}$ est sans doute plus conforme aux principes d'un point de vue théorique : elle compare la pénalité entropique de π par rapport à son maximum atteint lorsque $\pi = \alpha \otimes \beta$ et conduit à la divergence débiaisée S_{ε} avec toutes ses propriétés vertueuses. Mais en pratique, lorsque les mesures sont discrètes, $\mathrm{OT}_{\varepsilon}$ et $\mathrm{OT}^{\otimes}_{\varepsilon}$ sont-ils équivalents ? Est-ce que $\mathrm{OT}^{\otimes}_{\varepsilon}$ s'inscrit dans le cadre unifié (A.38) d'Chizat et al. (2018b) ? ?

La mesure uniforme et la mesure du produit et leurs variations de Sinkhorn. Par souci de clarté, reprenons les deux formulations dans le cas discret en utilisant une pénalité KL. Soit $\mathcal{X} = x_1, \ldots, x_p$ un ensemble fini, $\alpha, \beta \in \mathcal{P}(\mathcal{X})$. D'une part, jusqu'à la constante supplémentaire $\varepsilon(\log(p) - 1)$, l'OT discret discuté dans (A.30) est équivalent à :

$$\mathrm{OT}^{\mathcal{U}}_{\varepsilon}(\alpha, \beta) \overset{\text{def}}{=} \min_{\pi \in \mathbb{R}^{p\times p}_{+} \ \pi\mathbb{1}=\mathbf{a}, \pi^{\top}\mathbb{1}=\mathbf{b}} \langle \mathbf{C}, \pi \rangle + \varepsilon \mathrm{KL}(\pi \| \mathcal{U}) \ , \tag{A.55}$$

où $\mathcal{U}_{\mathcal{X}^2}$ est la mesure uniforme sur $\mathcal{X}^2$, en pondérant chaque x_i avec $\frac{1}{p^2}$. Remarquez que dans le cas discret, on peut toujours écrire pour un π faisable :

$$\begin{aligned}\mathrm{KL}(\pi\|\alpha\otimes\beta) &= \sum_{i,j}^{p}\pi_{ij}\log\left(\frac{\pi_{ij}}{\mathbf{a}_i\mathbf{b}_j}\right)\\ &= \sum_{i,j}^{p}\pi_{ij}\log\left(\frac{\pi_{ij}}{1/p^2}\right) - \sum_{i,j}^{p}\pi_{ij}(2\log(p)+\log(\mathbf{a}_i)+\log(\mathbf{b}_j))\\ &= \mathrm{KL}(\pi\|\mathcal{U}) - 2\log(p) - \langle\log(\mathbf{a}),\mathbf{a}\rangle + \langle\log(\mathbf{b}),\mathbf{b}\rangle\\ &= \mathrm{KL}(\pi\|\mathcal{U}) - \mathrm{KL}(\mathbf{a}\|\mathcal{U}_{\mathcal{X}}) - \mathrm{KL}(\mathbf{b}\|\mathcal{U}_{\mathcal{X}})\end{aligned} \tag{A.56}$$

où nous avons utilisé le fait que la somme de $\mathbf{a}$ et $\mathbf{b}$ est égale à 1. Ainsi, $\mathrm{OT}_\varepsilon^{\otimes}$ et $\mathrm{OT}_\varepsilon^{\mathcal{U}}$ sont équivalents jusqu'aux entropies additives de α et β :

$$\mathrm{OT}_\varepsilon^{\otimes}(\alpha,\beta) = \mathrm{OT}_\varepsilon^{\mathcal{U}}(\alpha,\beta) - \varepsilon\mathrm{KL}(\alpha\|\mathcal{U}_{\mathcal{X}}) - \varepsilon\mathrm{KL}(\beta\|\mathcal{U}_{\mathcal{X}}) \tag{A.57}$$

La dépendance de cette constante vis-à-vis de α et β induit cependant quelques modifications mineures à leur problème dual et aux itérations de Sinkhorn. Le problème dual équivalent de $\mathrm{OT}_\varepsilon^{\mathcal{U}}$ se lit comme suit :

$$\max_{\mathbf{f},\mathbf{g}\in\mathbb{R}^p} \langle\mathbf{f},\mathbf{a}\rangle + \langle\mathbf{g},\mathbf{b}\rangle - \varepsilon\langle e^{frac\mathbf{f}\oplus\mathbf{g}-C\varepsilon},1\rangle \ , \tag{A.58}$$

avec des conditions d'optimalité données par :

$$e^{\frac{\mathbf{f}}{\varepsilon}} = \frac{\mathbf{a}}{\mathbf{K}e^{\frac{\mathbf{g}}{\varepsilon}}}, \quad e^{\frac{\mathbf{g}}{\varepsilon}} = \frac{\mathbf{b}}{\mathbf{K}^\top e^{\frac{\mathbf{f}}{\varepsilon}}}, \quad \pi = \mathrm{diag}(e^{\frac{\mathbf{f}}{\varepsilon}})\mathbf{K}\,\mathrm{diag}(e^{\frac{\mathbf{g}}{\varepsilon}}) \tag{A.59}$$

alors que la formulation $\mathrm{OT}_\varepsilon^{\otimes}$ a un problème dual légèrement différent :

$$\begin{aligned}\mathrm{OT}_\varepsilon^{\otimes}(\alpha,\beta) &\overset{\text{def}}{=} \min_{\substack{\pi\in\mathbb{R}_+^{p\times p}\pi\\ mathds1=\mathbf{a},\pi^\top\mathbb{1}=\mathbf{b}}} \langle\mathbf{C},\pi\rangle + \varepsilon\mathrm{KL}(\pi\|\mathbf{a}\otimes\mathbf{b})\\ &= \max_{\mathbf{f},\mathbf{g}\in\mathbb{R}^p} \langle\mathbf{f},\mathbf{a}\rangle + \langle\mathbf{g},\mathbf{b}\rangle - \varepsilon\langle e^{frac\mathbf{f}\oplus\mathbf{g}-C\varepsilon},\mathbf{a}\otimes\mathbf{b}\rangle \ ,\end{aligned} \tag{A.60}$$

avec des conditions d'optimalité données par :

$$e^{\frac{\mathbf{f}}{\varepsilon}} = \frac{\mathbf{a}}{\mathbf{K}e^{\frac{\mathbf{g}}{\varepsilon}}}, \quad e^{\frac{\mathbf{g}}{\varepsilon}} = \frac{\mathbf{b}}{\mathbf{K}^\top e^{\frac{\mathbf{f}}{\varepsilon}}}, \quad \pi = \mathrm{diag}(e^{\frac{\mathbf{f}}{\varepsilon}})\mathbf{K}\,\mathrm{diag}(e^{\frac{\mathbf{g}}{\varepsilon}}) \tag{A.61}$$

Alors que l'algorithme de Sinkhorn reste presque inchangé, l'apparition de $\alpha\otimes\beta$ dans le problème dual (A.60) révèle une différence clé entre $\mathrm{OT}_\varepsilon^{\mathcal{U}}$ et $\mathrm{OT}_\varepsilon^{\otimes}$. En tant que supremum de fonctions linéaires dans $\mathbf{a}$ et $\mathbf{b}$, $\mathrm{OT}_\varepsilon^{\mathcal{U}}$ est **conjointement** convexe dans $(\mathbf{a},\mathbf{b})$ alors que le produit $\mathbf{a}\otimes\mathbf{b}$ dans le dual $\mathrm{OT}_\varepsilon^{\otimes}$ interdit la

OT	Non-negative	Convex	Jointly convex	$\arg\min_\alpha \mathrm{OT}(\alpha,\beta)=\beta$	Sinkhorn for barycenters
$\mathrm{OT}_\varepsilon^{\mathcal{U}}$	X	✓	✓	X	✓
$\mathrm{OT}_\varepsilon^{\otimes}$	X	✓	X	X	X
S_ε	✓	✓	X	✓	✓ (chapter 2).
$\mathrm{UOT}_{\varepsilon,\gamma}^{\mathcal{U}}$	X	✓	✓	X	✓
$\mathrm{UOT}_{\varepsilon,\gamma}^{\otimes}$	X	✓	X	X	X
$\mathrm{S}_{\varepsilon,\gamma}^{\otimes}$	✓	✓	X	✓	X
$\mathrm{S}_{\varepsilon,\gamma}^{\mathcal{U}}$	✓	X	X	✓	✓ (chapter 2).

Table A.3: Propriétés de différentes divergences OT **restrictives sur des mesures discrètes avec une matrice noyau symétrique semi-définie positive** $\mathbf{K} \stackrel{\text{def}}{=} e^{-\frac{\mathbf{C}}{\varepsilon}}$.

convexité conjointe de $\mathrm{OT}_\varepsilon^{\otimes}$. En effet, Feydy et al., 2019 a montré que $\mathrm{OT}_\varepsilon^{\otimes}$ est *concave* sur la diagonale c'est-à-dire que $\alpha \to \mathrm{OT}^{\otimes}(\alpha,\alpha)$ est concave, ce qui est pourtant utile pour prouver la convexité de S_ε. De plus, les problèmes de barycentre avec $\mathrm{OT}_\varepsilon^{\otimes}$ et S_ε ne peuvent pas être écrits comme une projection de KL, ainsi le cadre unifié de (Chizat et al., 2018b) est perdu.

Débarrassage des OT non équilibrés Des comparaisons similaires peuvent être faites pour les OT non équilibrés. Le débiaisage des OT non équilibrés à l'aide de la mesure du produit ($\mathrm{S}_{\varepsilon,\gamma}^{\otimes}$) conduit à des fonctions de perte - bien que présentant des propriétés intéressantes - pour lesquelles les barycentres ne peuvent pas tirer parti des algorithmes rapides compatibles avec les GPU offerts par la régularisation entropique. Pour les mesures discrètes sur supports fixes, nous pouvons conserver les propriétés attrayantes de Sinkhorn en définissant UOT par rapport à la mesure uniforme $\mathcal{U} \in \mathcal{P}(\mathcal{X}^2)$:

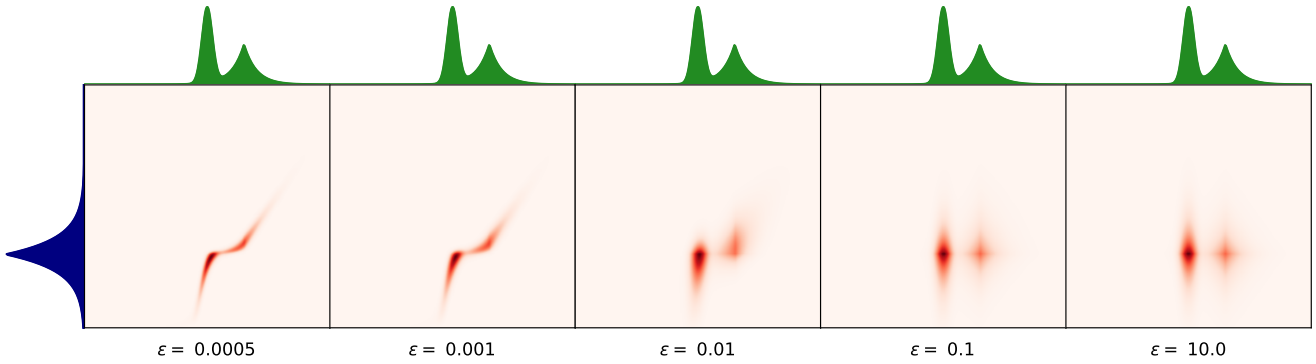
$$\underset{\varepsilon,\gamma}{\overset{\mathcal{U}}{\mathrm{UOT}}}(\alpha,\beta) \stackrel{\text{def}}{=} \inf_{\pi\in\mathcal{M}(\mathcal{X}\times\mathcal{X})} \int c\,\mathrm{d}\pi + \varepsilon\,\mathrm{KL}(\pi\|\mathcal{U}) + \gamma\,\mathrm{KL}(\pi_1\|\alpha) + \gamma\,\mathrm{KL}(\pi_2\|\beta), \tag{A.62}$$

et sa divergence abaissée :

$$\mathrm{S}_{\varepsilon,\gamma}^{\mathcal{U}}(\alpha,\beta) = \underset{\varepsilon,\gamma}{\overset{\mathcal{U}}{\mathrm{UOT}}}(\alpha,\beta) - \frac{1}{2}\left(\underset{\varepsilon,\gamma}{\overset{\mathcal{U}}{\mathrm{UOT}}}(\alpha,\alpha) + \underset{\varepsilon,\gamma}{\overset{\mathcal{U}}{\mathrm{UOT}}}(\beta,\beta)\right) \ . \tag{A.63}$$

Les propriétés de ces divergences pour les mesures discrètes sont résumées dans le tableau A.3 et seront examinées plus en détail dans le chapitre **??**.

Instabilité numérique, scalabilité et implémentations Sinkhorn L'un des effets secondaires les plus notoires de la régularisation entropique est peut-être le flou induit du plan de transport optimal. Au fur et à mesure que ε augmente, π_ε se rapproche du couplage indépendant $\alpha\otimes\beta$ qui a une entropie maximale et est illustré dans la Figure A.10. Pour apprivoiser ce comportement et conserver les propriétés attrayantes



(graphiques)

Fig. A.10. Flou entropique du plan de transport à mesure que ε augmente.

de l'OT, certaines applications peuvent nécessiter de petites valeurs de ε. Cependant, lorsque $\varepsilon \to 0$, la plupart des entrées du noyau $\mathbf{K} \overset{\text{def}}{=} e^{-\frac{\mathbf{c}}{\varepsilon}}$ disparaissent, ce qui entraîne des erreurs numériques lors de la division par $\mathbf{Ku}$ et $\mathbf{Kv}$. Au prix d'une perte de parallélisation, diverses implémentations "stabilisées" de Sinkhorn qui "absorbent" de grandes valeurs de $\mathbf{u}$ et $\mathbf{v}$ dans le domaine logarithmique ou qui sont calculées entièrement dans le domaine logarithmique à l'aide de routines logsumexp sont discutées dans (Schmitzer, 2016) avec d'autres procédures multi-échelles. Les praticiens intéressés peuvent trouver ces variantes de Sinkhorn dans la bibliothèque Python POT (Flamary and Courty, 2017).

Les GPU ont été l'ingrédient magique qui a ramené l'OT computationnelle sous le radar des mathématiciens appliqués. Si les itérations de Sinkhorn peuvent être simples et rapides sur les GPU, elles nécessitent de stocker en mémoire la matrice des coûts de base $\mathbf{C} \in \mathbb{R}_+^{p \times p}$, ce qui peut être problématique dès que p atteint quelques milliers. Cette limite d'évolutivité peut être surmontée en calculant $c(x, y)$ *à la volée* lors de l'application des routines logsumexp sur des données non-tensorisées. Cela nécessite d'importantes et non triviales modifications CUDA de bas niveau, qui, heureusement pour tout le monde, sont offertes sur un plateau d'argent dans la `KeOps` Python (Charlier et al., 2020) avec un package ultérieur spécifique aux fonctions de perte géométrique nommé `GeomLoss`[2](Feydy et al., 2019). Avec GeomLoss, le calcul de l'OT entropique entre des millions d'échantillons n'est pas un fardeau. Pour un aperçu complet des outils d'analyse de forme en géométrie et des diverses implémentations de Sinkhorn, nous ne saurions trop recommander le manuscrit de thèse de Jean Feydy (Feydy, 2020).

3 Plan et contributions

Après avoir établi toutes les connaissances de base nécessaires, nous pouvons maintenant énoncer nos contributions qui se situent à l'intersection du transport optimal, de l'imagerie cérébrale et des problèmes inverses. Notre objectif principal est d'utiliser l'OT pour construire un antécédent spatial P dans un cadre

[2] http://www.kernel-operations.io/geomloss/

régularisé de la forme :

$$\min_{\mathbf{x}} L(\mathbf{x}) + \mu P(\mathbf{x}) \ , \tag{A.64}$$

où L est un terme de fidélité des données et $\mu > 0$ un hyperparamètre fixe.

La minimisation des pertes entropiques d'OT induit cependant un biais dans le minimiseur appelé dans la littérature sur l'OT *biais entropique*.

Il peut être défini comme le cas simple du barycentre à 1 mesure : $\arg\min_\alpha \mathrm{OT}_\varepsilon(\alpha, \beta) \neq \beta$. L'une des vertus de S_ε est l'absence d'un tel biais, au prix de la perte de l'algorithme de Sinkhorn pour les barycentres et la convexité conjointe. Du point de vue pratique du problème (A.64), devons-nous tenter de débiaiser OT en premier ou utiliser le cadre unifié standard de Chizat et contrer le flou de l'entropie avec des pénalités supplémentaires ?

Chapitre 2 : Transport optimal entropique Ce chapitre comporte deux contributions majeures :

1. *Transport entropique optimal pour les gaussiens.* Avant de fournir une réponse pratique à la question susmentionnée, il est crucial de comprendre ce qu'est *exactement* le biais entropique. Le faire pour des mesures arbitraires n'est pas une tâche facile, nous nous concentrons donc sur les gaussiennes multivariées. Pour ce faire, il faut généraliser les résultats de convexité et de différentiabilité de l'OT entropique aux mesures à supports non compacts. Nous découvrons une forme fermée de l'OT entropique similaire à la métrique de Wasserstein-Bures. Cette forme fermée peut être généralisée aux *Unbalanced* gaussiennes, c'est-à-dire des gaussiennes non normalisées avec une masse arbitraire. Ces formes fermées fournissent le premier test-case pour les conjectures théoriques de l'OT entropique et peuvent servir de repères algorithmiques pour les algorithmes stochastiques de Sinkhorn. Afin de quantifier le biais entropique pour $\mathrm{OT}_\varepsilon^{\mathcal{L}}$, $\mathrm{OT}_\varepsilon^{\otimes}$ et S_ε, nous caractérisons les barycentres OT des gaussiennes multivariées. Nous montrons que (1) $\mathrm{OT}_\varepsilon^{\mathcal{U}}$ / $\mathrm{OT}_\varepsilon^{\mathcal{L}}$ induit un biais de flou (variance accrue), (2) $\mathrm{OT}_\varepsilon^{\otimes}$ produit un barycentre rétréci (variance diminuée) et (3) S_ε n'a (presque) aucun biais.

2. *Algorithmes pour les barycentres équilibrés et déséquilibrés débités.* Bien qu'il soit simple d'utiliser l'algorithme IBP pour calculer les barycentres avec $\mathrm{OT}_\varepsilon^{\mathcal{U}}$, faire de même pour les autres divergences n'est pas trivial. Nous proposons un schéma repondéré pour calculer le barycentre de $\mathrm{OT}_\varepsilon^{\otimes}$ et un algorithme rapide de type Sinkhorn pour calculer le barycentre débiaisé avec S_ε. Enfin, nous discutons des alternatives à la divergence déséquilibrée débitée $\mathrm{S}_{\varepsilon,\gamma}$ pour calculer les barycentres déséquilibrés débités en utilisant des itérations de type Sinkhorn.

Publications:

- H. Janati et al, *Debiased Sinkhorn barycenters*, ICML'20.
- H. Janati et al, *L'OT entropique entre gaussiens a une forme fermée*, NeurIPS'20.

Chapitre 3 : Régression multi-tâches avec un antécédent OT Armés des connaissances entropiques nécessaires en matière d'OT, nous pouvons maintenant nous intéresser à (A.64) dans le contexte de l'imagerie cérébrale inverse. Ce problème correspond à la localisation de sources neuronales à partir de mesures électro-magnétiques hors de la tête. Formellement, cela équivaut à un problème inverse linéaire mal conditionné. Notre objectif est d'apporter des informations spatiales au modèle en le résolvant conjointement pour plusieurs individus sains – appelés *subjets*. L'antériorité P agit comme un liant entre les sujets, conduisant la solution vers des modèles neuronaux plus cohérents dans l'espace. En commençant par le célèbre Group Lasso, plusieurs modèles basés sur des normes de spartialité par blocs sont discutés et comparés à notre modèle basé sur l'OT. Notre proposition est *aware* de la géométrie du cortex, ce qui la rend moins encline à produire des valeurs aberrantes. En pratique, nous montrons comment ce problème peut être résolu en utilisant la descente de coordonnées proximales avec l'algorithme de Sinkhorn pour refléter à la fois la sparsité des sources et leur proximité spatiale. Les expériences ont été menées sur des données synthétiques et réelles et confrontées à d'autres techniques d'imagerie cérébrale.

Publications:

- H. Janati et al, *Régularisation de Wasserstein pour la régression multi-tâches clairsemée*, AISTATS'19.
- H. Janati et al, *Estimations minimales de Wasserstein : imagerie des sources EEG-MEG au niveau du groupe via un transport optimal*, IPMI'19.
- H. Janati et al, *Imagerie de la source multi-sujets avec régression multi-tâches clairsemée*, Neuroimage 2020.

Chapitre 4 : Transport optimal spatio-temporel Analyser les données EEG et MEG sans tenir compte de l'information temporelle, c'est comme casser un œuf avec un marteau : aussi réussi soit-il, ce n'est pas pour cela que vous avez acheté le marteau en premier lieu. Contrairement aux autres technologies d'imagerie cérébrale, l'EEG et la MEG mesurent l'activité cérébrale jusqu'à la milliseconde. L'extension la plus simple de l'OT aux données spatio-temporelles est peut-être de considérer le temps comme une caractéristique supplémentaire. de considérer le temps comme une caractéristique supplémentaire. Toutefois, cette approche négligerait son ordre chronologique. La distorsion temporelle dynamique (DTW) offre une méthode de principe pour comparer des séries temporelles sur la base d'une fonction de coût prédéfinie tout en respectant la chronologie des données. En fixant cette fonction de coût à une perte OT, on pourrait théoriquement aligner les séries temporelles en faisant correspondre les trames temporelles individuelles qui sont spatialement similaires.

Cependant, le DTW a deux limitations majeures : il n'est pas différentiable et est aveugle aux décalages temporels. Nous montrons que sa variante lisse, le soft-DTW, est en fait non seulement différentiable, mais qu'elle augmente de façon quadratique avec les décalages temporels. En combinant le soft-DTW et une formulation sans biais d'entropie de l'UOT, nous définissons une perte pour les données spatio-temporelles et proposons une méthode prête à l'emploi pour calculer les barycentres spatio-temporels.

Publications:

- H. Janati et al, *Alignements spatio-temporels : transport optimal dans l'espace et le temps*, AISTATS'20.
- H. Janati et al, *Barycentres de transport optimal pour les données spatio-temporelles*, Soumis.

Bibliography

Abraham A., Pedregosa F., Eickenberg M., Gervais P., Mueller A., Kossaifi J., Gramfort A., Thirion B., and Varoquaux G. (2014). "Machine learning for neuroimaging with scikit-learn". In: *Frontiers in Neuroinformatics* 8, p. 14.

Agueh M. and Carlier G. (2011). "Barycenters in the Wasserstein space". In: *SIAM Journal on Mathematical Analysis* 43.2, pp. 904–924.

Ahlfors S. P., Ilmoniemi R. J., and Hämäläinen M. S. (1992). "Estimates of visually evoked cortical currents". In: *Electroencephalography and Clinical Neurophysiology* 82.3, pp. 225–236.

Allena E., Erhardta E. B., Weib Y., Eichelec T., and Calhoun V. D. (2012). "Capturing inter-subject variability with group independent component analysis of fMRI data: a simulation study". In: *NIH* 59.4, pp. 4141–4159.

Amari S.-i., Karakida R., Oizumi M., and Cuturi M. (2019). "Information geometry for regularized optimal transport and barycenters of patterns". In: *Neural computation* 31.5, pp. 827–848.

Arjovsky M., Chintala S., and Bottou L. (2017). "Wasserstein generative adversarial networks". In: *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 214–223.

Baillet S., Mosher J. C., and Leahy R. M. (2001a). "Electromagnetic brain mapping". In: *IEEE Signal Processing Magazine* 18.6, pp. 14–30.

Baillet S., Mosher J. C., and Leahy R. M. (2001b). "Electromagnetic brain mapping". In: *IEEE Signal Processing Magazine* 18.6, pp. 14–30.

Baillet S. (2017). "Magnetoencephalography for brain electrophysiology and imaging". In: *Nature Neuroscience* 20.

Becker H., Albera L., Comon P., Gribonval R., Wendling F., and Merlet I. (2015). "Brain-Source Imaging: From sparse to tensor models". In: *IEEE Signal Processing Magazine* 32.6, pp. 100–112.

Bellman R. (1952). "On the theory of dynamic programming". In: *Proceedings of the National Academy of Sciences*. Vol. 38, 716–719.

Benamou J., Carlier G., Cuturi M., Nenna L., and Peyré G. (2015). "Iterative Bregman Projections For Regularized Transportation Problems". In: *Society for Industrial and Applied Mathematics*.

Bentin S., Allison T., Puce A., Perez E., and McCarthy G. (1996). "Electrophysiological Studies of Face Perception in Humans". In: *Journal of Cognitive Neuroscience* 8.6, pp. 551–565.

Berg C., Christensen J. P. R., and Ressel P. (1984). *Harmonic Analysis on Semigroups*. Berlin: Springer.

Bhatia R. (2007). *Positive Definite Matrices*. Princeton Series in Applied Mathematics. Princeton, NJ, USA: Princeton University Press.

Bhatia R., Jain T., and Lim Y. (2018). "On the Bures-Wasserstein distance between positive definite matrices". In: *Expositiones Mathematicae*.

Blondel M., Seguy V., and Rolet A. (2018). "Smooth and Sparse Optimal Transport". In: *international conference on artificial intelligence and statistics*.

Bonneel N., Peyré G., and Cuturi M. (2016). "Wasserstein Barycentric Coordinates: Histogram Regression Using Optimal Transport". In: *ACM Trans. Graph.* 35.4. ISSN: 0730-0301.

Bonneel N., Rabin J., Peyré G., and Pfister H. (2015). "Sliced and radon wasserstein barycenters of measures". In: *Journal of Mathematical Imaging and Vision* 51.1, pp. 22–45.

Brown P. C., Roediger H. L., and McDaniel M. A. (2014). *Make it Stick: The Science of Successful Learning*. Harvard University Press - Belknap.

Bures D. (1969). "An extension of Kakutani's theorem on infinite product measures to the tensor product of semifinite w^*-algebras". In: *Transactions of the American Mathematical Society* 135, pp. 199–212.

Cai C., Hashemi A., Diwakar M., Haufe S., Sekihara K., and Nagarajan S. S. (2020). "Robust estimation of noise for electromagnetic brain imaging with the Champagne algorithm". In: *NeuroImage* 225, p. 117411.

Candès E. J., Wakin M. B., and Boyd S. P. (2008). "Enhancing Sparsity by Reweighted L1 Minimization". In: *Journal of Fourier Analysis and Applications* 14.5, pp. 877–905.

Caruana R. (1993). "Multitask Learning: A Knowledge-Based Source of Inductive Bias". In: *Proceedings of the Tenth International Conference on Machine Learning*.

Castaño-Candamil S., Höhne J., Martínez-Vargas J.-D., An X.-W., Castellanos-Domínguez G., and Haufe S. (2015). "Solving the EEG inverse problem based on space–time–frequency structured sparsity constraints". In: *NeuroImage* 118, pp. 598 –612.

Charlier B., Feydy J., Glaunès J. A., Collin F.-D., and Durif G. (2020). "Kernel operations on the GPU, with autodiff, without memory overflows". In: *arXiv preprint arXiv:2004.11127*.

Chen C.-P. and Qi F. (2003). "The best bounds of harmonic sequence". In: *arXiv preprint math/0306233*.

Chen Y., Georgiou T. T., and Tannenbaum A. (2018). "Optimal transport for Gaussian mixture models". In: *IEEE Access* 7, pp. 6269–6278.

Chizat L., Peyré G., Schmitzer B., and Vialard F.-X. (2018a). "An interpolating distance between optimal transport and Fisher–Rao metrics". In: *Foundations of Computational Mathematics* 18.1, pp. 1–44.

Chizat L., Peyré G., Schmitzer B., and Vialard F.-X. (2018b). "Scaling algorithms for unbalanced optimal transport problems". In: *Mathematics of Computation* 87.314, pp. 2563–2609.

Csiszár I. (1963). "Eine Informationstheoretisehe Ungleichung und ihre Anwendung auf den Beweis der Ergodizitat von Markoffschen Ketten". In: *Magyar. Tud. Akad. Mat Kutató Int. Kozl.* 8, pp. 85–108.

Cuturi M. (2013). "Sinkhorn Distances: Lightspeed Computation of Optimal Transport". In: *Neural Information Processing Systems*.

Cuturi M. (2011). "Fast Global Alignment Kernels". In: *Proceedings of the 28th International Conference on International Conference on Machine Learning*. ICML'11. USA: Omnipress, pp. 929–936.

Cuturi M. and Blondel M. (2017). "Soft-DTW: a Differentiable Loss Function for Time-Series". In: *International Conference on Machine Learning*.

Cuturi M. and Doucet A. (2014). "Fast Computation of Wasserstein Barycenters". In: *Proceedings of the 31st International Conference on Machine Learning*. Ed. by E. P. Xing and T. Jebara. Vol. 32. Proceedings of Machine Learning Research 2. Bejing, China: PMLR, pp. 685–693.

Cuturi M. and Peyré G. (2016). "A smoothed dual approach for variational Wasserstein problems". In: *SIAM Journal on Imaging Sciences* 9.1, pp. 320–343.

Cuturi M. and Peyré G. (2018). "Semidual regularized optimal transport". In: *SIAM Review* 60.4, pp. 941–965.

Dale A. M., Liu A. K., Fischl B. R., Buckner R. L., Belliveau J. W., Lewine J. D., and Halgren E. (2000). "Dynamic Statistical Parametric Mapping". In: *Neuron* 26.1, pp. 55–67.

Delorme A., Palmer J., Onton J., Oostenveld R., and Makeig S. (2012). "Independent EEG sources are dipolar". In: *PloS one*.

Deslauriers-Gauthier S., Lina J.-M., Butler R., Bernier P.-M., Whittingstall K., Deriche R., and Descoteaux M. (2017). "Inference and Visualization of Information Flow in the Visual Pathway using dMRI and EEG". In: *MICCAI 2017 Medical Image Computing and Computer Assisted Intervention*. Québec, Canada.

Destrieux C., Fischl B., Dale A., and Halgren E. (2010). "Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature". In: *NeuroImage* 53.1, pp. 1 –15.

Di Marino S. and Gerolin A. (2020). "An Optimal Transport Approach for the Schrödinger Bridge Problem and Convergence of Sinkhorn Algorithm". In: *Journal of Scientific Computing* 85.2, p. 27. DOI: 10.1007/s10915-020-01325-7.

Dongdong G., Haoyue W., Zikai X., and Yinyu Y. (2019). "Interior-point Methods Strike Back: Solving the Wasserstein Barycenter Problem". In: *NeurIPS 2019*.

Doucet A., Vo B., Andrieu C., and Davy M. (2002). "Particle filtering for multi-target tracking and sensor management". In: *Proceedings of the Fifth International Conference on Information Fusion.* Vol. 1, 474–481 vol.1.

Dowson D. and Landau B. (1982). "The Fréchet distance between multivariate normal distributions". In: *Journal of Multivariate Analysis* 12.3, pp. 450 –455.

Dudley R. M. (1969). "The Speed of Mean Glivenko-Cantelli Convergence". In: *Ann. Math. Statist.* 40.1, pp. 40–50.

Duyn J. H. (2012). "The future of ultra-high field MRI and fMRI for study of the human brain". In: *Neuroimage* 62.2, pp. 1241–1248.

Engemann D. A. and Gramfort A. (2015). "Automated model selection in covariance estimation and spatial whitening of MEG and EEG signals". In: *NeuroImage* 108, pp. 328 –342. ISSN: 1053-8119.

Feydy J. (2020). "Geometric data analysis, beyond convolutions". Theses. Université Paris-Saclay.

Feydy J., Charlier B., Vialard F.-X., and Peyré G. (2017). "Optimal Transport for Diffeomorphic Registration". In: pp. 291–299.

Feydy J., Séjourné T., Vialard F.-X., Amari S.-i., Trouvé A., and Peyré G. (2019). "Interpolating between Optimal Transport and MMD using Sinkhorn Divergences". In: *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*.

Fischl B. (2012). "FreeSurfer". In: *NeuroImage* 62.2, pp. 774 –781.

Fischl B., Sereno M. I., and Dale A. M. (1999). "Cortical Surface-Based Analysis: II: Inflation, Flattening, and a Surface-Based Coordinate System". In: *NeuroImage* 9. Mathematics in Brain Imaging, pp. 195 –207. ISSN: 1053-8119.

Flamary R. and Courty N. (2017). *POT Python Optimal Transport library*.

Fournier N. and Guillin A. (2015). "On the rate of convergence in Wasserstein distance of the empirical measure". In: *Probability Theory and Related Fields* 162.3-4, pp. 707–738.

Frogner C., Zhang C., Mobahi H., Araya M., and Poggio T. A. (2015). "Learning with a Wasserstein loss". In: *Advances in Neural Information Processing Systems*, pp. 2053–2061.

Fuchs M., Wagner M., Köhler T., and Wischmann H.-A. (1999). "Linear and Nonlinear Current Density Reconstructions". In: *Journal of Clinical Neurophysiology* 16.3.

Gasso G., Rakotomamonjy A., and Canu S. (2009). "Recovering Sparse Signals With a Certain Family of Nonconvex Penalties and DC Programming". In: *IEEE Transactions on Signal Processing* 57.12, pp. 4686–4698.

Gelbrich M. (1990). "On a formula for the L2 Wasserstein metric between measures on Euclidean and Hilbert spaces". In: *Mathematische Nachrichten* 147.1, pp. 185–203.

Genevay A. (2019). "Entropy-regularized optimal transport for machine learning". PhD thesis. Paris Sciences et Lettres.

Genevay A., Peyre G., and Cuturi M. (2018). "Learning Generative Models with Sinkhorn Divergences". In: *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*. Ed. by A. Storkey and F. Perez-Cruz. Vol. 84. Proceedings of Machine Learning Research. Playa Blanca, Lanzarote, Canary Islands: PMLR, pp. 1608–1617.

Genevay A., Cuturi M., Peyré G., and Bach F. (2016). "Stochastic optimization for large-scale optimal transport". In: *Advances in Neural Information Processing Systems*, pp. 3440–3448.

Genevay A., Chizat L., Bach F., Cuturi M., and Peyré G. (2019). "Sample Complexity of Sinkhorn Divergences". In: *Proceedings of Machine Learning Research*. Ed. by K. Chaudhuri and M. Sugiyama. Vol. 89. Proceedings of Machine Learning Research. PMLR, pp. 1574–1583.

Gramfort A., Peyré G., and Cuturi M. (2015). "Fast Optimal Transport Averaging of Neuroimaging Data". In: *Proceedings of the Information Processing in Medical Imaging conference*.

Gramfort A., Luessi M., Larson E., Engemann D. A., Strohmeier D., Brodbeck C., Parkkonen L., and Hämäläinen M. (2013a). "MNE software for processing MEG and EEG data". In: *NeuroImage* 86.

Gramfort A., Strohmeier D., Haueisen J., Hämäläinen M., and Kowalski M. (2013b). "Time-frequency mixed-norm estimates: Sparse M/EEG imaging with non-stationary source activations". In: *NeuroImage* 70.0, pp. 410 –422.

Gramfort A., Papadopoulo T., Baillet S., and Clerc M. (2011). "Tracking cortical activity from M/EEG using graph cuts with spatiotemporal constraints". In: *NeuroImage* 54.3, pp. 1930 –1941.

Gramfort A., Luessi M., Larson E., Engemann D., Strohmeier D., Brodbeck C., Goj R., Jas M., Brooks T., Parkkonen L., and Hämäläinen M. (2013c). "MEG and EEG data analysis with MNE-Python". In: *Frontiers in Neuroscience* 7, p. 267. ISSN: 1662-453X.

Gretton A., Borgwardt K., Rasch M., Schölkopf B., and Smola A. (2006). "A kernel method for the two-sample-problem". In: *Advances in neural information processing systems* 19, pp. 513–520.

Gross J., Kujala J., Hamalainen M., Timmermann L., Schnitzler A., and Salmelin R. (2001). "Dynamic imaging of coherent sources: Studying neural interactions in the human brain." In: *Proc Natl Acad Sci U S A* 98.2, pp. 694–699.

Hämäläinen M. S. and Ilmoniemi R. J. (1994). "Interpreting magnetic fields of the brain: minimum norm estimates". In: *Medical & Biological Engineering & Computing* 32.1, pp. 35–42.

Hämäläinen M. S. and Sarvas J (1987). "Feasibility of the homogeneous head model in the interpretation of neuromagnetic fields". In: *Physics in Medicine and Biology* 32.1, p. 91.

Haufe S., Nikulin V. V., Ziehe A., Müller K.-R., and Nolte G. (2008). "Combining sparsity and rotational invariance in EEG/MEG source reconstruction". In: *NeuroImage* 42.2, pp. 726 –738. ISSN: 1053-8119.

Haufe S., Nikulin V., Ziehe A., Müller K.-R., and Nolte G. (2009). "Estimating vector fields using sparse basis field expansions". In: *Advances in Neural Information Processing Systems*. Ed. by D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou. Vol. 21. Curran Associates, Inc., pp. 617–624.

Henson R. N., Wakeman D. G., Litvak V., and Friston K. J. (2011). "A Parametric Empirical Bayesian Framework for the EEG/MEG Inverse Problem: Generative Models for Multi-Subject and Multi-Modal Integration". In: *Frontiers in human neuroscience* 5, pp. 76; 76–76.

Heusel M., Ramsauer H., Unterthiner T., Nessler B., and Hochreiter S. (2017). "Gans trained by a two time-scale update rule converge to a local Nash equilibrium". In: *Advances in neural information processing systems*, pp. 6626–6637.

Higham N. J. (2008). *Functions of Matrices: Theory and Computation (Other Titles in Applied Mathematics)*. USA: Society for Industrial and Applied Mathematics. ISBN: 0898716462.

Huber P. J. (1981). *Robust Statistics*. Wiley.

Ivan Gentil Christian Léonard L. R. (2017). "About the analogy between optimal transport and minimal entropy". In: *Annales de la Faculté des Sciences de Toulouse, Mathématiques.*

Jalali A., Ravikumar P., Sanghavi S., and Ruan C. (2010). "A Dirty Model for Multi-task Learning". In: *Advances in Neural Information Processing Systems*.

Janati H., Bazeille T., Thirion B., Cuturi M., and Gramfort A. (2019). "Group level EEG/MEG source imaging via Optimal Transport: minimum Wasserstein estimates". In: *Proceedings of the Information Processing in Medical Imaging conference*.

Janati H., Cuturi M., and Gramfort A. (2019). "Wasserstein regularization for sparse multi-task regression". In: *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*. Vol. 89. Proceedings of Machine Learning Research. PMLR.

Janati H., Cuturi M., and Gramfort A. (2020a). "Debiased Sinkhorn barycenters". In: *Proceedings of the 34th International Conference on Machine Learning*.

Janati H., Cuturi M., and Gramfort A. (2020b). "Spatio-temporal Alignments: Optimal transport through space and time". In: *Proceedings of the Twenty-third International Conference on Artificial Intelligence and Statistics*. Proceedings of Machine Learning Research. PMLR.

Janati H., Muzellec B., Peyré G., and Cuturi M. (2020a). "Entropic Optimal Transport between (Unbalanced) Gaussian Measures has a Closed Form". In: *Advances in neural information processing systems*.

Janati H., Bazeille T., Thirion B., Cuturi M., and Gramfort A. (2020b). "Multi-subject MEG/EEG source imaging with sparse multi-task regression". In: *NeuroImage*, p. 116847.

Kantorovich L. (1942). "On the translocation of masses". In: *C.R. Acad. Sci. URSS*.

Kanwisher N., McDermott J., and Chun M. M. (1997a). "The Fusiform Face Area: A Module in Human Extrastriate Cortex Specialized for Face Perception". In: *Journal of Neuroscience* 17.11, pp. 4302–4311.

Kanwisher N., McDermott J., and Chun M. M. (1997b). "The Fusiform Face Area: A Module in Human Extrastriate Cortex Specialized for Face Perception". In: *Journal of Neuroscience* 17.11, pp. 4302–4311.

Karamzadeh N., Medvedev A., Azari A., Gandjbakhche A., and Najafizadeh L. (2013). "Capturing dynamic patterns of task-based functional connectivity with EEG". In: *NeuroImage* 66, pp. 311–317.

Knight P. A., Ruiz D., and Uçar B. (2014). "A Symmetry Preserving Algorithm for Matrix Scaling". In: *SIAM Journal on Matrix Analysis and Applications* 35.

Knopp P. and Sinkhorn R. (1967). "Concerning nonnegative matrices and doubly stochastic matrices. " In: *Pacific Journal of Mathematics* 1.2, pp. 343–348.

Kohler T., Wagner M., Fuchs M., Wischmann H., Drenckhahn R., and Theissen A. (1996). "Depth normalization in MEG/EEG current density imaging". In: *Proceedings of 18th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. Vol. 2, 812–813 vol.2.

Kolouri S., Nadjahi K., Simsekli U., Badeau R., and Rohde G. (2019). "Generalized sliced wasserstein distances". In: *Advances in Neural Information Processing Systems*, pp. 261–272.

Kozunov V. V. and Ossadtchi A. (2015). "GALA: group analysis leads to accuracy, a novel approach for solving the inverse problem in exploratory analysis of group MEG recordings". In: *Frontiers in Neuroscience* 9, p. 107.

Kujala J., Sudre G., Vartiainen J., Liljeström M., Mitchell T., and Salmelin R. (2014). "Multivariate analysis of correlation between electrophysiological and hemodynamic responses during cognitive processing". In: *NeuroImage* 92, pp. 207 –216.

Kybic J., Clerc M., Abboud T., Faugeras O., Keriven R., and Papadopoulo T. (2005). "A Common Formalism for the Integral Formulations of the Forward EEG Problem". In: *IEEE Transactions on Medical Imaging* 24.1, pp. 12–28.

Larson E., Maddox R. K., and Lee A. K. C. (2014). "Improving spatial localization in MEG inverse imaging by leveraging intersubject anatomical differences". In: *Frontiers in Neuroscience* 8, p. 330.

LeCun Y. and Cortes C. (2010). "MNIST handwritten digit database". In:

Li J. and Wang J. Z. (2006). "Real-Time Computerized Annotation of Pictures". In: *Proceedings of the 14th ACM International Conference on Multimedia*. MM '06. New York, NY, USA: Association for Computing Machinery, 911–920.

Liero M., Mielke A., and Savaré G. (2016). "Optimal transport in competition with reaction: the Hellinger–Kantorovich distance and geodesic curves". In: *SIAM Journal on Mathematical Analysis* 48.4, pp. 2869–2911.

Liero M., Mielke A., and Savaré G. (2018). "Optimal entropy-transport problems and a new Hellinger–Kantorovich distance between positive measures". In: *Inventiones Mathematicae* 211.3, pp. 969–1117.

Liese F. and Vajda I. (2006). "On Divergences and Informations in Statistics and Information Theory". In: *IEEE Transactions on Information Theory* 52.10, pp. 4394–4412.

Lim M., Ales J., Cottereau B. M., Hastie T., and Norcia A. M. (2017). "Sparse EEG/MEG source estimation via a group lasso". In: *PLOS*.

Lin F.-H., Witzel T., Ahlfors S. P., Stufflebeam S. M., Belliveau J. W., and Hämäläinen M. S. (2006). "Assessing and improving the spatial accuracy in MEG source localization by depth-weighted minimum-norm estimates". In: *Neuroimage* 31.1, pp. 160–171.

Litvak V. and Friston K. (2008). "Electromagnetic source reconstruction for group studies". In: *NeuroImage* 42.4, pp. 1490 –1498. ISSN: 1053-8119.

Lorenz D. A., Manns P., and Meyer C. (2019). "Quadratically regularized optimal transport". In: *Applied Mathematics & Optimization*, pp. 1–31.

Lozano A. and Swirszcz G. (2012). "Multi-level Lasso for Sparse Multi-task Regression". In: *ICML*.

Luise G., Rudi A., Pontil M., and Ciliberto C. (2018). "Differential properties of sinkhorn approximation for learning with wasserstein distance". In: *Advances in Neural Information Processing Systems*, pp. 5859–5870.

Luise G., Salzo S., Pontil M., and Ciliberto C. (2019). "Sinkhorn Barycenters with Free Support via Frank-Wolfe Algorithm". In: *Advances in Neural Information Processing Systems*.

Maaten L. and Hinton G. (2008). "Visualizing High-Dimensional Data using t-SNE". In: *Journal of Machine Learning Research* 9, pp. 2579–2605.

Mainini E. (2012). "A description of transport cost for signed measures". In: *Journal of Mathematical Sciences* 181.6, pp. 837–855.

Makeig S., Jung T.-P., Ghahremani A. J. B. andDara, and Sejnowski T. J. (1997). "Blind separation of auditory event-related brain responses into independent components". In: *Proceedings of the National Academy of Sciences (PNAS)*.

Malagò L., Montrucchio L., and Pistone G. (2018). "Wasserstein riemannian geometry of positive definite matrices". In: *arXiv preprint arXiv:1801.09269*.

Massias M., Fercoq O., Gramfort A., and Salmon J. (2018). "Generalized Concomitant Multi-Task Lasso for Sparse Multimodal Regression". In: *Proceedings of Machine Learning Research*. Ed. by A. Storkey and F. Perez-Cruz. Vol. 84. PMLR, pp. 998–1007.

Mena G. and Niles-Weed J. (2019). "Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem". In: *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., pp. 4541–4551.

Michel C. M., Murray M. M., Lantz G., Gonzalez S., Spinelli L., and Peralta R. G. de (2004). "EEG source imaging". In: *Clinical Neurophysiology* 115.10, pp. 2195–2222.

Monge G. (1781). "Mémoire sur la théorie des déblais et des remblais". In: *Histoire de l'Académie Royale des Sciences*, pp. 666 –704.

Mosher J., Leahy R., and Lewis P. (1999). "EEG and MEG: Forward Solutions for Inverse Methods". In: *IEEE Transactions on Biomedical Engineering* 46.3, pp. 245–259.

Mosher J. C. and Leahy R. M. (1999). "Source localization using recursively applied and projected (RAP) MUSIC". In: *IEEE Transactions on Signal Processing* 47.2, pp. 332–340.

Muzellec B. and Cuturi M. (2018). "Generalizing Point Embeddings using the Wasserstein Space of Elliptical Distributions". In: *Advances in Neural Information Processing Systems 31*. Curran Associates, Inc., pp. 10237–10248.

Mäkelä N., Stenroos M., Sarvas J., and Ilmoniemi R. J. (2018). "Truncated RAP-MUSIC (TRAP-MUSIC) for MEG and EEG source localization". In: *NeuroImage* 167, pp. 73 –83.

Ndiaye E., Fercoq O., Gramfort A., Leclère V., and Salmon J. (2017). "Efficient Smoothed Concomitant Lasso Estimation for High Dimensional Regression". In: *Journal of Physics: Conference Series* 904.1, p. 012006.

Negahban S. and Wainwright M. J. (2008). "Joint support recovery under high-dimensional scaling: Benefits and perils of l1,∞ regularization". In: *Advances in Neural Information Processing Systems*.

Niles-Weed J. and Rigollet P. (2019). "Estimation of wasserstein distances in the spiked transport model". In: *arXiv preprint arXiv:1909.07513*.

Nunez P. and Srinivasan R. (2006). *Electric fields of the brain: the neurophysics of EEG*. Oxford University Press.

Okada Y. (1993). "Empirical bases for constraints in current-imaging algorithms". In: *Brain Topography* 5, 373–377.

Olkin I. and Pukelsheim F. (1982). "The distance between two random vectors with given dispersion matrices". In: *Linear Algebra and its Applications* 48, pp. 257–263.

Ou W., Nummenmaa A., Ahveninen J., Belliveau J. W., Hämäläinen M. S., and Golland P. (2010). "Multimodal functional imaging using fMRI-informed regional EEG/MEG source estimation". In: *NeuroImage* 52.1, pp. 97 –108.

Owen A. B. (2007). "A robust hybrid of lasso and ridge regression". In: *Contemporary Mathematics* 443, pp. 59–72.

Pascual-Marqui R. (2002). "Standardized low-resolution brain electromagnetic tomography (sLORETA): technical details." In: *Methods Find Exp Clin Pharmacol* 24, D:5–12.

Paszke A., Gross S., Chintala S., Chanan G., Yang E., DeVito Z., Lin Z., Desmaison A., Antiga L., and Lerer A. (2017). "Automatic differentiation in PyTorch". In:

Paty F.-P. and Cuturi M. (2019). "Subspace Robust Wasserstein Distances". In: *International Conference on Machine Learning*, pp. 5072–5081.

Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., and Duchesnay E. (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Peyré G. and Cuturi M. (2018). *Computational Optimal Transport*.

Poincaré H. (1902). *La science et l'hypothèse*. Ed. by E. Flammarion.

Poline J.-B., Thirion B., Roche A., and Meriaux S. (2010). "Intersubject v[plariability in fMRI data: Causes, consequences, and related analysis strategies". In:

Profeta A. and Sturm K.-T. (2018). "Heat Flow with Dirichlet Boundary Conditions via Optimal Transport and Gluing of Metric Measure Spaces". arXiv Preprint 1809.00936.

Proudfoot M., Woolrich M. W., Nobre A. C., and Turner M. R. (2014). "Magnetoencephalography". In: *Practical Neurology* 14.5, pp. 336–343.

Rabin J., Peyré G., Delon J., and Bernot M. (2011). "Wasserstein barycenter and its application to texture mixing". In: *International Conference on Scale Space and Variational Methods in Computer Vision*. Springer, pp. 435–446.

Rakotomamonjy A., Gasso G., and Salmon J. (2019). "Screening rules for Lasso with non-convex Sparse Regularizers". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by K. Chaudhuri and R. Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. Long Beach, California, USA: PMLR, pp. 5341–5350.

Ramdas A., Trillos N., and Cuturi M. (2017). "On wasserstein two-sample testing and related families of nonparametric tests". In: *Entropy* 19.2, p. 47.

Ramón y Cajal S. (1899). *Comparative study of the sensory areas of the human cortex.* Harvard.

Rényi A. et al. (1961). "On measures of entropy and information". In: *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California.

Richtárik P. and Takáč M. (2014). "Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function". In: *Mathematical Programming* 144.1-2, pp. 1–38.

Rigollet P. and Weed J. (2018). "Entropic optimal transport is maximum-likelihood deconvolution". In: *Comptes Rendus Mathematique* 356.11, pp. 1228 –1235.

Robinson E. C., Jbabdi S., Glasser M. F., Andersson J., Burgess G. C., Harms M. P., Smith S. M., Essen D. C. V., and Jenkinson M. (2014). "MSM: A new flexible framework for Multimodal Surface Matching". In: *NeuroImage* 100, pp. 414 –426. ISSN: 1053-8119.

Rockafellar R. T. (1970). *Convex Analysis*. Princeton University Press. ISBN: 9780691015866.

Saigo H., Jean-Philippe, Vert, Ueda N., and Akutsu T. (2004). "Protein homology detection using string alignment kernels". In: *Bioinformatics* 20.11, 1682–1689.

Sakoe H. and Chiba S. (1978). "Dynamic programming algorithm optimization for spoken word recognition". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 26.1, pp. 43–49.

Santambrogio F. (2015). *Optimal transport for applied mathematicians*. Birkhauser.

Sato M., Yamashita O., Sato M.-A., and Miyawaki Y. (2018). "Information spreading by a combination of MEG source estimation and multivariate pattern classification". In: *PloS one* 13.6, e0198806–e0198806.

Schmitzer B. (2016). "Stabilized Sparse Scaling Algorithms for Entropy Regularized Transport Problems". In: *SIAM Journal on Scientific Computing* 41 (3), A1443–A1481.

Séjourné T., Feydy J., Vialard F.-X., Trouvé A., and Peyré G. (2019). "Sinkhorn Divergences for Unbalanced Optimal Transport". In: *arXiv preprint arXiv:1910.12958*.

Shirdhonkar S. and Jacobs D. W. (2008). "Approximate earth mover's distance in linear time". In: *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, pp. 1–8.

Slotnick S. D., Klein S. A., Carney T., Sutter E., and Dastmalchi S. (1999). "Using multi-stimulus VEP source localization to obtain a retinotopic map of human primary visual cortex". In: *Clinical Neurophysiology* 110.10, pp. 1793 –1800.

Sokol S (1983). "Abnormal evoked potential latencies in amblyopia". In: *The British journal of ophthalmology* 67.5, pp. 310–314. DOI: 10.1136/bjo.67.5.310.

Solomon J., Goes F. de, Peyré G., Cuturi M., Butscher A., Nguyen A., Du T., and Guibas L. (2015). "Convolutional Wasserstein Distances: Efficient Optimal Transportation on Geometric Domains". In: *ACM Trans. Graph.* 34.4, 66:1–66:11. ISSN: 0730-0301.

Sriperumbudur B. K., Fukumizu K., Gretton A., Schölkopf B., and Lanckriet G. R. G. (2012). "On the empirical estimation of integral probability metrics". In: *Electron. J. Statist.* 6, pp. 1550–1599.

Städler N., Bühlmann P., and Geer S. van de (2010). "L1-penalization for mixture regression models". In: *TEST* 19.2, pp. 209–256.

Stanley R. P. (2011). *Enumerative Combinatorics: Volume 1*. 2nd. New York, NY, USA: Cambridge University Press. ISBN: 1107602629, 9781107602625.

Strohmeier D., Bekhti, Y. Haueisen J., and Gramfort A. (2016). "The Iterative Reweighted Mixed-Norm Estimate for Spatio-Temporal MEG/EEG Source Reconstruction". In: *IEEE Transactions on Medical Imaging* 35.10, pp. 2218–2228.

Strohmeier D., Gramfort A., and Haueisen J. (2015). "MEG/EEG Source Imaging with a Non-Convex Penalty in the Time-Frequency Domain". In: *International Workshop on Pattern Recognition in NeuroImaging (PRNI)*, pp. 21–24.

Strohmeier D., Haueisen J., and Gramfort A. (2014). "Improved MEG/EEG source localization with reweighted mixed-norms". In: *2014 International Workshop on Pattern Recognition in Neuroimaging*. IEEE, pp. 1–4.

Sulanke R. A. (2003). "Objects Counted by the Central Delannoy Numbers". In: *Journal of Integer Sequences*.

Sullivan C. and Kaszynski A. (2019). "PyVista: 3D plotting and mesh analysis through a streamlined interface for the Visualization Toolkit (VTK)". In: *Journal of Open Source Software* 4.37, p. 1450.

Tadel F., Baillet S., Mosher J. C., Pantazis D., and Leahy R. M. (2011). "Brainstorm: A User-Friendly Application for MEG/EEG Analysis". In: *Computational Intelligence and Neuroscienc* vol. 2011.

Takatsu A. (2011). "Wasserstein geometry of Gaussian measures". In: *Osaka J. Math.* 48.4, pp. 1005–1026.

Takeda Y., Suzuki K., Kawato M., and Yamashita O. (2019). "MEG Source Imaging and Group Analysis Using VBMEG". In: *Frontiers in Neuroscience* 13, p. 241.

Talakoub O., Popovic M. R., Navaro J., Hamani C., Fonoff E. T., and Wong W. (2015). "Temporal alignment of electrocorticographic recordings for upper limb movement". In: *Frontiers in Neuroscience* 8, p. 431.

Taylor J. R., Williams N., Cusack R., Auer T., Shafto M. A., Dixon M., Tyler L. K., Cam-CAN, and Henson R. N. (2017). "The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) data repository: Structural and functionalMRI , MEG, and cognitive data from a cross-sectional adult lifespan sample". In: *NeuroImage* 144. Data Sharing Part II, pp. 262 –269. ISSN: 1053-8119.

Thorpe M., Park S., Kolouri S., Rohde G. K., and Slepčev D. (2017). "A Transportation L(p) Distance for Signal Analysis". In: *Journal of mathematical imaging and vision* 59.2, pp. 187–210.

Tibshirani R. (1996). "Regression shrinkage and selection via the Lasso". In: *J. Roy. Statist. Soc. Ser. B* 58.1, pp. 267–288.

Titouan V., Flamary R., Courty N., Tavenard R., and Chapel L. (2019). "Sliced Gromov-Wasserstein". In: *Advances in Neural Information Processing Systems*, pp. 14726–14736.

Uutela K., Hämäläinen M. S., and Somersalo E. (1999). "Visualization of Magnetoencephalographic Data Using Minimum Current Estimates". In: *NeuroImage* 10.2, pp. 173–180. ISSN: 1053-8119.

Van Veen B. D., Van Drongelen W., Yuchtman M., and Suzuki A. (1997). "Localization of brain electrical activity via linearly constrained minimum variance spatial filtering". In: *IEEE Transactions on Biomedical Engineering* 44.9, pp. 867–880.

Varoquaux G., Gramfort A., Pedregosa F., Michel V., and Thirion B. (2011). "Multi-subject dictionary learning to segment an atlas of brain spontaneous activity". In: *Information Processing in Medical Imaging*. Vol. 6801. Springer, pp. 562–573.

Vega-Hernández M., Martínez-Montes E., Sánchez-Bornot J. M., Lage-Castellanos A., and Valdés-Sosa P. A. (2008). "Penalized least squares methods for solving the EEG inverse problem". In: *Statistica Sinica* 18.4, p. 1535.

Wakeman D. and Henson R. (2015). "A multi-subject, multi-modal human neuroimaging dataset". In: *Scientific Data* 2.150001.

Whelan R., Lonergan R., Kiiski H., Nolan H., Kinsella K, Bramham J., O'Brien M, Reilly R., Hutchinson M., and Tubridy N (2010). "A high-density ERP study reveals latency, amplitude, and topographical

differences in multiple sclerosis patients versus controls". In: *Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology* 121, pp. 1420–6.

Williams B., M.Toussaint, and Storkey. A. (2006). "Extracting motion primitives from natural handwriting data". In: *ICANN*. Vol. 2, 634–643.

Wilson A. G. (1969). "The use of entropy maximising models, in the theory of trip distribution, mode split and route split". In: *Journal of transport economics and policy*, pp. 108–126.

Wipf D. and Nagarajan S. (2009). "A unified Bayesian framework for MEG/EEG source imaging". In: *NeuroImage* 44.3, pp. 947–966.

Yarkoni T. (2014). *Neurosynth core tools v0.3.1*.

Yuan M. and Lin Y. (2006). "Model selection and estimation in regression with grouped variables". In: *Journal of the Royal Statistical Society* 68.1, pp. 49–67.

Zhongming Liu, Lei Ding, and Bin He (2006). "Integration of EEG/MEG with MRI and fMRI". In: *IEEE Engineering in Medicine and Biology Magazine* 25.4, pp. 46–53.