Gromov-Wasserstein Distances between Gaussian Distributions

Antoine Salmona[1], Julie Delon[2], Agnès Desolneux[1] *

[1] ENS Paris-Saclay, CNRS, Centre Borelli UMR 9010
[2] Université de Paris, CNRS, MAP5 UMR 8145 and Institut Universitaire de France

April 19, 2021

Abstract

The Gromov-Wasserstein distances were proposed a few years ago to compare distributions which do not lie in the same space. In particular, they offer an interesting alternative to the Wasserstein distances for comparing probability measures living on Euclidean spaces of different dimensions. In this paper, we focus on the Gromov-Wasserstein distance with a ground cost defined as the squared Euclidean distance and we study the form of the optimal plan between Gaussian distributions. We show that when the optimal plan is restricted to Gaussian distributions, the problem has a very simple linear solution, which is also solution of the linear Gromov-Monge problem. We also study the problem without restriction on the optimal plan, and provide lower and upper bounds for the value of the Gromov-Wasserstein distance between Gaussian distributions.

Keywords— optimal transport, Wasserstein distance, Gromov-Wasserstein distance, Gaussian distributions.
MSC 2020 subject classifications : 60E99, 68T09, 62H25, 49Q22.

1 Introduction

Optimal transport (OT) theory has become nowadays a major tool to compare probability distributions. It has been increasingly used over the last past years in various applied fields such as economy [11], image processing [20, 21], machine learning [4, 5] or more generally data science [18], with applications to domain adaptation [9] or generative models [3, 12], to name just a few.

Given two probability distributions μ and ν on two Polish spaces $(\mathcal{X}, d_{\mathcal{X}})$ and $(\mathcal{Y}, d_{\mathcal{Y}})$ and a positive lower semi-continuous cost function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$, optimal transport focuses on solving the following optimization problem

$$\inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y) d\pi(x,y), \tag{1.1}$$

where $\Pi(\mu,\nu)$ is the set of measures on $\mathcal{X} \times \mathcal{Y}$ with marginals μ and ν. When $\mathcal{X}$ and $\mathcal{Y}$ are equal and Euclidean, typically $\mathbb{R}^d$, and $c(x,y) = \|x-y\|^p$ with $p \geq 1$, Equation (1.1) induces a distance over the set of measures with finite moment of order p, known as the p-Wasserstein distance W_p:

$$W_p(\mu,\nu) = \left(\inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x-y\|^p d\pi(x,y) \right)^{\frac{1}{p}}, \tag{1.2}$$

or equivalently

$$W_p^p(\mu,\nu) = \inf_{X\sim\mu, Y\sim\nu} \mathbb{E}[\|X-Y\|^p], \tag{1.3}$$

*The authors acknowledge support from the French Research Agency through the PostProdLEAP project (ANR-19-CE23-0027-01) and the MISTIC project (ANR-19-CE40-005).

where the notation $X \sim \mu$ means that X is a random variable with probability distribution μ. It is known that Equation (1.1) always admits a solution [27, 26, 22] , i.e. the infinimum is always reached. Moreover, in the case of W_2, it is known [6] that if μ is absolutely continuous, then the *optimal transport plan* π^* is unique and has the form $\pi^* = (Id, T)\#\mu$ where $\#$ is the push-forward operator and $T : \mathbb{R}^d \to \mathbb{R}^d$ is an application called *optimal transport map*, satisfying $T\#\mu = \nu$. The 2-Wasserstein distance W_2 admits a closed-form expression [10, 24] when $\mu = \mathcal{N}(m_0, \Sigma_0)$ and $\nu = \mathcal{N}(m_1, \Sigma_1)$ are two Gaussian measures with means $m_0 \in \mathbb{R}^d$, $m_1 \in \mathbb{R}^d$ and covariance matrices $\Sigma_0 \in \mathbb{R}^{d\times d}$ and $\Sigma_1 \in \mathbb{R}^{d\times d}$, that is given by

$$W_2^2(\mu,\nu) = \|m_1 - m_0\|^2 + \operatorname{tr}\left(\Sigma_0 + \Sigma_1 - 2\left(\Sigma_0^{\frac{1}{2}}\Sigma_1\Sigma_0^{\frac{1}{2}}\right)^{\frac{1}{2}}\right), \tag{1.4}$$

where for any symmetric semi-definite positive M, $M^{\frac{1}{2}}$ is the unique symmetric semi-definite positive squared root of M. Moreover, if Σ_0 is non-singular, then the optimal transport map T is affine and is given by

$$\forall x \in \mathbb{R}^d,\ T(x) = m_1 + \Sigma_0^{-\frac{1}{2}}\left(\Sigma_0^{\frac{1}{2}}\Sigma_1\Sigma_0^{\frac{1}{2}}\right)^{\frac{1}{2}}\Sigma_0^{-\frac{1}{2}}(x - m_0), \tag{1.5}$$

and the corresponding optimal transport plan π^* is a degenerate Gaussian measure.

For some applications such as shape matching or word embedding, an important limitation of classic OT lies in the fact that it is not invariant to rotations and translations and more generally to *isometries*. Moreover, OT implies that we can define a relevant cost function to compare spaces $\mathcal{X}$ and $\mathcal{Y}$. Thus, when for instance μ is a measure on $\mathbb{R}^2$ and ν a measure on $\mathbb{R}^3$, it is not straightforward to design a cost function $c : \mathbb{R}^2 \times \mathbb{R}^3 \to \mathbb{R}$ and so one cannot define easily an OT distance to compare μ with ν. To overcome these limitations, several extensions of OT have been proposed [1, 7, 17]. Among them, the most famous one is probably the Gromov-Wasserstein (GW) problem [16]: given two Polish spaces $(\mathcal{X}, d_{\mathcal{X}})$ and $(\mathcal{Y}, d_{\mathcal{Y}})$, each endowed respectively with probability measures μ and ν, and given two measurable functions $c_{\mathcal{X}} : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ and $c_{\mathcal{Y}} : \mathcal{Y}\times\mathcal{Y} \to \mathbb{R}$, it aims at finding

$$GW_p(c_{\mathcal{X}}, c_{\mathcal{Y}}, \mu, \nu) = \left(\inf_{\pi\in\Pi(\mu,\nu)} \int_{\mathcal{X}^2\times\mathcal{Y}^2} |c_{\mathcal{X}}(x,x') - c_{\mathcal{Y}}(y,y')|^p d\pi(x,y)d\pi(x',y')\right)^{\frac{1}{p}}, \tag{1.6}$$

with $p \geq 1$. As for classic OT, it can be shown that Equation (1.6) always admits a solution (see [25]). The GW problem can be seen as a quadratic optimization problem in π, as opposed to OT, which is a linear optimization problem in π. It induces a distance over the space of *metric measure spaces* (i.e. the triplets $(\mathcal{X}, d_{\mathcal{X}}, \mu)$) quotiented by the *strong isomorphisms* [18] [1]. The fundamental metric properties of GW_p have been studied in depth in [23, 16, 8]. In the Euclidean setting, when $\mathcal{X} = \mathbb{R}^m$, $\mathcal{Y} = \mathbb{R}^n$, with m not necessarily being equal to n, and for the natural choice of costs $c_{\mathcal{X}} = \|.\|^2_{\mathbb{R}^m}$ and $c_{\mathcal{Y}} = \|.\|^2_{\mathbb{R}^n}$, where $\|.\|_{\mathbb{R}^m}$ means the Euclidean norm on $\mathbb{R}^m$, it can be easily shown that $GW_2(c_{\mathcal{X}}, c_{\mathcal{Y}}, \mu, \nu)$ is invariant to isometries. With a slight abuse of notations, we will note in the following $GW_2(\mu,\nu)$ instead of $GW_2(\|.\|^2_{\mathbb{R}^m}, \|.\|^2_{\mathbb{R}^n}, \mu, \nu)$.

In this work, we focus on the problem of Gromov-Wasserstein between Gaussian measures. Given $\mu = \mathcal{N}(m_0, \Sigma_0)$, with $m_0 \in \mathbb{R}^m$ and with covariance matrix $\Sigma_0 \in \mathbb{R}^{m\times m}$, and $\nu = \mathcal{N}(m_1, \Sigma_1)$, with $m_1 \in \mathbb{R}^n$ and with covariance matrix $\Sigma_1 \in \mathbb{R}^{n\times n}$, we aim to solve

$$GW_2^2(\mu,\nu) = \inf_{\pi\in\Pi(\mu,\nu)} \int\int \left(\|x - x'\|^2_{\mathbb{R}^m} - \|y - y'\|^2_{\mathbb{R}^n}\right)^2 d\pi(x,y)d\pi(x',y'), \tag{GW}$$

or equivalently

$$GW_2^2(\mu,\nu) = \inf_{X,X',Y,Y'\sim\pi\otimes\pi} \mathbb{E}\left[\left(\|X - X'\|^2_{\mathbb{R}^m} - \|Y - Y'\|^2_{\mathbb{R}^n}\right)^2\right], \tag{1.7}$$

[1] We say that $(\mathcal{X}, d_{\mathcal{X}}, \mu)$ is strongly isomorphic to $(\mathcal{Y}, d_{\mathcal{Y}}, \nu)$ if it exists a bijection $\phi : \mathcal{X} \to \mathcal{Y}$ such that ϕ is an isometry $(d_{\mathcal{Y}}(\phi(y), \phi(y')) = d_{\mathcal{X}}(x,x'))$, and $\phi\#\mu = \nu$.

where for $x, x' \in \mathbb{R}^m$ and $y, y' \in \mathbb{R}^n$, $(\pi \otimes \pi)(x, x', y, y') = \pi(x, y)\pi(x', y')$. In particular, can we find equivalent formulas to (1.4) and (1.5) in the case of Gromov-Wasserstein? In Section 2, we derive an equivalent formulation of the Gromov-Wasserstein problem. This formulation is not specific to Gaussian measures but to all measures with finite order 4 moment. It takes the form of a sum of two terms depending respectively on co-moments of order 2 and 4 of π. Then in Section 3, we derive a lower bound by simply optimizing both terms separately. In Section 4, we show that the problem restricted to Gaussian optimal plans admits an explicit solution and this solution is closely related to Principal Components Analysis (PCA). In Section 5, we study the tightness of the bounds found in the previous sections and we exhibit a particular case where we are able to compute exactly the value of $GW_2^2(\mu, \nu)$ and the optimal plan π^* which achieves it. Finally, Section 6 discusses the form of the solution in the general case, and the possibility that the optimal plan between two Gaussian distributions is always Gaussian.

Notations

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

- The notation $Y \sim \mu$ means that Y is a random variable with probability distribution μ.
- If μ is a positive measure on $\mathcal{X}$ and $T : \mathcal{X} \to \mathcal{Y}$ is an application, $T\#\mu$ stands for the push-forward measure of μ by T, i.e. the measure on $\mathcal{Y}$ such that $\forall A \in \mathcal{Y}$, $(T\#\mu)(A) = \mu(T^{-1}(A))$.
- If X and Y are random vectors on $\mathbb{R}^m$ and $\mathbb{R}^n$, we denote $\text{Cov}(X, Y)$ the matrix of size $m \times n$ of the form $\mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])^T\right]$.
- the notation $\text{tr}(M)$ denotes the trace of a matrix M.
- $\|M\|_{\mathcal{F}}$ stands for the Frobenius norm of a matrix M, i.e. $\|M\|_{\mathcal{F}} = \sqrt{\text{tr}(M^T M)}$.
- $\text{rk}(M)$ stands for the rank of a matrix M.
- I_n is the identity matrix of size n.
- $\tilde{I}_n$ stands for any matrix of size n of the form $\text{diag}((\pm 1)_{i \leq n})$
- Suppose $n \leq m$. For $A \in \mathbb{R}^{m \times m}$, we denote $A^{(n)} \in \mathbb{R}^{n \times n}$ the submatrix containing the n first rows and the n first columns of A.
- Suppose $n \leq m$. For $A \in \mathbb{R}^{n \times n}$, we denote $A^{[m]} \in \mathbb{R}^{m \times m}$ the matrix of the form $\begin{pmatrix} A & 0 \\ 0 & 0 \end{pmatrix}$.
- We denote $S_n(\mathbb{R})$ the set of symmetric matrices of size n, $S_n^+(\mathbb{R})$ the set of semi-definite positive matrices, and $S_n^{++}(\mathbb{R})$ the set of definite positive matrices.
- $\mathbb{1}_{n,m} = (1)_{i \leq n, j \leq m}$ denotes the matrix of ones with n rows and m columns.
- $\|x\|_{\mathbb{R}^n}$ stands for the Euclidean norm of $x \in \mathbb{R}^n$. We will denote $\|x\|$ when there is no ambiguity about the dimension.
- $\langle x, x' \rangle_n$ stands for the Euclidean inner product in $\mathbb{R}^n$ between x and x'.

2 Derivation of the general problem

In this section, we derive an equivalent [2] formulation of problem (GW) which takes the form of a functional of co-moments of order 2 and 4 of π. This formulation is not specific to Gaussian measures but to all measures with finite 4^{th} order moment.

Theorem 2.1. *Let μ be a probability measure on $\mathbb{R}^m$ with mean vector $m_0 \in \mathbb{R}^m$ and covariance matrix $\Sigma_0 \in \mathbb{R}^{m\times m}$ such that $\int \|x\|^4 d\mu < +\infty$ and ν a probability measure on $\mathbb{R}^n$ with mean vector m_1 and covariance matrix $\Sigma_1 \in \mathbb{R}^{n\times n}$ such that $\int \|y\|^4 d\nu < +\infty$. Let P_0, D_0 and P_1, D_1 be respective diagonalizations of $\Sigma_0 (= P_0 D_0 P_0^T)$ and $\Sigma_1 (= P_1 D_1 P_1^T)$. Let us define $T_0 : x \in \mathbb{R}^m \mapsto P_0^T(x - m_0)$ and $T_1 : y \in \mathbb{R}^n \mapsto P_1^T(y - m_1)$. Then problem* (GW) *is equivalent to problem*

$$\sup_{X\sim T_0\#\mu, Y\sim T_1\#\nu} \sum_{i,j} \mathrm{Cov}(X_i^2, Y_j^2) + 2\|\mathrm{Cov}(X,Y)\|_{\mathcal{F}}^2, \tag{supCOV}$$

where $X = (X_1, X_2, \ldots, X_m)^T$, $Y = (Y_1, Y_2, \ldots, Y_n)^T$, and $\|.\|_{\mathcal{F}}$ is the Frobenius norm.

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

Lemma 2.1. *We denote $\mathbb{O}_m = \{O \in \mathbb{R}^{m\times m} | \, O^T O = I_m\}$ the set of orthogonal matrices of size m. Let μ and ν be two probability measures on $\mathbb{R}^m$ and $\mathbb{R}^n$. Let $T_m : x \mapsto O_m x + x_m$ and $T_n : y \mapsto O_n y + y_n$ be two affine applications with $x_m \in \mathbb{R}^m$, $O_m \in \mathbb{O}_m$, $y_n \in \mathbb{R}^n$, and $O_n \in \mathbb{O}_n$. Then $GW_2(T_m\#\mu, T_n\#\nu) = GW_2(\mu,\nu)$.*

Lemma 2.2 (**Vayer, 2020, [25]**)**.** *Suppose there exist some scalars a, b, c such that $c_{\mathcal{X}}(x,x') = a\|x\|_{\mathbb{R}^m}^2 + b\|x'\|_{\mathbb{R}^m}^2 + c\langle x, x'\rangle_m$, where $\langle .,.\rangle_m$ denotes the inner product on $\mathbb{R}^m$, and $c_{\mathcal{Y}}(y,y') = a\|y\|_{\mathbb{R}^n}^2 + b\|y'\|_{\mathbb{R}^n}^2 + c\langle y, y'\rangle_n$. Let μ and ν be two probability measures respectively on $\mathbb{R}^m$ and $\mathbb{R}^n$. Then*

$$GW_2^2(c_{\mathcal{X}}, c_{\mathcal{Y}}, \mu, \nu) = C_{\mu,\nu} - 2 \sup_{\pi\in\Pi(\mu,\nu)} Z(\pi), \tag{2.1}$$

where $C_{\mu,\nu} = \int c_{\mathcal{X}}^2 d\mu d\mu + \int c_{\mathcal{Y}}^2 d\nu d\nu - 4ab \int \|x\|_{\mathbb{R}^m}^2 \|y\|_{\mathbb{R}^n}^2 d\mu d\nu$ and

$$\begin{aligned} Z(\pi) = (a^2+b^2) \int \|x\|_{\mathbb{R}^m}^2 \|y\|_{\mathbb{R}^n}^2 d\pi(x,y) + c^2 \left\| \int xy^T d\pi(x,y) \right\|_{\mathcal{F}}^2 \\ + (a+b)c \int \left(\|x\|_{\mathbb{R}^m}^2 \langle \mathbb{E}_{Y\sim\nu}[Y], y\rangle_n + \|y\|_{\mathbb{R}^n}^2 \langle \mathbb{E}_{X\sim\mu}[X], x\rangle_m \right) d\pi(x,y). \end{aligned} \tag{2.2}$$

Proof of theorem 2.1. Using Lemma 2.1, we can focus without any loss of generality on centered Gaussian measures with diagonal covariance matrices. Thus, defining $T_0 : x \in \mathbb{R}^m \mapsto P_0^T(x - m_0)$ and $T_1 : y \in \mathbb{R}^n \mapsto P_1^T(y - m_1)$ and then applying Lemma 2.2 on $GW_2(T_0\#\mu, T_1\#\nu)$ with $a = 1$, $b = 1$, and $c = 2$ while remarking that the last term in Equation (2.2) is null because $\mathbb{E}_{X\sim T_0\#\mu}[X] = 0$ and $\mathbb{E}_{Y\sim T_1\#\nu}[Y] = 0$, it comes that problem (GW) is equivalent to

$$\sup_{\pi\in\Pi(T_0\#\mu, T_1\#\nu)} \int \|x\|_{\mathbb{R}^m}^2 \|y\|_{\mathbb{R}^n}^2 d\pi(x,y) + 2 \left\| \int xy^T d\pi(x,y) \right\|_{\mathcal{F}}^2. \tag{2.3}$$

Since $T_0\#\mu$ and $T_1\#\nu$ are centered, we have that $\int xy^T d\pi(x,y) = \mathrm{Cov}(X,Y)$ where $X \sim T_0\#\mu$ and $Y \sim T_1\#\nu$. Furthermore, it can be easily computed that

$$\int \|x\|_{\mathbb{R}^m}^2 \|y\|_{\mathbb{R}^n}^2 d\pi(x,y) = \sum_{i,j} \mathrm{Cov}(X_i^2, Y_j^2) + \sum_{i,j} \mathbb{E}[X_i^2]\mathbb{E}[Y_j^2]. \tag{2.4}$$

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

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

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

3 Study of the general problem

Since problem (supCOV) is hard to solve because of its dependence on co-moments of order 2 and 4 of π, one can optimize both terms seperately in order to find a lower bound of $GW_2(\mu,\nu)$. In the rest of the paper we suppose for convenience and without any loss of generality that $n \leq m$.

Theorem 3.1. *Suppose without any loss of generality that $n \leq m$. Let $\mu = \mathcal{N}(m_0, \Sigma_0)$ and $\nu = \mathcal{N}(m_1, \Sigma_1)$ be two Gaussian measures on $\mathbb{R}^m$ and $\mathbb{R}^n$. Let P_0, D_0 and P_1, D_1 be the respective diagonalizations of $\Sigma_0 (= P_0 D_0 P_0^T)$ and $\Sigma_1 (= P_1 D_1 P_1^T)$ which sort eigenvalues in decreasing order. We suppose that Σ_0 is non-singular. A lower bound for $GW_2(\mu,\nu)$ is then*

$$GW_2^2(\mu,\nu) \geq LGW_2^2(\mu,\nu), \tag{3.1}$$

where

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

The proof of this theorem is divided in smaller intermediary results. First we recall the Isserlis lemma (see [14]), which allows to derive the co-moments of order 4 of a Gaussian distribution as a fonction of its co-moments of order 2.

Lemma 3.1 (Isserlis, 1918, [14]). *Let X be a zero-mean Gaussian vector of size n. Then*

$$\forall i,j,k,l \leq n,\ \mathbb{E}[X_iX_jX_kX_l] = \mathbb{E}[X_iX_j]\mathbb{E}[X_kX_l] + \mathbb{E}[X_iX_k]\mathbb{E}[X_jX_l] + \mathbb{E}[X_iX_l]\mathbb{E}[X_jX_k]. \tag{3.2}$$

Then we derive the following general optimization lemmas. The proofs of these two lemmas are postponed to the Appendix (Section 8).

Lemma 3.2. *Suppose that $n \leq m$. Let Σ be a semi-definite positive matrix of size $m+n$ of the form*

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

with $\Sigma_0 \in S_m^{++}(\mathbb{R})$, $\Sigma_1 \in S_n^{+}(\mathbb{R})$ and $K \in \mathbb{R}^{m\times n}$. Let P_0, D_0 and P_1, D_1 be the respective diagonalisations of $\Sigma_0 (= P_0^T D_0 P_0)$ and $\Sigma_1 (= P_1^T D_1 P_1)$ which sort the eigenvalues in decreasing order. Then

$$\max_{\Sigma_1 - K^T\Sigma_0^{-1}K \in S_n^+(\mathbb{R})} \|K\|_{\mathcal{F}}^2 = \mathrm{tr}(D_0^{(n)} D_1), \tag{3.3}$$

and is achieved at any

$$K^* = P_0^T \begin{pmatrix} \tilde{I}_n (D_0^{(n)})^{\frac{1}{2}} D_1^{\frac{1}{2}} \\ 0_{m-n,n} \end{pmatrix} P_1, \tag{opKl2}$$

where $\tilde{I}_n$ is of the form $\mathrm{diag}((\pm 1)_{i\leq n})$.

Lemma 3.3. *Suppose that $n \leq m$. Let Σ be a semi-definite positive matrix of size $m+n$ of the form:*

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

where $\Sigma_0 \in S_m^{++}(\mathbb{R})$, $\Sigma_1 \in S_n^{+}(\mathbb{R})$, and $K \in \mathbb{R}^{m\times n}$. Let $A \in \mathbb{R}^{n\times m}$ be a matrix with rank 1. Then

$$\max_{\Sigma_1 - K^T\Sigma_0^{-1}K \in S_n^+(\mathbb{R})} \operatorname{tr}(KA) = \sqrt{\operatorname{tr}(A\Sigma_0 A^T \Sigma_1)}. \tag{3.4}$$

In particular, if $\Sigma_0 = \operatorname{diag}(\alpha)$ and $\Sigma_1 = \operatorname{diag}(\beta)$ with $\alpha \in \mathbb{R}^m$ and $\beta \in \mathbb{R}^n$, then

$$\max_{\Sigma_1 - K^T\Sigma_0^{-1}K \in S_n^+(\mathbb{R})} \operatorname{tr}(K\mathbb{1}_{n,m}) = \sqrt{\operatorname{tr}(\Sigma_0)\operatorname{tr}(\Sigma_1)}, \tag{3.5}$$

with $\mathbb{1}_{n,m} = (1)_{i\leq n, j\leq m}$, and is achieved at

$$K^* = \frac{\alpha\beta^T}{\sqrt{\operatorname{tr}(\Sigma_0)\operatorname{tr}(\Sigma_1)}}. \tag{opKl1}$$

Proof of theorem 3.1. For $\mu = \mathcal{N}(m_0, \Sigma_0)$ and $\nu = \mathcal{N}(m_1, \Sigma_1)$, we note P_0, D_0 and P_1, D_1 the respective diagonalizations of Σ_0 and Σ_1 which sort the eigenvalues in decreasing order. Let $T_0 : x \in \mathbb{R}^m \mapsto P_0^T(x - m_0)$ and $T_1 : y \in \mathbb{R}^n \mapsto P_1^T(y - m_1)$. For $\pi \in \Pi(T_0\#\mu, T_1\#\nu)$ and $(X, Y) \sim \pi$, we denote Σ the covariance matrix of π and $\tilde{\Sigma}$ the covariance matrix of (X^2, Y^2) with $X^2 := ([XX^T]_{i,i})_{i\leq m}$ and $Y^2 := ([YY^T]_{j,j})_{j\leq n}$. Using Isserlis lemma to compute $\operatorname{Cov}(X^2, X^2)$ and $\operatorname{Cov}(Y^2, Y^2)$, it comes that Σ and $\tilde{\Sigma}$ are of the form:

$$\Sigma = \begin{pmatrix} D_0 & K \\ K^T & D_1 \end{pmatrix} \quad \text{and} \quad \tilde{\Sigma} = \begin{pmatrix} 2D_0^2 & \tilde{K} \\ \tilde{K}^T & 2D_1^2 \end{pmatrix}. \tag{3.6}$$

In order to find a supremum for each term of (supCOV), we use a necessary condition for π to be in $\Pi(T_0\#\mu, T_1\#\nu)$ which is that Σ and $\tilde{\Sigma}$ must be semi-definite positive. To do so, we can use the equivalent condition that the Schur complements of Σ and $\tilde{\Sigma}$, namely $D_1 - K^T D_0^{-1} K$ and $2D_1^2 - \frac{1}{2}\tilde{K}^T D_0^{-2}\tilde{K}$, must also be semi-definite positive. Remarking that the left-hand term in (supCOV) can be rewritten $\operatorname{tr}(\tilde{K}\mathbb{1}_{m,n})$, we have the two following inequalities

$$\sup_{X\sim T_0\#\mu, Y\sim T_1\#\nu} \sum_{i,j} \operatorname{Cov}(X_i^2, Y_j^2) \leq \max_{2D_1^2 - \frac{1}{2}K^T D_0^{-2} K \in S_n^+(\mathbb{R})} \operatorname{tr}(\tilde{K}\mathbb{1}_{n,m}), \tag{3.7}$$

and

$$\sup_{X\sim T_0\#\mu, Y\sim T_1\#\nu} \|\operatorname{Cov}(X, Y)\|_{\mathcal{F}}^2 \leq \max_{D_1 - K^T D_0^{-1} K \in S_n^+(\mathbb{R})} \|K\|_{\mathcal{F}}^2. \tag{3.8}$$

Applying Lemmas 3.2 and 3.3 on both right-hand terms, we get on one hand:

$$\sup_{X\sim T_0\#\mu, Y\sim T_1\#\nu} \|\operatorname{Cov}(X, Y)\|_{\mathcal{F}}^2 \leq \operatorname{tr}(D_0^{(n)} D_1), \tag{3.9}$$

and on the other hand:

$$\begin{aligned} \sup_{X\sim T_0\#\mu, Y\sim T_1\#\nu} \sum_{i,j} \operatorname{Cov}(X_i^2, Y_j^2) &\leq 2\sqrt{\operatorname{tr}(D_0^2)\operatorname{tr}(D_1^2)} \\ &= 2\|D_0\|_{\mathcal{F}}\|D_1\|_{\mathcal{F}}. \end{aligned} \tag{3.10}$$

Furthermore, using Lemma 2.2, it comes that

$$\begin{aligned} GW_2^2(\mu,\nu) &= C_{\mu,\nu} \\ &\quad - 4 \sup_{X\sim T_0\#\mu, Y\sim T_1\#\nu} \left(\sum_{i,j} \mathrm{Cov}(X_i^2, Y_j^2) + \sum_{i,j} \mathbb{E}[X_i^2]\mathbb{E}[X_j^2] + 2\left\|\mathrm{Cov}(X,Y)\right\|_{\mathcal{F}}^2 \right) \\ &\geq C_{\mu,\nu} - 8\sqrt{\mathrm{tr}(D_0^2)\mathrm{tr}(D_1^2)} - 4\mathrm{tr}(D_0)\mathrm{tr}(D_1) - 8\mathrm{tr}(D_0^{(n)}D_1), \end{aligned} \tag{3.11}$$

where

$$\begin{aligned} C_{\mu,\nu} &= \mathbb{E}_{U\sim\mathcal{N}(0,2D_0)}[\|U\|_{\mathbb{R}^m}^4] + \mathbb{E}_{V\sim\mathcal{N}(0,2D_1)}[\|V\|_{\mathbb{R}^n}^4] - 4\mathbb{E}_{X\sim\mu}[\|X\|_{\mathbb{R}^m}^2]\mathbb{E}_{Y\sim\nu}[\|Y\|_{\mathbb{R}^n}^2] \\ &= 8\mathrm{tr}(D_0^2) + 4(\mathrm{tr}(D_0))^2 + 8\mathrm{tr}(D_1^2) + 4(\mathrm{tr}(D_1))^2 - 4\mathrm{tr}(D_0)\mathrm{tr}(D_1). \end{aligned} \tag{3.12}$$

Finally

$$\begin{aligned} GW_2^2(\mu,\nu) &\geq 4(\mathrm{tr}(D_0))^2 + 4(\mathrm{tr}(D_1))^2 - 8\mathrm{tr}(D_0)\mathrm{tr}(D_1) + 8\mathrm{tr}(D_0^2) + 8\mathrm{tr}(D_1^2) \\ &\quad - 8\sqrt{\mathrm{tr}(D_0^2)\mathrm{tr}(D_1^2)} - 8\mathrm{tr}(D_0^{(n)}D_1) \\ &= LGW_2^2(\mu,\nu). \end{aligned} \tag{3.13}$$

□

Inequalities (3.7) and (3.8) become equalities if one can exhibit a plan π such that Σ or $\tilde{\Sigma}$ are such that $\|K\|_{\mathcal{F}}^2$ or $\mathrm{tr}(\tilde{K}\mathbb{1}_{n,m})$ are maximized. This is the case for (3.8) where we can exhibit the Gaussian plan π^* such that K is of the form (opKl2) but it seems however more tricky to exhibit such a plan for inequality (3.7). Indeed, it can be shown that it doesn't exist a Gaussian plan such that $\tilde{K}$ is of the form (opKl1).

The lower bound LGW_2 is reached if it exists a plan π which optimizes both terms simultaneously. This seems rather unlikely because if a probability distribution has its covariance matrix such that K is of the form (opKl2), then it is necessarily Gaussian thanks to the equality case in Cauchy-Schwarz: if $D_0 = \mathrm{diag}(\alpha)$ and $D_1 = \mathrm{diag}(\beta)$ with $\alpha \in \mathbb{R}^m$ and $\beta \in \mathbb{R}^n$, and if π has its covariance matrix such that K is of the form (opKl2), then for all $i \leq n$, $\mathrm{Cov}(X_i, Y_i) = \pm\sqrt{\alpha_i\beta_i}$ and Y_i depends linearly in X_i. As an outcome, π is Gaussian and we can compute, using Isserlis lemma, that $\mathrm{tr}(\tilde{K}\mathbb{1}_{n,m}) = 2\mathrm{tr}(D_0D_1)$ and so $\tilde{K}$ cannot be of the form (opKl1). However, we didn't prove that the solution of the form (opKl2) is unique so it may exist another solution which doesn't imply that π has to be Gaussian.

4 Problem restricted to Gaussian transport plans

In this section, we study the following problem, where we constrain the optimal transport plan to be Gaussian.

$$GGW_2^2(\mu,\nu) = \inf_{\pi\in\Pi(\mu,\nu)\cap\mathcal{N}_{m+n}} \int\int \left(\|x-x'\|_{\mathbb{R}^m}^2 - \|y-y'\|_{\mathbb{R}^n}^2\right)^2 d\pi(x,y)d\pi(x',y'), \tag{GaussGW}$$

where $\mathcal{N}_{m+n}$ is the set of Gaussian measures on $\mathbb{R}^{m+n}$. We show the following main result.

Theorem 4.1. *Suppose without any loss of generality that $n \leq m$. Let $\mu = \mathcal{N}(m_0,\Sigma_0)$ and $\nu = \mathcal{N}(m_1,\Sigma_1)$ be two Gaussian measures on $\mathbb{R}^m$ and $\mathbb{R}^n$. Let P_0, D_0 and P_1, D_1 be the respective diagonalizations of $\Sigma_0(= P_0D_0P_0^T)$ and $\Sigma_1(= P_1D_1P_1^T)$ which sort eigenvalues in decreasing order. We suppose that Σ_0 is non-singular (μ is not degenerate). Then problem* (GaussGW) *admits a solution of the form $\pi^* = (I_m, T)\#\mu$ with T affine of the form*

$$\forall x \in \mathbb{R}^m,\ T(x) = m_1 + P_1AP_0^T(x-m_0). \tag{4.1}$$

where $A \in \mathbb{R}^{n\times m}$ is written

$$A = \begin{pmatrix} \tilde{I}_n D_1^{\frac{1}{2}} (D_0^{(n)})^{-\frac{1}{2}} & 0_{n,m-n} \end{pmatrix},$$

where $\tilde{I}_n$ is of the form $\mathrm{diag}((\pm 1)_{i\leq n})$*. Moreover*

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

Proof. This theorem is a direct consequence of Isserlis lemma 3.1: indeed, the left term in equation (supCOV) can be in that case rewritten $2\|\mathrm{Cov}(X,Y)\|_{\mathcal{F}}^2$ and so problem (GaussGW) is equivalent to

$$\sup_{X\sim T_0\#\mu, Y\sim T_1\#\nu} \|\mathrm{Cov}(X,Y)\|_{\mathcal{F}}^2. \tag{4.2}$$

Applying Lemma 3.2, we can exhibit a Gaussian optimal plan $\pi^* \in \Pi(\mu,\nu)$ with covariance matrix Σ of the form:

$$\Sigma = \begin{pmatrix} \Sigma_0 & K^* \\ K^{*T} & \Sigma_1 \end{pmatrix}, \tag{4.3}$$

with

$$K^* = P_0^T \begin{pmatrix} \tilde{I}_n (D_0^{(n)})^{\frac{1}{2}} D_1^{\frac{1}{2}} \\ 0_{m-n,n} \end{pmatrix} P_1. \tag{4.4}$$

Thus, using the equality case in Cauchy-Schwarz, we can exhibit an optimal transport map T of the form

$$\forall x \in \mathbb{R}^m,\ T(x) = m_1 + P_1 A P_0^T (x - m_0), \tag{4.5}$$

with

$$A = \begin{pmatrix} \tilde{I}_n D_1^{\frac{1}{2}} (D_0^{(n)})^{-\frac{1}{2}} & 0_{n,m-n} \end{pmatrix},$$

where $\tilde{I}_n$ is of the form $\mathrm{diag}((\pm 1)_{i\leq n})$. Moreover, using Lemmas 2.2 and 3.2, it comes that

$$\begin{aligned} GGW_2^2(\mu,\nu) &= C_{\mu,\nu} - 16 \sup_{X\sim T_0\#\mu, Y\sim T_1\#\nu} \|\mathrm{Cov}(X,Y)\|_{\mathcal{F}}^2 \\ &= 8\mathrm{tr}(D_0^2) + 4(\mathrm{tr}(D_0))^2 + 8\mathrm{tr}(D_1^2) + 4(\mathrm{tr}(D_1))^2 - 4\mathrm{tr}(D_0)\mathrm{tr}(D_1) - 16\mathrm{tr}(D_0^{(n)} D_1) \quad (4.6) \\ &= 4(\mathrm{tr}(D_0) - \mathrm{tr}(D_1))^2 + 8\mathrm{tr}\left((D_0^{(n)} - D_1)^2\right) + 8\left(\mathrm{tr}(D_0^2) - \mathrm{tr}((D_0^{(n)})^2)\right). \end{aligned}$$

□

Link with Gromov-Monge The previous result generalizes Theorem 4.2.6 in [25], which studies the solutions of the linear Gromov-Monge problem between Gaussian distributions

$$\inf_{T\#\mu=\nu, \text{T linear}} \int\int \left(\|x - x'\|_{\mathbb{R}^m}^2 - \|T(x) - T(x')\|_{\mathbb{R}^n}^2\right)^2 d\mu(x) d\nu(x'). \tag{4.7}$$

Indeed, solutions of (4.7) necessarily provide Gaussian transport plans $\pi = (I_m, T)\#\mu$ if T is linear. Conversely, Theorem 4.1 shows that restricting the optimal plan to be Gaussian in Gromov-Wasserstein between two Gaussian distributions yields an optimal plan of the form $\pi = (I_m, T)\#\mu$ with a linear T, whatever the dimensions m and n of the two Euclidean spaces.

Link with Principal Component Analysis We can easily draw connections between GGW_2^2 and PCA. Indeed, we can remark that the optimal plan can be derived by performing PCA on both distributions μ and ν in order to obtain distributions $\tilde{\mu}$ and $\tilde{\nu}$ with zero mean vectors and diagonal covariance matrices with eigenvalues in decreasing order ($\tilde{\mu} = T_0\#\mu$ and $\tilde{\nu} = T_1\#\nu$), then by keeping only the n first components in $\tilde{\mu}$ and finally by deriving the optimal transport plan which achieves W_2^2 between the obtained truncated distribution and $\tilde{\nu}$. In other terms, noting $P_n : \mathbb{R}^m \to \mathbb{R}^n$ the linear mapping which, for $x \in \mathbb{R}^m$ keeps only its n first components (n-frame), T_{W_2} the optimal transport map such that $\pi_{W_2} = (I_n, T_{W_2})\#P_n\#\tilde{\mu}$ achieves $W_2(P_n\#\tilde{\mu}, \tilde{\nu})$, it comes that the optimal plan π_{GGW_2} which achieves $GGW_2(\tilde{\mu}, \tilde{\nu})$ can be written

$$\pi_{GGW_2} = (I_m, \tilde{I}_n\#T_{W_2}\#P_n)\#\tilde{\mu}. \tag{4.8}$$

An example of π_{GGW_2} can be found in Figure 1 when $m = 2$ and $n = 1$.

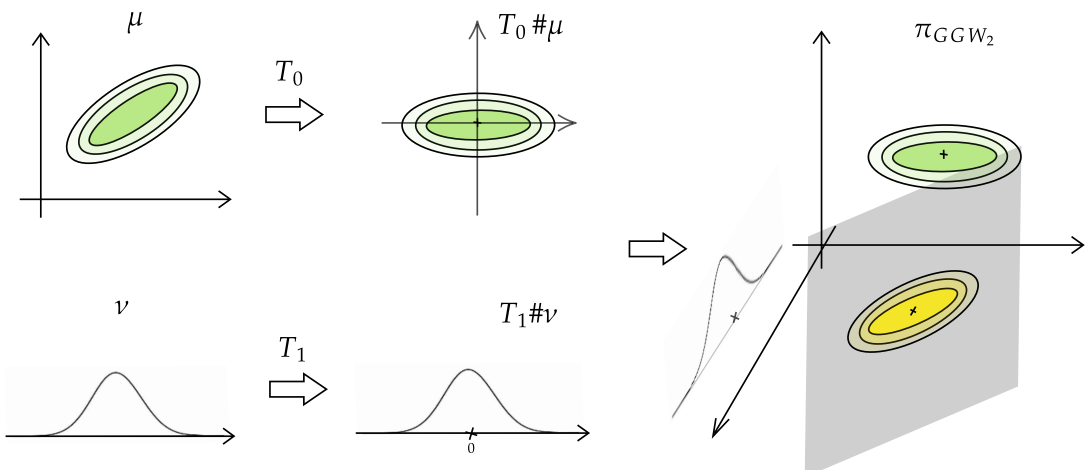

Figure 1: transport plan π_{GGW_2} solution of problem (GaussGW) with $m = 2$ and $n = 1$. In that case, π_{GGW_2} is the degenerate Gaussian distribution supported by the plan of equation $y = T_{W_2}(x)$, where T_{W_2} is the classic W_2 optimal transport map when the distributions are rotated and centered first.

Case of equal dimensions When $m = n$, the optimal plan π_{GGW_2} which achieves $GGW_2(\mu, \nu)$ is closely related to the optimal transport plan $\pi_{W_2} = (I_m, T_{W_2})\#T_0\#\mu$. Indeed, π_{GGW_2} can be simply derived by applying the transformations T_0 and T_1 to respectively μ and ν, then by computing π_{W_2} between $T_0\#\mu$ and $T_1\#\nu$, and finally by applying the inverse transformations T_0^{-1} and T_1^{-1}. In other terms, π_{GGW_2} can be written

$$\pi_{GGW_2} = (I_m, T_1^{-1}\#\tilde{I}_n\#T_{W_2}\#T_0)\#\mu. \tag{4.9}$$

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

As illustrated in Figure 3, the GGW_2 optimal transport map T_{GGW_2} defined in Equation (4.1) is not equivalent to the W_2 optimal transport map T_{W_2} defined in (1.5) even when the dimensions m and n are equal. More precisely, it Σ_0 and Σ_1 can be diagonalized in the same orthonormal basis with eigenvalues in the same order (decreasing or increasing), then T_{W_2} and T_{GGW_2} are equivalent (top of Figure 3). On the other hand, if Σ_0 and Σ_1 can be diagonalized in the same orthonormal basis but with eigenvalues not in the same order, T_{W_2} and T_{GGW_2} will have very different behaviors (bottom of Figure 3). Between those two extreme cases, we can say that the closer the columns of P_0 will be collinear to the columns of P_1 (with the eigenvalues in decreasing order), the more T_{W_2} and T_{GGW_2} will tend to have similar behaviors (middle of Figure 3).

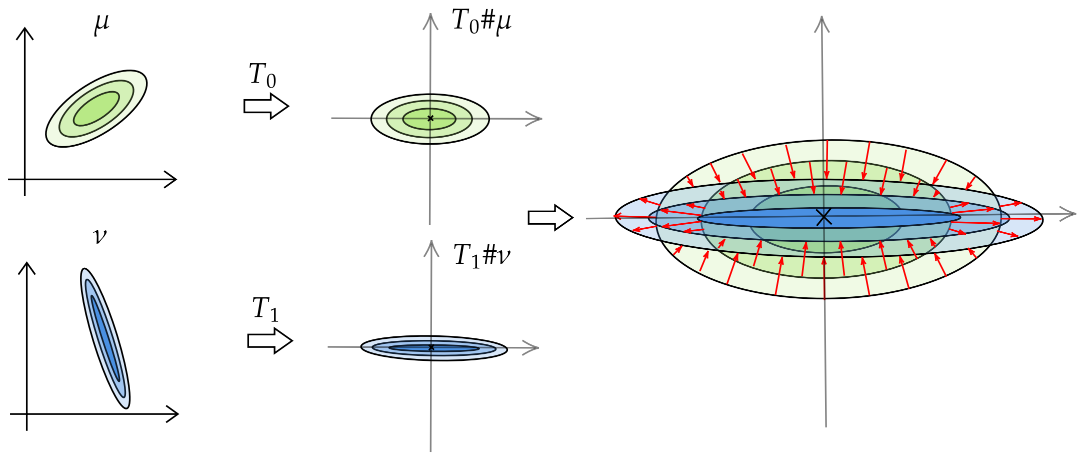

Figure 2: Solution of (GaussGW) between two Gaussians measures in dimension 2. First the distributions are centered and rotated. Then a classic W_2 transport is applied between the two aligned distributions.

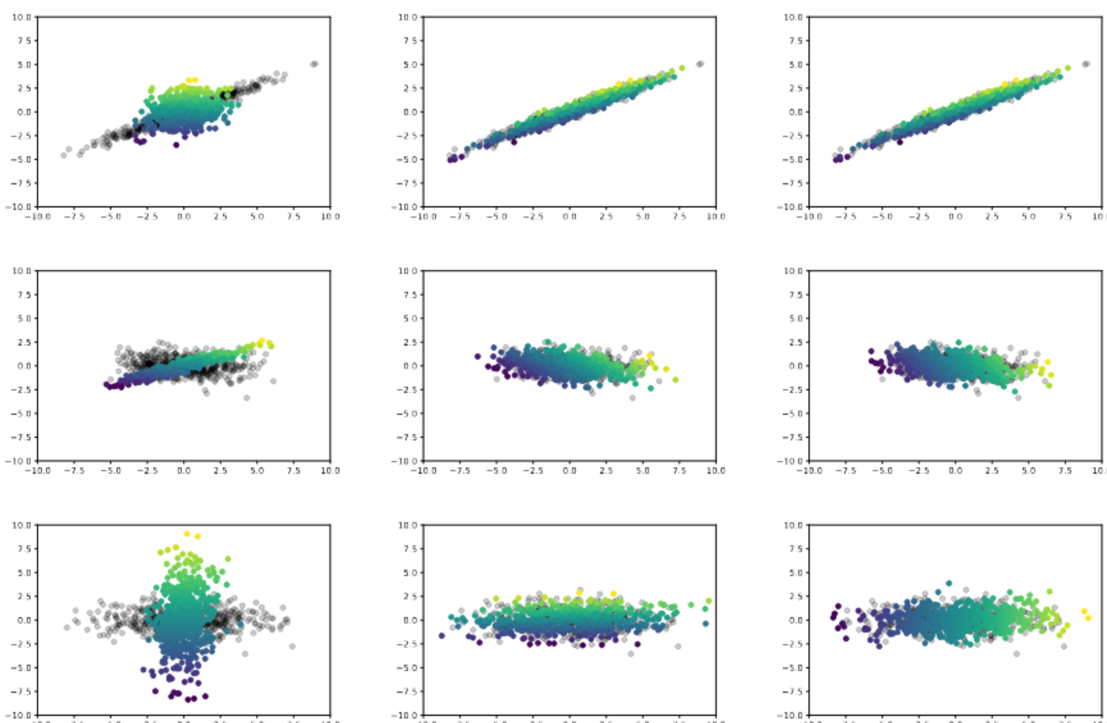

Figure 3: Comparison between W_2 and GGW_2 mappings between empirical distributions. Left: 2D source distribution (colored) and target distribution (transparent). Middle: resulting mapping of Wasserstein T_{W_2}. Right: resulting mapping of Gaussian Gromov-Wasserstein T_{GGW_2}. The colors are added in order to visualize where each sample has been sent.

Link with Gromov-Wasserstein with inner product as cost function If μ and ν are centered Gaussian measures, let us consider the following problem

$$GW_2^2(\langle.,.\rangle_m, \langle.,.\rangle_n, \mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \int\int (\langle x, x'\rangle_m - \langle y, y'\rangle_n)^2 d\pi(x,y) d\pi(x',y'). \tag{innerGW}$$

Notice that the above problem is not restricted to Gaussian plans, but the following proposition shows that in fact its solution is Gaussian.

Proposition 4.1. *Suppose $m \leq n$. Let $\mu = \mathcal{N}(0, \Sigma_0)$ and $\nu = \mathcal{N}(0, \Sigma_1)$ be two centered Gaussian measures respectively on $\mathbb{R}^m$ and $\mathbb{R}^n$. Then the solution of problem* (GaussGW) *exhibited in theorem 4.1 is also solution of problem* (innerGW)*.*

Proof. The proof of this proposition is a direct consequence of lemma 2.2: indeed, applying it with $a = 0$, $b = 0$, and $c = 1$, it comes that problem (innerGW) is equivalent to

$$\sup_{\pi \in \Pi(\mu,\nu)} \left\| \int xy^T d\pi(x,y) \right\|_{\mathcal{F}}^2. \tag{4.10}$$

Since μ and ν are centered, it comes that problem (innerGW) is equivalent to

$$\sup_{X \sim \mu, Y \sim \nu} \|\mathrm{Cov}(X,Y)\|_{\mathcal{F}}^2. \tag{4.11}$$

Applying Lemma 3.2, it comes that the solution exhibited in Theorem 4.1 is also solution of problem (innerGW). □

Since GGW_2 is the Gromov-Wasserstein problem restricted to Gaussian transport plan, it is clear that (GaussGW) is an upper bound of (GW). Combining this result with Theorem 3.1, we get the following simple but important result.

Proposition 4.2. *If $\mu = \mathcal{N}(m_0, \Sigma_0)$ and $\nu = \mathcal{N}(m_1, \Sigma_1)$ and Σ_0 is non-singular, then*

$$LGW_2^2(\mu,\nu) \leq GW_2^2(\mu,\nu) \leq GGW_2^2(\mu,\nu). \tag{4.12}$$

5 Tightness of the bounds and particular cases

5.1 Bound on the difference

Proposition 5.1. *Suppose without loss of generality that $n \leq m$, if $\mu = \mathcal{N}(m_0, \Sigma_0)$ and $\nu = \mathcal{N}(m_1, \Sigma_1)$, then*

$$GGW_2^2(\mu,\nu) - LGW_2^2(\mu,\nu) \leq 8\|\Sigma_0\|_{\mathcal{F}}\|\Sigma_1\|_{\mathcal{F}} \left(1 - \frac{1}{\sqrt{m}}\right). \tag{5.1}$$

To prove this proposition, we will use the following technical result (the proof is postponed to the Appendix (Section 8)):

Lemma 5.1. *Let $u \in \mathbb{R}^m$ and $v \in \mathbb{R}^m$ be two unit vectors with non-negative coordinates ordered in decreasing order. Then*

$$u^T v \geq \frac{1}{\sqrt{m}}, \tag{5.2}$$

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

Proof of Proposition 5.1. By subtracting (LGW) from (GGW), it comes that

$$\begin{aligned} GGW_2^2(\mu,\nu) - LGW_2^2(\mu,\nu) &= 8\left(\|D_0\|_{\mathcal{F}}\|D_1\|_{\mathcal{F}} - \mathrm{tr}(D_0^{(n)}D_1)\right) \\ &= 8\left(\|D_0\|_{\mathcal{F}}\|D_1^{[m]}\|_{\mathcal{F}} - \mathrm{tr}(D_0D_1^{[m]})\right), \end{aligned} \tag{5.3}$$

where $D_1^{[m]} = \begin{pmatrix} D_1 & 0 \\ 0 & 0 \end{pmatrix} \in \mathbb{R}^{m\times m}$. Noting $\alpha \in \mathbb{R}^m$ and $\beta \in \mathbb{R}^m$ the vectors of eigenvalues of D_0 and $D_1^{[m]}$, it comes

$$GGW_2^2(\mu,\nu) - LGW_2^2(\mu,\nu) = 8(\|\alpha\|\|\beta\| - \alpha^T\beta) = 8\|\alpha\|\|\beta\|(1-u^Tv), \tag{5.4}$$

where $u = \frac{\alpha}{\|\alpha\|}$ and $v = \frac{\beta}{\|\beta\|}$. Applying lemma 5.1, we get directly that

$$\begin{aligned} GGW_2^2(\mu,\nu) - LGW_2^2(\mu,\nu) &\leq 8\|D_0\|_{\mathcal{F}}\|D_1^{[m]}\|_{\mathcal{F}}\left(1-\frac{1}{\sqrt{m}}\right). \\ &= 8\|\Sigma_0\|_{\mathcal{F}}\|\Sigma_1\|_{\mathcal{F}}\left(1-\frac{1}{\sqrt{m}}\right). \end{aligned} \tag{5.5}$$

□

The difference between $GGW_2^2(\mu,\nu)$ and $LGW_2^2(\mu,\nu)$ can be seen as the difference between the right and left terms of the Cauchy-Schwarz inequality applied to the two vectors of eigenvalues $\alpha \in \mathbb{R}^m$ and $\beta \in \mathbb{R}^m$. The difference is maximized when the vectors α and β are the least collinear possible. This happens when the eigenvalues of D_0 are all equal and $n = 1$ or ν is degenerate of true dimension 1. On the other hand , this difference is null when α and β are collinear. Between those two extremal cases, we can say that the difference between $GGW_2^2(\mu,\nu)$ and $LGW_2^2(\mu,\nu)$ will be relatively small if the last $m-n$ eigenvalues D_0 are small compared to the n first eigenvalues and if the n first eigenvalues are close to be proportional to the eigenvalues of D_1. An example in the case where $m = 2$ and $n = 1$ can be found in Figure 4.

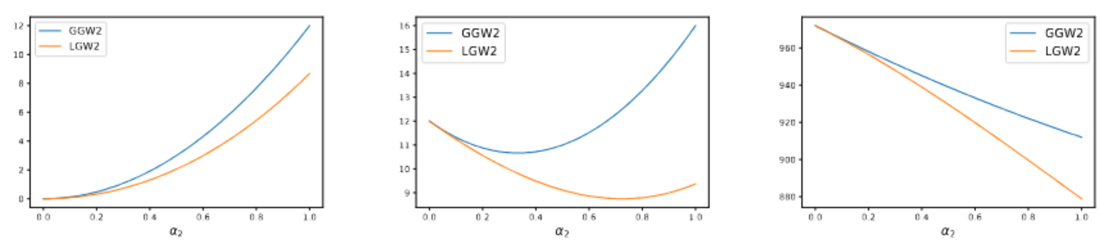

Figure 4: plot of $GGW_2^2(\mu,\nu)$ and $LGW_2^2(\mu,\nu)$ in function of α_2 for $\mu = \mathcal{N}(0,\mathrm{diag}(\alpha))$, $\nu = \mathcal{N}(0,\beta_1)$, $\alpha = (\alpha_1,\alpha_2)^T$, for $(\alpha_1,\beta_1) = (1,1)$ (left), $(\alpha_1,\beta_1) = (1,2)$ (middle), $(\alpha_1,\beta_1) = (1,10)$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

5.2 Explicit case

As seen before, the difference between $GGW_2^2(\mu,\nu)$ and $LGW_2^2(\mu,\nu)$, with $\mu = \mathcal{N}(m_0,\Sigma_0)$ and $\nu = \mathcal{N}(m_1,\Sigma_1)$, is null when the two vectors of eigenvalues of Σ_0 and Σ_1 (sorted in decreasing order) are collinear. When we suppose Σ_0 non-singular, it implies that $m = n$ and that the eigenvalues of Σ_1 are proportional to the eigenvalues of Σ_0 (rescaling). This case includes the more particular case where $m = n = 1$. In that case $\mu = \mathcal{N}(m_0,\sigma_0^2)$ and $\nu = \mathcal{N}(m_1,\sigma_1^2)$, because σ_1 is always proportional to σ_0.

Proposition 5.2. *Suppose $m = n$. Let $\mu = \mathcal{N}(m_0, \Sigma_0)$ and $\nu = \mathcal{N}(m_1, \Sigma_1)$ two Gaussian measures on $\mathbb{R}^m = \mathbb{R}^n$. Let P_0, D_0 and P_1, D_1 be the respective diagonalizations of $\Sigma_0(= P_0 D_0 P_0^T)$ and $\Sigma_1(= P_1 D_1 P_1^T)$ which sort eigenvalues in non-increasing order. Suppose Σ_0 is non-singular and that it exists a scalar $\lambda \geq 0$ such that $D_1 = \lambda D_0$. In that case, $GW_2^2(\mu,\nu) = GGW_2^2(\mu,\nu) = LGW_2^2(\mu,\nu)$ and the problem admits a solution of the form $(I_m, T)\#\mu$ with T affine of the form:*

$$\forall x \in \mathbb{R}^m,\ T(x) = m_1 + \sqrt{\lambda} P_1 \tilde{I}_m P_0^T (x - m_0), \tag{5.6}$$

where $\tilde{I}_m$ of the form $\mathrm{diag}((\pm 1)_{i \leq m})$. *Moreover*

$$GW_2^2(\mu,\nu) = (\lambda - 1)^2 \left(4(\mathrm{tr}(\Sigma_0))^2 + 8\|\Sigma_0\|_{\mathcal{F}}^2\right). \tag{5.7}$$

Proof. From (5.3), we have

$$GGW_2^2(\mu,\nu) - LGW2^2(\mu,\nu) = 8\left(\|D_0\|_{\mathcal{F}}\|D_1\|_{\mathcal{F}} - \mathrm{tr}(D_0 D_1)\right). \tag{5.8}$$

Noting $\alpha \in \mathbb{R}^m$ and $\beta \in \mathbb{R}^m$ the eigenvalues vectors of D_0 and D_1, it comes

$$GGW_2^2(\mu,\nu) - LGW2^2(\mu,\nu) = 8(\|\alpha\|\|\beta\| - \alpha^T\beta). \tag{5.9}$$

Since it exists $\lambda \geq 0$ such that $D_1 = \lambda D_0$, we have $\beta = \lambda\alpha$, and so $\alpha^T\beta = \|\alpha\|\|\beta\|$. Thus $GGW_2^2(\mu,\nu) - LGW_2^2(\mu,\nu) = 0$ and using Proposition 4.2, we get that $GW_2^2(\mu,\nu) = GGW_2^2(\mu,\nu) = LGW_2^2(\mu,\nu)$. We get (5.6) and (5.7) by simply reinjecting in (4.1) and (GGW). □

Corollary 5.1. *Let $\mu = \mathcal{N}(m_0, \sigma_0^2)$ and $\nu = \mathcal{N}(m_1, \sigma_1^2)$ be two Gaussian measures on $\mathbb{R}$. Then*

$$GW_2^2(\mu,\nu) = 12\left(\sigma_0^2 - \sigma_1^2\right)^2, \tag{5.10}$$

and the optimal transport plan π^* has the form $(I_1, T)\#\mu$ with T affine of the form:

$$\forall x \in \mathbb{R},\ T(x) = m_1 \pm \frac{\sigma_1}{\sigma_0}(x - m_0). \tag{5.11}$$

Thus, the solution of $W_2^2(\mu,\nu)$ is also solution of $GW_2^2(\mu,\nu)$.

5.3 Case of degenerate measures

In all the results exposed above, we have supposed Σ_0 non-singular, which means that μ is not degenerate. Yet, if Σ_0 is not full rank, one can easily extend the previous results thanks to the following proposition.

Proposition 5.3. *Let $\mu = \mathcal{N}(0, D_0)$ and $\nu = \mathcal{N}(0, D_1)$ be two centered Gaussian measures on $\mathbb{R}^m$ and $\mathbb{R}^n$ with diagonal covariance matrices D_0 and D_1 with eigenvalues in decreasing order. We denote $r = rk(D_0)$ the rank of D_0 and we suppose that $r < m$. Let us define $P_r = \begin{pmatrix} I_r & 0_{r,m-r}\end{pmatrix} \in \mathbb{R}^{r\times m}$. Then $GW_2^2(\mu,\nu) = GW_2^2(P_r\#\mu,\nu)$, $GGW_2^2(\mu,\nu) = GGW_2^2(P_r\#\mu,\nu)$, and $LGW_2^2(\mu,\nu) = LGW_2^2(P_r\#\mu,\nu)$.*

Proof. For $r < m$, we denote $\Gamma_r(\mathbb{R}^m)$ the set of vectors $x = (x_1,\ldots,x_m)^T$ of $\mathbb{R}^m$ such that $x_{r+1} = \cdots = x_m = 0$. For $\pi \in \Pi(\mu,\nu)$, one can remark that for any borel set $A \subset \mathbb{R}^m \smallsetminus \Gamma_r(\mathbb{R}^m)$, and any borel set $B \subset \mathbb{R}^n$, we have $\pi(A,B) = 0$ and so

$$\begin{aligned}
GW_2^2(\mu,\nu) &= \inf_{\pi\in\Pi(\mu,\nu)} \int_{\mathbb{R}^m\times\mathbb{R}^n}\int_{\mathbb{R}^m\times\mathbb{R}^n} (\|x - x'\|_{\mathbb{R}^m}^2 - \|y - y'\|_{\mathbb{R}^m}^2)^2 d\pi(x,y)d\pi(x',y') \\
&= \inf_{\pi\in\Pi(\mu,\nu)} \int_{\Gamma_r(\mathbb{R}^m)\times\mathbb{R}^n}\int_{\Gamma_r(\mathbb{R}^m)\times\mathbb{R}^n} (\|x - x'\|_{\mathbb{R}^m}^2 - \|y - y'\|_{\mathbb{R}^m}^2)^2 d\pi(x,y)d\pi(x',y') \\
&= \inf_{\pi\in\Pi(\mu,\nu)} \int_{\Gamma_r(\mathbb{R}^m)\times\mathbb{R}^n}\int_{\Gamma_r(\mathbb{R}^m)\times\mathbb{R}^n} (\|P_r(x - x')\|_{\mathbb{R}^r}^2 - \|y - y'\|_{\mathbb{R}^m}^2)^2 d\pi(x,y)d\pi(x',y')
\end{aligned} \tag{5.12}$$

Now, observe that for $\pi \in \Pi(\mu,\nu)$, $(P_r, I_n)\#\pi \in \Pi(P_r\#\mu,\nu)$. It follows that

$$\begin{aligned} GW_2^2(\mu,\nu) &\leq \inf_{\pi\in\Pi(P_r\#\mu,\nu)} \int_{\mathbb{R}^r\times\mathbb{R}^n}\int_{\mathbb{R}^r\times\mathbb{R}^n} (\|x-x'\|^2_{\mathbb{R}^r} - \|y-y'\|^2_{\mathbb{R}^m})^2 d\pi(x,y)d\pi(x',y') \\ &= GW_2^2(P_r\#\mu,\nu). \end{aligned} \tag{5.13}$$

Conversely, since μ has no mass outside of $\Gamma_r(\mathbb{R}^m)$, $P_r^T\#P_r\#\mu = \mu$, which implies that for $\pi \in \Pi(P_r\#\mu,\nu)$, $(P_r^T, I_n)\#\pi \in \Pi(\mu,\nu)$. It follows that

$$\begin{aligned} GW_2^2(P_r\#\mu,\nu) &= \inf_{\pi\in\Pi(P_r\#\mu,\nu)} \int_{\mathbb{R}^r\times\mathbb{R}^n}\int_{\mathbb{R}^r\times\mathbb{R}^n} (\|x-x'\|^2_{\mathbb{R}^r} - \|y-y'\|^2_{\mathbb{R}^m})^2 d\pi(x,y)d\pi(x',y') \\ &= \inf_{\pi\in\Pi(P_r\#\mu,\nu)} \int_{\mathbb{R}^r\times\mathbb{R}^n}\int_{\mathbb{R}^r\times\mathbb{R}^n} (\|P_r^T(x-x')\|^2_{\mathbb{R}^r} - \|y-y'\|^2_{\mathbb{R}^m})^2 d\pi(x,y)d\pi(x',y') \\ &\leq \inf_{\pi\in\Pi(\mu,\nu)} \int_{\mathbb{R}^m\times\mathbb{R}^n}\int_{\mathbb{R}^m\times\mathbb{R}^n} (\|x-x'\|^2_{\mathbb{R}^r} - \|y-y'\|^2_{\mathbb{R}^m})^2 d\pi(x,y)d\pi(x',y') \\ &\leq GW_2^2(\mu,\nu). \end{aligned} \tag{5.14}$$

The exact same reasoning can be made in the case of GGW_2. Morover, it can be easily seen when looking at (LGW) that $LGW_2^2(\mu,\nu) = LGW_2^2(P_r\#\mu,\nu)$. □

Thus, when Σ_0 is not full rank, one can apply Proposition 5.3 and consider directly the Gromov-Wasserstein distance between the projected (non-degenerate) measure $P_r\#\mu$ on $\mathbb{R}^r$ and ν and so Proposition 4.2 still holds when μ is degenerate.

In the case of GGW_2, an explicit optimal transport plan can still be exhibited. In the following, we denote r_0 and r_1 the ranks of Σ_0 and Σ_1, and we suppose without loss of generality that $r_0 \geq r_1$, but this time not necessarily that $m \geq n$. If $\mu = \mathcal{N}(m_0,\Sigma_0)$ and $\nu = \mathcal{N}(m_1,\Sigma_1)$ are two Gaussian measures on $\mathbb{R}^m$ and $\mathbb{R}^n$, and (P_0,D_0) and (P_1,D_1) are the respective diagonalizations of $\Sigma_0(= P_0D_0P_0^T)$ and $\Sigma_1(= P_1D_1P_1^T)$ which sort the eigenvalues in decreasing order, an optimal transport plan which achieves $GGW_2(\mu,\nu)$ is of the form $\pi^* = (I_m, T)\#\mu$ with

$$\forall x \in \mathbb{R}^m, T(x) = m_1 + P_1AP_0^T(x-m_0), \tag{5.15}$$

where $A \in \mathbb{R}^{n\times m}$ is of the form

$$A = \begin{pmatrix} \tilde{I}_{r_1}(D_1^{(r_1)})^{\frac{1}{2}}(D_0^{(r_1)})^{-\frac{1}{2}} & 0_{r_1,m-r_1} \\ 0_{n-r_1,r_1} & 0_{n-r_1,m-r_1} \end{pmatrix},$$

where $\tilde{I}_{r_1}$ is any matrix of the form $\text{diag}((\pm 1)_{i\leq r_1})$.

6 Behavior of the empirical solution

To complete the previous study, we perform a simple experiment to illustrate the behavior of the solution of the Gromov Wasserstein problem. In this experiment, we draw independently k samples $(X_j)_{j\leq k}$ and $(Y_i)_{i\leq k}$ from respectively $\mu = \mathcal{N}(0,\text{diag}(\alpha))$ and $\nu = \mathcal{N}(0,\text{diag}(\beta))$ with $\alpha \in \mathbb{R}^m$ and $\beta \in \mathbb{R}^n$. Then we compute the Gromov-Wasserstein distance between the two histograms X and Y with the algorithm proposed in [19] using the Python Optimal Transport library [3]. In Figure 5, we plot the first coordinates of the samples Y_i in fonction of the the first coordinate of the samples X_j they have been assigned to by the algorithm (blue dots). We draw also the line of equation $y = \pm\sqrt{\beta}x$ to compare with the theorical solution of the Gaussian restricted problem (orange line) for $k = 2000$,

[3]The library is accessible here: https://pythonot.github.io/index.html

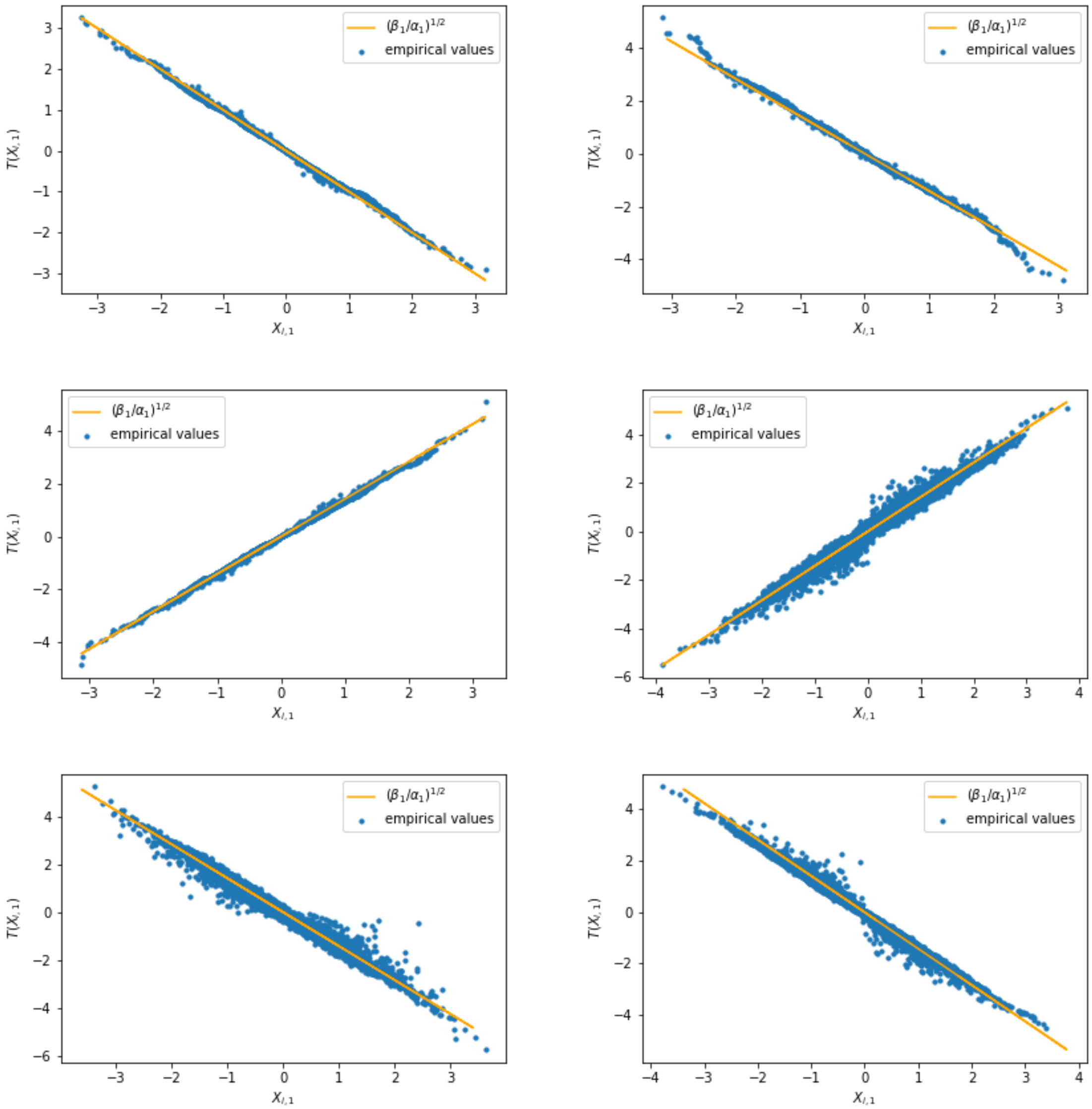

Figure 5: plot of the first coordinate of samples Y_i in fonction of the the first coordinate of their assigned samples X_j (blue dots) and line of equation $y = \pm\sqrt{\beta}x$ (orange line) for $k = 2000$, $\alpha = (1, 0.1)^T$ and $\beta = 2$ (top left), $k = 2000$, $\alpha = (1, 0.1)^T$ and $\beta = (2, 0.3)^T$ (top right), $k = 2000$, $\alpha = (1, 0.1, 0.01)^T$ and $\beta = 2$ (middle left), $k = 7000$, $\alpha = (1, 0.3)$ and $\beta = 2$ (middle right), $k = 7000$, $\alpha = (1, 0.1)^T$, and $\beta = (2, 1)^T$ (bottom left), and $k = 7000$ and $\alpha = (1, 0.3, 0.1)$ and $\beta = 2$ (bottom right).

$\alpha = (1, 0.1)^T$ and $\beta = 2$ (top left), $k = 2000$, $\alpha = (1, 0.1)^T$ and $\beta = (2, 0.3)^T$ (top right), $k = 2000$, $\alpha = (1, 0.1, 0.01)^T$ and $\beta = 2$ (middle left), $k = 7000$, $\alpha = (1, 0.3)$ and $\beta = 2$ (middle right), $k = 7000$, $\alpha = (1, 0.1)^T$, and $\beta = (2, 1)^T$ (bottom left), and $k = 7000$ and $\alpha = (1, 0.3, 0.1)$ and $\beta = 2$ (bottom right). Observe that the empirical solution seems to be behaving exactly in the same way as the theoretical solution exhibited in theorem 4.1 as soon as α and β are close to be collinear. However, when α and β are further away from collinearity, determining the behavior of the empirical solution becomes more complex. Solving Gromov-Wasserstein numerically, even approximately, is a particularly hard task, therefore we cannot conclude if the empirical solution does not behave in the same way as the theorical solution exhibited in theorem 4.1 or if the algorithm has not converged in these more complex cases. This second assumption seems to be more likely because it seems that increasing the number of points k reduces the gap between the blue dots and the orange line. Thus, we conjecture that the optimal plan which achieves $GGW_2(\mu,\nu)$ is also solution of the non-restricted problem $GW_2(\mu,\nu)$ and that $GW_2(\mu,\nu) = GGW_2(\mu,\nu)$.

7 Conclusion

In this paper, we have exhibited lower and upper bounds for the Gromov-Wasserstein distance (with a squared ground distance) between Gaussian measures living on different Euclidean spaces. We have also studied the tightness of the provided bounds, both theoretically and numerically. The upper bound is obtained through the study of the problem with the additional restriction that the optimal plan itself is Gaussian. We have shown that this particular case has a very simple closed-form solution, which can be described as first performing PCA on both distributions and then deriving the optimal linear plan between these aligned distributions. We conjecture that the linear solution exhibited when adding this restriction might also be the solution in more general cases.

8 Appendix: proof of the lemmas

8.1 Proof of Lemma 3.2

Proof. The proof is inspired from the proof of Equation (1.4) provided in [13]. We want to maximize $\mathrm{tr}(K^TK)$ with the constraint that Σ is semi-definite positive. Let $S = \Sigma_1 - K^T\Sigma_0^{-1}K$ (Schur complement). Problem (3.3) can be written in the following way

$$\min_{S\in S_n^+(\mathbb{R})} -\mathrm{tr}(K^TK). \tag{8.1}$$

For a given S, the set of feasible K is the set of K such that $K^T\Sigma_0^{-1}K = \Sigma_1 - S$. Since $\Sigma_0 \in S_m^{++}(\mathbb{R})$, $K^T\Sigma_0^{-1}K \in S_n^+(\mathbb{R})$ and so $\Sigma_1 - S \in S_n^+(\mathbb{R})$. We note r the rank of $K^T\Sigma_0^{-1}K$. One can observe that

$$r \leq n \leq m,$$

where the left-hand side inequality comes from the fact that $\mathrm{rk}(AB) \leq \min\{\mathrm{rk}(A), \mathrm{rk}(B)\}$. Then, $\Sigma_1 - S$ can be diagonalized

$$\Sigma_1 - S = K^T\Sigma_0^{-1}K = U\Lambda^2U^T = U_r\Lambda_r^2U_r^T, \tag{8.2}$$

with $\Lambda^2 = \mathrm{diag}(\lambda_1^2, ..., \lambda_r^2, 0, ..., 0)$, $\Lambda_r^2 = \mathrm{diag}(\lambda_1^2, ..., \lambda_r^2)$, and $U_r \in \mathbb{V}_r(\mathbb{R}^n) := \{M \in \mathbb{R}^{n\times r} | M^TM = I_r\}$ (Stiefel Manifold [15]) such that $U = \begin{pmatrix} U_r & U_{n-r}\end{pmatrix}$. From (8.2), we can deduce that

$$(\Sigma_0^{-\frac{1}{2}}KU_r\Lambda_r^{-1})^T\Sigma_0^{-\frac{1}{2}}KU_r\Lambda_r^{-1} = I_r. \tag{8.3}$$

We can set $B_r = \Sigma_0^{-\frac{1}{2}}KU_r\Lambda_r^{-1}$ such that $B_r \in \mathbb{V}_r(\mathbb{R}^m)$. One can deduce that

$$KU_r = \Sigma_0^{\frac{1}{2}}B_r\Lambda_r.$$

Moreover, since $U_{m-r}^TK^T\Sigma_0^{-1}KU_{m-r} = 0$ and $\Sigma_0 \in S_m^{++}(\mathbb{R})$, it comes that $KU_{n-r} = 0$ and so

$$K = KUU^T = KU_rU_r^T = \Sigma_0^{\frac{1}{2}}B_r\Lambda_rU_r^T. \tag{8.4}$$

We can write $\text{tr}(K^TK)$ as a function of B_r :

$$\begin{aligned}\text{tr}(K^TK) &= \text{tr}(U_r\Lambda_rB_r^T\Sigma_0B_r\Lambda_rU_r^T)\\ &= \text{tr}(U_r^TU_r\Lambda_rB_r^T\Sigma_0B_r\Lambda_r)\\ &= \text{tr}(\Lambda_r^2B_r^T\Sigma_0B_r),\end{aligned} \tag{8.5}$$

Thus, for a given S, the set of K such that $K^T\Sigma_0^{-1}K = \Sigma_1 - S$ is parametrized by the r-frame B_r. We want to find B_r which maximizes $\text{tr}(K^TK)$ for a given S. This problem can be rewritten:

$$\min_{B_r\in\mathbb{V}_r(\mathbb{R}^m)} -\text{tr}(\Lambda_r^2B_r^T\Sigma_0B_r). \tag{8.6}$$

The following is a readaptation of the proof of the Proposition (3.1) in [2] when B_r is not a squared matrix. The Lagrangian of problem (8.6) can be written

$$\mathcal{L}(B_r, C) = -\text{tr}(\Lambda_r^2B_r^T\Sigma_0B_r) + \text{tr}(C(B_r^TB_r - I_r)),$$

where $C \in S_r(\mathbb{R})$ is the Lagrange multiplier associated to the constraint $B_r^TB_r = I_r$ (C is symmetric because $B_r^TB_r - I_r$ is symmetric). We can then derive the first-order condition

$$-2\Sigma_0B_r\Lambda_r^2 + 2B_rC = 0,$$

or equivalently

$$\Sigma_0B_r\Lambda_r^2B_r^T = B_rCB_r^T. \tag{8.7}$$

Since $C \in S_r(\mathbb{R})$, $B_rCB_r^T \in S_m(\mathbb{R})$ and $\Sigma_0B_r\Lambda_r^2B_r^T \in S_m(\mathbb{R})$. We can deduce that Σ_0 and $B_r\Lambda_rB_r^T$ commute. Moreover, since Σ_0 and $B_r\Lambda_r^2B_r^T$ are both symmetric, they can be diagonalized in the same basis. Since $B_r \in \mathbb{V}_r(\mathbb{R}^m)$, it can be seen as the r first vectors of an orthogonal basis of $\mathbb{R}^m$. It means there exists a matrix B_{m-r} such that

$$B_r\Lambda_r^2B_r^T = B\Lambda_m^2B^T,$$

where $\Lambda_m^2 \in \mathbb{R}^{m\times m} = \text{diag}(\lambda_1^2, ..., \lambda_r^2, 0, ..., 0)$ and $B = \begin{pmatrix}B_r & B_{m-r}\end{pmatrix}$. Thus the eigenvalues of $B_r\Lambda_r^2B_r^T$ are exactly the eigenvalues of Λ_m^2. Since Σ_0 and $B_r\Lambda_r^2B_r^T$ can be diagonalized in the same basis, we get that $\text{tr}(\Lambda_r^2B_r^T\Sigma_0B_r) = \text{tr}(\Sigma_0B_r\Lambda_r^2B_r^T) = \text{tr}(D_0\tilde{\Lambda}_m)$ where $\tilde{\Lambda}_m$ is a diagonal matrix with the same eigenvalues as Λ_m, but in a different order. Now, it can be easily seen that the optimal value of (8.6) is reached when B_r is a permutation matrix which sorts the eigenvalues of Λ_m in decrasing order.

Thus, for a given S, the maximum value of $\text{tr}(K^TK)$ is $\text{tr}(D_0\tilde{\Lambda}_m(S))$. We can now establish for which S, $\text{tr}(D_0\tilde{\Lambda}_m(S))$ is optimal. For a given S, we denote $\lambda_1, ..., \lambda_n$ the eigenvalues of $\Sigma_1 - S$ and $\beta_1, ..., \beta_n$ the eigenvalues of Σ_1 ordered in decreasing order. Since $S \in S_n^+(\mathbb{R})$, $\forall x \in \mathbb{R}^n$, the following inequality holds:

$$x^T(\Sigma_1 - S)x \leq x^T\Sigma_1x. \tag{8.8}$$

This inequality still holds when restricted to any subspace of $\mathbb{R}^n$. Using the Courant-Fischer theorem, we can conclude that:

$$\forall i \leq n,\ \lambda_i \leq \beta_i. \tag{8.9}$$

Thus, the optimal value of $\text{tr}(D_0\tilde{\Lambda}_m(S))$ is reached when $S = 0$ and $\tilde{\Lambda}_m(0) = \begin{pmatrix}D_1 & 0\\ 0 & 0\end{pmatrix}$ and so $\text{tr}(D_0\tilde{\Lambda}_m(0)) = \text{tr}(D_0^{(n)}D_1)$. Let $A = \begin{pmatrix}\tilde{I}_n(D_0^{(n)})^{\frac{1}{2}}D_1^{\frac{1}{2}}\\ 0_{m-n,n}\end{pmatrix}$ with $\tilde{I}_n$ of the form $\text{diag}((\pm1)_{i\leq n}))$. It can be easily verified that $A^TD_0^{-1}A = D_1$ and if $K^* = P_0^TAP_1$, $K^{*T}\Sigma_0^{-1}K^* = P_1^TA^TP_0\Sigma_0^{-1}P_0^TAP_1 = P_1^TA^TD_0^{-1}AP_1 = P_1^TD_1P_1 = \Sigma_1$ and $K^{*T}K^*$ has the same eigenvalues as A^TA and $\text{tr}(A^TA) = \text{tr}(D_0^{(n)}D_1)$. □

8.2 Proof of Lemma 3.3

In order to prove lemma 3.3, we will use the following lemma, demonstrated by Antreicher and Wolkowicz [2].

Lemma 8.1. (Anstreicher and Wolkowicz, 1998, [2]) *Let Σ_0 and Σ_1 be two symmetric matrices of size n. We note $\Sigma_0 = P_0\Lambda_0P_0^T$ and $\Sigma_1 = P_1\Lambda_1P_1^T$ there respective diagonalization such that the eigenvalues of Λ_0 are sorted in non-increasing order and the eigenvalues of Λ_1 are sorted in increasing order. Then*

$$\min_{PP^T=I_n} \mathrm{tr}(\Sigma_0P\Sigma_1P^T) = \mathrm{tr}(\Lambda_0\Lambda_1), \tag{8.10}$$

and it is achieved for $P^* = P_0P_1^T$.

Proof of Lemma 3.3. We proceed in the same way as before: first, we derive the expression of the optimal value for a given $S = \Sigma_1 - K^T\Sigma_0^{-1}K$, then we determine for which S this expression is maximum. The start of the proof is exactly the same as the proof of (3.2) until formula (8.4). We diagonalize $\Sigma_1 - S = K^T\Sigma_0^{-1}K = U_r\Lambda_rU_r^T$ where r is the rank of $K^T\Sigma_0^{-1}K$, then we set $B_r = \Sigma_0^{-\frac{1}{2}}KU_r\Lambda_r^{-1}$ while observing that $B_r \in \mathbb{V}_r(\mathbb{R}^m)$ and we deduce that $K = \Sigma_0^{\frac{1}{2}}B_r\Lambda_rU_r^T$. By reinjecting this expression, it comes that

$$\mathrm{tr}(KA) = \mathrm{tr}(A^TK^T) = \mathrm{tr}(A^TU_r\Lambda_rB_r^T\Sigma_0^{\frac{1}{2}}) = \mathrm{tr}(\Sigma_0^{\frac{1}{2}}A^TU_r\Lambda_rB_r^T). \tag{8.11}$$

For a given S, the problem of finding the optimal value is parametrized by B_r and is:

$$\min_{B_r\in\mathbb{V}_r(\mathbb{R}^m)} -\mathrm{tr}(\Sigma_0^{\frac{1}{2}}A^TU_r\Lambda_rB_r^T). \tag{8.12}$$

The Lagrangian of this problem can be written:

$$\mathcal{L}(B_r, C) = -\mathrm{tr}(\Sigma_0^{\frac{1}{2}}A^TU_r\Lambda_rB_r^T) + \mathrm{tr}(C(B_r^TB_r - I_r)), \tag{8.13}$$

where $C \in S_r(\mathbb{R})$ is the Lagrangian multiplier associated to the constraint $B_r^TB_r = I_r$. We can then derive the first-order condition:

$$-\Sigma_0^{\frac{1}{2}}A^TU_r\Lambda_r + 2B_rC = 0,$$

or equivalently:

$$\Sigma_0^{\frac{1}{2}}A^TU_r\Lambda_rB_r^T = 2B_rCB_r^T.$$

Since $C \in S_r(\mathbb{R})$, $2B_rCB_r^T \in S_m(\mathbb{R})$ and $\Sigma_0^{\frac{1}{2}}A^TU_r\Lambda_rB_r^T \in S_m(\mathbb{R})$. Moreover, the rank of $\Sigma_0^{\frac{1}{2}}A^TU_r\Lambda_rB_r^T$ is equal to 1 because $\mathrm{rk}(A) = 1$ and $\mathrm{rk}(\Sigma_0^{\frac{1}{2}}A^TU_r\Lambda_rB_r^T) = 0$ would imply that $\mathrm{tr}(KA) = 0$, which cannot be the maximum value of our problem. So there exists a vector $u_m \in \mathbb{R}^m$ such that

$$\Sigma_0^{\frac{1}{2}}A^TU_r\Lambda_rB_r^T = u_mu_m^T. \tag{8.14}$$

Then we can reinject the value B_r in the expression:

$$\begin{aligned} \Sigma_0^{\frac{1}{2}}A^TU_r\Lambda_rB_r^T &= \Sigma_0^{\frac{1}{2}}A^TU_r\Lambda_r\Lambda_r^{-1}U_r^TK^T\Sigma_0^{-\frac{1}{2}} \\ &= \Sigma_0^{\frac{1}{2}}A^TU_rU_r^TK^T\Sigma_0^{-\frac{1}{2}} \\ &= \Sigma_0^{\frac{1}{2}}A^TK^T\Sigma_0^{-\frac{1}{2}}, \end{aligned} \tag{8.15}$$

where we used the fact that $K = KUU^T = KU_rU_r^T$ because $KU_{n-r} = 0$. We have so on one hand:

$$\mathrm{tr}(KA) = \mathrm{tr}(\Sigma_0^{\frac{1}{2}}A^TK^T\Sigma_0^{-\frac{1}{2}}) = \mathrm{tr}(u_mu_m^T) = u_m^Tu_m, \tag{8.16}$$

and on the other hand:

$$\begin{aligned}
\Sigma_0^{\frac{1}{2}} A^T K^T \Sigma_0^{-\frac{1}{2}} (\Sigma_0^{\frac{1}{2}} A^T K^T \Sigma_0^{-\frac{1}{2}})^T &= \Sigma_0^{\frac{1}{2}} A^T K^T \Sigma_0^{-1} K A D_0^{\frac{1}{2}} \\
&= \Sigma_0^{\frac{1}{2}} A^T (\Sigma_1 - S) A \Sigma_0^{\frac{1}{2}} \\
&= u_m u_m^T u_m u_m^T \\
&= u_m^T u_m u_m u_m^T,
\end{aligned} \tag{8.17}$$

and thus

$$\mathrm{tr}(\Sigma_0^{\frac{1}{2}} A^T (\Sigma_1 - S) A \Sigma_0^{\frac{1}{2}}) = u_m^T u_m \mathrm{tr}(u_m u_m^T) = (u_m^T u_m)^2 = (\mathrm{tr}(KA))^2. \tag{8.18}$$

Then we will determine for which S, $\mathrm{tr}(\Sigma_0^{\frac{1}{2}} A^T (\Sigma_1 - S) A \Sigma_0^{\frac{1}{2}})$ is maximum:

$$\begin{aligned}
\mathrm{tr}(\Sigma_0^{\frac{1}{2}} A^T (\Sigma_1 - S) A \Sigma_0^{\frac{1}{2}}) &= \mathrm{tr}(A \Sigma_0 A^T (\Sigma_1 - S)) \\
&= \mathrm{tr}(A \Sigma_0 A^T \Sigma_1) - \mathrm{tr}(A \Sigma_0 A^T S).
\end{aligned} \tag{8.19}$$

Let $B = A\Sigma_0 A^T$. We can observe that $B \in S_n^+(\mathbb{R})$ with rank 1. Moreover, since $S \in S_n^+(\mathbb{R})$, it can be diagonalized, and we denote $S = PDP^T$. As before, we will first determine the value of $\mathrm{tr}(BS)$ for a given D, then we will determine which D minimizes $\mathrm{tr}(BS)$. For a given D, we want the optimal value of

$$\min_{PP^T = I_b} \mathrm{tr}(BPDP^T). \tag{8.20}$$

Since B is symmetric with rank 1, it has only one non null eigenvalue which is equal to its trace. Using Lemma 8.1, we can deduce that

$$\min_{PP^T = I_n} \mathrm{tr}(BPDP^T) = \mathrm{tr}(B)\lambda_n, \tag{8.21}$$

where λ_n is the smallest eigenvalue of D. Since $S \in S_n^+(\mathbb{R})$, the smallest possible value for λ_n is 0.

If $\Sigma_0 = \mathrm{diag}(\alpha)$, $\Sigma_1 = \mathrm{diag}(\beta)$, it can be easily seen that $\mathrm{tr}(\mathbb{1}_{n,m}\Sigma_0\mathbb{1}_{m,n}\Sigma_1) = \mathrm{tr}(\Sigma_0)\mathrm{tr}(\Sigma_1)$. Thus, if $K = \frac{\alpha\beta^T}{\sqrt{\mathrm{tr}(\Sigma_0)\mathrm{tr}(\Sigma_1)}} = \frac{\Sigma_0 \mathbb{1}_{m,n}\Sigma_1}{\sqrt{\mathrm{tr}(\Sigma_0)\mathrm{tr}(\Sigma_1)}}$, we can observe that

$$\mathrm{tr}(K\mathbb{1}_{n,m}) = \mathrm{tr}(\mathbb{1}_{n,m}K) = \frac{\mathrm{tr}(\mathbb{1}_{n,m}\Sigma_0\mathbb{1}_{m,n}\Sigma_1)}{\sqrt{\mathrm{tr}(\Sigma_0)\mathrm{tr}(\Sigma_1)}} = \sqrt{\mathrm{tr}(\Sigma_0)\mathrm{tr}(\Sigma_1)}, \tag{8.22}$$

Now we must show that $S = \Sigma_1 - K^T\Sigma_0^{-1}K \in S_n^+(\mathbb{R})$. To do so, we will show that $\forall i \leq n$, the determinant of the principal minor $S^{(i)}$ is positive. We can derive that

$$S = \Sigma_1 - \frac{\beta\alpha^T\Sigma_0^{-1}\alpha\beta^T}{\mathrm{tr}(\Sigma_0)\mathrm{tr}(\Sigma_1)} = \Sigma_1 - \frac{\beta\beta^T\mathrm{tr}(\Sigma_0)}{\mathrm{tr}(\Sigma_0)\mathrm{tr}(\Sigma_1)} = \Sigma_1 - \frac{\beta\beta^T}{\mathrm{tr}(\Sigma_1)}. \tag{8.23}$$

Using the matrix determinant lemma, it comes that

$$\forall i \leq n,\ \det(S^{(i)}) = \prod_k^i \beta_k \left(1 - \frac{\mathrm{tr}(\Sigma_1^{(i)})}{\mathrm{tr}(\Sigma_1)}\right). \tag{8.24}$$

Thus, $\forall i < n$, $\det(S^{(i)}) > 0$, and $\det(S) = 0$. We conclude that $S \in S_n^+(\mathbb{R})$ and the smallest eigenvalue of S is 0. $\square$

8.3 Proof of Lemma 5.1

Proof. For $m \geq 1$, let Γ_m denote the set of vectors $v = (v_1, \ldots, v_m)$ of $\mathbb{R}^m$ such that $v_1 \geq v_2 \geq \ldots \geq v_m \geq 0$ and $\sum_{i=1}^m v_i^2 = 1$. We want to prove that

$$\forall u, v \in \Gamma_m, \quad \sum_{i=1}^m u_i v_i \geq \frac{1}{\sqrt{m}}. \tag{8.25}$$

We proceed by induction on m. For $m = 1$, it's obviously true since $\Gamma_1 = \{1\}$. Assume now $m > 1$, and the result true for $m - 1$. Let $u, v \in \Gamma_m$, then using the result for $(u_2, \ldots, u_m)/(\sum_{i=2}^m u_i^2)^{1/2}$ and $(v_2, \ldots, v_m)/(\sum_{i=2}^m v_i^2)^{1/2}$ that both belong to Γ_{m-1}, we have

$$\begin{aligned} \sum_{i=1}^m u_i v_i = u_1 v_1 + \sum_{i=2}^m u_i v_i &\geq u_1 v_1 + \frac{1}{\sqrt{m-1}} \left(\sum_{i=2}^m u_i^2 \right)^{1/2} \left(\sum_{i=2}^m v_i^2 \right)^{1/2} \\ &= u_1 v_1 + \frac{1}{\sqrt{m-1}} \sqrt{1 - u_1^2} \sqrt{1 - v_1^2}. \end{aligned} \tag{8.26}$$

Now since $u, v \in \Gamma_m$, we have $u_1, v_1 \in [\frac{1}{\sqrt{m}}, 1]$. Let us denote $F(u_1, v_1) = u_1 v_1 + \frac{1}{\sqrt{m-1}} \sqrt{1 - u_1^2} \sqrt{1 - v_1^2}$. We have for all $v_1 \in [\frac{1}{\sqrt{m}}, 1]$:

$$F(1, v_1) = v_1 \geq \frac{1}{\sqrt{m}} \quad \text{and} \quad F(\frac{1}{\sqrt{m}}, v_1) = \frac{\sqrt{1 - v_1^2} + v_1}{\sqrt{m}} \geq \frac{1 - v_1^2 + v_1}{\sqrt{m}} \geq \frac{1}{\sqrt{m}}. \tag{8.27}$$

And computing the partial derivative of F with respect to u_1, we get

$$\frac{\partial F}{\partial u_1}(u_1, v_1) = v_1 - \frac{u_1 \sqrt{1 - v_1^2}}{\sqrt{m-1} \sqrt{1 - u_1^2}}. \tag{8.28}$$

This is a decreasing function of u_1, with value v_1 at $u_1 = 0$ and value that goes to $-\infty$ when u_1 goes to 1. Therefore the function $F(\cdot, v_1)$ on $[0, 1]$ is first increasing and then decrasing, showing that

$$\forall u_1 \in [\frac{1}{\sqrt{m}}, 1], \quad F(u_1, v_1) \geq \min \left(F(\frac{1}{\sqrt{m}}, v_1), F(1, v_1) \right) \geq \frac{1}{\sqrt{m}}. \tag{8.29}$$

Finally we thus have proved that

$$\sum_{i=1}^m u_i v_i \geq \frac{1}{\sqrt{m}},$$

and moreover the equality is achieved when the vectors u and v are the vectors $(1, 0, \ldots, 0)$ and $(\frac{1}{\sqrt{m}}, \frac{1}{\sqrt{m}}, \ldots, \frac{1}{\sqrt{m}})$. ∎

References

[1] David Alvarez-Melis, Stefanie Jegelka, and Tommi S Jaakkola. Towards optimal transport with global invariances. In *International Conference on Artificial Intelligence and Statistics*, pages 1870–1879. PMLR, 2019.

[2] Kurt Anstreicher and Henry Wolkowicz. On Lagrangian relaxation of quadratic matrix constraints. In *Journal on Matrix Analysis and Applications*, volume 22, pages 41–55. SIAM, 2000.

[3] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223. PMLR, 2017.

[4] Jérémie Bigot, Raúl Gouet, Thierry Klein, Alfredo López, et al. Geodesic PCA in the Wasserstein space by convex PCA. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 53, pages 1–26. Institut Henri Poincaré, 2017.

[5] Jose Blanchet, Yang Kang, and Karthyek Murthy. Robust Wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56(3):830–857, 2019.

[6] Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. In *Communications on Pure and Applied Mathematics*, volume 44, pages 375–417. Wiley, 1991.

[7] Yuhang Cai and Lek-Heng Lim. Distances between probability distributions of different dimensions. In *arXiv preprint*, 2020.

[8] Samir Chowdhury and Tom Needham. Gromov–Wasserstein averaging in a Riemannian framework. In *Conference on Computer Vision and Pattern Recognition Workshops*, pages 842–843. IEEE/CVF, 2020.

[9] Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. In *Transactions on Pattern Analysis and Machine Intelligence*, volume 39, pages 1853–1865. IEEE, 2016.

[10] DC Dowson and BV Landau. The Fréchet distance between multivariate normal distributions. *Journal of multivariate analysis*, 12(3):450–455, 1982.

[11] Alfred Galichon, Pierre Henry-Labordere, Nizar Touzi, et al. A stochastic control approach to no-arbitrage bounds given marginals, with an application to lookback options. *The Annals of Applied Probability*, 24(1):312–336, 2014.

[12] Aude Genevay, Gabriel Peyre, and Marco Cuturi. Learning Generative Models with Sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, volume 84, pages 1608–1617. PMLR, 2018.

[13] Clark R Givens, Rae Michael Shortt, et al. A class of Wasserstein metrics for probability distributions. In *Michigan Mathematical Journal*, volume 31, pages 231–240. the University of Michigan, 1984.

[14] Leon Isserlis. On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables. In *Biometrika*, volume 12, pages 134–139. JSTOR, 1918.

[15] Ioan Mackenzie James. *The topology of Stiefel manifolds*, volume 24. Cambridge University Press, 1976.

[16] Facundo Mémoli. Gromov–Wasserstein distances and the metric approach to object matching. In *Foundations of Computational Mathematics*, volume 11, pages 417–487. Springer, 2011.

[17] Ofir Pele and Ben Taskar. The tangent earth mover's distance. In *International Conference on Geometric Science of Information*, pages 397–404. Springer, 2013.

[18] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: with applications to data science. In *Foundations and Trends in Machine Learning*, volume 11, pages 355–607. Now Publishers Inc., 2019.

[19] Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov-Wasserstein averaging of kernel and distance matrices. In *International Conference on Machine Learning*, pages 2664–2672. PMLR, 2016.

[20] Julien Rabin, Sira Ferradans, and Nicolas Papadakis. Adaptive color transfer with relaxed optimal transport. In *International Conference on Image Processing*, pages 4852–4856. IEEE, 2014.

[21] Julien Rabin, Gabriel Peyré, Julie Delon, and Marc Bernot. Wasserstein barycenter and its application to texture mixing. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 435–446. Springer, 2011.

[22] Filippo Santambrogio. Optimal transport for applied mathematicians. In *Birkäuser NY*, volume 55, page 94. Springer, 2015.

[23] Karl-Theodor Sturm. The space of spaces: curvature bounds and gradient flows on the space of metric measure spaces. In *arXiv preprint*, 2012.

[24] Asuka Takatsu. On Wasserstein geometry of Gaussian measures. In *Probabilistic approach to geometry*, pages 463–472. Mathematical Society of Japan, 2010.

[25] Titouan Vayer. A contribution to optimal transport on incomparable spaces. In *arXiv preprint*, 2020.

[26] C Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

[27] Cédric Villani. *Topics in optimal transportation.* Number 58. American Mathematical Soc., 2003.