J. Appl. Probab. **59,** 1178–1198 (2022)
*doi:*10.1017/jpr.2022.16

GROMOV–WASSERSTEIN DISTANCES BETWEEN GAUSSIAN DISTRIBUTIONS

JULIE DELON,* *Université de Paris*
AGNES DESOLNEUX,** *CNRS and ENS Paris-Saclay*
ANTOINE SALMONA,*** *ENS Paris-Saclay*

Abstract

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. 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 a 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

2020 Mathematics Subject Classification: Primary 60E99
Secondary 68T09; 62H25; 49Q22

1. Introduction

Optimal transport (OT) theory has nowadays become a major tool for comparing probability distributions. It has been increasingly used in recent years in various applied fields such as economics [11], image processing [20, 21], machine learning [4, 5], and, 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:

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

Received 15 April 2021; revision received 23 January 2022.
* Postal address: Université de Paris, CNRS, MAP5 UMR 8145 and Institut Universitaire de France, 45 rue des Saints-Pères, 75006 Paris, France. Email: julie.delon@parisdescartes.fr
** Postal address: ENS Paris-Saclay, CNRS, Centre Borelli, UMR 9010, 4 avenue des sciences, 91190 Gif-sur-Yvette, France. Email: agnes.desolneux@ens-paris-saclay.fr
*** Postal address: ENS Paris-Saclay, CNRS, Centre Borelli, UMR 9010, 4 avenue des sciences, 91190 Gif-sur-Yvette, France. Email: antoinesalmona2@gmail.com

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 $p \geq 1$, (1.1) induces a distance over the set of measures with finite moment of order p, known as the p-Wasserstein distance W_p:

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

or, equivalently, $W_p^p(\mu, \nu) = \inf_{X \sim \mu, Y \sim \nu} \mathbb{E}[\|X - Y\|^p]$, where the notation $X \sim \mu$ means that X is a random variable with probability distribution μ. It is known that (1.1) always admits a solution [22, 26, 27], 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 the *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 $\Sigma_1 \in \mathbb{R}^{d \times d}$; this is given by

$$W_2^2(\mu, \nu) = \|m_1 - m_0\|^2 + \mathrm{tr}\left(\Sigma_0 + \Sigma_1 - 2\left({\Sigma_0}^{1/2} \Sigma_1 {\Sigma_0}^{1/2} \right)^{1/2} \right), \tag{1.2}$$

where, for any symmetric semi-definite positive M, $M^{1/2}$ is the unique symmetric semi-definite positive square root of M. Moreover, if Σ_0 is non-singular, then the optimal transport map T is affine and is given, for all $x \in \mathbb{R}^d$, by

$$T(x) = m_1 + {\Sigma_0}^{-1/2} \left({\Sigma_0}^{1/2} \Sigma_1 {\Sigma_0}^{1/2} \right)^{1/2} {\Sigma_0}^{-1/2} (x - m_0), \tag{1.3}$$

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 classical OT lies in the fact that it is not invariant to rotations and translations, or, 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 easily define 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

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

with $p \geq 1$. As for classic OT, it can be shown that this equation 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. When $c_{\mathcal{X}} = d_{\mathcal{X}}^q$ and $c_{\mathcal{Y}} = d_{\mathcal{Y}}^q$ with $q \geq 1$, then GW_p 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, 25]. (We say that $(\mathcal{X}, d_{\mathcal{X}}, \mu)$ is strongly isomorphic to $(\mathcal{Y}, d_{\mathcal{Y}}, \nu)$ if there 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 $\phi\#\mu = \nu$.) The fundamental metric properties of GW_p have been studied in depth in [8, 16, 23]. In the Euclidean setting, when $\mathcal{X} = \mathbb{R}^m$, $\mathcal{Y} = \mathbb{R}^n$,

https://doi.org/10.1017/jpr.2022.16 Published online by Cambridge University Press

with m not necessarily being equal to n, and for the natural choice of costs $c_{\mathcal{X}} = \|\cdot\|^2_{\mathbb{R}^m}$ and $c_{\mathcal{Y}} = \|\cdot\|^2_{\mathbb{R}^n}$, where $\|\cdot\|_{\mathbb{R}^m}$ 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 notation, we will write 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

$$GW_2^2(\mu, \nu) = \inf_{\pi\in\Pi(\mu,\nu)} \int\int \left(\|x - x'\|^2_{\mathbb{R}^m} - \|y - y'\|^2_{\mathbb{R}^n}\right)^2 \mathrm{d}\pi(x, y)\,\mathrm{d}\pi(x', y'), \tag{1.4}$$

or, equivalently, $GW_2^2(\mu, \nu) = \inf_{X,Y,X',Y'\sim\pi\otimes\pi} \mathbb{E}\big[(\|X - X'\|^2_{\mathbb{R}^m} - \|Y - Y'\|^2_{\mathbb{R}^n})^2\big]$, where, for $x, x' \in \mathbb{R}^m$ and $y, y' \in \mathbb{R}^n$, $(\pi\otimes\pi)(x, y, x', y') = \pi(x, y)\pi(x', y')$. In particular, can we find equivalent formulas to (1.2) and (1.3) 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 component analysis (PCA). This yields an upper bound to the general problem. In Section 5 we study the tightness of the bounds found in the previous sections, and we exhibit a particular case where these upper and lower bounds coincide, which provides an exact computation 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.

1.1. Notation

We define in the following some of the notation 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, for all $A \in \mathcal{Y}$, $(T\#\mu)(A) = \mu(T^{-1}(A))$.
- If X and Y are random vectors on $\mathbb{R}^m$ and $\mathbb{R}^n$, we denote by $\mathrm{Cov}(X, Y)$ the matrix of size $m \times n$ of the form $\mathbb{E}\big[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])^\top\big]$.
- $\mathrm{tr}(M)$ denotes the trace of a matrix M.
- $\|M\|_{\mathcal{F}}$ stands for the Frobenius norm of a matrix M, i.e. $\|M\|_{\mathcal{F}} = \sqrt{\mathrm{tr}(M^\top M)}$.
- $\mathrm{rk}(M)$ 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 $\mathrm{diag}((\pm 1)_{i\le n})$.
- Suppose $n \le m$. For $A \in \mathbb{R}^{m\times m}$, we denote by $A^{(n)} \in \mathbb{R}^{n\times n}$ the submatrix containing the n first rows and the n first columns of A.
- Suppose $n \le m$. For $A \in \mathbb{R}^{n\times n}$, we denote by $A^{[m]} \in \mathbb{R}^{m\times m}$ the matrix of the form $\left(\begin{smallmatrix} A & 0 \\ 0 & 0 \end{smallmatrix}\right)$.

https://doi.org/10.1017/jpr.2022.16 Published online by Cambridge University Press

- We denote by $S_n(\mathbb{R})$ the set of symmetric matrices of size n, by $S_n^+(\mathbb{R})$ the set of semi-definite positive matrices, and by $S_n^{++}(\mathbb{R})$ the set of definite positive matrices.
- $\mathbf{1}_{n,m} = (1)_{i\le n, j\le m}$ 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 write $\|x\|$ 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 formulation of problem (1.4) which takes the form of a functional of co-moments of order 2 and 4 of π. (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). This formulation is not specific to Gaussian measures but to all measures with finite fourth-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 $\int \|x\|^4 \,\mathrm{d}\mu < +\infty$, 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 $\int \|y\|^4 \,\mathrm{d}\nu < +\infty$. Let P_0, D_0 and P_1, D_1 be respective diagonalizations of Σ_0 ($= P_0 D_0 P_0^\top$) and Σ_1 ($= P_1 D_1 P_1^\top$). Let us define $T_0 : x \in \mathbb{R}^m \mapsto P_0^\top(x - m_0)$ and $T_1 : y \in \mathbb{R}^n \mapsto P_1^\top(y - m_1)$. Then problem (1.4) is equivalent to*

$$\sup_{X\sim T_0\#\mu, Y\sim T_1\#\nu} \sum_{i,j} \mathrm{Cov}(X_i^2, Y_j^2) + 2\|\mathrm{Cov}(X, Y)\|_{\mathcal{F}}^2, \tag{2.1}$$

where $X = (X_1, X_2, \dots, X_m)^\top$, $Y = (Y_1, Y_2, \dots, Y_n)^\top$, and $\|\cdot\|_{\mathcal{F}}$ is the Frobenius norm. More precisely, (X,Y) is optimal for (2.1) if and only if the law of $(T_0^{-1}(X), T_1^{-1}(Y))$ is optimal for (1.4).

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

Lemma 2.1. *We denote by $\mathbb{O}_m = \{O \in \mathbb{R}^{m\times m} \mid O^\top O = I_m\}$ 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 $\langle\cdot,\cdot\rangle_m$ 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 $GW_2^2(c_{\mathcal{X}}, c_{\mathcal{Y}}, \mu, \nu) = C_{\mu,\nu} - 2\sup_{\pi\in\Pi(\mu,\nu)} Z(\pi)$, where $C_{\mu,\nu} = \int c_{\mathcal{X}}^2 \,\mathrm{d}\mu \,\mathrm{d}\mu + \int c_{\mathcal{Y}}^2 \,\mathrm{d}\nu \,\mathrm{d}\nu - 4ab \int \|x\|_{\mathbb{R}^m}^2 \|y\|_{\mathbb{R}^n}^2 \,\mathrm{d}\mu \,\mathrm{d}\nu$ and*

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

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

https://doi.org/10.1017/jpr.2022.16 Published online by Cambridge University Press

$T_0 : x \in \mathbb{R}^m \mapsto P_0^\top(x - m_0)$ and $T_1 : y \in \mathbb{R}^n \mapsto P_1^\top(y - m_1)$, 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 (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 follows that problem (1.4) is equivalent to

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

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

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

Since the second term does not depend on π, we get that problem (1.4) is equivalent to problem (2.1). □

The left-hand term of (2.1) 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 (2.1) is hard to solve because it involves simultaneously optimizing the co-moments of order 2 and 4 of π and so knowing the probabilistic rule which links them. This rule is well known when π is Gaussian (the 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 (2.1) to be Gaussian.

3. Study of the general problem

Since problem (2.1) is hard to solve because of its dependence on co-moments of order 2 and 4 of π, one can optimize both terms separately in order to find a lower bound for $GW_2(\mu, \nu)$. In the rest of the paper we suppose for convenience and without any loss of generality that $n \le m$.

Theorem 3.1. *Suppose, without any loss of generality, that $n \le m$. 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 $\Sigma_0 (= P_0D_0P_0^\top)$ and $\Sigma_1 (= P_1D_1P_1^\top)$ which sort the eigenvalues in decreasing order. We suppose that Σ_0 is non-singular. A lower bound for $GW_2(\mu, \nu)$ is then $GW_2^2(\mu, \nu) \ge LGW_2^2(\mu, \nu)$, where*

$$\begin{aligned} LGW_2^2(\mu, \nu) = {} & 4(\mathrm{tr}(D_0) - \mathrm{tr}(D_1))^2 + 4(\|D_0\|_{\mathcal{F}} - \|D_1\|_{\mathcal{F}})^2 + 4\|D_0^{(n)} - D_1\|^2_{\mathcal{F}} \\ & + 4\left(\|D_0\|^2_{\mathcal{F}} - \|D_0^{(n)}\|^2_{\mathcal{F}}\right). \end{aligned} \tag{3.1}$$

The proof of this theorem is divided into smaller intermediary results. We first recall the Isserlis lemma (see [14]), which allows us to derive the co-moments of order 4 of a Gaussian distribution as a function 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,for all $i, j, k, l \le n$, $\mathbb{E}[X_iX_jX_kX_l] = \mathbb{E}[X_iX_j]\mathbb{E}[X_kX_l] + \mathbb{E}[X_iX_k]\mathbb{E}[X_jX_l] + \mathbb{E}[X_iX_l]\mathbb{E}[X_jX_k]$.*

https://doi.org/10.1017/jpr.2022.16 Published online by Cambridge University Press

Then we derive the following general optimization lemmas. The proofs of these two lemmas are postponed to Appendix A.

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

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

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 $\Sigma_0 (= P_0^\top D_0 P_0)$ and $\Sigma_1 (= P_1^\top D_1 P_1)$ which sort the eigenvalues in decreasing order. Then

$$\max_{\Sigma_1 - K^\top \Sigma_0^{-1} K \in S_n^+(\mathbb{R})} \|K\|_{\mathcal{F}}^2 = \operatorname{tr}\big(D_0^{(n)} D_1\big), \tag{3.2}$$

and is achieved at any

$$K^* = P_0^\top \begin{pmatrix} \tilde{I}_n (D_0^{(n)})^{\frac{1}{2}} D_1^{\frac{1}{2}} \\ 0_{m-n,n} \end{pmatrix} P_1, \tag{3.3}$$

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

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

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

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 $\max_{\Sigma_1 - K^\top \Sigma_0^{-1} K \in S_n^+(\mathbb{R})} \operatorname{tr}(KA) = \big[\operatorname{tr}(A\Sigma_0 A^\top \Sigma_1)\big]^{1/2}$. 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

$$\max_{\Sigma_1 - K^\top \Sigma_0^{-1} K \in S_n^+(\mathbb{R})} \operatorname{tr}(K\mathbf{1}_{n,m}) = \big[\operatorname{tr}(\Sigma_0)\operatorname{tr}(\Sigma_1)\big]^{1/2},$$

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

$$K^* = \frac{\alpha\beta^\top}{\big[\operatorname{tr}(\Sigma_0)\operatorname{tr}(\Sigma_1)\big]^{1/2}}. \tag{3.4}$$

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

$$\Sigma = \begin{pmatrix} D_0 & K \\ K^\top & D_1 \end{pmatrix}, \qquad \tilde{\Sigma} = \begin{pmatrix} 2D_0^2 & \tilde{K} \\ \tilde{K}^\top & 2D_1^2 \end{pmatrix}.$$

In order to find a supremum for each term of (2.1), 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 $D_1 - K^\top D_0^{-1} K$ and

https://doi.org/10.1017/jpr.2022.16 Published online by Cambridge University Press

$2D_1^2 - \frac{1}{2}\tilde{K}^\top D_0^{-2}\tilde{K}$, must also be semi-definite positive. Remarking that the left-hand term in (2.1) can be rewritten as $\mathrm{tr}(\tilde{K}\mathbf{1}_{m,n})$, we have the following two inequalities:

$$\sup_{X\sim T_0\#\mu, Y\sim T_1\#\nu} \sum_{i,j} \mathrm{Cov}(X_i^2, Y_j^2) \leq \max_{2D_1^2-\frac{1}{2}K^\top D_0^{-2}K\in S_n^+(\mathbb{R})} \mathrm{tr}(\tilde{K}\mathbf{1}_{n,m}), \tag{3.5}$$

$$\sup_{X\sim T_0\#\mu, Y\sim T_1\#\nu} \|\mathrm{Cov}(X, Y)\|_{\mathcal{F}}^2 \leq \max_{D_1-K^\top D_0^{-1}K\in S_n^+(\mathbb{R})} \|K\|_{\mathcal{F}}^2. \tag{3.6}$$

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

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

and, on the other hand,

$$\sup_{X\sim T_0\#\mu, Y\sim T_1\#\nu} \sum_{i,j} \mathrm{Cov}(X_i^2, Y_j^2) \leq 2\big[\mathrm{tr}(D_0^2)\mathrm{tr}(D_1^2)\big]^{1/2} = 2\|D_0\|_{\mathcal{F}}\|D_1\|_{\mathcal{F}}.$$

Furthermore, using Lemma 2.2, it follows that

$$\begin{aligned} &GW_2^2(\mu, \nu) \\ &= C_{\mu,\nu} - 4 \sup_{X\sim T_0\#\mu, Y\sim T_1\#\nu} \left(\sum_{i,j} \mathrm{Cov}(X_i^2, Y_j^2) + \sum_{i,j} \mathbb{E}[X_i^2]\mathbb{E}[X_j^2] + 2\|\mathrm{Cov}(X, Y)|_{\mathcal{F}}^2 \right) \\ &\geq C_{\mu,\nu} - 8\big[\mathrm{tr}(D_0^2)\mathrm{tr}(D_1^2)\big]^{1/2} - 4\mathrm{tr}(D_0)\mathrm{tr}(D_1) - 8\mathrm{tr}(D_0^{(n)}D_1), \end{aligned}$$

where

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

Finally,

$$\begin{aligned} GW_2^2(\mu, \nu) &\geq 4(\mathrm{tr}(D_0))^2 + 4(\mathrm{tr}(D_1))^2 - 8\mathrm{tr}(D_0)\mathrm{tr}(D_1) + 8\mathrm{tr}(D_0^2) + 8\mathrm{tr}(D_1^2) \\ &\quad - 8\big[\mathrm{tr}(D_0^2)\mathrm{tr}(D_1^2)\big]^{1/2} - 8\mathrm{tr}(D_0^{(n)}D_1) \\ &= LGW_2^2(\mu, \nu). \end{aligned}$$

□

Inequality (3.5) becomes an equality if there exists a plan $\pi \in \Pi(T_0\#\mu, T_1\#\nu)$ such that for $(X, Y)\sim\pi$, $\sum \mathrm{Cov}(X_i^2, Y_j^2) = 2\|D_0\|_{\mathcal{F}}\|D_1\|_{\mathcal{F}}$. Thanks to Lemma 3.3, we know that $\mathrm{tr}(\tilde{K}^*\mathbf{1}_{n,m}) = 2\|D_0\|_{\mathcal{F}}\|D_1\|_{\mathcal{F}}$ for

$$\tilde{K}^* = \frac{2\alpha^2(\beta^2)^\top}{\|D_0\|_{\mathcal{F}}\|D_1\|_{\mathcal{F}}},$$

where $\alpha^2 = (\alpha_1^2, \alpha_2^2, \ldots, \alpha_m^2)$ and $\beta^2 = (\beta_1^2, \beta_2^2, \ldots, \beta_n^2)$ are the diagonal vectors of D_0^2 and D_1^2. However, it does not seem straightforward to exhibit a plan $\pi \in \Pi(T_0\#\mu, T_1\#\nu)$ such that for $(X, Y)\sim\pi$, $\mathrm{Cov}(X^2, Y^2) = \tilde{K}^*$. An important point to mention is that it can be shown,

https://doi.org/10.1017/jpr.2022.16 Published online by Cambridge University Press

thanks to Isserlis' lemma, that there does not exist a Gaussian plan in $\Pi(T_0\#\mu, T_1\#\nu)$ with such symmetric co-moments of order 4.

On the other hand, it can be easily seen that inequality (3.6) is in fact an equality since the maximal value of $\|\mathrm{Cov}(X, Y)\|_{\mathcal{F}}^2$ is reached when the law of (X,Y) is Gaussian and $\mathrm{Cov}(X, Y)$ is of the form (3.3). Moreover, the following lemma shows that if $\mathrm{Cov}(X, Y)$ is of the form (3.3), then the law of (X,Y) is necessarily Gaussian and (X,Y) is in general suboptimal for the left-hand term in (2.1) (the proof is postponed to Appendix A).

Lemma 3.4. *Suppose that $n \leq m$. Let $X \sim \mathcal{N}(0, D_0)$ and $Y \sim \mathcal{N}(0, D_1)$ be two Gaussian vectors of respective size m and n with diagonal covariance matrices. If* $\mathrm{Cov}(X, Y)$ *is of the form*

$$\mathrm{Cov}(X, Y) = \begin{pmatrix} \tilde{I}_n \big(D_0^{(n)}\big)^{1/2} D_1^{1/2} \\ 0_{m-n,n} \end{pmatrix},$$

then $Y = \begin{pmatrix} \tilde{I}_n D_1^{1/2} \big(D_0^{(n)}\big)^{-1/2} & 0_{n,m-n} \end{pmatrix} X$, *and so (X,Y) is a Gaussian vector and* $\sum_{i,j} \mathrm{Cov}(X_i^2, Y_j^2) = 2\mathrm{tr}(D_0^{(n)} D_1)$, *where* $X = (X_1, X_2, \ldots, X_m)^\top$ *and* $Y = (Y_1, Y_2, \ldots, Y_n)^\top$.

Thus, there does not exist a plan $\pi \in \Pi(T_0\#\mu, T_1\#\mu_1)$ with co-moments of order 2 of the form (3.3) and with symmetric co-moments of order 4 of the form (3.4), since the former requires π to be Gaussian and the latter requires π not to be Gaussian. However, the solutions exhibited in Lemmas 3.2 and 3.3 are not unique, so a plan π may exist which is optimal for both terms of (2.1) but with co-moments of order 2 of a different form than (3.3) or/and with symmetric co-moments of order 4 of a different form than (3.4). Thus, we cannot conclude whether $GW_2(\mu, \nu) = LGW_2(\mu, \nu)$ or $GW_2(\mu, \nu) > LGW_2(\mu, \nu)$.

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:

$$GGW_2^2(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu) \cap \mathcal{N}_{m+n}} \int \int \left(\|x - x'\|_{\mathbb{R}^m}^2 - \|y - y'\|_{\mathbb{R}^n}^2 \right)^2 \mathrm{d}\pi(x, y)\, \mathrm{d}\pi(x', y'), \quad (4.1)$$

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

Since GGW_2 is the Gromov–Wasserstein problem restricted to Gaussian transport plans, it is clear that (4.1) is an upper bound of (1.4). Combining this result with Theorem 3.1, we get the following simple but important proposition.

Proposition 4.1. *If $\mu = \mathcal{N}(m_0, \Sigma_0)$ and $\nu = \mathcal{N}(m_1, \Sigma_1)$, with Σ_0 non-singular, then* $LGW_2^2(\mu, \nu) \leq GW_2^2(\mu, \nu) \leq GGW_2^2(\mu, \nu)$.

We exhibit in the following a solution of (4.1) that yields an explicit form for the upper bound $GGW_2^2(\mu, \nu)$.

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 $\Sigma_0 (= P_0 D_0 P_0^\top)$ and $\Sigma_1 (= P_1 D_1 P_1^\top)$ which sort eigenvalues in decreasing order. We suppose that Σ_0 is non-singular (μ is not degenerate). Then, problem (4.1) admits a solution of the form $\pi^* = (I_m, T)\#\mu$ with T affine of the form, for all $x \in \mathbb{R}^m$,*

$$T(x) = m_1 + P_1 A P_0^\top (x - m_0), \quad (4.2)$$

https://doi.org/10.1017/jpr.2022.16 Published online by Cambridge University Press

where $A \in \mathbb{R}^{n\times m}$ *is written as* $A = \begin{pmatrix} \tilde{I}_n D_1^{1/2} (D_0^{(n)})^{-1/2} & 0_{n,m-n} \end{pmatrix}$, *where* $\tilde{I}_n$ *is of the form* $\mathrm{diag}((\pm 1)_{i\leq n})$. *Moreover,*

$$GGW_2^2(\mu, \nu) = 4(\mathrm{tr}(D_0) - \mathrm{tr}(D_1))^2 + 8\|D_0^{(n)} - D_1\|_{\mathcal{F}}^2 + 8\big(\|D_0\|_{\mathcal{F}}^2 - \|D_0^{(n)}\|_{\mathcal{F}}^2\big). \tag{4.3}$$

Proof. Since the problem is restricted to Gaussian plans, the left-hand term in (2.1) can be rewritten as $2\|\mathrm{Cov}(X, Y)\|_{\mathcal{F}}^2$ thanks to Lemma 3.1 (Isserlis), and so problem (2.1) in that case becomes $\sup_{X\sim T_0\#\mu, Y\sim T_1\#\nu} 4\|\mathrm{Cov}(X, Y)\|_{\mathcal{F}}^2$. Applying Lemma 3.2, we can exhibit a Gaussian optimal plan $\tilde{\pi}^* \in \Pi(T_0\#\mu, T_1\#\nu)$ with covariance matrix Σ of the form

$$\Sigma = \begin{pmatrix} D_0 & K^* \\ K^{*\top} & D_1 \end{pmatrix},$$

with

$$K^* = \begin{pmatrix} \tilde{I}_n \big(D_0^{(n)}\big)^{1/2} D_1^{1/2} \\ 0_{m-n,n} \end{pmatrix}.$$

Thus, applying Lemma 3.4, it follows directly that $\tilde{\pi}^*$ is of the form $(I_m, \tilde{T})\#T_0\#\mu$, with $\tilde{T}$ linear of the form $\tilde{T}(x) = Ax$ for all $x \in \mathbb{R}^m$, with A of the form $A = \begin{pmatrix} \tilde{I}_n D_1^{1/2} \big(D_0^{(n)}\big)^{-1/2} & 0_{n,m-n} \end{pmatrix}$. Then, we can deduce the form of the optimal Gaussian plan $\pi^* \in \Pi(\mu, \nu)$:

$$\begin{aligned} \pi^* &= (T_0^{-1}, T_1^{-1})\#\tilde{\pi}^* \\ &= (T_0^{-1}, T_1^{-1})\#(I_m, \tilde{T})\#T_0\#\mu \\ &= (I_m, T_1^{-1}\#\tilde{T}\#T_0)\#\mu = (I_m, T)\#\mu, \end{aligned}$$

where T is affine of the form, for all $x \in \mathbb{R}^m$, $T(x) = T_1^{-1} \circ \tilde{T} \circ T_0(x) = m_1 + P_1 A P_0^\top (x - m_0)$. Moreover, using Lemmas 2.2 and 3.2, it follows that

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

□

4.1. Link with Gromov–Monge

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

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

Indeed, solutions of this equation 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.

https://doi.org/10.1017/jpr.2022.16 Published online by Cambridge University Press

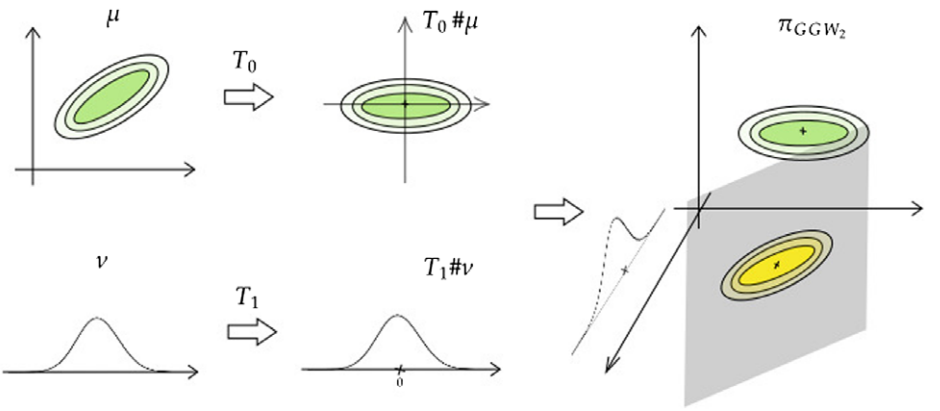

FIGURE 1. Transport plan π_{GGW_2} solution of problem (4.1) with $m = 2$ and $n = 1$. In this case, π_{GGW_2} is the degenerate Gaussian distribution supported by the affine plane of equation $y = T_{W_2}(x)$, where T_{W_2} is the classic W_2 optimal transport map when the distributions are rotated and centered first.

4.2. Link with PCA

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, writing $P_n : \mathbb{R}^m \to \mathbb{R}^n$ as the linear mapping which, for $x \in \mathbb{R}^m$, keeps only its n first components (n-frame), and T_{W_2} as 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 follows that the optimal plan π_{GGW_2} which achieves $GGW_2(\tilde{\mu}, \tilde{\nu})$ can be written as $\pi_{GGW_2} = (I_m, \tilde{I}_n\#T_{W_2}\#P_n)\#\tilde{\mu}$, where, with an abuse of notation, we write I_m (resp. $\tilde{I}_n$) as the map on $\mathbb{R}^m$ (resp. $\mathbb{R}^n$) represented by the similarly denoted matrix. An example of π_{GGW_2} can be found in Figure 1 when $m = 2$ and $n = 1$.

4.3. 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 derived simply by applying the transformations T_0 and T_1 to μ and ν respectively, 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 as $\pi_{GGW_2} = (I_m, T_1^{-1}\#\tilde{I}_n\#T_{W_2}\#T_0)\#\mu$. An example of transport between two Gaussian measures in dimension 2 is shown in Figure 2.

As illustrated in Figure 3, the GGW_2 optimal transport map T_{GGW_2} defined in (4.2) is not equivalent to the W_2 optimal transport map T_{W_2} defined in (1.3), even when the dimensions m and n are equal. More precisely, if 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

https://doi.org/10.1017/jpr.2022.16 Published online by Cambridge University Press

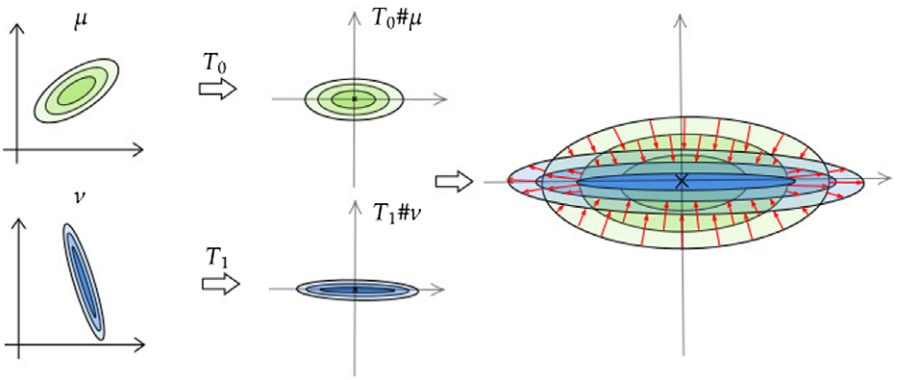

FIGURE 2. Solution of (4.1) between two Gaussians measures in dimension 2. First the distributions are centered and rotated. Then a classic W_2 transport is applied between the two aligned distributions.

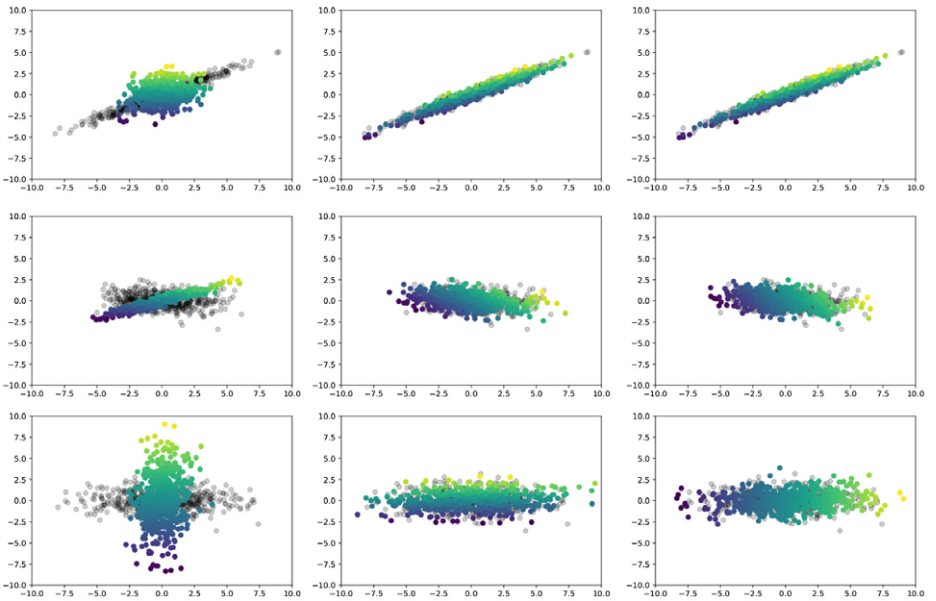

FIGURE 3. Comparison between W_2 and GGW_2 mappings between empirical distributions. Left: Two-dimensional source distribution (colored) and target distribution (transparent). Middle: Resulting mapping of Wasserstein T_{W_2}. Right: Resulting mapping of Gaussian Gromov–Wasserstein T_{GGW_2}. The colors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

different behaviors (bottom of Figure 3). Between those two extreme cases, we can say that the closer the columns of P_0 are to being collinear with 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).

https://doi.org/10.1017/jpr.2022.16 Published online by Cambridge University Press

4.4. Link with Gromov–Wasserstein with inner product as cost function

If μ and ν are centered Gaussian measures, let us consider the problem

$$GW_2^2(\langle\cdot,\cdot\rangle_m, \langle\cdot,\cdot\rangle_n, \mu, \nu) = \inf_{\pi\in\Pi(\mu,\nu)} \int\int (\langle x, x'\rangle_m - \langle y, y'\rangle_n)^2 \,\mathrm{d}\pi(x, y)\,\mathrm{d}\pi(x', y'). \tag{4.4}$$

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

Proposition 4.2. *Suppose $m \le n$. 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 (4.1) exhibited in Theorem 4.1 is also a solution of problem (4.4).*

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 follows that problem (4.4) is equivalent to $\sup_{\pi\in\Pi(\mu,\nu)} \left\| \int xy^\top \,\mathrm{d}\pi(x, y)\right\|_{\mathcal{F}}^2$. Since μ and ν are centered, it follows that problem (4.4) is equivalent to $\sup_{X\sim\mu, Y\sim\nu} \|\mathrm{Cov}(X, Y)\|_{\mathcal{F}}^2$. Applying Lemma 3.2, it follows that the solution exhibited in Theorem 4.1 is also a solution of problem (4.4). □

5. Tightness of the bounds and particular cases

5.1. Bound on the difference

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

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

To prove this proposition we will use the following technical result (the proof is postponed to Appendix A).

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 $u^\top v \ge 1/\sqrt{m}$, with equality if $u = (1/\sqrt{m}, 1/\sqrt{m}, \dots)^\top$ and $v = (1, 0, \dots)^\top$.*

Proof of Proposition 5.1. By subtracting (3.1) from (4.3), it follows that

$$\begin{aligned} GGW_2^2(\mu, \nu) - LGW_2^2(\mu, \nu) &= 8\big(\|D_0\|_{\mathcal{F}}\|D_1\|_{\mathcal{F}} - \mathrm{tr}(D_0^{(n)}D_1)\big) \\ &= 8\big(\|D_0\|_{\mathcal{F}}\|D_1^{[m]}\|_{\mathcal{F}} - \mathrm{tr}(D_0D_1^{[m]})\big), \end{aligned} \tag{5.1}$$

where $D_1^{[m]} = \left(\begin{smallmatrix} D_1 & 0 \\ 0 & 0 \end{smallmatrix}\right) \in \mathbb{R}^{m\times m}$. Writing $\alpha \in \mathbb{R}^m$ and $\beta \in \mathbb{R}^m$ as the vectors of eigenvalues of D_0 and $D_1^{[m]}$, it follows that $GGW_2^2(\mu, \nu) - LGW_2^2(\mu, \nu) = 8\big(\|\alpha\|\|\beta\| - \alpha^\top\beta\big) = 8\|\alpha\|\|\beta\|(1 - u^\top v)$, where $u = \alpha/\|\alpha\|$ and $v = \beta/\|\beta\|$. Applying Lemma 5.1, we get directly that

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

□

https://doi.org/10.1017/jpr.2022.16 Published online by Cambridge University Press

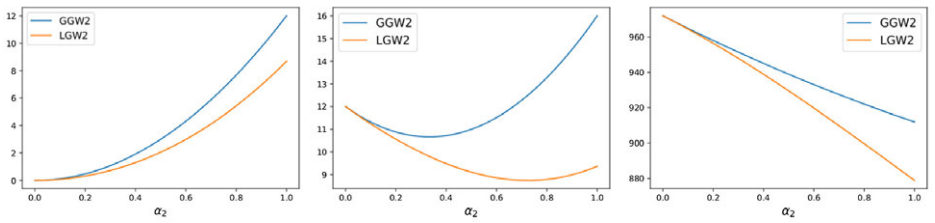

FIGURE 4. Plot of $GGW_2^2(\mu, \nu)$ and $LGW_2^2(\mu, \nu)$ as a function of α_2 for $\mu = \mathcal{N}(0, \mathrm{diag}(\alpha))$, $\nu = \mathcal{N}(0, \beta_1)$, $\alpha = (\alpha_1, \alpha_2)^\top$, for $(\alpha_1, \beta_1) = (1, 1)$ (left), $(\alpha_1, \beta_1) = (1, 2)$ (middle), and $(\alpha_1, \beta_1) = (1, 10)$ (right). We can easily compute using (4.3) and (3.1) that $GGW_2^2(\mu, \nu) = 12\alpha_2^2 + 8\alpha_2(\alpha_1 - \beta_1) + 12(\alpha_1 - \beta_1)^2$ and $LGW_2^2(\mu, \nu) = 12\alpha_2^2 + 8\alpha_2(\alpha_1 - \beta_1) - 4\left[\alpha_2^2 + \alpha_1^2\right]^{1/2}\beta_1 + 12(\alpha_1 - \beta_1)^2 + 8\alpha_1\beta_1$.

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 being proportional to the eigenvalues of D_1. An example in the case where $m = 2$ and $n = 1$ can be found in Figure 4.

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 $\nu = \mathcal{N}(m_1, \Sigma_1)$ be 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 Σ_0 $(= P_0 D_0 P_0^\top)$ and Σ_1 $(= P_1 D_1 P_1^\top)$ which sort the eigenvalues in non-increasing order. Suppose Σ_0 is non-singular and that a scalar $\lambda \geq 0$ exists 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, for all $x \in \mathbb{R}^m$,*

$$T(x) = m_1 + \sqrt{\lambda} P_1 \tilde{I}_m P_0^\top (x - m_0), \tag{5.2}$$

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

$$GW_2^2(\mu, \nu) = (\lambda - 1)^2 \left(4(\mathrm{tr}(\Sigma_0))^2 + 8\|\Sigma_0\|_{\mathcal{F}}^2\right). \tag{5.3}$$

Proof. From (5.1), $GGW_2^2(\mu, \nu) - LGW2^2(\mu, \nu) = 8(\|D_0\|_{\mathcal{F}}\|D_1\|_{\mathcal{F}} - \mathrm{tr}(D_0 D_1))$. Writing $\alpha \in \mathbb{R}^m$ and $\beta \in \mathbb{R}^m$ for the eigenvalue vectors of D_0 and D_1, it follows that

https://doi.org/10.1017/jpr.2022.16 Published online by Cambridge University Press

$GGW_2^2(\mu,\nu) - LGW2^2(\mu,\nu) = 8(\|\alpha\|\|\beta\| - \alpha^\top\beta)$. Since $\lambda \geq 0$ exists such that $D_1 = \lambda D_0$, we have $\beta = \lambda\alpha$, and so $\alpha^\top\beta = \|\alpha\|\|\beta\|$. Thus, $GGW_2^2(\mu,\nu) - LGW_2^2(\mu,\nu) = 0$ and, using Proposition 4.1, we get that $GW_2^2(\mu,\nu) = GGW_2^2(\mu,\nu) = LGW_2^2(\mu,\nu)$. We get (5.2) and (5.3) by simply reinjecting in (4.2) and (4.3). □

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 $GW_2^2(\mu,\nu) = 12\left(\sigma_0^2 - \sigma_1^2\right)^2$, and the optimal transport plan π^* has the form $(I_1, T)\#\mu$, with T affine of the form, for all $x \in \mathbb{R}$, $T(x) = m_1 \pm (\sigma_1/\sigma_0)(x - m_0)$. Thus, the solution of $W_2^2(\mu,\nu)$ is also a solution of $GW_2^2(\mu,\nu)$.*

5.3. Case of degenerate measures

In all the results described above, we have supposed Σ_0 to be non-singular, which means that μ is not degenerate. Yet, if Σ_0 is not full rank, we 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 by $r = \mathrm{rk}(D_0)$ 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 by $\Gamma_r(\mathbb{R}^m)$ the set of vectors $x = (x_1, \ldots, x_m)^\top$ of $\mathbb{R}^m$ such that $x_{r+1} = \cdots = x_m = 0$. For $\pi \in \Pi(\mu,\nu)$, we can remark that for any Borel set $A \subset \mathbb{R}^m \setminus \Gamma_r(\mathbb{R}^m)$, and any Borel set $B \subset \mathbb{R}^n$, we have $\pi(A, B) = 0$ and so

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

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

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

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

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

https://doi.org/10.1017/jpr.2022.16 Published online by Cambridge University Press

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

The exact same reasoning can be made in the case of GGW_2. Moreover, it can be easily seen when looking at (3.1) that $LGW_2^2(\mu,\nu) = LGW_2^2(P_r\#\mu,\nu)$. □

Thus, when Σ_0 is not full rank, we 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 4.1 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 by 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 Σ_0 ($= P_0 D_0 P_0^\top$) and Σ_1 ($= P_1 D_1 P_1^\top$) 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, for all $x \in \mathbb{R}^m$, $T(x) = m_1 + P_1 A P_0^\top (x - m_0)$, where $A \in \mathbb{R}^{n\times m}$ is of the form

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

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

6. Behavior of the empirical solution

In this section 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, $\mu = \mathcal{N}(0, \mathrm{diag}(\alpha))$ and $\nu = \mathcal{N}(0, \mathrm{diag}(\beta))$, 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 (accessible at https://pythonot.github.io/index.html). In Figure 5, we plot the first coordinates of the samples Y_i as a function of the the first coordinate of the samples X_j they have been assigned to by the algorithm (blue dots). We also draw the line $y = \pm\sqrt{\beta}x$ to compare with the theorical solution of the Gaussian restricted problem (orange line) for $k = 2000$, $\alpha = (1, 0.1)^\top$, and $\beta = 2$ (top left); $k = 2000$, $\alpha = (1, 0.1)^\top$, and $\beta = (2, 0.3)^\top$ (top right); $k = 2000$, $\alpha = (1, 0.1, 0.01)^\top$, and $\beta = 2$ (middle left); $k = 7000$, $\alpha = (1, 0.3)$, and $\beta = 2$ (middle right); $k = 7000$, $\alpha = (1, 0.1)^\top$, and $\beta = (2, 1)^\top$ (bottom left); and $k = 7000$, $\alpha = (1, 0.3, 0.1)$, and $\beta = 2$ (bottom right). Observe that the empirical solution seems to be behaving in exactly the same way as the theoretical solution exhibited in Theorem 4.1 as soon as α and β are close to being 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 whether the empirical solution does not behave in the same way as the theorical solution exhibited in Theorem 4.1 or whether 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

https://doi.org/10.1017/jpr.2022.16 Published online by Cambridge University Press

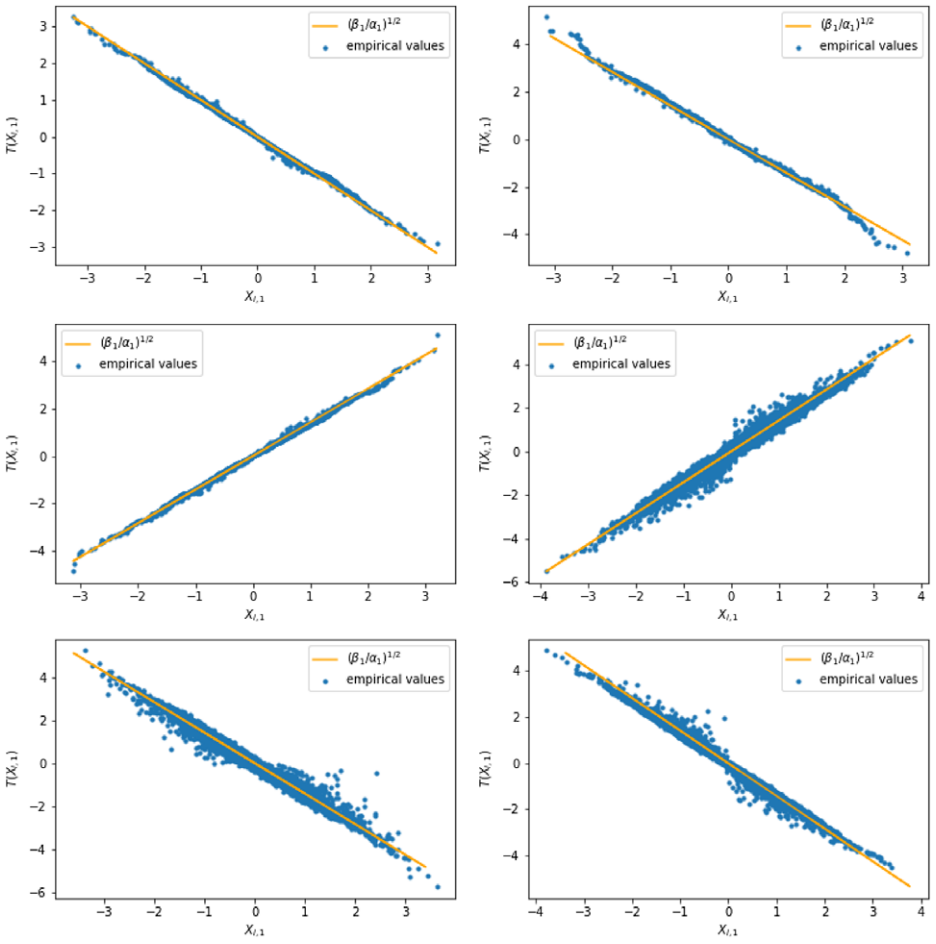

FIGURE 5. Plot of the first coordinate of samples Y_i as a function of the the first coordinate of their assigned samples X_j (blue dots) and the line $y = \pm\sqrt{\beta}x$ (orange line) for $k = 2000$, $\alpha = (1, 0.1)^\top$, and $\beta = 2$ (top left); $k = 2000$, $\alpha = (1, 0.1)^\top$, and $\beta = (2, 0.3)^\top$ (top right); $k = 2000$, $\alpha = (1, 0.1, 0.01)^\top$, and $\beta = 2$ (middle left); $k = 7000$, $\alpha = (1, 0.3)$, and $\beta = 2$ (middle right); $k = 7000$, $\alpha = (1, 0.1)^\top$, and $\beta = (2, 1)^\top$ (bottom left); and $k = 7000$, $\alpha = (1, 0.3, 0.1)$, and $\beta = 2$ (bottom right).

the optimal plan which achieves $GGW_2(\mu, \nu)$ is also a solution of the non-restricted problem $GW_2(\mu, \nu)$, and that $GW_2(\mu, \nu) = GGW_2(\mu, \nu)$.

7. Conclusion

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

https://doi.org/10.1017/jpr.2022.16 Published online by Cambridge University Press

that the linear solution exhibited when adding this restriction might also be the solution of the unconstrained problem.

Appendix A. Proofs of the lemmas

A.1. Proof of Lemma 3.2

Proof. The proof is inspired from the proof of (1.2) as provided in [13]. We want to maximize $\mathrm{tr}(K^\top K)$ with the constraint that Σ is semi-definite positive. Let $S = \Sigma_1 - K^\top \Sigma_0^{-1} K$ (Schur complement); problem (3.2) can be written as $\min_{S \in S_n^+(\mathbb{R})} -\mathrm{tr}(K^\top K)$. For a given S, the set of feasible K is the set of K such that $K^\top \Sigma_0^{-1} K = \Sigma_1 - S$. Since $\Sigma_0 \in S_m^{++}(\mathbb{R})$, $K^\top \Sigma_0^{-1} K \in S_n^+(\mathbb{R})$ and so $\Sigma_1 - S \in S_n^+(\mathbb{R})$. We denote by r the rank of $K^\top \Sigma_0^{-1} K$. We can observe that $r \le n \le m$, where the left-hand inequality follows from the fact that $\mathrm{rk}(AB) \le \min\{\mathrm{rk}(A), \mathrm{rk}(B)\}$. Then, $\Sigma_1 - S$ can be diagonalized,

$$\Sigma_1 - S = K^\top \Sigma_0^{-1} K = U \Lambda^2 U^\top = U_r \Lambda_r^2 U_r^\top, \tag{A.1}$$

with $\Lambda^2 = \mathrm{diag}(\lambda_1^2, \dots, \lambda_r^2, 0, \dots, 0)$, $\Lambda_r^2 = \mathrm{diag}(\lambda_1^2, \dots, \lambda_r^2)$, and $U_r \in \mathbb{V}_r(\mathbb{R}^n) := \{M \in \mathbb{R}^{n \times r} \mid M^\top M = I_r\}$ (Stiefel manifold [15]) such that $U = \begin{pmatrix} U_r & U_{n-r} \end{pmatrix}$. From (A.1), we can deduce that $(\Sigma_0^{-1/2} K U_r \Lambda_r^{-1})^\top \Sigma_0^{-1/2} K U_r \Lambda_r^{-1} = I_r$. We can set $B_r = \Sigma_0^{-1/2} K U_r \Lambda_r^{-1}$ such that $B_r \in \mathbb{V}_r(\mathbb{R}^m)$, and deduce that $K U_r = \Sigma_0^{1/2} B_r \Lambda_r$. Moreover, since $U_{m-r}^\top K^\top \Sigma_0^{-1} K U_{m-r} = 0$ and $\Sigma_0 \in S_m^{++}(\mathbb{R})$, it follows that $K U_{n-r} = 0$ and so

$$K = K U U^\top = K U_r U_r^\top = \Sigma_0^{\frac{1}{2}} B_r \Lambda_r U_r^\top. \tag{A.2}$$

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

$$\mathrm{tr}(K^\top K) = \mathrm{tr}(U_r \Lambda_r B_r^\top \Sigma_0 B_r \Lambda_r U_r^\top) = \mathrm{tr}(U_r^\top U_r \Lambda_r B_r^\top \Sigma_0 B_r \Lambda_r) = \mathrm{tr}(\Lambda_r^2 B_r^\top \Sigma_0 B_r).$$

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

$$\min_{B_r \in \mathbb{V}_r(\mathbb{R}^m)} -\mathrm{tr}\left(\Lambda_r^2 B_r^\top \Sigma_0 B_r\right). \tag{A.3}$$

The following is an adaptation of the proof of [2, Proposition (3.1)] for when B_r is not a squared matrix. The Lagrangian of problem (A.3) can be written as $\mathcal{L}(B_r, C) = -\mathrm{tr}(\Lambda_r^2 B_r^\top \Sigma_0 B_r) + \mathrm{tr}(C(B_r^\top B_r - I_r))$, where $C \in S_r(\mathbb{R})$ is the Lagrange multiplier associated with the constraint $B_r^\top B_r = I_r$ (C is symmetric because $B_r^\top B_r - I_r$ is symmetric). We can then derive the first-order condition $-2\Sigma_0 B_r \Lambda_r^2 + 2 B_r C = 0$, or, equivalently, $\Sigma_0 B_r \Lambda_r^2 B_r^\top = B_r C B_r^\top$. Since $C \in S_r(\mathbb{R})$, $B_r C B_r^\top \in S_m(\mathbb{R})$ and $\Sigma_0 B_r \Lambda_r^2 B_r^\top \in S_m(\mathbb{R})$. We can deduce that Σ_0 and $B_r \Lambda_r B_r^\top$ commute. Moreover, since Σ_0 and $B_r \Lambda_r^2 B_r^\top$ 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 $B_r \Lambda_r^2 B_r^\top = B \Lambda_m^2 B^\top$, where $\Lambda_m^2 \in \mathbb{R}^{m \times m} = \mathrm{diag}(\lambda_1^2, \dots, \lambda_r^2, 0, \dots, 0)$ and $B = \begin{pmatrix} B_r & B_{m-r} \end{pmatrix}$. Thus, the eigenvalues of $B_r \Lambda_r^2 B_r^\top$ are exactly the eigenvalues of Λ_m^2. Since Σ_0 and $B_r \Lambda_r^2 B_r^\top$ can be diagonalized in the same basis, we get that $\mathrm{tr}(\Lambda_r^2 B_r^\top \Sigma_0 B_r) = \mathrm{tr}(\Sigma_0 B_r \Lambda_r^2 B_r^\top) = \mathrm{tr}(D_0 \tilde{\Lambda}_m)$, where $\tilde{\Lambda}_m$ is a diagonal matrix with the same eigenvalues as Λ_m, but in a different order. Now, it can easily be seen that the optimal value of (A.3) is reached when B_r is a permutation matrix which sorts the eigenvalues of Λ_m in decreasing order.

https://doi.org/10.1017/jpr.2022.16 Published online by Cambridge University Press

Thus, for a given S, the maximum value of $\operatorname{tr}(K^\top K)$ is $\operatorname{tr}(D_0\tilde{\Lambda}_m(S))$. We can now establish for which S, $\operatorname{tr}(D_0\tilde{\Lambda}_m(S))$ is optimal. For a given S, we denote by $\lambda_1, \dots, \lambda_n$ the eigenvalues of $\Sigma_1 - S$ and by $\beta_1, \dots, \beta_n$ the eigenvalues of Σ_1 ordered in decreasing order. Since $S \in S_n^+(\mathbb{R})$, for all $x \in \mathbb{R}^n$, the following inequality holds: $x^\top(\Sigma_1 - S)x \le x^\top \Sigma_1 x$. This inequality still holds when restricted to any subspace of $\mathbb{R}^n$. Using the Courant–Fischer theorem, we can conclude that, for all $i \le n$, $\lambda_i \le \beta_i$. Thus, the optimal value of $\operatorname{tr}(D_0\tilde{\Lambda}_m(S))$ is reached when $S = 0$ and $\tilde{\Lambda}_m(0) = \left(\begin{smallmatrix} D_1 & 0 \\ 0 & 0 \end{smallmatrix}\right)$, and so $\operatorname{tr}(D_0\tilde{\Lambda}_m(0)) = \operatorname{tr}(D_0^{(n)}D_1)$. Let

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

with $\tilde{I}_n$ of the form $\operatorname{diag}((\pm 1)_{i \le n})$. It can be easily verified that $A^\top D_0^{-1} A = D_1$ and, if $K^* = P_0^\top A P_1$, $K^{*\top}\Sigma_0^{-1}K^* = P_1^\top A^\top P_0 \Sigma_0^{-1} P_0^\top A P_1 = P_1^\top A^\top D_0^{-1} A P_1 = P_1^\top D_1 P_1 = \Sigma_1$, and $K^{*\top}K^*$ has the same eigenvalues as $A^\top A$ and $\operatorname{tr}(A^\top A) = \operatorname{tr}(D_0^{(n)}D_1)$. □

A.2. Proof of Lemma 3.3

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

Lemma A.1. *(Anstreicher and Wolkowicz, 1998 [2])Let Σ_0 and Σ_1 be two symmetric matrices of size n. We denote by $\Sigma_0 = P_0\Lambda_0 P_0^\top$ and $\Sigma_1 = P_1\Lambda_1 P_1^\top$ their 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 $\min_{PP^\top = I_n} \operatorname{tr}\left(\Sigma_0 P \Sigma_1 P^\top\right) = \operatorname{tr}(\Lambda_0\Lambda_1)$, and it is achieved for $P^* = P_0 P_1^\top$.*

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 $S = \Sigma_1 - K^\top\Sigma_0^{-1}K$, 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), up until (A.2). We diagonalize $\Sigma_1 - S = K^\top\Sigma_0^{-1}K = U_r\Lambda_r U_r^\top$, where r is the rank of $K^\top\Sigma_0^{-1}K$, then we set $B_r = \Sigma_0^{-1/2}KU_r\Lambda_r^{-1}$ while observing that $B_r \in \mathbb{V}_r(\mathbb{R}^m)$, and we deduce that $K = \Sigma_0^{1/2}B_r\Lambda_r U_r^\top$. By reinjecting this expression, it follows that $\operatorname{tr}(KA) = \operatorname{tr}\left(A^\top K^\top\right) = \operatorname{tr}\left(A^\top U_r\Lambda_r B_r^\top\Sigma_0^{1/2}\right) = \operatorname{tr}\left(\Sigma_0^{1/2}A^\top U_r\Lambda_r B_r^\top\right)$. For a given S, the problem of finding the optimal value is parametrized by B_r and is $\min_{B_r \in \mathbb{V}_r(\mathbb{R}^m)} -\operatorname{tr}\left(\Sigma_0^{1/2}A^\top U_r\Lambda_r B_r^\top\right)$. The Lagrangian of this problem can be written as $\mathcal{L}(B_r, C) = -\operatorname{tr}\left(\Sigma_0^{1/2}A^\top U_r\Lambda_r B_r^\top\right) + \operatorname{tr}\left(C\left(B_r^\top B_r - I_r\right)\right)$, where $C \in S_r(\mathbb{R})$ is the Lagrangian multiplier associated with the constraint $B_r^\top B_r = I_r$. We can then derive the first-order condition, $-\Sigma_0^{1/2}A^\top U_r\Lambda_r + 2B_rC = 0$, or, equivalently, $\Sigma_0^{1/2}A^\top U_r\Lambda_r B_r^\top = 2B_rCB_r^\top$. Since $C \in S_r(\mathbb{R})$, $2B_rCB_r^\top \in S_m(\mathbb{R})$ and $\Sigma_0^{1/2}A^\top U_r\Lambda_r B_r^\top \in S_m(\mathbb{R})$. Moreover, the rank of $\Sigma_0^{1/2}A^\top U_r\Lambda_r B_r^\top$ is equal to 1 because $\operatorname{rk}(A) = 1$ and $\operatorname{rk}\left(\Sigma_0^{1/2}A^\top U_r\Lambda_r B_r^\top\right) = 0$ would imply that $\operatorname{tr}(KA) = 0$, which cannot be the maximum value of our problem. So there exists a vector $u_m \in \mathbb{R}^m$ such that $\Sigma_0^{1/2}A^\top U_r\Lambda_r B_r^\top = u_m u_m^\top$. Then we can reinject the value B_r in the expression:

$$\begin{aligned} \Sigma_0^{1/2}A^\top U_r\Lambda_r B_r^\top &= \Sigma_0^{1/2}A^\top U_r\Lambda_r\Lambda_r^{-1}U_r^\top K^\top\Sigma_0^{-1/2} \\ &= \Sigma_0^{1/2}A^\top U_r U_r^\top K^\top\Sigma_0^{-1/2} \\ &= \Sigma_0^{1/2}A^\top K^\top\Sigma_0^{-1/2}, \end{aligned}$$

where we used the fact that $K = KUU^\top = KU_rU_r^\top$ because $KU_{n-r} = 0$.

https://doi.org/10.1017/jpr.2022.16 Published online by Cambridge University Press

We therefore have, on one hand, $\mathrm{tr}(KA) = \mathrm{tr}\big(\Sigma_0^{1/2} A^\top K^\top \Sigma_0^{-1/2}\big) = \mathrm{tr}\big(u_m u_m^\top\big) = u_m^\top u_m$, and on the other hand,

$$\begin{aligned}\Sigma_0^{1/2} A^\top K^\top \Sigma_0^{-1/2}\big(\Sigma_0^{1/2} A^\top K^\top \Sigma_0^{-1/2}\big)^\top &= \Sigma_0^{1/2} A^\top K^\top \Sigma_0^{-1} K A D_0^{1/2} \\ &= \Sigma_0^{1/2} A^\top (\Sigma_1 - S) A \Sigma_0^{1/2} \\ &= u_m u_m^\top u_m u_m^\top = u_m^\top u_m u_m u_m^\top,\end{aligned}$$

and thus $\mathrm{tr}\big(\Sigma_0^{1/2} A^\top (\Sigma_1 - S) A \Sigma_0^{1/2}\big) = u_m^\top u_m \mathrm{tr}(u_m u_m^\top) = (u_m^\top u_m)^2 = (\mathrm{tr}(KA))^2$. Then, we can determine for which S, $\mathrm{tr}\big(\Sigma_0^{1/2} A^\top (\Sigma_1 - S) A \Sigma_0^{1/2}\big)$ is maximum:

$$\mathrm{tr}\big(\Sigma_0^{1/2} A^\top (\Sigma_1 - S) A \Sigma_0^{1/2}\big) = \mathrm{tr}\big(A \Sigma_0 A^\top (\Sigma_1 - S)\big) = \mathrm{tr}(A \Sigma_0 A^\top \Sigma_1) - \mathrm{tr}(A \Sigma_0 A^\top S).$$

Let $B = A \Sigma_0 A^\top$. 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 this as $S = PDP^\top$. As before, we first determine the value of $\mathrm{tr}(BS)$ for a given D, and then we determine which D minimizes $\mathrm{tr}(BS)$. For a given D, we want the optimal value of $\min_{PP^\top = I_b} \mathrm{tr}(BPDP^\top)$.

Since B is symmetric with rank 1, it has only one non null eigenvalue which is equal to its trace. Using Lemma A.1, we can deduce that $\min_{PP^\top = I_n} \mathrm{tr}(BPDP^\top) = \mathrm{tr}(B)\lambda_n$, 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 easily be seen that $\mathrm{tr}(\mathbf{1}_{n,m} \Sigma_0 \mathbf{1}_{m,n} \Sigma_1) = \mathrm{tr}(\Sigma_0)\mathrm{tr}(\Sigma_1)$. Thus, if

$$K = \frac{\alpha \beta^\top}{\big[\mathrm{tr}(\Sigma_0)\mathrm{tr}(\Sigma_1)\big]^{1/2}} = \frac{\Sigma_0 \mathbf{1}_{m,n} \Sigma_1}{\big[\mathrm{tr}(\Sigma_0)\mathrm{tr}(\Sigma_1)\big]^{1/2}},$$

we can observe that

$$\mathrm{tr}(K \mathbf{1}_{n,m}) = \mathrm{tr}(\mathbf{1}_{n,m} K) = \frac{\mathrm{tr}(\mathbf{1}_{n,m} \Sigma_0 \mathbf{1}_{m,n} \Sigma_1)}{\big[\mathrm{tr}(\Sigma_0)\mathrm{tr}(\Sigma_1)\big]^{1/2}} = \big[\mathrm{tr}(\Sigma_0)\mathrm{tr}(\Sigma_1)\big]^{1/2}.$$

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

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

Using the matrix determinant lemma, it follows that, for all $i \leq n$,

$$\det(S^{(i)}) = \prod_k^i \beta_k \left(1 - \frac{\mathrm{tr}(\Sigma_1^{(i)})}{\mathrm{tr}(\Sigma_1)}\right).$$

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

A.3. Proof of Lemma 3.4

Proof. for any $i \leq m$ and any $j \leq n$, the Cauchy–Schwarz inequality tells us that $|\mathrm{Cov}(X_i, Y_j)| \leq \big\{\mathbb{E}[X_i^2]\mathbb{E}[Y_j^2]\big\}^{1/2}$, with equality if and only if $Y_j = \lambda X_i$ with $\lambda \in \mathbb{R}$. If $\mathrm{Cov}(X, Y)$ is of the form

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

https://doi.org/10.1017/jpr.2022.16 Published online by Cambridge University Press

then, for all $i \le n$, $|\mathrm{Cov}(X_i, Y_i)| = \sqrt{\alpha_i \beta_i}$, where $\alpha_i = \mathbb{E}[X_i^2]$ and $\beta_i = \mathbb{E}[Y_i^2]$. Thus, for all $i \le n$, $Y_i = \lambda_i X_i$ with $\lambda_i \in \mathbb{R}$. Since $X_i \sim \mathcal{N}(0, \alpha_i)$ and $Y_i \sim \mathcal{N}(0, \beta_i)$, it follows that $\lambda_i = \pm\sqrt{\beta_i/\alpha_i}$ and that $Y = \big(\tilde{I}_n D_1^{1/2} (D_0^{(n)})^{-1/2} \quad 0_{n,m-n}\big) X$. Since Y depends linearly on X, it follows that (X,Y) is a Gaussian vector. Thus, using Isserlis' lemma, we can compute that

$$\mathrm{Cov}(X_i^2, Y_j^2) = \begin{cases} 2\alpha_i\beta_i & \text{if } i = j, \\ 0 & \text{otherwise,} \end{cases}$$

so that $\sum_{i,j} \mathrm{Cov}(X_i^2, Y_j^2) = 2\mathrm{tr}(D_0^{(n)} D_1)$. □

A.4. Proof of Lemma 5.1

Proof. For $m \ge 1$, let Γ_m denote the set of vectors $v = (v_1, \dots, v_m)$ of $\mathbb{R}^m$ such that $v_1 \ge v_2 \ge \cdots \ge v_m \ge 0$ and $\sum_{i=1}^m v_i^2 = 1$. We want to prove that, for all $u, v \in \Gamma_m$, $\sum_{i=1}^m u_i v_i \ge 1/\sqrt{m}$. We proceed by induction on m. For $m = 1$, it is obviously true since $\Gamma_1 = \{1\}$. Assume now that $m > 1$, and that the result is true for $m-1$. Let $u, v \in \Gamma_m$; then, using the result for $(u_2, \dots, u_m)/\big(\sum_{i=2}^m u_i^2\big)^{1/2}$ and $(v_2, \dots, v_m)/\big(\sum_{i=2}^m v_i^2\big)^{1/2}$ that both belong to Γ_{m-1}, we have

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

Now, since $u, v \in \Gamma_m$, we have $u_1, v_1 \in [1/\sqrt{m}, 1]$. Let us write $F(u_1, v_1) = u_1 v_1 + (m-1)^{-1/2} (1-u_1^2)^{1/2} (1-v_1^2)^{1/2}$. We have, for all $v_1 \in [1/\sqrt{m}, 1]$, $F(1, v_1) = v_1 \ge 1/\sqrt{m}$ and $F(1/\sqrt{m}, v_1) = \big(\sqrt{1-v_1^2} + v_1\big)/\sqrt{m} \ge (1 - v_1^2 + v_1)/\sqrt{m} \ge 1/\sqrt{m}$, and computing the partial derivative of F with respect to u_1, we get

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

This is a decreasing function of u_1, with value v_1 at $u_1 = 0$ and a 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 decreasing, showing that, for all $u_1 \in [1/\sqrt{m}, 1]$, $F(u_1, v_1) \ge \min\big(F(1/\sqrt{m}, v_1), F(1, v_1)\big) \ge 1/\sqrt{m}$. Finally, we have thus proved that $\sum_{i=1}^m u_i v_i \ge 1/\sqrt{m}$, and moreover the equality is achieved when the vectors u and v are the vectors $(1, 0, \dots, 0)$ and $(1/\sqrt{m}, 1/\sqrt{m}, \dots, 1/\sqrt{m})$. □

Funding information

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).

Competing interests

There were no competing interests to declare which arose during the preparation or publication process of this article.

https://doi.org/10.1017/jpr.2022.16 Published online by Cambridge University Press

References

[1] ALVAREZ-MELIS, D., JEGELKA, S. AND JAAKKOLA, T. S. (2019). Towards optimal transport with global invariances. *Proc. Mach. Learn. Res*. 89, 1870–1879.

[2] ANSTREICHER, K. AND WOLKOWICZ, H. (2000). On Lagrangian relaxation of quadratic matrix constraints. *J. Matrix Anal. Appl.* **22**, 41–55.

[3] ARJOVSKY, M., CHINTALA, S. AND BOTTOU, L. (2017). Wasserstein generative adversarial networks. *Proc. Mach. Learn. Res*. **70**, 214–223.

[4] BIGOT, J. *et al.* (2017). Geodesic PCA in the Wasserstein space by convex PCA. *Ann. Inst. H. Poincaré Prob. Statist.* **53**, 1–26.

[5] BLANCHET, J., KANG, Y. AND MURTHY, K. (2019). Robust Wasserstein profile inference and applications to machine learning. *J. Appl. Prob.* **56**, 830–857.

[6] BRENIER, Y. (1991). Polar factorization and monotone rearrangement of vector-valued functions. *Commun. Pure Appl. Math.* **44**, 375–417.

[7] CAI, Y. AND LIM, L.-H. (2020). Distances between probability distributions of different dimensions. Preprint, arXiv:2011.00629 [math.ST].

[8] CHOWDHURY, S. AND NEEDHAM, T. (2020). Gromov–Wasserstein averaging in a Riemannian framework. In *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition Workshops*, Curran Associates, Red Hook, NY, pp. 842–843.

[9] COURTY, N., FLAMARY, R., TUIA, D. AND RAKOTOMAMONJY, A. (2016). Optimal transport for domain adaptation. *IEEE Trans. Pattern Anal. Mach. Intellig.* **39**, 1853–1865.

[10] DOWSON, D. C. AND LANDAU, B. V. (1982). The Fréchet distance between multivariate normal distributions. *J. Multivar. Anal.* **12**, 450–455.

[11] GALICHON, A. *et al.* (2014). A stochastic control approach to no-arbitrage bounds given marginals, with an application to lookback options. *Ann. Appl. Prob.* **24**, 312–336.

[12] GENEVAY, A., PEYRÉ, G. AND CUTURI, M. (2018). Learning generative models with sinkhorn divergences. *Proc. Mach. Learn. Res*. 84, 1608–1617.

[13] GIVENS, C. R. *et al.* (1984). A class of Wasserstein metrics for probability distributions. *Mich. Math. J.* **31**, 231–240.

[14] ISSERLIS, L. (1918). On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables. *Biometrika* **12**, 134–139.

[15] JAMES, I. M. (1976). *The Topology of Stiefel Manifolds* (*London Math. Soc. Lecture Note Series* **24**). Cambridge University Press.

[16] MÉMOLI, F. (2011). Gromov–Wasserstein distances and the metric approach to object matching. *Found. Comput. Math.* **11**, 417–487.

[17] PELE, O. AND TASKAR, B. (2013). The tangent earth mover's distance. In *Proc. First Int. Conf. Geometric Science of Information*, eds F. Nielsen and F. Barbaresco. Springer, New York, pp. 397–404.

[18] PEYRÉ, G. AND CUTURI, M. (2019). Computational optimal transport, with applications to data science. *Found. Trends Mach. Learn.* **11**, 355–607.

[19] PEYRÉ, G., CUTURI, M. AND SOLOMON, J. (2016). Gromov–Wasserstein averaging of kernel and distance matrices. *J. Mach. Learn. Res.* **48**, 2664–2672.

[20] RABIN, J., FERRADANS, S. AND PAPADAKIS, N. (2014). Adaptive color transfer with relaxed optimal transport. In *Proc. IEEE Int. Conf. Image Processing*, pp. 4852–4856.

[21] RABIN, J., PEYRÉ, G., DELON, J. AND BERNOT, M. (2011). Wasserstein barycenter and its application to texture mixing. In *Proc. Third Int. Conf. Scale Space and Variational Methods in Computer Vision*, eds A. M. Bruckstein, B. M. Haar Romeny, A. M. Bronstein, and M. M. Bronstein. Springer, New York, pp. 435–446.

[22] SANTAMBROGIO, F. (2015). *Optimal Transport for Applied Mathematicians (Progress in Nonlinear Differential Equations and their Applications* **87**). Springer, New York.

[23] STURM, K.-T. (2012). The space of spaces: Curvature bounds and gradient flows on the space of metric measure spaces. Preprint, arXiv:1208.0434 [math.MG].

[24] TAKATSU, A. On Wasserstein geometry of Gaussian measures. In *Probabilistic Approach to Geometry*, eds M. Kotani, M. Hino, and T. Kumagai. Mathematical Society of Japan, Tokyo, pp. 463–472.

[25] VAYER, T. (2020). A contribution to optimal transport on incomparable spaces. Preprint, arXiv:2011.04447 [stat.ML].

[26] VILLANI, C. (2003). *Topics in Optimal Transportation* (Graduate Studies in Math. **58**). American Mathematical Society, Providence, RI.

[27] VILLANI, C. (2008). *Optimal Transport: Old and New (Grundlehren der mathematischen Wissenschaften* **338**). Springer, Berlin.

https://doi.org/10.1017/jpr.2022.16 Published online by Cambridge University Press