

Article

Subspace Detours Meet Gromov–Wasserstein

Clément Bonet ^{1,*}, Titouan Vayer ², Nicolas Courty ³, François Septier ¹ and Lucas Drumetz ⁴

¹ Laboratoire de Mathématiques de Bretagne Atlantique, Université Bretagne Sud, CNRS UMR 6205, 56000 Vannes, France; francois.septier@univ-ubs.fr

² ENS Lyon, CNRS UMR 5668, LIP, 69342 Lyon, France; titouan.vayer@ens-lyon.fr

³ Department of Computer Science, Université Bretagne Sud, CNRS UMR 6074, IRISA, 56000 Vannes, France; nicolas.courty@irisa.fr

⁴ IMT Atlantique, CNRS UMR 6285, Lab-STICC, 29238 Brest, France; lucas.drumetz@imt-atlantique.fr

* Correspondence: clement.bonet@univ-ubs.fr

Abstract: In the context of optimal transport (OT) methods, the subspace detour approach was recently proposed by Muzellec and Cuturi. It consists of first finding an optimal plan between the measures projected on a wisely chosen subspace and then completing it in a nearly optimal transport plan on the whole space. The contribution of this paper is to extend this category of methods to the Gromov–Wasserstein problem, which is a particular type of OT distance involving the specific geometry of each distribution. After deriving the associated formalism and properties, we give an experimental illustration on a shape matching problem. We also discuss a specific cost for which we can show connections with the Knothe–Rosenblatt rearrangement.

Keywords: optimal transport; Gromov–Wasserstein; subspace detours; Knothe–Rosenblatt



Citation: Bonet, C.; Vayer, T.; Courty, N.; Septier, F.; Drumetz, L. Subspace Detours Meet Gromov–Wasserstein. *Algorithms* **2021**, *14*, 366. <https://doi.org/10.3390/a14120366>

Academic Editor: Fabio Massimo Zanzotto

Received: 9 November 2021

Accepted: 14 December 2021

Published: 17 December 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Classical optimal transport (OT) has received lots of attention recently, in particular in Machine Learning for tasks such as generative networks [1] or domain adaptation [2] to name a few. It generally relies on the Wasserstein distance, which builds an optimal coupling between distributions given a notion of distance between their samples. Yet, this metric cannot be used directly whenever the distributions lie in different metric spaces and lacks from potentially important properties, such as translation or rotation invariance of the supports of the distributions, which can be useful when comparing shapes or meshes [3,4]. In order to alleviate those problems, custom solutions have been proposed, such as [5], in which invariances are enforced by optimizing over some class of transformations, or [6], in which distributions lying in different spaces are compared by optimizing over the Stiefel manifold to project or embed one of the measures.

Apart from these works, another meaningful OT distance to tackle these problems is the Gromov–Wasserstein (GW) distance, originally proposed in [3,7,8]. It is a distance between metric spaces and has several appealing properties such as geodesics or invariances [8]. Yet, the price to be paid lies in its computational complexity, which requires solving a nonconvex quadratic optimization problem with linear constraints. A recent line of work tends to compute approximations or relaxations of the original problem in order to spread its use in more data-intensive machine learning applications. For example, Peyré et al. [9] rely on entropic regularization and Sinkhorn iterations [10], while recent methods impose coupling with low-rank constraints [11] or rely on a sliced approach [12] or on mini-batch estimators [13] to approximate the Gromov–Wasserstein distance. In Chowdhury et al. [4], the authors propose to partition the space and to solve the optimal transport problem between a subset of points before finding a coupling between all the points.

In this work, we study the *subspace detour* approach for Gromov–Wasserstein. This class of method was first proposed for the Wasserstein setting in Muzellec and Cuturi [14] and consists of (1) projecting the measures onto a wisely chosen subspace and finding an

optimal coupling between them (2) and then constructing a nearly optimal plan of the measures on the whole space using disintegration (see Section 2.2). Our main contribution is to generalize the subspace detours approach on different subspaces and to apply it for the GW distance. We derive some useful properties as well as closed-form solutions of this transport plan between Gaussians distributions. From a practical side, we provide a novel closed-form expression of the one-dimensional GW problem that allows us to efficiently compute the subspace detours transport plan when the subspaces are one-dimensional. Illustrations of the method are given on a shape matching problem where we show good results with a cheaper computational cost compared to other GW-based methods. Interestingly enough, we also propose a separable quadratic cost for the GW problem that can be related with a triangular coupling [15], hence bridging the gap with Knothe–Rosenblatt (KR) rearrangements [16,17].

2. Background

In this section, we introduce all the necessary material to describe the subspace detours approach for classical optimal transport and relate it to the Knothe–Rosenblatt rearrangement. We show how to find couplings via the gluing lemma and measure disintegration. Then, we introduce the Gromov–Wasserstein problem for which we will derive the subspace detour in the next sections.

2.1. Classical Optimal Transport

Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ be two probability measures. The set of couplings between μ and ν is defined as:

$$\Pi(\mu, \nu) = \{\gamma \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d) \mid \pi_{\#}^1 \gamma = \mu, \pi_{\#}^2 \gamma = \nu\}$$

where π^1 and π^2 are the projections on the first and second coordinate (i.e., $\pi^1(x, y) = x$), respectively, and $\#$ is the push forward operator, defined such that:

$$\forall A \in \mathcal{B}(\mathbb{R}^d), T_{\#}\mu(A) = \mu(T^{-1}(A)).$$

2.1.1. Kantorovitch Problem

There exists several types of coupling between probability measures for which a non-exhaustive list can be found in [18] (Chapter 1). Among them, the so called optimal coupling is the minimizer of the following Kantorovitch problem:

$$\inf_{\gamma \in \Pi(\mu, \nu)} \int c(x, y) d\gamma(x, y) \tag{1}$$

with c being some cost function. The Kantorovitch problem (1) is known to admit a solution when c is non-negative and lower semi-continuous [19] (Theorem 1.7). When $c(x, y) = \|x - y\|_2^2$, it defines the so-called Wasserstein distance:

$$W_2^2(\mu, \nu) = \inf_{\gamma \in \Pi(\mu, \nu)} \int \|x - y\|_2^2 d\gamma(x, y). \tag{2}$$

When the optimal coupling is of the form $\gamma = (Id, T)_{\#}\mu$ with T , some deterministic map such that $T_{\#}\mu = \nu$, T is called the Monge map.

In one dimension, with μ atomless, the solution to (2) is a deterministic coupling of the form [19] (Theorem 2.5):

$$T = F_{\nu}^{-1} \circ F_{\mu} \tag{3}$$

where F_{μ} is the cumulative distribution function of μ , and F_{ν}^{-1} is the quantile function of ν . This map is also known as the increasing rearrangement map.

2.1.2. Knothe–Rosenblatt Rearrangement

Another interesting coupling is the Knothe–Rosenblatt (KR) rearrangement, which takes advantage of the increasing rearrangement in one dimension by iterating over the dimension and using the disintegration of the measures. Concatenating all the increasing rearrangements between the conditional probabilities, the KR rearrangement produces a nondecreasing triangular map (i.e., $T : \mathbb{R}^d \rightarrow \mathbb{R}^d$, for all $x \in \mathbb{R}^d$, $T(x) = (T_1(x_1), \dots, T_j(x_1, \dots, x_j), \dots, T_d(x))$, and for all j , T_j is nondecreasing with respect to x_j), and a deterministic coupling (i.e., $T_{\#}\mu = \nu$) [18–20].

Carlier et al. [21] made a connection between this coupling and optimal transport by showing that it can be obtained as the limit of optimal transport plans for a degenerated cost:

$$c_t(x, y) = \sum_{i=1}^d \lambda_i(t)(x_i - y_i)^2,$$

where for all $i \in \{1, \dots, d\}$, $t > 0$, $\lambda_i(t) > 0$, and for all $i \geq 2$, $\frac{\lambda_i(t)}{\lambda_{i-1}(t)} \xrightarrow{t \rightarrow 0} 0$. This cost can be recast as in [22] as $c_t(x, y) = (x - y)^T A_t(x - y)$, where $A_t = \text{diag}(\lambda_1(t), \dots, \lambda_d(t))$. This formalizes into the following Theorem:

Theorem 1 ([19,21]). *Let μ and ν be two absolutely continuous measures on \mathbb{R}^d , with compact supports. Let γ_t be an optimal transport plan for the cost c_t , let T_K be the Knothe–Rosenblatt map between μ and ν , and $\gamma_K = (Id \times T_K)_{\#}\mu$ the associated transport plan. Then, we have $\gamma_t \xrightarrow[t \rightarrow 0]{\mathcal{D}} \gamma_K$. Moreover, if γ_t are induced by transport maps T_t , then T_t converges in $L^2(\mu)$ when t tends to zero to the Knothe–Rosenblatt rearrangement.*

2.2. Subspace Detours and Disintegration

Muzellec and Cuturi [14] proposed another OT problem by optimizing over the couplings which share a measure on a subspace. More precisely, they defined subspace-optimal plans for which the shared measure is the OT plan between projected measures.

Definition 1 (Subspace-Optimal Plans [14] Definition 1). *Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ and let $E \subset \mathbb{R}^d$ be a k -dimensional subspace. Let γ_E^* be an OT plan for the Wasserstein distance between $\mu_E = \pi_E^{\#}\mu$ and $\nu_E = \pi_E^{\#}\nu$ (with π^E as the orthogonal projection on E). Then, the set of E -optimal plans between μ and ν is defined as $\Pi_E(\mu, \nu) = \{\gamma \in \Pi(\mu, \nu) \mid (\pi^E, \pi^E)_{\#}\gamma = \gamma_E^*\}$.*

In other words, the subspace OT plans are the transport plans of μ, ν that agree on the subspace E with the optimal transport plan γ_E^* on this subspace. To construct such coupling $\gamma \in \Pi(\mu, \nu)$, one can rely on the Gluing lemma [18] or use the disintegration of the measure as described in the following section.

2.2.1. Disintegration

Let (Y, \mathcal{Y}) and (Z, \mathcal{Z}) be measurable spaces, and $(X, \mathcal{X}) = (Y \times Z, \mathcal{Y} \otimes \mathcal{Z})$ the product measurable space. Then, for $\mu \in \mathcal{P}(X)$, we denote $\mu_Y = \pi_Y^{\#}\mu$ and $\mu_Z = \pi_Z^{\#}\mu$ as the marginals, where π^Y (respectively π^Z) is the projection on Y (respectively Z). Then, a family $(K(y, \cdot))_{y \in Y}$ is a disintegration of μ if for all $y \in Y$, $K(y, \cdot)$ is a measure on Z , for all $A \in \mathcal{Z}$, $K(\cdot, A)$ is measurable and:

$$\forall \phi \in C(X), \int_{Y \times Z} \phi(y, z) d\mu(y, z) = \int_Y \int_Z \phi(y, z) K(y, dz) d\mu_Y(y),$$

where $C(X)$ is the set of continuous functions on X . We can note $\mu = \mu_Y \otimes K$. K is a probability kernel if for all $y \in Y$, $K(y, Z) = 1$. The disintegration of a measure actually corresponds to conditional laws in the context of probabilities. This concept will allow us to obtain measures on the whole space from marginals on subspaces.

In the case where $X = \mathbb{R}^d$, which is our setting of interest, we have existence and uniqueness of the disintegration (see Box 2.2 of [19] or Chapter 5 of [23] for the more general case).

2.2.2. Coupling on the Whole Space

Let us note $\mu_{E^\perp|E}$ and $\nu_{E^\perp|E}$ as the disintegrated measures on the orthogonal spaces (i.e., such that $\mu = \mu_E \otimes \mu_{E^\perp|E}$ and $\nu = \nu_E \otimes \nu_{E^\perp|E}$ (if we have densities, $p(x_E, x_{E^\perp}) = p_E(x_E)p_{E^\perp|E}(x_{E^\perp}|x_E)$)). Then, to obtain a transport plan between the two original measures on the whole space, we can look for another coupling between disintegrated measures $\mu_{E^\perp|E}$ and $\nu_{E^\perp|E}$. In particular, two such couplings are proposed in [14], the Monge-Independent (MI) plan:

$$\pi_{\text{MI}} = \gamma_E^* \otimes (\mu_{E^\perp|E} \otimes \nu_{E^\perp|E})$$

where we take the independent coupling between $\mu_{E^\perp|E}(x_E, \cdot)$ and $\nu_{E^\perp|E}(y_E, \cdot)$ for γ_E^* almost every (x_E, y_E) , and the Monge-Knothe (MK) plan:

$$\pi_{\text{MK}} = \gamma_E^* \otimes \gamma_{E^\perp|E}^*$$

where $\gamma_{E^\perp|E}^*((x_E, y_E), \cdot)$ is an optimal plan between $\mu_{E^\perp|E}(x_E, \cdot)$ and $\nu_{E^\perp|E}(y_E, \cdot)$ for γ_E^* almost every (x_E, y_E) . Muzellec and Cuturi [14] observed that MI is more adapted to noisy environments since it only computes the OT plan of the subspace. MK is more suited for applications where we want to prioritize some subspace but where all the directions still contain relevant information [14].

This subspace detour approach can be of much interest following the popular assumption that two distributions on \mathbb{R}^d differ only in a low-dimensional subspace as in the Spiked transport model [24]. However, it is still required to find the adequate subspace. Muzellec and Cuturi [14] propose to either rely on a priori knowledge to select the subspace (by using, e.g., a reference dataset and a principal component analysis) or to optimize over the Stiefel manifold.

2.3. Gromov–Wasserstein

Formally, the Gromov–Wasserstein distance allows us to compare metric measure spaces (mm-space), triplets (X, d_X, μ_X) and (Y, d_Y, μ_Y) , where (X, d_X) and (Y, d_Y) are complete separable metric spaces and μ_X and μ_Y are Borel probability measures on X and Y [8], respectively, by computing:

$$GW(X, Y) = \inf_{\gamma \in \Pi(\mu_X, \mu_Y)} \iint L(d_X(x, x'), d_Y(y, y')) d\gamma(x, y) d\gamma(x', y')$$

where L is some loss on \mathbb{R} . It has actually been extended to other spaces by replacing the distances by cost functions c_X and c_Y , as, e.g., in [25]. Furthermore, it has many appealing properties such as having invariances (which depend on the costs).

Vayer [26] notably studied this problem in the setting where X and Y are Euclidean spaces, with $L(x, y) = (x - y)^2$ and $c(x, x') = \langle x, x' \rangle$ or $c(x, x') = \|x - x'\|_2^2$. In particular, let $\mu \in \mathcal{P}(\mathbb{R}^p)$ and $\nu \in \mathcal{P}(\mathbb{R}^q)$, and the inner-GW problem is defined as:

$$\text{InnerGW}(\mu, \nu) = \inf_{\gamma \in \Pi(\mu, \nu)} \iint (\langle x, x' \rangle_p - \langle y, y' \rangle_q)^2 d\gamma(x, y) d\gamma(x', y'). \tag{4}$$

For this problem, a closed form in one dimension can be found when one of the distributions admits a density *w.r.t.* the Lebesgue measure:

Theorem 2 ([26] Theorem 4.2.4). *Let $\mu, \nu \in \mathcal{P}(\mathbb{R})$, with μ being absolutely continuous with respect to the Lebesgue measure. Let $F_\mu^{\nearrow}(x) := F_\mu(x) = \mu([-\infty, x])$ be the cumulative distribution function and $F_\mu^{\searrow}(x) = \mu([x, +\infty])$ the anti-cumulative distribution function. Let*

$T_{asc}(x) = F_v^{-1}(F_\mu^{\nearrow}(x))$ and $T_{desc}(x) = F_v^{-1}(F_\mu^{\searrow}(-x))$. Then, an optimal solution of (4) is achieved either by $\gamma = (Id \times T_{asc})_{\#}\mu$ or by $\gamma = (Id \times T_{desc})_{\#}\mu$.

3. Subspace Detours for GW

In this section, we propose to extend subspace detours from Muzellec and Cuturi [14] with Gromov–Wasserstein costs. We show that we can even take subspaces of different dimensions and still obtain a coupling on the whole space using the Independent or the Monge–Knothe coupling. Then, we derive some properties analogously to Muzellec and Cuturi [14], as well as some closed-form solutions between Gaussians. We also provide a new closed-form expression of the inner-GW problem between one-dimensional discrete distributions and provide an illustration on a shape-matching problem.

3.1. Motivations

First, we adapt the definition of subspace optimal plans for *different* subspaces. Indeed, since the GW distance is adapted to distributions that have their own geometry, we argue that if we project on the *same* subspace, then it is likely that the resulting coupling would not be coherent with that of GW. To illustrate this point, we use as a source distribution μ one moon of the two moons dataset and obtain a target ν by rotating μ by an angle of $\frac{\pi}{2}$ (see Figure 1). As the GW with $c(x, x') = \|x - x'\|_2^2$ is invariant with respect to isometries, the optimal coupling is diagonal, as recovered on the left side of the figure. However, when choosing one subspace to project both the source and target distributions, we completely lose the optimal coupling between them. Nonetheless, by choosing one subspace for each measure more wisely (using here the first component of the principal component analysis (PCA) decomposition), we recover the diagonal coupling. This simple illustration underlines that the choice of both subspaces is important. A way of choosing the subspaces could be to project on the subspace containing the more information for each dataset using, e.g., PCA independently on each distribution. Muzellec and Cuturi [14] proposed to optimize the optimal transport cost with respect to an orthonormal matrix with a projected gradient descent, which could be extended to an optimization over two orthonormal matrices in our context.

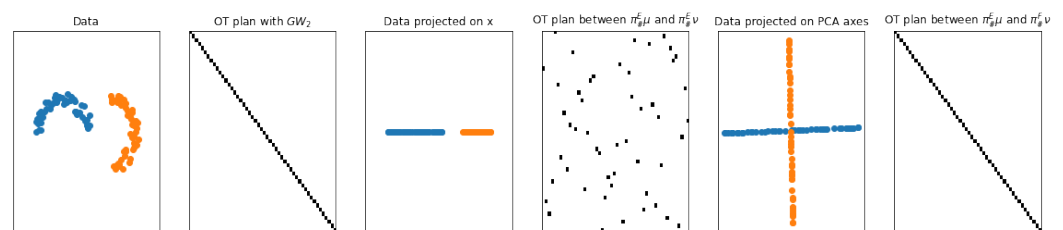


Figure 1. From left to right: Data (moons); OT plan obtained with GW for $c(x, x') = \|x - x'\|_2^2$; Data projected on the first axis; OT plan obtained between the projected measures; Data projected on their first PCA component; OT plan obtained between the the projected measures.

By allowing for different subspaces, we obtain the following definition of subspace optimal plans:

Definition 2. Let $\mu \in \mathcal{P}_2(\mathbb{R}^p)$, $\nu \in \mathcal{P}_2(\mathbb{R}^q)$, E be a k -dimensional subspace of \mathbb{R}^p and F a k' -dimensional subspace of \mathbb{R}^q . Let $\gamma_{E \times F}^*$ be an optimal transport plan for GW between $\mu_E = \pi_{\#}^E \mu$ and $\nu_F = \pi_{\#}^F \nu$ (with π^E (resp. π^F) the orthogonal projection on E (resp. F)). Then, the set of (E, F) -optimal plans between μ and ν is defined as $\Pi_{E, F}(\mu, \nu) = \{\gamma \in \Pi(\mu, \nu) \mid (\pi^E, \pi^F)_{\#}\gamma = \gamma_{E \times F}^*\}$.

Analogously to Muzellec and Cuturi [14] (Section 2.2), we can obtain from $\gamma_{E \times F}^*$ a coupling on the whole space by either defining the Monge–Independent plan $\pi_{MI} = \gamma_{E \times F}^* \otimes (\mu_{E^\perp|E} \otimes \nu_{F^\perp|F})$ or the Monge–Knothe plan $\pi_{MK} = \gamma_{E \times F}^* \otimes \gamma_{E^\perp \times F^\perp|E \times F}^*$ where OT plans are taken with some OT cost, e.g., GW.

3.2. Properties

Let $E \subset \mathbb{R}^p$ and $F \subset \mathbb{R}^q$ and denote:

$$GW_{E,F}(\mu, \nu) = \inf_{\gamma \in \Pi_{E,F}(\mu, \nu)} \iint L(x, x', y, y') d\gamma(x, y) d\gamma(x', y') \tag{5}$$

the Gromov–Wasserstein problem restricted to subspace optimal plans (2). In the following, we show that Monge–Knothe couplings are optimal plans of this problem, which is a direct transposition of Proposition 1 in [14].

Proposition 1. *Let $\mu \in \mathcal{P}(\mathbb{R}^p)$ and $\nu \in \mathcal{P}(\mathbb{R}^q)$, $E \subset \mathbb{R}^p$, $F \subset \mathbb{R}^q$, $\pi_{MK} = \gamma_{E \times F}^* \otimes \gamma_{E^\perp \times F^\perp |_{E \times F}}$, where $\gamma_{E \times F}^*$ is an optimal coupling between μ_E and ν_F , and for $\gamma_{E \times F}^*$, almost every (x_E, y_F) , $\gamma_{E^\perp \times F^\perp |_{E \times F}}^*((x_E, y_F), \cdot)$ is an optimal coupling between $\mu_{E^\perp | E}(x_E, \cdot)$ and $\nu_{F^\perp | F}(y_F, \cdot)$. Then we have:*

$$\pi_{MK} \in \operatorname{argmin}_{\gamma \in \Pi_{E,F}(\mu, \nu)} \iint L(x, x', y, y') d\gamma(x, y) d\gamma(x', y').$$

Proof. Let $\gamma \in \Pi_{E,F}(\mu, \nu)$, then:

$$\begin{aligned} & \iint L(x, x', y, y') d\gamma(x, y) d\gamma(x', y') \\ &= \iint \left(\iint L(x, x', y, y') \gamma_{E^\perp \times F^\perp |_{E \times F}}((x_E, y_F), (dx_{E^\perp}, dy_{F^\perp})) \right. \\ & \left. \gamma_{E^\perp \times F^\perp |_{E \times F}}((x'_E, y'_F), (dx'_{E^\perp}, dy'_{F^\perp})) \right) d\gamma_{E \times F}^*(x_E, y_F) d\gamma_{E \times F}^*(x'_E, y'_F). \end{aligned}$$

However, for $\gamma_{E \times F}^*$ a.e. $(x_E, y_F), (x'_E, y'_F)$,

$$\begin{aligned} & \iint L(x, x', y, y') \gamma_{E^\perp \times F^\perp |_{E \times F}}((x_E, y_F), (dx_{E^\perp}, dy_{F^\perp})) \gamma_{E^\perp \times F^\perp |_{E \times F}}((x'_E, y'_F), (dx'_{E^\perp}, dy'_{F^\perp})) \\ & \geq \iint L(x, x', y, y') \gamma_{E^\perp \times F^\perp |_{E \times F}}^*((x_E, y_F), (dx_{E^\perp}, dy_{F^\perp})) \gamma_{E^\perp \times F^\perp |_{E \times F}}^*((x'_E, y'_F), (dx'_{E^\perp}, dy'_{F^\perp})) \end{aligned}$$

by definition of the Monge–Knothe coupling. By integrating with respect to $\gamma_{E \times F}^*$, we obtain:

$$\iint L(x, x', y, y') d\gamma(x, y) d\gamma(x', y') \geq \iint L(x, x', y, y') d\pi_{MK}(x, y) d\pi_{MK}(x', y').$$

Therefore, π_{MK} is optimal for subspace optimal plans. \square

The key properties of GW that we would like to keep are its invariances. We show in two particular cases that we conserve them on the orthogonal spaces (since the measure on $E \times F$ is fixed).

Proposition 2. *Let $\mu \in \mathcal{P}(\mathbb{R}^p)$, $\nu \in \mathcal{P}(\mathbb{R}^q)$, $E \subset \mathbb{R}^p$, $F \subset \mathbb{R}^q$.*

For $L(x, x', y, y') = (\|x - x'\|_2^2 - \|y - y'\|_2^2)^2$ or $L(x, x', y, y') = (\langle x, x' \rangle_p - \langle y, y' \rangle_q)^2$, $GW_{E,F}$ (5) is invariant with respect to isometries of the form $f = (Id_E, f_{E^\perp})$ (resp. $g = (Id_F, g_{F^\perp})$) with f_{E^\perp} an isometry on E^\perp (resp. g_{F^\perp} an isometry on F^\perp) with respect to the corresponding cost ($c(x, x') = \|x - x'\|_2^2$ or $c(x, x') = \langle x, x' \rangle_p$).

Proof. We propose a sketch of the proof. The full proof can be found in Appendix A.1. Let $L(x, x', y, y') = (\|x - x'\|_2^2 - \|y - y'\|_2^2)^2$, let f_{E^\perp} be an isometry w.r.t $c(x_{E^\perp}, x'_{E^\perp}) = \|x_{E^\perp} - x'_{E^\perp}\|_2^2$, and let $f : \mathbb{R}^p \rightarrow \mathbb{R}^p$ be defined as such for all $x \in \mathbb{R}^p$, $f(x) = (x_E, f_{E^\perp}(x_{E^\perp}))$.

By using Lemma 6 of [27], we show that $\Pi_{E,F}(f\#\mu, \nu) = \{(f, Id)\#\gamma \mid \gamma \in \Pi_{E,F}(\mu, \nu)\}$. Hence, for all $\gamma \in \Pi_{E,F}(f\#\mu, \nu)$, there exists $\tilde{\gamma} \in \Pi_{E,F}(\mu, \nu)$ such that $\gamma = (f, Id)\#\tilde{\gamma}$.

By disintegrating $\tilde{\gamma}$ with respect to $\gamma_{E \times F}^*$ and using the properties of the pushforward, we can show that:

$$\begin{aligned} & \iint (\|x - x'\|_2^2 - \|y - y'\|_2^2)^2 d(f, Id)_\# \tilde{\gamma}(x, y) d(f, Id)_\# \tilde{\gamma}(x', y') \\ &= \iint (\|x - x'\|_2^2 - \|y - y'\|_2^2)^2 d\tilde{\gamma}(x, y) d\tilde{\gamma}(x', y'). \end{aligned}$$

Finally, by taking the infimum with respect to $\tilde{\gamma} \in \Pi_{E, F}(\mu, \nu)$, we find:

$$GW_{E, F}(f_\# \mu, \nu) = GW_{E, F}(\mu, \nu).$$

□

3.3. Closed-Form between Gaussians

We can also derive explicit formulas between Gaussians in particular cases. Let $q \leq p$, $\mu = \mathcal{N}(m_\mu, \Sigma) \in \mathcal{P}(\mathbb{R}^p)$, $\nu = \mathcal{N}(m_\nu, \Lambda) \in \mathcal{P}(\mathbb{R}^q)$ two Gaussian measures with $\Sigma = P_\mu D_\mu P_\mu^T$ and $\Lambda = P_\nu D_\nu P_\nu^T$. As previously, let $E \subset \mathbb{R}^p$ and $F \subset \mathbb{R}^q$ be k and k' dimensional subspaces, respectively. Following Muzellec and Cuturi [14], we represent Σ in an orthonormal basis of $E \oplus E^\perp$ and Λ in an orthonormal basis of $F \oplus F^\perp$, i.e., $\Sigma = \begin{pmatrix} \Sigma_E & \Sigma_{EE^\perp} \\ \Sigma_{E^\perp E} & \Sigma_{E^\perp} \end{pmatrix}$. Now, let us denote the following:

$$\Sigma / \Sigma_E = \Sigma_{E^\perp} - \Sigma_{EE^\perp}^T \Sigma_E^{-1} \Sigma_{EE^\perp}$$

as the Schur complement of Σ with respect to Σ_E . We know that the conditionals of Gaussians are Gaussians and that their covariances are the Schur complements (see, e.g., [28,29]).

For $L(x, x', y, y') = (\|x - x'\|_2^2 - \|y - y'\|_2^2)^2$, we have for now no certainty that the optimal transport plan is Gaussian. Let \mathcal{N}_{p+q} denote the set of Gaussians in \mathbb{R}^{p+q} . By restricting the minimization problem to Gaussian couplings, i.e., by solving:

$$GGW(\mu, \nu) = \inf_{\gamma \in \Pi(\mu, \nu) \cap \mathcal{N}_{p+q}} \iint (\|x - x'\|_2^2 - \|y - y'\|_2^2)^2 d\gamma(x, y) d\gamma(x', y'), \tag{6}$$

Salmona et al. [30] showed that there is a solution $\gamma^* = (Id, T)_\# \mu \in \Pi(\mu, \nu)$ with $\mu = \mathcal{N}(m_\mu, \Sigma)$, $\nu = \mathcal{N}(m_\nu, \Lambda)$ and

$$\forall x \in \mathbb{R}^d, T(x) = m_\nu + P_\nu A P_\mu^T (x - m_\mu) \tag{7}$$

where $A = \begin{pmatrix} \tilde{I}_q D_\nu^{\frac{1}{2}} (D_\mu^{(q)})^{-\frac{1}{2}} & 0_{q, p-q} \end{pmatrix} \in \mathbb{R}^{q \times p}$, and \tilde{I}_q is of the form $\text{diag}((\pm 1)_{i \leq q})$.

By combining the results of Muzellec and Cuturi [14] and Salmona et al. [30], we obtain the following closed-form for Monge–Knothe couplings:

Proposition 3. *Suppose $p \geq q$ and $k = k'$. For the Gaussian-restricted GW problem (6), a Monge–Knothe transport map between $\mu = \mathcal{N}(m_\mu, \Sigma) \in \mathcal{P}(\mathbb{R}^p)$ and $\nu = \mathcal{N}(m_\nu, \Lambda) \in \mathcal{P}(\mathbb{R}^q)$ is, for all $x \in \mathbb{R}^p$, $T_{MK}(x) = m_\nu + B(x - m_\mu)$ where:*

$$B = \begin{pmatrix} T_{E, F} & 0 \\ C & T_{E^\perp, F^\perp | E, F} \end{pmatrix}$$

with $T_{E, F}$ being an optimal transport map between $\mathcal{N}(0_E, \Sigma_E)$ and $\mathcal{N}(0_F, \Lambda_F)$ (of the form (7)), $T_{E^\perp, F^\perp | E, F}$ an optimal transport map between $\mathcal{N}(0_{E^\perp}, \Sigma / \Sigma_E)$ and $\mathcal{N}(0_{F^\perp}, \Lambda / \Lambda_F)$, and C satisfies:

$$C = (\Lambda_{F^\perp F} (T_{E, F}^T)^{-1} - T_{E^\perp, F^\perp | E, F} \Sigma_{E^\perp E}) \Sigma_E^{-1}.$$

Proof. See Appendix A.2.1. □

Suppose that $k \geq k'$, $m_\mu = 0$, and $m_\nu = 0$ and let $T_{E,F}$ be an optimal transport map between μ_E and ν_F (of the form (7)). We can derive a formula for the Monge–Independent coupling for the inner-GW problem and the Gaussian restricted GW problem.

Proposition 4. $\pi_{MI} = \mathcal{N}(0_{p+q}, \Gamma)$ where $\Gamma = \begin{pmatrix} \Sigma & C \\ C^T & \Lambda \end{pmatrix}$ with

$$C = (V_E \Sigma_E + V_{E^\perp} \Sigma_{E^\perp E}) T_{E,F}^T (V_F^T + \Lambda_F^{-1} \Lambda_{F^\perp F}^T V_{F^\perp}^T)$$

where $T_{E,F}$ is an optimal transport map, either for the inner-GW problem or the Gaussian restricted problem.

Proof. See Appendix A.2.2. \square

3.4. Computation of Inner-GW between One-Dimensional Empirical Measures

In practice, computing the Gromov–Wasserstein distance from samples of the distributions is costly. From a computational point of view, the subspace detour approach provides an interesting method with better computational complexity when choosing 1D subspaces. Moreover, we have the intuition that the GW problem between measures lying on smaller dimensional subspaces has a better sample complexity than between the original measures, as it is the case for the Wasserstein distance [31,32].

Below, we show that when both E and F are one-dimensional subspaces, then the resulting GW problem between the projected measures can be solved in linear time. This will rely on a new closed-form expression of the GW problem in 1D. Vayer et al. [12] provided a closed-form for GW with $c(x, x') = \|x - x'\|_2^2$ in one dimension between discrete measures containing the same number of points and with uniform weights. However, in our framework, the 1D projection of E, F may not have uniform weights, and we also would like to be able to compare distributions with different numbers of points. We provide in the next proposition a closed-form expression for the inner-GW problem between any unidimensional discrete probability distributions:

Proposition 5. Consider $\Sigma_n = \{a \in \mathbb{R}_+^n, \sum_{i=1}^n a_i = 1\}$ the n probability simplex. For a vector $a \in \mathbb{R}^n$, we denote a^- as the vector with values sorted decreasingly, i.e., $a_1^- \geq \dots \geq a_n^-$. Let $\mu = \sum_{i=1}^n a_i \delta_{x_i}, \nu = \sum_{j=1}^m b_j \delta_{y_j} \in \mathcal{P}(\mathbb{R}) \times \mathcal{P}(\mathbb{R})$ with $a, b \in \Sigma_n \times \Sigma_m$. Suppose that $x_1 \leq \dots \leq x_n$ and $y_1 \leq \dots \leq y_m$. Consider the problem:

$$\min_{\gamma \in \Pi(a,b)} \sum_{ijkl} (x_i x_k - y_j y_l)^2 \gamma_{ij} \gamma_{kl} \tag{8}$$

Then, there exists $\gamma \in \{NW(a, b), NW(a^-, b)\}$ such that γ is an optimal solution of (8) where NW is the North-West corner rule defined in Algorithm 1. As a corollary, an optimal solution of (8) can be found in $O(n + m)$.

Algorithm 1 North-West corner rule $NW(a, b)$

```

a ∈ Σn, b ∈ Σm
while i ≤ n, j ≤ m do
    γij = min{ai, bj}
    ai = ai - γij
    bj = bj - γij
    If ai = 0, i = i + 1, if bj = 0, j = j + 1
end while
return γ ∈ Π(a, b)
    
```

Proof. Let $\gamma \in \Pi(a, b)$. Then:

$$\sum_{ijkl} (x_i x_k - y_j y_l)^2 \gamma_{ij} \gamma_{kl} = \sum_{ijkl} (x_i x_k)^2 \gamma_{ij} \gamma_{kl} + \sum_{ijkl} (y_j y_l)^2 \gamma_{ij} \gamma_{kl} - 2 \sum_{ijkl} x_i x_k y_j y_l \gamma_{ij} \gamma_{kl}$$

However, $\sum_{ijkl} (x_i x_k)^2 \gamma_{ij} \gamma_{kl} = \sum_{ik} (x_i x_k)^2 a_i a_k$, and $\sum_{ijkl} (y_j y_l)^2 \gamma_{ij} \gamma_{kl} = \sum_{jl} (y_j y_l)^2 b_j b_l$, so this does not depend on γ . Moreover $2 \sum_{ijkl} x_i x_k y_j y_l \gamma_{ij} \gamma_{kl} = 2 (\sum_{ij} x_i y_j \gamma_{ij})^2$. Hence, the problem (8) is equivalent to $\max_{\gamma \in \Pi(a,b)} (\sum_{ij} x_i y_j \gamma_{ij})^2$ (in terms of the OT plan), which is also equivalent to solving $\max_{\gamma \in \Pi(a,b)} |\sum_{ij} x_i y_j \gamma_{ij}|$ or equivalently:

$$\max_{\gamma \in \Pi(a,b)} \pm 1 \sum_{ij} x_i y_j \gamma_{ij} \tag{9}$$

We have two cases to consider: If $\pm 1 = 1$, we have to solve $-\min_{\gamma \in \Pi(a,b)} \sum_{ij} (-x_i) y_j \gamma_{ij}$. Since the points are sorted, the matrix $c_{ij} = -x_i y_j$ satisfies the Monge property [33]:

$$\forall (i, j) \in \{1, \dots, n-1\} \times \{1, \dots, m-1\}, c_{ij} + c_{i+1, j+1} \leq c_{i+1, j} + c_{i, j+1} \tag{10}$$

To see this, check that:

$$\begin{aligned} & (-x_i) y_j + (-x_{i+1}) y_{j+1} - (-x_{i+1}) y_j - (-x_i) y_{j+1} \\ & = (-x_i) (y_j - y_{j+1}) + (-x_{i+1}) (y_{j+1} - y_j) = (y_j - y_{j+1}) (x_{i+1} - x_i) \leq 0 \end{aligned} \tag{11}$$

In this case, the North-West corner rule $NW(a, b)$ defined in Alg. 1 is known to produce an optimal solution to the linear problem (9) [33]. If $\pm = -1$, then changing x_i to $-x_i$ concludes. \square

We emphasize that this result is novel and generalizes [12] in the sense that the distributions do not need to have uniform weights and the same number of points. In addition, Theorem 2 is not directly applicable to this setting since it requires having absolutely regular distributions, which is not the case here. Both results are, however, related, as the solution obtained by using the NW corner rule on the sorted samples is the same as that obtained by considering the coupling obtained from the quantile functions. The previous result could also be used to define tractable alternatives to GW in the same manner as the Sliced Gromov–Wasserstein [12].

3.5. Illustrations

We use the Python Optimal Transport (POT) library [34] to compute the different optimal transport problems involved in this illustration. We are interested here in solving a 3D mesh registration problem, which is a natural application of Gromov–Wasserstein [3] since it enjoys invariances with respect to isometries such as permutations and can also naturally exploit the topology of the meshes. For this purpose, we selected two base meshes from the FAUST dataset [35], which provides ground truth correspondences between shapes. The information available from those meshes are geometrical (6890 vertices positions) and topological (mesh connectivity). These two meshes are represented, along with the visual results of the registration, in Figure 2. In order to visually depict the quality of the assignment induced by the transport map, we propagate through it a color code of the source vertices toward their associated counterpart vertices in the target mesh. Both the original color-coded source and the associated target ground truth are available on the first line of the illustration. To compute our method, we simply use as a natural subspace for both meshes the algebraic connectivity of the mesh’s topological information, also known as the Fiedler vector [36] (eigenvector associated to the second smallest eigenvalue of the un-normalized Laplacian matrix). Fiedler vectors are computed in practice using NetworkX [37] but could also be obtained by using power methods [38]. Reduced to a 1D optimal transport problem (8), we used the Proposition 5 to compute the optimal coupling in $O(n + m)$. Consequently, the computation time is very low (~ 5 s. on a standard laptop),

and the associated matching is very good, with more than 98% of correct assignments. We qualitatively compare this result to Gromov–Wasserstein mappings induced by different cost functions, in the second line of Figure 2: adjacency [39], weighted adjacency (weights are given by distances between vertices), heat kernel (derived from the un-normalized Laplacian) [40], and, finally, geodesic distances over the meshes. On average, computing the Gromov–Wasserstein mapping using POT took around 10 min of time. Both methods based on adjacency fail to recover a meaningful mapping. Heat kernel allows us to map continuous areas of the source mesh but fails in recovering a global structure. Finally, the geodesic distance gives a much more coherent mapping but has inverted left and right of the human figure. Notably, a significant extra computation time was induced by the computation of the geodesic distances (~ 1 h/mesh using the NetworkX [37] shortest path procedure). As a conclusion, and despite the simplification of the original problem, our method performs best with a speed-up of two-orders of magnitude.

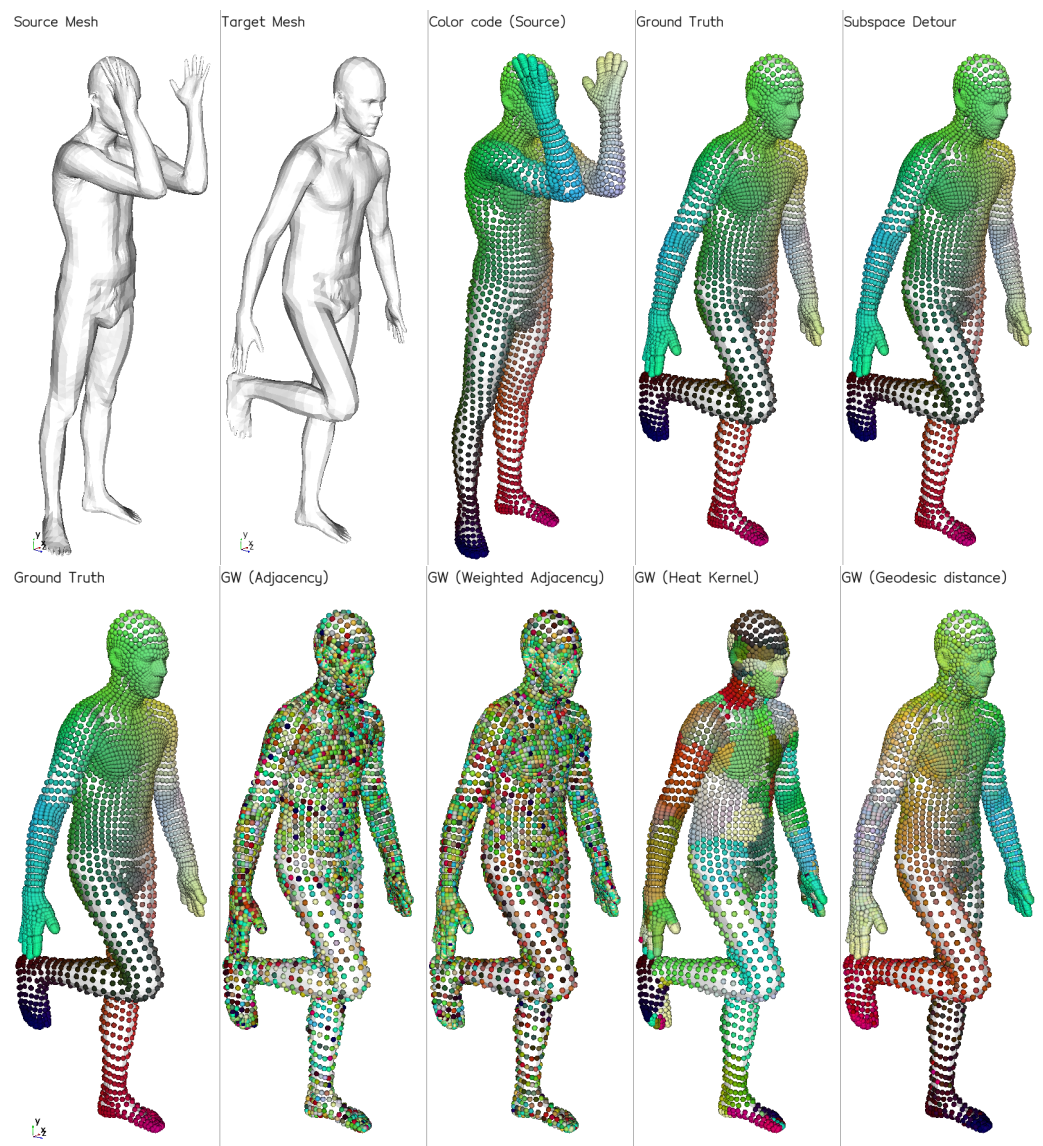


Figure 2. Three-dimensional mesh registration. (First row) source and target meshes, color code of the source, ground truth color code on the target, result of subspace detour using Fiedler vectors as subspace. (Second row) After recalling the expected ground truth for ease of comparison, we present results of different Gromov–Wasserstein mappings obtained with metrics based on adjacency, heat kernel, and geodesic distances.

4. Triangular Coupling as Limit of Optimal Transport Plans for Quadratic Cost

Another interesting property derived in Muzellec and Cuturi [14] of the Monge–Knothe coupling is that it can be obtained as the limit of classic optimal transport plans, similar to Theorem 1, using a separable cost of the form:

$$c_t(x, y) = (x - y)^T P_t (x - y)$$

with $P_t = V_E V_E^T + t V_{E^\perp} V_{E^\perp}^T$ and (V_E, V_{E^\perp}) as an orthonormal basis of \mathbb{R}^p .

However, this property is not valid for the classical Gromov–Wasserstein cost (e.g., $L(x, x', y, y') = (d_X(x, x')^2 - d_Y(y, y')^2)^2$ or $L(x, x', y, y') = (\langle x, x' \rangle_p - \langle y, y' \rangle_q)^2$) as the cost is not separable. Motivated by this question, we ask ourselves in the following if we can derive a quadratic optimal transport cost for which we would have this property.

Formally, we derive a new quadratic optimal transport problem using the Hadamard product. We show that this problem is well-defined and that it has interesting properties such as invariance with respect to axis. We also show that it can be related to a triangular coupling in a similar fashion than the classical optimal transport problem with the Knothe–Rosenblatt rearrangement.

4.1. Construction of the Hadamard–Wasserstein Problem

In this part, we define the “Hadamard–Wasserstein” problem between $\mu \in \mathcal{P}(\mathbb{R}^d)$ and $\nu \in \mathcal{P}(\mathbb{R}^d)$ as:

$$\mathcal{HW}^2(\mu, \nu) = \inf_{\gamma \in \Pi(\mu, \nu)} \iint \|x \odot x' - y \odot y'\|_2^2 d\gamma(x, y) d\gamma(x', y'), \tag{12}$$

where \odot is the Hadamard product (element-wise product). This problem is different than the Gromov–Wasserstein problem in the sense that we do not compare intradistance anymore but rather the Hadamard products between vectors of the two spaces (in the same fashion as the classical Wasserstein distance). Hence, we need the two measures to belong in the same Euclidean space. Let us note L as the cost defined as:

$$\forall x, x', y, y' \in \mathbb{R}^d, L(x, x', y, y') = \sum_{k=1}^d (x_k x'_k - y_k y'_k)^2 = \|x \odot x' - y \odot y'\|_2^2. \tag{13}$$

We observe that it coincides with the inner-GW (4) loss in one dimension. Therefore, by Theorem 2, we know that we have a closed-form solution in 1D.

4.2. Properties

First, we derive some useful properties of (12) which are usual for the regular Gromov–Wasserstein problem. Formally, we show that the problem is well defined and that it is a pseudometric with invariances with respect to axes.

Proposition 6. Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$.

1. The problem (12) always admits a minimizer.
2. \mathcal{HW} is a pseudometric (i.e., it is symmetric, non-negative, $\mathcal{HW}(\mu, \mu) = 0$, and it satisfies the triangle inequality).
3. \mathcal{HW} is invariant to reflection with respect to axes.

Proof. Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$,

1. $(x, x') \mapsto x \odot x'$ is a continuous map, therefore, L is less semi-continuous. Hence, by applying Lemma 2.2.1 of [26], we observe that $\gamma \mapsto \iint L(x, x', y, y') d\gamma(x, y) d\gamma(x', y')$ is less semi-continuous for the weak convergence of measures. Now, as $\Pi(\mu, \nu)$ is a compact set (see the proof of Theorem 1.7 in [19] for the Polish space case and of Theorem 1.4 for the compact metric space) and $\gamma \mapsto \iint L d\gamma d\gamma$ is

less semi-continuous for the weak convergence, we can apply the Weierstrass theorem (Memo 2.2.1 in [26]), which states that (12) always admits a minimizer.

2. See Theorem 16 in [25].
3. For invariances, we first look at the properties that must be satisfied by T in order to have: $\forall x, x', f(x, x') = f(T(x), T(x'))$ where $f : (x, x') \mapsto x \odot x'$.

We find that $\forall x \in \mathbb{R}^d, \forall 1 \leq i \leq d, |[T(x)]_i| = |x_i|$ because, denoting $(e_i)_{i=1}^d$ as the canonical basis, we have:

$$x \odot e_i = T(x) \odot T(e_i),$$

which implies that:

$$x_i = [T(x)]_i [T(e_i)]_i.$$

However, $f(e_i, e_i) = f(T(e_i), T(e_i))$ implies $[T(e_i)]_i^2 = 1$, and therefore:

$$|[T(x)]_i| = |x_i|.$$

If we take for T the reflection with respect to axis, then it satisfies $f(x, x') = f(T(x), T(x'))$ well. Moreover, it is a good equivalence relation, and therefore, we have a distance on the quotient space.

□

\mathcal{HW} loses some properties compared to GW . Indeed, it is only invariant with respect to axes, and it can only compare measures lying in the same Euclidean space in order for the distance to be well defined. Nonetheless, we show in the following that we can derive some links with triangular couplings in the same way as the Wasserstein distance with KR .

Indeed, the cost L (13) is separable and reduces to the inner- GW loss in 1D, for which we have a closed-form solution. We can therefore define a degenerated version of it:

$$\begin{aligned} \forall x, x', y, y' \in \mathbb{R}^d, L_t(x, x', y, y') &= \sum_{k=1}^d \left(\prod_{i=1}^{k-1} \lambda_t^{(i)} \right) (x_k x'_k - y_k y'_k)^2 \\ &= (x \odot x' - y \odot y')^T A_t (x \odot x' - y \odot y') \end{aligned} \tag{14}$$

with $A_t = \text{diag}(1, \lambda_t^{(1)}, \lambda_t^{(1)} \lambda_t^{(2)}, \dots, \prod_{i=1}^{d-1} \lambda_t^{(i)})$, such as for all $t > 0$, and for all $i \in \{1, \dots, d-1\}$, $\lambda_t^{(i)} > 0$, and $\lambda_t^{(i)} \xrightarrow[t \rightarrow 0]{} 0$. We denote \mathcal{HW}_t the problem (12) with the degenerated cost (14). Therefore, we will be able to decompose the objective as:

$$\begin{aligned} &\iint L_t(x, x', y, y') d\gamma(x, y) d\gamma(x', y') \\ &= \iint (x_1 x'_1 - y_1 y'_1)^2 d\gamma(x, y) d\gamma(x', y') \\ &+ \iint \sum_{k=2}^d \left(\prod_{i=1}^{k-1} \lambda_t^{(i)} \right) (x_k x'_k - y_k y'_k)^2 d\gamma(x, y) d\gamma(x', y') \end{aligned}$$

and to use the same induction reasoning as [21].

Then, we can define a triangular coupling different from the Knothe–Rosenblatt rearrangement in the sense that each map will not be nondecreasing. Indeed, following Theorem 2, the solution of each 1D problem:

$$\operatorname{argmin}_{\gamma \in \Pi(\mu, \nu)} \iint (xx' - yy')^2 d\gamma(x, y) d\gamma(x', y')$$

is either $(Id \times T_{asc})\#\mu$ or $(Id \times T_{desc})\#\mu$. Hence, at each step $k \geq 1$, if we disintegrate the joint law of the k first variables as $\mu^{1:k} = \mu^{1:k-1} \otimes \mu^{k|1:k-1}$, the optimal transport map $T(\cdot|x_1, \dots, x_{k-1})$ will be the solution of:

$$\operatorname{argmin}_{T \in \{T_{asc}, T_{desc}\}} \iint (x_k x'_k - T(x_k)T(x'_k))^2 \mu^{k|1:k-1}(dx_k | x_{1:k-1}) \mu^{k|1:k-1}(dx'_k | x'_{1:k-1}).$$

We now state the main theorem, where we show that the limit of the OT plans obtained with the degenerated cost will be the triangular coupling we just defined.

Theorem 3. *Let μ and ν be two absolutely continuous measures on \mathbb{R}^d such that $\int \|x\|_2^4 \mu(dx) < +\infty$, $\int \|y\|_2^4 \nu(dy) < +\infty$ and with compact support. Let γ_t be an optimal transport plan for \mathcal{HW}_t , let T_K be the alternate Knothe–Rosenblatt map between μ and ν as defined in the last paragraph, and let $\gamma_K = (Id \times T_K)\#\mu$ be the associated transport plan. Then, we have $\gamma_t \xrightarrow[t \rightarrow 0]{\mathcal{D}} \gamma_K$. Moreover, if γ_t are induced by transport maps T_t , then $T_t \xrightarrow[t \rightarrow 0]{L^2(\mu)} T_K$.*

Proof. See Appendix B.2. \square

However, we cannot extend this Theorem to the subspace detour approach. Indeed, by choosing $A_t = V_E V_E^T + t V_{E^\perp} V_{E^\perp}^T$ with (V_E, V_{E^\perp}) an orthonormal basis of \mathbb{R}^d , then we project $x \odot x' - y \odot y'$ on E (respectively on E^\perp), which is generally different from $x_E \odot x'_E - y_E \odot y'_E$ (respectively $x_{E^\perp} \odot x'_{E^\perp} - y_{E^\perp} \odot y'_{E^\perp}$).

4.3. Solving Hadamard–Wasserstein in the Discrete Setting

In this part, we derive formulas to solve numerically \mathcal{HW} (12). Let $x_1, \dots, x_n \in \mathbb{R}^d$, $y_1, \dots, y_m \in \mathbb{R}^d$, $\alpha \in \Sigma_n$, $\beta \in \Sigma_m$, $p = \sum_{i=1}^n \alpha_i \delta_{x_i}$ and $q = \sum_{j=1}^m \beta_j \delta_{y_j}$ two discrete measures in \mathbb{R}^d . The Hadamard Wasserstein problem (12) becomes in the discrete setting:

$$\begin{aligned} \mathcal{HW}^2(p, q) &= \inf_{\gamma \in \Pi(p, q)} \sum_{i,j} \sum_{k,\ell} \|x_i \odot x_k - y_j \odot y_\ell\|_2^2 \gamma_{i,j} \gamma_{k,\ell} \\ &= \inf_{\gamma \in \Pi(p, q)} \mathcal{E}(\gamma) \end{aligned}$$

with $\mathcal{E}(\gamma) = \sum_{i,j} \sum_{k,\ell} \|x_i \odot x_k - y_j \odot y_\ell\|_2^2 \gamma_{i,j} \gamma_{k,\ell}$. As denoted in [9], if we note:

$$\mathcal{L}_{i,j,k,\ell} = \|x_i \odot x_k - y_j \odot y_\ell\|_2^2,$$

then we have:

$$\mathcal{E}(\gamma) = \langle \mathcal{L} \otimes \gamma, \gamma \rangle,$$

where \otimes is defined as:

$$\mathcal{L} \otimes \gamma = \left(\sum_{k,\ell} \mathcal{L}_{i,j,k,\ell} \gamma_{k,\ell} \right)_{i,j} \in \mathbb{R}^{n \times m}.$$

We show in the next proposition a decomposition of $\mathcal{L} \otimes \gamma$, which allows us to compute this tensor product more efficiently.

Proposition 7. *Let $\gamma \in \Pi(p, q) = \{M \in (\mathbb{R}_+)^{n \times m}, M \mathbb{1}_m = p, M^T \mathbb{1}_n = q\}$, where $\mathbb{1}_n = (1, \dots, 1)^T \in \mathbb{R}^n$. Let us note $X = (x_i \odot x_k)_{i,k} \in \mathbb{R}^{n \times n \times d}$, $Y = (y_j \odot y_\ell)_{j,\ell} \in \mathbb{R}^{m \times m \times d}$, $X^{(2)} = (\|X_{i,k}\|_2^2)_{i,k} \in \mathbb{R}^{n \times n}$, $Y^{(2)} = (\|Y_{j,\ell}\|_2^2)_{j,\ell} \in \mathbb{R}^{m \times m}$, and $\forall t \in \{1, \dots, d\}$, $X_t = (X_{i,k,t})_{i,k} \in \mathbb{R}^{n \times n}$ and $Y_t = (Y_{j,\ell,t})_{j,\ell} \in \mathbb{R}^{m \times m}$. Then:*

$$\mathcal{L} \otimes \gamma = X^{(2)} p \mathbb{1}_m^T + \mathbb{1}_n q^T (Y^{(2)})^T - 2 \sum_{t=1}^d X_t \gamma Y_t^T.$$

Proof. First, we can start by writing:

$$\begin{aligned} \mathcal{L}_{i,j,k,\ell} &= \|x_i \odot x_k - y_j \odot y_\ell\|_2^2 \\ &= \|X_{i,k} - Y_{j,\ell}\|_2^2 \\ &= \|X_{i,k}\|_2^2 + \|Y_{j,\ell}\|_2^2 - 2\langle X_{i,k}, Y_{j,\ell} \rangle \\ &= [X^{(2)}]_{i,k} + [Y^{(2)}]_{j,\ell} - 2\langle X_{i,k}, Y_{j,\ell} \rangle. \end{aligned}$$

We cannot directly apply proposition 1 from [9] (as the third term is a scalar product), but by performing the same type of computation, we obtain:

$$\mathcal{L} \otimes \gamma = A + B + C$$

with

$$A_{i,j} = \sum_{k,\ell} [X^{(2)}]_{i,k} \gamma_{k,\ell} = \sum_k [X^{(2)}]_{i,k} \sum_\ell \gamma_{k,\ell} = \sum_k [X^{(2)}]_{i,k} [\gamma \mathbb{1}_m]_{k,1} = [X^{(2)} \gamma \mathbb{1}_m]_{i,1} = [X^{(2)} p]_{i,1}$$

$$B_{i,j} = \sum_{k,\ell} [Y^{(2)}]_{j,\ell} \gamma_{k,\ell} = \sum_\ell [Y^{(2)}]_{j,\ell} \sum_k \gamma_{k,\ell} = \sum_\ell [Y^{(2)}]_{j,\ell} [\gamma^T \mathbb{1}_n]_{\ell,1} = [Y^{(2)} \gamma^T \mathbb{1}_n]_{j,1} = [Y^{(2)} q]_{j,1}$$

and

$$\begin{aligned} C_{i,j} &= -2 \sum_{k,\ell} \langle X_{i,k}, Y_{j,\ell} \rangle \gamma_{k,\ell} = -2 \sum_{k,\ell} \sum_{t=1}^d X_{i,k,t} Y_{j,\ell,t} \gamma_{k,\ell} \\ &= -2 \sum_{t=1}^d \sum_k [X_t]_{i,k} \sum_\ell [Y_t]_{j,\ell} \gamma_{\ell,k} \\ &= -2 \sum_{t=1}^d \sum_k [X_t]_{i,k} [Y_t \gamma^T]_{j,k} \\ &= -2 \sum_{t=1}^d [X_t (Y_t \gamma^T)^T]_{i,j}. \end{aligned}$$

Finally, we have:

$$\mathcal{L} \otimes \gamma = X^{(2)} p \mathbb{1}_m^T + \mathbb{1}_n q^T (Y^{(2)})^T - 2 \sum_{t=1}^d X_t \gamma Y_t^T.$$

□

From this decomposition, we can compute the tensor product $\mathcal{L} \otimes \gamma$ with a complexity of $O(d(n^2m + m^2n))$ using only multiplications of matrices (instead of $O(dn^2m^2)$ for a naive computation).

Remark 1. For the degenerated cost function (14), we just need to replace X and Y by $\tilde{X}_t = A_t^{\frac{1}{2}} X$ and $\tilde{Y}_t = A_t^{\frac{1}{2}} Y$ in the previous proposition.

To solve this problem numerically, we can use the conditional gradient algorithm (Algorithm 2 in [41]). This algorithm only requires to compute the gradient:

$$\nabla \mathcal{E}(\gamma) = 2(A + B + C) = 2(\mathcal{L} \otimes \gamma)$$

at each step and a classical OT problem. This algorithm is more efficient than solving the quadratic problem directly. Moreover, while it is a non-convex problem, it actually converges to a local stationary point [42].

On Figure 3, we generated 30 points of 2 Gaussian distributions, and computed the optimal coupling of \mathcal{HW}_t for several t . These points have the same uniform weight. We plot

the couplings between the points on the second row, and between the projected points on their first coordinate on the first row. Note that for discrete points, the Knothe–Rosenblatt coupling amounts to sorting the points with respect to the first coordinate if there is no ambiguity (i.e., $x_1^{(1)} < \dots < x_n^{(1)}$) as it comes back to perform the optimal transport in one dimension [43] (Remark 2.28). For our cost, the optimal coupling in 1D can either be the increasing or the decreasing rearrangement. We observe on the first row of Figure 3 that the optimal coupling when t is close to 0 corresponds to the decreasing rearrangement, which corresponds well to the alternate Knothe–Rosenblatt map we defined in Section 4.2. It underlines the results provided in Theorem 3.

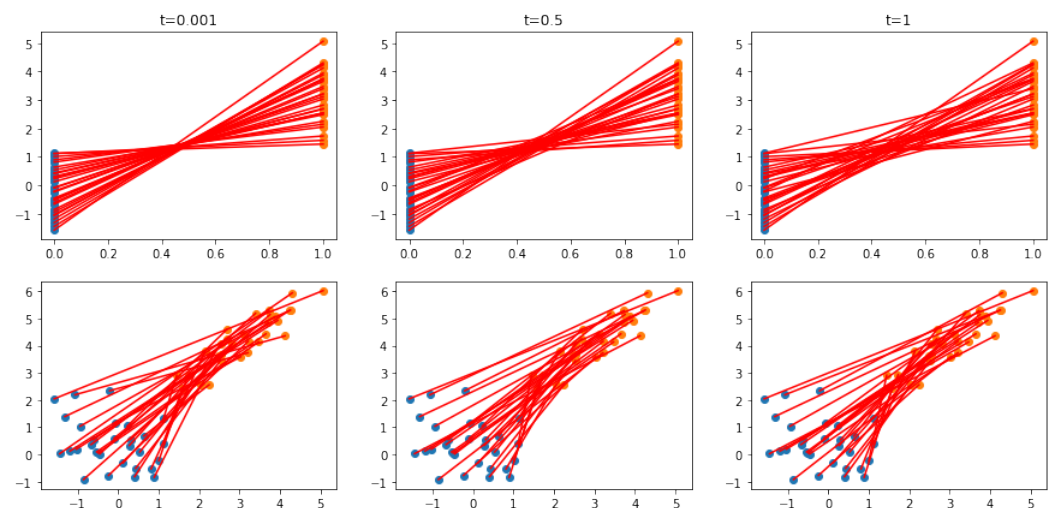


Figure 3. Degenerated coupling. On the **first row**, the points are projected on their first coordinate and we plot the optimal coupling. On the **second row**, we plot the optimal coupling between the original points.

5. Discussion

We proposed in this work to extend the subspace detour approach to different subspaces, and to other optimal transport costs such as Gromov–Wasserstein. Being able to project on different subspaces can be useful when the data are not aligned and do not share the same axes of interest, as well as when we are working between different metric spaces as it is the case, for example, with graphs. However, a question that arises is how to choose these subspaces. Since the method is mostly interesting when we choose one-dimensional subspaces, we proposed to use a PCA and to project on the first directions for data embedded in Euclidean spaces. For more complicated data such as graphs, we projected onto the Fiedler vector and obtained good results in an efficient way on a 3D mesh registration problem. More generally, Muzellec and Cuturi [14] proposed to perform a gradient descent on the loss with respect to orthonormal matrices. This approach is non-convex and is only guaranteed to converge to a local minimum. Designing such an algorithm, which would minimize alternatively between two transformations in the Stiefel manifold, is left for future works.

The subspace detour approach for transport problem is meaningful whenever one can identify subspaces that gather most of the information from the original distributions, while making the estimate more robust and with a better sample complexity as far as dimensions are lower. On the computational complexity side, and when we have only access to discrete data, the subspace detour approach brings better computational complexity solely when the subspaces are chosen as one dimensional. Indeed, otherwise, we have the same complexity for solving the subspace detour and solving the OT problem directly (since the complexity only depends on the number of samples). In this case, the 1D projection often gives distinct values for all the samples (for continuous valued data) and hence the Monge–Knothe coupling is exactly the coupling in 1D. As such, information is lost on

the orthogonal spaces. It can be artificially recovered by quantizing the 1D values (as experimented in practice in [14]), but the added value is not clear and deserves broader studies. If given absolutely continuous distributions *wrt.* the Lebesgue measure, however, this limit does not exist but comes with the extra cost of being able to compute efficiently the projected measure onto the subspace, which might require discretization of the space and is therefore not practical in high dimensions.

We also proposed a new quadratic cost \mathcal{HW} that we call Hadamard–Wasserstein, which allows us to define a degenerated cost for which the optimal transport plan converges to a triangular coupling. However, this cost loses many properties compared to W_2 or GW , for which we are inclined to use these problems. Indeed, while \mathcal{HW} is a quadratic cost, it uses a Euclidean norm between the Hadamard product of vectors and requires the two spaces to be the same (in order to have the distance well defined). A work around in the case $X = \mathbb{R}^p$ and $Y = \mathbb{R}^q$ with $p \leq q$ would be to “lift” the vectors in \mathbb{R}^p into vectors in \mathbb{R}^q with padding as it is proposed in [12] or to project the vectors in \mathbb{R}^q on \mathbb{R}^p as in [6]. Yet, for some applications where only the distance/similarity matrices are available, a different strategy still needs to be found. Another concern is the limited invariance properties (only with respect to axial symmetry symmetry in our case). Nevertheless, we expect that such a cost can be of interest in cases where invariance to symmetry is a desired property, such as in [44].

Author Contributions: Methodology, C.B., N.C., T.V., F.S., and L.D.; software, C.B. and N.C.; writing—original draft, C.B., N.C., T.V., F.S., and L.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by project DynaLearn from Labex CominLabs and Région Bretagne ARED DLearnMe. N.C. acknowledges fundings from the ANR OTTOPIA AI chair (ANR-20-CHIA-0030). T.V. was supported in part by the AllegroAssai ANR project (ANR-19-CHIA-0009) and by the ACADEMICS grant of the IDEXLYON, project of the Université de Lyon, PIA operated by ANR-16-IDEX-0005.

Data Availability Statement: The FAUST dataset might be found at <http://faust.is.tue.mpg.de> (accessed in 1 September 2021).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

OT	Optimal Transport
GW	Gromov–Wasserstein
KR	Knothe–Rosenblatt
MI	Monge–Independent
MK	Monge–Knothe
PCA	Principal Component Analysis
POT	Python Optimal Transport

Appendix A. Subspace Detours

Appendix A.1. Proofs

Proof of Proposition 2. We first deal with $L(x, x', y, y') = (\|x - x'\|_2^2 - \|y - y'\|_2^2)^2$. Let f_{E^\perp} be an isometry *w.r.t* $c(x_{E^\perp}, x'_{E^\perp}) = \|x_{E^\perp} - x'_{E^\perp}\|_2^2$, and let $f : \mathbb{R}^p \rightarrow \mathbb{R}^p$ be defined such as for all $x \in \mathbb{R}^p$, $f(x) = (x_E, f_{E^\perp}(x_{E^\perp}))$.

From Lemma 6 of Paty and Cuturi [27], we know that $\Pi(f\#\mu, \nu) = \{(f, Id)\#\gamma \mid \gamma \in \Pi(\mu, \nu)\}$. We can rewrite:

$$\begin{aligned} \Pi_{E,F}(f\#\mu, \nu) &= \{\gamma \in \Pi(f\#\mu, \nu) \mid (\pi^E, \pi^F)\#\gamma = \gamma_{E \times F}^*\} \\ &= \{(f, Id)\#\gamma \mid \gamma \in \Pi(\mu, \nu), (\pi^E, \pi^F)\#\gamma = \gamma_{E \times F}^*\} \\ &= \{(f, Id)\#\gamma \mid \gamma \in \Pi(\mu, \nu), (\pi^E, \pi^F)\#\gamma = \gamma_{E \times F}^*\} \\ &= \{(f, Id)\#\gamma \mid \gamma \in \Pi_{E,F}(\mu, \nu)\} \end{aligned}$$

using $f = (Id_E, f_{E^\perp})$, $\pi^E \circ f = Id_E$ and $(\pi^E, \pi^F)\#\gamma = (\pi^E, \pi^F)\#\gamma$.

Now, for all $\gamma \in \Pi_{E,F}(f\#\mu, \nu)$, there exists $\tilde{\gamma} \in \Pi_{E,F}(\mu, \nu)$ such that $\gamma = (f, Id)\#\tilde{\gamma}$, and we can disintegrate $\tilde{\gamma}$ with respect to $\gamma_{E \times F}^*$:

$$\tilde{\gamma} = \gamma_{E \times F}^* \otimes K$$

with K a probability kernel on $(E \times F, \mathcal{B}(E^\perp) \otimes \mathcal{B}(F^\perp))$.

For $\gamma_{E \times F}^*$ almost every $(x_E, y_F), (x'_{E^\perp}, y'_{F^\perp})$, we have:

$$\begin{aligned} &\iint (\|x_E - x'_{E^\perp}\|_2^2 + \|x_{E^\perp} - x'_{E^\perp}\|_2^2 - \|y_F - y'_{F^\perp}\|_2^2 - \|y_{F^\perp} - y'_{F^\perp}\|_2^2)^2 \\ &(f_{E^\perp}, Id)\#K((x_E, y_F), (dx_{E^\perp}, dy_{F^\perp}))(f_{E^\perp}, Id)\#K((x'_{E^\perp}, y'_{F^\perp}), (dx'_{E^\perp}, dy'_{F^\perp})) \\ &= \iint (\|x_E - x'_{E^\perp}\|_2^2 + \|f_{E^\perp}(x_{E^\perp}) - f_{E^\perp}(x'_{E^\perp})\|_2^2 - \|y_F - y'_{F^\perp}\|_2^2 - \|y_{F^\perp} - y'_{F^\perp}\|_2^2)^2 \\ &K((x_E, y_F), (dx_{E^\perp}, dy_{F^\perp}))K((x'_{E^\perp}, y'_{F^\perp}), (dx'_{E^\perp}, dy'_{F^\perp})) \\ &= \iint (\|x_E - x'_{E^\perp}\|_2^2 + \|x_{E^\perp} - x'_{E^\perp}\|_2^2 - \|y_F - y'_{F^\perp}\|_2^2 - \|y_{F^\perp} - y'_{F^\perp}\|_2^2)^2 \\ &K((x_E, y_F), (dx_{E^\perp}, dy_{F^\perp}))K((x'_{E^\perp}, y'_{F^\perp}), (dx'_{E^\perp}, dy'_{F^\perp})) \end{aligned}$$

using in the last line that $\|f_{E^\perp}(x_{E^\perp}) - f_{E^\perp}(x'_{E^\perp})\|_2 = \|x_{E^\perp} - x'_{E^\perp}\|_2$ since f_{E^\perp} is an isometry. By integrating with respect to $\gamma_{E \times F}^*$, we obtain:

$$\begin{aligned} &\iint \left(\iint (\|x - x'\|_2^2 - \|y - y'\|_2^2)^2 \right. \\ &(f_{E^\perp}, Id)\#K((x_E, y_F), (dx_{E^\perp}, dy_{F^\perp}))(f_{E^\perp}, Id)\#K((x'_{E^\perp}, y'_{F^\perp}), (dx'_{E^\perp}, dy'_{F^\perp})) \tag{A1} \\ &d\gamma_{E \times F}^*(x_E, y_F) d\gamma_{E \times F}^*(x'_{E^\perp}, y'_{F^\perp}) \\ &= \iint (\|x - x'\|_2^2 - \|y - y'\|_2^2)^2 d\tilde{\gamma}(x, y) d\tilde{\gamma}(x', y'). \end{aligned}$$

Now, we show that $\gamma = (f, Id)\#\tilde{\gamma} = \gamma_{E \times F}^* \otimes (f_{E^\perp}, Id)\#K$. Let ϕ be some bounded measurable function on $\mathbb{R}^p \times \mathbb{R}^q$:

$$\begin{aligned} \int \phi(x, y) d\gamma(x, y) &= \int \phi(x, y) d((f, Id)\#\tilde{\gamma}(x, y)) \\ &= \int \phi(f(x), y) d\tilde{\gamma}(x, y) \\ &= \iint \phi(f(x), y) K((x_E, y_F), (dx_{E^\perp}, dy_{F^\perp})) d\gamma_{E \times F}^*(x_E, y_F) \\ &= \iint \phi((x_E, f_{E^\perp}(x_{E^\perp})), y) K((x_E, y_F), (dx_{E^\perp}, dy_{F^\perp})) d\gamma_{E \times F}^*(x_E, y_F) \\ &= \iint \phi(x, y) (f_{E^\perp}, Id)\#K((x_E, y_F), (dx_{E^\perp}, dy_{F^\perp})) d\gamma_{E \times F}^*(x_E, y_F). \end{aligned}$$

Hence, we can rewrite (A1) as:

$$\begin{aligned} & \iint (\|x - x'\|_2^2 - \|y - y'\|_2^2)^2 d(f, Id)_\# \tilde{\gamma}(x, y) d(f, Id)_\# \tilde{\gamma}(x', y') \\ &= \iint (\|x - x'\|_2^2 - \|y - y'\|_2^2)^2 d\tilde{\gamma}(x, y) d\tilde{\gamma}(x', y'). \end{aligned}$$

Now, by taking the infimum with respect to $\tilde{\gamma} \in \Pi_{E,F}(\mu, \nu)$, we find:

$$GW_{E,F}(f\#\mu, \nu) = GW_{E,F}(\mu, \nu).$$

For the inner product case, we can do the same proof for linear isometries on E^\perp . \square

Appendix A.2. Closed-Form between Gaussians

Let $q \leq p$, $\mu = \mathcal{N}(m_\mu, \Sigma) \in \mathcal{P}(\mathbb{R}^p)$, and $\nu = \mathcal{N}(m_\nu, \Lambda) \in \mathcal{P}(\mathbb{R}^q)$ be two Gaussian measures with $\Sigma = P_\mu D_\mu P_\mu^T$ and $\Lambda = P_\nu D_\nu P_\nu^T$.

Let $E \subset \mathbb{R}^p$ be a subspace of dimension k and $F \subset \mathbb{R}^q$ a subspace of dimension k' .

We represent Σ in an orthonormal basis of $E \oplus E^\perp$ and Λ in an orthonormal basis of $F \oplus F^\perp$, i.e., $\Sigma = \begin{pmatrix} \Sigma_E & \Sigma_{EE^\perp} \\ \Sigma_{E^\perp E} & \Sigma_{E^\perp} \end{pmatrix}$. We denote $\Sigma/\Sigma_E = \Sigma_{E^\perp} - \Sigma_{EE^\perp}^T \Sigma_E^{-1} \Sigma_{EE^\perp}$ as the Schur complement of Σ with respect to Σ_E . We know that the conditionals of Gaussians are Gaussians and of covariance, the Schur complement (see e.g., Rasmussen [28] or Von Mises [29]).

Appendix A.2.1. Quadratic GW Problem

For GW with $c(x, x') = \|x - x'\|_2^2$, we have for now no guarantee that there exists an optimal coupling which is a transport map. Salmona et al. [30] proposed to restrict the problem to the set of Gaussian couplings $\pi(\mu, \nu) \cap \mathcal{N}_{p+q}$ where \mathcal{N}_{p+q} denotes the set of Gaussians in \mathbb{R}^{p+q} . In that case, the problem becomes:

$$GGW(\mu, \nu) = \inf_{\gamma \in \Pi(\mu, \nu) \cap \mathcal{N}_{p+q}} \iint (\|x - x'\|_2^2 - \|y - y'\|_2^2)^2 d\gamma(x, y) d\gamma(x', y'). \tag{A2}$$

In that case, they showed that an optimal solution is of the form $T(x) = m_\nu + P_\nu A P_\mu^T(x - m_\mu)$ with $A = \begin{pmatrix} \tilde{I}_q D_\nu^{\frac{1}{2}} (D_\mu^{(q)})^{-\frac{1}{2}} & 0_{q, p-q} \end{pmatrix}$ and \tilde{I}_q of the form $\text{diag}((\pm 1)_{i \leq q})$.

Since the problem is translation invariant, we can always solve the problem between the centered measures.

In the following, we suppose that $k = k'$. Let us denote $T_{E,F}$ as the optimal transport map for (A2) between $\mathcal{N}(0, \Sigma_E)$ and $\mathcal{N}(0, \Lambda_F)$. According to Theorem 4.1 in Salmona et al. [30], such a solution exists and is of the form (7). We also denote T_{E^\perp, F^\perp} as the optimal transport map between $\mathcal{N}(0, \Sigma/\Sigma_E)$ and $\mathcal{N}(0, \Lambda/\Lambda_F)$ (which is well defined since we assumed $p \geq q$ and hence $p - k \geq q - k'$ since $k = k'$).

We know that the Monge–Knothe transport map will be a linear map $T_{MK}(x) = Bx$ with B a block triangular matrix of the form:

$$B = \begin{pmatrix} T_{E,F} & 0_{k', p-k} \\ C & T_{E^\perp, F^\perp} \end{pmatrix} \in \mathbb{R}^{q \times p},$$

with $C \in \mathbb{R}^{(q-k') \times k}$ and such that $B\Sigma B^T = \Lambda$ (to have well a transport map between μ and ν).

Actually,

$$B\Sigma B^T = \begin{pmatrix} T_{E,F} \Sigma_E T_{E,F}^T & T_{E,F} \Sigma_E C^T + T_{E,F} \Sigma_{EE^\perp} T_{E^\perp, F^\perp}^T \\ (C \Sigma_E + T_{E^\perp, F^\perp} \Sigma_{E^\perp E}) T_{E,F}^T & (C \Sigma_E + T_{E^\perp, F^\perp} \Sigma_{E^\perp E}) C^T + (C \Sigma_{EE^\perp} + T_{E^\perp, F^\perp} \Sigma_{E^\perp}) T_{E^\perp, F^\perp}^T \end{pmatrix}.$$

First, we have well $T_{E,F}\Sigma_E T_{E,F}^T = \Lambda_F$, as $T_{E,F}$ is a transport map between μ_E and ν_F . Then:

$$B\Sigma B^T = \Lambda \iff \begin{cases} T_{E,F}\Sigma_E T_{E,F}^T = \Lambda_F \\ T_{E,F}\Sigma_E C^T + T_{E,F}\Sigma_{EE^\perp} T_{E^\perp,F^\perp}^T = \Lambda_{FF^\perp} \\ (C\Sigma_E + T_{E^\perp,F^\perp}\Sigma_{E^\perp E})T_{E,F}^T = \Lambda_{F^\perp F} \\ (C\Sigma_E + T_{E^\perp,F^\perp}\Sigma_{E^\perp E})C^T + (C\Sigma_{EE^\perp} + T_{E^\perp,F^\perp}\Sigma_{E^\perp})T_{E^\perp,F^\perp}^T = \Lambda_{F^\perp}. \end{cases}$$

We have:

$$(C\Sigma_E + T_{E^\perp,F^\perp}\Sigma_{E^\perp E})T_{E,F}^T = \Lambda_{F^\perp F} \iff C\Sigma_E T_{E,F}^T = \Lambda_{F^\perp F} - T_{E^\perp,F^\perp}\Sigma_{E^\perp E}T_{E,F}^T.$$

As $k = k'$, $\Sigma_E T_{E,F}^T \in \mathbb{R}^{k \times k}$ and is invertible (as Σ_E and Λ_F are positive definite and $T_{E,F} = P_{\mu_E} A_{E,F} P_{\nu_F}$ with $A_{E,F} = (\tilde{I}_k D_{\nu_F}^{\frac{1}{2}} D_{\mu_E}^{-\frac{1}{2}})$ with positive values on the diagonals. Hence, we have:

$$C = (\Lambda_{F^\perp F} (T_{E,F}^T)^{-1} - T_{E^\perp,F^\perp}\Sigma_{E^\perp E})\Sigma_E^{-1}.$$

Now, we still have to check the last two equations. First:

$$\begin{aligned} T_{E,F}\Sigma_E C^T + T_{E,F}\Sigma_{EE^\perp} T_{E^\perp,F^\perp}^T &= T_{E,F}\Sigma_E \Sigma_E^{-1} T_{E,F}^{-1} \Lambda_{F^\perp F}^T - T_{E,F}\Sigma_E \Sigma_E^{-1} \Sigma_{E^\perp E}^T T_{E^\perp,F^\perp}^T + T_{E,F}\Sigma_{EE^\perp} T_{E^\perp,F^\perp}^T \\ &= \Lambda_{FF^\perp}. \end{aligned}$$

For the last equation:

$$\begin{aligned} &(C\Sigma_E + T_{E^\perp,F^\perp}\Sigma_{E^\perp E})C^T + (C\Sigma_{EE^\perp} + T_{E^\perp,F^\perp}\Sigma_{E^\perp})T_{E^\perp,F^\perp}^T \\ &= (\Lambda_{F^\perp F} (T_{E,F}^T)^{-1} - T_{E^\perp,F^\perp}\Sigma_{E^\perp E} + T_{E^\perp,F^\perp}\Sigma_{E^\perp E})\Sigma_E^{-1} (T_{E,F}^{-1} \Lambda_{F^\perp F}^T - \Sigma_{E^\perp E}^T T_{E^\perp,F^\perp}^T) \\ &+ \Lambda_{F^\perp F} (T_{E,F}^T)^{-1} \Sigma_E^{-1} \Sigma_{EE^\perp}^T T_{E^\perp,F^\perp}^T - T_{E^\perp,F^\perp}\Sigma_{E^\perp E} \Sigma_E^{-1} \Sigma_{EE^\perp}^T T_{E^\perp,F^\perp}^T + T_{E^\perp,F^\perp}\Sigma_{E^\perp} T_{E^\perp,F^\perp}^T \\ &= \Lambda_{F^\perp F} (T_{E,F}^T)^{-1} \Sigma_E^{-1} T_{E,F}^{-1} \Lambda_{F^\perp F}^T - \Lambda_{F^\perp F} (T_{E,F}^T)^{-1} \Sigma_E^{-1} \Sigma_{E^\perp E}^T T_{E^\perp,F^\perp}^T - T_{E^\perp,F^\perp}\Sigma_{E^\perp E} \Sigma_E^{-1} T_{E,F}^{-1} \Lambda_{F^\perp F}^T \\ &+ T_{E^\perp,F^\perp}\Sigma_{E^\perp E} \Sigma_E^{-1} \Sigma_{E^\perp E}^T T_{E^\perp,F^\perp}^T + T_{E^\perp,F^\perp}\Sigma_{E^\perp E} \Sigma_E^{-1} T_{E,F}^{-1} \Lambda_{F^\perp F}^T - T_{E^\perp,F^\perp}\Sigma_{E^\perp E} \Sigma_E^{-1} \Sigma_{E^\perp E}^T T_{E^\perp,F^\perp}^T \\ &+ \Lambda_{F^\perp F} (T_{E,F}^T)^{-1} \Sigma_E^{-1} \Sigma_{EE^\perp}^T T_{E^\perp,F^\perp}^T - T_{E^\perp,F^\perp}\Sigma_{E^\perp E} \Sigma_E^{-1} \Sigma_{E^\perp E}^T T_{E^\perp,F^\perp}^T + T_{E^\perp,F^\perp}\Sigma_{E^\perp} T_{E^\perp,F^\perp}^T \\ &= \Lambda_{F^\perp F} (T_{E,F}^T)^{-1} \Sigma_E^{-1} T_{E,F}^{-1} \Lambda_{F^\perp F}^T - T_{E^\perp,F^\perp}\Sigma_{E^\perp E} \Sigma_E^{-1} \Sigma_{E^\perp E}^T T_{E^\perp,F^\perp}^T + T_{E^\perp,F^\perp}\Sigma_{E^\perp} T_{E^\perp,F^\perp}^T \end{aligned}$$

Now, using that $(T_{E,F}^T)^{-1} \Sigma_E^{-1} T_{E,F}^{-1} = (T_{E,F}\Sigma_E T_{E,F}^T)^{-1} = \Lambda_F^{-1}$ and $\Sigma_{E^\perp} - \Sigma_{E^\perp E} \Sigma_E^{-1} \Sigma_{E^\perp E}^T = \Sigma / \Sigma_E$, we have:

$$\begin{aligned} &(C\Sigma_E + T_{E^\perp,F^\perp}\Sigma_{E^\perp E})C^T + (C\Sigma_{EE^\perp} + T_{E^\perp,F^\perp}\Sigma_{E^\perp})T_{E^\perp,F^\perp}^T \\ &= \Lambda_{F^\perp F} \Lambda_F^{-1} \Lambda_{F^\perp F}^T + T_{E^\perp,F^\perp} (\Sigma_{E^\perp} - \Sigma_{E^\perp E} \Sigma_E^{-1} \Sigma_{E^\perp E}^T) T_{E^\perp,F^\perp}^T \\ &= \Lambda_{F^\perp F} \Lambda_F^{-1} \Lambda_{F^\perp F}^T + \Lambda / \Lambda_F \\ &= \Lambda_{F^\perp} \end{aligned}$$

Then, π_{MK} is of the form $(Id, T_{MK})\# \mu$ with:

$$T_{MK}(x) = m_\nu + B(x - m_\mu).$$

Appendix A.2.2. Closed-Form between Gaussians for Monge-Independent

Suppose is $k \geq k'$ in order to be able to define the OT map between μ_E and ν_F .

For the Monge–Independent plan, $\pi_{MI} = \gamma_{E \times F}^* \otimes (\mu_{E^\perp|E} \otimes \nu_{F^\perp|F})$, let $(X, Y) \sim \pi_{MI}$. We know that π_{MI} is a degenerate Gaussian with a covariance of the form:

$$\text{Cov}(X, Y) = \begin{pmatrix} \text{Cov}(X) & C \\ C^T & \text{Cov}(Y) \end{pmatrix}$$

where $\text{Cov}(X) = \Sigma$ and $\text{Cov}(Y) = \Lambda$. Moreover, we know that C is of the form:

$$\begin{pmatrix} \text{Cov}(X_E, Y_F) & \text{Cov}(X_E, Y_{F^\perp}) \\ \text{Cov}(X_{E^\perp}, Y_F) & \text{Cov}(X_{E^\perp}, Y_{F^\perp}) \end{pmatrix}.$$

Let us assume that $m_\mu = m_\nu = 0$, then:

$$\text{Cov}(X_E, Y_F) = \text{Cov}(X_E, T_{E,F}X_E) = \mathbb{E}[X_E X_E^T] T_{E,F}^T = \Sigma_E T_{E,F}^T,$$

$$\begin{aligned} \text{Cov}(X_E, Y_{F^\perp}) &= \mathbb{E}[X_E Y_{F^\perp}^T] \\ &= \mathbb{E}[\mathbb{E}[X_E Y_{F^\perp}^T | X_E, Y_F]] \\ &= \mathbb{E}[X_E \mathbb{E}[Y_{F^\perp}^T | Y_F]] \end{aligned}$$

since $Y_F = T_{E,F}X_E$, X_E is $\sigma(Y_F)$ -measurable. Now, using the equation (A.6) from Rasmussen [28], we have:

$$\begin{aligned} \mathbb{E}[Y_{F^\perp} | Y_F] &= \Lambda_{F^\perp F} \Lambda_F^{-1} Y_F \\ &= \Lambda_{F^\perp F} \Lambda_F^{-1} T_{E,F} X_E \end{aligned}$$

and

$$\mathbb{E}[X_{E^\perp} | X_E] = \Sigma_{E^\perp E} \Sigma_E^{-1} X_E.$$

Hence:

$$\begin{aligned} \text{Cov}(X_E, Y_{F^\perp}) &= \mathbb{E}[X_E \mathbb{E}[Y_{F^\perp}^T | Y_F]] \\ &= \mathbb{E}[X_E X_E^T] T_{E,F}^T \Lambda_F^{-1} \Lambda_{F^\perp F}^T \\ &= \Sigma_E T_{E,F}^T \Lambda_F^{-1} \Lambda_{F^\perp F}^T. \end{aligned}$$

We also have:

$$\text{Cov}(X_{E^\perp}, Y_F) = \mathbb{E}[X_{E^\perp} X_E^T T_{E,F}^T] = \Sigma_{E^\perp E} T_{E,F}^T,$$

and

$$\begin{aligned} \text{Cov}(X_{E^\perp}, Y_{F^\perp}) &= \mathbb{E}[X_{E^\perp} Y_{F^\perp}^T] \\ &= \mathbb{E}[\mathbb{E}[X_{E^\perp} Y_{F^\perp}^T | X_E, Y_F]] \\ &= \mathbb{E}[\mathbb{E}[X_{E^\perp} | X_E] \mathbb{E}[Y_{F^\perp}^T | Y_F]] \text{ by independence} \\ &= \mathbb{E}[\Sigma_{E^\perp E} \Sigma_E^{-1} X_E X_E^T T_{E,F}^T \Lambda_F^{-1} \Lambda_{F^\perp F}^T] \\ &= \Sigma_{E^\perp E} T_{E,F}^T \Lambda_F^{-1} \Lambda_{F^\perp F}^T. \end{aligned}$$

Finally, we find:

$$C = \begin{pmatrix} \Sigma_E T_{E,F}^T & \Sigma_E T_{E,F}^T \Lambda_F^{-1} \Lambda_{F^\perp F}^T \\ \Sigma_{E^\perp E} T_{E,F}^T & \Sigma_{E^\perp E} T_{E,F}^T \Lambda_F^{-1} \Lambda_{F^\perp F}^T \end{pmatrix}.$$

By taking orthogonal bases (V_E, V_{E^\perp}) and (V_F, V_{F^\perp}) , we can put it in a more compact way, such as in Proposition 4 in Muzellec and Cuturi [14]:

$$C = (V_E \Sigma_E + V_{E^\perp} \Sigma_{E^\perp E}) T_{E,F}^T (V_F^T + \Lambda_F^{-1} \Lambda_{F^\perp F}^T V_{F^\perp}^T).$$

To check it, just expand the terms and see that $C_{E,F} = V_E C V_F^T$.

Appendix B. Knothe–Rosenblatt

Appendix B.1. Properties of (12)

Proposition A1. *In a slightly more general setting, let $X_0 = X_1 = \mathbb{R}^d$, functions f_0, f_1 from $\mathbb{R}^d \times \mathbb{R}^d$ to \mathbb{R}^d , and measures $\mu_0 \in \mathcal{P}(X_0), \mu_1 \in \mathcal{P}(X_1)$. Then, the family $\mathcal{X}_t = (X_0 \times X_1, f_t, \gamma^*)$ defines a geodesic between \mathcal{X}_0 and \mathcal{X}_1 , where γ^* is the optimal coupling of \mathcal{HW} between μ_0 and μ_1 , and*

$$f_t((x_0, x'_0), (x_1, x'_1)) = (1 - t)f_0(x_0, x'_0) + tf_1(x_1, x'_1).$$

Proof. See Theorem 3.1 in [8]. \square

Appendix B.2. Proof of Theorem 3

We first recall a useful theorem.

Theorem A1 (Theorem 2.8 in Billingsley [45]). *Let $\Omega = X \times Y$ be a separable space, and let $P, P_n \in \mathcal{P}(\Omega)$ with marginals P_X (respectively $P_{n,X}$) and P_Y (respectively $P_{n,Y}$). Then, $P_{n,X} \otimes P_{n,Y} \xrightarrow{\mathcal{D}} P$ if and only if $P_{n,X} \xrightarrow{\mathcal{D}} P_X, P_{n,Y} \xrightarrow{\mathcal{D}} P_Y$ and $P = P_X \otimes P_Y$.*

Proof of Theorem 3. The following proof is mainly inspired by the proof of Theorem 1 in [21] (Theorem 2.1), [22] (Theorem 3.1.6) and [19] (Theorem 2.23).

Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$, absolutely continuous, with finite fourth moments and compact supports. We recall the problem \mathcal{HW}_t :

$$\mathcal{HW}_t^2(\mu, \nu) = \inf_{\gamma \in \Pi(\mu, \nu)} \iint \sum_{k=1}^d \left(\prod_{i=1}^{k-1} \lambda_t^{(i)} \right) (x_k x'_k - y_k y'_k)^2 d\gamma_t(x, y) d\gamma_t(x', y'),$$

with $\forall t > 0, \forall i \in \{1, \dots, d - 1\}, \lambda_t^{(i)} > 0$ and $\lambda_t^{(i)} \xrightarrow[t \rightarrow 0]{} 0$.

First, let us denote γ_t the optimal coupling for \mathcal{HW}_t for all $t > 0$. We want to show that $\gamma_t \xrightarrow[t \rightarrow 0]{\mathcal{D}} \gamma_K$ with $\gamma_K = (Id \times T_K) \# \mu$ and T_K our alternate Knothe-Rosenblatt rearrangement.

Let $\gamma \in \Pi(\mu, \nu)$ such that $\gamma_t \xrightarrow[t \rightarrow 0]{\mathcal{D}} \gamma$ (true up to subsequence as $\{\mu\}$ and $\{\nu\}$ are tight in $\mathcal{P}(X)$ and $\mathcal{P}(Y)$ if X and Y are polish space, therefore, by [18] (Lemma 4.4), $\Pi(\mu, \nu)$ is a tight set, and we can apply the Prokhorov theorem [19] (Box 1.4) on $(\gamma_t)_t$ and extract a subsequence)).

Part 1:

First, let us notice that:

$$\begin{aligned} \mathcal{HW}_t^2(\mu, \nu) &= \iint \sum_{k=1}^d \left(\prod_{i=1}^{k-1} \lambda_t^{(i)} \right) (x_k x'_k - y_k y'_k)^2 d\gamma_t(x, y) d\gamma_t(x', y') \\ &= \iint (x_1 x'_1 - y_1 y'_1)^2 d\gamma_t(x, y) d\gamma_t(x', y') \\ &\quad + \iint \sum_{k=2}^d \left(\prod_{i=1}^{k-1} \lambda_t^{(i)} \right) (x_k x'_k - y_k y'_k)^2 d\gamma_t(x, y) d\gamma_t(x', y'). \end{aligned}$$

Moreover, as γ_t is the optimal coupling between μ and ν , and $\gamma_K \in \Pi(\mu, \nu)$,

$$\begin{aligned} \mathcal{HW}_t^2(\mu, \nu) &\leq \iint \sum_{k=1}^d \left(\prod_{i=1}^{k-1} \lambda_t^{(i)} \right) (x_k x'_k - y_k y'_k)^2 \, d\gamma_K(x, y) d\gamma_K(x', y') \\ &= \iint (x_1 x'_1 - y_1 y'_1)^2 \, d\gamma_K(x, y) d\gamma_K(x', y') \\ &\quad + \iint \sum_{k=2}^d \left(\prod_{i=1}^{k-1} \lambda_t^{(i)} \right) (x_k x'_k - y_k y'_k)^2 \, d\gamma_K(x, y) d\gamma_K(x', y'). \end{aligned}$$

In our case, we have $\gamma_t \xrightarrow[t \rightarrow 0]{\mathcal{D}} \gamma$, thus, by Theorem A1, we have $\gamma_t \otimes \gamma_t \xrightarrow[t \rightarrow 0]{\mathcal{D}} \gamma \otimes \gamma$. Using the fact that $\forall i, \lambda_t^{(i)} \xrightarrow[t \rightarrow 0]{} 0$ (and Lemma 1.8 of Santambrogio [19], since we are on compact support, we can bound the cost (which is continuous) by its max), we obtain the following inequality

$$\iint (x_1 x'_1 - y_1 y'_1)^2 \, d\gamma(x, y) d\gamma(x', y') \leq \iint (x_1 x'_1 - y_1 y'_1)^2 \, d\gamma_K(x, y) d\gamma_K(x', y').$$

By denoting γ^1 and γ_K^1 the marginals on the first variables, we can use the projection $\pi^1(x, y) = (x_1, y_1)$, such as $\gamma^1 = \pi_{\#}^1 \gamma$ and $\gamma_K^1 = \pi_{\#}^1 \gamma_K$. Hence, we get

$$\iint (x_1 x'_1 - y_1 y'_1)^2 \, d\gamma^1(x_1, y_1) d\gamma^1(x'_1, y'_1) \leq \iint (x_1 x'_1 - y_1 y'_1)^2 \, d\gamma_K^1(x_1, y_1) d\gamma_K^1(x'_1, y'_1).$$

However, γ_K^1 was constructed in order to be the unique optimal map for this cost (either T_{asc} or T_{desc} according to theorem [26] (Theorem 4.2.4)). Thus, we can deduce that $\gamma^1 = (Id \times T_K^1)_{\#} \mu^1 = \gamma_K^1$.

Part 2:

We know that for any $t > 0$, γ_t and γ_K share the same marginals. Thus, as previously, $\pi_{\#}^1 \gamma_t$ should have a cost worse than $\pi_{\#}^1 \gamma_K$, which translates to

$$\begin{aligned} \iint (x_1 x'_1 - y_1 y'_1)^2 \, d\gamma_K^1(x_1, y_1) d\gamma_K^1(x'_1, y'_1) &= \iint (x_1 x'_1 - y_1 y'_1)^2 \, d\gamma^1(x_1, y_1) d\gamma^1(x'_1, y'_1) \\ &\leq \iint (x_1 x'_1 - y_1 y'_1)^2 \, d\gamma_t^1(x_1, y_1) d\gamma_t^1(x'_1, y'_1). \end{aligned}$$

Therefore, we have the following inequality,

$$\begin{aligned} &\iint (x_1 x'_1 - y_1 y'_1)^2 \, d\gamma^1(x, y) d\gamma^1(x', y') + \iint \sum_{k=2}^d \left(\prod_{i=1}^{k-1} \lambda_t^{(i)} \right) (x_k x'_k - y_k y'_k)^2 \, d\gamma_t(x, y) d\gamma_t(x', y') \\ &\leq \mathcal{HW}_t^2(\mu, \nu) \\ &\leq \iint (x_1 x'_1 - y_1 y'_1)^2 \, d\gamma^1(x, y) d\gamma^1(x', y') \\ &\quad + \iint \sum_{k=2}^d \left(\prod_{i=1}^{k-1} \lambda_t^{(i)} \right) (x_k x'_k - y_k y'_k)^2 \, d\gamma_K(x, y) d\gamma_K(x', y'). \end{aligned}$$

We can subtract the first term and factorize by $\lambda_t^{(1)} > 0$,

$$\begin{aligned} & \iint \sum_{k=2}^d \left(\prod_{i=1}^{k-1} \lambda_t^{(i)} \right) (x_k x'_k - y_k y'_k)^2 \, d\gamma_t(x, y) d\gamma_t(x', y') \\ &= \lambda_t^{(1)} \left(\iint (x_2 x'_2 - y_2 y'_2)^2 \, d\gamma_t(x, y) d\gamma_t(x', y') \right. \\ &+ \iint \sum_{k=3}^d \left(\prod_{i=2}^{k-1} \lambda_t^{(i)} \right) (x_k x'_k - y_k y'_k)^2 \, d\gamma_t(x, y) d\gamma_t(x', y') \left. \right) \\ &\leq \lambda_t^{(1)} \left(\iint (x_2 x'_2 - y_2 y'_2)^2 \, d\gamma_K(x, y) d\gamma_K(x', y') \right. \\ &+ \iint \sum_{k=3}^d \left(\prod_{i=2}^{k-1} \lambda_t^{(i)} \right) (x_k x'_k - y_k y'_k)^2 \, d\gamma_K(x, y) d\gamma_K(x', y') \left. \right). \end{aligned}$$

By dividing by $\lambda_t^{(1)}$ and by taking the limit $t \rightarrow 0$ as in the first part, we get

$$\iint (x_2 x'_2 - y_2 y'_2)^2 \, d\gamma(x, y) d\gamma(x', y') \leq \iint (x_2 x'_2 - y_2 y'_2)^2 \, d\gamma_K(x, y) d\gamma_K(x', y'). \tag{A3}$$

Now, the 2 terms depend only on (x_2, y_2) and (x'_2, y'_2) . We will project on the two first coordinates, i.e., let $\pi^{1,2}(x, y) = ((x_1, x_2), (y_1, y_2))$ and $\gamma^{1,2} = \pi_{\#}^{1,2} \gamma$, $\gamma_K^{1,2} = \pi_{\#}^{1,2} \gamma_K$. Using the disintegration of measures, we know that there exist kernels $\gamma^{2|1}$ and $\gamma_K^{2|1}$ such that $\gamma^{1,2} = \gamma^1 \otimes \gamma^{2|1}$ and $\gamma_K^{1,2} = \gamma_K^1 \otimes \gamma_K^{2|1}$, where

$$\forall A \in \mathcal{B}(X \times Y), \mu \otimes K(A) = \iint \mathbb{1}_A(x, y) K(x, dy) \mu(dx).$$

We can rewrite the previous Equation (A3) as

$$\begin{aligned} & \iint (x_2 x'_2 - y_2 y'_2)^2 \, d\gamma(x, y) d\gamma(x', y') \\ &= \iiint \iint (x_2 x'_2 - y_2 y'_2)^2 \, \gamma^{2|1}((x_1, y_1), (dx_2, dy_2)) \gamma^{2|1}((x'_1, y'_1), (dx'_2, dy'_2)) \\ & \, d\gamma^1(x_1, y_1) d\gamma^1(x'_1, y'_1) \tag{A4} \\ &\leq \iiint \iint (x_2 x'_2 - y_2 y'_2)^2 \, \gamma_K^{2|1}((x_1, y_1), (dx_2, dy_2)) \gamma_K^{2|1}((x'_1, y'_1), (dx'_2, dy'_2)) \\ & \, d\gamma_K^1(x_1, y_1) d\gamma_K^1(x'_1, y'_1). \end{aligned}$$

Now, we will assume at first that the marginals of $\gamma^{2|1}((x_1, y_1), \cdot)$ are well $\mu^{2|1}(x_1, \cdot)$ and $\nu^{2|1}(y_1, \cdot)$. Then, by definition of $\gamma_K^{2|1}$, as it is optimal for the GW cost with inner products, we have for all $(x_1, y_1), (x'_1, y'_1)$,

$$\begin{aligned} & \iint (x_2 x'_2 - y_2 y'_2)^2 \, \gamma_K^{2|1}((x_1, y_1), (dx_2, dy_2)) \gamma_K^{2|1}((x'_1, y'_1), (dx'_2, dy'_2)) \\ &\leq \iint (x_2 x'_2 - y_2 y'_2)^2 \, \gamma^{2|1}((x_1, y_1), (dx_2, dy_2)) \gamma^{2|1}((x'_1, y'_1), (dx'_2, dy'_2)). \tag{A5} \end{aligned}$$

Moreover, we know from the first part that $\gamma^1 = \gamma_K^1$, then by integrating with respect to (x_1, y_1) and (x'_1, y'_1) , we have

$$\begin{aligned} & \iiint (x_2 x'_2 - y_2 y'_2)^2 \gamma_K^{2|1}((x_1, y_1), (dx_2, dy_2)) \gamma_K^{2|1}((x'_1, y'_1), (dx'_2, dy'_2)) \\ & d\gamma^1(x_1, y_1) d\gamma^1(x'_1, y'_1) \\ & \leq \iiint (x_2 x'_2 - y_2 y'_2)^2 \gamma^{2|1}((x_1, y_1), (dx_2, dy_2)) \gamma^{2|1}((x'_1, y'_1), (dx'_2, dy'_2)) \\ & d\gamma^1(x_1, y_1) d\gamma^1(x'_1, y'_1). \end{aligned} \tag{A6}$$

By (A4) and (A6), we deduce that we have an equality and we get

$$\begin{aligned} & \iint \left(\iint (x_2 x'_2 - y_2 y'_2)^2 \gamma^{2|1}((x_1, y_1), (dx_2, dy_2)) \gamma^{2|1}((x'_1, y'_1), (dx'_2, dy'_2)) \right. \\ & \left. - \iint (x_2 x'_2 - y_2 y'_2)^2 \gamma_K^{2|1}((x_1, y_1), (dx_2, dy_2)) \gamma_K^{2|1}((x'_1, y'_1), (dx'_2, dy'_2)) \right) \\ & d\gamma^1(x_1, y_1) d\gamma^1(x'_1, y'_1) = 0. \end{aligned} \tag{A7}$$

However, we know by (A5) that the middle part of (A7) is nonnegative, thus we have for γ^1 -a.e. $(x_1, y_1), (x'_1, y'_1)$,

$$\begin{aligned} & \iint (x_2 x'_2 - y_2 y'_2)^2 \gamma_K^{2|1}((x_1, y_1), (dx_2, dy_2)) \gamma_K^{2|1}((x'_1, y'_1), (dx'_2, dy'_2)) \\ & = \iint (x_2 x'_2 - y_2 y'_2)^2 \gamma^{2|1}((x_1, y_1), (dx_2, dy_2)) \gamma^{2|1}((x'_1, y'_1), (dx'_2, dy'_2)). \end{aligned}$$

From that, we can conclude as in the first part that $\gamma^{2|1} = \gamma_K^{2|1}$ (by unicity of the optimal map). And thus $\gamma^{1,2} = \gamma_K^{1,2}$.

Now, we still have to show that the marginals of $\gamma^{2|1}((x_1, y_1), \cdot)$ and $\gamma_K^{2|1}((x_1, y_1), \cdot)$ are well the same, i.e., $\mu^{2|1}(x_1, \cdot)$ and $\nu^{2|1}(y_1, \cdot)$. Let ϕ and ψ be continuous functions, then we have to show that for γ^1 -a.e. (x_1, y_1) , we have

$$\begin{cases} \int \phi(x_2) \gamma^{2|1}((x_1, y_1), (dx_2, dy_2)) = \int \phi(x_2) \mu^{2|1}(x_1, dx_2) \\ \int \psi(y_2) \gamma^{2|1}((x_1, y_1), (dx_2, dy_2)) = \int \psi(y_2) \nu^{2|1}(y_1, dy_2). \end{cases}$$

As we want to prove it for γ^1 -a.e. (x_1, y_1) , it is sufficient to prove that for all continuous function ξ ,

$$\begin{cases} \iint \xi(x_1, y_1) \phi(x_2) \gamma^{2|1}((x_1, y_1), (dx_2, dy_2)) d\gamma^1(x_1, y_1) \\ = \iint \xi(x_1, y_1) \phi(x_2) \mu^{2|1}(x_1, dx_2) d\gamma^1(x_1, y_1) \\ \iint \xi(x_1, y_1) \psi(y_2) \gamma^{2|1}((x_1, y_1), (dx_2, dy_2)) d\gamma^1(x_1, y_1) \\ = \iint \xi(x_1, y_1) \psi(y_2) \nu^{2|1}(y_1, dy_2) d\gamma^1(x_1, y_1). \end{cases}$$

First, we can use the projections $\pi_x(x, y) = x$ and $\pi_y(x, y) = y$. Moreover, we know that $\gamma^1 = (Id \times T_K^1) \# \mu^1$. The alternate Knothe–Rosenblatt rearrangement is, as the usual one, bijective (because μ and ν are absolutely continuous), and thus, as we suppose that ν satisfies the same hypothesis than μ , we also have $\gamma^1 = ((T_K^1)^{-1}, Id) \# \nu^1$. Let us note $\tilde{T}_K^1 = (T_K^1)^{-1}$. Then, the equalities that we want to show are:

$$\begin{cases} \iint \xi(x_1, T_K^1(x_1)) \phi(x_2) \gamma_x^{2|1}((x_1, T_K^1(x_1)), dx_2) d\mu^1(x_1) \\ = \iint \xi(x_1, T_K^1(x_1)) \phi(x_2) \mu^{2|1}(x_1, dx_2) d\mu^1(x_1) \\ \iint \xi(\tilde{T}_K^1(y_1), y_1) \psi(y_2) \gamma_y^{2|1}((\tilde{T}_K^1(y_1), y_1), dy_2) d\nu^1(y_1) \\ = \iint \xi(\tilde{T}_K^1(y_1), y_1) \psi(y_2) \nu^{2|1}(y_1, dy_2) d\nu^1(y_1). \end{cases}$$

In addition, we have indeed

$$\begin{aligned} & \iint \zeta(x_1, T_K^1(x_1)) \phi(x_2) \gamma_x^{2|1}((x_1, T_K^1(x_1)), dx_2) d\mu^1(x_1) \\ &= \iint \zeta(x_1, T_K^1(x_1)) \phi(x_2) d\gamma^{1,2}((x_1, x_2), (y_1, y_2)) \\ &= \iint \zeta(x_1, T_K^1(x_1)) \phi(x_2) d\gamma_x^{1,2}(x_1, x_2) \\ &= \iint \zeta(x_1, T_K^1(x_1)) \phi(x_2) \mu^{2|1}(x_1, dx_2) d\mu^1(x_1). \end{aligned}$$

We can do the same for the ν part by symmetry.

Part 3:

Now, we can proceed the same way by induction. Let $\ell \in \{2, \dots, d\}$ and suppose that the result is true in dimension $\ell - 1$ (i.e., $\gamma^{1:\ell-1} = \pi_{\#}^{1:\ell-1} \gamma = \gamma_K^{1:\ell-1}$).

For this part of the proof, we rely on [19] (Theorem 2.23). We can build a measure $\gamma_K^t \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$ such that:

$$\begin{cases} \pi_{\#}^x \gamma_K^t = \mu \\ \pi_{\#}^y \gamma_K^t = \nu \\ \pi_{\#}^{1:\ell-1} \gamma_K^t = \eta_{t,\ell} \end{cases} \tag{A8}$$

where $\eta_{t,\ell}$ is the optimal transport plan between $\mu^\ell = \pi_{\#}^{1:\ell-1} \mu$ and $\nu^\ell = \pi_{\#}^{1:\ell-1} \nu$ for the objective:

$$\iint \sum_{k=1}^{\ell-1} \left(\prod_{i=1}^{k-1} \lambda_t^{(i)} \right) (x_k x'_k - y_k y'_k)^2 d\gamma(x, y) d\gamma(x', y').$$

By induction hypothesis, we have $\eta_{t,\ell} \xrightarrow[t \rightarrow 0]{\mathcal{D}} \pi_{\#}^{1:\ell-1} \gamma_K$. To build such a measure, we can first disintegrate μ and ν :

$$\begin{cases} \mu = \mu^{1:\ell-1} \otimes \mu^{\ell:d|1:\ell-1} \\ \nu = \nu^{1:\ell-1} \otimes \nu^{\ell:d|1:\ell-1}, \end{cases}$$

then we pick the Knothe transport $\gamma_K^{\ell:d|1:\ell-1}$ between $\mu^{\ell:d|1:\ell-1}$ and $\nu^{\ell:d|1:\ell-1}$. Thus, by taking $\gamma_K^T = \eta_{t,\ell} \otimes \gamma_K^{\ell:d|1:\ell-1}$, γ_K^T satisfies the conditions well (A8).

Hence, we have:

$$\begin{aligned} & \iint \sum_{k=1}^{\ell-1} \left(\prod_{i=1}^{k-1} \lambda_t^{(i)} \right) (x_k x'_k - y_k y'_k)^2 d\gamma_K^t(x, y) d\gamma_K^t(x', y') \\ &= \iint \sum_{k=1}^{\ell-1} \left(\prod_{i=1}^{k-1} \lambda_t^{(i)} \right) (x_k x'_k - y_k y'_k)^2 d\eta_{t,\ell}(x_{1:\ell-1}, y_{1:\ell-1}) d\eta_{t,\ell}(x'_{1:\ell-1}, y'_{1:\ell-1}) \\ &\leq \iint \sum_{k=1}^{\ell-1} \left(\prod_{i=1}^{k-1} \lambda_t^{(i)} \right) (x_k x'_k - y_k y'_k)^2 d\gamma_t(x, y) d\gamma_t(x', y'), \end{aligned}$$

and therefore:

$$\begin{aligned} & \iint \sum_{k=1}^{\ell-1} \left(\prod_{i=1}^{k-1} \lambda_t^{(i)} \right) (x_k x'_k - y_k y'_k)^2 d\gamma_K^t(x, y) d\gamma_K^t(x', y') \\ & + \iint \sum_{k=\ell}^d \left(\prod_{i=1}^{k-1} \lambda_t^{(i)} \right) (x_k x'_k - y_k y'_k)^2 d\gamma_t(x, y) d\gamma_t(x', y') \\ & \leq \mathcal{HW}_t^2(\mu, \nu) \\ & \leq \iint \sum_{k=1}^{\ell-1} \left(\prod_{i=1}^{k-1} \lambda_t^{(i)} \right) (x_k x'_k - y_k y'_k)^2 d\gamma_K^t(x, y) d\gamma_K^t(x', y') \\ & + \iint \sum_{k=\ell}^d \left(\prod_{i=1}^{k-1} \lambda_t^{(i)} \right) (x_k x'_k - y_k y'_k)^2 d\gamma_K^t(x, y) d\gamma_K^t(x', y'). \end{aligned}$$

As before, by subtracting the first term, dividing by $\prod_{i=1}^{\ell-1} \lambda_t^{(i)}$ and taking the limit, we obtain:

$$\iint (x_\ell x'_\ell - y_\ell y'_\ell)^2 d\gamma_t(x, y) d\gamma_t(x', y') \leq \iint (x_\ell x'_\ell - y_\ell y'_\ell)^2 d\gamma_K^t(x, y) d\gamma_K^t(x', y').$$

For the right hand side, using that $\gamma_K^t = \eta_{t,\ell} \otimes \gamma_K^{\ell:d|1:\ell-1}$, we have:

$$\begin{aligned} & \iint (x_\ell x'_\ell - y_\ell y'_\ell)^2 d\gamma_K^t(x, y) d\gamma_K^t(x', y') \\ & = \iiint \int (x_\ell x'_\ell - y_\ell y'_\ell)^2 \gamma_K^{\ell:d|1:\ell-1}((x_{1:\ell-1}, y_{1:\ell-1}), (dx_{\ell:d}, dy_{\ell:d})) \\ & \gamma_K^{\ell:d|1:\ell-1}((x'_{1:\ell-1}, y'_{1:\ell-1}), (dx'_{\ell:d}, dy'_{\ell:d})) d\eta_{t,\ell}(x_{1:\ell-1}, y_{1:\ell-1}) d\eta_{t,\ell}(x'_{1:\ell-1}, y'_{1:\ell-1}) \\ & = \iiint \int (x_\ell x'_\ell - y_\ell y'_\ell)^2 \gamma_K^{\ell|1:\ell-1}((x_{1:\ell-1}, y_{1:\ell-1}), (dx_\ell, dy_\ell)) \\ & \gamma_K^{\ell|1:\ell-1}((x'_{1:\ell-1}, y'_{1:\ell-1}), (dx'_\ell, dy'_\ell)) d\eta_{t,\ell}(x_{1:\ell-1}, y_{1:\ell-1}) d\eta_{t,\ell}(x'_{1:\ell-1}, y'_{1:\ell-1}). \end{aligned}$$

Let us note for $\eta_{t,\ell}$ almost every $(x_{1:\ell-1}, y_{1:\ell-1}), (x'_{1:\ell-1}, y'_{1:\ell-1})$

$$\begin{aligned} & \text{GW}(\mu^{\ell|1:\ell-1}, \nu^{\ell|1:\ell-1}) \\ & = \iint (x_\ell x'_\ell - y_\ell y'_\ell)^2 \gamma_K^{\ell|1:\ell-1}((x_{1:\ell-1}, y_{1:\ell-1}), (dx_\ell, dy_\ell)) \gamma_K^{\ell|1:\ell-1}((x'_{1:\ell-1}, y'_{1:\ell-1}), (dx'_\ell, dy'_\ell)), \end{aligned}$$

then

$$\begin{aligned} & \iint (x_\ell x'_\ell - y_\ell y'_\ell)^2 d\gamma_K^t(x, y) d\gamma_K^t(x', y') \\ & = \iint \text{GW}(\mu^{\ell|1:\ell-1}, \nu^{\ell|1:\ell-1}) d\eta_{t,\ell}(x_{1:\ell-1}, y_{1:\ell-1}) d\eta_{t,\ell}(x'_{1:\ell-1}, y'_{1:\ell-1}). \end{aligned}$$

By Theorem A1, we have $\eta_{t,\ell} \otimes \eta_{t,\ell} \xrightarrow[t \rightarrow 0]{\mathcal{D}} \pi_{\#}^{1:\ell-1} \gamma_K \otimes \pi_{\#}^{1:\ell-1} \gamma_K$. So, if

$$\eta \mapsto \iint \text{GW}(\mu^{\ell|1:\ell-1}, \nu^{\ell|1:\ell-1}) d\eta d\eta \tag{A9}$$

is continuous over the transport plans between $\mu^{1:\ell-1}$ and $\nu^{1:\ell-1}$, we have

$$\begin{aligned} & \iint (x_\ell x'_\ell - y_\ell y'_\ell)^2 d\gamma_K^t(x, y) d\gamma_K^t(x', y') \\ & \xrightarrow[t \rightarrow 0]{} \iint \text{GW}(\mu^{\ell|1:\ell-1}, \nu^{\ell|1:\ell-1}) \pi_{\#}^{1:\ell-1} \gamma_K(dx_{1:\ell-1}, dy_{1:\ell-1}) \pi_{\#}^{1:\ell-1} \gamma_K(dx'_{1:\ell-1}, dy'_{1:\ell-1}) \end{aligned}$$

and

$$\begin{aligned} & \iint GW(\mu^{\ell|1:\ell-1}, \nu^{\ell|1:\ell-1}) \pi_{\#}^{1:\ell-1} \gamma_K(dx_{1:\ell-1}, dy_{1:\ell-1}) \pi_{\#}^{1:\ell-1} \gamma_K(dx'_{1:\ell-1}, dy'_{1:\ell-1}) \\ &= \iint (x_{\ell} x'_{\ell} - y_{\ell} y'_{\ell})^2 d\gamma_K(x, y) d\gamma_K(x', y') \end{aligned}$$

by replacing the true expression of GW and using the disintegration $\gamma_K = (\pi_K^{1:\ell-1})_{\#} \gamma_K \otimes \gamma_K^{\ell|1:\ell-1}$.

For the continuity, we can apply [19] (Lemma 1.8) (as in the [19] (Corollary 2.24)) with $X = Y = \mathbb{R}^{\ell-1} \times \mathbb{R}^{\ell-1}$, $\tilde{X} = \tilde{Y} = \mathcal{P}(\Omega)$ with $\Omega \subset \mathbb{R}^{d-\ell+1} \times \mathbb{R}^{d-\ell+1}$ and $c(a, b) = GW(a, b)$, which can be bounded on compact supports by $\max |c|$. Moreover, we use Theorem A1 and the fact that $\eta_t \otimes \eta_t \xrightarrow[t \rightarrow 0]{\mathcal{D}} \gamma_K^{1:\ell-1} \otimes \gamma_K^{1:\ell-1}$.

By taking the limit $t \rightarrow 0$, we now obtain:

$$\iint (x_{\ell} x'_{\ell} - y_{\ell} y'_{\ell})^2 d\gamma(x, y) d\gamma(x', y') \leq \iint (x_{\ell} x'_{\ell} - y_{\ell} y'_{\ell})^2 d\gamma_K(x, y) d\gamma_K(x', y').$$

We can now disintegrate with respect to $\gamma^{1:\ell-1}$ as before. We just need to prove that the marginals coincide, which is performed by taking for test functions:

$$\begin{cases} \xi(x_1, \dots, x_{\ell-1}, y_1, \dots, y_{\ell-1}) \phi(x_{\ell}) \\ \xi(x_1, \dots, x_{\ell-1}, y_1, \dots, y_{\ell-1}) \psi(y_{\ell}) \end{cases}$$

and using the fact that the measures are concentrated on $y_k = T_K(x_k)$.

Part 4:

Therefore, we have well $\gamma_t \xrightarrow[t \rightarrow 0]{\mathcal{D}} \gamma_K$. Finally, for the L^2 convergence, we have:

$$\int \|T_t(x) - T_K(x)\|_2^2 \mu(dx) = \int \|y - T_K(x)\|_2^2 d\gamma_t(x, y) \rightarrow \int \|y - T_K(x)\|_2^2 d\gamma_K(x, y) = 0$$

as $\gamma_t = (Id \times T_t)_{\#} \mu$ and $\gamma_K = (Id \times T_K)_{\#} \mu$. Hence, $T_t \xrightarrow[t \rightarrow 0]{L^2} T_K$. \square

References

1. Arjovsky, M.; Chintala, S.; Bottou, L. Wasserstein generative adversarial networks. In Proceedings of the International Conference on Machine Learning, PMLR, Sydney, Australia, 6–11 August 2017; pp. 214–223.
2. Courty, N.; Flamary, R.; Tuia, D.; Rakotomamonjy, A. Optimal transport for domain adaptation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 1853–1865. [[CrossRef](#)] [[PubMed](#)]
3. Mémoli, F. Gromov–Wasserstein distances and the metric approach to object matching. *Found. Comput. Math.* **2011**, *11*, 417–487. [[CrossRef](#)]
4. Chowdhury, S.; Miller, D.; Needham, T. Quantized Gromov–Wasserstein. In Proceedings of the Machine Learning and Knowledge Discovery in Databases, Research Track—European Conference, ECML PKDD 2021, Bilbao, Spain, 13–17 September 2021; Proceedings, Part III; Oliver, N., Pérez-Cruz, F., Kramer, S., Read, J., Lozano, J.A., Eds.; Springer: Berlin/Heidelberg, Germany, 2021; Volume 12977, pp. 811–827. [[CrossRef](#)]
5. Alvarez-Melis, D.; Jegelka, S.; Jaakkola, T.S. Towards optimal transport with global invariances. In Proceedings of the The 22nd International Conference on Artificial Intelligence and Statistics, PMLR, Naha, Japan, 16–18 April 2019; pp. 1870–1879.
6. Cai, Y.; Lim, L.H. Distances between probability distributions of different dimensions. *arXiv* **2020**, arXiv:2011.00629.
7. Mémoli, F. On the use of Gromov–Hausdorff Distances for Shape Comparison. In Proceedings of the 4th Symposium on Point Based Graphics, PBG@Eurographics 2007, Prague, Czech Republic, 2–3 September 2007; Botsch, M., Pajarola, R., Chen, B., Zwicker, M., Eds.; Eurographics Association: Switzerland, Geneva, 2007; pp. 81–90. [[CrossRef](#)]
8. Sturm, K.T. The space of spaces: Curvature bounds and gradient flows on the space of metric measure spaces. *arXiv* **2012**, arXiv:1208.0434.
9. Peyré, G.; Cuturi, M.; Solomon, J. Gromov–Wasserstein averaging of kernel and distance matrices. In Proceedings of the International Conference on Machine Learning, PMLR, New York, NY, USA, 19–24 June 2016; pp. 2664–2672.
10. Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. *Adv. Neural Inf. Process. Syst.* **2013**, *26*, 2292–2300.

11. Scetbon, M.; Peyré, G.; Cuturi, M. Linear-Time Gromov Wasserstein Distances using Low Rank Couplings and Costs. *arXiv* **2021**, arXiv:2106.01128.
12. Vayer, T.; Flamary, R.; Courty, N.; Tavenard, R.; Chapel, L. Sliced Gromov–Wasserstein. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 14753–14763.
13. Fatras, K.; Zine, Y.; Majewski, S.; Flamary, R.; Gribonval, R.; Courty, N. Minibatch optimal transport distances; analysis and applications. *arXiv* **2021**, arXiv:2101.01792.
14. Muzellec, B.; Cuturi, M. Subspace Detours: Building Transport Plans that are Optimal on Subspace Projections. In Proceedings of the Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, Vancouver, BC, Canada, 8–14 December 2019; Wallach, H.M., Larochelle, H., Beygelzimer, A., d’Alché-Buc, F., Fox, E.B., Garnett, R., Eds.; 2019; pp. 6914–6925.
15. Bogachev, V.I.; Kolesnikov, A.V.; Medvedev, K.V. Triangular transformations of measures. *Sb. Math.* **2005**, *196*, 309. [[CrossRef](#)]
16. Knothe, H. Contributions to the theory of convex bodies. *Mich. Math. J.* **1957**, *4*, 39–52. [[CrossRef](#)]
17. Rosenblatt, M. Remarks on a multivariate transformation. *Ann. Math. Stat.* **1952**, *23*, 470–472. [[CrossRef](#)]
18. Villani, C. *Optimal Transport: Old and New*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2008; Volume 338.
19. Santambrogio, F. Optimal transport for applied mathematicians. *Birkhäuser NY* **2015**, *55*, 94.
20. Jaini, P.; Selby, K.A.; Yu, Y. Sum-of-Squares Polynomial Flow. In Proceedings of the 36th International Conference on Machine Learning, PMLR, Long Beach, CA, USA, 9–15 June 2019; pp. 3009–3018.
21. Carlier, G.; Galichon, A.; Santambrogio, F. From Knothe’s transport to Brenier’s map and a continuation method for optimal transport. *SIAM J. Math. Anal.* **2010**, *41*, 2554–2576. [[CrossRef](#)]
22. Bonnotte, N. Unidimensional and Evolution Methods for Optimal Transportation. Ph.D. Thesis, Université Paris-Sud, Paris, France, 2013.
23. Ambrosio, L.; Gigli, N.; Savaré, G. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2008.
24. Niles-Weed, J.; Rigollet, P. Estimation of wasserstein distances in the spiked transport model. *arXiv* **2019**, arXiv:1909.07513.
25. Chowdhury, S.; Mémoli, F. The gromov–wasserstein distance between networks and stable network invariants. *Inf. Inference A J. IMA* **2019**, *8*, 757–787. [[CrossRef](#)]
26. Vayer, T. A Contribution to Optimal Transport on Incomparable Spaces. Ph.D. Thesis, Université de Bretagne Sud, Vannes, France, 2020.
27. Paty, F.P.; Cuturi, M. Subspace robust wasserstein distances. In Proceedings of the 36th International Conference on Machine Learning, PMLR, Long Beach, CA, USA, 9–15 June 2019; pp. 5072–5081.
28. Rasmussen, C.E. Gaussian processes in machine learning. In *Summer School on Machine Learning*; Springer: Berlin/Heidelberg, Germany, 2003; pp. 63–71.
29. Von Mises, R. *Mathematical Theory of Probability and Statistics*; Academic Press: Cambridge, MA, USA, 1964.
30. Salmona, A.; Delon, J.; Desolneux, A. Gromov–Wasserstein Distances between Gaussian Distributions. *arXiv* **2021**, arXiv:2104.07970.
31. Weed, J.; Bach, F. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli* **2019**, *25*, 2620–2648. [[CrossRef](#)]
32. Lin, T.; Zheng, Z.; Chen, E.; Cuturi, M.; Jordan, M. On projection robust optimal transport: Sample complexity and model misspecification. In Proceedings of the International Conference on Artificial Intelligence and Statistics, PMLR, Virtual, 13–15 April 2021; pp. 262–270.
33. Burkard, R.E.; Klinz, B.; Rudolf, R. Perspectives of Monge Properties in Optimization. *Discret. Appl. Math.* **1996**, *70*, 95–161. [[CrossRef](#)]
34. Flamary, R.; Courty, N.; Gramfort, A.; Alaya, M.Z.; Boisbunon, A.; Chambon, S.; Chapel, L.; Corenflos, A.; Fatras, K.; Fournier, N.; et al. POT: Python Optimal Transport. *J. Mach. Learn. Res.* **2021**, *22*, 1–8.
35. Bogo, F.; Romero, J.; Loper, M.; Black, M.J. FAUST: Dataset and evaluation for 3D mesh registration. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 3794–3801.
36. Fiedler, M. Algebraic connectivity of graphs. *Czechoslov. Math. J.* **1973**, *23*, 298–305. [[CrossRef](#)]
37. Hagberg, A.; Swart, P.; Chult, D.S. *Exploring Network Structure, Dynamics, and Function Using NetworkX*; Technical Report; Los Alamos National Lab. (LANL): Los Alamos, NM, USA, 2008.
38. Wu, J.P.; Song, J.Q.; Zhang, W.M. An efficient and accurate method to compute the Fiedler vector based on Householder deflation and inverse power iteration. *J. Comput. Appl. Math.* **2014**, *269*, 101–108. [[CrossRef](#)]
39. Xu, H.; Luo, D.; Carin, L. Scalable Gromov–Wasserstein learning for graph partitioning and matching. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 3052–3062.
40. Chowdhury, S.; Needham, T. Generalized spectral clustering via Gromov–Wasserstein learning. In Proceedings of the International Conference on Artificial Intelligence and Statistics, PMLR, Virtual, 13–15 April 2021; pp. 712–720.
41. Vayer, T.; Courty, N.; Tavenard, R.; Flamary, R. Optimal transport for structured data with application on graphs. In Proceedings of the 36th International Conference on Machine Learning, PMLR, Long Beach, CA, USA, 9–15 June 2019; pp. 6275–6284.
42. Lacoste-Julien, S. Convergence rate of frank-wolfe for non-convex objectives. *arXiv* **2016**, arXiv:1607.00345.

-
43. Peyré, G.; Cuturi, M. Computational optimal transport: With applications to data science. *Found. Trends[®] Mach. Learn.* **2019**, *11*, 355–607. [[CrossRef](#)]
 44. Nagar, R.; Raman, S. Detecting approximate reflection symmetry in a point set using optimization on manifold. *IEEE Trans. Signal Process.* **2019**, *67*, 1582–1595. [[CrossRef](#)]
 45. Billingsley, P. *Convergence of Probability Measures*; John Wiley & Sons: Hoboken, NJ, USA, 2013.