Busemann Functions in the Wasserstein Space: Existence, Closed-Forms, and Applications to Slicing

Clément Bonet
CMAP
Ecole Polytechnique

Elsa Cazelles
CNRS, IRIT
Université de Toulouse

Lucas Drumetz
Lab-STICC
IMT Atlantique

Nicolas Courty
IRISA
Université Bretagne Sud

Abstract

The Busemann function has recently found much interest in a variety of geometric machine learning problems, as it naturally defines projections onto geodesic rays of Riemannian manifolds and generalizes the notion of hyperplanes. As several sources of data can be conveniently modeled as probability distributions, it is natural to study this function in the Wasserstein space, which carries a rich formal Riemannian structure induced by Optimal Transport metrics. In this work, we investigate the existence and computation of Busemann functions in Wasserstein space, which admits geodesic rays. We establish closed-form expressions in two important cases: one-dimensional distributions and Gaussian measures. These results enable explicit projection schemes for probability distributions on $\mathbb{R}$, which in turn allow us to define novel Sliced-Wasserstein distances over Gaussian mixtures and labeled datasets. We demonstrate the efficiency of those original schemes on synthetic datasets as well as transfer learning problems.

1 INTRODUCTION

The Busemann function, introduced by Busemann (1955), provides a natural generalization of affine functions on non-compact metric spaces admitting geodesics which can be extended to infinity. As such, its level sets generalize the notion of affine hyperplanes and so it provides a reliable means of projecting onto geodesics. Thus, it has recently received a lot of attention in geometric Machine Learning approaches, which aim at extending classical Euclidean algorithms towards different spaces such as manifolds (Bronstein et al., 2017; Papillon et al., 2025) for data analysis purposes.

In particular, the Busemann function is well defined on geodesically complete spaces, on which geodesics can be extended to infinity in both directions. Although this rules out compact manifolds, such spaces include for instance hyperbolic manifolds, on which the Busemann function has been widely used to perform Principal Component Analysis (Chami et al., 2021), to characterize directions and perform classification with prototypes (Ghadimi Atigh et al., 2021; Durrant and Leontidis, 2023; Berg et al., 2024, 2025), to define decision boundaries for classification (Fan et al., 2023; Doorenbos et al., 2024), to define layers of neural networks (Wang, 2021; Sonoda et al., 2022; Nguyen et al., 2025b), or as a projection operator on geodesics in order to define a Sliced-Wasserstein distance (Bonet et al., 2023a, 2025b).

These successes in hyperbolic geometry suggest exploring the role of the Busemann function in other non-Euclidean settings. Many real-world data are best modeled as probability distributions. This is the case for instance for documents that are distributions of words (Kusner et al., 2015), single-cells (Bellazzi et al., 2021; Haviv et al., 2025b), point clouds (Haviv et al., 2025a; Geuter et al., 2025), or even images (Seguy and Cuturi, 2015). Moreover, Gaussian mixtures (Chen et al., 2018; Delon and Desolneux, 2020) or datasets with discrete labels that can be represented as mixtures of discrete distributions (Alvarez-Melis and Fusi, 2020; Bonet et al., 2025a), can be seen as datasets of probability distributions. One powerful way to endow the space of probability distributions with a metric consists in using Optimal Transport (OT) and, in particular, the Wasserstein distance (Villani, 2009). This distance allows to define the Wasserstein space, which enjoys a very rich geometry thoroughly studied in the last decades, see *e.g.* (Ambrosio et al., 2008; Vil-

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

lani, 2009; Santambrogio, 2015). Notably, it carries a formal Riemannian structure, and admits geodesics. Thus, the study of Busemann functions on the Wasserstein space is especially compelling for data analysis. A key challenge, however, is that the Wasserstein space is not geodesically complete, which prevents defining the Busemann function along every geodesic. Fortunately, for any base measure μ_0, there is always at least one geodesic starting from μ_0 that can be extended to infinity in one direction (Zhu et al., 2021).

Contributions. In this work, we first provide sufficient conditions to characterize geodesic rays on the Wasserstein space, *i.e* geodesics that can be extended in one direction. We then investigate how to compute the Busemann function along such geodesics. In full generality, we show that this computation reduces to solving an OT problem. In specific cases such as one-dimensional or Gaussian distributions, the Busemann function also admits closed-form expressions. Leveraging these closed-forms, we introduce new sliced distances between labeled datasets. Our results show a strong correlation with classical distances between datasets while being more computationally efficient. Finally, minimizing these distances allows to flow datasets, which we apply in a transfer learning setting.

2 WASSERSTEIN SPACE

In this section, we introduce the Wasserstein space and its associated Riemannian structure. Then, we study characterizations of geodesic rays on this space, *i.e.* geodesics that can be extended to infinity in one direction. In particular, we focus on absolutely continuous probability measures, one dimensional probability measures, and Gaussians.

2.1 Wasserstein Distance

Let $\mathcal{P}_2(\mathbb{R}^d) = \{\mu \in \mathcal{P}(\mathbb{R}^d),\ \int \|x\|_2^2 \,\mathrm{d}\mu(x) < \infty\}$ be the space of probability measures in $\mathbb{R}^d$ with finite second moments. Optimal Transport (OT) provides a principled way to define a distance on this space through the 2-Wasserstein distance (Villani, 2009), defined for all $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ as

$$\mathrm{W}_2^2(\mu,\nu) = \inf_{\gamma\in\Pi(\mu,\nu)} \int \|x-y\|_2^2 \,\mathrm{d}\gamma(x,y), \qquad (1)$$

where $\Pi(\mu,\nu) = \{\gamma \in \mathcal{P}(\mathbb{R}^d\times\mathbb{R}^d),\ \pi^1_\#\gamma = \mu,\ \pi^2_\#\gamma = \nu\}$ is the set of couplings between μ and ν, $\pi^1 : (x,y) \mapsto x$ and $\pi^2 : (x,y) \mapsto y$ are the projections on the coordinates, and $\#$ is the push-forward operator defined such that $\mathrm{T}_\#\mu(A) = \mu\big(\mathrm{T}^{-1}(A)\big)$ for a measurable map $\mathrm{T} : \mathbb{R}^d \to \mathbb{R}^{d'}$, $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ and Borelian $A \in \mathcal{B}(\mathbb{R}^{d'})$. In particular, W_2 is a well defined distance, and $(\mathcal{P}_2(\mathbb{R}^d), \mathrm{W}_2)$ is called the Wasserstein space.

The OT problem introduces two objects of interest: the distance W_2 and the optimal coupling $\gamma^* \in \Pi_o(\mu,\nu)$ solving (1). In the particular case where $\mu \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ is absolutely continuous with respect to the Lebesgue measure, it is well known by Brenier's theorem (Brenier, 1991) that the optimal coupling γ^* is unique and supported on the graph of a map T, called the Monge map, *i.e.* $\gamma^* = (\mathrm{Id}, \mathrm{T})_\#\mu$ with T satisfying $\mathrm{T}_\#\mu = \nu$ and $\mathrm{Id} : x \mapsto x$ the identity function.

Additionally, there are a few specific cases for which the Wasserstein distance can be computed in closed-form. In dimension $d = 1$, the Wasserstein distance is only the L^2 norm between the quantile functions, *i.e.* for all $\mu, \nu \in \mathcal{P}_2(\mathbb{R})$,

$$\mathrm{W}_2^2(\mu,\nu) = \int_0^1 \left|F_\mu^{-1}(u) - F_\nu^{-1}(u)\right|^2 \mathrm{d}u, \qquad (2)$$

where F_μ^{-1} and F_ν^{-1} denote the quantile functions of μ and ν. The optimal coupling is then obtained as $\gamma^* = (F_\mu^{-1}, F_\nu^{-1})_\#\mathrm{Unif}([0,1])$, and if μ is absolutely continuous, the OT map is the increasing rearrangement $\mathrm{T} = F_\nu^{-1} \circ F_\mu$ (Santambrogio, 2015, Theorem 2.9). In higher dimensions, we usually do not have a closed-form, except in particular cases such as the Wasserstein distance between two Gaussian distributions (Givens and Shortt, 1984; Gelbrich, 1990). Namely, for $\mu = \mathcal{N}(m_\mu, \Sigma_\mu)$, $\nu = \mathcal{N}(m_\nu, \Sigma_\nu)$ two Gaussian distributions with respective means $m_\mu, m_\nu \in \mathbb{R}^d$ and positive definite covariance matrices $\Sigma_\mu, \Sigma_\nu \in S_d^{++}(\mathbb{R})$, we get,

$$\mathrm{W}_2^2(\mu,\nu) = \|m_\mu - m_\nu\|_2^2 + \mathcal{B}^2(\Sigma_\mu, \Sigma_\nu), \qquad (3)$$

where $\mathcal{B}$ defines a distance between positive semi-definite matrices, known in the literature of quantum information as the Bures distance (Bhatia et al., 2019), and is of the form

$$\mathcal{B}^2(\Sigma_\mu, \Sigma_\nu) = \mathrm{Tr}\left(\Sigma_\mu + \Sigma_\nu - 2(\Sigma_\mu^{\frac{1}{2}}\Sigma_\nu\Sigma_\mu^{\frac{1}{2}})^{\frac{1}{2}}\right). \qquad (4)$$

Thus, we refer to the Wasserstein distance between Gaussians as the Bures-Wasserstein distance BW, and the space of Gaussians endowed with BW is called the Bures-Wasserstein space $\mathrm{BW}(\mathbb{R}^d)$. Furthermore, the OT map between μ and ν is of the form $\mathrm{T} : x \mapsto m_\nu + A_\mu^\nu(x - m_\mu)$ (Peyré and Cuturi, 2019, Remark 2.31) where

$$A_\mu^\nu = \Sigma_\mu^{-\frac{1}{2}}(\Sigma_\mu^{\frac{1}{2}}\Sigma_\nu\Sigma_\mu^{\frac{1}{2}})^{\frac{1}{2}}\Sigma_\mu^{-\frac{1}{2}}. \qquad (5)$$

Riemannian Structure. It is well known that the Wasserstein space has a formal Riemannian structure (Otto, 2001). In particular, it is a geodesic space: for any measures $\mu_0, \mu_1 \in \mathcal{P}_2(\mathbb{R}^d)$, there is at least one constant-speed geodesic, *i.e.* a continuous curve

$t \in [0,1] \mapsto \mu_t \in \mathcal{P}_2(\mathbb{R}^d)$ interpolating between μ_0 and μ_1, and satisfying

$$\forall s,t \in [0,1],\ \mathrm{W}_2(\mu_t,\mu_s) = |t-s|\mathrm{W}_2(\mu_0,\mu_1). \quad (6)$$

We call $\kappa_\mu = \mathrm{W}_2(\mu_0,\mu_1)$ the speed of the geodesic $(\mu_t)_{t\in[0,1]}$. Such a curve is always a displacement interpolation (Bertrand and Kloeckner, 2012, Proposition 2.9), *i.e.* it is of the form (McCann, 1997)

$$\forall t \in [0,1],\ \mu_t = \big((1-t)\pi^1 + t\pi^2\big)_\# \gamma^*, \quad (7)$$

where $\gamma^* \in \Pi_o(\mu_0,\mu_1)$, and is fully characterized by μ_0, μ_1 and γ^*. In the case where an OT map T exists between μ_0 and μ_1, *e.g.* if $\mu_0 \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$, then the geodesic curve can be further written as

$$\forall t \in [0,1],\ \mu_t = \big((1-t)\mathrm{Id} + t\mathrm{T}\big)_\# \mu_0. \quad (8)$$

If the geodesic can be extended for any $t \in \mathbb{R}$, *i.e.* (6) is satisfied for any $s,t \in \mathbb{R}$, it is called a geodesic line. If (6) holds for any $s,t \in \mathbb{R}_+$, it is called a geodesic ray (Bridson and Haefliger, 2013).

2.2 Geodesic Rays on $(\mathcal{P}_2(\mathbb{R}^d), \mathrm{W}_2)$

Kloeckner (2010) first studied the conditions on the measures μ_0 and μ_1 under which the geodesics connecting them can be extended. For instance, in (Kloeckner, 2010, Proposition 3.6), it was shown that the geodesic curve $t \mapsto \mu_t$ is a geodesic line if and only if μ_1 is a translation of μ_0. Consequently, constructing geodesic lines is very restrictive. Geodesic rays are more flexible, and are also the appropriate object that allows defining Busemann functions: in (Zhu et al., 2021), it is proved that for any $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$, there exists at least one geodesic ray originating from it.

In this paper, we discuss how to characterize geodesic rays on $(\mathcal{P}_2(\mathbb{R}^d), \mathrm{W}_2)$. First, if $\mu_0 = \delta_{x_0}$ for some $x_0 \in \mathbb{R}^d$, then by (Bertrand and Kloeckner, 2016, Lemma 2.1), we can extend any geodesic starting from μ_0 and passing through $\mu_1 \in \mathcal{P}_2(\mathbb{R}^d)$ as a geodesic ray of the form $\mu_t = \big((1-t)x_0 + t\mathrm{Id}\big)_\#\mu_1$ for any $t \geq 0$, since the optimal coupling in this case is the independent coupling $\gamma = \mu_0 \otimes \mu_1$. However, for an arbitrary $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$, not all geodesics can be extended as geodesic rays.

In the setting of Brenier's theorem, we show that geodesics are rays if and only if the Monge map between μ_0 and μ_1 is the gradient of a 1-convex Brenier potential function u, that is $x \mapsto u(x) - \frac{\|x\|_2^2}{2}$ is convex.

Proposition 1. *Let $\mu_0 \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$, $\mu_1 \in \mathcal{P}_2(\mathbb{R}^d)$, and T the Monge map between μ_0 and μ_1. The curve $t \mapsto \mu_t = \big((1-t)\mathrm{Id} + t\mathrm{T}\big)_\#\mu_0$ is a geodesic ray if and only if T is the gradient of a 1-convex function u.*

This result is strongly related to (Gallouët et al., 2024, Section 4) in which it is stated that a geodesic can be extended on a segment $[0,\alpha]$ for $\alpha \geq 1$ if and only if $x \mapsto \alpha u(x) - (\alpha-1)\frac{\|x\|_2^2}{2}$ is convex (or equivalently, $x \mapsto u(x) - (1-\frac{1}{\alpha})\frac{\|x\|_2^2}{2}$ is convex). Taking the limit $\alpha \to +\infty$, we recover the result of Proposition 1.

Note that $\mu_0 \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ in Proposition 1 allows leveraging Brenier's theorem and guarantees that there exists an OT map. In the one dimensional case, we can further characterize geodesic rays starting from any $\mu_0 \in \mathcal{P}_2(\mathbb{R})$ with quantile functions. Indeed, denoting F_0^{-1} and F_1^{-1} the quantile functions of $\mu_0, \mu_1 \in \mathcal{P}_2(\mathbb{R})$, the geodesic between μ_0 and μ_1 at time $t \in [0,1]$ is defined by $\mu_t = \big((1-t)\pi^1 + t\pi^2\big)_\#\gamma^*$ with $\gamma^* = (F_0^{-1}, F_1^{-1})_\#\mathrm{Unif}([0,1])$ the optimal coupling between μ_0 and μ_1. Then, for F_t^{-1} the quantile of the geodesic at time $t \in [0,1]$, it is well known (see *e.g.* (Ambrosio et al., 2008, Equation 7.2.8)) that

$$\forall t \in [0,1],\ F_t^{-1} = (1-t)F_0^{-1} + tF_1^{-1}. \quad (9)$$

As observed by Kloeckner (2010), non-decreasing left-continuous functions are the inverse cumulative distribution function of a probability distribution. We can thus extend the geodesic as long as F_t^{-1} is non-decreasing, which gives a condition on $F_1^{-1} - F_0^{-1}$.

Proposition 2. *Let μ_0, $\mu_1 \in \mathcal{P}_2(\mathbb{R})$ and F_0^{-1}, F_1^{-1} their quantile functions. The geodesic between μ_0 and μ_1 is a ray if and only if $F_1^{-1} - F_0^{-1}$ is non-decreasing.*

As an application of Proposition 2, we get the following results on discrete 1D distributions with the same number of samples, and on 1D Gaussian distributions.

Corollary 1. *Let $x_1 < \cdots < x_n \in \mathbb{R}$, $y_1 < \cdots < y_n \in \mathbb{R}$, $\mu_0 = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ and $\mu_1 = \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$. Then, the geodesic between μ_0 and μ_1 is a ray if and only if for all $j > i$, $y_i - x_i \leq y_j - x_j$.*

Note that in the setting of Corollary 1, the geodesic is given by $\mu_t = \frac{1}{n}\sum_{i=1}^n \delta_{(1-t)x_i + ty_i}$ as points are sorted. Hence, the condition $y_i - x_i \leq y_j - x_j$ for all $j > i$ ensures that there are no crossings between particles for all time $t \geq 0$ since $x_i + t(y_i - x_i) < x_j + t(y_j - x_j)$. The optimal assignment between μ_t and μ_s therefore remains the same (the identity) for any $s,t \geq 0$.

Corollary 2. *Let $\mu_0 = \mathcal{N}(m_0, \sigma_0^2)$, $\mu_1 = \mathcal{N}(m_1, \sigma_1^2)$ with $m_0, m_1 \in \mathbb{R}$, $\sigma_0, \sigma_1 \in \mathbb{R}_+$. Then, the geodesic between μ_0 and μ_1 is a ray if and only if $\sigma_1 \geq \sigma_0$.*

Note that if $\sigma_0^2 = \sigma_1^2$ in Corollary 2, *i.e.* if the measures are translated, we recover that the geodesic is indeed a line, as it can be extended to infinity in both directions. In the case of arbitrary 1D Gaussian distributions, we can actually obtain the largest interval over which the geodesic can be extended, see Appendix C.

For Gaussians of any dimension, we have that the Bures-Wasserstein space $\mathrm{BW}(\mathbb{R}^d)$ is geodesically convex, *i.e.* the geodesic between two Gaussian distributions stays in $\mathrm{BW}(\mathbb{R}^d)$ at each time $t \in [0,1]$. In particular, the geodesic between $\mu = \mathcal{N}(m_\mu, \Sigma_\mu)$ and $\nu = \mathcal{N}(m_\nu, \Sigma_\nu)$ is given, for all $t \in [0,1]$, by $\mu_t = \mathcal{N}(m_t, \Sigma_t)$ (Altschuler et al., 2021) with,

$$\begin{cases} m_t = (1-t)m_\mu + t m_\nu \\ \Sigma_t = \big((1-t)I_d + tA_\mu^\nu\big)\Sigma_\mu\big((1-t)I_d + tA_\mu^\nu\big). \end{cases} \tag{10}$$

As an application of Proposition 1, we can extend Corollary 2 to Gaussian distributions for $d \geq 1$. In particular, the condition $\sigma_0 \leq \sigma_1$ is extended to Σ_0 and Σ_1 through the partial (Loewner) ordering $\preceq$ on $S_d^+(\mathbb{R})$, the space of positive semi-definite matrices, defined by $A \succeq B$ if and only if $A - B \in S_d^+(\mathbb{R})$.

Corollary 3. *Let $\mu_0 = \mathcal{N}(m_0, \Sigma_0)$, $\mu_1 = \mathcal{N}(m_1, \Sigma_1)$ with $m_0, m_1 \in \mathbb{R}^d$ and $\Sigma_0, \Sigma_1 \in S_d^{++}(\mathbb{R})$. Then, the geodesic between μ_0 and μ_1 is a geodesic ray if and only if $(\Sigma_0^{\frac{1}{2}}\Sigma_1\Sigma_0^{\frac{1}{2}})^{\frac{1}{2}} \succeq \Sigma_0$.*

Note that this condition is implied by $\Sigma_1^{\frac{1}{2}} \succeq \Sigma_0^{\frac{1}{2}}$ in general by Furata's inequality (Fujii, 2010, Theorem 1.3), and equivalent whenever Σ_0 and Σ_1 commute.

The conditions on μ_0 and μ_1 to obtain geodesic rays can be seen as guaranteeing a certain regularity along the geodesic. For instance, when $\mu_0 \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$, choosing $\mu_1 = \nabla u_{\#}\mu_0$ for a 1-convex function u (see Proposition 1) guarantees that the geodesic always stays in $\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$. While it does not hold for an arbitrary $\mu_1 \in \mathcal{P}_2(\mathbb{R}^d)$, it would be possible to find the closest geodesic ray *e.g.* by minimizing $\mathrm{W}_2(\nabla u_{\#}\mu_0, \mu_1)$ with u 1-convex (Paty et al., 2020). In the discrete case, these conditions can be interpreted in terms of particle dynamics: failure to satisfy the conditions corresponds to particle crossings.

An alternative extrapolation of geodesic curves beyond $t = 1$ was recently proposed by Gallouët et al. (2025) by solving a suitable variational problem. Their extrapolation coincides with the constant-speed geodesic as long as the curve remains geodesic, and therefore coincides with a true geodesic for all $t \geq 0$ when considering rays. Nonetheless, it is not the case for geodesics that are not rays. In particular for μ_1 a Dirac, their solution merges particles after crossing at the Dirac. In this work, we only consider true geodesic rays.

We also note that the conditions to have geodesic rays are very similar to those where μ_0 is smaller than μ_1 in the convex order when their first moments coincide, *i.e.* satisfying for all $f : \mathbb{R}^d \to \mathbb{R}^d$ convex, $\int f \mathrm{d}\mu_0 \leq \int f \mathrm{d}\mu_1$, see *e.g.* (Müller, 2001, Theorem 4 and 6) for the Gaussian case, and (Shu, 2020) for the 1D case. These conditions are also connected to projections in the convex order, see *e.g.* (Alfonsi and Jourdain, 2025, Proposition 3.3). We defer a more comprehensive study of these relations to future work.

3 BUSEMANN FUNCTION

In this section, we first introduce the Busemann function in geodesic metric spaces. We then discuss how to compute it in the Wasserstein space in general settings, and in specific cases where a closed-form exists.

3.1 Background on the Busemann Function

In any geodesic metric space (X, d) that admits geodesic rays, the Busemann function B^γ associated to a geodesic ray γ can be defined for any $x \in \mathrm{X}$, as in (Bridson and Haefliger, 2013, II.8.17),

$$B^\gamma(x) = \lim_{t\to\infty} \; d\big(\gamma(t), x\big) - t \cdot d\big(\gamma(0), \gamma(1)\big). \tag{11}$$

This function has attracted particular interest in geometric Machine Learning as it provides a natural generalization of hyperplanes on metric spaces. Indeed, in the particular case of Euclidean spaces, geodesic rays are of the form $\gamma(t) = t\theta$ for $\theta \in S^{d-1} = \{\theta \in \mathbb{R}^d, \; \|\theta\|_2^2 = 1\}$, $t \in \mathbb{R}$, and the Busemann function is given, for any $x \in \mathbb{R}^d$, by $B^\gamma(x) = -\langle x, \theta\rangle$. Therefore, its level sets are (affine) hyperplanes. Moreover, the Busemann function provides a principled way to project a point $x \in \mathrm{X}$ on the geodesic ray γ. In fact, noticing that for any $s \in \mathbb{R}_+$, $B^\gamma\big(\gamma(s)\big) = -s$, the projection of $x \in \mathrm{X}$ on γ is given by $P^\gamma(x) = \gamma\big(-B^\gamma(x)\big)$. In particular, all points on a level set of B^γ are projected on the same point. Note however that when $-B^\gamma(x) < 0$, there is, in general, no guarantee that $\gamma\big(-B^\gamma(x)\big)$ belongs to the geodesic, but $B^\gamma(x) \in \mathbb{R}$ is always well defined, and provides a projection on $\mathbb{R}$. In a Hilbertian space (*i.e.* of null curvature), the Busemann projection is actually equivalent to the coordinate of the metric projection, *i.e.* $-B^\gamma(x) = \mathrm{argmin}_t \; d\big(x, \gamma(t)\big)$.

3.2 Busemann on the Wasserstein Space

As the Wasserstein space is not geodesically complete, not all geodesic can be extended as a ray, and thus the Busemann function is not defined along every geodesic. Fortunately, any $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$ is the starting point of at least one geodesic ray (Zhu et al., 2021, Theorem 1.1), and in some particular cases, we can characterize them as described in the previous section. Let $(\mu_t)_{t\geq 0}$ be a geodesic ray and $\kappa_\mu = \mathrm{W}_2(\mu_0, \mu_1)$ its speed. Let us define the Busemann function B^μ associated to $(\mu_t)_{t\geq 0}$ by, for any $\nu \in \mathcal{P}_2(\mathbb{R}^d)$,

$$B^\mu(\nu) = \lim_{t\to\infty} \; \mathrm{W}_2(\mu_t, \nu) - \kappa_\mu t. \tag{12}$$

Thanks to the Riemannian structure of $(\mathcal{P}_2(\mathbb{R}^d), \mathrm{W}_2)$, we can always assume that geodesics have unit speed, *i.e.* $\kappa_\mu = 1$, see Appendix A.1. In the following formulas of B^μ, this translates as a renormalization by κ_μ.

First, we show that B^μ admits a more convenient form as an infimum over a suitable set of couplings $\Gamma(\mu_0, \mu_1, \nu) = \{\tilde{\gamma} \in \Pi(\mu_0, \mu_1, \nu),\ \pi^{1,2}_{\#}\tilde{\gamma} \in \Pi_o(\mu_0, \mu_1)\}$ between μ_0, μ_1 and ν, and such that the coupling between the two first marginals is optimal for (1).

Proposition 3. *Let $(\mu_t)_{t\geq 0}$ be a geodesic ray on $\mathcal{P}_2(\mathbb{R}^d)$. Let $\nu \in \mathcal{P}_2(\mathbb{R}^d)$, then*

$$B^\mu(\nu) = \inf_{\tilde{\gamma}\in\Gamma(\mu_0,\mu_1,\nu)} -\kappa_\mu^{-1} \int \langle x_1 - x_0, y - x_0\rangle \, \mathrm{d}\tilde{\gamma}(x_0, x_1, y). \tag{13}$$

We can refine the result in the case where the geodesic ray is given by an OT map, *i.e.*, when the OT map is the gradient of a 1-convex function by Proposition 1.

Corollary 4. *Let $\mu_0 \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$, $\mu_1 \in \mathcal{P}_2(\mathbb{R}^d)$, and assume the OT map T between μ_0 and μ_1 is the gradient of a 1-convex function. Let $\nu \in \mathcal{P}_2(\mathbb{R}^d)$, then,*

$$B^\mu(\nu) = \inf_{\gamma\in\Pi(\mu_0,\nu)} -\kappa_\mu^{-1} \int \langle \mathrm{T}(x_0) - x_0, y - x_0\rangle \, \mathrm{d}\gamma(x_0, y). \tag{14}$$

This problem is equivalent to the OT problem

$$\inf_{\gamma\in\Pi(\mu_0,\nu)} \int \|\mathrm{T}(x_0) - x_0 - y\|_2^2 \, \mathrm{d}\gamma(x_0, y), \tag{15}$$

and can thus be solved using classical OT solvers.

In the specific case where μ_0 is a Dirac, we can also leverage that any geodesic is a ray, and that the optimal coupling is of the form $\gamma = \mu_0 \otimes \mu_1$.

Corollary 5. *Let $\mu_0 = \delta_{x_0}$ where $x_0 \in \mathbb{R}^d$, and $\mu_1 \in \mathcal{P}_2(\mathbb{R}^d)$. Let $\nu \in \mathcal{P}_2(\mathbb{R}^d)$, then*

$$B^\mu(\nu) = \inf_{\gamma\in\Pi(\mu_1,\nu)} -\kappa_\mu^{-1} \int \langle x_1 - x_0, y - x_0\rangle \, \mathrm{d}\gamma(x_1, y). \tag{16}$$

Note that (16) is equivalent to the OT problem (1) between μ_1 and ν. Letting $\mu_1 = \delta_{x_1}$ with $\theta := x_1 - x_0 \in S^{d-1}$ and $\gamma(t) = x_0 + t\theta$, (16) is equal to $\int B^\gamma(y)\mathrm{d}\nu(y)$ and thus the Busemann function lifts from $\mathbb{R}^d$ to $\mathcal{P}_2(\mathbb{R}^d)$.

3.3 Closed-forms of the Busemann Function

In the cases mentioned above, we don't have a closed-form for the OT problem, and must therefore solve an optimization problem to compute the corresponding Busemann functions. Nonetheless, we can compute it in closed-form whenever closed-forms for the Wasserstein distance and the geodesics are available. First, we consider 1D distributions leveraging (2) and (9).

Proposition 4. *Let $(\mu_t)_{t\geq 0}$ be a unit-speed geodesic ray in $\mathcal{P}_2(\mathbb{R})$ (*i.e. *$\kappa_\mu = 1$), then for any $\nu \in \mathcal{P}_2(\mathbb{R})$,*

$$B^\mu(\nu) = -\langle F_1^{-1} - F_0^{-1}, F_\nu^{-1} - F_0^{-1}\rangle_{L^2([0,1])}. \tag{17}$$

We observe that, up to a sign, (17) corresponds to the inner product in $L^2([0,1])$ between $F_1^{-1} - F_0^{-1}$ and $F_\nu^{-1} - F_0^{-1}$, which are the quantiles centered around F_0^{-1}, and is directly obtained from the Hilbert structure of the one dimensional Wasserstein space. Consequently, the Busemann function between 1D Gaussians is only an inner product on the product space $\mathbb{R} \times \mathbb{R}_+^*$ of the (centered) means and standard deviations.

Corollary 6. *Let $\mu_0 = \mathcal{N}(m_0, \sigma_0^2)$, $\mu_1 = \mathcal{N}(m_1, \sigma_1^2)$, $\nu = \mathcal{N}(m, \sigma^2)$ with $m_0, m_1, m \in \mathbb{R}$, $\sigma_0, \sigma_1, \sigma \in \mathbb{R}_+^*$ such that $\sigma_1 \geq \sigma_0$ and $\mathrm{W}_2^2(\mu_0, \mu_1) = 1$. Then,*

$$B^\mu(\nu) = -(m_1 - m_0)(m - m_0) - (\sigma_1 - \sigma_0)(\sigma - \sigma_0). \tag{18}$$

More generally, on $\mathrm{BW}(\mathbb{R}^d)$, we leverage the closed-forms of the Wasserstein distance (3) and geodesics (10), which remain Gaussian at all time.

Proposition 5. *Let $(\mu_t)_{t\geq 0}$ be a geodesic ray characterized by $\mu_0 = \mathcal{N}(m_0, \Sigma_0)$ and $\mu_1 = \mathcal{N}(m_1, \Sigma_1)$, and such that $\kappa_\mu = 1$. Then, for any $\nu = \mathcal{N}(m, \Sigma)$,*

$$B^\mu(\nu) = -\langle m_1 - m_0, m - m_0\rangle + \mathrm{Tr}\big(\Sigma_0(A^{\mu_1}_{\mu_0} - I_d)\big) - \mathrm{Tr}\big((\Sigma^{\frac{1}{2}}(\Sigma_0 - \Sigma_0 A^{\mu_1}_{\mu_0} - A^{\mu_1}_{\mu_0}\Sigma_0 + \Sigma_1)\Sigma^{\frac{1}{2}})^{\frac{1}{2}}\big). \tag{19}$$

When all covariance matrices commute, *e.g.* if they are diagonal matrices, (19) simplifies as

$$B^\mu(\nu) = -\langle m_1 - m_0, m - m_0\rangle - \langle \Sigma_1^{\frac{1}{2}} - \Sigma_0^{\frac{1}{2}}, \Sigma^{\frac{1}{2}} - \Sigma_0^{\frac{1}{2}}\rangle_F. \tag{20}$$

This corresponds to the inner product in the space $\mathbb{R}^d \times S_d(\mathbb{R})$. Moreover, we recover (18) in one dimension.

4 SLICING DATASETS

Building on the Sliced-Wasserstein distance, a computationally efficient alternative to the Wasserstein distance, we use the Busemann function to construct new sliced distances to compare labeled datasets.

4.1 Sliced-Wasserstein Distance

Given two discrete distributions $\mu^n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ and $\nu^n = \frac{1}{n}\sum_{j=1}^n \delta_{y_j} \in \mathcal{P}_2(\mathbb{R}^d)$, the Wasserstein distance

between μ^n and ν^n can be computed in $\mathcal{O}(n^3 \log n)$, which is too costly for many applications. Thus, several variants have been proposed, such as adding an entropic regularization and using Sinkhorn's algorithm (Cuturi, 2013), mini-batchs (Fatras et al., 2020, 2021) or low-rank solvers (Scetbon et al., 2021).

Another very popular alternative to the Wasserstein distance, which enjoys much better computational properties, is the Sliced-Wasserstein (SW) distance (Rabin et al., 2012; Bonneel et al., 2015). It is based on the attractive closed-form of the Wasserstein distance in 1D (2), which can be computed in practice between discrete distributions by sorting the samples and therefore has a complexity of $\mathcal{O}(n \log n)$. Given a parametric one dimensional projection $P^\theta : \mathbb{R}^d \to \mathbb{R}$ with $\theta \in \Theta$, SW is then defined between $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ as the average of the 1D Wasserstein distances between the projected distributions, *i.e.*,

$$\mathrm{SW}_2^2(\mu, \nu) = \int \mathrm{W}_2^2(P^\theta_\# \mu, P^\theta_\# \nu) \, \mathrm{d}\lambda(\theta), \tag{21}$$

for $\lambda \in \mathcal{P}(\Theta)$. In its original formulation, SW is set for $\Theta = S^{d-1}$, *i.e.* the unit hypersphere in dimension d, $\lambda = \mathcal{U}(S^{d-1})$, the uniform measure on S^{d-1} and for any $\theta \in S^{d-1}$, $P^\theta(x) = \langle x, \theta \rangle$. However, there exist other variants with different projection schemes, *e.g.* convolutions for image data (Nguyen and Ho, 2022), manifold-aware projections (Bonet et al., 2023a,b, 2025b), or more general non linear projections (Kolouri et al., 2019; Chen et al., 2022).

4.2 Comparing Labeled Datasets

We consider a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ of pairs of samples $x_i \in \mathbb{R}^d$ associated to a label y_i from a set of C classes $\mathcal{Y} = \{1, \dots, C\}$. Class conditional distributions are given for a specific class $y \in \mathcal{Y}$ by $\varphi(y) = \frac{1}{n_y} \sum_{i=1}^n \delta_{x_i} \mathbb{1}_{\{y_i = y\}}$, with $n_y = \sum_{i=1}^n \mathbb{1}_{\{y_i = y\}}$ the number of samples in the class y. A dataset can then be represented by a probability distribution over the product space $\mathbb{R}^d \times \mathcal{P}_2(\mathbb{R}^d)$, *i.e.* as $\frac{1}{n} \sum_{i=1}^n \delta_{(x_i, \varphi(y_i))} \in \mathcal{P}_2\big(\mathbb{R}^d \times \mathcal{P}_2(\mathbb{R}^d)\big)$.

Alvarez-Melis and Fusi (2020) proposed to use

$$d_{\mathcal{D}}\big((x, y), (x', y')\big)^2 = \|x - x'\|_2^2 + \mathrm{W}_2^2\big(\varphi(y), \varphi(y')\big). \tag{22}$$

as groundcost of an optimal transport problem on $\mathcal{P}_2\big(\mathbb{R}^d \times \mathcal{P}_2(\mathbb{R}^d)\big)$, defining the Optimal Transport Dataset Distance (OTDD), see Appendix A.4. However, OTDD is costly to compute as it requires to solve $\mathcal{O}(C^2)$ OT problems with n_y samples to compute the groundcost (22), and a global OT problem with $n = \sum_{y=1}^C n_y$ samples. This has led to several approximations and variants aimed at reducing the computational cost, see *e.g.* (Hua et al., 2023; Liu et al., 2025; Nguyen et al., 2025a; Bonet et al., 2025a).

Nguyen et al. (2025a) recently proposed a sliced distance for labeled datasets on $\mathcal{P}_2\big(\mathbb{R}^d \times \mathcal{P}_2(\mathbb{R}^d)\big)$. This requires building a projection from $\mathbb{R}^d \times \mathcal{P}_2(\mathbb{R}^d)$ to $\mathbb{R}$ to project the distribution of pairs $(x_i, \varphi(y_i))$ onto a distribution in $\mathcal{P}_2(\mathbb{R})$. Their construction can be broken down into combining two projections from $\mathbb{R}^d \to \mathbb{R}$ and from $\mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ using the Hierarchical Hybrid projection (Nguyen and Ho, 2024), which consists of a random linear combination with weights in the sphere. Especially, for any labeled sample $(x, y) \in \mathbb{R}^d \times \mathcal{Y}$, their projection is of the form

$$P^{\alpha,\theta,\lambda}(x, y) = \alpha_1 P^\theta(x) + \sum_{i=1}^k \alpha_{i+1} \mathcal{M}^{\lambda_i}\big(P^\theta_\# \varphi(y)\big), \tag{23}$$

with $P^\theta : \mathbb{R}^d \to \mathbb{R}$ any projection on a line, $\mathcal{M}^\lambda : \mathcal{P}_2(\mathbb{R}) \to \mathbb{R}$ the moment transform projection and $\alpha \in S^k$. Plugging (23) into the Wasserstein distance term on $\mathcal{P}_2(\mathbb{R})$ in (21) defines the sliced OTDD distance (SOTDD) on $\mathcal{P}_2\big(\mathbb{R}^d \times \mathcal{P}_2(\mathbb{R}^d)\big)$.

4.3 Slicing Datasets with Busemann

As the Busemann function allows to project any probability distribution onto $\mathbb{R}$, it is natural to use it as a projection to define Sliced-Wasserstein distances for the purpose of comparing labeled datasets.

From a computational perspective, we want to avoid solving additional OT problems to compute the Busemann functions. Therefore, we propose two new discrepancies based on the closed-forms of the Busemann function for 1D probability distributions (17) and for Gaussians (19). In both cases, as in Nguyen et al. (2025a), we use the Hierarchical Hybrid projection from (Nguyen and Ho, 2024) to combine the projection $P^\theta : \mathbb{R}^d \to \mathbb{R}$ of the features and $Q^\eta : \mathcal{Y} \to \mathbb{R}$ of the labels, *i.e.* for $\alpha \in S^1$, we define for $(x, y) \in \mathbb{R}^d \times \mathcal{Y}$,

$$P^{\alpha,\theta,\eta}\big((x, y)\big) = \alpha_1 P^\theta(x) + \alpha_2 Q^\eta(y). \tag{24}$$

Gaussian Approximation. To leverage the closed-form between Gaussian (19), we use a Gaussian approximation of the classes, with a possible dimension reduction of the features beforehand as in (Hua et al., 2023). Let us denote $\psi : \mathbb{R}^d \to \mathbb{R}^{d'}$ a dimension reduction operator with $d' \ll d$, and, for $\mu \in \mathcal{P}_2(\mathbb{R}^{d'})$, the Gaussian approximation $\Xi(\mu) = \mathcal{N}\big(m(\mu), \Sigma(\mu)\big)$ with $m(\mu) = \int x \mathrm{d}\mu(x)$ and $\Sigma(\mu) = \int (x - m(\mu)) \otimes (x - m(\mu)) \, \mathrm{d}\mu(x)$ the mean and covariance operators. The label projections are then given by $Q^\eta(y) = B^\eta\big(\Xi\big(\psi_\# \varphi(y)\big)\big)$ with η a geodesic ray on $\mathrm{BW}(\mathbb{R}^{d'})$.

To define a sliced distance, we also need to sample a valid ray η so that the Busemann function is well

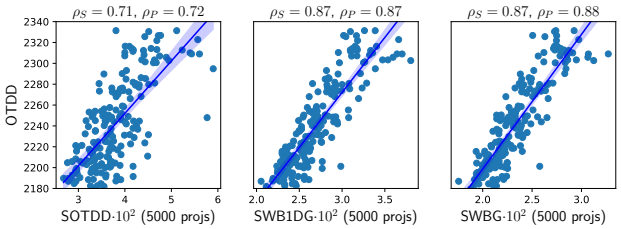


Figure 1: Spearman (ρ_S) and Pearson (ρ_P) correlation between SOTDD, SWB1DG, SWBG and OTDD between subdatasets of CIFAR10.

defined. To do so, we choose to fix $\eta_0 = \mathcal{N}(0, I_d)$, and sample $\eta_1 = \mathcal{N}(m_1, \Sigma_1)$ such that $m_1 \in S^{d-1}$, $\Sigma_1 \in S_d^{++}(\mathbb{R})$ with $\Sigma_1^{\frac{1}{2}} \succeq I_d$ and $\mathrm{W}_2^2(\eta_0, \eta_1) = 1$. To enforce $\Sigma_1^{\frac{1}{2}} \succeq I_d$, we remark that it is equivalent to consider $S = \log_{I_d}(\Sigma_1) = \Sigma_1^{\frac{1}{2}} - I_d \succeq 0$, where $\log_{I_d}$ is the logarithm map in $S_d^{++}(\mathbb{R})$. Thus, we sample uniformly $\Delta \in O_d(\mathbb{R})$ an orthogonal matrix and $\tilde{\theta} \in S^{d-1}$, and define $S := \Delta \mathrm{diag}(|\tilde{\theta}|)\Delta^T$ and $\Sigma_1 := \exp_{I_d}(S) = (I_d + S)^2$. To enforce $\mathrm{W}_2^2(\eta_0, \eta_1) = 1$, we normalize $(m_1 - m_0, S)$ in the tangent space to have a unit-speed geodesic ray, see Appendix B. Given two datasets $\mathbb{P}, \mathbb{Q} \in \mathcal{P}_2\big(\mathbb{R}^d \times \mathcal{P}_2(\mathbb{R}^d)\big)$, we define the Sliced-Wasserstein Busemann Gaussian distance (SWBG) as

$$\mathrm{SWBG}^2(\mathbb{P}, \mathbb{Q}) = \int \mathrm{W}_2^2(P^\vartheta_\# \mathbb{P}, P^\vartheta_\# \mathbb{Q}) \, \mathrm{d}\lambda(\vartheta), \quad (25)$$

with $\vartheta = (\alpha, \theta, \tilde{\theta}, \Delta, m_1)$ and λ the uniform measure on the resulting product space.

1D Projections. To leverage the 1D closed-form of the Busemann function (17), we can first project the class conditional distributions in 1D, and define for $y \in \mathcal{Y}$, $Q^{\eta,\theta}(y) = B^\eta\big(P^\theta_\# \varphi(y)\big)$. Regarding η, setting $\eta_0 = \delta_0$, we get geodesic rays for any η_1, and thus set $\eta_1 = \mathcal{N}(m_1, \sigma_1^2)$ such that the speed $\kappa_\eta = \mathrm{W}_2^2(\eta_0, \eta_1) = m_1^2 + \sigma_1^2 = 1$. Then, for $\mu \in \mathcal{P}_2(\mathbb{R})$, (17) writes

$$B^\eta(\mu) = -m_1 m(\mu) - \sigma_1 \int_0^1 \phi^{-1}(u) F_\mu^{-1}(u) \, \mathrm{d}u, \quad (26)$$

with ϕ the cumulative distribution function of $\mathcal{N}(0, 1)$. In practice, the geodesic ray can be sampled using $m_1 \sim \mathcal{U}([-1, 1])$ and setting $\sigma_1 = \sqrt{1 - m_1^2}$. Given $\mathbb{P}, \mathbb{Q} \in \mathcal{P}_2\big(\mathbb{R}^d \times \mathcal{P}_2(\mathbb{R}^d)\big)$, we define the Sliced-Wasserstein Busemann 1D Gaussian distance (SWB1DG) as

$$\mathrm{SWB1DG}^2(\mathbb{P}, \mathbb{Q}) = \int \mathrm{W}_2^2(P^\vartheta_\# \mathbb{P}, P^\vartheta_\# \mathbb{Q}) \, \mathrm{d}\lambda(\vartheta), \quad (27)$$

with $\vartheta = (\alpha, \theta, m_1)$ and $\lambda = \mathcal{U}(S^1 \times S^{d-1} \times [-1, 1])$.

Table 1: Correlation averaged over 10 sets of 50 bootstrapped pairs of subdatasets of CIFAR10, for different number of projections L.

L	Spearman correlation (ρ_S)			Pearson correlation (ρ_P)		
	SOTDD	SWB1DG	SWBG	SOTDD	SWB1DG	SWBG
10	$14.0_{\pm 11.3}$	$\mathbf{44.3}_{\pm 10.8}$	$40.2_{\pm 12.2}$	$16.0_{\pm 12.9}$	$38.6_{\pm 14.6}$	$\mathbf{42.7}_{\pm 9.5}$
50	$30.5_{\pm 12.9}$	$\mathbf{62.6}_{\pm 6.4}$	$40.4_{\pm 9.8}$	$25.2_{\pm 11.4}$	$\mathbf{63.6}_{\pm 6.3}$	$42.8_{\pm 8.5}$
100	$15.5_{\pm 11.8}$	$\mathbf{71.9}_{\pm 6.4}$	$68.1_{\pm 7.2}$	$21.0_{\pm 11.4}$	$\mathbf{73.9}_{\pm 5.5}$	$72.8_{\pm 5.4}$
500	$52.1_{\pm 8.1}$	$\mathbf{82.3}_{\pm 2.2}$	$78.4_{\pm 6.0}$	$54.6_{\pm 8.8}$	$\mathbf{83.5}_{\pm 2.1}$	$79.4_{\pm 7.7}$
1000	$52.0_{\pm 10.9}$	$83.6_{\pm 4.8}$	$\mathbf{83.7}_{\pm 5.0}$	$53.1_{\pm 11.3}$	$\mathbf{85.6}_{\pm 3.5}$	$84.9_{\pm 4.8}$
5000	$72.2_{\pm 7.5}$	$88.5_{\pm 4.8}$	$\mathbf{89.3}_{\pm 3.8}$	$75.4_{\pm 5.5}$	$87.8_{\pm 2.8}$	$\mathbf{89.0}_{\pm 2.4}$
10000	$72.6_{\pm 6.1}$	$82.7_{\pm 4.8}$	$\mathbf{86.7}_{\pm 3.0}$	$77.1_{\pm 4.3}$	$87.3_{\pm 2.8}$	$\mathbf{90.2}_{\pm 2.3}$

At the time of submission, we noticed the concurrent work (Piening and Beinert, 2026) whose sliced construction on $\mathcal{P}_2\big(\mathcal{P}_2(\mathbb{R}^d)\big)$ is close to SWB1DG, even though it is obtained from a different viewpoint. We detail in Appendix A.3 the relation between the two sliced distances.

Computational Properties. The sliced distances can be approximated using Monte-Carlo projections. Given L projections, the complexity of SWB1DG is $\mathcal{O}\big(Ln(\log n + d)\big)$, similarly to SOTDD. SWBG is more costly as it requires to compute square roots of matrices, which gives a complexity of $\mathcal{O}\big(LCd'^3 + Ln(\log n + d') + Cd'^2 N_C\big)$ with $N_C = \max_y n_y$. We refer to Appendix E for numerical comparisons.

Slicing Mixtures. The constructions can also be used to compare distributions on $\mathcal{P}_2\big(\mathcal{P}_2(\mathbb{R}^d)\big)$ by setting $\alpha_1 = 0$. In Appendix B, we investigate such constructions to compare Gaussian mixtures.

5 EXPERIMENTS

In this Section, we compare the sliced-based distances on labeled datasets. We first show that SWB1DG and SWBG are better replacements for OTDD than SOTDD, as they are more correlated with it. Then, we show that these distances can be used to flow datasets (Alvarez-Melis and Fusi, 2021), for instance to perform transfer learning. We refer to Appendix E for details[1].

5.1 Correlation with OTDD

To show that the sliced distances are suitable proxies to the costly OTDD distance (Alvarez-Melis and Fusi, 2020), we measure the correlation between OTDD and the sliced distances on image datasets.

Following (Nguyen et al., 2025a), we randomly split the CIFAR10 dataset (Krizhevsky et al., 2009) to get subdataset pairs, with sizes ranging from 5000 to 10000 samples, obtaining 200 pairs. Between each

[1]Code available at `https://github.com/clbonet/Busemann_Functions_in_the_Wasserstein_Space`

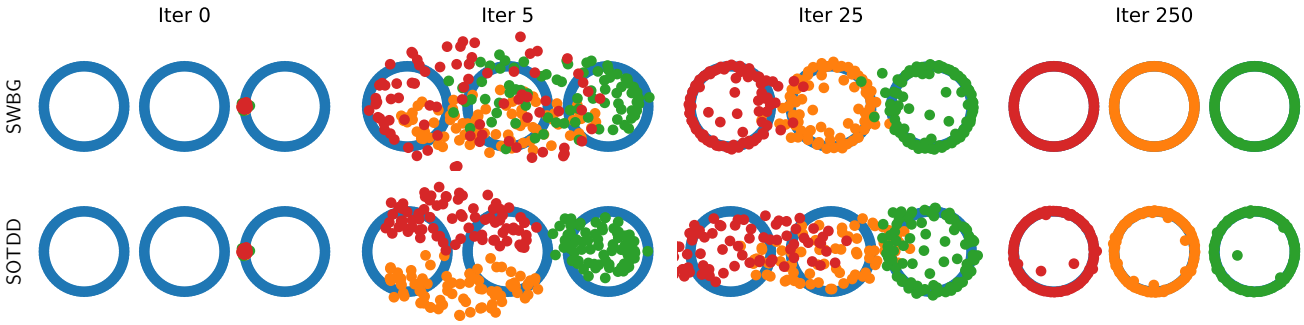


Figure 2: Evolution of the WoW gradient flow of SWBG and SOTDD with the 3-rings dataset as target.

pair, we compute OTDD, SOTDD, SWB1DG and SWBG. OTDD and SOTDD are computed using the code shared by Nguyen et al. (2025a)[2]. For SWBG, we used a TSNE in dimension 10 for ψ using the `TorchDR` library (Van Assel et al., 2024). The projections are done using convolution projections, as they better capture the spatial structure and are more suitable to compare distributions of images (Nguyen and Ho, 2022).

We report the results in Figure 1, where we first scatter the values obtained in ordinate for OTDD, and in abscissa for the sliced distances with 5000 projections. Then, we report the values of the Spearman and Pearson correlations.The Pearson correlation is equal to ± 1 when both quantities are linearly correlated while Spearman correlation is equal to ± 1 if the quantities are monotonically related. We observe that both SWB1DG and SWBG are better correlated to OTDD than SOTDD. We note that the results do not match those in (Nguyen et al., 2025a), where they used only 10 pairs. We hypothesize that for such small number of samples, the Spearman and Pearson correlation are very sensitive to randomness. To verify the robustness of our results, we report in Table 1 the results obtained by bootstrapping 50 pairs of sub-datasets, and averaging over 10 experiments. Using this process for different numbers of projections, we see that the Busemann based sliced distances always outperform SOTDD, and often under a much smaller number of projections.

5.2 Flowing Labeled Datasets

Flowing one dataset onto another is useful to solve tasks ranging from domain adaptation to transfer learning (Alvarez-Melis and Fusi, 2021; Hua et al., 2023) or dataset distillation (Bonet et al., 2025a). This is achieved by minimizing a discrepancy on the space of datasets with respect to a target dataset, and starting from the source dataset. Alvarez-Melis and Fusi (2021) proposed to minimize OTDD while Hua et al. (2023) minimized a Maximum Mean Discrepancy.

[2]at `https://github.com/hainn2803/s-OTDD`

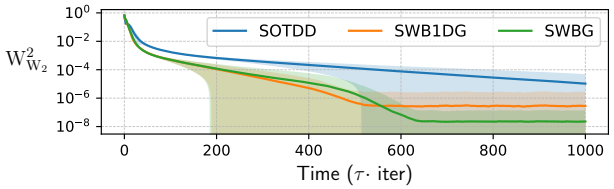


Figure 3: Convergence of the flow towards the 3-rings, averaged over 100 random batches of the target.

Bonet et al. (2025a) modeled this task as a minimization problem over the space $\mathcal{P}_2(\mathcal{P}_2(\mathbb{R}^d))$, representing datasets with n samples by class as $\mathbb{P} = \frac{1}{C}\sum_{c=1}^{C}\delta_{\mu_c} \in \mathcal{P}_2(\mathcal{P}_2(\mathbb{R}^d))$ with $\mu_c = \frac{1}{n}\sum_{i=1}^{n}\delta_{x_{i,c}}\mathbb{1}_{\{y_i=c\}}$. A discrepancy $\mathbb{F}(\mathbb{P}) = D(\mathbb{P},\mathbb{Q})$ with $\mathbb{P},\mathbb{Q} \in \mathcal{P}_2(\mathcal{P}_2(\mathbb{R}^d))$ can be minimized by a Wasserstein over Wasserstein (WoW) gradient descent on this space. In particular, for $\mathbb{P} = \frac{1}{C}\sum_{c=1}^{C}\delta_{\mu_c}$, the WoW gradient $\nabla_{\mathrm{W}_{\mathrm{W}_2}}\mathbb{F}(\mathbb{P})$ can be recovered by rescaling the Euclidean gradient of $F(\mathbf{x}) = \mathbb{F}(\mathbb{P})$ for $\mathbf{x} := (x_{i,c})_{i,c}$ by nC, *i.e.* $\nabla_{\mathrm{W}_{\mathrm{W}_2}}\mathbb{F}(\mathbb{P})(\mu_c)(x_{i,c}) = nC\nabla F(\mathbf{x})_{i,c}$, see (Bonet et al., 2025a, Proposition B.7). Given $\mathbb{P}^k = \frac{1}{C}\sum_{c=1}^{C}\delta_{\mu_c^k}$, $\mu_c^k = \frac{1}{n}\sum_{i=1}^{n}\delta_{x_{i,c}}$, the WoW gradient descent corresponds to updating each particle $x_{i,c}^k$ as

$$\forall k \geq 0,\ x_{i,c}^{k+1} = x_{i,c}^k - \tau\nabla_{\mathrm{W}_{\mathrm{W}_2}}\mathbb{F}(\mathbb{P}^k)(\mu_c^k)(x_{i,c}^k). \quad (28)$$

Thus, we propose to minimize the sliced distances using the bijection from $\mathcal{P}_2(\mathcal{P}_2(\mathbb{R}^d)) \to \mathcal{P}_2(\mathbb{R}^d \times \mathcal{P}_2(\mathbb{R}^d))$, and performing gradient descent on $\mathcal{P}_2(\mathcal{P}_2(\mathbb{R}^d))$.

Rings Dataset. We first focus on a toy example, where the target dataset contains $C = 3$ classes, and each class forms a ring of $n = 80$ samples (Glaser et al., 2021; Bonet et al., 2025a). We learn a distribution $\mathbb{P}$ of the same form and with the same number of particles. In Figure 3, we show the convergence of the WoW gradient flows of SOTDD, SWB1DG and SWBG with the 3-rings dataset as target, averaged over 100 different random batches of the target, and measured with

Table 2: Accuracy of a classifier on augmented datasets for $k \in \{1, 5, 10, 100\}$. M refers to MNIST, F to Fashion MNIST and U to USPS.

Dataset	k	Trained on $\mathbb{Q}$	OTDD	SWB1DG	SOTDD
M to F	1	$26.0_{\pm 5.3}$	$30.5_{\pm 4.2}$	$41.3_{\pm 3.4}$	$\mathbf{43.4}_{\pm 2.6}$
	5	$38.5_{\pm 6.7}$	$59.7_{\pm 1.8}$	$\mathbf{65.5}_{\pm 1.6}$	$64.5_{\pm 1.2}$
	10	$53.9_{\pm 7.9}$	$64.0_{\pm 1.4}$	$66.0_{\pm 0.9}$	$\mathbf{67.7}_{\pm 0.6}$
	100	$71.1_{\pm 1.5}$	-	$\mathbf{74.1}_{\pm 0.6}$	$72.0_{\pm 1.9}$
M to U	1	$32.4_{\pm 7.9}$	$39.5_{\pm 7.9}$	$45.4_{\pm 4.3}$	$\mathbf{50.1}_{\pm 2.6}$
	5	$51.4_{\pm 9.8}$	$73.3_{\pm 1.4}$	$73.5_{\pm 1.4}$	$\mathbf{75.7}_{\pm 0.8}$
	10	$60.3_{\pm 10.1}$	$72.7_{\pm 2.7}$	$77.8_{\pm 1.4}$	$\mathbf{80.1}_{\pm 1.0}$
	100	$87.5_{\pm 0.7}$	-	$\mathbf{90.1}_{\pm 0.4}$	$89.6_{\pm 0.3}$

the WoW distance (*i.e.* the OT problem with W_2^2 as ground cost, which we denote $\mathrm{W}_{\mathrm{W}_2}$, see Appendix A.3). We use a step size of $\tau = 1$ and 1000 iterations. The best performing distance appears to be SWBG. We also report the particles along the flows of SWBG and SOTDD on Figure 2, clearly showing that the flow of SWBG converges faster than the one of SOTDD.

Transfer Learning. We now consider flowing image datasets to solve a k-shot transfer learning task (Alvarez-Melis and Fusi, 2021; Hua et al., 2023). In this experiment, we consider a target dataset $\mathbb{Q}$ with k samples by class, where k is typically small. To improve the classification results, we augment the dataset $\mathbb{Q}$ by concatenating it with samples flowed from another dataset $\mathbb{P}_0$, from which we have n samples by class. This is done by minimizing a distance on the space of datasets with respect to $\mathbb{Q}$, *i.e.* by minimizing $\mathbb{F}(\mathbb{P}) = \mathrm{D}(\mathbb{P}, \mathbb{Q})$ starting from $\mathbb{P}_0$. Any divergence on $\mathcal{P}_2\big(\mathbb{R}^d \times \mathcal{P}_2(\mathbb{R}^d)\big)$ may be chosen for D. We focus here on comparing OTDD, SOTDD and SWB1DG for D.

We take $\mathbb{P}_0$ as the MNIST dataset (LeCun and Cortes, 2010), and $\mathbb{Q}$ as Fashion MNIST (Xiao et al., 2017) or USPS (Hull, 1994), and set $k \in \{1, 5, 10, 100\}$ and $n = 200$. We minimize SWB1DG and SOTDD using the WoW gradient descent with step size $\tau = 1$ and momentum $m = 0.9$. For the choice of the number of projections for the Monte-Carlo approximation of the sliced distances, and of the number of iterations for the gradient descent, we perform a grid search. We refer to Appendix E.4 for more details about the experiment, and for the best parameters selected with the grid search.

Once the dataset $\mathbb{P}_0$ has been flowed towards $\mathbb{Q}$, and each class of the flowed dataset $\mathbb{P}_T$ has been matched to a class of the target dataset $\mathbb{Q}$ by solving the WoW OT problem, we train a LeNet5 on the augmented concatenated dataset. On Table 2, we report the accuracy of the classifier evaluated on a test set. We report the best results among the grid search for the classifier trained on $\mathbb{Q}$ augmented with data flowed by minimizing OTDD, SWB1DG and SOTDD. We also report a baseline where the network was only trained on $\mathbb{Q}$. The results are averaged over 5 training of the networks, and 3 outputs of the flows for SWB1DG and SOTDD, and are taken from (Bonet et al., 2025a) for OTDD and the baseline.

Table 3: Runtime in seconds for the transfer learning experiment from MNIST to Fashion MNIST.

Dataset	k-shot	OTDD	SWB1DG	SOTDD
M to F	1	294.53 ± 5.21	13.53 ± 0.42	14.07 ± 0.37
	5	1130.89 ± 108	13.84 ± 0.34	14.11 ± 0.21
	10	2294.13 ± 48	14.00 ± 0.37	14.15 ± 0.18
	100	-	15.19 ± 0.49	15.31 ± 0.49

The results are overall comparable between SWB1DG and SOTDD. Additional details, as well as examples of generated images, are provided in Appendix E.4. In particular, for a large number of iterations, both flows converge to good images. However, note that improved image quality does not necessarily translate into better performance on the transfer task.

Additionally, we report on Table 3 the runtimes for the transfer learning experiment, averaged over 3 outputs of the flows and trained for 5K epochs. We observe that SOTDD and SWB1DG almost have the same runtime, which is expected as they have the same computational complexity up to constants, and both are much more efficient than OTDD.

6 CONCLUSION

We studied in this work in which cases the Busemann function is well defined on the Wasserstein space, and how to compute it in practice. More precisely, we identified conditions to define geodesic rays on the Wasserstein space, showed that the Busemann function can be computed in general by solving an optimal transport problem, and derived closed-form formulas under which it can be computed efficiently. Then, we leveraged these closed-forms to define new efficient Sliced-Wasserstein distances on the space of datasets. Future works will include improving the scalability of SWBG, *e.g.* by using Gaussian approximations with low-rank covariances (Bouveyron and Corneli, 2026), investigating other applications for the Busemann function such as Principal Component Analysis on the Wasserstein space (Cazelles et al., 2018; Vesseron et al., 2026), or its computation on probabilities over manifolds.

Acknowledgements

We thank the anonymous reviewers for their valuable comments. CB thanks Pierre-Cyril Aubin-Frankowski for feedbacks on an earlier version of the draft. This

work was granted access to the HPC resources of IDRIS under the allocation 2024-AD011015891 made by GENCI. CB and EC acknowledge the support of the Agence nationale de la recherche, through the PEPR PDE-AI project (ANR-23-PEIA-0004). NC was supported by the ANR chair OTTOPIA ANR-20-CHIA-0030 and contributes to AI Excellence Cluster SequoIA (grant ANR-23-IACL-0009).

References

Aurélien Alfonsi and Benjamin Jourdain. Wasserstein projections in the convex order: regularity and characterization in the quadratic Gaussian case. *arXiv preprint arXiv:2506.23981*, 2025. (Cited on p. 4)

Jason Altschuler, Sinho Chewi, Patrik R Gerber, and Austin Stromme. Averaging on the Bures-Wasserstein manifold: dimension-free convergence of gradient descent. *Advances in Neural Information Processing Systems*, 34:22132–22145, 2021. (Cited on p. 4, 18, 32)

David Alvarez-Melis and Nicolo Fusi. Geometric Dataset Distances via Optimal Transport. *Advances in Neural Information Processing Systems*, 33:21428–21439, 2020. (Cited on p. 1, 6, 7, 20)

David Alvarez-Melis and Nicolò Fusi. Dataset Dynamics via Gradient Glows in Probability Space. In *International conference on machine learning*, pages 219–230. PMLR, 2021. (Cited on p. 7, 8, 9, 37)

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows: in Metric Spaces and in the Space of Probability Measures*. Springer Science & Business Media, 2008. (Cited on p. 1, 3, 17, 18, 25)

Ali Baouan, Mathieu Rosenbaum, and Sergio Pulido. An optimal transport based embedding to quantify the distance between playing styles in collective sports. *Journal of Quantitative Analysis in Sports*, 2025. (Cited on p. 19)

Mathias Beiglböck, Gudmund Pammer, and Stefan Schrott. A Brenier Theorem on $(\mathcal{P}_2(\mathcal{P}_2(\mathbb{R}^d)), W_2)$ and Applications to Adapted Transport. *arXiv preprint arXiv:2509.03506*, 2025. (Cited on p. 18)

Riccardo Bellazzi, Andrea Codegoni, Stefano Gualandi, Giovanna Nicora, and Eleonora Vercesi. The Gene Mover's Distance: Single-cell similarity via Optimal Transport. *arXiv preprint arXiv:2102.01218*, 2021. (Cited on p. 1)

Paul Berg, Bjoern Michele, Minh-Tan Pham, Laetitia Chapel, and Nicolas Courty. Horospherical Learning with Smart Prototypes. In *British Machine Vision Conference (BMVC)*, 2024. (Cited on p. 1, 17)

Paul Berg, Léo Buecher, Björn Michele, Minh-Tan Pham, Laetitia Chapel, and Nicolas Courty. Multi-Prototype Hyperbolic Learning Guided by Class Hierarchy. *International Journal of Computer Vision*, pages 1–16, 2025. (Cited on p. 1, 17)

Jérôme Bertrand and Benoît Kloeckner. A geometric study of Wasserstein spaces: Hadamard spaces. *Journal of Topology and Analysis*, 4(04):515–542, 2012. (Cited on p. 3)

Jérôme Bertrand and Benoît Kloeckner. A geometric study of Wasserstein spaces: isometric rigidity in negative curvature. *International Mathematics Research Notices*, 2016(5):1368–1386, 2016. (Cited on p. 3)

Rajendra Bhatia. *Matrix Analysis*, volume 169. Springer Science & Business Media, 2013. (Cited on p. 26)

Rajendra Bhatia, Tanvi Jain, and Yongdo Lim. On the Bures–Wasserstein distance between positive definite matrices. *Expositiones Mathematicae*, 37(2): 165–191, 2019. (Cited on p. 2, 18)

Clément Bonet, Christophe Vauthier, and Anna Korba. Flowing Datasets with Wasserstein over Wasserstein Gradient Flows. In *Forty-second International Conference on Machine Learning*, 2025a. (Cited on p. 1, 6, 8, 9, 18, 19, 20, 37, 38, 40)

Clément Bonet, Laetitia Chapel, Lucas Drumetz, and Nicolas Courty. Hyperbolic Sliced-Wasserstein via Geodesic and Horospherical Projections. In *Proceedings of 2nd Annual Workshop on Topology, Algebra, and Geometry in Machine Learning (TAG-ML)*, pages 334–370. PMLR, 2023a. (Cited on p. 1, 6, 16, 17)

Clément Bonet, Benoît Malézieux, Alain Rakotomamonjy, Lucas Drumetz, Thomas Moreau, Matthieu Kowalski, and Nicolas Courty. Sliced-Wasserstein on Symmetric Positive Definite Matrices for M/EEG Signals. In *International Conference on Machine Learning*, pages 2777–2805. PMLR, 2023b. (Cited on p. 6, 21)

Clément Bonet, Lucas Drumetz, and Nicolas Courty. Sliced-Wasserstein Distances and Flows on Cartan-Hadamard Manifolds. *Journal of Machine Learning Research*, 26(32):1–76, 2025b. (Cited on p. 1, 6, 16, 17, 21, 34)

Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. Sliced and Radon Wasserstein Barycenters of Measures. *Journal of Mathematical Imaging and Vision*, 51(1):22–45, 2015. (Cited on p. 6)

Charles Bouveyron and Marco Corneli. Scaling Optimal Transport to High-Dimensional Gaussian Distributions with Application to Domain Adaptation. *Statistics and Computing*, 36(2), 2026. ISSN 0960-3174. (Cited on p. 9)

Yann Brenier. Polar Factorization and Monotone Rearrangement of Vector-Valued Functions. *Communications on pure and applied mathematics*, 44(4): 375–417, 1991. (Cited on p. 2, 17, 24)

Martin R Bridson and André Haefliger. *Metric Spaces of Non-Positive Curvature*, volume 319. Springer Science & Business Media, 2013. (Cited on p. 3, 4, 16, 32)

Michael M Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: going beyond Euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017. (Cited on p. 1)

Herbert Busemann. *The Geometry of Geodesics*. Academic Press, New York, 1955. (Cited on p. 1)

Elsa Cazelles, Vivien Seguy, Jérémie Bigot, Marco Cuturi, and Nicolas Papadakis. Geodesic PCA versus Log-PCA of histograms in the Wasserstein space. *SIAM Journal on Scientific Computing*, 40 (2):B429–B456, 2018. (Cited on p. 9)

Ines Chami, Albert Gu, Dat P Nguyen, and Christopher Ré. HoroPCA: Hyperbolic Dimensionality Reduction via Horospherical Projections. In *International Conference on Machine Learning*, pages 1419–1429. PMLR, 2021. (Cited on p. 1, 16, 17)

Xiongjie Chen, Yongxin Yang, and Yunpeng Li. Augmented Sliced Wasserstein Distances. In *International Conference on Learning Representations*, 2022. (Cited on p. 6)

Yongxin Chen, Tryphon T Georgiou, and Allen Tannenbaum. Optimal transport for Gaussian mixture models. *IEEE Access*, 7:6269–6278, 2018. (Cited on p. 1, 21)

Christopher Criscitiello and Jungbin Kim. Horospherically Convex Optimization on Hadamard Manifolds Part I: Analysis and Algorithms. *arXiv preprint arXiv:2505.16970*, 2025. (Cited on p. 16, 17)

Marco Cuturi. Sinkhorn Distances: Lightspeed Computation of Optimal Transport. *Advances in neural information processing systems*, 26, 2013. (Cited on p. 6, 20)

Julie Delon and Agnes Desolneux. A Wasserstein-Type Distance in the space of Gaussian Mixture Models. *SIAM Journal on Imaging Sciences*, 13(2):936–970, 2020. (Cited on p. 1, 21)

Michael Ziyang Diao, Krishna Balasubramanian, Sinho Chewi, and Adil Salim. Forward-Backward Gaussian Variational Inference via JKO in the Bures-Wasserstein Space. In *International Conference on Machine Learning*, pages 7960–7991. PMLR, 2023. (Cited on p. 18)

Lars Doorenbos, Pablo Márquez Neila, Raphael Sznitman, and Pascal Mettes. Hyperbolic Random Forests. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. (Cited on p. 1, 17)

Aiden Durrant and Georgios Leontidis. HMSN: Hyperbolic Self-Supervised Learning by Clustering with Ideal Prototypes. *arXiv preprint arXiv:2305.10926*, 2023. (Cited on p. 1, 17)

Pavel Dvurechensky, Alexander Gasnikov, and Alexey Kroshnin. Computational Optimal Transport: Complexity by Accelerated Gradient Descent is Better than by Sinkhorn's Algorithm. In *International conference on machine learning*, pages 1367–1376. PMLR, 2018. (Cited on p. 20)

Pedram Emami and Brendan Pass. Optimal transport with optimal transport cost: the Monge–Kantorovich problem on Wasserstein spaces. *Calculus of Variations and Partial Differential Equations*, 64(2):43, 2025. (Cited on p. 18)

Xiran Fan, Chun-Hao Yang, and Baba Vemuri. Horocycle Decision Boundaries for Large Margin Classification in Hyperbolic Space. *Advances in neural information processing systems*, 36:11194–11204, 2023. (Cited on p. 1, 17)

Kilian Fatras, Younes Zine, Rémi Flamary, Remi Gribonval, and Nicolas Courty. Learning with minibatch Wasserstein : asymptotic and gradient properties. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 2131–2141. PMLR, 26–28 Aug 2020. (Cited on p. 6)

Kilian Fatras, Younes Zine, Szymon Majewski, Rémi Flamary, Rémi Gribonval, and Nicolas Courty. Minibatch optimal transport distances; analysis and applications. *arXiv preprint arXiv:2101.01792*, 2021. (Cited on p. 6)

Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, et al. POT: Python Optimal Transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021. (Cited on p. 35, 39)

Rémi Flamary, Cédric Vincent-Cuaz, Nicolas Courty, Alexandre Gramfort, Oleksii Kachaiev, Huy Quang Tran, Laurène David, Clément Bonet, Nathan Cassereau, Théo Gnassounou, Eloi Tanguy, Julie Delon, Antoine Collas, Sonia Mazelet, Laetitia Chapel, Tanguy Kerdoncuff, Xizheng Yu, Matthew Feickert, Paul Krzakala, Tianlin Liu, and Eduardo Fernandes Montesuma. POT Python Op-

timal Transport (version 0.9.5), 2024. URL `https://github.com/PythonOT/POT`. (Cited on p. 35)

P Thomas Fletcher, John Moeller, Jeff M Phillips, and Suresh Venkatasubramanian. Computing Hulls and Centerpoints in Positive Definite Space. *arXiv preprint arXiv:0912.1580*, 2009. (Cited on p. 16)

P Thomas Fletcher, John Moeller, Jeff M Phillips, and Suresh Venkatasubramanian. Horoball Hulls and Extents in Positive Definite Space. In *Workshop on Algorithms and Data Structures*, pages 386–398. Springer, 2011. (Cited on p. 16)

Masatoshi Fujii. Furuta Inequality and its Related Topics. *Annals of Functional Analysis*, 1(2):24–45, 2010. (Cited on p. 4)

Thomas Gallouët, Andrea Natale, and Gabriele Todeschi. From geodesic extrapolation to a variational BDF2 scheme for Wasserstein gradient flows. *Mathematics of Computation*, 2024. (Cited on p. 3, 34)

Thomas O Gallouët, Andrea Natale, and Gabriele Todeschi. Metric extrapolation in the Wasserstein space. *Calculus of Variations and Partial Differential Equations*, 64(5):147, 2025. (Cited on p. 4)

Matthias Gelbrich. On a Formula for the L2 Wasserstein Metric between Measures on Euclidean and Hilbert Spaces. *Mathematische Nachrichten*, 147(1): 185–203, 1990. (Cited on p. 2)

Jonathan Geuter, Clément Bonet, Anna Korba, and David Alvarez-Melis. DDEQs: Distributional Deep Equilibrium Models through Wasserstein Gradient Flows. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025. (Cited on p. 1)

Mina Ghadimi Atigh, Martin Keller-Ressel, and Pascal Mettes. Hyperbolic Busemann Learning with Ideal Prototypes. *Advances in Neural Information Processing Systems*, 34:103–115, 2021. (Cited on p. 1, 17)

Clark R Givens and Rae Michael Shortt. A Class of Wasserstein Metrics for Probability Distributions. *Michigan Mathematical Journal*, 31(2):231–240, 1984. (Cited on p. 2)

Pierre Glaser, Michael Arbel, and Arthur Gretton. KALE flow: A Relaxed KL Gradient Flow for Probabilities with Disjoint Support. *Advances in Neural Information Processing Systems*, 34:8018–8031, 2021. (Cited on p. 8)

Ariel Goodwin, Adrian S Lewis, Genaro López-Acedo, and Adriana Nicolae. A subgradient splitting algorithm for optimization on nonpositively curved metric spaces. *arXiv preprint arXiv:2412.06730*, 2024. (Cited on p. 17)

Ruiyu Han. Sliced Wasserstein distance between probability measures on Hilbert spaces. *arXiv preprint arXiv:2307.05802*, 2023. (Cited on p. 19)

Doron Haviv, Aram-Alexandre Pooladian, Dana Pe'er, and Brandon Amos. Wasserstein Flow Matching: Generative Modeling Over Families of Distributions. In *Forty-second International Conference on Machine Learning*, 2025a. (Cited on p. 1)

Doron Haviv, Ján Remšík, Mohamed Gatie, Catherine Snopkowski, Meril Takizawa, Nathan Pereira, John Bashkin, Stevan Jovanovich, Tal Nawy, Ronan Chaligne, et al. The covariance environment defines cellular niches for spatial inference. *Nature Biotechnology*, 43(2):269–280, 2025b. (Cited on p. 1)

Xinru Hua, Truyen Nguyen, Tam Le, Jose Blanchet, and Viet Anh Nguyen. Dynamic Flows on Curved Space Generated by Labeled Data. In Edith Elkind, editor, *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 3803–3811. International Joint Conferences on Artificial Intelligence Organization, 8 2023. Main Track. (Cited on p. 6, 8, 9, 20, 37)

Jonathan J. Hull. A Database for Handwritten Text Recognition Research. *IEEE Transactions on pattern analysis and machine intelligence*, 16(5):550–554, 1994. (Cited on p. 9)

Benoit Kloeckner. A geometric study of Wasserstein spaces: Euclidean spaces. *Annali della Scuola Normale Superiore di Pisa-Classe di Scienze*, 9(2):297–323, 2010. (Cited on p. 3)

Soheil Kolouri, Kimia Nadjahi, Umut Simsekli, Roland Badeau, and Gustavo Rohde. Generalized Sliced Wasserstein Distances. *Advances in neural information processing systems*, 32, 2019. (Cited on p. 6)

Alex Krizhevsky, Geoffrey Hinton, et al. Learning Multiple Layers of Features from Tiny Images. 2009. (Cited on p. 7)

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From Word Embeddings to Document Distances. In *International conference on machine learning*, pages 957–966. PMLR, 2015. (Cited on p. 1)

Marc Lambert, Sinho Chewi, Francis Bach, Silvère Bonnabel, and Philippe Rigollet. Variational inference via Wasserstein gradient flows. *Advances in Neural Information Processing Systems*, 35:14434–14447, 2022. (Cited on p. 18)

Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010. (Cited on p. 9)

Xinran Liu, Yikun Bai, Yuzhe Lu, Andrea Soltoggio, and Soheil Kolouri. Wasserstein Task Embedding for Mmeasuring Task Similarities. *Neural Networks*, 181:106796, 2025. (Cited on p. 6, 19, 20)

Robert J McCann. A Convexity Principle for Interacting Gases. *Advances in mathematics*, 128(1):153–179, 1997. (Cited on p. 3, 17)

Alfred Müller. Stochastic Ordering of Multivariate Normal Distributions. *Annals of the Institute of Statistical Mathematics*, 53(3):567–575, 2001. (Cited on p. 4, 23)

Khai Nguyen and Nhat Ho. Revisiting Sliced Wasserstein on Images: From Vectorization to Convolution. *Advances in Neural Information Processing Systems*, 35:17788–17801, 2022. (Cited on p. 6, 8, 35)

Khai Nguyen and Nhat Ho. Hierarchical Hybrid Sliced Wasserstein: A Scalable Metric for Heterogeneous Joint Distributions. *Advances in Neural Information Processing Systems*, 37:108140–108166, 2024. (Cited on p. 6, 20)

Khai Nguyen and Peter Mueller. Summarizing Nonparametric Bayesian Mixture Posteriors – Sliced Optimal Transport Metrics for Gaussian Mixtures. *Journal of Computational and Graphical Statistics*, pages 1–22, 2026. (Cited on p. 21)

Khai Nguyen, Hai Nguyen, Tuan Pham, and Nhat Ho. Lightspeed Geometric Dataset Distance via Sliced Optimal Transport. In *Forty-second International Conference on Machine Learning*, 2025a. (Cited on p. 6, 7, 8, 20, 35)

Xuan Son Nguyen, Shuo Yang, and Aymeric Histace. Neural networks on symmetric spaces of noncompact type. In *The Thirteenth International Conference on Learning Representations*, 2025b. (Cited on p. 1, 17)

XuanLong Nguyen. Borrowing strengh in hierarchical Bayes: Posterior concentration of the Dirichlet base measure. *Bernoulli*, 22(3):1535 – 1571, 2016. (Cited on p. 18)

Felix Otto. The geometry of dissipative evolution equations: the porous medium equation. *Communications in Partial Differential Equations*, 26(1-2): 101–174, 2001. (Cited on p. 2)

Mathilde Papillon, Sophia Sanborn, Johan Mathe, Louisa Cornelis, Abby Bertics, Domas Buracas, Hansen J Lillemark, Christian Shewmake, Fatih Dinc, Xavier Pennec, et al. Beyond Euclid: An Illustrated Guide to Modern Machine Learning with Geometric, Topological, and Algebraic Structures. *Machine Learning: Science and Technology*, 6(3): 031002, 2025. (Cited on p. 1)

François-Pierre Paty, Alexandre d'Aspremont, and Marco Cuturi. Regularity as Regularization: Smooth and Strongly Convex Brenier Potentials in Optimal Transport. In *International Conference on Artificial Intelligence and Statistics*, pages 1222–1232. PMLR, 2020. (Cited on p. 4)

Gabriel Peyré and Marco Cuturi. Computational Optimal Transport: With Applications to Data Science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019. (Cited on p. 2, 26)

Moritz Piening and Robert Beinert. Slicing the Gaussian Mixture Wasserstein Distance. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. (Cited on p. 19, 21, 39)

Moritz Piening and Robert Beinert. Slicing Wasserstein over Wasserstein via Functional Optimal Transport. In *The Fourteenth International Conference on Learning Representations*, 2026. (Cited on p. 7, 19)

Alessandro Pinzi and Giuseppe Savaré. Totally convex functions, L^2-Optimal transport for laws of random measures, and solution to the Monge problem. *arXiv preprint arXiv:2509.01768*, 2025. (Cited on p. 18)

Julien Rabin, Gabriel Peyré, Julie Delon, and Marc Bernot. Wasserstein Barycenter and its Application to Texture Mixing. In *Scale Space and Variational Methods in Computer Vision: Third International Conference, SSVM 2011, Ein-Gedi, Israel, May 29–June 2, 2011, Revised Selected Papers 3*, pages 435–446. Springer, 2012. (Cited on p. 6)

Filippo Santambrogio. *Optimal Transport for Applied Mathematicians*, volume 55. Springer, 2015. (Cited on p. 2, 17, 18, 28)

Meyer Scetbon, Marco Cuturi, and Gabriel Peyré. Low-Rank Sinkhorn Factorization. In *International Conference on Machine Learning*, pages 9344–9354. PMLR, 2021. (Cited on p. 6)

Vivien Seguy and Marco Cuturi. Principal Geodesic Analysis for Probability Measures under the Optimal Transport Metric. *Advances in Neural Information Processing Systems*, 28, 2015. (Cited on p. 1)

Yan Shu. From Hopf–Lax Formula to Optimal Weak Transfer Plan. *SIAM Journal on Mathematical Analysis*, 52(3):3052–3072, 2020. (Cited on p. 4, 23)

Sho Sonoda, Isao Ishikawa, and Masahiro Ikeda. Fully-Connected Network on Noncompact Symmetric Space and Ridgelet Transform based on Helgason-Fourier Analysis. In *International Conference on Machine Learning*, pages 20405–20422. PMLR, 2022. (Cited on p. 1, 17)

Asuka Takatsu. Wasserstein geometry of Gaussian measures. *Osaka Journal of Mathematics*, 48(4): 1005 – 1026, 2011. (Cited on p. 18, 22)

Hugues Van Assel, Nicolas Courty, Rémi Flamary, Aurélien Garivier, Mathurin Massias, Titouan Vayer, and Cédric Vincent-Cuaz. TorchDR : A PyTorch library for Dimensionality Reduction. `https://torchdr.github.io/`, 2024. (Cited on p. 8, 35)

Nina Vesseron, Elsa Cazelles, Alice Le Brigant, and Klein. On the Wasserstein Geodesic Principal Component Analysis of probability measures. In *The Fourteenth International Conference on Learning Representations*, 2026. (Cited on p. 9)

Cédric Villani. *Optimal Transport: Old and New*, volume 338. Springer, 2009. (Cited on p. 1, 2)

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2. (Cited on p. 35)

Ming-Xi Wang. Laplacian Eigenspaces, Horocycles and Neuron Models on Hyperbolic Spaces, 2021. (Cited on p. 1, 17)

Wei Wang, Dejan Slepčev, Saurav Basu, John A Ozolek, and Gustavo K Rohde. A Linear Optimal Transportation Framework for Quantifying and Visualizing Variations in Sets of Images. *International journal of computer vision*, 101(2):254–269, 2013. (Cited on p. 19, 20)

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. *arXiv preprint arXiv:1708.07747*, 2017. (Cited on p. 9)

Guomin Zhu, Wen-Long Li, and Xiaojun Cui. Busemann functions on the Wasserstein space. *Calculus of Variations and Partial Differential Equations*, 60 (3):1–16, 2021. (Cited on p. 2, 3, 4)

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. Yes
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. Yes
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. Yes

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. Yes
   (b) Complete proofs of all theoretical results. Yes
   (c) Clear explanations of any assumptions. Yes

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). Yes
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). Yes
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). Yes
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). Yes

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. Yes
   (b) The license information of the assets, if applicable. Not Applicable
   (c) New assets either in the supplemental material or as a URL, if applicable. Yes
   (d) Information about consent from data providers/curators. Not Applicable
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. Not Applicable

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. Not Applicable

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. Not Applicable

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. Not Applicable

Supplementary Materials

A BACKGROUND

We provide in this section additional background on the Busemann function, on the Wasserstein space, on the Wasserstein over Wasserstein space, and finally on Optimal Transport distances to compare labeled datasets.

A.1 Background on the Busemann Function

Let (X, d) be a geodesic metric space, *i.e.* a metric space, where each $x, y \in \mathrm{X}$ can be linked by a continuous curve $\gamma : [0,1] \to \mathrm{X}$ such that $\gamma(0) = x$, $\gamma(1) = y$ and which satisfies for all $s, t \in [0,1]$, $d\big(\gamma(s), \gamma(t)\big) = |t-s| d\big(\gamma(0), \gamma(1)\big)$.

Suppose that (X, d) admits geodesic rays, *i.e.* geodesic curves $\gamma : \mathbb{R}_+ \to \mathrm{X}$ such that for all $t, s \geq 0$, $d\big(\gamma(t), \gamma(s)\big) = |t-s| d\big(\gamma(0), \gamma(1)\big)$. Well-known spaces in which any geodesic can be extended as a geodesic ray are Hadamard spaces (Bridson and Haefliger, 2013), which are metric space of non-positive curvature, including Hadamard manifolds such as Euclidean spaces, Hyperbolic spaces or the space of Symmetric Positive Definite matrices (SPDs) with appropriate metrics.

The Busemann function B^γ associated to the geodesic ray γ is defined, for all $x \in \mathrm{X}$, as (see *e.g.* (Bridson and Haefliger, 2013, II. 8.17))

$$B^\gamma(x) = \lim_{t \to \infty} \; d\big(x, \gamma(t)\big) - d\big(\gamma(0), \gamma(t)\big) = d\big(x, \gamma(t)\big) - t d\big(\gamma(0), \gamma(1)\big). \tag{29}$$

This function has attracted a lot of attention as it provides a natural generalization of affine functions, and thus of hyperplanes through its level sets. Indeed, for $\mathrm{X} = \mathbb{R}^d$, $v \in \mathbb{R}^d$ and $\gamma(t) = x + tv$ for all $t \in \mathbb{R}$, the Busemann function is equal to

$$\forall y \in \mathbb{R}^d, \; B^\gamma(y) = -\left\langle y - x, \frac{v}{\|v\|_2} \right\rangle. \tag{30}$$

Its level sets $(B^\gamma)^{-1}(\{t\})$ for $t \in \mathbb{R}$ are called horospheres, and allow to define a generalization of affine hyperplanes beyond the Euclidean space.

The Busemann function can also be computed in closed-form in many spaces, including hyperbolic spaces (Chami et al., 2021; Bonet et al., 2023a), the space of SPDs with the Affine-Invariant metric (Fletcher et al., 2009, 2011) or with pullback Euclidean metrics (Bonet et al., 2025b). However, it has attracted the most attention in spaces where any geodesic is a ray. Thus, in this work, we provide an analysis of this function on the Wasserstein space, which has non-negative curvature, and in which not all geodesics can therefore be extended as rays.

In its original formulation, the Busemann function does not depend on the speed of the geodesic $d\big(\gamma(0), \gamma(1)\big)$ (Criscitiello and Kim, 2025). For instance, let $\mathrm{X} = \mathcal{M}$ be a Hadamard manifold. If we consider two geodesic rays γ and $\tilde{\gamma}$ both starting from $x \in \mathcal{M}$ and with respective speed $v \in T_x\mathcal{M}$ and $\frac{v}{\|v\|_x} \in T_x\mathcal{M}$, *i.e.* $\gamma(t) = \exp_x(tv)$ and $\tilde{\gamma}(t) = \exp_x(tv/\|v\|_x)$, then $B^\gamma(y) = B^{\tilde{\gamma}}(y)$ for any $y \in \mathcal{M}$. Indeed, let $y \in \mathcal{M}$,

$$\begin{aligned} B^{\tilde{\gamma}}(y) &= \lim_{t \to +\infty} \; d\left(y, \exp_x\left(t \frac{v}{\|v\|_x}\right)\right) - t \\ &= \lim_{s \to +\infty} \; d\big(y, \exp_x(sv)\big) - s\|v\|_x \qquad\qquad (s \leftarrow t/\|v\|_x) \\ &= \lim_{s \to +\infty} \; d\big(y, \exp_x(sv)\big) - s d\big(\gamma(0), \gamma(1)\big) \\ &= B^\gamma(y). \end{aligned} \tag{31}$$

Thus, it can always be assumed that the geodesics are of unit speed, up to a normalization of its speed. This is in particular the case on the Wasserstein space, leveraging its Riemannian structure.

In term of applications, the Busemann function has been used to perform classification with prototypes (Ghadimi Atigh et al., 2021; Durrant and Leontidis, 2023; Berg et al., 2024, 2025), to define boundary conditions on manifolds for classification (Fan et al., 2023; Doorenbos et al., 2024) or define layers of neural networks (Wang, 2021; Sonoda et al., 2022; Nguyen et al., 2025b). It has also been used to define projections on geodesics subspaces to perform Principal Component Analysis on Hyperbolic spaces (Chami et al., 2021) or on geodesics to define Sliced-Wasserstein distances on manifolds (Bonet et al., 2023a, 2025b). Recently, it has also received attention to define notions of convexity and gradients on Hadamard manifolds (Criscitiello and Kim, 2025) and more generally on Hadamard spaces (Goodwin et al., 2024).

A.2 Background on the Wasserstein space

We provide here some additional background on the Wasserstein space and on measure theory by recalling the disintegration of a measure on product spaces.

Optimal Transport. We recall that $\mathcal{P}_2(\mathbb{R}^d) = \{\mu \in \mathcal{P}(\mathbb{R}^d),\ \int \|x\|_2^2 \ \mathrm{d}\mu(x) < \infty\}$. Then, the Wasserstein distance between $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ is defined as

$$\mathrm{W}_2^2(\mu,\nu) = \inf_{\gamma \in \Pi(\mu,\nu)} \int \|x-y\|_2^2 \ \mathrm{d}\gamma(x,y), \tag{32}$$

with $\Pi(\mu,\nu)$ the set of couplings between μ and ν. Defining $\pi^1 : (x,y) \mapsto x$ and $\pi^2 : (x,y) \mapsto y$ the projections on the coordinates, and $\#$ the push forward operator which satisfies for any measurable map $\mathrm{T} : \mathbb{R}^d \to \mathbb{R}^{d'}$ and Borelian $A \in \mathcal{B}(\mathbb{R}^{d'})$, $(\mathrm{T}_\# \mu)(A) = \mu\big(\mathrm{T}^{-1}(A)\big)$, $\Pi(\mu,\nu) = \{\gamma \in \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d),\ \pi^1_\# \gamma = \mu,\ \pi^2_\# \gamma = \nu\}$.

Note that Optimal Transport costs can be more generally defined between measures on any measurable spaces $\mathcal{X}, \mathcal{Y}$ and for any cost $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ lower semi-continuous, using the Kantorovich formulation, *i.e.* for $\mu \in \mathcal{P}(\mathcal{X})$, $\nu \in \mathcal{P}(\mathcal{Y})$,

$$\mathrm{W}_c(\mu,\nu) = \inf_{\gamma \in \Pi(\mu,\nu)} \int c(x,y) \ \mathrm{d}\gamma(x,y). \tag{33}$$

For $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$ and $c(x,y) = \|x-y\|_2^2$, W_2 defines a distance, and $(\mathcal{P}_2(\mathbb{R}^d), \mathrm{W}_2)$ has a formal Riemannian structure. Between $\mu_0, \mu_1 \in \mathcal{P}_2(\mathbb{R}^d)$, we can always define a constant-speed geodesic $t \in [0,1] \mapsto \mu_t$, which satisfies for all $s,t \in [0,1]$, $\mathrm{W}_2(\mu_s,\mu_t) = |t-s|\mathrm{W}_2(\mu_0,\mu_1)$ (see *e.g.* (Santambrogio, 2015, Theorem 5.27)). In particular, these curves can be written as McCann's displacement interpolation (McCann, 1997)

$$\forall t \in [0,1], \mu_t = \big((1-t)\pi^1 + t\pi^2\big)_\# \gamma, \tag{34}$$

with $\gamma \in \Pi_o(\mu,\nu)$ an optimal coupling between μ and ν.

When $\mu \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ is absolutely continuous *w.r.t* the Lebesgue measure, Brenier's theorem (Brenier, 1991) states that there is a unique optimal coupling γ between μ and $\nu \in \mathcal{P}_2(\mathbb{R}^d)$, and that this optimal coupling is supported on a graph of a function, *i.e.* there exists $\mathrm{T} : \mathbb{R}^d \to \mathbb{R}^d$ such that $\mathrm{T}_\# \mu = \nu$ and $\gamma = (\mathrm{Id}, \mathrm{T})_\# \mu$. In this case, the geodesic between $\mu \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ and $\nu \in \mathcal{P}_2(\mathbb{R}^d)$ is of the form

$$\forall t \in [0,1],\ \mu_t = \big((1-t)\mathrm{Id} + t\mathrm{T}\big)_\# \mu. \tag{35}$$

We can also define the notion of exponential map as $\exp_\mu(v) = (\mathrm{Id} + v)_\# \mu$ for any $v \in L^2(\mu)$. Using this map, the geodesic can be written as $\mu_t = \exp_\mu\big(t(\mathrm{T} - \mathrm{Id})\big)$ for all $t \in [0,1]$. We can also define its inverse, the logarithm map, as $\log_\mu(\nu) = \mathrm{T}_\mu^\nu - \mathrm{Id}$ with T_μ^ν the OT map between μ and ν.

Thanks to the formal Riemannian structure, we can also define notions of gradients. We refer *e.g.* to (Ambrosio et al., 2008) for details on their definition and properties.

Bures-Wasserstein Space. The Wasserstein distance between two Gaussian has a closed-form, and is named the Bures-Wasserstein distance, *i.e.* for $\mu = \mathcal{N}(m_\mu, \Sigma_\mu)$, $\nu = \mathcal{N}(m_\nu, \Sigma_\nu)$ with $m_\mu, m_\nu \in \mathbb{R}^d$ and $\Sigma_\mu, \Sigma_\nu \in S_d^{++}(\mathbb{R})$,

$$\mathrm{W}_2^2(\mu,\nu) = \mathrm{BW}^2(\mu,\nu) = \|m_\mu - m_\nu\|_2^2 + \mathrm{Tr}\big(\Sigma_\mu + \Sigma_\nu - 2(\Sigma_\mu^{\frac{1}{2}} \Sigma_\nu \Sigma_\mu^{\frac{1}{2}})^{\frac{1}{2}}\big). \tag{36}$$

Gaussian being absolutely continuous measures, there is also a unique OT map which is given by

$$\forall x \in \mathbb{R}^d,\ \mathrm{T}(x) = m_\nu + A_\mu^\nu(x - m_\mu), \quad \text{with} \quad A_\mu^\nu = \Sigma_\mu^{-\frac{1}{2}}(\Sigma_\mu^{\frac{1}{2}}\Sigma_\nu\Sigma_\mu^{\frac{1}{2}})^{\frac{1}{2}}\Sigma_\mu^{-\frac{1}{2}}. \tag{37}$$

In particular, since $\mathrm{T}_\#\mu = \nu$, we also have the relation between the covariance matrices $A_\mu^\nu\Sigma_\mu(A_\mu^\nu)^T = \Sigma_\nu$. Moreover, the geodesics are of the form, for any $t \in [0,1]$, $\mu_t = \big((1-t)\mathrm{Id} + t\mathrm{T}\big)_\#\mu$. Since the map $x \mapsto (1-t)x + t\mathrm{T}(x)$ is affine for any $t \in [0,1]$, geodesics stay Gaussian at each time t, *i.e.* $\mu_t = \mathcal{N}(m_t, \Sigma_t)$. Moreover, their closed-forms can be computed, and are given by (Altschuler et al., 2021)

$$\begin{cases} m_t = (1-t)m_\mu + tm_\nu \\ \Sigma_t = \big((1-t)I_d + tA_\mu^\nu\big)\Sigma_\mu\big((1-t)I_d + tA_\mu^\nu\big). \end{cases} \tag{38}$$

When endowing the space of Gaussian $\mathrm{BW}(\mathbb{R}^d) = \{\mathcal{N}(m,\Sigma),\ m \in \mathbb{R}^d,\ \Sigma \in S_d^{++}(\mathbb{R})\}$ with the Bures-Wasserstein distance, the space $(\mathrm{BW}(\mathbb{R}^d), \mathrm{BW})$ is actually a real Riemannian manifold (Bhatia et al., 2019), and not just formally. Its tangent space at any $\mu = \mathcal{N}(m,\Sigma)$ is the space of affine functions with symmetric linear term (Diao et al., 2023, Appendix A.1). We identify it here as $T_\mu\mathrm{BW}(\mathbb{R}^d) = \mathbb{R}^d \times S_d(\mathbb{R})$ with $S_d(\mathbb{R})$ the space of symmetric matrices in $\mathbb{R}^{d\times d}$. Using this identification, the Riemannian metric is at any $(m_\mu, \Sigma_\mu) \in \mathbb{R}^d \times S_d^{++}(\mathbb{R})$, $(m,S) \in \mathbb{R}^d \times S_d(\mathbb{R})$, $\|(m,S)\|^2_{m_\mu,\Sigma_\mu} = \|m\|_2^2 + \|S\|^2_{\Sigma_\mu}$ where $\|S\|^2_{\Sigma_\mu} = \mathrm{Tr}(S\Sigma_\mu S)$ (Takatsu, 2011). We can also define the notion of exponential map at $\mu = \mathcal{N}(m_\mu,\Sigma_\mu)$ as, for any $(m,S) \in \mathbb{R}^d \times S_d(\mathbb{R})$,

$$\exp_\mu(m,S) = \mathcal{N}\big(m_\mu + m, (I_d+S)\Sigma_\mu(I_d+S)\big). \tag{39}$$

As the mean part is Euclidean, we will often just focus on the covariance part, and write $\exp_\Sigma(S) = (I_d + S)\Sigma(I_d+S)$. We can also define the logarithm map, for $\mu = \mathcal{N}(m_\mu,\Sigma_\mu)$ and $\nu = \mathcal{N}(m_\nu,\Sigma_\nu)$, as

$$\log_\mu(\nu) = (m_\nu - m_\mu, A_\mu^\nu - I_d). \tag{40}$$

Similarly, we can write $S = \log_{\Sigma_\mu}(\Sigma_\nu) = A_\mu^\nu - I_d$. As it is a Riemannian manifold, we can also define notions of Bures-Wasserstein gradients, see *e.g.* (Lambert et al., 2022; Diao et al., 2023).

This space is in particular of non-negative curvature (Takatsu, 2011), and thus not geodesically complete. Therefore, not any geodesic can be extended towards infinity.

Disintegration. We also recall the definition of the disintegration, see *e.g.* (Ambrosio et al., 2008, Theorem 5.3.1), which will be useful in subsequent proofs.

Definition 1 (Disintegration of a measure)**.** *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 the marginals as $\mu_Y = \pi^Y_\#\mu$ and $\mu_Z = \pi^Z_\#\mu$, where π^Y (respectively π^Z) is the projection on Y (respectively Z). Then, a family $\big(K(y,\cdot)\big)_{y\in\mathcal{Y}}$ 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:*

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

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. In the case where $X = \mathbb{R}^d$, we have existence and uniqueness of the disintegration (see (Santambrogio, 2015, Box 2.2) or (Ambrosio et al., 2008, Chapter 5) for the more general case).

A.3 Backgound on the Wasserstein over Wasserstein Space

When working with probability over probability distributions $\mathcal{P}_2\big(\mathcal{P}_2(\mathbb{R}^d)\big)$, a natural metric is to use the OT distance with W_2 as groundcost, which we call the Wasserstein over Wasserstein (WoW) distance, *i.e.* for any $\mathbb{P},\mathbb{Q} \in \mathcal{P}_2\big(\mathcal{P}_2(\mathbb{R}^d)\big)$,

$$\mathrm{W}^2_{\mathrm{W}_2}(\mathbb{P},\mathbb{Q}) = \inf_{\Gamma\in\Pi(\mathbb{P},\mathbb{Q})} \int \mathrm{W}_2^2(\mu,\nu)\ \mathrm{d}\Gamma(\mu,\nu). \tag{41}$$

This defines a distance on $\mathcal{P}_2\big(\mathcal{P}_2(\mathbb{R}^d)\big)$ (Nguyen, 2016), and the resulting space has also a geodesic structure (Bonet et al., 2025a; Pinzi and Savaré, 2025). Moreover, several recent works have investigated the analog of Brenier's theorem on this space (Emami and Pass, 2025; Pinzi and Savaré, 2025; Beiglböck et al., 2025).

WoW Gradients. Bonet et al. (2025a) recently defined a notion of gradient on $\big(\mathcal{P}_2\big(\mathcal{P}_2(\mathbb{R}^d)\big), \mathrm{W}_{\mathrm{W}_2}\big)$, see (Bonet et al., 2025a, Definition 3.3). In their paper, they give a more general definition on manifolds, for clarity, we report here the definition on $\mathcal{P}_2\big(\mathcal{P}_2(\mathbb{R}^d)\big)$. For any $\gamma \in \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d)$, denote $\phi^1(\gamma) = \pi^1_\# \gamma$ and $\phi^2(\gamma) = \pi^2_\# \gamma$. Moreover, let $L^2\big(\mathbb{P}, T\mathcal{P}_2(\mathbb{R}^d)\big) = \{\xi : \mathcal{P}_2(\mathbb{R}^d) \to T\mathcal{P}_2(\mathbb{R}^d),\ \int \|\xi(\mu)\|^2_{L^2(\mu)}\ \mathrm{d}\mathbb{P}(\mu) < \infty\}$.

Definition 2 (WoW Gradient on $\mathcal{P}_2\big(\mathcal{P}_2(\mathbb{R}^d)\big)$)**.** *Let $\mathbb{F} : \mathcal{P}_2\big(\mathcal{P}_2(\mathbb{R}^d)\big) \to \mathbb{R}$. The WoW gradient of $\mathbb{F}$ at $\mathbb{P} \in \mathcal{P}_2\big(\mathcal{P}_2(\mathbb{R}^d)\big)$, if it exists, is defined as the map $\xi \in L^2\big(\mathbb{P}, T\mathcal{P}_2(\mathbb{R}^d)\big)$, which satisfies for any $\mathbb{Q} \in \mathcal{P}_2\big(\mathcal{P}_2(\mathbb{R}^d)\big)$ and $\mathbb{\Gamma} \in \{\mathbb{\Gamma} \in \mathcal{P}_2\big(\mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d)\big),\ \phi^1_\# \mathbb{\Gamma} = \mathbb{P},\ \phi^2_\# \mathbb{\Gamma} = \mathbb{Q},\ \iint \|x - y\|_2^2\ \mathrm{d}\gamma(x,y)\mathrm{d}\mathbb{\Gamma}(\gamma) = \mathrm{W}^2_{\mathrm{W}_2}(\mathbb{P}, \mathbb{Q})\}$,*

$$\mathbb{F}(\mathbb{Q}) = \mathbb{F}(\mathbb{P}) + \iint \langle \xi(\pi^1_\# \gamma)(x), y - x\rangle\ \mathrm{d}\gamma(x,y)\mathrm{d}\mathbb{\Gamma}(\gamma) + o\big(\mathrm{W}_{\mathrm{W}_2}(\mathbb{P}, \mathbb{Q})\big). \tag{42}$$

In the following, we note $\nabla_{\mathrm{W}_{\mathrm{W}_2}} \mathbb{F}(\mathbb{P}) = \xi$ such a gradient.

Let $\mathbb{F} : \mathcal{P}_2\big(\mathcal{P}_2(\mathbb{R}^d)\big) \to \mathbb{R}$ be a function such that there exists for any discrete measure of the form $\mathbb{P} = \frac{1}{C}\sum_{c=1}^C \delta_{\mu_c^n}$ with $\mu_c^n = \frac{1}{n}\sum_{i=1}^n \delta_{x_{i,c}}$, a function $F : \mathbb{R}^{n\times C} \to \mathbb{R}$ satisfying $\mathbb{F}(\mathbb{P}) = F(\mathbf{x})$ with $\mathbf{x} = (x_{i,c})_{i,c}$. The WoW gradient of $\mathbb{F}$, if well defined, can be obtained by rescaling the Euclidean gradient of F. More precisely, $\nabla_{\mathrm{W}_{\mathrm{W}_2}} \mathbb{F}(\mathbb{P})(\mu_c^n)(x_{i,c}) = nC\nabla F(\mathbf{x})_{i,c}$ (Bonet et al., 2025a, Proposition B.7). In practice, ∇F can be obtained using backpropagation.

This gradient allows to perform gradient descent on $(\mathcal{P}_2\big(\mathcal{P}_2(\mathbb{R}^d)\big), \mathrm{W}_{\mathrm{W}_2})$, by the scheme, for any $\tau > 0$,

$$\forall k \geq 0,\ \mathbb{P}_{k+1} = \exp_{\mathbb{P}_k}\big(-\tau \nabla_{\mathrm{W}_{\mathrm{W}_2}} \mathbb{F}(\mathbb{P}_k)\big). \tag{43}$$

For $\mathbb{P} = \frac{1}{C}\sum_{c=1}^C \delta_{\mu_c}$, the scheme can be obtained by applying to each particle $x_{i,c}^k$ the update,

$$\forall i \in \{1, \dots, n\},\ c \in \{1, \dots, C\},\ k \geq 0,\ x_{i,c}^{k+1} = x_{i,c}^k - \tau \nabla_{\mathrm{W}_{\mathrm{W}_2}} \mathbb{F}(\mathbb{P}_k)(\mu_c^k)(x_{i,c}^k). \tag{44}$$

Computational Properties and Variants. The computation of $\mathrm{W}_{\mathrm{W}_2}$ can be costly. Indeed, for $\mathbb{P} = \frac{1}{C}\sum_{c=1}^C \delta_{\mu_c^n}$ and $\mathbb{Q} = \frac{1}{C}\sum_{c=1}^C \delta_{\nu_c^n}$ two discrete distributions with $\mu_c^n, \nu_c^n \in \mathcal{P}_2(\mathbb{R}^d)$ empirical distributions with n samples, it is required to first compute $\mathcal{O}(C^2)$ OT distances with n samples, and a final OT distance with C samples. In general, $C \ll n$, and thus the computational complexity is $\mathcal{O}(C^2 n^3 \log n)$.

To alleviate this computational burden, several approximations can be used. On one hand, it is possible to use a less costly distance as groundcost, such as the Sliced-Wasserstein distance (Baouan et al., 2025; Piening and Beinert, 2025), reducing the complexity to $\mathcal{O}(C^2 Ln \log n + C^3 \log C)$. One could also use Linear OT (Wang et al., 2013; Liu et al., 2025) as groundcost, hence allowing to compute only $2C$ OT problems instead of $\mathcal{O}(C^2)$ and reducing the complexity to $\mathcal{O}(Cn^3 \log n)$. This is particularly appealing when there are lots of classes.

Piening and Beinert (2026) very recently proposed a doubled slicing distance to compare distributions on $\mathcal{P}_2\big(\mathcal{P}_2(\mathbb{R}^d)\big)$. Given a projection $P^\theta : \mathbb{R}^d \to \mathbb{R}$ for $\theta \in \Theta$, they first project in 1D the distributions $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ using $\varphi^\theta(\mu) = P^\theta_\# \mu \in \mathcal{P}_2(\mathbb{R})$, *i.e.* for $\mathbb{P} \in \mathcal{P}_2\big(\mathcal{P}_2(\mathbb{R}^d)\big)$, $\varphi^\theta_\# \mathbb{P} \in \mathcal{P}_2\big(\mathcal{P}_2(\mathbb{R})\big)$. Then, they project $\mathbb{P}^\theta := \varphi^\theta_\# \mathbb{P}$ in $\mathcal{P}_2(L^2([0,1])$ using, for any $\mu \in \mathcal{P}_2(\mathbb{R})$, $\phi(\mu) = F_\mu^{-1}$, and use the Sliced-Wasserstein distance on the Hilbert space $L^2([0,1])$ (Han, 2023). More precisely, they define for $\mathbb{P}, \mathbb{Q} \in \mathcal{P}_2\big(\mathcal{P}_2(\mathbb{R}^d)\big)$

$$\mathrm{DSW}_2^2(\mathbb{P}, \mathbb{Q}) = \int_{S^{d-1}} \mathrm{SW}_2^2(\phi_\# \varphi^\theta_\# \mathbb{P}, \phi_\# \varphi^\theta_\# \mathbb{Q})\ \mathrm{d}\lambda(\theta), \tag{45}$$

with the Sliced-Wasserstein distance defined on $\mathcal{P}_2\big(L^2([0,1])\big)$. For SW on this space, they use Gaussian on $\mathcal{P}_2\big(L^2([0,1])\big)$ as slicing measure, which are sampled using Gaussian processes.

Note that given $f \in L^2([0,1])$ and $\theta \in \Theta$, $\mu \sim \mathbb{P}$ is projected on $\mathbb{R}$ by the map $Q^{f,\theta}(\mu) = \langle f, F^{-1}_{P^\theta_\# \mu}\rangle_{L^2([0,1])}$. While for SWB1DG, given a geodesic ray η on $\mathcal{P}(\mathbb{R})$ and $\theta \in \Theta$, the projection is $Q^{\eta,\theta}(\mu) = B^\eta(P^\theta_\# \mu) = -\langle F_1^{-1} - F_0^{-1}, F^{-1}_{P^\theta_\# \mu} - F_0^{-1}\rangle_{L^2([0,1])}$ using (17) and noting F_0^{-1} and F_1^{-1} the quantile functions of η_0 and η_1. Thus both DSW and SWB1DG are very similar as they use an inner product on $L^2([0,1])$ for their projection. However, they differ on how to sample the directions and on the projections. In particular, in SWB1DG, we only sample directions which produce valid geodesic rays, and take elements from $L^2([0,1])$ which are difference of left continuous and non-decreasing functions, while DSW can sample on the full space $L^2([0,1])$.

A.4 Background on Optimal Transport Dataset Distances

A labeled dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ with $x_1, \dots, x_n \in \mathcal{X} = \mathbb{R}^d$ the features and $y_1, \dots, y_n \in \mathcal{Y} = \{1, \dots, C\}$ their associated labels, which we suppose here to be discrete, can be represented as a probability distribution $\mu_{\mathcal{D}}$ over $\mathcal{X} \times \mathcal{Y}$, *i.e.* $\mu_{\mathcal{D}} = \frac{1}{n}\sum_{i=1}^n \delta_{(x_i,y_i)}$. Thus, a natural way to compare labeled datasets is through distances on $\mathcal{P}(\mathcal{X} \times \mathcal{Y})$.

OTDD. Optimal transport distances can be defined on this space if provided a suitable groundcost on $\mathcal{X} \times \mathcal{Y}$. While there is usually a natural distance on $\mathcal{X}$, it is less clear which cost to use on $\mathcal{Y}$ as the labels of the classes might be chosen arbitrarily in practice. Thus, Alvarez-Melis and Fusi (2020) proposed to embed labels $y \in \mathcal{Y}$ in the space $\mathcal{P}_2(\mathcal{X})$ through their conditional distributions, *i.e.* using an embedding $\varphi : \mathcal{Y} \to \mathcal{P}_2(\mathcal{X})$ defined as $\varphi(y) = \frac{1}{n_y}\sum_{i=1}^n \delta_{x_i}\mathbb{1}_{\{y_i=y\}}$ where $n_y = \sum_{i=1}^n \mathbb{1}_{\{y_i=y\}}$ is the cardinal of the class y, and to represent labeled datasets as distributions on $\mathcal{X} \times \mathcal{P}_2(\mathcal{X})$. One natural groundcost in this space is then the geodesic distance on the product space, defined for any $(x, y), (x', y') \in \mathcal{X} \times \mathcal{Y}$, as

$$d_{\mathcal{D}}\big((x, y), (x', y')\big)^2 = \|x - x'\|_2^2 + \mathrm{W}_2^2\big(\varphi(y), \varphi(y')\big). \tag{46}$$

Then, Alvarez-Melis and Fusi (2020) proposed to compare two datasets $\mathcal{D}, \mathcal{D}'$ using optimal transport with this ground cost, called the Optimal Transport Dataset Distance (OTDD):

$$\mathrm{OTDD}^2(\mathcal{D}, \mathcal{D}') = \inf_{\gamma \in \Pi(\mu_{\mathcal{D}}, \mu_{\mathcal{D}'})} \int d_{\mathcal{D}}\big((x, y), (x', y')\big)^2 \,\mathrm{d}\gamma\big((x, y), (x', y')\big). \tag{47}$$

For C classes, and a maximum of n_C samples by class, OTDD requires solving $\mathcal{O}(C^2)$ OT problems with n_C samples, and a final OT problem of Cn_C samples, which leads to the complete complexity of $\mathcal{O}(C^3 n_C^3)$. It is thus a very costly distance to compute. Alvarez-Melis and Fusi (2020) hence proposed to approximate it using an entropic regularization for the final OT problem (Cuturi, 2013), and a Gaussian approximation for the C^2 smaller OT problems, reducing the complexity to $\mathcal{O}\big(Cn_C d^2 + C^2 d^3 + \varepsilon^{-2} n_C^2 C^2 \log(n_C C)\big)$ (Dvurechensky et al., 2018), which remains quite costly.

Variants of OTDD. This prohibitive computational cost motivated the introduction of variants of OTDD. For instance, Liu et al. (2025) proposed to embed the labels in $\mathbb{R}^d$ using Multidimensional Scaling methods, and to use Linear Optimal Transport (Wang et al., 2013), which allows computing only $2C$ OT problems with n_C samples. Hua et al. (2023) proposed to use dimension reduction on the labels to do the Gaussian approximation in a lower dimensional space, and to compare the datasets with a Maximum Mean Discrepancy (MMD), while Bonet et al. (2025a) proposed to represent datasets on $\mathcal{P}_2\big(\mathcal{P}_2(\mathcal{X})\big)$ and to compare them with a suitable MMD on this space. Note that from this point of view, any distance introduced in Appendix A.3 could be used to compare datasets.

Slicing OTDD. Nguyen et al. (2025a) recently proposed a Sliced-Wasserstein distance on the space $\mathcal{P}_2\big(\mathcal{X} \times \mathcal{P}_2(\mathcal{X})\big)$ to compare labeled datasets. This requires to construct a projection from $\mathcal{X} \times \mathcal{P}_2(\mathcal{X})$ to $\mathbb{R}$ to be able to project the distribution of pairs $(x_i, \varphi(y_i))$ onto a distribution in $\mathcal{P}_2(\mathbb{R})$. Their construction is based on a projection of the form, for any $(x, y) \in \mathcal{X} \times \mathcal{Y}$,

$$P^{\alpha,\theta,\lambda}(x, y) = \alpha_1 P^\theta(x) + \sum_{i=1}^k \alpha_{i+1} \mathcal{M}^{\lambda_i}\big(P^\theta_\# \varphi(y)\big), \tag{48}$$

with $\alpha \in S^k$, $P^\theta : \mathcal{X} \to \mathbb{R}$ and $\mathcal{M}^\lambda : \mathcal{P}_2(\mathbb{R}) \to \mathbb{R}$ the moment transform projection, defined for $\lambda \in \mathbb{N}$ and $\mu \in \mathcal{P}_\lambda(\mathbb{R})$ as

$$\mathcal{M}^\lambda(\mu) = \int \frac{x^\lambda}{\lambda!} \,\mathrm{d}\mu(x). \tag{49}$$

The random linear combination corresponds to the Hierarchical Hybrid projection (Nguyen and Ho, 2024), and allows to combine projection on different space in order to define a projection on a product space. In practice, the λ are sampled using a zero-truncated Poisson distribution, but this projection can be numerically unstable when λ is too big because of the $\lambda!$.

B SLICING GAUSSIAN MIXTURES

B.1 Background on the Wasserstein over Bures-Wasserstein Space

Gaussian mixtures can be represented as discrete probability distributions on the space of Gaussian distributions (Chen et al., 2018; Delon and Desolneux, 2020), *i.e.* as discrete distributions $\mathbb{P} \in \mathcal{P}_2\big(\mathrm{BW}(\mathbb{R}^d)\big)$ of the form $\mathbb{P} = \sum_{k=1}^K \alpha_k \delta_{\mu_k}$ with $\mu_k = \mathcal{N}(m_k, \Sigma_k)$, $m_k \in \mathbb{R}^d$, $\Sigma_k \in S_d^{++}(\mathbb{R})$. Thus, a natural distance to compare Gaussian mixtures is given by the OT distance with BW^2 as groundcost:

$$\mathrm{W}_{\mathrm{BW}}^2(\mathbb{P}, \mathbb{Q}) = \inf_{\Gamma \in \Pi(\mathbb{P}, \mathbb{Q})} \int \mathrm{BW}^2(\mu, \nu) \, \mathrm{d}\Gamma(\mu, \nu). \tag{50}$$

This defines a distance on $\mathcal{P}_2\big(\mathrm{BW}(\mathbb{R}^d)\big)$. As $\mathrm{BW}(\mathbb{R}^d)$ is a Riemannian manifold, $\mathcal{P}_2\big(\mathrm{BW}(\mathbb{R}^d)\big)$ has a Riemannian structure, and we can define notions such as geodesics or gradients.

Computational Properties. To compute it between two discrete Gaussian mixtures $\mathbb{P} = \frac{1}{K}\sum_{k=1}^K \alpha_k \delta_{\mu_k}$ and $\mathbb{Q} = \sum_{k=1}^K \beta_k \delta_{\nu_k}$ with $\mu_k = \mathcal{N}(m_k^\mu, \Sigma_k^\mu)$ and $\nu_k = \mathcal{N}(m_k^\nu, \Sigma_k^\nu)$, it is required to compute $\mathcal{O}(K^2)$ BW distances, which has a complexity of $\mathcal{O}(K^2 d^3)$.

Variants. To alleviate this computational burden, several methods were proposed. First, Nguyen and Mueller (2026) proposed to compare Gaussian mixtures seeing them as distributions over the product space $\mathbb{R}^d \times S_d^{++}(\mathbb{R})$, and endowing $S_d^{++}(\mathbb{R})$ with the Log-Euclidean metric as in (Bonet et al., 2023b, 2025b). However, it is not specifically designed for Gaussian mixtures, and thus they also proposed to use a doubly SW distance.

Let $\mathbb{P} \in \mathcal{P}_2\big(\mathrm{BW}(\mathbb{R}^d)\big)$, $\theta \in S^{d-1}$ and $\varphi^\theta : \mathcal{P}_2(\mathbb{R}^d) \to \mathcal{P}_2(\mathbb{R})$ defined as $\varphi^\theta(\mu) = P^\theta_\# \mu$. For any $\mu = \mathcal{N}(m_\mu, \Sigma_\mu) \sim \mathbb{P}$, $\varphi^\theta(\mu) = \mathcal{N}(\langle m_\mu, \theta\rangle, \theta^T \Sigma_\mu \theta)$. Thus, $\mathbb{P}^\theta = \varphi^\theta_\# \mathbb{P} \in \mathcal{P}_2\big(\mathrm{BW}(\mathbb{R})\big)$. Moreover, the space of 1D Gaussian can be identified as a product space over the means and standard deviations $\mathbb{R} \times \mathbb{R}_+^*$. For $\mu_\theta = \mathcal{N}(m_\theta, \Sigma_\theta) \in \mathrm{BW}(\mathbb{R})$, let $\Xi : \mathrm{BW}(\mathbb{R}) \to \mathbb{R}^2$ such that $\Xi(\mu_\theta) = (m_\theta, \sigma_\theta)$. Then $\Xi_\# \mathbb{P}^\theta \in \mathcal{P}_2(\mathbb{R}^2)$. Piening and Beinert (2025); Nguyen and Mueller (2026) proposed to define the Doubly Mixture Sliced-Wasserstein distance (DMSW) between $\mathbb{P}, \mathbb{Q} \in \mathcal{P}_2\big(\mathrm{BW}(\mathbb{R}^d)\big)$ as

$$\mathrm{DMSW}^2(\mathbb{P}, \mathbb{Q}) = \int_{S^{d-1}} \mathrm{SW}_2^2\left(\Xi_\# \varphi^\theta_\# \mathbb{P}, \Xi_\# \varphi^\theta_\# \mathbb{Q}\right) \mathrm{d}\lambda(\theta), \tag{51}$$

where the inner SW is between the distributions of the mean and covariances in $\mathbb{R}^2$.

B.2 Slicing Gaussian Mixtures with Busemann

As the Busemann function allows to project any probability distribution onto $\mathbb{R}$, it is natural to use it as a projection to define Sliced-Wasserstein distances for the purpose of comparing mixtures, and in particular mixtures of Gaussian. We discuss here how we can define Sliced-Wasserstein distances based on the closed-forms of the Busemann function in 1D (17) and between Gaussian distributions (19). Note that both constructions are very close to the one presented in Section 4.3, but specialized to mixtures of Gaussian. We refer to Appendix F for numerical experiments.

Busemann on Gaussian. Leveraging the closed-form of the Busemann function between Gaussians (19), we can project any atom of a Gaussian mixture on $\mathbb{R}$. To define a sliced distance, we only need to construct geodesic rays on which the Busemann function is well defined. We choose to fix $\eta_0 = \mathcal{N}(0, I_d)$, and sample $\eta_1 = \mathcal{N}(m_1, \Sigma_1)$ such that $m_1 \in S^{d-1}$, $\Sigma_1 \in S_d^{++}(\mathbb{R})$ with $\Sigma_1^{\frac{1}{2}} \succeq I_d$ and $\mathrm{W}_2^2(\eta_0, \eta_1) = 1$.

To enforce $\Sigma_1^{\frac{1}{2}} \succeq I_d$, we remark that it is equivalent to $\log_{I_d}(\Sigma_1) = \Sigma_1^{\frac{1}{2}} - I_d \succeq 0$, where $\log_{I_d}$ is the logarithm map in $S_d^{++}(\mathbb{R})$ (see Appendix A.2 for the definitions of exponential and logarithm map on the Bures-Wasserstein space). Thus we sample $S = \Delta \mathrm{diag}(|\theta|) \Delta^T \in S_d^{++}(\mathbb{R})$ with $\Delta \in O_d(\mathbb{R})$ an orthogonal matrix, $\theta \in S^{d-1}$ and define $\Sigma_1 := \exp_{I_d}(S) = (I_d + S)^2$. To enforce $\mathrm{W}_2^2(\eta_0, \eta_1) = 1$, we first observe that

$$\mathrm{W}_2^2(\eta_0, \eta_1) = \|\log_{\eta_0}(\eta_1)\|^2_{L^2(\eta_0)} = \|m_1 - m_0\|_2^2 + \|\log_{\Sigma_0}(\Sigma_1)\|^2_{\Sigma_0} = \|m_1\|_2^2 + \|S\|^2_{\Sigma_0}, \tag{52}$$

where $\|S\|^2_{\Sigma_0} = \mathrm{Tr}(S\Sigma_0 S)$ (Takatsu, 2011). Therefore, we simply normalize the vectors in the tangent space to obtain $\mathrm{W}_2^2(\eta_0, \eta_1) = 1$, *i.e.*

$$\begin{cases} m_1 = \frac{m_1}{\sqrt{\|m_1\|_2^2 + \|S\|^2_{\Sigma_0}}} \\ S = \frac{S}{\sqrt{\|m_1\|_2^2 + \|S\|^2_{\Sigma_0}}}. \end{cases} \tag{53}$$

Defining $\lambda = \mathcal{U}(S^{d-1} \times O_d(\mathbb{R}) \times S^{d-1})$, $\vartheta = (m_1, \Delta, \theta)$ and η^ϑ the resulting geodesic ray starting from η_0 and passing through η_1, we define for $\mathbb{P}, \mathbb{Q}$ Gaussian mixtures, the Busemann Gaussian Mixture Sliced-Wasserstein distance (BGMSW) as

$$\mathrm{BGMSW}^2(\mathbb{P}, \mathbb{Q}) = \int \mathrm{W}_2^2(B^{\eta^\vartheta}_\# \mathbb{P}, B^{\eta^\vartheta}_\# \mathbb{Q}) \, \mathrm{d}\lambda(\vartheta). \tag{54}$$

Busemann on 1D Gaussian. BGMSW requires to compute the Busemann function between Gaussian distributions, which might be computationally heavy in high dimension. Thus, we also propose a second distance, by first projecting the Gaussian in 1D, and then leveraging the closed-form of the Busemann function for 1D Gaussians (18). More precisely, we consider $\mathbb{P} = \sum_{k=1}^K \alpha_k \delta_{\mu^k}$ with $\mu^k = \mathcal{N}(m_k, \Sigma_k)$. Denoting $\varphi^\theta(\mu) = P^\theta_\# \mu$ where $P^\theta = \langle \theta, \cdot \rangle$, we have $\mathbb{P}^\theta := \varphi^\theta_\# \mathbb{P} = \sum_{k=1}^K \alpha_k \delta_{P^\theta_\# \mu_k}$ with $P^\theta_\# \mu_k = \mathcal{N}(\langle m_k, \theta \rangle, \theta^T \Sigma_k \theta)$ a one dimensional Gaussian distribution.

For the choice of the geodesic ray, we set $\eta_0 = \mathcal{N}(0, 1)$ and $\eta_1 = \mathcal{N}(m_1, \sigma_1^2)$, where we sample $m_1 \in [-1, 1]$ uniformly, and fix $\sigma_1 = 1 + \sqrt{1 - m_1^2}$ to enforce $\sigma_1 \geq \sigma_0$ and $\mathrm{W}_2^2(\eta_0, \eta_1) = 1$. Defining $\lambda = \mathcal{U}(S^{d-1} \times [-1, 1])$, we define the Busemann 1D Gaussian Mixture Sliced-Wasserstein distance (B1DGMSW) as

$$\mathrm{B1DGMSW}^2(\mathbb{P}, \mathbb{Q}) = \int \mathrm{W}_2^2(B^{\eta^{m_1}}_\# \varphi^\theta_\# \mathbb{P}, B^{\eta^{m_1}}_\# \varphi^\theta_\# \mathbb{Q}) \, \mathrm{d}\lambda(\theta, m_1). \tag{55}$$

We notice that this distance resembles DMSW. Indeed, for DMSW, the projection of $\mu = \mathcal{N}(m, \Sigma) \sim \mathbb{P}$ on $\mathbb{R}$ is given, for $\theta_d \in S^{d-1}, \theta_2 \in S^1$, by

$$Q^{\theta_d, \theta_1}(\mu) = \left\langle \theta_2, \begin{pmatrix} \langle m, \theta_d \rangle \\ \theta_d^T \Sigma \theta_d \end{pmatrix} \right\rangle, \tag{56}$$

and θ_2, θ_d are sampled uniformly on the sphere. On the other hand, for B1DGMSW, the projection is, for η a geodesic ray on $\mathrm{BW}(\mathbb{R})$ and $\theta \in S^{d-1}$,

$$\begin{aligned} Q^{\eta, \theta}(\mu) = B^\eta(P^\theta_\# \mu) &= B^\theta\big(\mathcal{N}(\langle m, \theta \rangle, \theta^T \Sigma \theta)\big) \\ &= -\left\langle \begin{pmatrix} m_1 \\ \sigma_1 - 1 \end{pmatrix}, \begin{pmatrix} \langle m, \theta \rangle \\ \theta^T \Sigma \theta - 1 \end{pmatrix} \right\rangle. \end{aligned} \tag{57}$$

Therefore DMSW and B1DGMSW are very similar. The main difference is that the directions in $\mathbb{R}^2$ for B1DGMSW are centered around $(0, 1)$.

B.3 Properties of our Proposed Distances

The proofs of this section can be found in Appendix D.3.

Theoretical Properties. We first have that BGMSW is a pseudo distance.

Proposition 6. BGMSW *is a pseudo-distance on* $\mathrm{BW}(\mathbb{R}^d)$.

Showing that BGMSW is a distance would require to show that the Busemann function on $\mathrm{BW}(\mathbb{R}^d)$ allows defining an injective Radon transform, which is outside the scope of this work.

Concerning B1DGMSW, we can exploit that the 1D Gaussian space is actually Euclidean, and thus we can show that it is a well-defined distance, which is also bounded by W_{BW}.

Proposition 7. *For any* $\mathbb{P}, \mathbb{Q} \in \mathcal{P}_2\big(\mathrm{BW}(\mathbb{R}^d)\big)$,

$$\mathrm{B1DGMSW}^2(\mathbb{P}, \mathbb{Q}) \leq \mathrm{W}^2_{\mathrm{BW}}(\mathbb{P}, \mathbb{Q}). \tag{58}$$

Proposition 8. B1DGMSW *is a distance on the space of discrete Gaussian mixtures* $\bigcup_{K>0} \mathrm{GMM}_d(K)$, *with* $\mathrm{GMM}_d(K)$ *the set of Gaussian mixtures on* $\mathbb{R}^d$ *with* K *components.*

Computational Properties. On a computational point of view, we approximate BGMSW by a Monte-Carlo approximation with L projections. Given discrete mixtures $\mathbb{P}_k = \sum_{k=1}^K \alpha_k \delta_{\mu_k}$ and $\mathbb{Q}_k = \sum_{k=1}^K \beta_k \delta_{\nu_k}$, BGMSW requires to project K Gaussian with the Busemann function, which has a complexity of $\mathcal{O}(LKd^3)$. Then, it is also required to solve a 1D OT transport problem with K samples. Thus, the full complexity is $\mathcal{O}\big(LK(d^3 + \log K)\big)$ which can be costly in high dimension. On the other hand, B1DGMSW can be approximated with a complexity of $\mathcal{O}\big(LK(\log K + d)\big)$, which is much cheaper that BGMSW.

C ADDITIONAL RESULTS

In the case of arbitrary 1D Gaussian distributions, in addition to Corollary 2, we can obtain the largest interval over which the geodesic can be extended, even in the direction $t < 0$. The proof can be found in Appendix D.3.

Proposition 9. *Let $\mu_0 = \mathcal{N}(m_0, \sigma_0^2)$ and $\mu_1 = \mathcal{N}(m_1, \sigma_1^2)$ two Gaussian distributions such that $\sigma_1 > \sigma_0$. Then, the geodesic $t \mapsto \mu_t$ is well defined on $]-\frac{\sigma_0}{\sigma_1-\sigma_0}, +\infty[$. By symmetry, if $\sigma_1 < \sigma_0$, the geodesic is well defined on $]-\infty, \frac{\sigma_0}{\sigma_0-\sigma_1}[$.*

In the setting of Proposition 9, the geodesic is given by $\mu_t = \mathcal{N}\big((1-t)m_0 + tm_1, ((1-t)\sigma_0 + t\sigma_1)^2\big)$. Thus, we observe that the geodesic property breaks at time $t = -\sigma_0/(\sigma_1 - \sigma_0)$, for which $(1-t)\sigma_0 + t\sigma_1 = 0$, *i.e.* the geodesic reaches a Dirac. This time is given by Proposition 9 for $\sigma_1 \neq \sigma_0$. Note that in the limit case $\sigma_0 = \sigma_1$, the geodesic is defined on $\mathbb{R}$ and is a translation.

We also notice that the curve $t \mapsto \mathcal{N}\big((1-t)m_0 + tm_1, ((1-t)\sigma_0 + t\sigma_1)^2\big)$ could be extended on $\mathbb{R}$, by allowing $(1-t)\sigma_0 + t\sigma_1 \leq 0$. This curve is geodesic on the right hand side and left hand side of $t_e = \text{sign}(\sigma_1 - \sigma_0)\frac{\sigma_0}{\sigma_1-\sigma_0}$. Thus, it is piecewise geodesic on both sides of the Dirac.

Projections on Geodesic Rays in the 1D Gaussian Case. Let $\mu_0 = \mathcal{N}(m_0, \sigma_0^2)$, $\mu_1 = \mathcal{N}(m_1, \sigma_1^2)$ with $\sigma_1 > \sigma_0$ and such that $\mathrm{W}_2^2(\mu_0, \mu_1) = 1$. As discussed in Section 3.1, the projection of $\nu = \mathcal{N}(m, \sigma^2)$ on the geodesic given by $\mu_t = \mathcal{N}\big((1-t)m_0 + tm_1, ((1-t)\sigma_0 + t\sigma_1)^2\big)$ is $P^\mu(\nu) = \mu_{-B^\mu(\nu)}$. However, if $-B^\mu(\nu) < 0$, this projection might be out of the original geodesic ray passing through μ_0 and μ_1. In particular, by Proposition 9, we know that this happens when $B^\mu(\nu) > \frac{\sigma_0}{\sigma_1-\sigma_0}$.

Here we work with a unit-speed geodesic ray, that is $\mathrm{W}_2^2(\mu_0, \mu_1) = (m_1 - m_0)^2 + (\sigma_1 - \sigma_0)^2 = 1$ and thus $m_1 - m_0, \sigma_1 - \sigma_0 \in [-1, 1]$. We have two limiting cases. The first one is $\sigma_0 = \sigma_1$ for which the geodesic ray is actually a line and can be extended to $\mathbb{R}$, which we recover here as $-\frac{\sigma_0}{\sigma_1-\sigma_0} \xrightarrow[\sigma_1 \to \sigma_0^+]{} -\infty$. The second one is $\sigma_1 = 1 + \sigma_0$ for which the ray can be extended to $[-\sigma_0, +\infty[$ and corresponds to a dilation. Moreover, in this case, since $\sigma_1 = 1 + \sigma_0$ and $m_1 = m_0$, we note that any 1D Gaussian will be projected on the geodesic since, for any $\nu = \mathcal{N}(m, \sigma^2)$,

$$B^\mu(\nu) = -(m - m_0)(m_1 - m_0) - (\sigma - \sigma_0)(\sigma_1 - \sigma_0) = -(\sigma - \sigma_0), \tag{59}$$

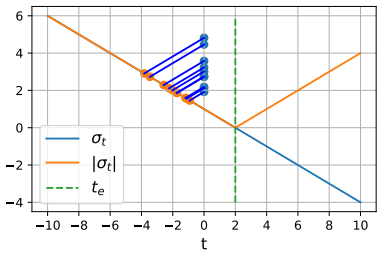
and thus the projection coordinate is $-B^\mu(\nu) = \sigma - \sigma_0 < -\sigma_0 \iff \sigma < 0$, which is not possible.

Likewise, for $\sigma_1 < \sigma_0$, the geodesic can be extended towards $-\infty$ and in the case of $m_0 = m_1$, the distributions are also necessarily well projected on the geodesic since $\sigma_1 = \sigma_0 - 1$ and $B^\mu(\nu) = \sigma - \sigma_0$. Thus, $-B^\mu(\nu) = \sigma_0 - \sigma > \sigma_0 \iff -\sigma > 0$.

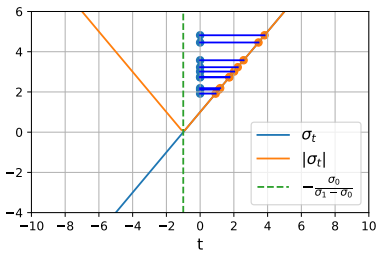
We illustrate these observations on Figure 4. We choose $\mu_0 = \mathcal{N}(0, 1)$ and $\mu_1 = \mathcal{N}(0, \sigma_1^2)$ with $\sigma_1 = \frac{1}{2}$ or $\mu_1 = \mathcal{N}(0, \sigma_1^2)$ with $\sigma_1 = \frac{3}{2}$. In the first case, this does not define a geodesic ray (even if the geodesic can be extended toward $-\infty$), but it does in the second case. We plot the projections of several Gaussian $\nu_i = \mathcal{N}(0, \sigma_i^2)$ (where the σ_i are plotted in the line $t = 0$). We see that every points are projected on the geodesic as expected by Proposition 9.

On Figure 5, we choose $\mu_0 = \mathcal{N}(0, 1)$ and $\mu_1 = \mathcal{N}(m_1, \sigma_1^2)$ with $m_1 = -\sqrt{1 - (\sigma_1 - \sigma_0)^2}$. We observe that some points projected onto the extended part of the geodesic are less consistent with the geometry.

We finally notice that when the first moments of μ_0 and μ_1 coincide, then the conditions to have geodesic rays are similar to conditions to have μ_0 smaller than μ_1 in the convex order, see *e.g.* (Müller, 2001, Theorem 4 and 6) for Gaussians or (Shu, 2020) in 1D for $\mu_0 \in \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$.

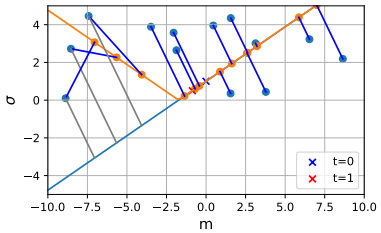


(a) $\sigma_0 = 1$, $\sigma_1 = \frac{1}{2}$

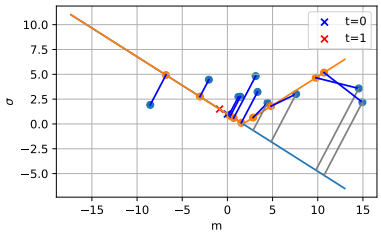


(b) $\sigma_0 = 1$, $\sigma_1 = \frac{3}{2}$

Figure 4: Projections of centered 1D Gaussian $\mathcal{N}(0, \sigma_i^2)$ (blue points) on the geodesic $\mathcal{N}(0, \sigma_t)$ starting at $\sigma_0 = 1$ and passing through $\sigma_1 = \frac{1}{2}$ (**Left**) and $\sigma_1 = \frac{3}{2}$ (**Right**).



(a) $\sigma_0 = 1$, $\sigma_1 = \frac{1}{2}$



(b) $\sigma_0 = 1$, $\sigma_1 = \frac{3}{2}$

Figure 5: Projections of 1D Gaussian $\mathcal{N}(m_i, \sigma_i^2)$ (blue points) on the geodesic $\mathcal{N}(m_t, \sigma_t)$ starting at $\sigma_0 = 1$, $m_0 = 0$ and passing through $\sigma_1 = \frac{1}{2}$ (**Left**) and $\sigma_1 = \frac{3}{2}$ (**Right**) with $m_1 = -\sqrt{1 - (\sigma_1 - \sigma_0)^2}$.

D PROOFS

D.1 Proofs of Section 2

D.1.1 Proof of Proposition 1

Proof of Proposition 1. Thanks to Brenier's theorem (Brenier, 1991) and since μ_0 is absolutely continuous with respect to the Lebesgue measure, there is a unique OT map T between μ_0 and μ_1, and T is the gradient of a convex function, *i.e.* $T = \nabla u$ with u convex.

First, let us suppose that the OT map T between μ_0 and μ_1 is the gradient of a 1-convex function u. Let $\mu : t \mapsto \mu_t := \big((1-t)\mathrm{Id} + tT\big)_{\#}\mu_0 = \big((1-t)\mathrm{Id} + t\nabla u\big)_{\#}\mu_0$. Then, on one hand, we have

$$\mathrm{W}_2^2(\mu_s, \mu_t) \leq (t-s)^2 \mathrm{W}_2^2(\mu_0, \mu_1). \tag{60}$$

Indeed, let $\gamma^* \in \Pi_o(\mu_0, \mu_1)$ be an optimal coupling. Then, necessarily, denoting $\pi^s(x, y) = (1-s)x + sy$, we have for any $s, t \in \mathbb{R}$, $(\pi^s, \pi^t)_{\#}\gamma^* \in \Pi(\mu_s, \mu_t)$. Therefore,

$$\begin{aligned} \mathrm{W}_2^2(\mu_s, \mu_t) &\leq \int \|x - y\|_2^2 \, \mathrm{d}(\pi^s, \pi^t)_{\#}\gamma^*(x, y) \\ &= \int \|(1-s)x + sy - (1-t)x - ty\|_2^2 \, \mathrm{d}\gamma^*(x, y) \\ &= (s-t)^2 \mathrm{W}_2^2(\mu_0, \mu_1). \end{aligned} \tag{61}$$

Then, let $\alpha \geq 1$ and $0 \leq s < t \leq \alpha$. By the triangular inequality and the previous inequality, we have

$$\begin{aligned} \mathrm{W}_2(\mu_0, \mu_\alpha) &\leq \mathrm{W}_2(\mu_0, \mu_s) + \mathrm{W}_2(\mu_s, \mu_t) + \mathrm{W}_2(\mu_t, \mu_\alpha) \\ &= (s + \alpha - t)\mathrm{W}_2(\mu_0, \mu_1) + \mathrm{W}_2(\mu_s, \mu_t). \end{aligned} \tag{62}$$

If $x \mapsto (1-\alpha)\frac{\|x\|_2^2}{2} + \alpha u(x)$ is convex (*i.e.* u is $\frac{\alpha-1}{\alpha}$-convex), then its gradient $x \mapsto (1-\alpha)x + \alpha\nabla u(x)$ is the Monge map between μ_0 and μ_α as $\mu_\alpha = \big((1-\alpha)\mathrm{Id} + \alpha\nabla u\big)_{\#}\mu_0$, and thus $\mathrm{W}_2^2(\mu_0,\mu_\alpha) = \alpha^2\mathrm{W}_2(\mu_0,\mu_1)$. Hence, we obtain

$$\mathrm{W}_2(\mu_0,\mu_\alpha) = \alpha\mathrm{W}_2(\mu_0,\mu_1) \leq (s+\alpha-t)\mathrm{W}_2(\mu_0,\mu_1) + \mathrm{W}_2(\mu_s,\mu_t) \iff (t-s)\mathrm{W}_2(\mu_0,\mu_1) \leq \mathrm{W}_2(\mu_s,\mu_t). \quad (63)$$

It allows to conclude that $\mathrm{W}_2(\mu_s,\mu_t) = |t-s|\mathrm{W}_2(\mu_0,\mu_1)$ for all $s,t \in [0,\alpha]$. In order to extend the result on $\mathbb{R}_+$, it has to be true for any $\alpha \geq 1$, which corresponds to u being 1-convex. Thus, we can conclude that $t \mapsto \mu_t$ is a geodesic ray.

For the inverse implication, suppose that $\mu_t = \big((1-t)\mathrm{Id} + tT\big)_{\#}\mu_0$ is a geodesic ray. Then, for all $s \geq 0$,

$$\begin{aligned}
\mathrm{W}_2^2(\mu_s,\mu_0) &= s^2\mathrm{W}_2^2(\mu_0,\mu_1) \\
&= \int \|s(x-\nabla u(x))\|_2^2 \,\mathrm{d}\mu_0(x) \\
&= \int \|x-(1-s)x - s\nabla u(x)\|_2^2 \,\mathrm{d}\mu_0(x) \\
&= \int \|x - T_s(x)\|_2^2 \,\mathrm{d}\mu_0(x),
\end{aligned} \quad (64)$$

where $(T_s)_{\#}\mu_0 = \mu_s$ with $T_s : x \mapsto (1-s)x + s\nabla u(x)$. By Brenier's theorem, since the OT map is unique and necessarily the gradient of a convex functions, we have that $T_s = \nabla u_s$ with $u_s : x \mapsto (1-s)\frac{\|x\|_2^2}{2} + su(x) = \frac{\|x\|_2^2}{2} + s\left(u(x) - \frac{\|x\|_2^2}{2}\right)$ convex. Thus, for all $s \geq 0$,

$$I_d + s(\nabla^2 u - I_d) \succeq 0 \iff \nabla^2 u - I_d \succeq -\frac{1}{s}I_d. \quad (65)$$

It is true for all $s \geq 0$, hence taking the limit $s \to \infty$, we obtain $\nabla^2 u - I \succeq 0$, *i.e.* u is 1-convex. □

D.1.2 Proof of Proposition 2

Proof of Proposition 2. By (Ambrosio et al., 2008, Equation 7.2.8), the quantile of μ_t is $F_t^{-1} = (1-t)F_0^{-1} + tF_1^{-1}$. Then, we know that F_t^{-1} is a quantile function if and only if it is non-decreasing and left-continuous. As a linear combination of left-continuous function, it is always left-continuous. It suffices then to find conditions under which F_t^{-1} is non-decreasing for all $t \geq 0$. Let $0 < m < m' < 1$, then

$$F_t^{-1}(m) - F_t^{-1}(m') = F_0^{-1}(m) - F_0^{-1}(m') + t\big(F_1^{-1}(m) - F_0^{-1}(m) - F_1^{-1}(m') + F_0^{-1}(m')\big), \quad (66)$$

and hence, $\forall t \geq 0$

$$\begin{aligned}
\forall m' > m,\ F_t^{-1}(m) - F_t^{-1}(m') \leq 0 &\iff \forall m' > m,\ F_1^{-1}(m) - F_0^{-1}(m) \leq F_1^{-1}(m') - F_0^{-1}(m') \\
&\iff F_1^{-1} - F_0^{-1} \text{ non-decreasing.}
\end{aligned} \quad (67)$$

□

D.1.3 Proof of Corollary 1

Proof of Corollary 1. We apply Proposition 2, and thus the resulting geodesic is a ray if and only if $F_1^{-1} - F_0^{-1}$ is non-decreasing, which is true if and only if for all $j > i$,

$$\begin{aligned}
F_1^{-1}\left(\frac{i}{n}\right) - F_0^{-1}\left(\frac{i}{n}\right) \leq F_1^{-1}\left(\frac{j}{n}\right) - F_0^{-1}\left(\frac{j}{n}\right) &\iff F_1^{-1}\left(\frac{i}{n}\right) - F_1^{-1}\left(\frac{j}{n}\right) \leq F_0^{-1}\left(\frac{i}{n}\right) - F_0^{-1}\left(\frac{j}{n}\right) \\
&\iff y_i - y_j \leq x_i - x_j.
\end{aligned} \quad (68)$$

□

D.1.4 Proof of Corollary 2

Proof of Corollary 2. Let $\mu_0 = \mathcal{N}(m_0, \sigma_0^2)$ and $\mu_1 = \mathcal{N}(m_1, \sigma_1^2)$ with $m_0, m_1 \in \mathbb{R}$ and $\sigma_0, \sigma_1 \in \mathbb{R}_+$. It is well known that for $p \in [0,1]$, $F_0^{-1}(p) = m_0 + \sigma_0 \phi^{-1}(p)$ where ϕ^{-1} denotes the quantile function of the standard Gaussian distribution $\mathcal{N}(0,1)$. In this case, for $0 < p < p' < 1$, we observe that

$$F_0^{-1}(p') - F_0^{-1}(p) = \sigma_0\big(\phi^{-1}(p') - \phi^{-1}(p)\big), \tag{69}$$

and therefore

$$\begin{aligned}(F_1^{-1} - F_0^{-1})(p') - (F_1^{-1} - F_0^{-1})(p) &= \big(F_1^{-1}(p') - F_1^{-1}(p)\big) - \big(F_0^{-1}(p') - F_0^{-1}(p)\big) \\ &= (\sigma_1 - \sigma_0)\big(\phi^{-1}(p') - \phi^{-1}(p)\big).\end{aligned} \tag{70}$$

Since ϕ^{-1} is non-decreasing, $F_1^{-1} - F_0^{-1}$ is non-decreasing if and only if $\sigma_0 \leq \sigma_1$. Thus, by Proposition 2, $\sigma_0 \leq \sigma_1$ is a sufficient condition to define a geodesic ray starting from μ_0 and passing through μ_1. □

D.1.5 Proof of Corollary 3

Proof of Corollary 3. Let $\mu_0 = \mathcal{N}(m_0, \Sigma_0)$ and $\mu_1 = \mathcal{N}(m_1, \Sigma_1)$ with $m_0, m_1 \in \mathbb{R}^d$ and Σ_0, Σ_1 symmetric positive definite matrices. The Monge map between μ_0 and μ_1 is (Peyré and Cuturi, 2019, Remark 2.31)

$$\forall x \in \mathbb{R}^d,\ \mathrm{T}(x) = A(x - m_0) + m_1, \tag{71}$$

where $A = \Sigma_0^{-\frac{1}{2}}\big(\Sigma_0^{\frac{1}{2}}\Sigma_1\Sigma_0^{\frac{1}{2}}\big)^{\frac{1}{2}}\Sigma_0^{-\frac{1}{2}}$. Let $u : x \mapsto \frac{1}{2}\langle Ax, x\rangle + \langle m_1 - Am_0, x\rangle = \frac{1}{2}\|A^{\frac{1}{2}}x\|_2^2 + \langle m_1 - Am_0, x\rangle$. Note that we have $\nabla u = \mathrm{T}$. Let us denote $g : x \mapsto u(x) - \frac{\|x\|_2^2}{2}$. Then, u is 1-convex if and only if $\nabla^2 g \succeq 0$ (with $\succeq$ the partial order, also called the Loewner order), *i.e.*

$$\begin{aligned}\nabla^2 g(x) = A - I_d \succeq 0 &\iff A \succeq I_d \\ &\iff \Sigma_0^{\frac{1}{2}} A \Sigma_0^{\frac{1}{2}} \succeq \Sigma_0^{\frac{1}{2}} I_d \Sigma_0^{\frac{1}{2}} \quad \textit{e.g.} \text{ by (Bhatia, 2013, Lemma V.1.5)} \\ &\iff \big(\Sigma_0^{\frac{1}{2}}\Sigma_1\Sigma_0^{\frac{1}{2}}\big)^{\frac{1}{2}} \succeq \Sigma_0.\end{aligned} \tag{72}$$

□

D.2 Proofs of Section 3

D.2.1 Proof of Proposition 3

First, we prove the following lemma relating the OT problem between the measures μ_t and ν, and the problem over couplings of (μ_0, μ_1, ν).

Lemma 1. *Let $(\mu_t)_{t\geq 0}$ be a geodesic ray. Let $t \geq 0$ and $\nu \in \mathcal{P}_2(\mathbb{R}^d)$. Then,*

$$\mathrm{W}_2^2(\mu_t, \nu) = \inf_{\gamma \in \Pi(\mu_t, \nu)} \int \|x - y\|_2^2 \,\mathrm{d}\gamma(x,y) = \inf_{\tilde{\gamma} \in \Gamma(\mu_0, \mu_1, \nu)} \int \|(1-t)x_0 + tx_1 - y\|_2^2 \,\mathrm{d}\tilde{\gamma}(x_0, x_1, y), \tag{73}$$

where $\Gamma(\mu_0, \mu_1, \nu) = \{\tilde{\gamma} \in \Pi(\mu_0, \mu_1, \nu),\ \pi^{1,2}_{\#}\tilde{\gamma} \in \Pi_o(\mu_0, \mu_1)\}$ and $\pi^{1,2} : (x_0, x_1, y) \mapsto (x_0, x_1)$ is the projection onto the first two coordinates.

Proof. On one hand, let $\tilde{\gamma} \in \Gamma(\mu_0, \mu_1, \nu)$. Then, $\gamma_t := \big((1-t)\pi^1 + t\pi^2, \pi^3\big)_{\#}\tilde{\gamma} \in \Pi(\mu_t, \nu)$, and we have

$$\begin{aligned}\mathrm{W}_2^2(\mu_t, \nu) &= \inf_{\gamma \in \Pi(\mu_t, \nu)} \int \|x - y\|_2^2 \,\mathrm{d}\gamma(x,y) \\ &\leq \int \|x - y\|_2^2 \,\mathrm{d}\big((1-t)\pi^1 + t\pi^2, \pi^3\big)_{\#}\tilde{\gamma}(x,y) \\ &= \int \|(1-t)x_0 + tx_1 - y\|_2^2 \,\mathrm{d}\tilde{\gamma}(x_0, x_1, y),\end{aligned} \tag{74}$$

and therefore, by taking the infimum over $\tilde{\gamma} \in \Gamma(\mu_0, \mu_1, \nu)$ on the right term,

$$\mathrm{W}_2^2(\mu_t, \nu) \leq \inf_{\tilde{\gamma} \in \Gamma(\mu_0, \mu_1, \nu)} \int \|(1-t)x_0 + tx_1 - y\|_2^2 \, \mathrm{d}\tilde{\gamma}(x_0, x_1, y). \tag{75}$$

On the other hand, let $\sigma_t \in \Pi(\mu_t, \nu)$. By disintegration of σ_t with respect to its first marginal, there exists a probability kernel K_t such that $\sigma_t = \mu_t \otimes K_t$, *i.e.*, such that for any test function h,

$$\int h(x, y) \, \mathrm{d}\sigma_t(x, y) = \iint h(x, y) \, K_t(x, \mathrm{d}y) \mathrm{d}\mu_t(x). \tag{76}$$

Let $\gamma^* \in \Pi_o(\mu_0, \mu_1)$ an optimal plan, and define π as the measure verifying for any test function h

$$\int h(x_0, x_1, y) \, \mathrm{d}\pi(x_0, x_1, y) = \iint h(x_0, x_1, y) \, K_t\big((1-t)x_0 + tx_1, \mathrm{d}y\big) \, \mathrm{d}\gamma^*(x_0, x_1). \tag{77}$$

We now verify that $\pi \in \Gamma(\mu_0, \mu_1, \nu)$. On one hand,

$$\begin{aligned}
\int h(y) \mathrm{d}\pi(x_0, x_1, y) &= \iint h(y) \, K_t\big((1-t)x_0 + tx_1, \mathrm{d}y\big) \, \mathrm{d}\gamma^*(x_0, x_1) \\
&= \iint h(y) \, K_t(x_t, \mathrm{d}y) \, \mathrm{d}\mu_t(x_t) \quad \text{since } \mu_t = \big((1-t)\pi^1 + t\pi^2\big)_{\#}\gamma^* \\
&= \int h(y) \, \mathrm{d}\sigma_t(x, y) \quad \text{by definition of the disintegration} \\
&= \int h(y) \, \mathrm{d}\nu(y),
\end{aligned} \tag{78}$$

and thus $\pi^3_{\#}\pi = \nu$. Moreover,

$$\begin{aligned}
\int h(x_0, x_1) \, \mathrm{d}\pi(x_0, x_1, y) &= \iint h(x_0, x_1) \, K_t\big((1-t)x_0 + tx_1, \mathrm{d}y\big) \, \mathrm{d}\gamma^*(x_0, x_1) \\
&= \int h(x_0, x_1) \, \mathrm{d}\gamma^*(x_0, x_1),
\end{aligned} \tag{79}$$

and thus $\pi^{1,2}_{\#}\pi = \gamma^* \in \Pi_o(\mu_0, \mu_1)$. Therefore, we can write

$$\begin{aligned}
&\inf_{\tilde{\gamma} \in \Gamma(\mu_0, \mu_1, \nu)} \int \|(1-t)x_0 + tx_1 - y\|_2^2 \, \mathrm{d}\tilde{\gamma}(x_0, x_1, y) \\
&\quad \leq \int \|(1-t)x_0 + tx_1 - y\|_2^2 \, \mathrm{d}\pi(x_0, x_1, y) \\
&\quad = \iint \|(1-t)x_0 + tx_1 - y\|_2^2 \, K_t\big((1-t)x_0 + tx_1, \mathrm{d}y\big) \, \mathrm{d}\gamma^*(x_0, x_1) \\
&\quad = \iint \|x_t - y\|_2^2 \, K_t(x_t, \mathrm{d}y) \mathrm{d}\mu_t(x_t) \\
&\quad = \int \|x - y\|_2^2 \, \mathrm{d}\sigma_t(x, y) \quad \text{by definition of the disintegration.}
\end{aligned} \tag{80}$$

Hence, taking the infimum on the right hand side, we deduce

$$\inf_{\tilde{\gamma} \in \Gamma(\mu_0, \mu_1, \nu)} \int \|(1-t)x_0 + tx_1 - y\|_2^2 \, \mathrm{d}\tilde{\gamma}(x_0, x_1, y) \leq \mathrm{W}_2^2(\mu_t, \nu), \tag{81}$$

and we can conclude. □

Let us now show that the infimum over $\Gamma(\mu_0, \mu_1, \nu)$ is attained and thus is a minimum. For this purpose, we need prove the following technical results on $\Gamma(\mu_0, \mu_, \nu)$.

Lemma 2. *Let $\mu_0, \mu_1, \nu \in \mathcal{P}_2(\mathbb{R}^d)$. Any sequence $(\tilde{\gamma}_n)_n$ in the space $\Gamma(\mu_0, \mu_1, \nu) = \{\tilde{\gamma} \in \Pi(\mu_0, \mu_1, \nu),\ \pi^{1,2}_{\#}\tilde{\gamma} \in \Pi_o(\mu_0, \mu_1)\}$ is tight.*

Proof. Following the proof of (Santambrogio, 2015, Theorem 1.7), the singleton $\{\mu_0\}$, $\{\mu_1\}$ and $\{\nu\}$ are tight and for all $\varepsilon > 0$, there exist compacts K_0, K_1, K_ν such that $\mu_0(\mathbb{R}^d \setminus K_0) < \frac{\varepsilon}{3}$, $\mu_1(\mathbb{R}^d \setminus K_1) < \frac{\varepsilon}{3}$ and $\nu(\mathbb{R}^d \setminus K_\nu) < \frac{\varepsilon}{3}$. Therefore,

$$\begin{aligned}\tilde{\gamma}_n^*\big((\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d) \setminus (K_0 \times K_1 \times K_\nu)\big) \\ \leq \tilde{\gamma}_n^*\big((\mathbb{R}^d \setminus K_0) \times \mathbb{R}^d \times \mathbb{R}^d\big) + \tilde{\gamma}_n^*\big(\mathbb{R}^d \times (\mathbb{R}^d \setminus K_1) \times \mathbb{R}^d\big) + \tilde{\gamma}_n^*\big(\mathbb{R}^d \times \mathbb{R}^d \times (\mathbb{R}^d \setminus K_\nu)\big) < \varepsilon. \quad (82)\end{aligned}$$

□

Lemma 3. *Let $\mu_0, \mu_1, \nu \in \Gamma(\mu_0, \mu_1, \nu)$. Then,*

$$\inf_{\tilde{\gamma} \in \Gamma(\mu_0, \mu_1, \nu)} \int \|(1-t)x_0 + tx_1 - y\|_2^2 \, \mathrm{d}\tilde{\gamma}(x_0, x_1, y) = \min_{\tilde{\gamma} \in \Gamma(\mu_0, \mu_1, \nu)} \int \|(1-t)x_0 + tx_1 - y\|_2^2 \, \mathrm{d}\tilde{\gamma}(x_0, x_1, y). \quad (83)$$

Proof. First, since any sequence in $\Gamma(\mu_0, \mu_1, \nu)$ is tight from Lemma 2, we have that $\Gamma(\mu_0, \mu_1, \nu)$ is a compact subset of $\Pi(\mu_0, \mu_1, \nu)$ (see *e.g.* (Santambrogio, 2015, Proof of Theorem 1.4)). Then, by (Santambrogio, 2015, Lemma 1.6), $J : \tilde{\gamma} \mapsto \int \|(1-t)x_0 + tx_1 - y\|_2^2 \, \mathrm{d}\tilde{\gamma}(x_0, x_1, y)$ is lower semi-continuous for the weak convergence of measures and by Weierstrass theorem (see *e.g.* (Santambrogio, 2015, Box 1.1)), the infimum is attained. □

Proof of Proposition 3. Let $t \geq 0$ and $\nu \in \mathcal{P}_2(\mathbb{R}^d)$. Using Lemma 1, we have that

$$\begin{aligned}\mathrm{W}_2^2(\mu_t, \nu) &= \inf_{\tilde{\gamma} \in \Gamma(\mu_0, \mu_1, \nu)} \int \|(1-t)x_0 + tx_1 - y\|_2^2 \, \mathrm{d}\tilde{\gamma}(x_0, x_1, y) \\ &= \inf_{\tilde{\gamma} \in \Gamma(\mu_0, \mu_1, \nu)} \int \|x_0 - y\|_2^2 \, \mathrm{d}\tilde{\gamma}(x_0, x_1, y) + t^2 \mathrm{W}_2^2(\mu_0, \mu_1) - 2t \int \langle x_1 - x_0, y - x_0 \rangle \, \mathrm{d}\tilde{\gamma}(x_0, x_1, y) \\ &= t^2 \mathrm{W}_2^2(\mu_0, \mu_1) \left(1 + \inf_{\tilde{\gamma} \in \Gamma(\mu_0, \mu_1, \nu)} \left[\frac{1}{t^2 \mathrm{W}_2^2(\mu_0, \mu_1)} \int \|x_0 - y\|_2^2 \, \mathrm{d}\tilde{\gamma}(x_0, x_1, y) \right.\right. \\ &\quad \left.\left. - \frac{2}{t\mathrm{W}_2^2(\mu_0, \mu_1)} \int \langle x_1 - x_0, y - x_0 \rangle \, \mathrm{d}\tilde{\gamma}(x_0, x_1, y)\right]\right).\end{aligned} \quad (84)$$

Thus, we have:

$$\begin{aligned}&\mathrm{W}_2(\mu_t, \nu) - t\mathrm{W}_2(\mu_0, \mu_1) \\ &= t\mathrm{W}_2(\mu_0, \mu_1) \\ &\quad \sqrt{1 + \frac{2}{t\mathrm{W}_2^2(\mu_0, \mu_1)} \inf_{\tilde{\gamma} \in \Gamma(\mu_0, \mu_1, \nu)} \left[\frac{1}{2t} \int \|x_0 - y\|_2^2 \, \mathrm{d}\tilde{\gamma}(x_0, x_1, y) - \int \langle x_1 - x_0, y - x_0 \rangle \, \mathrm{d}\tilde{\gamma}(x_0, x_1, y)\right]} \\ &\quad - t\mathrm{W}_2(\mu_0, \mu_1) \\ &\underset{t \to \infty}{=} t\mathrm{W}_2(\mu_0, \mu_1) \\ &\quad \left(1 + \frac{1}{t\mathrm{W}_2^2(\mu_0, \mu_1)} \inf_{\tilde{\gamma} \in \Gamma(\mu_0, \mu_1, \nu)} \left[\frac{1}{2t} \int \|x_0 - y\|_2^2 \, \mathrm{d}\tilde{\gamma}(x_0, x_1, y) - \int \langle x_1 - x_0, y - x_0 \rangle \, \mathrm{d}\tilde{\gamma}(x_0, x_1, y)\right] + o(t^{-1})\right) \\ &\quad - t\mathrm{W}_2(\mu_0, \mu_1) \\ &\underset{t \to \infty}{=} \frac{1}{\mathrm{W}_2(\mu_0, \mu_1)} \inf_{\tilde{\gamma} \in \Gamma(\mu_0, \mu_1, \nu)} \left(\frac{1}{2t} \int \|x_0 - y\|_2^2 \, \mathrm{d}\tilde{\gamma}(x_0, x_1, y) - \int \langle x_1 - x_0, y - x_0 \rangle \, \mathrm{d}\tilde{\gamma}(x_0, x_1, y)\right).\end{aligned} \quad (85)$$

To conclude, we need to show that we can pass to the limit. First, let $\tilde{\gamma}_t^*$ be defined as

$$\tilde{\gamma}_t^* \in \operatorname*{argmin}_{\tilde{\gamma} \in \Gamma(\mu_0, \mu_1, \nu)} \left(\frac{1}{2t} \int \|x_0 - y\|_2^2 \, \mathrm{d}\tilde{\gamma}(x_0, x_1, y) - \int \langle x_1 - x_0, y - x_0 \rangle \, \mathrm{d}\tilde{\gamma}(x_0, x_1, y)\right), \quad (86)$$

and let

$$\tilde{\gamma}^* \in \underset{\tilde{\gamma}\in\Gamma(\mu_0,\mu_1,\nu)}{\operatorname{argmin}} \; -\int \langle x_1 - x_0, y - x_0\rangle \; \mathrm{d}\tilde{\gamma}(x_0, x_1, y). \tag{87}$$

By definition of $\tilde{\gamma}_t^*$ and $\tilde{\gamma}^*$, we have the following inequality:

$$\begin{aligned}\frac{1}{2t}\int \|x_0 - y\|_2^2 \; \mathrm{d}\tilde{\gamma}_t^*(x_0, x_1, y) - \int \langle x_1 - x_0, y - x_0\rangle \; \mathrm{d}\tilde{\gamma}_t^*(x_0, x_1, y) \\ \leq \frac{1}{2t}\int \|x_0 - y\|_2^2 \; \mathrm{d}\tilde{\gamma}^*(x_0, x_1, y) - \int \langle x_1 - x_0, y - x_0\rangle \; \mathrm{d}\tilde{\gamma}^*(x_0, x_1, y).\end{aligned} \tag{88}$$

Let $(t_n)_n$ be a sequence such that $t_n \to \infty$. Any sequence in $\Gamma(\mu_0, \mu_1, \nu)$ is tight by Lemma 2. Hence, by Prokhorov's theorem, we can extract a subsequence $\tilde{\gamma}^*_{t_{\varphi(n)}}$ converging in law towards $\gamma_\infty \in \Gamma(\mu_0, \mu_1, \nu)$. Thus, passing to the limit in (88), we have

$$-\int \langle x_1 - x_0, y - x_0\rangle \; \mathrm{d}\gamma_\infty(x_0, x_1, y) \leq -\int \langle x_1 - x_0, y - x_0\rangle \; \mathrm{d}\tilde{\gamma}^*(x_0, x_1, y). \tag{89}$$

But by definition, $\tilde{\gamma}^*$ is optimal in (87), therefore (89) is an equality. We can conclude that

$$B^\mu(\nu) = \inf_{\tilde{\gamma}\in\Gamma(\mu_0,\mu_1,\nu)} \; -\frac{1}{\mathrm{W}_2(\mu_0, \mu_1)}\int \langle x_1 - x_0, y - x_0\rangle \; \mathrm{d}\tilde{\gamma}(x_0, x_1, y). \tag{90}$$

□

D.2.2 Proof of Corollary 4

Proof of Corollary 4. Let μ_0 absolutely continuous *w.r.t.* the Lebesgue measure, and T the gradient of a 1-convex function such that $\mu_1 = T_\# \mu_0$. Let us show that in this case $\Gamma(\mu_0, \mu_1, \nu) = \{(\pi^1, T\circ\pi^1, \pi^2)_\#\gamma, \; \gamma \in \Pi(\mu_0, \nu)\}$, where we recall that $\Gamma(\mu_0, \mu_1, \nu) = \{\tilde{\gamma} \in \Pi(\mu_0, \mu_1, \nu), \; \pi^{1,2}_\#\tilde{\gamma} \in \Pi_o(\mu_0, \mu_1)\}$.

On one hand, let $\gamma \in \Pi(\mu_0, \nu)$ and $\tilde{\gamma} = (\pi^1, T\circ\pi^1, \pi^2)_\#\gamma$. Then, we verify easily that the marginals are satisfied, *i.e.*, $\pi^1_\#\tilde{\gamma} = \mu_0$, $\pi^2_\#\tilde{\gamma} = \mu_1$ and $\pi^3_\#\tilde{\gamma} = \nu$. Moreover,

$$\int h(x_0, x_1) \; \mathrm{d}\tilde{\gamma}(x_0, x_1, y) = \int h\big(x, T(x)\big) \; \mathrm{d}\gamma(x_0, y) = \int h\big(x, T(x)\big) \; \mathrm{d}\mu_0(x), \tag{91}$$

and hence $\pi^{1,2}_\#\tilde{\gamma} = (\mathrm{Id}, T)_\#\mu_0 \in \Pi_o(\mu_0, \mu_1)$. Thus, $\{(\pi^1, T\circ\pi^1, \pi^2)_\#\gamma, \; \gamma \in \Pi(\mu_0, \nu)\} \subset \Gamma(\mu_0, \mu_1, \nu)$.

On the other hand, let $\tilde{\gamma} \in \Gamma(\mu_0, \mu_1, \nu)$. Thus, we know that $\pi^{1,2}_\#\tilde{\gamma} = (\mathrm{Id}, T)_\#\mu_0$. Additionally, by the disintegration theorem, there exists a probability kernel K such that $\tilde{\gamma} = \pi^{1,2}_\#\tilde{\gamma} \otimes K = (\mathrm{Id}, T)_\#\mu_0 \otimes K$, *i.e.*

$$\begin{aligned}\int h(x_0, x_1, y) \; \mathrm{d}\tilde{\gamma}(x_0, x_1, y) &= \iint h(x_0, x_1, y) \; K\big((x_0, x_1), \mathrm{d}y\big) \; \mathrm{d}(\pi^{1,2}_\#\tilde{\gamma})(x_0, x_1) \\ &= \iint h(x_0, T(x_0), y) \; K\big((x_0, T(x_0)), \mathrm{d}y\big) \; \mathrm{d}\mu_0(x_0).\end{aligned} \tag{92}$$

Denoting $\tilde{K}(x_0, \mathrm{d}y) = K\big((x_0, T(x_0)), \mathrm{d}y\big)$ and defining $\gamma = \mu_0 \otimes \tilde{K}$, we obtain

$$\begin{aligned}\int h(x_0, x_1, y) \; \mathrm{d}\tilde{\gamma}(x_0, x_1, y) &= \iint h(x_0, T(x_0), y) \; \tilde{K}(x_0, \mathrm{d}y)\mathrm{d}\mu_0(x_0) \\ &= \int h(x_0, T(x_0), y) \; \mathrm{d}\gamma(x_0, y) \\ &= \int h(x_0, x_1, y) \; \mathrm{d}(\pi^1, T\circ\pi^1, \pi^2)_\#\gamma(x_0, x_1, y).\end{aligned} \tag{93}$$

Thus, $\tilde{\gamma} = (\pi^1, T \circ \pi^1, \pi^2)_{\#}\gamma$. Moreover, we can verify that $\gamma \in \Pi(\mu_0, \nu)$ as $\pi^1_{\#}\gamma = \mu_0$ by definition of γ and

$$\begin{aligned}\int h(y)\ \mathrm{d}\gamma(x,y) &= \iint h(y)\ \tilde{K}(x,\mathrm{d}y)\mathrm{d}\mu_0(x)\\ &= \iint h(y)\ K\big((x,T(x)),\mathrm{d}y\big)\mathrm{d}\mu_0(x)\\ &= \int h(y)\ \mathrm{d}\tilde{\gamma}(x_0,x_1,y)\\ &= \int h(y)\ \mathrm{d}\nu(y).\end{aligned} \tag{94}$$

Therefore, we can conclude that the two sets are equals, and rewrite the Busemann function from Proposition 3 in this case as

$$B^{\mu}(\nu) = \inf_{\gamma\in\Pi(\mu_0,\nu)} -\frac{1}{\mathrm{W}_2(\mu_0,\mu_1)}\int \langle T(x_0)-x_0, y-x_0\rangle\ \mathrm{d}\gamma(x_0,y). \tag{95}$$

□

D.2.3 Proof of Corollary 5

Proof of Corollary 5. Let $\mu_0 = \delta_{x_0}$ with $x_0 \in \mathbb{R}^d$. Recall that $\Gamma(\mu_0,\mu_1,\nu) = \{\tilde{\gamma} \in \Pi(\mu_0,\mu_1,\nu),\ \pi^{1,2}_{\#}\tilde{\gamma} \in \Pi_o(\mu_0,\mu_1)\} = \{\tilde{\gamma} \in \Pi(\mu_0,\mu_1,\nu),\ \pi^{1,2}_{\#}\tilde{\gamma} = \mu_0\otimes\mu_1\}$ since the optimal coupling between $\mu_0 = \delta_{x_0}$ and μ_1 is $\mu_0\otimes\mu_1$. Let us show that $\Gamma(\mu_0,\mu_1,\nu) = \{\mu_0\otimes\gamma,\ \gamma\in\Pi(\mu_1,\nu)\}$ in this case.

On one hand, let $\gamma\in\Pi(\mu_1,\nu)$ and define $\tilde{\gamma} = \mu_0\otimes\gamma$. Then, trivially, we have that $\tilde{\gamma}\in\Pi(\mu_0,\mu_1,\nu)$. Moreover, let us verify that $\pi^{1,2}_{\#}\tilde{\gamma} = \mu_0\otimes\mu_1$. For any continuous bounded function h,

$$\begin{aligned}\int h(x_0,x_1)\ \mathrm{d}(\pi^{1,2}_{\#}\tilde{\gamma})(x_0,x_1) &= \int h(x_0,x_1)\ \mathrm{d}\tilde{\gamma}(x_0,x_1,y)\\ &= \iint h(x_0,x_1)\ \mathrm{d}\gamma(x_1,y)\mathrm{d}\mu_0(x_0)\\ &= \iint h(x_0,x_1)\ \mathrm{d}\mu_1(x_1)\mathrm{d}\mu_0(x_0).\end{aligned} \tag{96}$$

Thus, $\{\mu_0\otimes\gamma,\ \gamma\in\Pi(\mu_1,\nu)\}\subset\Gamma(\mu_0,\mu_1,\nu)$.

On the other hand, let $\tilde{\gamma}\in\Gamma(\mu_0,\mu_1,\nu)$. Since $\pi^{1,2}_{\#}\tilde{\gamma} = \mu_0\otimes\mu_1$, we can disintegrate $\tilde{\gamma}$ as $\tilde{\gamma} = (\mu_0\otimes\mu_1)\otimes K$, *i.e.* for any h,

$$\int h(x_0,x_1,y)\ \mathrm{d}\tilde{\gamma}(x_0,x_1,y) = \iint h(x_0,x_1,y)\ K\big((x_0,x_1),\mathrm{d}y\big)\ \mathrm{d}\mu_1(x_1)\mathrm{d}\mu_0(x_0). \tag{97}$$

Let us define the distribution γ_{x_0} satisfying for μ_0-a.e. x_0,

$$\int h(x_0,x_1,y)\ \mathrm{d}\gamma_{x_0}(x_1,y) = \iint h(x_0,x_1,y)\ K\big((x_0,x_1),\mathrm{d}y\big)\mathrm{d}\mu_1(x_1). \tag{98}$$

First, we can verify that for μ_0-a.e. x_0, $\pi^1_{\#}\gamma_{x_0} = \mu_1$ as

$$\int h(x_1)\mathrm{d}\gamma_{x_0}(x_1,y) = \iint h(x_1)K\big((x_0,x_1),\mathrm{d}y\big)\mathrm{d}\mu_1(x_1) = \int h(x_1)\mathrm{d}\mu_1(x_1). \tag{99}$$

Moreover, we can also disintegrate $\tilde{\gamma}$ *w.r.t.* μ_0 as $\tilde{\gamma} = \mu_0\otimes\tilde{K}$. By uniqueness of the disintegration, we have for μ_0-a.e. x_0,

$$\iint h(x_1,y)K\big((x_0,x_1),\mathrm{d}y\big)\mathrm{d}\mu_1(x_1) = \int h(x_1,y)\tilde{K}\big(x_0,(\mathrm{d}x_1,\mathrm{d}y)\big), \tag{100}$$

i.e. $\gamma_{x_0} = \tilde{K}(x_0,\cdot)$. Integrating *w.r.t* μ_0, we get that the left hand side is equal to $\int h\mathrm{d}\tilde{\gamma}$ by (97). But we also get

$$\begin{aligned}\iiint h(x_1,y)K\big((x_0,x_1),\mathrm{d}y\big)\mathrm{d}\mu_1(x_1)\mathrm{d}\mu_0(x_0) &= \iint h(x_1,y)\left(\int K\big((x_0,x_1),\mathrm{d}y\big)\mathrm{d}\mu_0(x_0)\right)\ \mathrm{d}\mu_1(x_1)\\ &= \iint h(x_1,y)\ \bar{K}(x_1,\mathrm{d}y)\mathrm{d}\mu_1(x_1),\end{aligned} \tag{101}$$

where $\bar{K}(x_1, \mathrm{d}y) := \int K\big((x_0, x_1), \mathrm{d}y\big)\mathrm{d}\mu_0(x_0)$. Thus, we deduce that $\pi^{2,3}_{\#}\tilde{\gamma} = \mu_1 \otimes \bar{K}$. Integrating the right hand side of (100), we get for any bounded measurable function h

$$\int h(x_1, y)\ \mathrm{d}\tilde{\gamma}(x_0, x_1, y) = \int h(x_1, y) \int \mathrm{d}\gamma_{x_0}(x_1, y)\mathrm{d}\mu_0(x_0) = \iint h(x_1, y)\ \bar{K}(x_1, \mathrm{d}y)\mathrm{d}\mu_1(x_1). \tag{102}$$

It implies that $\gamma_{x_0} = \mu_1 \otimes \bar{K}$ for μ_0-a.e. x_0 and thus that K does not depend on x_0. Finally, we can conclude that $\tilde{\gamma} = \mu_0 \otimes \gamma$ with $\gamma = \mu_1 \otimes \bar{K} \in \Pi(\mu_1, \nu)$. □

D.2.4 Proof of Proposition 4

Proof of Proposition 4. Let $(\mu_t)_{t\geq 0}$ be a geodesic ray. Recall that $\mathrm{W}_2^2(\mu_0, \mu_1) = \int_0^1 \left|F_0^{-1}(u) - F_1^{-1}(u)\right|^2 \mathrm{d}u = \|F_0^{-1} - F_1^{-1}\|^2_{L^2([0,1])}$. Moreover, the quantile functions of any measure on the geodesic ray is of the form,

$$\forall t \geq 0,\ F_t^{-1} = (1-t)F_0^{-1} + tF_1^{-1}. \tag{103}$$

Thus, we have, for any $\nu \in \mathcal{P}_2(\mathbb{R})$, $t \geq 0$,

$$\begin{aligned}
\mathrm{W}_2(\nu, \mu_t) - t\mathrm{W}_2(\mu_1, \mu_0) &= \|F_\nu^{-1} - F_t^{-1}\|_{L^2([0,1])} - t\mathrm{W}_2(\mu_1, \mu_0) \\
&= \|F_\nu^{-1} - (1-t)F_0^{-1} - tF_1^{-1}\|_{L^2([0,1])} - t\mathrm{W}_2(\mu_1, \mu_0) \\
&= \|F_\nu^{-1} - F_0^{-1} - t(F_1^{-1} - F_0^{-1})\|_{L^2([0,1])} - t\mathrm{W}_2(\mu_1, \mu_0) \\
&= \sqrt{\|F_\nu^{-1} - F_0^{-1}\|^2_{L^2([0,1])} - 2t\langle F_\nu^{-1} - F_0^{-1}, F_1^{-1} - F_0^{-1}\rangle_{L^2([0,1])} + t^2\mathrm{W}_2^2(\mu_1, \mu_0)} \\
&\quad - t\mathrm{W}_2(\mu_1, \mu_0) \\
&= t\mathrm{W}_2(\mu_1, \mu_0)\sqrt{1 - \frac{2}{t\mathrm{W}_2^2(\mu_1, \mu_0)}\langle F_\nu^{-1} - F_0^{-1}, F_1^{-1} - F_0^{-1}\rangle_{L^2([0,1])} + o\left(\frac{1}{t}\right)} \\
&\quad - t\mathrm{W}_2(\mu_1, \mu_0) \\
&\underset{t\to\infty}{=} t\mathrm{W}_2(\mu_1, \mu_0)\left(1 - \frac{1}{t\mathrm{W}_2^2(\mu_1, \mu_0)}\langle F_\nu^{-1} - F_0^{-1}, F_1^{-1} - F_0^{-1}\rangle_{L^2([0,1])} + o\left(\frac{1}{t}\right)\right) \\
&\quad - t\mathrm{W}_2(\mu_1, \mu_0) \\
&\xrightarrow[t\to\infty]{} - \left\langle F_\nu^{-1} - F_0^{-1}, \frac{F_1^{-1} - F_0^{-1}}{\|F_1^{-1} - F_0^{-1}\|_{L^2([0,1])}} \right\rangle_{L^2([0,1])}.
\end{aligned} \tag{104}$$

Thus, we can conclude that

$$B^\mu(\nu) = - \left\langle F_\nu^{-1} - F_0^{-1}, \frac{F_1^{-1} - F_0^{-1}}{\|F_1^{-1} - F_0^{-1}\|_{L^2([0,1])}} \right\rangle_{L^2([0,1])}. \tag{105}$$

For a unit-speed geodesic ray $(\mu_t)_{t\geq 0}$, we have $\|F_1^{-1} - F_0^{-1}\|_{L^2([0,1])} = 1$, and we then recover (17). □

D.2.5 Proof of Corollary 6

Proof of Corollary 6. Recall that for any $\eta = \mathcal{N}(m, \sigma^2)$, for all $u \in [0,1]$, $F_\eta^{-1}(u) = m + \sigma\phi^{-1}(u)$ where ϕ^{-1} is the quantile function of $\mathcal{N}(0,1)$, therefore satisfying $\int_0^1 \phi^{-1}(u)\ \mathrm{d}u = 0$ and $\int_0^1 \phi^{-1}(u)^2\ \mathrm{d}u = 1$.

Thus, let $\mu_0 = \mathcal{N}(m_0, \sigma_0^2)$, $\mu_1 = \mathcal{N}(m_1, \sigma_1^2)$, $\nu = \mathcal{N}(m, \sigma^2)$ with $m_0, m_1, m \in \mathbb{R}$, $\sigma_0, \sigma_1, \sigma \in \mathbb{R}_+^*$ and $\sigma_1 \geq \sigma_0$,

$\mathrm{W}_2^2(\mu_0,\mu_1)=1$. Then, by Proposition 4, we have

$$\begin{aligned}
B^\mu(\nu) &= -\langle F_1^{-1}-F_0^{-1}, F_\nu^{-1}-F_0^{-1}\rangle_{L^2([0,1])}\\
&= -\langle (m_1-m_0)+(\sigma_1-\sigma_0)\phi^{-1}, (m-m_0)+(\sigma-\sigma_0)\phi^{-1}\rangle_{L^2([0,1])}\\
&= -(m_1-m_0)(m-m_0)-(m_1-m_0)(\sigma-\sigma_0)\int_0^1\phi^{-1}(u)\mathrm{d}u\\
&\quad -(m-m_0)(\sigma_1-\sigma_0)\int_0^1\phi^{-1}(u)\mathrm{d}u-(\sigma_1-\sigma_0)(\sigma-\sigma_0)\int_0^1\phi^{-1}(u)^2\,\mathrm{d}u\\
&= -(m_1-m_0)(m-m_0)-(\sigma_1-\sigma_0)(\sigma-\sigma_0)\\
&= -\left\langle \begin{pmatrix} m_1-m_0\\ \sigma_1-\sigma_0\end{pmatrix}, \begin{pmatrix} m-m_0\\ \sigma-\sigma_0\end{pmatrix}\right\rangle.
\end{aligned} \tag{106}$$

More generally, if $\mathrm{W}_2(\mu_0,\mu_1)>0$, we have

$$B^\mu(\nu) = -\frac{(m_1-m_0)(m-m_0)-(\sigma_1-\sigma_0)(\sigma-\sigma_0)}{\sqrt{(m_1-m_0)^2+(\sigma_1-\sigma_0)^2}}. \tag{107}$$

□

D.2.6 Proof of Proposition 5

Proof of Proposition 5. We will use here that for any geodesic ray γ, $\lim_{t\to\infty}\frac{d(x,\gamma(t))+t}{2t}=1$ (cf (Bridson and Haefliger, 2013, II. 8.24). Then we know that

$$\lim_{t\to\infty}\frac{d(x,\gamma(t))^2-t^2}{2t}=\lim_{t\to\infty}\big(d(x,\gamma(t))-t\big), \tag{108}$$

since

$$\frac{d(x,\gamma(t))^2-t^2}{2t}=\frac{(d(x,\gamma(t))-t)(d(x,\gamma(t))+t)}{2t}=\big(d(x,\gamma(t))-t\big)\frac{d(x,\gamma(t))+t}{2t}\xrightarrow[t\to\infty]{}B^\gamma(x). \tag{109}$$

In our case, we have for any $t\geq 0$, $\mu_t=\mathcal{N}(m_t,\Sigma_t)$ where

$$\begin{cases} m_t=(1-t)m_0+tm_1\\ \Sigma_t=\big((1-t)I_d+tA\big)\Sigma_0\big((1-t)I_d+tA\big),\end{cases} \tag{110}$$

with $A=\Sigma_0^{-\frac12}(\Sigma_0^{\frac12}\Sigma_1\Sigma_0^{\frac12})^{\frac12}\Sigma_0^{-\frac12}$ (see *e.g.* (Altschuler et al., 2021, Appendix A.1)). Then, using $A\Sigma_0A=\Sigma_1$, we have for any $t\geq 0$,

$$\frac{\|m_t-m\|_2^2}{2t}=\frac{t}{2}\|m_1-m_0\|_2^2+\langle m_1-m_0,m_0-m\rangle+O\left(\frac1t\right), \tag{111}$$

$$\frac{\mathrm{Tr}(\Sigma_t)}{2t}=\frac{t}{2}\mathrm{Tr}\big(\Sigma_0-2\Sigma_0A+\Sigma_1\big)+\mathrm{Tr}\big(\Sigma_0A-\Sigma_0\big)+O\left(\frac1t\right) \tag{112}$$

$$\frac{\mathrm{Tr}\big((\Sigma^{\frac12}\Sigma_t\Sigma^{\frac12})^{\frac12}\big)}{2t}=\frac12\mathrm{Tr}\left(\left(\Sigma^{\frac12}(\Sigma_0-\Sigma_0A-A\Sigma_0+\Sigma_1)\Sigma^{\frac12}+O\left(\frac1t\right)\right)^{\frac12}\right). \tag{113}$$

Additionally, by hypothesis,

$$\mathrm{W}_2^2(\mu_0,\mu_1)=\|m_1-m_0\|_2^2+\mathrm{Tr}(\Sigma_0+\Sigma_1-2(\Sigma_0^{\frac12}\Sigma_1\Sigma_0^{\frac12})^{\frac12})=1, \tag{114}$$

and since

$$\Sigma_0A=\Sigma_0^{\frac12}(\Sigma_0^{\frac12}\Sigma_1\Sigma_0^{\frac12})^{\frac12}\Sigma_0^{-\frac12}, \tag{115}$$

we get

$$\mathrm{Tr}(\Sigma_0 A) = \mathrm{Tr}\big((\Sigma_0^{\frac{1}{2}}\Sigma_1\Sigma_0^{\frac{1}{2}})^{\frac{1}{2}}\big). \tag{116}$$

Therefore, we obtain

$$\begin{aligned}
&\frac{\mathrm{W}_2^2(\nu,\mu_t)-t^2}{2t}\\
&=\frac{\|m_t-m\|_2^2+\mathrm{Tr}\left(\Sigma_t+\Sigma-2(\Sigma^{\frac{1}{2}}\Sigma_t\Sigma^{\frac{1}{2}})^{\frac{1}{2}}\right)-t^2}{2t}\\
&=\frac{t}{2}\left(\|m_1-m_0\|_2^2+\mathrm{Tr}(\Sigma_0+\Sigma_1-2\Sigma_0 A)\right)+\langle m_1-m_0,m_0-m\rangle+\mathrm{Tr}\big(\Sigma_0 A-\Sigma_0\big)\\
&\quad-\mathrm{Tr}\left(\left(\Sigma^{\frac{1}{2}}(\Sigma_0-\Sigma_0 A-A\Sigma_0+\Sigma_1)\Sigma^{\frac{1}{2}}+O\left(\frac{1}{t}\right)\right)^{\frac{1}{2}}\right)-\frac{t}{2}+O\left(\frac{1}{t}\right)\\
&=\frac{t}{2}\mathrm{W}_2^2(\mu_0,\mu_1)+\langle m_1-m_0,m_0-m\rangle+\mathrm{Tr}\big(\Sigma_0 A-\Sigma_0\big)\\
&\quad-\mathrm{Tr}\left(\left(\Sigma^{\frac{1}{2}}(\Sigma_0-\Sigma_0 A-A\Sigma_0+\Sigma_1)\Sigma^{\frac{1}{2}}+O\left(\frac{1}{t}\right)\right)^{\frac{1}{2}}\right)-\frac{t}{2}+O\left(\frac{1}{t}\right)\\
&=\langle m_1-m_0,m_0-m\rangle+\mathrm{Tr}\big(\Sigma_0 A-\Sigma_0\big)\\
&\quad-\mathrm{Tr}\left(\left(\Sigma^{\frac{1}{2}}(\Sigma_0-\Sigma_0 A-A\Sigma_0+\Sigma_1)\Sigma^{\frac{1}{2}}+O\left(\frac{1}{t}\right)\right)^{\frac{1}{2}}\right)+O\left(\frac{1}{t}\right)\\
&\xrightarrow[t\to\infty]{}\langle m_1-m_0,m_0-m\rangle+\mathrm{Tr}\big(\Sigma_0(A-I_d)\big)-\mathrm{Tr}\big((\Sigma^{\frac{1}{2}}(\Sigma_0-\Sigma_0 A-A\Sigma_0+\Sigma_1)\Sigma^{\frac{1}{2}})^{\frac{1}{2}}\big).
\end{aligned} \tag{117}$$

□

D.3 Proofs of Appendix

D.3.1 Proof of Proposition 7

Proof of Proposition 7. Let $\mathbb{P},\mathbb{Q}\in\mathcal{P}_2\big(\mathrm{BW}(\mathbb{R}^d)\big)$. Recall that for any $\theta\in S^{d-1}$, $\mathbb{P}^\theta=\varphi^\theta_\#\mathbb{P}$, and for any $\mu=\mathcal{N}(m,\Sigma)\sim\mathbb{P}$, $\varphi^\theta(\mu)=P^\theta_\#\mu=\mathcal{N}(\langle m,\theta\rangle,\theta^T\Sigma\theta)=\mathcal{N}(m_\theta,\sigma_\theta^2)$ where we note $m_{\mu_\theta}=\langle m,\theta\rangle$ and $\sigma^2_{\mu_\theta}=\theta^T\Sigma\theta$.

Then, for any $\theta\in S^{d-1}$ and $\eta_1=\mathcal{N}(m_1,\sigma_1^2)$, let $\Gamma\in\Pi_o(\mathbb{P},\mathbb{Q})$, $\Gamma_\theta=(\varphi^\theta,\varphi^\theta)_\#\Gamma\in\Pi(\mathbb{P}^\theta,\mathbb{Q}^\theta)$ and $\gamma^\eta_\theta=(B^\eta,B^\eta)_\#\Gamma_\theta\in\Pi(B^\eta_\#\mathbb{P}^\theta,B^\eta_\#\mathbb{Q}^\theta)$. Recall that $\mathrm{W}_2^2(\eta_0,\eta_1)=(m_1-m_0)^2+(\sigma_1-\sigma_0)^2=1$.

Then, we have

$$\begin{aligned}
\mathrm{W}_2^2(B^\eta_\#\mathbb{P}^\theta,B^\eta_\#\mathbb{Q}^\theta)&=\inf_{\gamma\in\Pi(B^\eta_\#\mathbb{P}^\theta,B^\eta_\#\mathbb{Q}^\theta)}\int|x-y|^2\,\mathrm{d}\gamma(x,y)\\
&\le\int|x-y|\,\mathrm{d}\gamma^\eta_\theta(x,y)\\
&=\int|B^\eta(\mu_\theta)-B^\eta(\nu_\theta)|^2\,\mathrm{d}\Gamma_\theta(\mu_\theta,\nu_\theta)\\
&=\int\left|\left\langle\begin{pmatrix}m_1-m_0\\ \sigma_1-\sigma_0\end{pmatrix},\begin{pmatrix}m_{\mu_\theta}-m_{\nu_\theta}\\ \sigma_{\mu_\theta}-\sigma_{\nu_\theta}\end{pmatrix}\right\rangle\right|^2\,\mathrm{d}\Gamma_\theta(\mu_\theta,\nu_\theta)\\
&\le\mathrm{W}_2^2(\eta_0,\eta_1)\cdot\int\mathrm{W}_2^2(\mu_\theta,\nu_\theta)\,\mathrm{d}\Gamma(\mu_\theta,\nu_\theta)\quad\text{by Cauchy-Schwartz}\\
&=\int\mathrm{W}_2^2\big(\varphi^\theta(\mu),\varphi^\theta(\nu)\big)\,\mathrm{d}\Gamma(\mu,\nu)\\
&=\int\mathrm{W}_2^2(P^\theta_\#\mu,P^\theta_\#\nu)\,\mathrm{d}\Gamma(\mu,\nu)\\
&\le\int\mathrm{W}_2^2(\mu,\nu)\,\mathrm{d}\Gamma(\mu,\nu)\quad\text{since }P^\theta\text{ 1-Lipschitz}\\
&=\mathrm{W}_{\mathrm{BW}}^2(\mathbb{P},\mathbb{Q})\quad\text{since }\Gamma\in\Pi_o(\mathbb{P},\mathbb{Q}).
\end{aligned} \tag{118}$$

By integrating *w.r.t.* λ, we can conclude that

$$\text{B1DGMSW}^2(\mathbb{P},\mathbb{Q}) \leq \text{W}_{\text{BW}}^2(\mathbb{P},\mathbb{Q}). \tag{119}$$

□

D.3.2 Proof of Proposition 8

Proof of Proposition 8. The symmetric, positivity, and triangular inequality are clear by classical arguments.

For positive definiteness, let $\mathbb{P},\mathbb{Q} \in \bigcup_{K\geq 0} \text{GMM}_d(K)$ such that $\text{B1DGMSW}(\mathbb{P},\mathbb{Q}) = 0$. Then, there exists $K \geq 0$ such that $\mathbb{P} = \sum_{k=1}^K \alpha_k \delta_{\mu_k}$ and $\mathbb{Q} = \sum_{k=1}^K \beta_k \delta_{\nu_k}$.

First, note that we can rewrite B1DGMSW as

$$\text{B1DGMSW}^2(\mathbb{P},\mathbb{Q}) = \int_{S^{d-1}} \text{SW}_2^2(\Xi_\# \mathbb{P}_\theta, \Xi_\# \mathbb{Q}_\theta))\, \mathrm{d}\lambda(\theta), \tag{120}$$

with $\Xi(\mathcal{N}(m,\sigma^2)) = (m,\sigma)$, and $\mathbb{P}^\theta = \varphi_\#^\theta \mathbb{P}$.

Thus, $\text{B1DGMSW}(\mathbb{P},\mathbb{Q}) = 0$ implies that for λ-almost all $\theta \in S^{d-1}$, $\text{SW}_2^2(\mathbb{P}^\theta,\mathbb{Q}^\theta) = 0$. However, SW is a Pullback-Euclidean Sliced-Wasserstein distance. Thus, by (Bonet et al., 2025b, Proposition 26), it is a distance and thus $\mathbb{P}_\theta = \mathbb{Q}_\theta$. Moreover, this also implies that $\text{W}_2^2(\sum_{k=1}^K \alpha_k \delta_{P^\theta_\# \mu_k}, \sum_{k=1}^K \beta_k \delta_{P^\theta_\# \nu_k}) = 0$, and thus by integrating *w.r.t* $\theta \in S^{d-1}$, the Sliced-Wasserstein distance between the mixtures seen in $\mathcal{P}_2(\mathbb{R}^d)$ is equal to 0. Thus, as SW is a distance, we can conclude that $\mathbb{P} = \mathbb{Q}$. □

D.3.3 Proof of Proposition 9

Proof of Proposition 9. Let $\mu_0 = \mathcal{N}(m_0,\sigma_0^2)$ and $\mu_1 = \mathcal{N}(m_1,\sigma_1^2)$ such that $\sigma_1 > \sigma_0$. Extending the geodesic between μ_0 and μ_1 on $]1-\alpha,0]$ for $\alpha > 1$ is equivalent to extending the geodesic between μ_1 and μ_0 on $[0,\alpha[$. Thus, we first find a condition to extend the geodesic between μ_1 and μ_0.

The Monge map $\tilde{T}$ between μ_1 and μ_0 is defined for all $x \in \mathbb{R}$ as $\tilde{T}(x) = \frac{\sigma_0}{\sigma_1}(x-m_1)+m_0 = h'(x)$ with $h : x \mapsto \frac{\sigma_0}{2\sigma_1}(x-m_1)^2 + m_0 x$. Then, by (Gallouët et al., 2024, Section 4), we know that we can extend the geodesic linking μ_1 to μ_0 on $[0,\alpha[$ for $\alpha \geq 1$ if and only if h is $\frac{\alpha-1}{\alpha}$-convex, *i.e.* if and only if

$$h''(x) - \frac{\alpha-1}{\alpha} \geq 0 \iff \frac{\sigma_0}{\sigma_1} \geq \frac{\alpha-1}{\alpha} \iff \frac{\sigma_1}{\sigma_0} \leq \frac{\alpha}{\alpha-1}. \tag{121}$$

Therefore, we deduce that we can extend the geodesic ray starting from μ_0 and passing through μ_1 at $t=1$ on $]-(\alpha-1),+\infty[$ if and only if $\frac{\alpha}{\alpha-1} \geq \frac{\sigma_1}{\sigma_0} \geq 1$ (the last inequality results from the geodesic ray condition $\sigma_1 \geq \sigma_0$). Thus, we find that the largest possible value $\alpha > 1$ satisfying inequality (121) is $\frac{\sigma_1}{\sigma_1-\sigma_0}$ as

$$\frac{\alpha}{\alpha-1} \geq \frac{\sigma_1}{\sigma_0} \iff \alpha\frac{\sigma_0-\sigma_1}{\sigma_0} \geq -\frac{\sigma_1}{\sigma_0} \iff \alpha \leq \frac{\sigma_1}{\sigma_1-\sigma_0}, \tag{122}$$

and for $\alpha = \frac{\sigma_1}{\sigma_1-\sigma_0}$, $1-\alpha = -\frac{\sigma_0}{\sigma_1-\sigma_0}$. Hence the geodesic ray can be extended at least over the interval $]-\frac{\sigma_0}{\sigma_1-\sigma_0},+\infty[$. □

E EXPERIMENTS ON LABELED DATASETS

We report here experimental details of the experiments done in Section 5. We begin with comparing the runtime of the different sliced distances. Then, we detail the experiment of correlation between the sliced distances and OTDD, as well as the gradient flows and transfer learning experiments. All the experiments are done on a Nvidia v100 GPU.

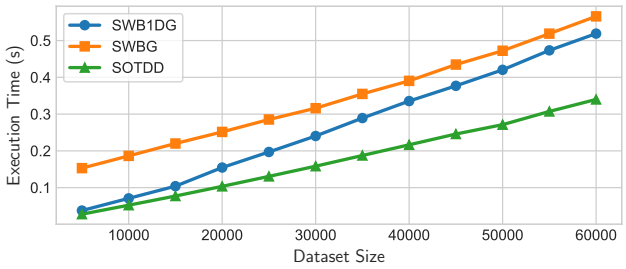


Figure 6: Comparison of the runtime in second between SOTDD, SWB1DG and SWBG on subsets of MNIST.

E.1 Runtimes

The theoretical computational complexity between SOTDD, SWB1DG and SWBG is about the same *w.r.t* the total number of samples n, *i.e.* $\mathcal{O}\big(Ln(\log n + d)\big)$. We verify this on Figure 6 by plotting the runtime between subsets of MNIST with different number of samples. We observe the same asymptotic runtime, which are super linear. The runtimes reported are averaged over 10 tries, but appear stable. For SWBG, we used a dimension reduction in $\mathbb{R}^{10}$ with TSNE.

E.2 Correlation

Drawing inspiration from the experiment of (Nguyen et al., 2025a), we consider OTDD to be the ideal distance between datasets and we aim to approximate it, or at least, we want to obtain bahavior similar to that of OTDD while being more efficient to compute. To assess the similarity between the sliced distances and OTDD, we measure their correlation.

Our protocol is the following, we first subsample a dataset, with random batchs of size between 5000 and 10000 samples. Then, we compute both the OTDD distance and the sliced distances between the pairs of batchs. Finally, we compute the Pearson and Spearman correlations to have a quantified value of the correlation. For pairs of data $(x_i, y_i)_{i=1}^n$, noting $\bar{x}^n = \frac{1}{n}\sum_{i=1}^n x_i$ and $\bar{y}^n = \frac{1}{n}\sum_{i=1}^n y_i$, the Pearson correlation ρ_P is defined as

$$\rho_P = \frac{\sum_{i=1}^n (x_i - \bar{x}^n)(y_i - \bar{y}^n)}{\sqrt{\sum_{i=1}^n (x_i - \bar{x}^n)^2}\sqrt{\sum_{i=1}^n (y_i - \bar{y}^n)^2}} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}, \tag{123}$$

and is equal to ± 1 when both quantities are linearly correlated. The Spearman correlation ρ_S is defined similarly, but between the order statistics, *i.e.* $(x_{\sigma_X(i)}, y_{\sigma_Y(i)})_{i=1}^n$ with $x_{\sigma_X(1)} \leq \cdots \leq x_{\sigma_X(n)}$ and $y_{\sigma_Y(1)} \leq \cdots \leq y_{\sigma_Y(n)}$ the sorted samples. This allows to recover if the quantities are correlated also in a non affine way. We use `SciPy` (Virtanen et al., 2020) for their computation in practice.

OTDD and SOTDD are computed using the code from Nguyen et al. (2025a) available at `https://github.com/hainn2803/s-OTDD`. For OTDD, the inner distances between the classes are computed by using the `POT` solver (Flamary et al., 2021, 2024) of the Wasserstein distance. The outer loss is computed using Sinkhorn with an entropic regularization of $\varepsilon = 10^{-3}$ and is debiased. SOTDD is computed using 5 moments. We report the results of all the sliced distances with 5000 projections on CIFAR10 on Figure 1, and with $L \in \{10, 50, 100, 500, 1000, 5000\}$ projections on CIFAR10 on Figure 8 and on MNIST on Figure 7. On MNIST, we use in every cases linear projections, while on CIFAR10, we use convolution projections (Nguyen and Ho, 2022). We use the code of (Nguyen et al., 2025a) for the choice of the random convolutions. For SWBG, we embed labels as Gaussian in $\mathbb{R}^{10}$ using TSNE, with the `TorchDR` library (Van Assel et al., 2024).

To obtain reasonably meaningful and statistically significant results, we report the results for 200 pairs of datasets on Figures 1, 7 and 8. In each case, we observe that SWB1DG and SWBG outperform in general SOTDD. Moreover, both the Spearman correlation ρ_S and the Pearson correlation ρ_P are close to 1, indicating an affine correlation. To emphasize this linear correlation, we also fit a linear regression on each of these figures.

Finally, we report the results by averaging over 10 sets of 50 bootstrapped pairs over the 200 initial ones. We report the results for different number of projections on Table 4 on MNIST and FashionMNIST. We observe again the superior results of SWB1DG and SWBG over Sliced-OTDD for a much smaller number of projections.

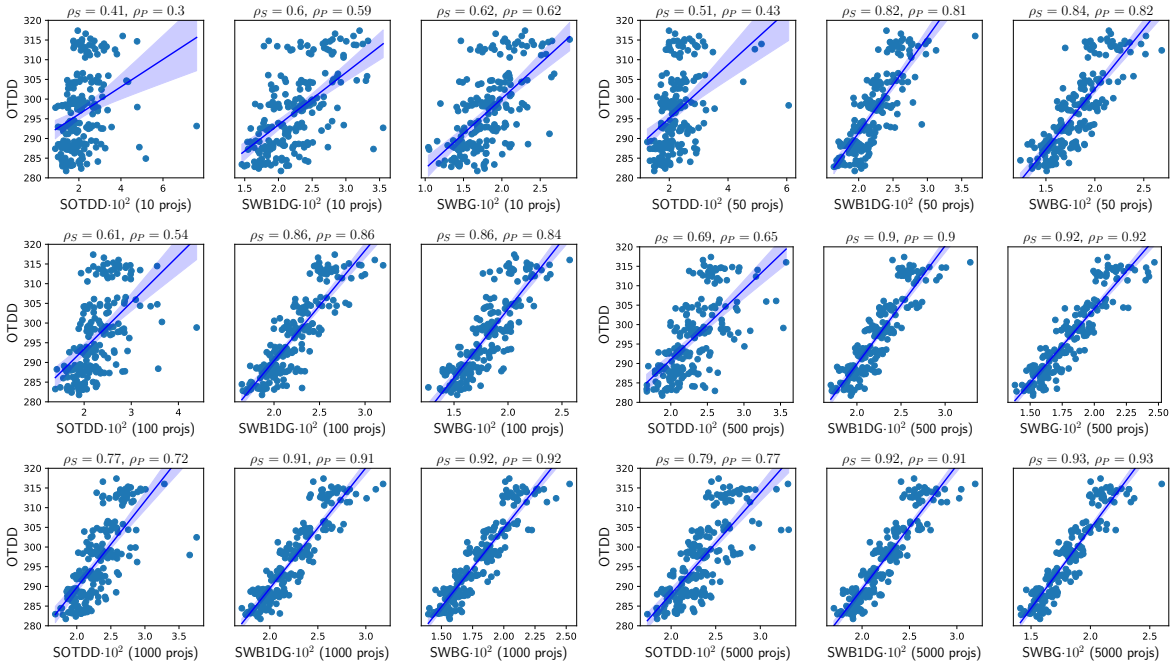


Figure 7: Distance correlation between s-OTDD, SWB1DG, SWBG and OTDD (exact) between subdatasets of MNIST.

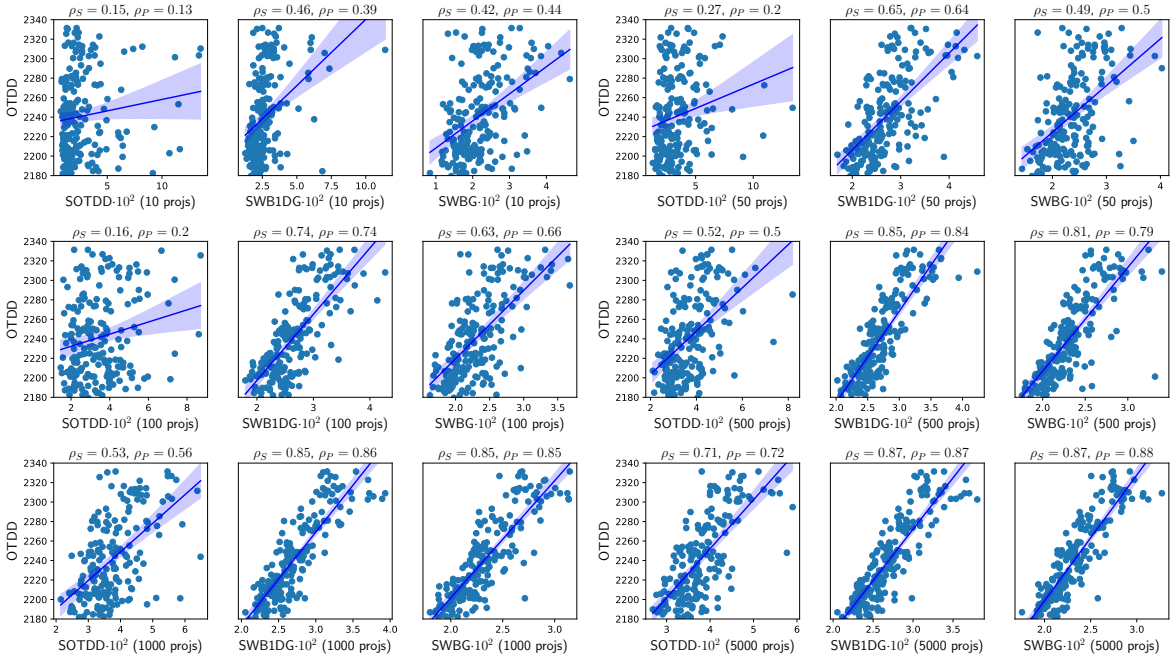


Figure 8: Distance correlation between s-OTDD, SWB1DG, SWBG and OTDD (exact) between subdatasets of CIFAR10.

E.3 Gradient Flows on Rings

The ring dataset is composed of 3 rings, where each ring is seen as a class of 80 samples. Thus, the target is $\mathbb{Q} = \frac{1}{3}\delta_{\hat{\nu}_1} + \frac{1}{3}\delta_{\hat{\nu}_2} + \frac{1}{3}\delta_{\hat{\nu}_3}$ with $\hat{\nu}_c = \frac{1}{n}\sum_{i=1}^{n}\delta_{x_{i,c}}$ and $n = 80$. On Figure 2, we evenly sample the points on each

Table 4: Correlation for different number of projections averaged over 10 datasets of 50 bootstrapped pairs.

Projections	MNIST						CIFAR10					
	Spearman correlation (ρ_S)			Pearson correlation (ρ_P)			Spearman correlation (ρ_S)			Pearson correlation (ρ_P)		
	SOTDD	SWB1DG	SWBG	SOTDD	SWB1DG	SWBG	SOTDD	SWB1DG	SWBG	SOTDD	SWB1DG	SWBG
10	$44.4_{\pm 6.4}$	$57.8_{\pm 11.3}$	$\mathbf{61.5}_{\pm 10.4}$	$36.2_{\pm 12.2}$	$60.2_{\pm 12.2}$	$\mathbf{61.3}_{\pm 13.1}$	$14.0_{\pm 11.3}$	$\mathbf{44.3}_{\pm 10.8}$	$40.2_{\pm 12.2}$	$16.0_{\pm 12.9}$	$38.6_{\pm 14.6}$	$\mathbf{42.7}_{\pm 9.5}$
50	$42.5_{\pm 8.9}$	$81.6_{\pm 4.5}$	$\mathbf{83.4}_{\pm 2.4}$	$39.8_{\pm 11.5}$	$81.7_{\pm 4.2}$	$\mathbf{82.8}_{\pm 2.0}$	$30.5_{\pm 12.9}$	$\mathbf{62.6}_{\pm 6.4}$	$40.4_{\pm 9.8}$	$25.2_{\pm 11.4}$	$\mathbf{63.6}_{\pm 6.3}$	$42.8_{\pm 8.5}$
100	$62.6_{\pm 7.4}$	$84.0_{\pm 6.2}$	$\mathbf{84.3}_{\pm 3.5}$	$55.9_{\pm 9.5}$	$\mathbf{86.0}_{\pm 3.8}$	$83.7_{\pm 3.4}$	$15.5_{\pm 11.8}$	$\mathbf{71.9}_{\pm 6.4}$	$68.1_{\pm 7.2}$	$21.0_{\pm 11.4}$	$\mathbf{73.9}_{\pm 5.5}$	$72.8_{\pm 5.4}$
500	$67.0_{\pm 7.5}$	$90.4_{\pm 1.7}$	$\mathbf{91.2}_{\pm 1.6}$	$64.6_{\pm 7.0}$	$90.4_{\pm 2.1}$	$\mathbf{91.6}_{\pm 2.0}$	$52.1_{\pm 8.1}$	$\mathbf{82.3}_{\pm 2.2}$	$78.4_{\pm 6.0}$	$54.6_{\pm 8.8}$	$\mathbf{83.5}_{\pm 2.1}$	$79.4_{\pm 7.7}$
1000	$77.7_{\pm 4.1}$	$89.6_{\pm 1.5}$	$\mathbf{91.0}_{\pm 1.5}$	$75.6_{\pm 8.3}$	$91.5_{\pm 1.3}$	$\mathbf{92.1}_{\pm 1.5}$	$52.0_{\pm 10.9}$	$83.6_{\pm 4.8}$	$\mathbf{83.7}_{\pm 5.0}$	$53.1_{\pm 11.3}$	$\mathbf{85.6}_{\pm 3.5}$	$84.9_{\pm 4.8}$
5000	$78.8_{\pm 6.2}$	$91.4_{\pm 1.8}$	$\mathbf{92.6}_{\pm 1.6}$	$77.8_{\pm 5.8}$	$91.4_{\pm 1.5}$	$\mathbf{93.0}_{\pm 1.6}$	$72.2_{\pm 7.5}$	$88.5_{\pm 4.8}$	$\mathbf{89.3}_{\pm 3.8}$	$75.4_{\pm 5.5}$	$87.8_{\pm 2.8}$	$\mathbf{89.0}_{\pm 2.4}$
10000	$78.7_{\pm 3.9}$	$89.8_{\pm 0.2}$	$\mathbf{91.3}_{\pm 1.5}$	$78.9_{\pm 4.3}$	$91.1_{\pm 2.0}$	$\mathbf{92.6}_{\pm 1.7}$	$72.6_{\pm 6.1}$	$82.7_{\pm 4.8}$	$\mathbf{86.7}_{\pm 3.0}$	$77.1_{\pm 4.3}$	$87.3_{\pm 2.8}$	$\mathbf{90.2}_{\pm 2.3}$

Table 5: Best hyperparameters on the k-shot transfer learning experiments for SWB1DG and SOTDD.

	SWB1DG								SOTDD							
	MNIST to FMNIST				MNIST to USPS				MNIST to FMNIST				MNIST to USPS			
	$k=1$	$k=5$	$k=10$	$k=100$	$k=1$	$k=5$	$k=10$	$k=100$	$k=1$	$k=5$	$k=10$	$k=100$	$k=1$	$k=5$	$k=10$	$k=100$
Projections	5K	5K	1K	10K	10K	10K	1K	10K	10K	5K	10K	1K	5K	10K	10K	10K
Gradient steps	1K	10K	10K	20K	1K	20K	20K	20K	10K	10K	10K	20K	1K	10K	10K	20K

ring, and thus have always the same target. On Figure 3, we sampled 80 points uniformly on each ring, which means they may not be evenly spaced. In addition, we averaged the results over 100 different samples of the target.

We compare on Figure 3 the convergence of the WoW gradient flows of SOTDD, SWB1DG and SWBG towards the target. These flows are approximated by performing a WoW gradient descent (see Appendix A.3 for details) over $\mathbb{P} = \frac{1}{3}\delta_{\mu_1} + \frac{1}{3}\delta_{\mu_2} + \frac{1}{3}\delta_{\mu_3}$ with each μ_c of the form $\mu_c = \frac{1}{n}\sum_{i=1}^n \delta_{x_{i,c}}$ for $n = 80$, with 1000 iterations, and a step size of $\tau = 1$.

We also tried minimizing SOTDD, SWB1DG and SWBG with $\alpha_1 = 0$, *i.e.* comparing the measures on $\mathcal{P}_2\big(\mathcal{P}_2(\mathbb{R}^d)\big)$ instead of on $\mathcal{P}_2\big(\mathbb{R}^d \times \mathcal{P}_2(\mathbb{R}^d)\big)$. However, while it may be enough to compare the distributions, the sliced distances were not capable in this case to flow the points towards the rings. The flows only recovered roughly the means, and the first moments, but the particles did not exactly match the target rings. Thus, it is important to use the representations on $\mathcal{P}_2\big(\mathbb{R}^d \times \mathcal{P}_2(\mathbb{R}^d)\big)$ to flow datasets with these distances.

E.4 k-shot Learning

We provide the details of the k-shot learning experiment, inspired from (Alvarez-Melis and Fusi, 2021). In this task, we want to learn a classifier on a dataset from which we have access to k samples by class, where k is typically small. Alvarez-Melis and Fusi (2021) proposed to solve this task by flowing a source dataset, from which we have access to more samples by class, towards the dataset of interest, hence augmenting each class with new additional images.

Let $\mathcal{D}^*$ be the dataset of interest and $C \in \mathbb{N}^*$ its number of classes. Let us denote $\mathbb{Q} = \frac{1}{C}\sum_{c=1}^C \delta_{\nu_c^k} = \psi(\mu_{\mathcal{D}^*})$ with $\nu_c^k = \frac{1}{k}\sum_{\ell=1}^k \delta_{y_\ell^c}$ the target dataset, considered here on $\mathcal{P}_2\big(\mathcal{P}_2(\mathbb{R}^d)\big)$. Alvarez-Melis and Fusi (2021) proposed to flow the datasets by minimizing $\mathbb{F}(\mathbb{P}) = \mathrm{OTDD}(\mathbb{P}, \mathbb{Q})$ starting from $\mathbb{P}_0 = \psi(\mu_{\mathcal{D}_0})$ with $\mu_{\mathcal{D}_0}$ an initial dataset with $n \gg k$ samples by class, *i.e.* $\mathbb{P}_0 = \frac{1}{C}\sum_{c=1}^C \delta_{\mu_c^{n,0}}$ and $\mu_c^{n,0} = \frac{1}{n}\sum_{i=1}^n \delta_{x_i^{c,0}}$. Alvarez-Melis and Fusi (2021) proposed several strategies to minimize OTDD. For instance, approximating the label distributions by Gaussian, they minimized it using gradient flows on the product space $\mathbb{R}^d \times \mathbb{R}^d \times S_d^{++}(\mathbb{R})$, by flowing simultaneously the samples, the means and the covariances in a decouple way.

Following (Alvarez-Melis and Fusi, 2021), several works proposed to solve this task. In particular, Hua et al. (2023) solved it by minimizing the MMD with a Gaussian kernel over the product space $\mathbb{R}^d \times \mathbb{R}^p \times S_p^{++}(\mathbb{R})$, using a dimension reduction to obtain embedding of the labels in a space of dimension $p \ll d$, and using a Riemannian gradient descent on the Bures-Wasserstein space for the covariance part. More recently, Bonet et al. (2025a) proposed to minimize an MMD on $\mathcal{P}_2\big(\mathcal{P}_2(\mathbb{R}^d)\big)$ with kernels on $\mathcal{P}_2(\mathbb{R}^d)$. In particular, they used the SW-Riesz kernel $K(\mu, \nu) = -\mathrm{SW}_2(\mu, \nu)$, and used a Wasserstein over Wasserstein Gradient descent to minimize it, endowing $\mathcal{P}_2\big(\mathcal{P}_2(\mathbb{R}^d)\big)$ with $\mathrm{W}_{\mathrm{W}_2}$.

In this work, we also perform the gradient descent in $(\mathrm{W}_{\mathrm{W}_2}, \mathcal{P}_2\big(\mathcal{P}_2(\mathbb{R}^d)\big))$ following the theory derived in (Bonet et al., 2025a), and minimize SOTDD and SWB1DG. The scheme to minimize $\mathbb{F}(\mathbb{P}) = \mathrm{D}(\mathbb{P}, \mathbb{Q})$ with D any

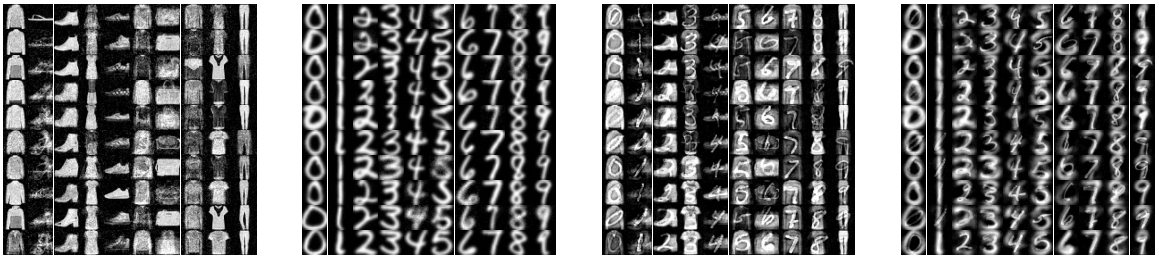


Figure 9: Examples of images output by the flows for the transfer learning task on Fashion MNIST and USPS with $k = 10$ and the best performing hyperparameters, for SWB1DG (**Left**) and SOTDD (**Right**).

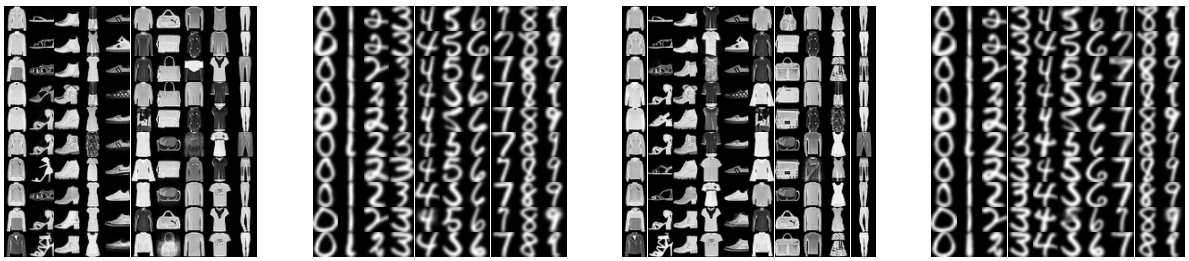


Figure 10: Examples of images output by the flows for the transfer learning task on Fashion MNIST and USPS with $k = 10$, 100K epochs and 10K projections, for SWB1DG (**Left**) and SOTDD (**Right**).

differentiable divergence on $\mathcal{P}_2\big(\mathcal{P}_2(\mathbb{R}^d)\big)$ is, for any $\ell \geq 0$, $c \in \{1, \dots, C\}$ and $i \in \{1, \dots, n\}$, and step size $\tau > 0$,

$$x_i^{c,\ell+1} = x_i^{c,\ell} - \tau \nabla_{\mathrm{W}_{\mathrm{W}_2}} \mathbb{F}(\mathbb{P}^\ell)(\mu_c^{n,\ell})(x_i^{c,\ell}), \tag{124}$$

where the Wasserstein over Wasserstein gradient $\nabla_{\mathrm{W}_{\mathrm{W}_2}} \mathbb{F}(\mathbb{P}^\ell)(\mu_c^{n,\ell})(x_i^{c,\ell})$ is obtained by rescaling the Euclidean gradient of $F(\mathbf{x}^\ell) = \mathbb{F}(\mathbb{P}^\ell)$ for $\mathbf{x}_c^\ell = (x_i^{c,\ell})_{i,c}$, *i.e.* $\nabla_{\mathrm{W}_{\mathrm{W}_2}} \mathbb{F}(\mathbb{P}^\ell)(\mu_c^{n,\ell})(x_i^{c,\ell}) = nC\nabla F(\mathbf{x}_c^\ell)$ (Bonet et al., 2025a, Proposition B.7). In practice, ∇F is obtained by backpropagation.

Details of the Experiments. The datasets of interest on which we learn classifiers are Fashion-MNIST and USPS. Thus, the number of class is always $C = 10$, and we use $k \in \{1, 5, 10, 100\}$. For the source dataset $\mathbb{P}_0$, we always use MNIST with $n = 200$ samples by class. In Table 2, we report the accuracy obtained by training a LeNet-5 neural network for 50 epochs, with a AdamW optimizer and a learning rate of $3 \cdot 10^{-4}$. We also average the results for 5 trainings of the neural network, and 3 outputs of the flows. The code to set-up the experiment is taken from the github of (Bonet et al., 2025a) available at `https://github.com/clbonet/Flowing_Datasets_with_WoW_Gradient_Flows`. We compare the results between the baseline where the neural network is trained directly on the target dataset $\mathcal{D}^*$ with k samples by class, and minimizing OTDD. These results are taken from (Bonet et al., 2025a, Table 2).

For the minimization of SOTDD and SWB1DG, we choose a step size of $\tau = 1$, a momentum of $m = 0.9$, and do a grid search over the number of gradient steps $T \in \{100, 200, 500, 1000, 10000, 20000\}$ and the number of projections $L \in \{500, 1000, 5000, 10000\}$ to approximate the integrals. We report in Table 2 the best results over this grid search, and on Table 5 the values of the hyperparameters giving the best results. We also show on Figure 9 examples of images in each class obtained after minimizing respectively SWBG and SOTDD, for $k = 10$. We observe that these images are not always very clean. Thus, to show that minimizing these distances allows to obtain good looking images, we also report on Figure 10 results for $k = 10$, $L = 10K$ and $T = 100K$, which however gave worse results than the ones reported in Table 2. We hypothesize that the presence of noise can help the classifier to better generalize and thus to improve the accuracy results.

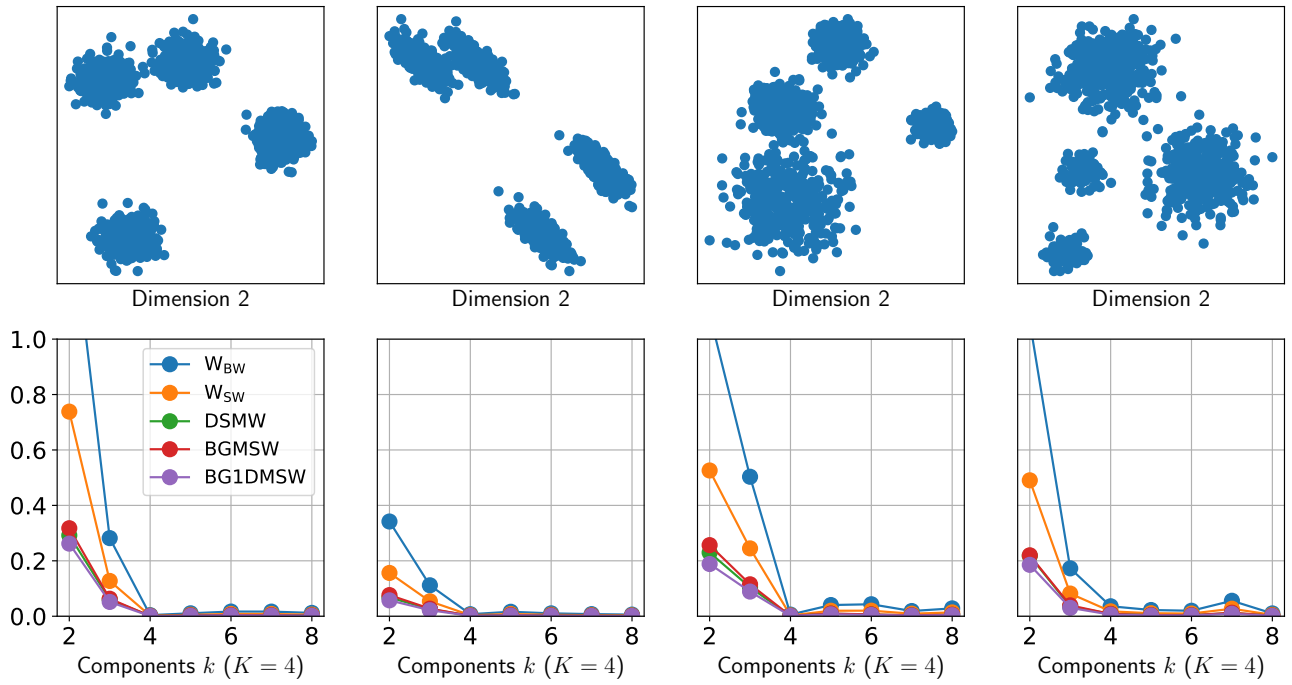


Figure 11: Detection of the number of clusters in dimension $d = 2$.

F EXPERIMENTS ON GAUSSIAN MIXTURES

We show how the Gaussian Busemann Sliced-Wasserstein distances compare with W_{BW}, W_{SW} and DSMW on the tasks of clustering detection and gradient flows.

F.1 Clustering Detection

We perform the same experiment as in (Piening and Beinert, 2025, Section 4.3). We want to fit a Gaussian mixture on a dataset. A common technique is to use an EM. However, the number of clusters in the mixture needs to be specified and is not necessarily known beforehand. For K components, denote $\mathbb{P}_K \in \mathcal{P}_2\big(\mathrm{BW}(\mathbb{R}^d)\big)$ the corresponding mixture. Piening and Beinert (2025) propose to compute the distances between $\mathbb{P}_k$ and $\mathbb{P}_{k+1}$ for $k \geq 1$, and to find the smallest $k \in \mathbb{N}^*$ such that adding more components does not change the fitted model. To do so, they increase k until the distances vanish.

We report on Figure 11 the results using the same setting as (Piening and Beinert, 2025), *i.e.* we have 1500 samples of a mixture in dimension $d = 2$ with 4 clusters. On Figure 12, we add an experiment in dimensions $d \in \{3, 5, 10, 20\}$ with $K \in \{4, 5, 6\}$ components, with means distributed uniformly on a sphere with radius $r = 100$ and random covariance matrices. In every cases, we observe that the results between DSMW, BGMSW and B1DGMSW are very close.

For DMSW, we use the code of (Piening and Beinert, 2025), available at `https://github.com/MoePien/sliced_OT_for_GMMs` and for W_{BW}, we use the implementation of `POT` (Flamary et al., 2021).

F.2 Flows

As a proof of concept, we show on Figure 13 the trajectories of flows minimizing W_{BW}, DMSW, BGMSW and B1DGMSW over $\mathcal{P}_2\big(\mathrm{BW}(\mathbb{R}^d)\big)$. We use the code from `POT` (Flamary et al., 2021) of an example minimizing W_{BW} and available at `https://pythonot.github.io/auto_examples/gaussian_gmm/plot_GMM_flow.html#sphx-glr-auto-examples-gaussian-gmm-plot-gmm-flow-py`.

In this experiment, we start from a mixture with 3 Gaussian, with weights, means, and covariances randomly sampled. The target is a mixture with 2 Gaussian with also weights, means and covariances randomly sampled.

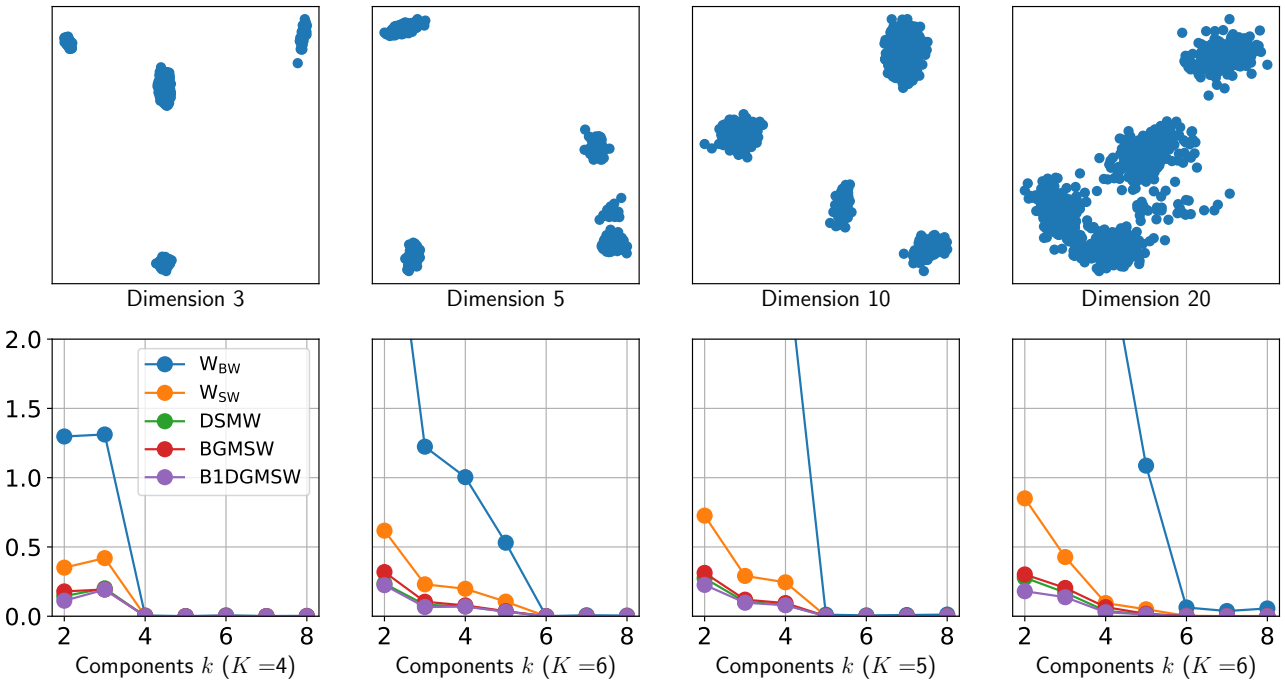


Figure 12: Detection of the number of clusters in dimension $d > 2$.

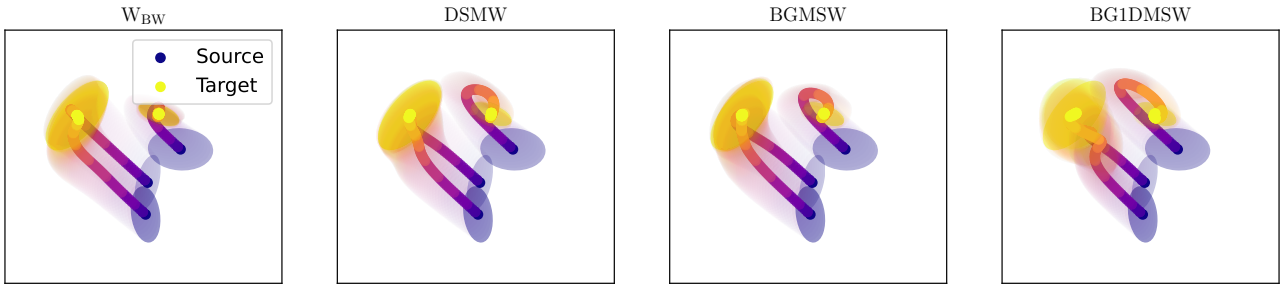


Figure 13: Flows of Gaussian Mixtures minimizing W_{BW}, DMSW, BGMSW or B1DGMSW.

Then, we optimize over the weights, means and covariances using the Adam optimizer and with projection steps to stay on the space of Gaussian mixtures, *i.e.* we use a softmax to project the weights and clip the eigenvalues of the matrices to project the covariances in the space of positive definite matrices. The gradients are obtained using backpropagation. Note that this way of minimizing does not correspond to a gradient descent in $\mathcal{P}_2(\mathrm{BW}(\mathbb{R}^d))$, which would not allow to change the weights, see *e.g.* (Bonet et al., 2025a, Appendix D.7) for a discussion on how to handle different number of components in the mixture.