Linearized optimal transport on manifolds

Clément Sarrazin* and Bernhard Schmitzer†

March 27, 2023

Abstract

Optimal transport is a geometrically intuitive, robust and flexible metric for sample comparison in data analysis and machine learning. Its formal Riemannian structure allows for a local linearization via a tangent space approximation. This in turn leads to a reduction of computational complexity and simplifies combination with other methods that require a linear structure. Recently this approach has been extended to the unbalanced Hellinger–Kantorovich (HK) distance. In this article we further extend the framework in various ways, including measures on manifolds, the spherical HK distance, a study of the consistency of discretization via the barycentric projection, and the continuity properties of the logarithmic map for the HK distance.

1 Introduction

1.1 Motivation

Optimal transport in data analysis. Optimal transport (OT) induces a geometrically intuitive and robust metric on the set of probability measures over a metric space, with many applications in geometry, stochastics and PDE analysis. OT is also becoming an increasingly popular tool in data analysis and machine learning, including numerical applications in image classification [19], inverse problems [18], and deep learning [5, 12]. This is also made possible by the development of increasingly efficient corresponding algorithms. We refer to [45, 42, 21] for general introductions to the topic of optimal transport and to [40] for an overview on computational techniques.

Consider a collection of probability measures $\{\nu_1, \dots, \nu_N\} \subset \mathcal{P}(\mathbb{R}^n)$, each ν_i representing one data sample, and assume that the 2-Wasserstein distance W_2 is a meaningful metric for their comparison. Our task is to analyze this dataset. This means we might have to train a classifier, according to some given labels $(l_i, \dots, l_N)$; to find a meaningful clustering; or to perform a dimensionality reduction and to visualize the dominant types of variation within the samples.

To solve this task, we could, in principle, start by computing the full matrix of pairwise Wasserstein distances $\mathrm{W}_2(\nu_i, \nu_j)$ between all samples. However, this would require the solution of $N(N-1)/2 = O(N^2)$ optimal transport problems. Due to the quadratic growth in N this will quickly become a computational problem, despite recent progress in algorithmic efficiency. Moreover, any of the aforementioned tasks is difficult to perform in a general (non-linear) metric space, based solely on the knowledge of the pairwise distances.

*Campus Institute Data Science, Göttingen University, Germany
†Campus Institute Data Science, Göttingen University, Germany

Linearized optimal transport. Instead, in [46] it was proposed to leverage the weak Riemannian structure of the 2-Wasserstein distance over $\mathbb{R}^n$ (see for instance [37, 4, 33]) and to formally conduct a local linearization by a tangent space approximation. For this purpose, one chooses a suitable reference measure $\mu \in \mathcal{P}(\mathbb{R}^n)$ and maps all samples ν_i to the tangent space at μ via the logarithmic map, $v_i := \mathrm{Log}_\mu(\nu_i)$. The tangent space can be identified with (curl free) vector fields in $\mathbb{L}^2(\mathbb{R}^n, \mathbb{R}^n; \mu)$. Pairwise distances between the embeddings v_i in the tangent space are then used to approximate distances between samples. This is called the linearized optimal transport distance (LOT). Beyond mere approximation of pairwise distances, the Hilbertian structure of $\mathbb{L}^2(\mathbb{R}^n, \mathbb{R}^n; \mu)$ allows the application of a wide variety of standard data analysis techniques to tackle any of the aforementioned potential analysis tasks. For instance, principal component analysis (PCA) can be used for dimensionality reduction and to extract the dominant modes of variation. Interpolation between samples ν_i is replaced by linear interpolation between the embeddings v_i and these trajectories can be visualized by employing the exponential map (a left-inverse of the logarithmic map). Such curves are called generalized geodesics, introduced in [4], we also refer the reader to [36] for a recent article on the subject. In one dimension, $n = 1$, this embedding reduces to the cumulative distance transform [38], which drastically reduces the computational complexity. This can still be applied to higher dimensional datasets via the slicing trick [27].

The empirical success in applications has also lead to an increased interest in the theoretical properties of the method. How good is the approximation of the pairwise distances by local linearization? In the conventional, finite-dimensional Riemannian setting this question is tied to the curvature of the manifold. The results do not apply to W_2 as here the Riemannian structure is merely formal and additional assumptions are necessary to establish some approximation guarantees. For relatively simple families of samples this is addressed in [26]. More general continuity properties of the logarithmic map are established, for instance, in [25, 17]. Another relevant question is the choice of the reference measure μ. From a Riemannian perspective it should be chosen as the Riemannian center of mass (or barycenter) of the sample set $(\nu_i)_i$, which has been established for the Wasserstein distance in [1]. While this can, in principle, be approximated numerically, for instance via [6], this is numerically not feasible for large number of samples. In practice an approximation by (a few or even a single iteration of) Lloyd's algorithm works well, as highlighted in [16] and [2], and even naive linear averaging of the measures (e.g. as densities in some $\mathbb{L}^p$ space) tends to work acceptably. [8] studies the consistency of the barycenter as the number of samples tends to infinity, and [7] studies the consistency of tangent space principal component analysis in the one-dimensional, $n = 1$, setting.

Unbalanced optimal transport. Recently, optimal transport has also been extended to comparing measures of varying mass, the most prominent example maybe being the Hellinger–Kantorovich distance (HK) [29, 13, 31]. From the perspective of data analysis this can explicitly model growth or shrinkage, but can also provide additional robustness towards general mass fluctuations. The linearized optimal transport framework has been extended to the Hellinger–Kantorovich distance in [11].

1.2 Contribution and outline

In this article we further extend the toolbox of linearized optimal transport and provide some new theoretical results. In particular we consider the case where the base space X is no longer flat $\mathbb{R}^n$, but itself a non-linear Riemannian manifold.

Sections 2.1 and 2.2 recall basic properties of the W_2 distance and its local linearization. In Section 2.3 we show how to extend this to the case of manifold base spaces X. In Section 3.1 we recall basic properties of the HK metric. In Section 3.2 we give a slight extension of a result from [23] concerning the existence of optimal HK transport maps on manifold base spaces X. In Section 3.3 we introduce the local linearization for manifold base spaces X. Beyond the extension to base manifolds, compared to [11] the formulas have been simplified by using primal-dual optimality conditions for the HK transport problem. In addition we give a formula for the logarithmic map based on a suitable dual variable. The latter will be useful when determining the range of the logarithmic map (see Section 4). In Section 3.4 we introduce the local linearization of the spherical Hellinger–Kantorovich (SHK) distance, a projection of the HK metric structure which was introduced in [30]. Working with SHK might be preferable, when all samples are indeed probability measures, but still present non-local mass fluctuations. The latter make it difficult to work with the standard Wasserstein distance, but working with HK will introduce a bias towards smaller masses. The SHK metric may be a suitable way to attain robustness to local mass fluctuations without the bias induced by the HK metric, see Section 6 for an illustration. Results from [30] show how SHK can be computed directly from HK and we extend these results to the logarithmic and exponential maps.

In Section 4 we provide a discussion on the range of the logarithmic map on the (formal) optimal transport manifolds. Along most directions in tangent space the range of the logarithmic map cannot be unbounded as the induced transport map will eventually run into particle collisions and cusps, violating optimality. From a practical perspective a convex range is desirable, as it guarantees that operations such as averaging between samples remains in the range. We discuss how this property is related to the convexity of the set of c-concave functions and thus to the Ma–Trudinger–Wang (MTW) tensor [34], the Loeper condition [32], and the curvature of the base manifold (see [45, 22] for an introduction to these concepts). We discuss the range of the logarithmic map for HK when X is a sphere. This is related to results from [23] concerning the regularity of unbalanced Monge maps, but our discussion suggests that a different MTW tensor than the one in [23] has to be studied to answer the question.

In Section 5 we give several approximation results. First (Section 5.1) we show that linearized HK converges to standard linearized W_2, as the length scale parameter in HK tends to infinity, in agreement with the fact that HK converges to W_2 [31]. In Section 5.2 we show that 'barycentric projection', a popular technique for numerical approximation of LOT for W_2, does actually converge to the true logarithmic map in the limit of increasing discretization resolution. In Section 5.3 we give the corresponding result for the HK distance. This extends beyond discrete approximation and generally provides a continuity (or lack thereof) result for the logarithmic map of the HK metric. We illustrate this result with several examples.

Section 6 provides some numerical experiments, in particular a comparison between linearized HK and SHK, and an example of linearized optimal transport over a sphere as base manifold.

1.3 Notation

Riemannian manifolds. Throughout the article X will be a connected complete n-dimensional Riemannian manifold. For $x \in X$ denote by T_xX the tangent space of X at x, and by $TX := \bigcup_{x\in X}\{x\} \times T_xX$ the tangent bundle of X. For $v, w \in T_xX$ we denote by $\langle v, w\rangle_x$ the corresponding Riemannian inner product and by $\|v\|_x$ the induced norm.

We will denote by Log_x^X the logarithmic map on X at x and likewise, by Exp_x^X the exponential map at x. Intuitively, $\mathrm{Exp}_x^X(v)$ yields the point on X that one reaches at time 1 when one embarks on a constant speed geodesic starting at x with velocity vector v. Conversely, $\mathrm{Log}_x^X(y)$

yields the smallest such initial velocity vector v such that one ends up at y, when starting from x along this geodesic. In particular, $\mathrm{Exp}_x^X(\mathrm{Log}_x^X(y)) = y$, i.e. the exponential map is a left inverse of the logarithmic map (however, the image of Log_x^X might be strictly smaller than T_xX). Because we are working on a complete and connected manifold, the Hopf-Rinow theorem guarantees that the exponential map is defined on the entire tangent bundle TX (geodesics can be extended indefinitely before $t = 0$ or after $t = 1$) and that any two points in X can be connected by such a geodesic (so that it is surjective).

Example 1.1. When X is the 'flat space' $\mathbb{R}^n$ (with the Euclidean metric), then $T_xX = \mathbb{R}^n$, $\mathrm{Log}_x^X(y) = y - x$ is well defined for any x and y and $\mathrm{Exp}_x^X(v) = x + v$.

Measures on metric spaces. For a Polish metric space X, we will denote by $\mathcal{M}(X)$ the signed Radon measures on X, by $\mathcal{M}_+(X)$ the non-negative Radon measures and by $\mathcal{P}(X)$ the Radon probability measures.

When X is a Riemannian manifold, denote by Vol its volume measure. We will write $\mathcal{M}_{+,\mathrm{Vol}}(X)$ for the non-negative measures on X that are dominated by Vol and similarly $\mathcal{P}_{\mathrm{Vol}}(X)$ for the respective probability measures. Intuitively, the volume measure will play the role of the Lebesgue measure on the curved space X. In particular, when X is the flat space $\mathbb{R}^d$, $\mathrm{Vol} = \mathcal{L}^d$ is the Lebesgue measure itself.

For a non-negative measure μ on X we will denote by $\mathbb{L}^2(X;\mu)$ the space of real functions on X whose square is μ-integrable (up to equality μ-almost everywhere) and $\mathbb{L}^2(X, TX;\mu)$ the space of vector fields on X whose squared norm (for the metric on TX) is μ-integrable (again, up to equality μ-almost everywhere). We may sometimes merely write $\mathbb{L}^2(\mu)$ when the context is clear.

Finally, for $\mu \in \mathcal{M}(X)$ and a measurable map $T : X \to Y$ (where Y is a Polish space equipped with its Borel σ-algebra), the *push-forward measure* of μ along T is the measure $T_\#\mu$ on Y characterized by

$$\int_Y \Phi(x_1)\mathrm{d}(T_\#\mu)(x_1) = \int_X \Phi(T(x_0))\mathrm{d}\mu(x_0) \tag{1.1}$$

for all bounded measurable $\Phi : Y \to \mathbb{R}$.

Finally, let K be a compact subset of X. For the sake of simplicity, we will frequently assume that the support of measures is contained in K to avoid technical compactness arguments.

Optimal transport. We refer to [45, 42, 21, 40] for detailed monographs on optimal transport, applications, and corresponding computational methods. In the following paragraph we collect some fundamental definitions and concepts that are required for the manuscript. For two measures $\mu_0, \mu_1 \in \mathcal{M}_+(X)$ and a lower-semicontinuous cost function $c : X \times X \to \mathbb{R} \cup \{\infty\}$ the corresponding optimal transport problem is defined as

$$\mathcal{I}_c(\mu_0, \mu_1) := \inf \left\{ \int_{X\times X} c(x_0, x_1)\,\mathrm{d}\pi(x_0, x_1) \;\middle|\; \pi \in \Pi(\mu_0, \mu_1) \right\} \tag{1.2}$$

were the infimum is taken among all *transport plans* between μ_0 and μ_1

$$\Pi(\mu_0, \mu_1) := \left\{ \pi \in \mathcal{M}_+(X \times X) \;\middle|\; \mathrm{p}_{i\#}\,\pi = \mu_i \text{ for } i = 0, 1 \right\} \tag{1.3}$$

and $\mathrm{p}_i : X \times X \to X$ is the projection onto the i-th coordinate, $(x_0, x_1) \mapsto x_i$, and therefore $\mathrm{p}_{i\#}\,\pi$ yields the corresponding marginal of π, see (1.1). When $\mu_0(X) \neq \mu_1(X)$ then $\Pi(\mu_0, \mu_1) = \emptyset$ and

therefore by convention $\mathcal{I}_c(\mu_0,\mu_1) = +\infty$. Sometimes the optimal plan π is induced by a Borel map $T : X \to X$ via $\pi = (\mathrm{id}, T)_{\#}\mu_0$, where $(\mathrm{id}, T) : x \mapsto (x, T(x))$. In this case, T is called an optimal transport map for the associated OT problem.

Since $\mathcal{I}_c(\mu_0,\mu_1)$ is defined via a convex minimization problem of a linear objective over an affine set, one can derive a corresponding concave dual maximization problem:

$$\mathcal{I}_c(\mu_0,\mu_1) = \sup\left\{ \int_X \Phi_0(x_0)\,\mathrm{d}\mu_0(x_0) + \int_X \Phi_1(y)\,\mathrm{d}\mu_1(y) \;\middle|\; (\Phi_0,\Phi_1) \in \mathrm{C}_c(X) \right\} \tag{1.4}$$

with the admissible set given by

$$\mathrm{C}_c(X) := \left\{ (\Phi_0,\Phi_1) \in \mathrm{C}(X)\times\mathrm{C}(X) \;\middle|\; \forall (x_0,x_1)\in X^2,\ \Phi_0(x_0)+\Phi_0(y) \le c(x_0,x_1) \right\}. \tag{1.5}$$

Dual maximizers are sometimes referred to as (Kantorovich) potentials. Depending on the regularity of c or compactness of the suppport, existence of such Kantorovich potentials may require the space $\mathrm{C}(X)$ to be relaxed, e.g. to $\mathbb{L}^1(X,\mu_i)$ (see for instance [42, Theorem 1.40]). Since μ_0 and μ_1 are non-negative measures, for fixed Φ_0 it is trivial to maximize over Φ_1. One obtains

$$\Phi_1 = \Phi_0^c := \inf_{x_0\in X} c(x_0,.) - \Phi_0(x_0).$$

The function Φ_0^c is called the c-transform of Φ_0 and a function obtained in that way is said to be c-concave. Likewise, for fixed Φ_1 we can maximize over Φ_0. When c is symmetric, as all costs considered throughout this article are, both c-transforms coincide. In (1.4) one may therefore impose the additional constraint that the Φ_i are c-transforms of each other and therefore c-concave, and we will make that assumption for simplicity in the future.

2 2-Wasserstein distance

2.1 Basic properties

Setting $c = d^2$ as the squared distance in (1.2), $\mathcal{I}_c$ becomes the squared 2-Wasserstein distance on suitable subsets of probability measures, e.g. those with support contained in some compact set $K \subset X$.

Definition 2.1 (2-Wasserstein distance)**.** For $\mu_0,\mu_1 \in \mathcal{P}(K)$, the corresponding 2-Wasserstein distance between μ_0 and μ_1 is defined as

$$\mathrm{W}_2(\mu_0,\mu_1) := \inf\left\{ \int_{X\times X} d^2(x_0,x_1)\,\mathrm{d}\pi(x_0,x_1) \middle| \pi \in \Pi(\mu_0,\mu_1) \right\}^{1/2}. \tag{2.1}$$

It is a distance on $\mathcal{P}(K)$ and the resulting metric space $(\mathcal{P}(K), W_2)$ is often called the 2-Wasserstein space (over K).

The construction can be generalized to any $p \in [1,\infty)$. For a set K that would be non-compact, an additional bound on the p-th moments of the μ_i needs to be imposed, so as to guarantee a finite transport cost.

Existence of a transport plan π minimizing (2.1) is a corollary of the general existence result for (1.2).

Proposition 2.2. *Minimizing π for Problem (2.1) exist.*

Furthermore, compactness of K, together with the continuity of $c = d_X^2$ yields the existence of optimal Kantorovich potentials (which can then be taken to be c-concave).

Proposition 2.3 (Dual transport problem)**.** *For $\mu_0, \mu_1 \in \mathcal{P}(K)$,*

$$\mathrm{W}_2^2(\mu_0,\mu_1) = \sup\left\{ \int_X \Phi_0 \,\mathrm{d}\mu_0 + \int_X \Phi_1 \,\mathrm{d}\mu_1 \middle| \Phi_0, \Phi_1 \in \mathrm{C}(X), \right. \\ \left. \Phi_0(x_0) + \Phi_1(x_1) \le d^2(x_0,x_1) \text{ for all } (x_0,x_1) \in X \times X \right\} \quad (2.2)$$

Maximizing Φ_0, Φ_1 exist.

For $X = \mathbb{R}^n$ and $\mu_0 \in \mathcal{P}_{\mathrm{Vol}}(X)$ a celebrated result by Brenier [10] shows that the minimizing π in (2.1) is unique and concentrated on the graph of the gradient of a convex function.

Theorem 2.4 (Brenier)**.** *Let $X = \mathbb{R}^n$ endowed with its flat metric, $\mu_0 \in \mathcal{P}_{\mathrm{Vol}}(K)$ and $\mu_1 \in \mathcal{P}(K)$. Then,*

(i) The minimizer π of (2.1) is unique and concentrated on the graph of a map $T : K \to K$, i.e. it is given by $\pi = (\mathrm{id}, T)_{\#}\mu_0$ where $(\mathrm{id}, T) : x \mapsto (x, T(x))$.

(ii) The map T satisfies $T = \mathrm{id} - \frac{1}{2}\nabla\Phi_0$ μ_0-almost everywhere, where Φ_0 is a corresponding dual optimizer of (2.2). T is the gradient of the function $x \mapsto \frac{1}{2}(\|x\|^2 - \Phi_0(x))$, which is convex.

(iii) Conversely, if $\Phi_0 \in \mathrm{C}(X)$ such that $x \mapsto \frac{1}{2}(\|x\|^2 - \Phi_0(x))$ is convex, then $T = \mathrm{id} - \frac{1}{2}\nabla\Phi_0$ is the optimal transport map between μ_0 and $T_{\#}\mu_0$.

It is easy to check that convexity of $x \mapsto \frac{1}{2}(\|x\|^2 - \Phi_0(x))$ is equivalent to c-concavity of Φ_0 in that case. At this point, the reader's choice between Rademacher's theorem (which still holds for more general costs c) or Alexandrov's theorem guarantees the differentiability almost-everywhere allowing to state the theorem.

This result was later generalized to Riemannian manifolds by McCann [35] (see Section 2.3) and to the Hellinger–Kantorovich distances, on $\mathbb{R}^n$ by Liero et al. [31] and on Riemannian manifolds by Gallouët et al. [23] (we give a further extension of this last result in Section 3.3).

Finally, when $X \subset \mathbb{R}^n$ is (geodesically) convex, then $(\mathcal{P}(X), \mathrm{W}_2)$ is also a geodesic space. Geodesics in the latter space are obtained by lifting geodesics from the base space X as formalized in the next result.

Proposition 2.5. *Under the assumptions of Theorem 2.4, for $\mu_0 \in \mathcal{P}_{\mathrm{Vol}}(K)$, $\mu_1 \in \mathcal{P}(X)$, let T be the corresponding optimal transport map. Then the curve*

$$[0,1] \ni t \mapsto \mu_t := ((1-t)\cdot \mathrm{id} + t \cdot T)_{\#}\mu_0 \in \mathcal{P}(X) \quad (2.3)$$

is the unique geodesic in $(\mathcal{P}(X), \mathrm{W}_2)$ connecting μ_0 to μ_1.

Again, this result will readily generalize to Riemannian manifolds (Section 2.3).

2.2 Riemannian structure and linearization

The space $(\mathcal{P}(X \subset \mathbb{R}^n), \mathrm{W}_2)$ exhibits a pseudo-Riemannian structure which can be intuited from the celebrated Benamou–Brenier formula, giving the 2-Wasserstein distance as the length of a path in the Wasserstein space:

$$\mathrm{W}_2^2(\mu_0, \mu_1) = \inf_{(\mu_t, v_t)_{t\in[0,1]}} \int_0^1 \int_X \|v_t(x)\|^2 \mathrm{d}\mu_t(x)\,\mathrm{d}t \tag{2.4}$$

where the infimum is taken over solutions (in the sense of distributions) of the continuity equation $\partial_t \mu_t + \mathrm{div}(v_t \cdot \mu_t) = 0$ on $t \in [0,1]$ with $\mu_t \geq 0$, and boundary conditions $\mu_{t=0} = \mu_0$ and $\mu_{t=1} = \mu_1$.

Formula (2.4) may formally be interpreted as energy functional on the Riemannian manifold of probability measures $(\mathcal{P}(X), \mathrm{W}_2)$, where v_t represents an element of the tangent space at μ_t, the infinitesimal change of μ_t under v_t being encoded by the continuity equation. The Riemannian inner product at the point μ_t is given by the $\mathbb{L}^2(\mu_t)$-inner product, such that $\|v_t\|^2_{\mathbb{L}^2(\mu_t)} = \int_X \|v_t(x)\|^2 \mathrm{d}\mu_t$ can be interpreted as squared norm of v_t in the tangent space at μ_t.

Using the vocabulary of Riemannian geometry, the logarithmic map of $(\mathcal{P}(X), W_2)$ at support point $\mu_0 \in \mathcal{P}_{\mathrm{Vol}}(X)$ is then the map that associates to a measure $\mu_1 \in \mathcal{P}(X)$, the tangent vector of the constant speed geodesic from μ_0 to μ_1 at $t = 0$, which corresponds to the initial velocity field $v_{t=0}$ of a minimizer in (2.4). From (2.3), this velocity field actually turns out to be simply $T - \mathrm{id}$ where T is the corresponding optimal transport map from μ_0 to μ_1:

$$\mathrm{Log}^{\mathrm{W}_2}_{\mu_0} : \mathcal{P}(X) \to \mathbb{L}^2(X, \mathbb{R}^n; \mu_0), \qquad \mu_1 \mapsto T - \mathrm{id}, \tag{2.5}$$

where T is the optimal transport map from μ_0 to μ_1. We can then define the corresponding *exponential map* as

$$\mathrm{Exp}^{\mathrm{W}_2}_{\mu_0} : \mathbb{L}^2(X, TX = \mathbb{R}^n; \mu_0) \to \mathcal{P}(\mathbb{R}^n), \qquad v \mapsto (\mathrm{id} + v)_{\#}\mu_0. \tag{2.6}$$

We make two observations. First, note that we set the co-domain of $\mathrm{Exp}^{\mathrm{W}_2}_{\mu_0}$ to $\mathcal{P}(\mathbb{R}^n)$, since an arbitrary vector field v might easily push mass outside of X, if $x + v(x) \notin X$. Second, it is easy to see that $\mathrm{Exp}^{\mathrm{W}_2}_{\mu_0}$ is a left-inverse of $\mathrm{Log}^{\mathrm{W}_2}_{\mu_0}$. It is not a general inverse however, because while any measurable vector field v yields some image measure under $\mathrm{Exp}^{\mathrm{W}_2}_{\mu_0}$, the corresponding transport map $T = \mathrm{id} + v$ is in general not optimal, see Theorem 2.4. The logarithmic map will therefore then extract a different velocity field.

Finally, note that by Brenier's theorem (Theorem 2.4) the logarithmic map can be rewritten as

$$\mathrm{Log}^{\mathrm{W}_2}_{\mu_0}(\mu_1) = -\frac{1}{2}\nabla\Phi_0$$

where Φ_0 is a c-concave optimal Kantorovich potential in (2.2). This implies a better understanding of the range of the logarithmic map (see Section 4).

The Riemannian structure of W_2 has first been highlighted in [37], more details are given in [4]. As in conventional Riemannian geometry the logarithmic map can be used for local linearization of the Wasserstein distance. For a reference measure $\mu_0 \in \mathcal{P}(K)$ we can map a set of samples $(\nu_i)_{i=1}^M$ in $\mathcal{P}(X)$ to the tangent space of μ_0 via $v_i := \mathrm{Log}^{\mathrm{W}_2}_{\mu_0}(\nu_i)$, given by $\mathbb{L}^2(X, TX; \mu_0)$. Intuitively, one should have

$$\|v_i - v_j\|_{\mathbb{L}^2(\mu_0)} \approx \mathrm{W}_2(\nu_i, \nu_j)$$

as long as ν_i and ν_j are 'close' to μ_0. This approximation was proposed for (medical) image analysis in [46]. More applications and details are given, for instance, in [28, 39, 26]. Since the Riemannian structure of W_2 is merely formal, it is difficult to quantify what 'close' in the above sense means. Important contributions to that question are given in [25, 17].

2.3 Linearized 2-Wasserstein distance for manifolds

We now illustrate how the definition of logarithmic and exponential map for W_2 can be extended to the case where X is a connected complete Riemannian manifold. The key ingredient is a generalization of Brenier's theorem (Theorem 2.4) to manifolds due to McCann [35]. We will use the slightly extended version of [45, Theorem 10.41].

Theorem 2.6 (McCann)**.** *Let X be a n-dimensional Riemannian manifold and K a compact subset of X. Let $\mu_0 \in \mathcal{P}_{\mathrm{Vol}}(K)$, $\mu_1 \in \mathcal{P}(K)$. Furthermore, let Φ_0 be a maximizer of (2.2) for these measures. Then:*

(i) The minimizer π of (2.1) is unique and can be written as $\pi = (\mathrm{id}, T)_{\#}\mu_0$ for a corresponding optimal transport map T.

(ii) Φ_0 is differentiable μ_0-almost everywhere and the map T satisfies

$$T(x) = \mathrm{Exp}_x^X\left(-\frac{1}{2}\nabla\Phi_0(x)\right) \quad \mu_0(x)\text{-almost everywhere.} \tag{2.7}$$

Logarithmic and exponential maps (2.5) and (2.6) can then be extended in the natural way:

Definition 2.7 (Logarithmic and exponential map for W_2 over manifolds)**.** In the setting of the above Theorem we define the *logarithmic and exponential maps* at base point μ_0 as

$$\mathrm{Log}_{\mu_0}^{\mathrm{W}_2} : \mathcal{P}(X) \to \mathbb{L}^2(X, TX; \mu_0), \qquad \mu_1 \mapsto (x \mapsto \mathrm{Log}_x^X(T(x))) \tag{2.8}$$

with T the optimal transport map from μ_0 to μ_1, provided by Theorem 2.6, and

$$\mathrm{Exp}_{\mu_0}^{\mathrm{W}_2} : \mathbb{L}^2(X, TX; \mu_0) \to \mathcal{P}(X), \qquad v \mapsto [x \mapsto \mathrm{Exp}_x^X(v(x))]_{\#}\mu_0. \tag{2.9}$$

For $X \subset \mathbb{R}^n$ these formulas reduce again to (2.5) and (2.6). The following proposition shows that these maps do indeed mimic the behaviour of logarithmic and exponential maps in the Riemannian sense.

Proposition 2.8. *For $\mathrm{Log}_{\mu_0}^{\mathrm{W}_2}$ and $\mathrm{Exp}_{\mu_0}^{\mathrm{W}_2}$ as given in Definition 2.7 one has:*

(i) $\mathrm{W}_2^2(\mu_0, \mu_1) = \|v\|^2_{\mathbb{L}^2(\mu)}$ when $v = \mathrm{Log}_{\mu_0}^{\mathrm{W}_2}(\mu_1)$.

(ii) $\mathrm{Exp}_{\mu_0}^{\mathrm{W}^2}$ is a left-inverse of $\mathrm{Log}_{\mu_0}^{\mathrm{W}_2}$.

(iii) For the same v as in (i), $[0,1] \ni t \mapsto \mathrm{Exp}_{\mu_0}^{\mathrm{W}_2}(t \cdot v)$ yields the constant speed geodesic from μ_0 to μ_1.

Proof. (i) is an immediate consequence of the fact that the map T in the definition of $\mathrm{Log}_{\mu_0}^{\mathrm{W}_2}$ is an optimal transport map from μ_0 to μ_1 for the quadratic cost, and the properties of Log_x^X on X. One finds

$$\|v\|^2_{\mathbb{L}^2(\mu_0)} = \int_X \|\mathrm{Log}_{x_0}^X(T(x_0))\|^2_{x_0}\,\mathrm{d}\mu_0(x_0) = \int_X d(x_0, T(x_0))^2\,\mathrm{d}\mu_0(x_0) = \mathrm{W}_2^2(\mu_0, \mu_1).$$

Likewise, for (ii) the claim follows from

$$\mathrm{Exp}_x^X(v(x)) = \mathrm{Exp}_x^X(\mathrm{Log}_x^X(T(x))) = T(x).$$

(iii): For $t \in [0,1]$ let $p_t : X \times X \to X$ be a (measurable, also w.r.t. t) map that takes (x_0, x_1) to a point on a constant speed geodesic from x_0 to x_1 at time t. Let π be a minimizer in (2.1). Then $t \mapsto p_{t\#}\pi$ yields a geodesic between μ_0 and μ_1. See, for instance, [45, Section 7] for details. In our case, π is given by $(\mathrm{id}, T)_\# \mu_0$ and by our assumptions a map p_t as above can be given by $(x_0, x_1) \mapsto \mathrm{Exp}^X_{x_0}(t \cdot \mathrm{Log}^X_{x_0}(x_1))$. For these choices of π and p_t, $t \mapsto p_{t\#}\pi$ yields the expression in (iii) Uniqueness follows from uniqueness of π (Theorem 2.6) and essential uniqueness of p_t (Section 1.3). □

Finally, formula (2.7) of Theorem 2.6 yields again a dual expression for the logarithmic map, which will provide more insight into its range (Section 4).

Proposition 2.9. *In the setting of Theorem 2.6 one has* $\mathrm{Log}^{W_2}_{\mu_0}(\mu_1) = -\frac{1}{2}\nabla\Phi_0$ *where* Φ_0 *is a c-concave optimal Kantorovich potential in* (2.2).

3 Hellinger–Kantorovich distance

3.1 Basic properties

For a meaningful comparison of non-negative measures with varying mass various unbalanced transport models have been introduced. A prominent example is the Hellinger–Kantorovich distance. Similar to W_2 it can be formulated as a static and a dynamic optimization problem, and it exhibits a weak Riemannian structure. We refer to [31] and [15] for more details and further formulations. We start by recalling a formulation that is obtained by a relaxation of classical balanced transport, (1.2).

Definition 3.1 (Hellinger–Kantorovich distance)**.** For a parameter $\kappa > 0$ we set

$$\mathrm{HK}^2_\kappa(\mu_0, \mu_1) := \inf \left\{ E_\kappa(\pi|\mu_0, \mu_1) \middle| \pi \in \mathcal{M}_+(X \times X) \right\} \tag{3.1}$$

where

$$E_\kappa(\pi|\mu_0, \mu_1) := \int_{X\times X} c^{\mathrm{HK}}_\kappa(x_0, x_1)\mathrm{d}\pi(x_0, x_1) + \kappa^2 \cdot \mathrm{KL}(\mathrm{p}_{0\#}\,\pi|\mu_0) + \kappa^2 \cdot \mathrm{KL}(\mathrm{p}_{1\#}\,\pi|\mu_1) \tag{3.2}$$

with the cost

$$c^{\mathrm{HK}}_\kappa(x_0, x_1) := \begin{cases} -2\kappa^2 \log\left(\cos\left(\frac{d(x_0,x_1)}{\kappa}\right)\right) & \text{if } d(x_0, x_1) < \kappa\frac{\pi}{2}, \\ +\infty & \text{otherwise} \end{cases} \tag{3.3}$$

and the Kullback–Leibler divergence is defined as

$$\mathrm{KL}(\rho|\mu) := \begin{cases} \int_X \log\left(\frac{\mathrm{d}\rho}{\mathrm{d}\mu}\right)\mathrm{d}\rho - |\rho| + |\mu| & \text{if } \rho \ll \mu,\ \rho \geq 0, \mu \geq 0, \\ +\infty & \text{otherwise} \end{cases} \tag{3.4}$$

for $\rho, \mu \in \mathcal{M}(X)$.

By relaxing the marginal conditions $\mathrm{p}_{i\#}\,\pi = \mu_i$ this optimization problem remains meaningful even when measures of unequal mass are compared. It turns out that this particular relaxation in combination with the cost function c^{HK}_κ indeed yields a metric on $\mathcal{M}_+(X)$. We collect several fundamental properties in the following proposition.

Proposition 3.2.

(i) HK_κ *is a metric on* $\mathcal{M}_+(X)$.

(ii) Minimizing π *in* (3.1) *exist. All minimizers have the same marginals* $\mathrm{p}_{i\#}\,\pi$, $i = 0, 1$.

(iii) Any minimizing π *is an optimal transport plan for the balanced OT problem with cost* c_κ^{HK}, *between its marginals* $p_{0\#}\pi$, $p_{1\#}\pi$ *in the sense of* (1.2).

Proof. (i) is one of the key results of [31], see in particular [31, Theorem 7.20]. Existence in (ii) is established in [31, Theorem 6.2], uniqueness of the marginals in [31, Theorem 6.6]. (iii) is a trivial observation: if it were not true, the alternative minimizer $\tilde{\pi}$ for (1.2) (with the same marginals as π) would also yield a better objective in (3.1). □

Remark 3.3. Note that for a parameter $\kappa > 0$, an optimal π for (3.1) will not transport as far or further than distance $\kappa \cdot \pi/2$. The choice of κ therefore controls the trade-off between transport and mass creation/annihilation in (3.1). [31] focuses on the case $\kappa = 1$, the other cases can be described by re-scaling the metric on X. See also [11, Remark 3.3].

Remark 3.4 (Mass transport and generation)**.** Let π be an optimal plan for the soft-marginal formulation (3.1) of $\mathrm{HK}_\kappa^2(\mu_0, \mu_1)$. We will denote its marginals by $\pi_0 := p_{0\#}\pi$ and $\pi_1 := p_{1\#}\pi$. π_0 and π_1 do not depend on the choice of the minimizer π (Proposition 3.2 (ii)). We can then consider the Lebesgue decompositions of μ_i with respect to the π_i, i.e.

$$\mu_0 = \frac{\mathrm{d}\mu_0}{\mathrm{d}\pi_0}\pi_0 + \mu_0^\perp, \qquad\qquad \mu_1 = \frac{\mathrm{d}\mu_1}{\mathrm{d}\pi_1}\pi_1 + \mu_1^\perp,$$

where $\mu_i^\perp$ is the part of μ_i that is singular with respect to π_i. The parts of μ_i that are dominated by π_i should be interpreted as undergoing transport (and gradual mass change along the transport), whereas the singular parts $\mu_i^\perp$ undergo complete destruction ($i = 0$) and creation ($i = 1$) respectively, without any movement. This is illustrated in more detail in [11, Section 3.3]. If possible, the HK_κ metric always prefers transport over pure destruction/creation, i.e. $d(x_0, x_1) \geq \kappa\frac{\pi}{2}$ for $\mu_0^\perp \times \mu_1$-a.e. (x_0, x_1) (and symmetrically, exchanging the roles of μ_0 and μ_1), see [11, Lemma 3.13].

Duality for HK_κ^2, (3.1), is slightly more involved than for W_2, since the cost c_κ^{HK} can become $+\infty$ at finite range, and due to the non-linearity of the KL-terms. But (3.1) is still a convex optimization problem and with Fenchel–Rockafellar duality one can obtain, for instance, the following dual problem.

Proposition 3.5. *For* $\mu_0, \mu_1 \in \mathcal{M}_+(X)$,

$$\mathrm{HK}_\kappa^2(\mu_0, \mu_1) = \kappa^2 \sup\left\{ \int_X 1 - e^{-\phi_0(x_0)/\kappa^2}\mathrm{d}\mu_0(x_0) + \int_X 1 - e^{-\phi_1(x_1)/\kappa^2}\mathrm{d}\mu_1(x_1) \;\middle|\; (\phi_0, \phi_1) \in \mathrm{C}_{c_\kappa^{\mathrm{HK}}}(X) \right\} \quad (3.5)$$

with the definition of $\mathrm{C}_{c_\kappa^{\mathrm{HK}}}(X)$ *from balanced transport,* (1.3).

By a (monotonous) change of variables, $\Phi_i = \kappa^2\left(1 - e^{-\phi_i/\kappa^2}\right)$, this can be rewritten as a problem with linear objective, but somewhat more complicated constraints.

Proposition 3.6. *For any positive measures μ_0, μ_1 on X,*

$$\mathrm{HK}^2_\kappa(\mu_0,\mu_1) = \sup\left\{ \int_X \Phi_0(x_0)\mathrm{d}\mu_0(x_0) + \int_X \Phi_1(x_1)\mathrm{d}\mu_1(x_1) \;\middle|\; (\Phi_0,\Phi_1)\in \mathcal{Q}_\kappa(X)\right\} \tag{3.6}$$

where $\mathcal{Q}_\kappa(X)$ is now the (convex) set

$$\mathcal{Q}_\kappa(X) := \left\{ \Phi_0,\Phi_1 \in \mathrm{C}(X) \;\middle|\; \left(1-\frac{\Phi_0(x_0)}{\kappa^2}\right)\left(1-\frac{\Phi_1(x_1)}{\kappa^2}\right) \geq \mathrm{Cos}\left(\frac{d(x_0,x_1)}{\kappa}\right), \right.$$
$$\left. \Phi_i \leq \kappa^2,\ i=0,1 \right\}. \tag{3.7}$$

where we use Cos *to denote the 'truncated'* cos *function:* $\mathrm{Cos}(s) := \cos(\min\{|s|,\pi/2\})$.

Both duality formulas are covered by [31, Theorem 6.3]. For $X\subset\mathbb{R}^n$ the latter is also given in [15, Corollary 5.8]. We will focus on the formulation (3.6) in the following, as it has a more explicit connection with the logarithmic map for HK_κ (Proposition 3.17).

Remark 3.7. If $\operatorname{spt}\mu_i \subset K$ for some compact $K\subset X$ with $\operatorname{diam} K < \kappa\cdot\pi/2$, then c^{HK} is continuous on the compact set $K\times K$ and continuous maximizers of (3.5) can be shown to exist with the same arguments as for W_2, (2.2). (This does not hold for (3.6) if one of the two measures is zero.)

This can be generalized to the case where the transport distance of minimal π in (3.1) is bounded strictly away from $\kappa\pi/2$. A sufficient condition for this is that the distance between two points, in $\operatorname{spt}\mu_0$ and $\operatorname{spt}\mu_1$ respectively, has to be strictly less than $\kappa\pi/2$. This 'admissibility condition' was introduced in [23] where further regularity properties are deduced from this assumption. We do not use this assumption in our article.

In general, existence of dual maximizers can be established by relaxing the maximization space $\mathrm{C}(X)$. The following proposition is a subset of [31, Theorem 6.3] and will be usefull in order to state the existence of a transport map in our fairly general setting:

Proposition 3.8 (Existence of weak dual potentials for HK)**.** *For $\mu_0,\mu_1\in\mathcal{P}(X)$,*

$$\mathrm{HK}^2_\kappa(\mu_0,\mu_1) = \sup\left\{ \int_X \Phi_0\,\mathrm{d}\mu_0 + \int_X \Phi_1\,\mathrm{d}\mu_1 \middle| \Phi_0\in\mathbb{L}^1(X,\mu_0),\ \Phi_1\in\mathbb{L}^1(X,\mu_1),\right.$$
$$\left(1-\frac{\Phi_0(x_0)}{\kappa^2}\right)\left(1-\frac{\Phi_1(x_1)}{\kappa^2}\right) \geq \mathrm{Cos}^2\left(\frac{d(x_0,x_1)}{\kappa}\right),\ \Phi_i(x_i)\leq\kappa^2,\ i=0,1$$
$$\left. \text{for } \mu_0\otimes\mu_1\text{-a.e. } (x_0,x_1)\right\} \tag{3.8}$$

Maximizing (Φ_0,Φ_1) in (3.8) exist. A plan $\pi\in\mathcal{M}_+(X\times X)$ and a pair (Φ_0,Φ_1) are optimal in (3.1) and (3.8) respectively if and only if

(i) for π-almost every $(x_0,x_1)\in X^2$, $d(x_0,x_1)<\kappa\frac{\pi}{2}$ and

$$\left(1-\frac{\Phi_0(x_0)}{\kappa^2}\right)\left(1-\frac{\Phi_1(x_1)}{\kappa^2}\right) = \cos^2\left(\frac{d(x_0,x_1)}{\kappa}\right) \quad\text{and}\quad \Phi_i(x_i)<\kappa^2,\ i=0,1,$$

(ii) for $i=0,1$ and $(\mu_i^\perp\otimes\mu_{1-i})$-almost every (x_i,x_{1-i}),

$$d(x_i,x_{1-i})\geq\kappa\frac{\pi}{2},\quad\text{and}\quad \Phi_i(x_i)=\kappa^2,$$

(iii) the marginal densities of π verify

$$\frac{\mathrm{d}\pi_i}{\mathrm{d}\mu_i}(x_i) = 1 - \frac{\Phi_i(x_i)}{\kappa^2} \qquad \textit{for } \mu_i\textit{-almost every } x_i,\ i = 0, 1. \tag{3.9}$$

3.2 Transport maps for HK

Similar as for W_2, to state exponential and logarithmic maps for HK_κ, it will be helpful to establish the existence of optimal transport maps first. While existence of optimal transport maps for a particular class of convex cost functions has been established by [24] this does not apply to HK_κ directly, as c_κ^{HK} can become $+\infty$. For $X \subset \mathbb{R}^n$ existence of optimal transport maps is shown by [31, Theorem 6.6]. This was generalized to Riemannian manifolds in [23, Proposition 16] under the admissibility condition, see Remark 3.7. The loss of regularity of c_κ^{HK} requires to relax the notion of a gradient for the corresponding dual potentials, in order to obtain an expression coherent with the result by McCann. The two following definitions serve this purpose.

Definition 3.9 (Approximate limit)**.** For a Borel function f, we say that f admits an approximate limit l at x when Lebesgue-almost any sequence converges to l through f in the following sense:

$$\text{For any } \epsilon > 0, \text{ the set } \Delta_\epsilon = \{y \mid |f(y) - l| \leq \epsilon\} \text{ has Lebesgue-density 1.}$$

One can then state a notion of approximate differentiability for a function as admitting both a weak limit and a weak limit for its growth rate:

Definition 3.10 (Approximate gradient)**.** For a Borel function $f : X \to \mathbb{R}$, we say that f admits an approximate gradient $\tilde{\nabla} f(x) = u$ at $x \in X$ when

- f admits an approximate limit l at x.
- for any $\epsilon > 0$, the set

$$\Delta_\epsilon = \left\{ y \in X \;\middle|\; \mathrm{Log}_x^X \text{ defined at y, } \frac{\left|f(y) - l - u \cdot \mathrm{Log}_x^X(y)\right|}{\left\|\mathrm{Log}_x^X(y)\right\|} \leq \epsilon \right\} \tag{3.10}$$

  has density 1 at x.

One can show that this last definition is equivalent to finding a function g differentiable at the chosen point $x \in X$ such that the set $\{f = g\}$ has density 1 at x, and we shall use this reformulation for convenience in the upcomming proof, but we still give definition 3.10 as it is more concrete in our opinion. For a more extensive introductions to these weaker notions of limit and gradient (albeit only in $\mathbb{R}^d$), we refer the reader to [4, Section 5.5].

We now state a generalization of [23, Proposition 16] with some adaptations for our purposes: First of, we will write the transport map corresponding to an optimal plan in the sense of Definition 3.1 using the dual variables of (3.6) instead of these of (3.5) as they will be more natural in the context of the logarithmic map. Furthermore, and most importantly, we drop the admissibility assumption of [23] (see Remark 3.7), at the price that the transport map is only defined π_0-almost everywhere. However, since we are not concerned with regularity here, this is not an actual restriction, as we know that no transport can happen for the mass of μ_0 outside of $\mathrm{spt}(\pi_0)$ (only destruction or creation, see Remark 3.4).

Theorem 3.11. *Let $\mu_0 \in \mathcal{M}_{+,\mathrm{Vol}}(X)$, $\mu_1 \in \mathcal{M}_+(X)$ and Φ_0 be a maximizer of (3.8) for these measures. Then:*

(i) The minimizer π of (3.1) is unique and can be written as $\pi = (\mathrm{id}, T)_\# \pi_0$ for a corresponding optimal transport map T and π_0 is the first marginal of π.

(ii) The function Φ_0 is μ_0-almost everywhere approximately differentiable and the map T satisfies

$$T(x) = \mathrm{Exp}_x^X \left(-\kappa \arctan\left(\frac{1}{2\kappa} \frac{\left\| \tilde{\nabla} \Phi_0(x) \right\|}{1 - \Phi_0(x)/\kappa^2} \right) \frac{\tilde{\nabla} \Phi_0(x)}{\left\| \tilde{\nabla} \Phi_0(x) \right\|} \right) \quad \text{for } \pi_0\text{−almost every } x, \tag{3.11}$$

while $\tilde{\nabla} \Phi_0(x_0) = 0$ for $\mu_0^\perp$-a.e. $x_0 \in X$, where $\mu_0^\perp$ is the part of μ_0 that is singular with respect to π_0.

Proof. As mentioned above, the result and its proof are essentially minor adaptations of [23, Proposition 16] or [31, Theorem 6.6 (iii)]. For transparency we still give the full presentation here.

Let us consider a dual optimizer Φ_0 for (3.8). By the optimality conditions of proposition 3.8, there exist two sets $A_0 \subset \mathrm{spt}(\mu_0)$, $A_1 \subset \mathrm{spt}(\mu_1)$ such that $\mu_i(X \setminus A_i) = 0$, $i = 0, 1$ and

$$(1 - \frac{\Phi_0(x_0)}{\kappa^2})(1 - \frac{\Phi_1(x_1)}{\kappa^2}) \geq \mathrm{Cos}^2(d(x_0, x_1)/\kappa) \qquad \text{for every } (x_0, x_1) \in A_0 \times A_1, \tag{3.12a}$$

$$(1 - \frac{\Phi_0(x_0)}{\kappa^2})(1 - \frac{\Phi_1(x_1)}{\kappa^2}) = \cos^2(d(x_0, x_1)/\kappa) \qquad \text{for } \pi\text{-a.e. } (x_0, x_1) \in A_0 \times A_1, \tag{3.12b}$$

$$\Phi_i = \kappa^2 \qquad \text{on } \mathrm{spt}(\mu_i^\perp) \cap A_i, \ i = 0, 1. \tag{3.12c}$$

Let us first quickly notice that since Φ_0 is constant equal to κ^2 $\mu_0^\perp$-almost-everywhere on $\mathrm{spt}(\mu_0^\perp) \cap A_0$, it has approximate gradient 0 at each point of density 1 of this set, which is μ_0-almost every point since $\mu_0 \ll \mathrm{Vol}$.

We define the sub-level sets for $n \in \mathbb{N} \setminus \{0\}$,

$$A_{1,n} = \left\{ x_1 \in A_1 \;\middle|\; 1 - \frac{\Phi_1(x_1)}{\kappa^2} \geq \frac{1}{n} \right\}$$

as well as the (Lipschitz-continuous) function on X,

$$f_{0,n} : x_0 \mapsto \sup_{x_1 \in A_{1,n}} \frac{\mathrm{Cos}^2(d(x_0, x_1)/\kappa)}{1 - \frac{\Phi_1(x_1)}{\kappa^2}}.$$

It is immediate that $(1 - \frac{\Phi_0}{\kappa^2}) \geq f_{0,n}$ on A_0. Furthermore, on $\mathrm{spt}(\mu_0^\perp) \cap A_0$, both of these functions are identically 0 (because any x_0 in that set is at distance at least $\kappa \frac{\pi}{2}$ from any point in $A_{1,n}$). On the other hand, π_0-almost any x_0 is associated to some $x_1 \in A_{1,n}$ for some n, in the sense that $1 - \Phi_0(x_0)/\kappa^2 = \frac{\cos^2(d(x_0,x_1)/\kappa)}{1 - \Phi_1(x_1)/\kappa^2}$ (by (3.12b)). Therefore:

$$\pi_0 \left(X \setminus \bigcup_{n=1}^{\infty} \left\{ x_0 \in A_0 \;\middle|\; 1 - \frac{\Phi_0(x_0)}{\kappa^2} = f_{0,n}(x_0) \right\} \right) = 0.$$

Furthermore, since μ_0 and therefore π_0 are absolutely continuous w.r.t. the Lebesgue measure, denoting by A_0' the subset of all points $x \in A_0$ with Lebesgue density 1 and such that $f_{0,n}$ be

differentiable at x, Rademacher's theorem guarantees that π_0-almost any $x \in A_0$ is in A_0' and therefore also

$$\pi_0\left(X \setminus \bigcup_{n=1}^{\infty} A_{0,n}\right) = 0$$

with the definition

$$A_{0,n} := \left\{ x_0 \in A_0' \;\middle|\; 1 - \frac{\Phi_0(x_0)}{\kappa^2} = \frac{\mathrm{Cos}^2(d(x_0,x_1)/\kappa)}{1-\frac{\Phi_1(x_1)}{\kappa^2}} (= f_{0,n}(x_0)) \text{ for some } x_1 \in A_{1,n} \right\}$$

Now, any $x_0 \in A_{0,n}$ is a point of strong differentiability for $f_{0,n}$ and therefore of approximate differentiability of Φ_0. It follows that, for π_0-almost every $x_0 \in X$, there exists $n \in \mathbb{N} \setminus \{0\}$ s.t. $\tilde{\nabla}\Phi_0(x_0) = \nabla f_{0,n}(x_0)$.

Let us now compute the gradient of the Lipschitz function $f_{0,n}$ (where it is defined). Consider $x_0 \in A_{0,n}$ and $x_1 \in A_{1,n}$, such that

$$f_{0,n} = 1 - \frac{\Phi_0(x_0)}{\kappa^2} = \frac{\cos^2(d(x_0,x_1)/\kappa)}{1-\frac{\Phi_1(x_1)}{\kappa^2}}.$$

But for any other $x \in X$, $\kappa^2 - \Phi_1(x_1) \geq \frac{\cos^2(d(x,x_1))}{f_{0,n}(x)} =: g_n(x)$ (by definition of $f_{0,n}$) and therefore, x_0 is a maximum of g_n i.e. $0 \in \nabla^+ g_n(x_0)$ where $\nabla^+ g_n$ denotes the supergradient of the function g_n. Let us recall that this function is also sub-differentiable at x_0 giving us that it actually is differentiable at that point, with gradient 0. By hypothesis, $f_{0,n}$ is differentiable and non-zero at x_0 and $\mathrm{Cos}^2(d(.,x_1))$ is sub-differentiable (this is straightforward by super-differentiability of the squared distance on X, see [23, Lemma 15]), therefore g_n is indeed sub-differentiable and therefore differentiable at x_0 (and its gradient is its only super-gradient, 0). This implies, using the chain rule, that $d^2(.,x_1)$ is also differentiable at x_0, with gradient $-2\,\mathrm{Log}_{x_0}^X(x_1)$ and:

$$0 = \frac{2}{\kappa} \frac{\sin\left(\frac{d(x_0,x_1)}{\kappa}\right)\cos\left(\frac{d(x_0,x_1)}{\kappa}\right)}{f_{0,n}(x_0)} \frac{\mathrm{Log}_{x_0}^X(x_1)}{\left\|\mathrm{Log}_{x_0}^X(x_1)\right\|} - \frac{\cos^2\left(\frac{d(x_0,x_1)}{\kappa}\right)}{f_{0,n}(x_0)^2} \nabla f_{0,n}(x_0)$$

But since $x_0 \in A_{0,n}$,

$$\tilde{\nabla}\Phi_0(x_0) = -\kappa^2 \nabla f_{0,n}(x_0) = -2\kappa \tan\left(\frac{d(x_0,x_1)}{\kappa}\right) f_{0,n}(x_0) \frac{\mathrm{Log}_{x_0}^X(x_1)}{\left\|\mathrm{Log}_{x_0}^X(x_1)\right\|}$$

and using the fact that $d(x_0,x_1) < \kappa\frac{\pi}{2}$, we can conclude

$$\mathrm{Log}_{x_0}(x_1) = -\kappa \arctan\left(\frac{1}{2\kappa} \frac{\left\|\tilde{\nabla}\Phi_0(x_0)\right\|}{1-\Phi_0(x_0)/\kappa^2}\right) \frac{\tilde{\nabla}\Phi_0(x_0)}{\left\|\tilde{\nabla}\Phi_0(x_0)\right\|}$$

and x_1 is in fact uniquely determined by x_0 (and the dual solution Φ_0). Any optimal transport plan π is therefore supported on the graph of the map T given in the statement and, since the marginal densities $1-\Phi_0/\kappa^2$, $1-\Phi_1/\kappa^2$ (Proposition 3.8 (iii)) do not depend on the chosen optimal transport plan, by strict convexity of the KL divergence w.r.t. its first argument (see Proposition 3.2 (ii)), the map T and therefore also the optimal plan π are unique as well. □

Remark 3.12. Under the admissibility assumption of [23], the function Φ_0 is actually Lipschitz-continuous and one can then replace the approximate gradients by regular gradients (still almost everywhere) and the map T is then defined μ_0-almost everywhere, see also Remark 3.7.

3.3 Riemannian structure and linearization

Similar to W_2 the pseudo-Riemannian structure of $(\mathcal{M}_+(X), \mathrm{HK}_\kappa)$ can be intuited from a dynamical formulation of the distance in the spirit of the Benamou–Brenier formulation,

$$\mathrm{HK}_\kappa^2(\mu_0,\mu_1) = \inf_{(\mu_t,v_t,\alpha_t)_{t\in[0,1]}} \int_0^1 \int_X \left[\|v_t(x)\|^2 + \frac{\kappa^2}{4}\alpha_t(x)^2\right] \mathrm{d}\mu_t(x)\,\mathrm{d}t \tag{3.13}$$

where this time the infimum is taken over (distributional) solutions of the continuity equation *with source term*,

$$\partial_t \mu_t + \operatorname{div}(v_t \cdot \mu_t) = \alpha_t \cdot \mu_t \tag{3.14}$$

on $t \in [0,1]$ with $\mu_t \geq 0$, again with boundary conditions $\mu_{t=0} = \mu_0$ and $\mu_{t=1} = \mu_1$. We refer to [29, 13, 31] for more details on this formulation.

Similar to the 2-Wasserstein distance, (3.13) can be thought to induce a formal Riemannian structure on $(\mathcal{M}_+(X), \mathrm{HK}_\kappa)$, where tangent vectors now have an additional mass component α describing the rate of mass destroyed/created in order to go from μ_0 to μ_1. In the Euclidean case $X \subset \mathbb{R}^n$, a logarithmic and exponential map for this HK_κ metric space have been introduced in [11] and therefore, we will only give the extensions and basic properties in the case were X is a general Riemannian manifold to minimize redundancy.

Definition 3.13 (Logarithmic map)**.** Let $\mu_0 \in \mathcal{M}_{+,\mathrm{Vol}}(X)$ and $\mu_1 \in \mathcal{M}_+(X)$, and the optimal transport plan minimizing (3.1), $\pi = (\mathrm{id}, T)_\# \pi_0$ (uniqueness and existence of T provided by Theorem 3.11), let $\pi_1 := T_\# \pi_0 = \mathrm{p}_{1\#}\,\pi$. Consider the following Lebesgue decompositions of μ_0 and μ_1,

$$\mu_0 = \tfrac{\mathrm{d}\mu_0}{\mathrm{d}\pi_0}\pi_0 + \mu_0^\perp, \qquad\qquad \mu_1 = \tfrac{\mathrm{d}\mu_1}{\mathrm{d}\pi_1}\pi_1 + \mu_1^\perp.$$

Then we define the logarithmic map of μ_1 at μ_0 as $\mathrm{Log}_{\mu_0}^{\mathrm{HK}_\kappa}(\mu_1) = (v_0, \alpha_0, \mu_1^\perp)$, where

$$v_0(x_0) = \begin{cases} \kappa \frac{\mathrm{d}\pi_0}{\mathrm{d}\mu_0}(x_0) \tan\left(\frac{\left\|\mathrm{Log}_{x_0}(T(x_0))\right\|}{\kappa}\right) \frac{\mathrm{Log}_{x_0}(T(x_0))}{\left\|\mathrm{Log}_{x_0}(T(x_0))\right\|} & \text{for } \pi_0\text{-a.e. } x_0 \\ 0 & \text{for } \mu_0^\perp\text{-a.e. } x_0 \end{cases} \tag{3.15}$$

$$\alpha_0(x_0) = \begin{cases} 2\left(\frac{\mathrm{d}\pi_0}{\mathrm{d}\mu_0}(x_0) - 1\right) & \text{for } \pi_0\text{-a.e. } x_0 \\ -2 & \text{for } \mu_0^\perp\text{-a.e. } x_0 \end{cases} \tag{3.16}$$

These tangent components live in the set of *feasible tangent vectors for* μ_0:

$$\mathrm{TFeas}_\kappa(\mu_0) := \Big\{(v,\alpha,\mu^\perp) \in \mathbb{L}^2(X, TX; \mu_0) \times \mathbb{L}^2(X;\mu_0) \times \mathcal{M}_+(X) \;\Big|\; \alpha \geq -2 \ \mu_0\text{-a.e.},\ d(x_0,x_1) \geq \kappa\frac{\pi}{2} \text{ for } \mu_0 \otimes \mu^\perp\text{-a.e. } (x_0,x_1)\Big\} \tag{3.17}$$

Remark 3.14. Relative to [11, Proposition 4.1, Definition 4.5] the above expressions have been adjusted to the Riemannian context, simplified by using optimality conditions from Proposition 3.8 (which will be more useful in the present article), and the formal square root for the $\mu_1^\perp$ component has been omitted for simplicity as this component will only play a minor role in this article (see [11, Remark 4.6]).

The corresponding exponential map can once again be defined as a left-inverse of this logarithmic map. The following generalizes [11, Proposition 4.8].

Proposition 3.15. *For $\kappa > 0$, $\mu_0 \in \mathcal{M}_+(X)$, define the* exponential map *at μ_0 for HK_κ as*

$$\mathrm{Exp}_{\mu_0}^{\mathrm{HK}_\kappa} : \mathrm{TFeas}_\kappa(\mu_0) \ni (v, \alpha, \mu^\perp) \mapsto T_\# u^2 \mu_0 + \mu^\perp \tag{3.18}$$

where

$$T : x \mapsto \exp_x \left(\kappa \arctan\left(\frac{1}{\kappa} \frac{\|v(x)\|}{1 + \alpha(x)/2} \right) \frac{v(x)}{\|v(x)\|} \right) \tag{3.19}$$

$$u : x \mapsto \sqrt{\left(1 + \frac{\alpha(x)}{2}\right)^2 + \frac{1}{\kappa^2} \|v(x)\|^2} \tag{3.20}$$

Then $\mathrm{Exp}_{\mu_0}^{\mathrm{HK}_\kappa}$ is a left inverse of $\mathrm{Log}_{\mu_0}^{\mathrm{HK}_\kappa}$,

$$\mathrm{Exp}_{\mu_0}^{\mathrm{HK}_\kappa} \circ \mathrm{Log}_{\mu_0}^{\mathrm{HK}_\kappa} = \mathrm{id}\,.$$

Next, we give the expression for $\mathrm{HK}_\kappa^2(\mu_0, \mu_1)$ in terms of $\mathrm{Log}_{\mu_0}^{\mathrm{HK}_\kappa}(\mu_1)$, generalizing [11, Proposition 4.1, (4.4)]. As expected in Riemannian geometry, this is a quadratic form in v and α, but since we discarded the formal square root for the $\mu_1^\perp$ component (see Remark 3.14) it is 1-homogeneous in this component.

Proposition 3.16. *Let $\mu_0 \in \mathcal{M}_{+,\mathrm{Vol}}(X)$, $\mu_1 \in \mathcal{M}_+(X)$ and let $(v_0, \alpha_0, \mu_1^\perp) = \mathrm{Log}_{\mu_0}^{\mathrm{HK}_\kappa}(\mu_1)$. Then,*

$$\mathrm{HK}_\kappa^2(\mu_0, \mu_1) = \|v_0\|^2_{\mathbb{L}^2(\mu_0)} + \frac{\kappa^2}{4} \|\alpha_0\|^2_{\mathbb{L}^2(\mu_0)} + \kappa^2 \left|\mu_1^\perp\right|. \tag{3.21}$$

Proof. For the optimal transport plan π for the soft-marginal formulation, one has the following expression for HK^2 as a cumulated mass discrepancy (see for instance [31, Theorem 6.3, eq. (6.20)]):

$$\mathrm{HK}_\kappa^2(\mu_0, \mu_1) = \kappa^2 \left(|\mu_0| + |\mu_1| - 2\,|\pi| \right)$$

Now we evaluate the right-hand side of (3.21) (using the optimality conditions from Proposition 3.8)

$$\begin{aligned}
\int_X \|v_0(x_0)\|^2 \,\mathrm{d}\mu_0(x_0) &= \kappa^2 \int_X \tfrac{\mathrm{d}\pi_0}{\mathrm{d}\mu_0}(x_0) \frac{\sin^2\left(\frac{d(x_0, T(x_0))}{\kappa^2}\right)}{\cos^2\left(\frac{d(x_0, T(x_0))}{\kappa^2}\right)} \mathrm{d}\pi_0(x_0) \\
&= \kappa^2 \int_{X\times X} \tfrac{\mathrm{d}\mu_1}{\mathrm{d}\pi_1}(x_1) \sin^2\left(\frac{d(x_0, x_1)}{\kappa}\right) \mathrm{d}\pi(x_0, x_1) \\
\tfrac{1}{4}\int_X \alpha_0(x_0)^2 \mathrm{d}\mu_0(x_0) &= \int_X \left(\tfrac{\mathrm{d}\pi_0}{\mathrm{d}\mu_0} - 1\right)^2 \mathrm{d}\mu_0 = \int_X \left(1 - 2\tfrac{\mathrm{d}\pi_0}{\mathrm{d}\mu_0} + \left(\tfrac{\mathrm{d}\pi_0}{\mathrm{d}\mu_0}\right)^2\right) \mathrm{d}\mu_0 \\
&= |\mu_0| - 2\,|\pi| + \int_X \tfrac{\mathrm{d}\mu_1}{\mathrm{d}\pi_1}(x_1) \cos^2\left(\frac{d(x_0, x_1)}{\kappa}\right) \mathrm{d}\pi(x_0, x_1)
\end{aligned}$$

and therefore,

$$\begin{aligned}
&\|v_0\|^2_{\mathbb{L}^2(\mu_0)} + \frac{\kappa^2}{4} \|\alpha_0\|^2_{\mathbb{L}^2(\mu_0)} + \kappa^2 \left|\mu_1^\perp\right| \\
&\qquad = \kappa^2 \left(|\mu_0| - 2\,|\pi| + \int_X \tfrac{\mathrm{d}\mu_1}{\mathrm{d}\pi_1}(x_1) \mathrm{d}\pi(x_0, x_1) + \left|\mu_1^\perp\right| \right) = \mathrm{HK}_\kappa^2(\mu_0, \mu_1)
\end{aligned}$$

and we obtain the claim. □

Finally, the expression for the optimal transport map T, in the style of McCann using the dual potential, (3.11) allows to give simpler expressions for the logarithmic map in terms of the dual potential. This form of the logarithmic map is closely related to the polar factorization result for HK given in [23, Theorem 18].

Proposition 3.17. *For $\mu_0 \in \mathcal{M}_{+,\mathrm{Vol}}(X)$ and $\mu_1 \in \mathcal{M}_+(X)$, let (Φ_0, Φ_1) be optimal dual solutions of (3.8). Then the components $(v_0, \alpha_0, \mu_1^\perp) = \mathrm{Log}_{\mu_0}^{\mathrm{HK}_\kappa}(\mu_1)$ of the logarithmic map (definition 3.13) can be written as*

$$v_0 = -\frac{\tilde{\nabla}\Phi_0}{2}, \qquad \alpha_0 = -\frac{2}{\kappa^2}\Phi_0, \qquad \mu_1^\perp = \mu_1 \llcorner \{\Phi_1 = \kappa^2\}. \tag{3.22}$$

Proof. We plug the expression for the optimal transport map T from theorem 3.11 into the definition of v_0 and α_0 of Definition 3.13 and find for π_0-almost every x,

$$\tan\left(\frac{\|\mathrm{Log}_x(T(x))\|}{\kappa}\right)\frac{\mathrm{Log}_x(T(x))}{\|\mathrm{Log}_x(T(x))\|} = -\frac{1}{2\kappa}\frac{\tilde{\nabla}\Phi_0(x)}{1-\Phi_0(x)/\kappa^2},$$

and therefore for π_0-a.e. x,

$$\begin{aligned} v_0(x) &= \kappa\frac{\mathrm{Log}_x(T(x))}{\|\mathrm{Log}_x(T(x))\|}\frac{\mathrm{d}\pi_0}{\mathrm{d}\mu_0}(x)\tan\left(\frac{\|\mathrm{Log}_x(T(x))\|}{\kappa}\right) \\ &= \kappa\frac{\mathrm{Log}_x(T(x))}{\|\mathrm{Log}_x(T(x))\|}(1-\Phi_0(x)/\kappa^2)\tan\left(\frac{\|\mathrm{Log}_x(T(x))\|}{\kappa}\right) \\ &= \kappa(1-\Phi_0(x)/\kappa^2)\left(-\frac{1}{2\kappa}\frac{\tilde{\nabla}\Phi_0(x)}{1-\Phi_0(x)/\kappa^2}\right) = -\frac{1}{2}\tilde{\nabla}\Phi_0(x) \end{aligned}$$

and

$$\alpha_0(x) = 2(1-\frac{\Phi_0(x)}{\kappa^2}-1) = -\frac{2}{\kappa^2}\Phi_0(x).$$

On the other hand, for $\mu_0^\perp$-a.e. $x \in X$, $\Phi_0(x) = \kappa^2$ so that $\tilde{\nabla}\Phi_0(x) = 0$ (up to a $\mu_0^\perp$-negligible set of points) and so the expressions for $v_0(x)$ and $\alpha_0(x)$ are still valid.

Finally, by the optimality conditions of proposition 3.8 (ii), $\Phi_1 < \kappa^2$ π_1-a.e., whereas $\Phi_1 = \kappa^2$ $\mu_1^\perp$-a.e. which yields the expression for $\mu_1^\perp$. □

Remark 3.18 (Relation to Hamilton–Jacobi equation)**.** Formally the dual problem to (3.13) is given by

$$\sup_{\phi:[0,1]\times X\to\mathbb{R}} \int_X \phi(1,\cdot)\mathrm{d}\mu_1 - \int_X \phi(0,\cdot)\mathrm{d}\mu_0$$

over sufficiently regular functions ϕ subject to the constraint

$$\partial_t\phi(t,x) + \frac{\|\nabla\phi(t,x)\|^2}{4} + \frac{\phi(t,x)^2}{\kappa^2} \le 0$$

and with formal primal-dual optimality conditions

$$v_t = \nabla\phi(t,\cdot)/2 \qquad \text{and} \qquad \alpha_t = 2\phi(t,\cdot)/\kappa^2. \tag{3.23}$$

This dynamic dual problem can subsequently be identified with the static dual (3.8) with the correspondence $(\Phi_0, \Phi_1) = (-\phi(0,\cdot), \phi(1,\cdot))$. Formulas (3.22) and (3.23) can then seen to be consistent. We refer to [31, Section 8.4] for more details on the dynamic dual perspective.

3.4 Linearized Spherical Hellinger–Kantorovich distance

A motivation for the introduction of the Hellinger–Kantorovich distance is its ability to compare measures of different total mass. For one, this allows the description of growth and destruction processes, but it also makes the metric more robust with respect to small mass fluctuations and measurement errors when comparing measures of (almost) identical mass. It turns out that the HK distance between measures of different mass can be reduced to the comparison of probability measures by a simple scaling formula. Let $\mu_0, \mu_1 \in \mathcal{P}(X)$, $m_0, m_1 \geq 0$, then [30, Theorem 3.3]

$$\mathrm{HK}_1(m_0 \cdot \mu_0, m_1 \cdot \mu_1)^2 = \sqrt{m_0 \cdot m_1}\,\mathrm{HK}_1(\mu_0, \mu_1)^2 + (\sqrt{m_0} - \sqrt{m_1})^2.$$

So the expected key advantage of HK over W_2 in data analysis applications may not be so much the ability to deal with global mass differences, but rather the robustness with respect to local fluctuations, i.e. being able to slightly increase mass in one area of the sample while decreasing it in another to allow for a better matching.

Of course, by restriction HK still induces a metric on the set of probability measures, but it will no longer be geodesic, since shortest paths in the full space $(\mathcal{M}_+(X), \mathrm{HK})$ between $\mu_0, \mu_1 \in \mathcal{P}(X)$ will have mass less than 1 for times $t \in (0, 1)$. In fact, one will have [30, Equation (1.1)]

$$M(t) := |\mu_t| = 1 - t(1-t)\,\mathrm{HK}_\kappa^2(\mu_0, \mu_1)/\kappa^2. \tag{3.24}$$

Thus, when working with HK_κ on probability measures, we will always see this bias in the proposed interpolations and it will also affect the linearized analysis.

To obtain a geodesic distance within $\mathcal{P}(X)$, one can restrict admissible paths in (3.13) to those with constant unit mass. That is, one still admits a growth field α_t that can locally create or destroy mass, but on average the mass must be preserved, i.e. one adds the constraint to the optimization problem that $\int_X \alpha_t \mathrm{d}\mu_t = 0$ for all t. The resulting metric is called the spherical Hellinger–Kantorovich distance (SHK_1), introduced in [30]. A key observation of [30] is that SHK_1 and HK_1 are even more intimately related than the preceding discussion might suggest. Indeed, $(\mathcal{M}_+(X), \mathrm{HK}_1)$ turns out to be a cone space over $(\mathcal{P}(X), \mathrm{SHK}_1)$. This means that This means that $\mathrm{SHK}(\mu_0, \mu_1)$ and $\mathrm{HK}(\mu_0, \mu_1)$ can be directly computed from each other, as can be their geodesics, by a simple geometric intuition, sketched in Figure 1.

Definition 3.19 (Spherical Hellinger-Kantorovich distance)**.** For $\mu_0, \mu_1 \in \mathcal{P}(X)$, $\kappa > 0$, the spherical Hellinger–Kantorovich distance between μ_0 and μ_1 (with scale κ) is defined as

$$\mathrm{SHK}_\kappa(\mu_0, \mu_1) := \kappa \cdot \arccos\left(1 - \frac{\mathrm{HK}_\kappa(\mu, \nu)^2}{2\kappa^2}\right). \tag{3.25}$$

For $\kappa = 1$ this formula is given by [30, Equation (1.1)] for $\alpha = 1$, $\beta = 4$. The proper generalization to $\kappa \neq 1$ can be deduced from a simple scaling argument, see [11, Remark 3.3].

Moreover, if $[0, 1] \ni t \mapsto \mu_t \in \mathcal{M}_+(X)$ is a constant speed geodesic between μ_0 and μ_1 for HK_κ, then a constant speed geodesic $(\mu_t^S)_{t\in[0,1]}$ in $(\mathcal{P}(X), \mathrm{SHK}_\kappa)$ can be obtained by re-normalizing to unit mass and a reparametrization to account for the discrepancy between the arc and the chord in Figure 1. One finds [30, Theorem 2.7],

$$\mu_t^S := \mu_{s(t)}/M(s(t))$$

with $M(s) := |\mu_s| = 1 - s(1-s)\,\mathrm{HK}_\kappa^2(\mu_0, \mu_1)/\kappa^2$ as in (3.24) and

$$s(t) := \frac{\sin(t\,\mathrm{SHK}_\kappa(\mu_0, \mu_1)/\kappa)}{\sin((1-t)\,\mathrm{SHK}_\kappa(\mu_0, \mu_1)/\kappa) + \sin(t\,\mathrm{SHK}_\kappa(\mu_0, \mu_1)/\kappa)} \tag{3.26}$$

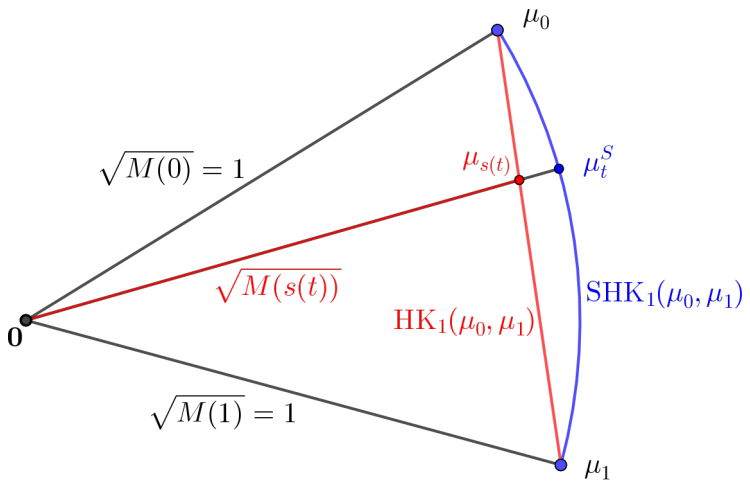


Figure 1: Geometric intuition for the cone structure and relation between HK_1 and SHK_1. See text for definition of symbols.

The formula for s can be deduced from Figure 1, it is also a special case of [30, Theorem 2.7].

The reparametrization and rescaling of μ_t to μ_t^S imply corresponding transformations on the velocity and mass change fields. Let $(v_t, \alpha_t)_{t\in[0,1]}$ be the fields for the HK_κ-geodesic. Then the ones for the SHK_κ-geodesic are given by

$$v_t^S := v_{s(t)} \cdot s'(t), \qquad \alpha_t^S := (\alpha_{s(t)} - \overline{\alpha}_{s(t)}) \cdot s'(t), \qquad \overline{\alpha}_s := \tfrac{1}{M(s)} \int_X \alpha_s \mathrm{d}\mu_s = \tfrac{M'(s)}{M(s)} \tag{3.27}$$

where for $t \in [0;1]$, $\overline{\alpha}_{s(t)}$ is a constant shift in the growth field to maintain the total mass of μ_t^S constant. One can show that $(\mu_t^S, v_t^S, \alpha_t^S)_{t\in[0,1]}$ do indeed solve the continuity equation (3.14), and that plugging them into (3.13) yields the expected value $\mathrm{SHK}_\kappa^2(\mu_0, \mu_1)$, as the integral of a constant squared speed.

One can then introduce the corresponding notion of logarithmic map for SHK_κ, similar to HK_κ.

Definition 3.20 (Logarithmic map for SHK_κ)**.** Let $\mu_0 \in \mathcal{P}_{\mathrm{Vol}}(X)$ and $\mu_1 \in \mathcal{P}(X)$. Let $v_0, \alpha_0, \mu_1^\perp = \mathrm{Log}_{\mu_0}^{\mathrm{HK}_\kappa}(\mu_1)$. We define the Logarithmic map for SHK_κ at base point μ_0 as

$$\mathrm{Log}_{\mu_0}^{\mathrm{SHK}_\kappa}(\mu_1) := (v_0^S, \alpha_0^S, (\mu_1^S)^\perp)$$

with

$$v_0^S := v_0 \cdot s'(0), \qquad \alpha_0^S := (\alpha_0 + \mathrm{HK}_\kappa^2(\mu_0, \mu_1)/\kappa^2) \cdot s'(0), \qquad (\mu_1^S)^\perp = \mu_1^\perp \cdot s'(0)^2 \tag{3.28}$$

and

$$s'(0) = \frac{\mathrm{SHK}_\kappa(\mu_0, \mu_1)/\kappa}{\sin(\mathrm{SHK}_\kappa(\mu_0, \mu_1)/\kappa)}. \tag{3.29}$$

Expressions for v_0^S and α_0^S follow directly from (3.27). Evaluating s', (3.26) at $t = 0$ is a simple algebraic exercise. The value for $\overline{\alpha}_0$ is computed in the following Lemma 3.21. The appropriate rescaling of μ_1^S then becomes apparent, for example by requiring that the norm of $\mathrm{Log}_{\mu_0}^{\mathrm{SHK}_\kappa}(\mu_1)$ equals $\mathrm{SHK}_\kappa(\mu_0, \mu_1)$, as shown in Proposition 3.22 below.

Lemma 3.21. *Let $\mu_0 \in \mathcal{P}_{\mathrm{Vol}}(X)$, $\mu_1 \in \mathcal{P}(X)$ and $v_0, \alpha_0, \mu_1^\perp = \mathrm{Log}_{\mu_0}^{\mathrm{HK}_\kappa}(\mu_1)$.*
Then,

$$\bar{\alpha}_0 = \int_X \alpha_0 \mathrm{d}\mu_0 = -\frac{1}{\kappa^2}\,\mathrm{HK}_\kappa^2(\mu_0,\mu_1) = 2\left[\cos\left(\frac{\mathrm{SHK}_\kappa(\mu_0,\mu_1)}{\kappa}\right) - 1\right]$$

Proof. This is straightforward from the expressions in proposition 3.17 and the fact that μ_0 and μ_1 have same mass. Indeed, recall that the marginals of the optimal transport plan π defining the HK_κ distance are given by the dual potentials Φ_0, Φ_1 and also have same mass (that of π):

$$\int_X \mathrm{d}\pi_0 = \int_X \left(1 - \frac{\Phi_0}{\kappa^2}\right)\mathrm{d}\mu_0 = \int_X \left(1 - \frac{\Phi_1}{\kappa^2}\right)\mathrm{d}\mu_1 = \int_X \mathrm{d}\pi_1$$

Simplifying the middle equality (and recalling $|\mu_0| = |\mu_1| = 1$), we get

$$\mathrm{HK}_\kappa^2(\mu_0,\mu_1) = \int_X \Phi_0 \mathrm{d}\mu_0 + \int_X \Phi_1 \mathrm{d}\mu_1 = 2\int_X \Phi_0 \mathrm{d}\mu_0$$

and one can conclude by replacing Φ_0 by $-\kappa^2\alpha_0/2$. □

Proposition 3.22. *In the setting of Definition 3.20 one has*

$$\mathrm{SHK}_\kappa^2(\mu_0,\mu_1) = \left\|v_0^S\right\|^2_{\mathbb{L}^2(\mu_0)} + \frac{\kappa^2}{4}\left\|\alpha_0^S\right\|^2_{\mathbb{L}^2(\mu_0)} + \kappa^2\left|(\mu_1^S)^\perp\right|. \tag{3.30}$$

Proof. First observe that using $\overline{\alpha}_0 = \int_X \alpha_0 \mathrm{d}\mu_0$, one finds

$$\|\alpha_0^S\|^2_{\mathbb{L}^2(\mu_0)} = (s'(0))^2 \int_X (\alpha_0 - \overline{\alpha}_0)^2 \mathrm{d}\mu_0 = (s'(0))^2\left(\|\alpha_0\|^2_{\mathbb{L}^2(\mu_0)} - \overline{\alpha}_0^2\right)$$

where $\overline{\alpha}_0^2 = \mathrm{HK}_\kappa(\mu_0,\mu_1)^4/\kappa^4$, and consequently with Proposition 3.16, (3.29) and finally (3.25),

$$\begin{aligned}
&\left\|v_0^S\right\|^2_{\mathbb{L}^2(\mu_0)} + \frac{\kappa^2}{4}\left\|\alpha_0^S\right\|^2_{\mathbb{L}^2(\mu_0)} + \kappa^2\left|(\mu_1^S)^\perp\right| \\
&= (s'(0))^2 \cdot \left(\|v_0\|^2_{\mathbb{L}^2(\mu_0)} + \frac{\kappa^2}{4}\|\alpha_0\|^2_{\mathbb{L}^2(\mu_0)} - \frac{\mathrm{HK}_\kappa^4(\mu_0,\mu_1)}{4\kappa^2} + \kappa^2\left|\mu_1^\perp\right|\right) \\
&= \frac{\mathrm{SHK}_\kappa^2(\mu_0,\mu_1)}{\kappa^2\sin^2\left(\mathrm{SHK}_\kappa(\mu_0,\mu_1)/\kappa\right)}\left(\mathrm{HK}_\kappa^2(\mu_0,\mu_1) - \frac{\mathrm{HK}_\kappa^4(\mu_0,\mu_1)}{4\kappa^2}\right) \\
&= \frac{\mathrm{SHK}_\kappa^2(\mu_0,\mu_1)}{\kappa^2 - \kappa^2\left(1 - \frac{\mathrm{HK}_\kappa^2(\mu_0,\mu_1)}{2\kappa^2}\right)^2}\left(\mathrm{HK}_\kappa^2(\mu_0,\mu_1) - \frac{\mathrm{HK}_\kappa^4(\mu_0,\mu_1)}{4\kappa^2}\right) \\
&= \mathrm{SHK}_\kappa^2(\mu_0,\mu_1).
\end{aligned}$$

□

Remark 3.23 (Exponential map for SHK_κ)**.** The formulas that provide the logarithmic map for SHK_κ, based on that of HK_κ are invertible (using that (3.25) is also invertible). Consequently, using this inversion and then using the exponential map for HK_κ, one can define the exponential map for SHK_κ and by construction it is again a left-inverse of the logarithmic map, yielding

$$\mathrm{Exp}_{\mu_0}^{\mathrm{SHK}_\kappa} \circ \mathrm{Log}_{\mu_0}^{\mathrm{SHK}_\kappa} = \mathrm{id}\,.$$

We point out here that the slope $s'(0)$ did depend on the target measure μ_1 and therefore, should be computed from the arguments of the exponential, using (3.30).

Figure 2: Interpolation between two discrete measures by mapping to tangent space, linear interpolation, and application of the exponential map for W_2 on a sphere (left), and the paraboloid model of the hyperbolic space (right). In each case, $\nu_0 := \delta_{y_0}$ and $\nu_1 := \delta_{y_1}$ are indicated by red triangles. $\mu := \frac{1}{2}(\delta_{x_0} + \delta_{x_1})$ is indicated by the isolated cross and dot. The interpolated measures ν_t (for some $t \in [0;1]$) are indicated by a crosses and dots respectively, indicating from which x_i the corresponding mass was transported. One can see that in the case of the sphere, the interpolated transport is optimal between the initial measure and the target one, whereas for the hyperbolic space, it is not as the order of the Dirac masses is reversed.

4 Convexity of the range of logarithmic map

A natural question in the context of linearized optimal transport is whether or not taking averages of embeddings (or interpolating between them) in the tangent space yields another 'valid' tangent vector, i.e. a vector that corresponds to an embedding of a measure via the logarithmic map. In other words: Is the range of the logarithmic map convex?

4.1 For the 2-Wasserstein distance

For balanced W_2 over convex subsets $X \subset \mathbb{R}^n$ this can be answered positively by Brenier's theorem (Theorem 2.4). Optimal transport maps can be written as gradients of convex potentials. By (2.5), interpolating between tangent vectors corresponds to interpolating between the transport maps, and the set of convex functions is itself convex.

By Proposition 2.9 for general manifolds X convexity of the range of the logarithmic map is equivalent to the convexity of the set of c-concave functions. This is a non-trivial question, related to the Ma-Trudinger-Wang condition (in this case, MTW(0,0)). We refer to [45, Definition 12.27] for several formulations of this condition, as the definite positivity of the Ma-Trudinger-Wang tensor, and [20, Theorem 3.2] and [22, Theorem 1.5.4] for the fact that this condition yields the convexity of the set of c-concave functions. For the quadratic cost of the 2-Wasserstein distance, the Riemannian manifolds verifying MTW(0,0) are scarce, most notably the convex compact subsets of $\mathbb{R}^n$ with the Euclidean metric (consistent with the previous paragraph) and the spheres of $\mathbb{R}^n$ with their spherical metric. On the other hand, one can show that no manifold having negative sectional curvature somewhere verifies this condition (this last negative result is a consequence of Loeper's identity for the Ma-Trudinger-Wang tensor).

We illustrate this (non-)convexity result with a simple example. Let X be the 2-dimensional unit sphere, let $\mu := \frac{1}{2}\delta_{x_0} + \frac{1}{2}\delta_{x_1}$, $\nu_0 := \delta_{y_0}$ and $\nu_1 := \delta_{y_1}$ for points $x_0, x_1, y_0, y_1 \in X$. We will interpolate between ν_0 and ν_1 in the tangent space of μ. This means, we set $\nu_t := \mathrm{Exp}_\mu^{\mathrm{W}_2}(v_t)$

where $v_i := \operatorname{Log}_\mu^{W_2}(\nu_i)$ and $v_t := v_0 + t(v_1 - v_0)$ for $t \in [0;1]$. This is visualized in Figure 2, left. Since the sphere satisfies the MTW(0,0) condition, the interpolated v_t are indeed valid tangent vectors and thus the implied transport maps from μ to ν_t are optimal which is represented in the figure by the fact that the interpolated mass positions of ν_t associated to x_0 are closer to x_0 than to x_1 (and vice versa for the roles of x_0 and x_1 swapped). Figure 2, right, shows the same experiment for hyperbolic space with negative curvature (hence MTW(0,0) is violated) and indeed the order of the points is reversed by the exponential map and therefore v_t does not correspond to the optimal transport plan between μ and ν_t anymore (except for $t = 0, 1$) and the range of the logarithmic map is not convex in that case. Note that even in the sphere example, ν_t is not the Wasserstein geodesic between ν_0 and ν_1, but instead the generalized geodesic, with respect to the measure μ. The actual geodesic is obtained as a single traveling Dirac mass on the geodesic between y_0 and y_1.

4.2 For the HK and SHK distances

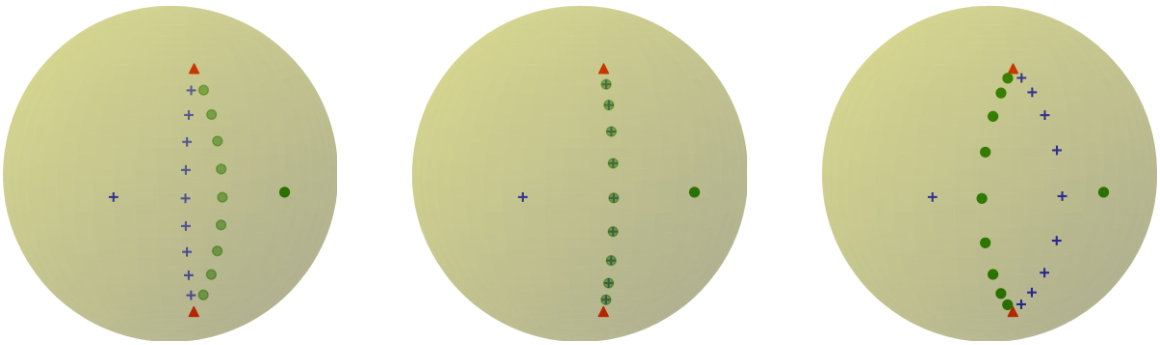


Figure 3: Same setup as in Figure 2 for HK_1 on spheres with radii $r = 0.1$,1 and 1.5 (left to right). For $r \leq 1$, the interpolation of tangent the tangent vectors lies in the range of the logarithmic map, for $r > 1$ it does not.

For the Hellinger–Kantorovich distance Proposition 3.17 shows that the convexity of the range of the logarithmic map is equivalent to the convexity of the set of optimal Φ_0 in (3.8) or (3.6). The dual constraint in (3.6) is non-linear, but as discussed in Section 3.1, it can be translated to a linear constraint as in (3.5) by a change of variables. So the appropriate notion of concavity is the following: We say that Φ is log-c_κ^{HK}-concave iff

$$\exists \Psi \in \mathrm{C}(X,]-\infty; 1[) : \ \Phi = 1 - \sup_{y \in X} \frac{\cos^2(d(\cdot, y)/\kappa)}{1 - \Psi(y)/\kappa^2} \quad \Longleftrightarrow \quad -\ln(1 - \Phi/\kappa^2) \text{ is } c_\kappa^{\mathrm{HK}}\text{-concave.}$$

For simplicity, here we have assumed that our measures are admissible in the sense of Gallouët et al., [23], so that Lipschitz-continuous solutions to (3.6) exist and the truncation of the cos function in the duality conditions can be ignored.

Gallouët et al. have shown a weaker MTW condition (weaker than MTW(0,0)) for c_κ^{HK}, implying in particular C^1-regularity of c_κ^{HK}-concave dual solutions to the squared HK_κ distance in some cases, and that this regularity, in turn, carries over to dual solutions to (3.6) (because these are obtained via composition with a smooth function under the admissibility assumption). However, even if one were to show MTW(0,0) for c_κ^{HK}, it would not directly translate to stability under convex combinations of admissible functions for the cone formulation (3.6) due to the non-linear change of variables.

Conversely, one can explicitly verify that the range of the logarithmic map is not convex in some settings, using the same setup of μ, ν_0 and ν_1 as in the previous section. This includes, for instance, the flat space $\mathbb{R}^n$ and spheres of radius strictly greater than 1. For the unit sphere convexity can be shown to hold (which stems from the fact that the cone over the unit sphere is locally isometric to flat Euclidean space, see [31, Section 7] for more details). For radii less than one numerical examples suggest that convexity is also satisfied. Some illustrations for spheres of different radii are given in Figure 3, in analogy to Figure 2. These observations seem to be consistent with the results on c_κ^{HK} for the sphere obtained in [23], albeit the precise relation is as of now unclear due to the aforementioned non-linear change of variables.

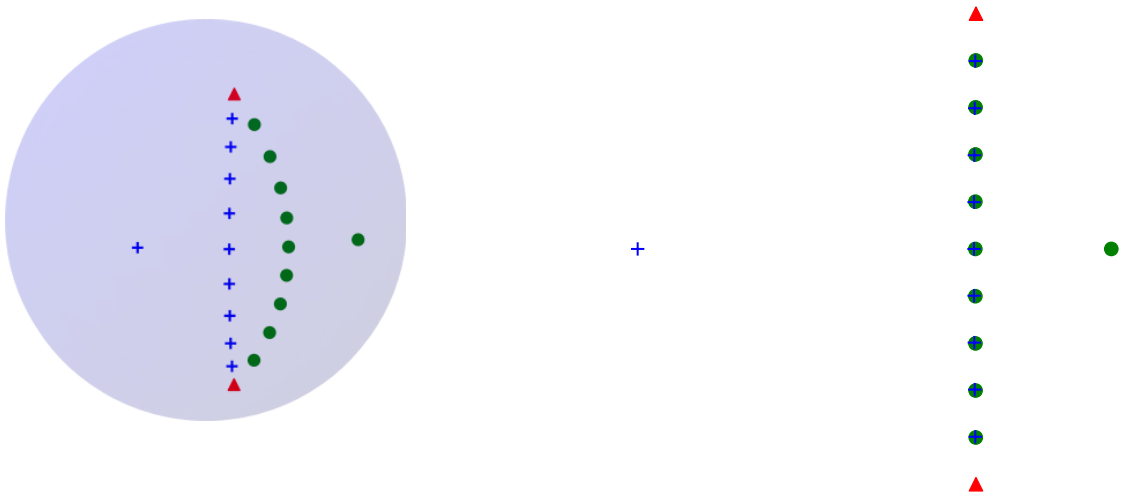


Figure 4: Same setup as in Figures 2 and 3 for SHK_1 on the sphere of radius 1.5 (left) and flat $\mathbb{R}^2$ (right). In both cases convexity of the range of the log map appears to be satisfied, with the interpolating measures being given by single Dirac masses in the flat case.

For the spherical Hellinger–Kantorovich distance the situation is even less clear due to the additional non-linear transformation to turn a HK tangent vector into the corresponding SHK tangent vector, see Definition 3.20. Preliminary numerical examples, see Figure 4, suggest that SHK is better behaved than HK in terms of convexity.

A deeper study of this convexity property for the HK and SHK distances will be subject of future work.

5 Asymptotic stability of the logarithmic maps

5.1 Convergence of the HK tangent structure to the W_2 one as $\kappa \to \infty$

It is well-known that, as $\kappa \to \infty$, the distance HK_κ converges to the 2-Wasserstein distance, see for instance [31, Theorem 7.24] and [15, Theorem 5.10]. A very natural 'sanity check' for our definition of their tangent structure is to verify whether or not the logarithmic map of HK_κ does also converge (in a suitable sense) to that of W_2.

To establish this result it is convenient to exploit the convergence of the optimal plans π_κ in (3.1) to a corresponding optimizer of (2.1) as $\kappa \to \infty$. To the best of our knowledge, this result is only implicit in previous statements of the convergence of the distances that are references above. We will therefore give a brief sketch here.

Lemma 5.1. *Let $\mu_0, \mu_1 \in \mathcal{M}_+(K)$ for some compact $K \subset X$ be probability measures and $(\kappa_n)_{n\in\mathbb{N}}$ be a sequence with $\kappa_n \to +\infty$ as $n \to \infty$. For $n \in \mathbb{N}$, let π_n be an optimal transport plan for $\mathrm{HK}^2_{\kappa_n}(\mu_0, \mu_1)$, minimizing (3.1).*

Then (up to subsequences) $(\pi_n)_{n\in\mathbb{N}}$ converges weakly-$\star$ to some π which is an optimal transport plan for $\mathrm{W}_2^2(\mu_0, \mu_1)$, minimizing (2.1). The optimal values also converge, i.e.

$$\lim_{n\to\infty} \mathrm{HK}^2_{\kappa_n}(\mu_0, \mu_1) = \mathrm{W}_2^2(\mu_0, \mu_1).$$

Proof. This is a simple Gamma-convergence result on the primal formulation of $\mathrm{HK}^2_{\kappa_n}(\mu_0, \mu_1)$ as $n \to \infty$ and we merely give a sketch.

Note that by compactness of K and for κ_n large enough, $c^{\mathrm{HK}}_{\kappa_n}$ is uniformly continuous on K^2 (since eventually $\mathrm{diam}(K) < \kappa_n \frac{\pi}{2}$) and converges uniformly to the quadratic cost $c = d^2$ of W_2. Hence, for any converging sequence $(\pi_n)_n$ with limit π, one has $\int_{K\times K} c^{\mathrm{HK}}_{\kappa_n} \mathrm{d}\pi_n \to \int_{K\times K} c \mathrm{d}\pi$.

For the liminf condition, in addition to the above convergence of the cost we need to consider the marginal constraints. When $\kappa_n^2\, \mathrm{KL}(\mathrm{p}_{i\#}\, \pi_n)$ remains bounded as $n \to \infty$, this implies that the limit π lies in $\Pi(\mu_0, \mu_1)$ as required. The limsup condition is readily verified by taking the constant sequence $\pi_n = \pi$ as recovery sequence. In this case the Kullback–Leibler terms always yield zero and the cost term converges, as discussed above.

By boundedness of $\mathrm{HK}^2_{\kappa_n}(\mu_0, \mu_1)$ we obtain that the sequence $(\pi_n)_n$ of optimal plans has bounded mass and tight marginals and is therefore weak* precompact. The previous Gamma-convergence then establishes that any cluster point π solves (2.1). □

Proposition 5.2. *Let $K \subset X$ be compact, let $\mu_0 \in \mathcal{P}_{\mathrm{Vol}}(K)$, $\mu_1 \in \mathcal{P}(K)$ and $(\kappa_n)_{n\in\mathbb{N}}$ be a positive sequence with $\kappa_n \to \infty$ as $n \to \infty$. Let $(v_0^N, \alpha_0^N, (\mu_1^N)^\perp) := \mathrm{Log}^{HK_{\kappa_n}}_{\mu_0}(\mu_1)$.*

Then, there will be some $N \in \mathbb{N}$ such that $(\mu_1^N)^\perp = 0$ for $n \geq N$. Furthermore, as $n \to \infty$, v_0^N converges strongly in $\mathbb{L}^2(X, \mathbb{R}^n, \mu_0)$ to $v_0 := \mathrm{Log}^{\mathrm{W}_2}_{\mu_0}(\mu_1)$ and α_0^N converges strongly in $\mathbb{L}^2(X; \mu_0)$ to 0.

Proof. Let us take $N \in \mathbb{N}$ large enough, such that $\kappa_n \pi/2 > \mathrm{diam}(K)$ for $n \geq N$. By Proposition 3.8 (ii) one has $d(x_0, x_1) \geq \kappa_N \pi/2$ for $\mu_0 \otimes (\mu_1^N)^\perp$ almost all (x_0, x_1). Therefore, for measurable $A \subset K$ one has

$$(\mu_1^N)^\perp(A) = \mu_0 \otimes (\mu_1^N)^\perp(K \times A) = \mu_0 \otimes (\mu_1^N)^\perp\left(\{(x_0, x_1) \in K \times A \,:\, d(x_0, x_1) \geq \kappa_n \pi/2\}\right) = 0$$

where the last equality follows because the considered subset of $K \times A$ is empty.

For N large enough, by Proposition 2.8 (i) and Proposition 3.16 (where we use $(\mu_1^N)^\perp = 0$) one has

$$\mathrm{W}_2^2(\mu_0, \mu_1) = \|v_0\|^2_{\mathbb{L}^2(X;\mu_0)} \quad \text{and} \quad \mathrm{HK}^2_{\kappa_N}(\mu_0, \mu_1) = \|v_0^N\|^2_{\mathbb{L}^2(X;\mu_0)} + \tfrac{\kappa_N^2}{4}\|\alpha_0^N\|^2_{\mathbb{L}^2(X;\mu_0)}$$

and since $\mathrm{HK}^2_{\kappa_N}(\mu_0, \mu_1) \uparrow \mathrm{W}_2^2(\mu_0, \mu_1)$ as $N \to \infty$ [31, Theorem 7.24], we deduce that $\alpha_N \to 0$ strongly in $\mathbb{L}^2(X; \mu_0)$ and that

$$\limsup_{N\to\infty} \|v_0^N\|^2_{\mathbb{L}^2(X;\mu_0)} \leq \|v_0\|^2_{\mathbb{L}^2(X;\mu_0)}. \tag{5.1}$$

We now show the strong convergence of v_0^N to v_0. For $n \geq N$, let π^n be optimal for the primal soft-marginal formulation (3.1) of HK_{κ_N}. By Theorem 3.11 there is a map T^N such that $\pi^N = (\mathrm{id}, T^N)_\# \pi_0^N$ where π_0^N is the first marginal of π^N. Analogously let $\pi = (\mathrm{id}, T)_\# \mu_0$ be the optimal transport plan for $\mathrm{W}_2^2(\mu_0, \mu_1)$ in (2.1), uniqueness and existence of the map T provided by Theorem 2.6. By Lemma 5.1, $\pi^N \xrightarrow[N\to\infty]{} \pi$. By the bound (5.1) the sequence $(v_0^N)_{N\in\mathbb{N}}$ must (up to selection of a subsequence) converge weakly in $\mathbb{L}^2(X;\mu_0)$ to some $\tilde{v}$. Below we will show that indeed $\tilde{v} = v_0$ (i.e. the whole sequence converges weakly to this unique cluster point). By weak lower-semicontinuity of the norm and the bound (5.1), we then get that $\lim_{n\to\infty} \|v_0^N\|_{\mathbb{L}^2(X;\mu_0)} = \|v_0\|_{\mathbb{L}^2(X;\mu_0)}$, which in combination with weak convergence will imply strong convergence of $v_0^N \to v_0$.

To show $\tilde{v} = v_0$, let $\xi \in \mathrm{C}(X)$. Then, recalling the expression for v_0^N from Definition 3.13,

$$
\begin{aligned}
&\int_X \xi(x_0)\cdot\tilde{v}(x_0)\mathrm{d}\mu_0(x_0) \\
&\qquad = \lim_{N\to\infty} \int_X \xi(x_0)\cdot v_0^N(x_0)\,\mathrm{d}\mu_0(x_0) \\
&\qquad = \lim_{N\to\infty} \int_{X\times X} \kappa_N \xi(x_0)\cdot \frac{\mathrm{Log}_{x_0}^X(T^N(x_0))}{\left\|\mathrm{Log}_{x_0}^X(T^N(x_0))\right\|} \tan\left(\frac{\left\|\mathrm{Log}_{x_0}^X(T^N(x_0))\right\|}{\kappa_N}\right) \frac{\mathrm{d}\pi_0^N}{\mathrm{d}\mu_0}(x_0)\mathrm{d}\mu_0(x_0) \\
&\qquad = \lim_{N\to\infty} \int_{X\times X} \xi(x_0)\cdot \mathrm{Log}_{x_0}^X(x_1) \frac{\tan\left(\frac{\left\|\mathrm{Log}_{x_0}^X(x_1)\right\|}{\kappa_N}\right)}{\left\|\mathrm{Log}_{x_0}^X(x_1)\right\|/\kappa_N} \mathrm{d}\pi^N(x_0,x_1) \\
&\qquad = \int_{X\times X} \xi(x_0)\cdot \mathrm{Log}_{x_0}^X(x_1)\,\mathrm{d}\pi(x_0,x_1) \\
&\qquad = \int_{X\times X} \xi(x_0)\cdot v_0(x_0)\,\mathrm{d}\mu_0(x_0)
\end{aligned}
$$

The second to last equality here would simply be the narrow convergence of π^N and uniform convergence of the integrated functions as $N \to 0$ if not for the fact that the logarithmic map $(x_0, x_1) \mapsto \mathrm{Log}_{x_0}^X(x_1)$ might not be continuous everywhere on $K \times K$. Fortunately, a byproduct of the proof of Theorem 2.6 is that for π-almost any (x_0, x_1), the squared distance function $d^2(., x_1)$ is differentiable at x_0. This is actually equivalent to the function d^2 being smooth at (x_0, x_1), and in particular, the logarithmic map being continuous at (x_0, x_1) (this is due to the link between the differentiability of the squared distance and the cut locus at x_1, see for instance [41, Proposition 4.8]). Furthermore, we have a trivial global (essential) bound on the norm of the Log_x^X map, namely $\mathrm{diam}(K)$. This is sufficient to apply the continuous mapping theorem (see for instance [44, Theorem 2.3]) to be able to replace the limit integral with the integral of the limit. Finally, by the density of continuous functions in $\mathbb{L}^2(X, TX; \mu_0)$, we have $\tilde{v} = v_0$ is the strong $\mathbb{L}^2(\mu_0)$ limit of $(v_0^N)_{N\in\mathbb{N}}$. □

5.2 Convergence of barycentric projection for Wasserstein distance

For the theoretical introduction of the logarithmic map for W_2, HK_κ and SHK_κ it was necessary to assume that the reference measure μ_0 had a Lebesgue density such that the optimal transport from μ_0 to the sample μ_1 is deterministic, i.e. induced by an optimal map (in fact, slightly weaker assumptions on μ_0 suffice, see e.g. [9]). Only then can the information about the transport be fully encoded in a velocity field (and an additional growth field for the unbalanced metrics). In

numerical practice however measures are often approximated by discretization and will then be concentrated on a finite number of points. The optimal transport plans are then in general not deterministic, i.e. no map T exists. On top of this, in practice one often uses entropic regularization for numerical approximation, which introduces additional non-determinism. A popular remedy for the linearized W_2 metric is 'barycentric projection' where the mean velocity of all particles leaving a point x_0 is used as approximation [46]. A similar heuristic has been proposed in [11] for the linearized HK_κ metric. In this section we show convergence of the approximation of [46] as the approximation of the marginal measures and transport plans improves, thus providing a rigorous foundation for their use. In the next section we show the analogous result for a slightly adjusted variant of [11]. As a corollary we gain some insight on the continuity (or lack thereof) of the logarithmic map for HK. Our results are not restricted to $X \subset \mathbb{R}^n$, but also apply in the manifold setting.

We now introduce a generalized notion of the logarithmic maps for W_2 and HK_κ that corresponds to the barycentric projection used as a stand-in for a transport map in [11], but can also be used to construct "tangent components" based on slightly sub-optimal transport plans (see Remark 5.20). As a consequence, we define them directly as a function of the transport plan instead of second measure μ_1. Since such a plan does not have to be deterministic, the expressions for the Logarithmic components will be obtained as averages, according to a disintegration of the sub-optimal plan w.r.t. its first marginal (the one corresponding to the base measure μ_0):

Definition 5.3. for any measure $\pi \in \mathcal{M}_+(X \times X)$ we denote by $(\pi(\cdot|x_0))_{x_0 \in X}$ its disintegration with respect to its first marginal $\pi_0 = \mathrm{p}_{0\#}\,\pi$, i.e. the π_0-a.e. unique family of measures verifying

$$\int_{X\times X} \phi(x_0, x_1)\,\mathrm{d}\pi(x_0, x_1) = \int_X \left[\int_X \phi(x_0, x_1)\,\mathrm{d}\pi(x_1|x_0)\right] \mathrm{d}\pi_0(x_0)$$

for any measurable test functions ϕ.

This is a specific case to a theorem generalizing the notion of Radon-Nikodim derivatives, see for instance [4, Theorem 5.3.1] for the general statement. We will simply point out that, for a deterministic plan $\pi = (\mathrm{id}, T)_\#\sigma$, one has $\pi(\cdot|x) = \delta_{T(x)}$ for σ-almost all $x \in X$. In the non-deterministic case $\pi(\cdot|x)$ is (σ-almost everywhere) a probability measure on X which gives us the distribution of targets of particles starting at x.

Definition 5.4 (Barycentric logarithmic map for W_2)**.** For $\mu_0 \in \mathcal{P}(K)$ and $\pi \in \mathcal{P}(K \times K)$ with $\mathrm{p}_{0\#}\,\pi = \mu_0$ and $(x, y) \mapsto \mathrm{Log}_x(y)$ defined $\pi - a.e.$, we set

$$\mathrm{Log}^{W_2}_{\mu_0}(\pi)(x) := \int_X \mathrm{Log}^X_x(y)\mathrm{d}\pi(y|x).$$

When $\mu_0 \in \mathcal{P}_{\mathrm{Vol}}(K)$ and π is an optimal plan for $W_2^2(\mu_0, \mu_1)$ to its second marginal $\mu_1 := \mathrm{p}_{1\#}$ then this definition reduces to Definition 2.7.

The next result then provides convergence of the barycentric projection approximation for the Wasserstein-2 case, when the approximated measure is absolutely continuous and as the discretization or approximation becomes increasingly refined.

Proposition 5.5. *Let $(\mu_0^N)_N$, $(\mu_1^N)_N$ be two sequences in $\mathcal{P}(K)$, converging weakly* to $\mu_0 \in \mathcal{P}_{\mathrm{Vol}}(K)$ and $\mu_1 \in \mathcal{P}(K)$ respectively. Let $(\pi^N)_N$ be a sequence with $\pi^N \in \Pi(\mu_0^N, \mu_1^N)$ such that $\pi^N \xrightarrow[N\to\infty]{} \pi$ where π is the optimal plan for $\mathrm{W}_2(\mu_0, \mu_1)$. Set $v_0^N := \mathrm{Log}^{\mathrm{W}_2}_{\mu_0^N}(\pi^N)$ as given by Definition 5.4 and $v_0 := \mathrm{Log}^{\mathrm{W}_2}_{\mu_0}(\pi) (= \mathrm{Log}^{\mathrm{W}_2}_{\mu_0}(\mu_1))$, using either Definition 5.4 or Definition 2.7. Then the momentum measures $v_0^N \mu_0^N$ converge weakly* to $v_0\mu_0$ as $N \to \infty$.*

Above the π^N could, for instance, be approximately optimal plans for $W_2(\mu_0^N, \mu_1^N)$ with vanishing sub-optimality as $N \to \infty$, e.g. computed with entropic regularization with vanishing entropy as $N \to \infty$. This narrow convergence of momentum measures then implies strong $\mathbb{L}^2(\mu_0)$ convergence of a suitable small transformation of the v_0^N to the limit v_0, as the next result shows. In particular, for large N, values of v_0^N on μ_0^N are on average good approximations of values of v_0 on nearby points of μ_0 and thus the barycentric projection works as intended.

Corollary 5.6. *For the setup of Proposition 5.5, let S^N be the optimal transport map for $W_2(\mu_0, \mu_0^N)$. Then $v_0^N \circ S^N$ converges to v_0 strongly in $\mathbb{L}^2(\mu_0)$.*

Proof of Proposition 5.5. The result is closely related to the stability of optimal transport plans under perturbations of the marginals [45, Theorem 5.20, Corollary 5.23]. We give it mainly as preparation for the more involved proof for the HK-metric.

For a test function $\phi \in \mathrm{C}(K)$ we find that

$$\begin{aligned}\int_K \phi(x) v_0^N(x) \mathrm{d}\mu_0^N(x) &= \int_K \int_K \phi(x) \operatorname{Log}_x^X(y) \mathrm{d}\pi^N(y|x) \mathrm{d}\mu_0^N(x) \\ &= \int_{K\times K} \phi(x) \operatorname{Log}_x^X(y) \mathrm{d}\pi^N(x_0, x_1) \xrightarrow[N\to\infty]{} \int_{K\times K} \phi(x) \operatorname{Log}_x^X(y) \mathrm{d}\pi(x_0, x_1)\end{aligned}$$

as $N \to \infty$ where we have once more applied the continuous mapping theorem ([44, Theorem 2.3], see proof of Lemma 5.1 for the same argument and more details) for the transition to the limit. Therefore, we can conclude that $v_0^N \mu_0^N$ converges narrowly towards $v_0 \mu_0$. □

Proof of Corollary 5.6. First observe that,

$$\begin{aligned}\left\| v_0^N \circ S^N \right\|_{\mathbb{L}^2(\mu_0)}^2 = \left\| v_0^N \right\|_{\mathbb{L}^2(\mu_0^N)}^2 &= \int_K \left\| v_0^N(x_0) \right\|^2 \mathrm{d}\mu_0^N(x_0) \\ &\le \int_K \int_K \left\| \operatorname{Log}_{x_0}(x_1) \right\|^2 \mathrm{d}\pi^N(x_1|x_0) \mathrm{d}\mu_0^N(x_0) \\ &= \int_{K\times K} d(x_0, x_1)^2 \mathrm{d}\pi^N(x_0, x_1) \to \mathrm{W}_2^2(\mu_0, \mu_1) \end{aligned} \tag{5.2}$$

where the inequality is due to Jensen for the probability measures $\pi^N(\cdot|x_0)$ and the convergence is due to $\pi^N \xrightarrow[N\to\infty]{} \pi$.

Now wet $\gamma^N := (\mathrm{id}, S^N)_\# \mu_0$ the optimal transport plan associated with S^N. We first show that $v_0^N \circ S^N$ converges weakly to v_0 in $\mathbb{L}^2(\mu_0)$. Indeed for any continuous vector field supported on K, $\xi : K \to TX$,

$$\begin{aligned}&\left| \int_K \xi(x) \cdot v_0^N(S^N(x)) \mathrm{d}\mu_0(x) - \int_K \xi(S^N(x)) \cdot v_0^N(S^N(x)) \mathrm{d}\mu_0(x) \right| \\ &= \int_{K\times K} \left| (\xi(x) - \xi(y)) \cdot v_0^N(y) \right| \mathrm{d}\gamma^N(x, y) \\ &\le \sqrt{\int_{K\times K} \|\xi(x) - \xi(y)\|^2 \mathrm{d}\gamma^N(x, y)} \sqrt{\int_{K\times K} \left\| v_0^N(y) \right\|^2 \mathrm{d}\gamma^N(x, y)}\end{aligned}$$

by the Cauchy–Schwarz inequality. The second factor equals $\|v_0^N\|_{\mathbb{L}^2(\mu_0^N)}$, which is bounded due to (5.2). The first term tends to zero since γ^N has to converge narrowly to the unique optimal

plan from μ_0 to μ_0 for the quadratic cost, i.e. $(\mathrm{id},\mathrm{id})_\#\mu_0$, and so

$$\begin{aligned}\lim_{N\to\infty}\int_K \xi(x)\cdot v_0^N(S^N(x))\mathrm{d}\mu_0(x) &= \lim_{N\to\infty}\int_K \xi(S^N(x))\cdot v_0^N(S^N(x))\mathrm{d}\mu_0(x)\\ &= \lim_{N\to\infty}\int_K \xi(y)\cdot v_0^N(y)\mathrm{d}\mu_0^N(y) = \int_K \xi(y)\cdot v_0(y)\mathrm{d}\mu_0(y)\end{aligned}$$

and we have the weak convergence $v_0^N\circ S^N \xrightarrow[N\to\infty]{\mathbb{L}^2(\mu_0)} v_0$.

We augment this to strong convergence by showing that $\|v_0^N\circ S^N\|_{\mathbb{L}^2(\mu_0)}$ also converges to $\|v\|_{\mathbb{L}^2(\mu_0)}$. From (5.2) one has

$$\limsup_{n\to\infty}\left\|v_0^N\circ S^N\right\|^2_{\mathbb{L}^2(\mu_0)} \le W_2^2(\mu_0,\mu_1) = \|v\|^2_{\mathbb{L}^2(\mu_0)}$$

where the last equality is due to Proposition 2.8 (i). By weak lower-semicontinuity of the norm this implies convergence of the norm and thus strong convergence. □

5.3 Convergence of barycentric projection for Hellinger–Kantorovich distance

Now we show the corresponding results for the Hellinger–Kantorovich distance. Here the situation is more involved due to the cut-off of transport at $\kappa\pi/2$. We will find that the singular part of the logarithmic map $\mu_1^\perp$ will not be continuous (Example 5.17), that mere asymptotic optimality of the plans π^N is not sufficient for convergence of the logarithmic map (Example 5.18) and additional conditions must therefore be imposed (Proposition 5.10 and Remark 5.20), and that for singular μ_0 the logarithmic map may not converge even when optimal transport maps would exist (Example 5.19). The following is the analogue of Definition 5.4.

Definition 5.7 (Barycentric logarithmic map for HK_κ)**.** Let $\mu_0\in\mathcal{M}_+(K)$, $\pi\in\mathcal{M}_+(K\times K)$ with $\mathrm{KL}(\mathrm{p}_{0\#}\,\pi|\mu_0)<\infty$ and $\int_{X\times X} c_\kappa^{\mathrm{HK}}\mathrm{d}\pi<\infty$. Then we set

$$\mathrm{Log}^{\mathrm{HK}_\kappa}_{\mu_0}(\pi) := (v_0,\alpha_0)$$

where the components are defined for μ_0-almost any $x\in X$ by

$$v_0(x) := \kappa\tfrac{\mathrm{d}\pi_0}{\mathrm{d}\mu_0}(x)\int_X \frac{\mathrm{Log}_x^X(y)}{\|\mathrm{Log}_x(y)\|}\tan\left(\frac{\|\mathrm{Log}_x(y)\|}{\kappa}\right)\mathrm{d}\pi(y|x),\qquad \alpha_0(x) := 2\left(\tfrac{\mathrm{d}\pi_0}{\mathrm{d}\mu_0}(x)-1\right).$$

When $\mu_0\in\mathcal{M}_{+,\mathrm{Vol}}(K)$ and π is an optimal plan for $\mathrm{HK}^2_\kappa(\mu_0,\mu_1)$ for the formulation (3.1) then this definition reduces to Definition 3.13, for v_0 and α_0. We drop the singular part $\mu_1^\perp$ from the generalized definition since it does not exhibit meaningful continuity properties (see for instance Example 5.17). Note that the expression for $\alpha_0(x)$ does not depend on the transport target(s) y associated to x by π (since π_0 is uniquely defined by μ_0 and μ_1) and therefore, averaging over the possible targets is unnecessary. We then have the following stability result.

Proposition 5.8. *Let $(\mu_0^N)_N$, $(\mu_1^N)_N$ be two sequences in $\mathcal{M}_+(K)$, narrowly converging to $\mu_0\in\mathcal{M}_{+,\mathrm{Vol}}(K)$ and $\mu_1\in\mathcal{M}_+(K)$ respectively. Let $(\pi^N)_N$ be a sequence in $\mathcal{M}_+(K\times K)$ such that $\pi^N/\cos(d/\kappa)\xrightarrow[N\to\infty]{}\pi/\cos(d/\kappa)$ where π is the optimal plan for $\mathrm{HK}^2_\kappa(\mu_0,\mu_1)$ in (3.1). Set $(v_0^N,\alpha_0^N) := \mathrm{Log}^{\mathrm{HK}_\kappa}_{\mu_0^N}(\pi^N)$, and $(v_0,\alpha_0) := \mathrm{Log}^{\mathrm{HK}_\kappa}_{\mu_0}(\pi)$, as given by Definition 5.7. Then the momentum measures $v_0^N\mu_0^N$ and $\alpha_0^N\mu_0^N$ narrowly converge to $v_0\mu_0$ and $\alpha_0\mu_0$ as $N\to\infty$.*

Necessity of the assumption about the convergence of $\pi^N/\cos(d/\kappa)$ is demonstrated in Example 5.18. We show in Proposition 5.10 that it is satisfied when the π^N are actually optimal, and discuss the case of entropic approximation in Remark 5.20.

Remark 5.9. We quickly mention here for clarity that, provided π is the optimal plan for for $\mathrm{HK}^2_\kappa(\mu_0,\mu_1)$, the measure $\pi/\cos(d/\kappa)$ is finite. Indeed, due to optimality conditions Proposition 3.8, item iii, one has

$$\int_{K\times K}\frac{\mathrm{d}\pi(x,y)}{\cos\left(\frac{d(x,y)}{\kappa}\right)}=\int_{K\times K}\frac{\mathrm{d}\pi(x,y)}{\sqrt{\frac{\mathrm{d}\pi_0}{\mathrm{d}\mu_0}(x)\frac{\mathrm{d}\pi_1}{\mathrm{d}\mu_1}(y)}}\leq\sqrt{|\mu_0|\,|\mu_1|}.$$

Proof of Proposition 5.8. We start with the convergence of $\alpha^N\cdot\mu_0^N$. For a test function $\phi\in\mathrm{C}(X)$ one finds that

$$\int_X\phi(x)\alpha_0^N(x)\mathrm{d}\mu_0^N(x)=2\left(\int_{X\times X}\phi(x)\mathrm{d}\pi^N(x,y)-\int_X\phi(x)\mathrm{d}\mu_0^N(x)\right)$$

and the same expression holds for the limit. As by assumption, $\pi^N/\cos(d/\kappa)\xrightarrow[N\to\infty]{}\pi/\cos(d/\kappa)$, one also has $\pi^N\xrightarrow[N\to\infty]{}\pi$, from which we deduce that $\alpha_0^N\mu_0^N\xrightarrow[N\to\infty]{}\alpha_0\mu_0$. As for the convergence of $v_0^N\cdot\mu_0^N$, consider a test function $\phi\in\mathrm{C}(X,TX)$. One has

$$\begin{aligned}&\int_X\phi(x_0)\cdot v_0^N(x_0)\mathrm{d}\mu_0^N(x_0)\\&=\kappa\int_X\int_X\tan\left(\frac{d(x_0,x_1)}{\kappa}\right)\phi(x_0)\cdot\frac{\mathrm{Log}^X_{x_0}(x_1)}{\left\|\mathrm{Log}^X_{x_0}(x_1)\right\|}\mathrm{d}\pi^N(x_1|x_0)\frac{\mathrm{d}\pi_0^N}{\mathrm{d}\mu_0^N}(x_0)\,\mathrm{d}\mu_0^N(x_0)\\&=\int_{X\times X}F(x_0,x_1)\frac{1}{\cos\left(\frac{d(x_0,x_1)}{\kappa}\right)}\mathrm{d}\pi^N(x_0,x_1)\end{aligned}$$

where we grouped some terms into the function F, which is independent of N, bounded, and continuous except on a π-negligible set (see related discussion in the proof of Proposition 5.2). By the assumption that $\pi^N/\cos(d/\kappa)\xrightarrow[N\to\infty]{}\pi/\cos(d/\kappa)$ and the continuous mapping theorem (again, see proof of Proposition 5.2) this last integral converges to

$$\int_{X\times X}F(x_0,x_1)\frac{1}{\cos\left(\frac{d(x_0,x_1)}{\kappa}\right)}\mathrm{d}\pi(x_0,x_1)=\int_X\phi(x)\cdot v_0(x)\mathrm{d}\mu_0(x).\qquad\square$$

The rest of this section deals mainly with the investigation of the strong convergence assumption for the plans from Proposition 5.8, namely $\pi^N/\cos(d/\kappa)\xrightarrow[N\to\infty]{}\pi/\cos(d/\kappa)$. First, we show that it is satisfied when $\mu_0\ll\mathrm{Vol}$ and the π^N are optimal plans for each N.

Proposition 5.10. *Let $(\mu_0^N)_N$, $(\mu_1^N)_N$ be two sequences in $\mathcal{M}_+(K)$, converging narrowly to $\mu_0\in\mathcal{M}_{+,\mathrm{Vol}}(K)$ and $\mu_1\in\mathcal{M}_+(K)$ respectively. Let $(\pi^N)_N$ be an optimal unbalanced plan for $\mathrm{HK}^2_\kappa(\mu_0^N,\mu_1^N)$ in (3.1) and likewise let π be optimal for $\mathrm{HK}^2_\kappa(\mu_0,\mu_1)$. Then $\pi^N/\cos(d/\kappa)$ narrowly converges to $\pi/\cos(d/\kappa)$ as $N\to\infty$.*

The proof of Proposition 5.10 uses the following *semi-coupling* formulation of HK_κ, introduced in [15].

Remark 5.11 (Square-root measure)**.** For $\rho, \sigma \in \mathcal{M}_+(K)$ (and similarly for general compact metric spaces) the square root measure $\sqrt{\rho\sigma}$ is characterized by

$$\int_K \phi \mathrm{d}\sqrt{\rho\sigma} := \int_K \phi \sqrt{\tfrac{\mathrm{d}\rho}{\mathrm{d}\tau}\tfrac{\mathrm{d}\sigma}{\mathrm{d}\tau}}\mathrm{d}\tau$$

for all $\phi \in \mathrm{C}(K)$, where $\tau \in \mathcal{M}_+(K)$ is such that $\rho, \sigma \ll \tau$. Due to the joint 1-homogeneity of the function $(r,s) \mapsto \sqrt{rs}$ the right integral does not depend on the choice of τ.

Using that the function $(s,t) \mapsto s + t - 2\sqrt{rs}$ is non-negative, convex, lower-semicontinuous and positively 1-homogeneous on $\mathbb{R}^2_+$, by integrating against continuous test functions, and with [3, Theorem 2.38], we obtain that if $(\rho_n, \sigma_n) \xrightarrow[N\to\infty]{} (\rho, \tau)$ (in the sense of narrow convergence of vector measures), then

$$\int_K \phi \mathrm{d}\sqrt{\rho\sigma} \geq \limsup_n \int_K \phi \mathrm{d}\sqrt{\rho_n\sigma_n}.$$

In particular $\sqrt{\rho\sigma} \geq \widehat{\sqrt{\rho\sigma}}$ holds for any narrow cluster point $\widehat{\sqrt{\rho\sigma}}$ of $\sqrt{\rho_n\sigma_n}$.

Proposition 5.12 (Semi-coupling formulation)**.** *For $\mu_0, \mu_1 \in \mathcal{M}_+(K)$ one has*

$$\mathrm{HK}^2_\kappa(\mu_0,\mu_1) = \kappa^2 \inf \left\{ \|\gamma_0\| + \|\gamma_1\| - 2\int_{X\times X} \mathrm{Cos}\left(\frac{d(x_0,x_1)}{\kappa}\right) \mathrm{d}\sqrt{\gamma_0\gamma_1}(x_0,x_1) \;\middle|\; \gamma_0,\gamma_1 \in \mathcal{M}_+(K\times K),\ p_{i\#}\gamma_i = \mu_i \text{ for } i=0,1 \right\} \tag{5.3}$$

and minimizers for the right-hand side exist. Furthermore, with π denoting an optimal transport plan for the soft-marginal formulation (3.1), optimal semi-couplings (γ_0,γ_1) for (5.3) are given by:

$$\gamma_0 := \tfrac{\mathrm{d}\mu_0}{\mathrm{d}\pi_0}\pi + (\mathrm{id},\mathrm{id})_\#\mu_0^\perp, \qquad \gamma_1 := \tfrac{\mathrm{d}\mu_1}{\mathrm{d}\pi_1}\pi + (\mathrm{id},\mathrm{id})_\#\mu_1^\perp \tag{5.4}$$

with the notations of Remark 3.3. Conversely, if (γ_0,γ_1) are optimal for (5.3), then

$$\pi := \mathrm{Cos}\left(\frac{d(x_0,x_1)}{\kappa}\right)\sqrt{\gamma_0\gamma_1} \tag{5.5}$$

is a minimizer of the soft-marginal formulation (3.1).

Proof. Existence of minimizers and equivalence with (3.1) was established in [15, Corollary 5.9].

Formula (5.4) follows from the following inequality (see [11, Lemma 3.16]) between the costs featured in the two problems: for $x_0, x_1 \in X$ and $u_0, u_1 > 0$,

$$c_\kappa(x_0,x_1) + \kappa^2 \cdot \sum_{i\in\{0,1\}} \varphi\left(\frac{1}{u_i}\right) u_i \geq \kappa^2 \left(u_0 + u_1 - 2\sqrt{u_0u_1}\,\mathrm{Cos}\left(\frac{d(x_0,x_1)}{\kappa}\right)\right) \tag{5.6}$$

where $\varphi(s) = s\log(s) - s + 1$ is the integrand of the KL-divergence, (3.4). Furthermore, equality happens if and only if

$$\sqrt{u_0u_1}\,\mathrm{Cos}\left(\frac{d(x_0,x_1)}{\kappa}\right) = 1.$$

Integrating (5.6) with $u_i := \frac{\mathrm{d}\mu_i}{\mathrm{d}\pi_i}(x_i)$ against the optimal plan π yields optimality of the specific semi-couplings in (5.4).

Formula (5.5) is established for the constructed π as follows:

$$\begin{aligned} E_\kappa(\pi|\mu_0,\mu_1) &= \int_{K\times K} c_\kappa^{\mathrm{HK}}\mathrm{d}\pi + \kappa^2 \sum_{i\in\{0,1\}} \int_K \varphi(\tfrac{\mathrm{d}\pi_i}{\mathrm{d}\mu_i})\mathrm{d}\mu_i \\ &\leq \int_{K\times K} c_\kappa^{\mathrm{HK}}\mathrm{d}\pi + \kappa^2 \sum_{i\in\{0,1\}} \int_{K\times K} \varphi(\tfrac{\mathrm{d}\pi}{\mathrm{d}\gamma_i})\mathrm{d}\gamma_i \end{aligned}$$

where we used [11, Lemma 3.15] with $T = \mathrm{p}_i$ and then finally expanding the formulae for φ and c_κ^{HK} (or equivalently leveraging the equality case in (5.6)), the reader can check that the expression simplifies into:

$$= \kappa^2 \left[|\gamma_0| + |\gamma_1| - 2\int_{X\times X} \mathrm{Cos}\left(\frac{d(x_0,x_1)}{\kappa}\right) \mathrm{d}\sqrt{\gamma_0\gamma_1}(x_0,x_1)\right]$$

which equals $\mathrm{HK}_\kappa^2(\mu_0,\mu_1)$ and thus completes the proof. □

Proof of Proposition 5.10. The narrow convergence of π^N to π would follow directly from the joint lower-semicontinuity of $E_\kappa(\pi|\mu_0,\mu_1)$, (3.1), with respect to all three measures and the triangle inequality for HK_κ. The convergence of $\pi^N/\cos(d/\kappa)$ is however more involved.

In the following for $i \in \{0,1\}$, let

$$\gamma_i^N := \tfrac{\mathrm{d}\mu_i^N}{\mathrm{d}\pi_i^N}\cdot\pi^N + (\mathrm{id},\mathrm{id})_\#(\mu_i^N)^\perp, \qquad \gamma_i := \tfrac{\mathrm{d}\mu_i}{\mathrm{d}\pi_i}\cdot\pi + (\mathrm{id},\mathrm{id})_\#(\mu_i)^\perp, \tag{5.7}$$

where $\pi_i^N = \mathrm{p}_{i\#}\,\pi^N$ (in the spirit of Remark 3.4). By Proposition 3.8 one has

$$\frac{\pi^N}{\cos(d/\kappa)} = \sqrt{\gamma_0^N\gamma_1^{(N)}},$$

and by Proposition 5.12 $(\gamma_0^{(N)},\gamma_1^{(N)})$ are optimal semi-couplings for $\mathrm{HK}_\kappa^2(\mu_0^{(N)},\mu_1^{(N)})$. Finally, by narrow compactness we select a common subsequence and respective cluster points such that narrowly as $N\to\infty$ (on this subsequence)

$$\frac{\pi^N}{\cos(d/\kappa)} = \sqrt{\gamma_0^N\gamma_1^N} \to \rho, \qquad \gamma_0^N \to \tilde\gamma_0, \qquad \gamma_1^N \to \tilde\gamma_1.$$

By standard lower-semicontinuity and the triangle inequality the cluster point $(\tilde\gamma_0,\tilde\gamma_1)$ are also optimal semi-couplings for $\mathrm{HK}_\kappa^2(\mu_0,\mu_1)$. In addition, we set

$$\tilde\rho := \sqrt{\tilde\gamma_0\tilde\gamma_1},$$

and by Proposition 5.12 and uniqueness of the optimal π in (3.1) one has $\pi = \cos(d/\kappa)\cdot\tilde\rho$.

Let now

$$A_= := \{(x_0,x_1)\in X^2 | d(x_0,x_1) = \kappa\pi/2\}$$

and in the same way define $A_<$ and $A_>$ for the point pairs closer and further than $\kappa\pi/2$ respectively. From the above considerations we conclude

$$\pi/\cos(d/\kappa)\llcorner_{A_<} = \rho\llcorner_{A_<} = \tilde\rho\llcorner_{A_<}.$$

This can be verified by integrating against continuous test functions ϕ with compact support in $A_<$. These do not see the singularity of $1/\cos(d/\kappa)$, and thus $\phi/\cos(d/\kappa)$ is still continuous test function and the above convergence is reduced to the already established convergence of π^N to π. By optimality of π for (3.1) one has

$$\pi/\cos(d/\kappa)\llcorner_{A_=} = \pi/\cos(d/\kappa)\llcorner_{A_>} = 0,$$

and by the Portmanteau theorem one has

$$\rho\llcorner_{A_>} = \tilde{\rho}\llcorner_{A_>} = 0.$$

By Remark 5.11 one has furthermore that $\tilde{\rho} \geq \rho$. So in conclusion one has $\tilde{\rho} \geq \rho \geq \pi/\cos(d/\kappa)$ and the differences between the measures (if any) are concentrated on $A_=$.

Next, we prove that the components of γ_i and $\tilde{\gamma}_i$ that are dominated by π are identical, for $i = \{0, 1\}$. We have already established that both (γ_0, γ_1) and $(\tilde{\gamma}_0, \tilde{\gamma}_1)$ are optimal semi-couplings for $\mathrm{HK}^2_\kappa(\mu_0, \mu_1)$. By convexity, any convex combination between the two, denoted by (γ_0^t, γ_1^t) for $t \in [0, 1]$, must also be optimal, and once more by Proposition 5.12 and uniqueness of π one must have $\sqrt{\gamma_0^t \gamma_1^t} = \pi$ is constant on $t \in [0, 1]$. The function $\mathbb{R}^2_+ \ni (a, b) \mapsto \sqrt{ab}$ is constant along line segments (of non-zero length) if and only if $a = 0$ or $b = 0$ along the whole segment. This implies that the γ_i^t must be constant in t, π-almost everywhere. Recalling the Lebesgue decomposition of γ_i with respect to π from (5.7) we can therefore now write for $i \in \{0, 1\}$,

$$\gamma_i := \tfrac{\mathrm{d}\mu_i}{\mathrm{d}\pi_i} \cdot \pi + (\mathrm{id}, \mathrm{id})_\#(\mu_i)^\perp, \qquad \tilde{\gamma}_i := \tfrac{\mathrm{d}\mu_i}{\mathrm{d}\pi_i} \cdot \pi + \tilde{\gamma}_i^\perp.$$

We recall from Proposition 3.8 that $\mu_0^\perp$ and $\mu_1^\perp$ are mutually singular and in fact even $d(\operatorname{spt}\mu_0, \operatorname{spt}\mu_1) \geq \kappa\pi/2$. From this decomposition (and $\mathrm{p}_{i\#}\,\tilde{\gamma}_i = \mu_i$) we also deduce that $\mathrm{p}_{i\#}\,\tilde{\gamma}_i^\perp = \mu_i^\perp$. We then observe that

$$\Delta\rho := \tilde{\rho} - \pi/\cos(d/\kappa) = \sqrt{\tilde{\gamma}_0^\perp \tilde{\gamma}_1^\perp}$$

from which we also infer that $\mathrm{p}_{i\#}\,\Delta\rho \ll \mu_i^\perp$, which implies

$$d(\operatorname{spt}\mathrm{p}_{0\#}\,\Delta\rho, \operatorname{spt}\mu_1^\perp) \geq \kappa\pi/2,$$

Also, we have shown earlier that $\Delta\rho$ is concentrated on $A_=$, and therefore

$$d(\operatorname{spt}\mathrm{p}_{0\#}\,\Delta\rho, \operatorname{spt}\mu_1^\perp) \leq d(\operatorname{spt}\mathrm{p}_{0\#}\,\Delta\rho, \operatorname{spt}\mathrm{p}_{1\#}\,\Delta\rho) \leq \kappa\pi/2$$

and in combination

$$\operatorname{spt}\mathrm{p}_{0\#}\,\Delta\rho \subset B = \{x \in X \,|\, d(x, \operatorname{spt}\mu_1^\perp) = \kappa\pi/2\}.$$

Since

$$\operatorname{spt}\mathrm{p}_{0\#}\,\Delta\rho \ll \mu_0^\perp \ll \mu_0 \ll \mathrm{Vol}$$

and $\mathrm{Vol}(B) = 0$ we conclude that $\Delta\rho = 0$ and thus $\tilde{\rho} = \rho = \pi/\cos(d/\kappa)$. □

Similarly to the balanced case, one can augment the narrow convergence of the momentum measures from Proposition 5.8 to a strong $\mathbb{L}^2(\mu_0)$ convergence, up to composition with a transport map from $\mu_0/|\mu_0|$ to $\mu_0^N/|\mu_0^N|$. However, in the unbalanced case an additional regularity assumption is required. A counter-example, in the absence of the assumption, is given in Example 5.17.

Corollary 5.13. *With the notations and under the assumptions of Proposition 5.8, assume furthermore that* $\mu_0 \neq 0$ *and*

$$\limsup_{N\to\infty} \int_{X\times X} \frac{\mathrm{d}\pi_0^N}{\mathrm{d}\mu_0^N}(x_0) \frac{\mathrm{d}\pi^N}{\cos\left(\frac{d(x_0,x_1)}{\kappa}\right)} \leq \int_X \frac{\mathrm{d}\mu_1}{\mathrm{d}\pi_1}(x_1)\mathrm{d}\pi_1(x_1). \tag{5.8}$$

Then, denoting by S^N *the optimal transport map for* W_2 *from* $\mu_0/|\mu_0|$ *to* $\mu_0^N/|\mu_0^N|$, $(v_0^N \circ S^N)_{N\in\mathbb{N}}$ *and* $(\alpha_0^N \circ S^N)_{N\in\mathbb{N}}$ *converge strongly in* $\mathbb{L}^2(\mu_0)$ *towards respectively* v_0 *and* α_0.

Remark 5.14. The additional condition for Corollary 5.13 might seem unexpected at first. However, in the case were, for any N, π^N is an optimal plan for the soft-marginal formulation of $\mathrm{HK}^2_\kappa(\mu_0^N, \mu_1^N)$, inequality (5.8) simplifies to $\liminf_{N\to\infty} \left|(\mu_1^N)^\perp\right| \geq \left|\mu_1^\perp\right|$. Indeed, using the optimality conditions item iii of Proposition 3.8, eq. (5.8) simplifies to

$$\limsup_{N\to\infty} \int_{X\times X} \frac{\mathrm{d}\mu_1{}^N}{\mathrm{d}\pi_1^N}(x_1)\mathrm{d}\pi^N(x_0,x_1) = \limsup_{N\to\infty} \int_X \frac{\mathrm{d}\mu_1{}^N}{\mathrm{d}\pi_1^N}(x_1)\mathrm{d}\pi_1^N(x_1) \leq \int_X \frac{\mathrm{d}\mu_1}{\mathrm{d}\pi_1}\mathrm{d}\pi_1(x)$$

and using the fact that $\lim_{N\to\infty} \left|\mu_1^N\right| = |\mu_1|$ and the definition of $(\mu_1^N)^\perp, \mu_1^\perp$, the former limsup inequality is equivalent to the latter liminf one.

Note that we always have $\limsup_{N\to\infty} \left|(\mu_1^N)^\perp\right| \leq \left|\mu_1^\perp\right|$ from the narrow continuity of the squared HK_κ distance, $\lim_{N\to\infty} \mathrm{HK}^2_\kappa(\mu_0^N, \mu_1^N) = \mathrm{HK}^2_\kappa(\mu_0, \mu_1)$: Indeed, leveraging the weak $\mathbb{L}^2(\mu_0)$ convergence of $v_0^N \circ S_N$ and $\alpha_0^N \circ S_N$ to v_0 and α_0 (which does not require to assume (5.8)),

$$\begin{aligned}
\mathrm{HK}^2(\mu_0,\mu_1) &= \|v_0\|^2_{\mathbb{L}^2(\mu_0)} + \frac{\kappa^2}{4}\|\alpha_0\|^2_{\mathbb{L}^2(\mu_0)} + \kappa^2\left|\mu_1^\perp\right| \\
&\leq \liminf_{N\to\infty} \left\|v_0^N \circ S_N\right\|^2_{\mathbb{L}^2(\mu_0)} + \frac{\kappa^2}{4}\|\alpha_0 \circ S_N\|^2_{\mathbb{L}^2(\mu_0)} + \kappa^2\left|\mu_1^\perp\right|
\end{aligned}$$

and rewritting

$$\begin{aligned}
\mathrm{HK}^2(\mu_0,\mu_1) &= \lim_{N\to\infty} \mathrm{HK}^2_\kappa(\mu_0^N,\mu_1^N) \\
&\geq \limsup_{N\to\infty} \left\|v_0^N \circ S_N\right\|^2_{\mathbb{L}^2(\mu_0)} + \frac{\kappa^2}{4}\|\alpha_0 \circ S_N\|^2_{\mathbb{L}^2(\mu_0)} + \kappa^2\left|(\mu_1^N)^\perp\right|,
\end{aligned}$$

we have, *a priori*, $\limsup_{N\to\infty} \left|(\mu_1^N)^\perp\right| \leq \left|\mu_1^\perp\right|$. Therefore, the assumption of Corollary 5.13 is equivalent to $\lim_{N\to\infty} \left|(\mu_1^N)^\perp\right| = \left|\mu_1^\perp\right|$ (which may not always be verified, see Example 5.17).

Remark 5.15 (Extension to SHK)**.** We remark without proof that by leveraging the relationship between the logarithmic maps for the HK and SHK metrics, one can obtain the same stability results, under the same assumptions as Proposition 5.8 (resp. Corollary 5.13) for the SHK logarithmic map. For instance, the continuity of the scaling factors

$$s'(0) = \frac{\mathrm{SHK}_\kappa(\mu_0^N,\mu_1^N)/\kappa}{\sin(\mathrm{SHK}_\kappa(\mu_0^N,\mu_1^N)/\kappa)}$$

is a consequence of the continuity of the SHK distance with respect to the narrow convergence of measures.

Remark 5.16 (Re-scaling of HK tangent vectors for SHK conversion)**.** The barycentric logarithmic map of Definition 5.7 is not exactly invertible due to the averaging of the velocity field. As this section shows, in the limit the true logarithmic map is recovered. However, when working numerically with SHK, small deviations at finite N can already lead to undesirable errors. Let $\mu_0, \mu_1 \in \mathcal{P}(X)$, let π be an (approximately) optimal non-deterministic plan. Computing (v_0, α_0) as in Definition 5.7 one finds in general that $\tilde{\mu}_1 := \mathrm{Exp}^{\mathrm{HK}_\kappa}_{\mu_0}(v_0, \alpha_0) \neq \mu_1$ and $\tilde{\mu}_1 \notin \mathcal{P}(X)$. In this case, the formulas of Section 3.4 for the conversion to SHK tangent vectors become inconsistent as they rely on the assumption $\tilde{\mu}_1 \in \mathcal{P}(X)$. This can be remedied, by re-scaling the vectors (v_0, α_0) first. Indeed, using Proposition 3.15 one can show for re-scaled tangent vector $(\tilde{v}_0, \tilde{\alpha}_0) := (q \cdot v_0, q \cdot \alpha_0 + 2(q-1))$ for $q = 1/\sqrt{|\tilde{\mu}_1|}$ that $\mathrm{Exp}^{\mathrm{HK}_\kappa}_{\mu_0}(\tilde{v}_0, \tilde{\alpha}_0) = \tilde{\mu}_1/|\tilde{\mu}_1| \in \mathcal{P}(X)$.

Example 5.17 (Discontinuity of $\mu_1^\perp$)**.** The singular component $\mu_1^\perp$ of the logarithmic map does not necessarily converge, even under the assumptions of Proposition 5.8 (despite it being uniquely defined for any N, unlike the velocity v_0^N which potentially depends on the choice the optimal plan π^N solving (3.1)). Indeed, let $X = \mathbb{R}$, $\kappa = 1$ and consider for some $L > \pi/2 + 1$,

$$\mu_0^N := (1 - 1/N)\mathcal{L}\llcorner_{[0,1]} + 1/N\mathcal{L}\llcorner_{[L,L+1]} \overset{*}{\rightharpoonup} \mu_0 := \mathcal{L}\llcorner[0,1],$$

$$\mu_1^N := 1/N\mathcal{L}\llcorner_{[0,1]} + (1 - 1/N)\mathcal{L}\llcorner_{[L,L+1]} \overset{*}{\rightharpoonup} \mu_1 := \mathcal{L}\llcorner[L, L+1].$$

One can easily see that for every N one has

$$\pi^N = \sqrt{(1-1/N)1/N}(\mathrm{id}, \mathrm{id})_\#[\mathcal{L}\llcorner_{[0,1]} + \mathcal{L}\llcorner_{[L,L+1]}]$$

and so $(\mu_i^N)^\perp = 0$, whereas for the limit one has $\pi = 0$ and thus $\mu_i^\perp = \mu_i \neq 0$.

This also readily gives an example where the strong $\mathbb{L}^2$-convergence is not verified (although we make no claims as to the necessity of the assumption $\left|(\mu_1^N)^\perp\right| \xrightarrow[N\to\infty]{} \left|\mu_1^\perp\right|$ in Corollary 5.13). Indeed, since $(\mu_i^N)^\perp = 0$, and $\mu_0^N \ll \mathcal{L}$, one has (after some straightforward calculations),

$$\mathrm{HK}_1^2(\mu_0^N, \mu_1^N) = \left\|v_0^N\right\|^2_{\mathbb{L}^2(\mu_0^N)} + \frac{1}{4}\left\|\alpha_0^N\right\|^2_{\mathbb{L}^2(\mu_0^N)} \xrightarrow[N\to\infty]{} \mathrm{HK}_1^2(\mu_0, \mu_1)$$

whereas

$$\|v_0\|^2_{\mathbb{L}^2(\mu_0)} + \frac{\kappa^2}{4}\|\alpha_0\|^2_{\mathbb{L}^2(\mu_0)} = |\mu_0| < \mathrm{HK}_1^2(\mu_0, \mu_1) = |\mu_0| + |\mu_1|\,,$$

forbidding the strong $\mathbb{L}^2$ convergence. In fact, one can even check that $v_0^N = v_0 = 0$ as neither π^N nor π represent any movement of mass particles but

$$\alpha_0^N = \begin{cases} 2\left(\sqrt{\frac{1/N}{1-1/N}} - 1\right) & \text{on } [0;1] \\ 2\left(\sqrt{\frac{1-1/N}{1/N}} - 1\right) & \text{on } [L; L+1] \end{cases}$$

and the expression on $[L; L+1]$ diverges to $+\infty$ as $N \to \infty$ (however, this diverging expression is exactly compensated by the density of μ_0^N in the moment measure $\alpha_0^N\mu_0^N$, allowing us to still have the claim of Proposition 5.8).

Example 5.18 (Non-convergence of $\pi^N/\cos(d/\kappa)$ due to sub-optimality of π^N)**.** For $\mu_1 \ll \mathrm{Vol}$ and π^N optimal, Proposition 5.10 implies convergence of $\pi^N/\cos(d/\kappa)$ to $\pi/\cos(d/\kappa)$. We show that slightly sub-optimality of π^N, such that still $\pi^N \to \pi$, can already hinder the convergence

$\pi^N/\cos(d/\kappa)$ to $\pi/\cos(d/\kappa)$, even when $\mu_1 \ll \mathrm{Vol}$. Let $X = \mathbb{R}$, for simplicity let $\kappa = 1$, and consider the identical sequences of measures

$$\mu_0^N := \mu_1^N := \underbrace{\mathcal{L}\llcorner_{[0;1]}}_{=:\mu_{00}} + \mathcal{L}\llcorner_{[\frac{\pi}{2}-\frac{1}{N};1+\frac{\pi}{2}-\frac{1}{N}]}$$

narrowly converging to the limit measures $\mu_0 := \mu_1 := \mathcal{L}\llcorner_{[0;1]} + \mathcal{L}\llcorner_{[\frac{\pi}{2};1+\frac{\pi}{2}]}$. The obvious optimal plans π^N for the soft marginal formulation of $\mathrm{HK}^2(\mu_0^N, \mu_1^N)$ are induced by the identity map, $\pi^N := (\mathrm{id}, \mathrm{id})_\# \mu_0^N$, and similarly at the limit, $\pi := (\mathrm{id}, \mathrm{id})_\# \mu_0$. Consider now the sequence of perturbed plans $\tilde{\pi}^N := \left(1 - \frac{1}{N}\right) \pi^N + \frac{1}{N} (\mathrm{id}, T^N)_\# \mu_{00}$ where $T^N : x \in \mathbb{R} \mapsto x + \frac{\pi}{2} - \frac{1}{N}$. We notice that both π^N and $\tilde{\pi}^N$ narrowly converge to π and in fact, one even has

$$\begin{aligned} E(\tilde{\pi}^N | \mu_0^N, \mu_1^N) &= \left(1 - \frac{1}{N}\right) \times 0 - \frac{2}{N} \ln\left(\cos\left(\frac{\pi}{2} - \frac{1}{N}\right)\right) + \mathrm{KL}\left(\left(1 - \frac{1}{N}\right) \mu_0^N + \frac{1}{N} \mu_{00} \;\middle|\; \mu_0^N\right) \\ &\quad + \mathrm{KL}\left(\left(1 - \frac{1}{N}\right) \mu_0^N + \frac{1}{N} (T^N_\# \mu_{00}) \;\middle|\; \mu_0^N\right) \\ &\xrightarrow[N\to\infty]{} 0 = \mathrm{HK}^2(\mu_0, \mu_1). \end{aligned}$$

However,

$$\frac{\tilde{\pi}^N}{\cos(d)} = \left(1 - \frac{1}{N}\right) \pi^N + \frac{1}{N} \frac{(\mathrm{id}, T^N)_\# \mu_{00}}{\cos\left(\frac{\pi}{2} - \frac{1}{N}\right)} \xrightarrow[N\to\infty]{} \pi + (\mathrm{id}, T)_\# \mu_{00} \neq \frac{\pi}{\cos(d)} = \pi$$

where $T : x \in \mathbb{R} \mapsto x + \frac{\pi}{2}$. And indeed, for the logarithmic maps, $v_0^N, \alpha_0^N := \mathrm{Log}_{\mu_0^N}(\tilde{\pi}^N)$ and $v_0, \alpha_0 := \mathrm{Log}_{\mu_0}(\mu_1) = (0, 0)$, we obtain that

$$v_0^N \mu_0^N \xrightarrow[N\to\infty]{} 0 + \frac{T - \mathrm{id}}{\|T - \mathrm{id}\|} \mu_{00} \neq v_0 \mu_0.$$

Example 5.19 (Non-convergence of $\pi^N/\cos(d/\kappa)$ due to singular μ_1)**.** In this example, the π^N are optimal, but assumption $\mu_0 \ll \mathrm{Vol}$ is violated, resulting in non-convergence of $\pi^N/\cos(d/\kappa)$. Let $X = \mathbb{R}$, $\kappa = 1$, and set $\mu_0^N := \mu_0 := \delta_0$, as well as $\mu_1^N := \delta_{\pi/2 - 1/N}$, converging narrowly to $\mu_1 := \delta_{\pi/2}$. One quickly finds that the unique optimal plan for the soft-marginal formulation of $\mathrm{HK}_1^2(\mu_0^N, \mu_1^N)$ is given by $\pi^N := \cos(\pi/2 - 1/N) \cdot \delta_0 \otimes \delta_{\pi/2 - 1/N}$. Between μ_0 and μ_1 the optimal plan is $\pi = 0$. Therefore, we have $\pi^N \to \pi$, but $\pi^N/\cos(d) \to \delta_0 \otimes \delta_{\pi/2} \neq \pi$. At the level of logarithmic maps one obtains for $(v_0^N, \alpha_0^N) := \mathrm{Log}_{\mu_0}^{\mathrm{HK}_1}(\pi^N)$ (defined only at $x_0 = 0$) that

$$v_0^N = \sin\left(\frac{\pi}{2} - \frac{1}{N}\right), \qquad \alpha_0^N = 2\left[\mathrm{Cos}\left(\frac{\pi}{2} - \frac{1}{N}\right) - 1\right]$$

for which $\lim_{N\to\infty}(v_0^N, \alpha_0^N) = (1, -2)$. On the other hand, $v_0, \alpha_0 := \mathrm{Log}_{\mu_0}^{\mathrm{HK}_1}(\pi) = (0, -2)$ and we have neither the narrow convergence of $v_0^N \mu_0^N$ to $v_0 \mu_0$ nor the strong $\mathbb{L}^2(\mu_0)$-convergence $v_0^N \circ S_N \xrightarrow[N\to\infty]{} v_0$, with the notations of Corollary 5.13.

Remark 5.20 (Entropic regularization)**.** A relevant sequence of plans π^N is given by the solution of the entropic regularization of the soft-marginal formulation,

$$\mathrm{HK}^2_{\kappa,\epsilon}(\mu_0, \mu_1) := \min_\pi E_\kappa(\pi | \mu_0, \mu_1) + \epsilon\, \mathrm{KL}(\pi | \mu_0 \otimes \mu_1).$$

We refer the reader to [14] for a more extensive introduction to this regularized unbalanced transport problem and its numerical resolution. This problem admits a dual formulation,

$$\begin{aligned}\mathrm{HK}^2_{\kappa,\epsilon}(\mu_0,\mu_1) = \sup_{u_0,u_1\in \mathrm{C}(K)} \kappa^2 \left(\int_X 1 - e^{-u_0(x_0)/\kappa^2}\mathrm{d}\mu_0(x_0) + \int_X 1 - e^{-u_1(x_1)/\kappa^2}\mathrm{d}\mu_1(x_1)\right) \\ + \epsilon \int_X 1 - e^{\frac{u_0(x_0)+u_1(x_1)-c^{\mathrm{HK}}_\kappa(x_0,x_1)}{\epsilon}}\mathrm{d}\mu_0(x_0)\otimes\mu_1(x_1) \qquad (5.9)\end{aligned}$$

and a corresponding primal-dual optimality condition for $u_{\varepsilon,0}, u_{1,\varepsilon} \in \mathrm{C}(K)$ and $\pi_\varepsilon \in \mathcal{M}(K\times K)$,

$$\pi_\varepsilon = e^{\frac{u_{\varepsilon,0}\oplus u_{\varepsilon,1} - c^{\mathrm{HK}}_\kappa}{\epsilon}} \cdot \mu_0\otimes\mu_1.$$

For simplicity, assume that optimal $u_{\varepsilon,i} \in \mathrm{C}(K)$ exist for all $\varepsilon \geq 0$ and that one has uniform convergence of $u_{\varepsilon,i}$ to unregularized limit solutions $u_{0,i}$. This implies that the entropic plans π_ϵ narrowly converge to some π that minimizes the unregularized problem. Indeed in that case, the entropic term in the dual formulation (5.9) vanishes as $\epsilon \to 0$ and therefore, any cluster point of the (tight) family $(\pi_\epsilon)_\epsilon$ has to be optimal. Furthermore, one finds that

$$\frac{\pi_\varepsilon}{\cos(d/\kappa)} = \exp\left(\left[2\kappa^2 \log \mathrm{Cos}(d/\kappa) - \varepsilon \log \mathrm{Cos}(d/\kappa) + u_{\varepsilon,0}\oplus u_{\varepsilon,1}\right]/\varepsilon\right)\cdot\mu_0\otimes\mu_1$$

with the convention that the exponential function evaluates to 0 for $d(x,y) > \kappa\pi/2$. Then, using the uniform boundedness of the sequences $u_{\varepsilon,i}$, $i = 0,1$, one gets that there exist $\eta_0 > 0$ and $C > 0$ such that for any $\eta < \eta_0$, $\varepsilon > 0$ and $d(x,y)/\kappa > \pi/2 - \eta$,

$$2\kappa^2 \log\cos(d(x,y)/\kappa) - \varepsilon\log\cos(d(x,y)/\kappa) + u_{\varepsilon,0}(x) + u_{\varepsilon,1}(y) < C\log\cos(d(x,y)/\kappa).$$

Consequently, for any $\phi \in \mathrm{C}(K^2)$ and $\eta < \eta_0$,

$$\begin{aligned}&\limsup_{\epsilon\to 0}\left|\int_{K^2}\phi(x,y)\mathrm{d}\frac{\pi_\epsilon(x,y)}{\cos\left(d(x,y)/\kappa\right)} - \int_{K^2}\phi(x,y)\mathrm{d}\frac{\pi(x,y)}{\cos\left(d(x,y)/\kappa\right)}\right| \\ \leq &\limsup_{\epsilon\to 0}\int_{d(x,y)/\kappa>\pi/2-\eta}|\phi(x,y)|\, e^{C\log\cos(d(x,y)/\kappa)/\epsilon}\mathrm{d}\mu_0(x)\mathrm{d}\mu_1(y) \\ &+ \int_{d(x,y)/\kappa>\pi/2-\eta}|\phi(x,y)|\,\mathrm{d}\frac{\pi(x,y)}{\cos\left(d(x,y)/\kappa\right)} \\ &+ \left|\int_{d(x,y)/\kappa\leq\pi/2-\eta}\phi(x,y)\mathrm{d}\frac{\pi_\epsilon(x,y)}{\cos\left(d(x,y)/\kappa\right)} - \int_{d(x,y)/\kappa\leq\pi/2-\eta}\phi(x,y)\mathrm{d}\frac{\pi(x,y)}{\cos\left(d(x,y)/\kappa\right)}\right| \\ \leq &\int_{d(x,y)/\kappa>\pi/2-\eta}|\phi(x,y)|\,\mathrm{d}\frac{\pi(x,y)}{\cos\left(d(x,y)/\kappa\right)}\end{aligned}$$

and the desired convergence is obtained by taking $\eta \to 0$.

6 Numerical examples

6.1 Preliminaries

For our numerical experiments we discretize measures as weighted point clouds. The discrete (unbalanced) optimal transport problems are solved via entropic regularization and a generalized

Sinkhorn algorithm [14] (see [43] for some algorithmic details). The length scale of the entropic blur is set to be comparable to the nearest-neighbour distance of the discrete point clouds to somewhat dampen discretization artefacts. The code (Python/Numpy) for the discrete linearized optimal transport analysis is an updated version of that used in [11], accounting for some extensions introduced in the present article. This includes support for the SHK metric, Section 3.4; an improved implementation of the barycentric projection for the HK metric, as analyzed in Section 5.3; and support for manifolds as base spaces. The updated code is available at `https://gitlab.gwdg.de/bernhard.schmitzer/linot`.

In each of the following experiments, a set of sample measures is mapped into the tangent space of a reference measure. Principal component analysis is then applied to the obtained set of tangent vectors, i.e. the set is centered and its empirical covariance matrix is diagonalized. To simplify this step, all discrete tangent vectors are transformed such that the respective Riemannian inner products ($\mathbb{L}^2(\mu)$ of the velocity and growth fields, see Proposition 2.8, Proposition 3.16, and Proposition 3.22) are simply given by the standard Euclidean inner product of the transformed vectors. One can then study the spectrum of the covariance matrix, and low-dimensional PCA embeddings, i.e. projecting the sample tangent vectors on the dominating eigenvectors. Movements in tangent space can be visualized by applying the exponential map (relative to the mean of the tangent vector samples), for instance to obtain an interpretation of the eigenvectors.

Applying the exponential map to a tangent vector associated to a measure given by a weighted point cloud, one will obtain a Lagrangian deformation of the point cloud. If the reference point cloud were a rigid Cartesian grid (or an equivalent structure on a manifold), after applying the exponential map, this is in general no longer true. Therefore, to deal with point density fluctuations we rasterize the outcomes of the exponential map back to a reference grid for better visualization and comparison. To dampen Moiré artefacts, a small additional blur kernel is applied to the rasterized images.

6.2 Linearized HK and SHK in the plane

The numerical experiments in [11, Section 5.2] focused on the comparison between the linearization of W_2 and HK. In particular they demonstrated the susceptibility of balanced transport to misinterpret small mass fluctuations happening over large ranges as transport, whereas the Hellinger–Kantorovich can explain such mass discrepancies purely as mass vanishing/creation thus better capturing the local transport variations in the data set. In this article we focus on a comparison between HK and SHK. We demonstrate that the flexibility of HK to change the total mass of measures introduces a systematic bias towards lower masses when averaging between samples. On one hand, this bias implies that averaged measures have too little mass, and also it increases the apparent dimension of the dataset, leading to a 'widened' PCA spectrum with additional eigenvectors corresponding to pure mass changes that are not present in the samples. The SHK metric retains the ability to create and destroy mass locally, but is subjected to the constraint that the total mass must be preserved. It is therefore still able to deal gracefully with long range mass fluctuations but avoids the bias of HK.

Dirac measures along a line. The aforementioned bias can be studied analytically in the case of single Dirac measures on the interval $[0; L]$, for $0 < L < \kappa\pi/2$ (assuming the upper bound

to avoid the pure Hellinger regime for simplicity). In this case one has that

$$\begin{aligned} \mathrm{HK}_\kappa^2(m_0\delta_{x_0}, m_1\delta_{x_1}) &= \kappa^2\left(m_0 + m_1 - 2\sqrt{m_0 m_1}\,\mathrm{Cos}(|x_0 - x_1|/\kappa)\right) \\ &= \kappa^2 \left\|\sqrt{m_0}\exp(ix_0/\kappa) - \sqrt{m_1}\exp(ix_1/\kappa)\right\|_{\mathbb{C}}^2 \end{aligned} \tag{6.1}$$

which can be obtained by checking that the optimal transport plan in (3.1) is given by

$$\pi = \mathrm{Cos}\left(|x_1 - x_0|/\kappa\right)\sqrt{m_0 m_1}\delta_{x_0}\otimes\delta_{x_1}.$$

As can be seen from (6.1), on the subset of Dirac measures $\{m\delta_x | m \geq 0, x \in [0; L]\}$ the HK_κ metric is flat (since it can be isometrically embedded into $\mathbb{C}$) and the tangent space approximation around, for instance, $\mu_0 := \delta_{x_0}$, $x_0 := L/2$ is exact. The one-dimensional parametric family of probability Dirac measures $\{\delta_x | x \in [0; L]\}$ corresponds to a circle segment in $\mathbb{C}$ and therefore also in the tangent space of μ_0. So a PCA of the embedded measures will yield two non-zero eigenvalues, and the corresponding PCA projection will show the circle shape. The mean of the embedded measures will then lie strictly within the disk, which is the aforementioned bias towards lower masses. The argument via the local isometry with $\mathbb{C}$ can be confirmed explicitly by evaluating the logarithmic map $(v_0, \alpha_0, \mu_1^\perp) = \mathrm{Log}_{\mu_0}^{\mathrm{HK}_\kappa}(\mu_1)$ for $\mu_1 := \delta_{x_1}$, $x_1 \in [0; L]$, (3.15) one finds

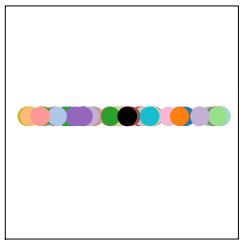
$$v_0(x_0) = \kappa \sin\left(\frac{x_1 - x_0}{\kappa}\right), \qquad \alpha_0(x_0) = 2\left(\cos\left(\frac{|x_1 - x_0|}{\kappa}\right) - 1\right), \qquad \mu_1^\perp = 0. \tag{6.2}$$

Indeed, one sees that the embedding describes a circle segment as x_1 is varied.

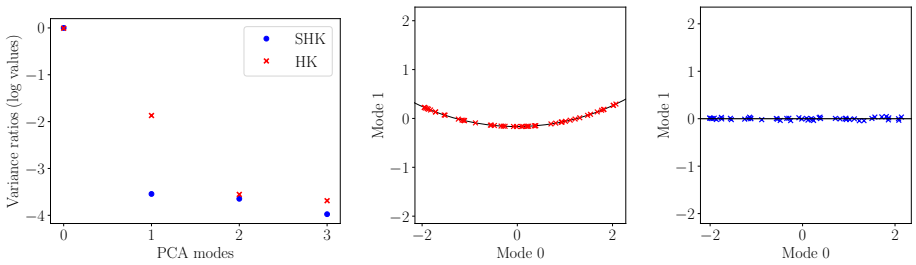
In the SHK metric, the creation and destruction of mass is still allowed, but the total mass must be preserved. In the case of single Dirac measures, this means that the metric structure reduces to the standard W_2 distance. Indeed, plugging (6.1) for $m_0 = m_1 = 1$ into (3.25) yields $\mathrm{SHK}_\kappa(\delta_{x_0}, \delta_{x_1}) = |x_0 - x_1|$. With this and (6.2) one then obtains from (3.28)

$$v_0^S(x_0) = x_1 - x_0, \qquad \alpha_0^S(x_0) = 0, \qquad (\mu_1^S)^\perp = 0.$$

So in this case, the *SHK* and W_2 metrics yield the same embedding, one in which the one-dimensional structure of the dataset is more faithfully represented.

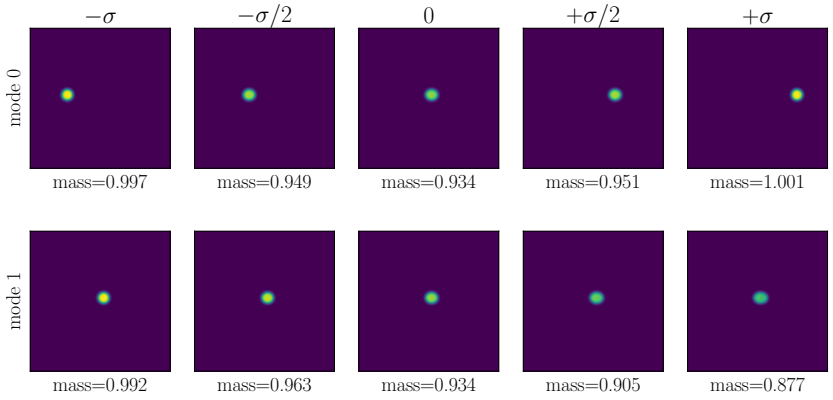


(a) Generated data, reference measure supported on black disk

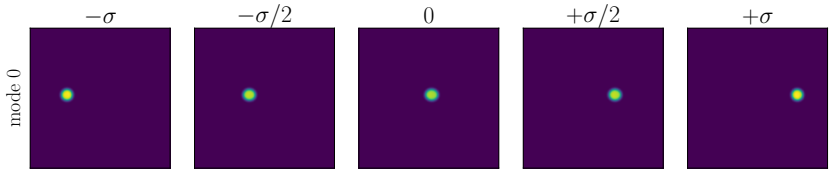


(b) Ratios of explained variance for first 4 PCA modes (log scale) and 2-dimensional PCA projection for HK and SHK embeddings.

Figure 5: Principal component analysis for the HK and SHK embeddings of the 1-dimensional disk data. In b one can observe the circular structure for HK.



(a) Exponential map along first two PCA modes for HK.



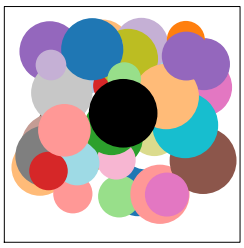
(b) Exponential map along first PCA mode for SHK.

Figure 6: Visualizing the dominant PCA eigenvectors via the exponential map for HK and SHK on the 1-dimensional disk data.

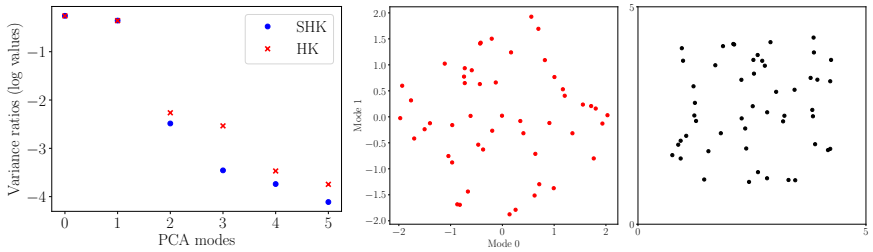
Small disks (with fixed radius) along a line. Next, we consider uniform probability measures on disks of radius $R := 0.2$ with centers sampled at random uniformly from the segment $[R; L-R] \times \{0\}$ for $L := 5$ and set $\kappa := 6$. The reference measure μ_0 is taken to be the disk centered in the middle of the segment (see Figure 5a). For these choices of R, L and κ, the pure Hellinger parts in both HK and SHK metrics are zero.

The stability results from Proposition 5.8 and Corollary 5.13 suggest that the embedded data will be close to the Dirac example from the previous paragraph. Indeed, the PCA spectra and two-dimensional embeddings for HK and SHK, shown in Figure 5b, confirm this. For both metrics the dominant mode corresponds to the translations along the chosen horizontal axis (cf. Figure 6a and Figure 6b). For HK this translation is combined with a change in mass, as discussed in the previous paragraph. Accordingly, in the PCA spectrum for HK the second largest eigenvalue is substantially larger than for SHK and it corresponds to the 'radial' direction of pure mass changes to compensate for the bias. The respective PCA embeddings feature a circle for HK and an (approximately) straight line for SHK.

Disks with random radii and positions in a box In the next experiment the data consists of uniform probability measures on disks of varying radii and centers, with radii and centers sampled uniformly from the interval $[R_{\min}; R_{\max}] := [0.3; 0.7]$) and $[R_{\max}, L - R_{\max}]^2$ respectively, where $L := 5$ and we set again $\kappa := 6$. The reference measure for the tangent space approximation corresponds to the disk of radius $R = 0.5$, centered at $(L/2, L/2)$, see for instance Figure 5a. Applying PCA to the tangent space embeddings of these measures for both HK and SHK yields the spectra and projection shown in Figure 7b. Shooting along the modes via the exponential map is shown in Figure 8. From these Figures we conclude that for HK and SHK the first two modes encode essentially the positions of the centers of the disk (up to an exchange, rescaling and small rotation of the axes, compare with the right-most Plot of Figure 7b). Moreover, mode



(a) Generated data, reference measure supported on black disk.



(b) Ratios of explained variance for first PCA modes for HK and SHK (log scale), 2d PCA embedding for HK (the 2d SHK embedding looks virtually identical), and scatter plot of the sample disk centers.

Figure 7: Principal component analysis for the HK and SHK embeddings of the 3-dimensional disk data. Up to reflection and rotation the 2d PCA projection and the scatter plot of the sample disk centers look almost identical suggesting that the first two modes capture the translation of the samples.

2 in SHK captures the radius variations. However, for HK mode 2 corresponds instead mostly to the mass bias, as in the previous experiments, whereas the radius variations are captured by mode 3.

6.3 Linearized HK and SHK metrics on the sphere

In this section we give an example for linearized optimal transport when the base space X is the 2-dimensional unit sphere. We set $\kappa = 4$. As samples we pick uniform probability measures on balls (for the sphere metric) centered on points along the equator, whereas the reference measure is a ball centered on the north pole, see Figure 9a). In this experiment the first four PCA eigenvalues for HK and SHK are essentially identical and the first two components capture approximately the position of the samples from the perspective of the north pole, resulting in a circular structure, see Figure 9b. Exponential shooting confirms this interpretation (see fig. 10) showing the reference measure being pushed along two orthogonal directions relative to the north pole. This results in a strong deformation of the balls which has to be corrected by the two subsequent modes, albeit carrying a much lower variance.

Acknowledgements. The authors gratefully acknowledge support from the DFG CRC 1456 Mathematics of the Experiment A03.

7 References

[1] M. Agueh and G. Carlier. Barycenters in the Wasserstein space. *SIAM J. Math. Anal.*, 43(2):904–924, 2011.

[2] Pedro C Álvarez-Esteban, E Del Barrio, JA Cuesta-Albertos, and C Matrán. A fixed-point approach to barycenters in wasserstein space. *Journal of Mathematical Analysis and Applications*, 441(2):744–762, 2016.

[3] L. Ambrosio, N. Fusco, and D. Pallara. *Functions of Bounded Variation and Free Discontinuity Problems*. Oxford Science Publications. Clarendon Press, 2000.

[4] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows in Metric Spaces and in the Space of Probability Measures*. Springer Science & Business Media, 2005.

[5] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223. PMLR, 06–11 Aug 2017.

[6] Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. Iterative Bregman projections for regularized transportation problems. *SIAM J. Imaging Sci.*, 37(2):A1111–A1138, 2015.

[7] Jérémie Bigot, Raùl Gouet, Thierry Klein, and Alfredo López. Geodesic PCA in the Wasserstein space by convex PCA. *Ann. Inst. H. Poincaré Probab. Statist.*, 53(1):1–26, 2017.

[8] Jérémie Bigot and Thierry Klein. Characterization of barycenters in the Wasserstein space by averaging optimal transport maps. *ESAIM: PS*, 22:35–57, 2018.

[9] Y. Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Comm. Pure Appl. Math.*, 44(4):375–417, 1991.

[10] Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on pure and applied mathematics*, 44(4):375–417, 1991.

[11] Tianji Cai, Junyi Cheng, Bernhard Schmitzer, and Matthew Thorpe. The linearized Hellinger–Kantorovich distance. *SIAM Journal on Imaging Sciences*, 15(1):45–83, 2022.

[12] Lénaïc Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in Neural Information Processing Systems 31 (NIPS 2018)*, 2018.

[13] Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. An interpolating distance between optimal transport and Fisher–Rao metrics. *Found. Comp. Math.*, 18(1):1–44, 2018.

[14] Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of Computation*, 87(314):2563–2609, 2018.

[15] Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Unbalanced optimal transport: Dynamic and Kantorovich formulations. *J. Funct. Anal.*, 274(11):3090–3123, 2018.

[16] Marco Cuturi and Arnaud Doucet. Fast computation of wasserstein barycenters. In *International conference on machine learning*, pages 685–693. PMLR, 2014.

[17] A. Delalande and Q. Merigot. Quantitative stability of optimal transport maps under variations of the target measure. arXiv:2103.05934, 2021.

[18] Björn Engquist, Brittany D. Froese, and Yunan Yang. Optimal transport for seismic full waveform inversion. *Communications in Mathematical Sciences*, 14(8):2309–2330, 2016.

[19] Jean Feydy, Benjamin Charlier, François-Xavier Vialard, and Gabriel Peyré. Optimal transport for diffeomorphic registration. In *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part I 20*, pages 291–299. Springer, 2017.

[20] Alessio Figalli, Young-Heon Kim, and Robert J McCann. When is multidimensional screening a convex program? *Journal of Economic Theory*, 146(2):454–478, 2011.

[21] Alfred Galichon. *Optimal Transport Methods in Economics*. Princeton University Press, 2017.

[22] Thomas Gallouët. *Optimal transport: Regularity and Applications*. PhD thesis, Ecole normale supérieure de lyon-ENS LYON, 2012.

[23] Thomas Gallouët, Roberta Ghezzi, and François-Xavier Vialard. Regularity theory and geometry of unbalanced optimal transport. arXiv:2112.11056, 2021.

[24] Wilfrid Gangbo and Robert J. McCann. The geometry of optimal transportation. *Acta Math.*, 177(2):113–161, 1996.

[25] N. Gigli. On Holder continuity-in-time of the optimal transport map towards measures along a curve. *Proceedings of the Edinburgh Mathematical Society*, 54(2):401–409, 2011.

[26] V. Khurana, H. Kannan, A. Cloninger, and C. Moosmüller. Supervised learning of sheared distributions using linearized optimal transport. arXiv:2201.10590, 2022.

[27] Soheil Kolouri, Se Rim Park, and Gustavo K. Rohde. The radon cumulative distribution transform and its application to image classification. *IEEE Transactions on Image Processing*, 25(2):920–934, 2016.

[28] Soheil Kolouri, Akif B Tosun, John A Ozolek, and Gustavo K Rohde. A continuous linear optimal transport approach for pattern analysis in image datasets. *Pattern recognition*, 51:453–462, 2016.

[29] Stanislav Kondratyev, Léonard Monsaingeon, and Dmitry Vorotnikov. A new optimal transport distance on the space of finite Radon measures. *Adv. Differential Equations*, 21(11-12):1117–1164, 2016.

[30] Vaios Laschos and Alexander Mielke. Geometric properties of cones with applications on the hellinger–kantorovich space, and a new distance on the space of probability measures. *Journal of Functional Analysis*, 276(11):3529–3576, 2019.

[31] Matthias Liero, Alexander Mielke, and Giuseppe Savaré. Optimal entropy-transport problems and a new hellinger–kantorovich distance between positive measures. *Inventiones mathematicae*, 211(3):969–1117, 2018.

[32] G. Loeper. On the regularity of solutions of optimal transportation problems. *Acta Math*, 202:241–283, 2009.

[33] John Lott. Some geometric calculations on Wasserstein space. *Comm. Math. Phys.*, 277:423–437, 2008.

[34] Xi-Nan Ma, Neil S. Trudinger, and Xu-Jia Wang. Regularity of potential functions of the optimal transportation problem. *Arch. Rat. Mech. Analysis*, 177:151–183, 2005.

[35] Robert J McCann. Polar factorization of maps on riemannian manifolds. *Geometric & Functional Analysis GAFA*, 11(3):589–608, 2001.

[36] Luca Nenna and Brendan Pass. Transport type metrics on the space of probability measures involving singular base measures. arXiv:2201.00875, 2022.

[37] F. Otto. The geometry of dissipative evolution equations: the porous medium equation. *Comm. Partial Differential Equations*, 26(1-2):101–174, 2001.

[38] Se Rim Park, Soheil Kolouri, Shinjini Kundu, and Gustavo K. Rohde. The cumulative distribution transform and linear pattern classification. *Applied and Computational Harmonic Analysis*, 45(3):616–641, 2018.

[39] Serim Park and Matthew Thorpe. Representing and learning high dimensional data with the optimal transport map from a probabilistic viewpoint. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7864–7872, 2018.

[40] Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5–6):355–607, 2019.

[41] Takashi Sakai. *Riemannian geometry*, volume 149. American Mathematical Soc., 1996.

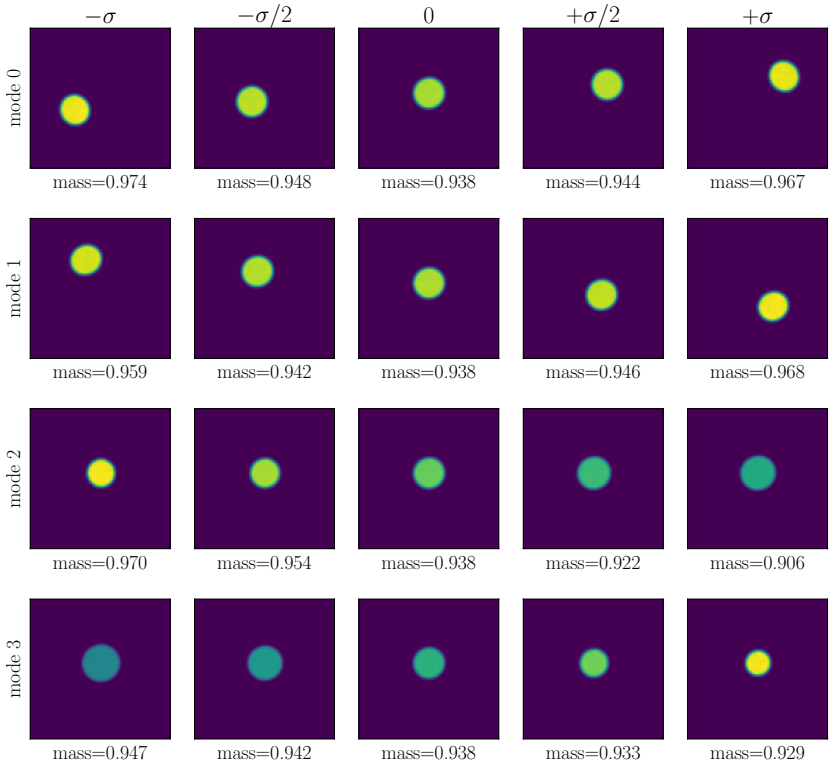
[42] Filippo Santambrogio. *Optimal transport for applied mathematicians*, volume 87 of *Progress in Nonlinear Differential Equations and Their Applications*. Birkhäuser Boston, 2015.

[43] Bernhard Schmitzer. Stabilized sparse scaling algorithms for entropy regularized transport problems. *SIAM J. Sci. Comput.*, 41(3):A1443–A1481, 2019.

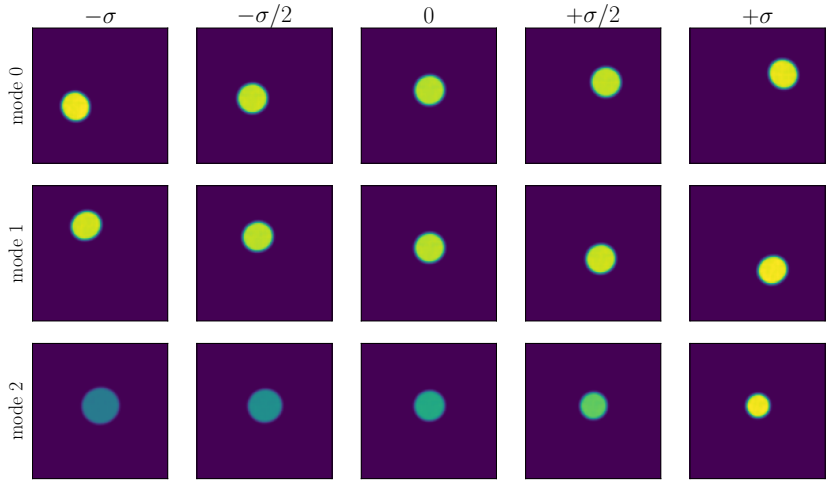
[44] A. W. van der Vaart. *Asymptotic statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

[45] Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, 2009.

[46] 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.

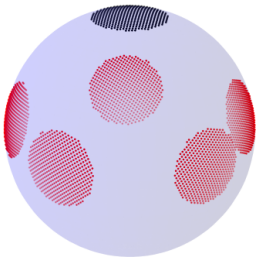


(a) Exponential map along first four PCA modes for HK.

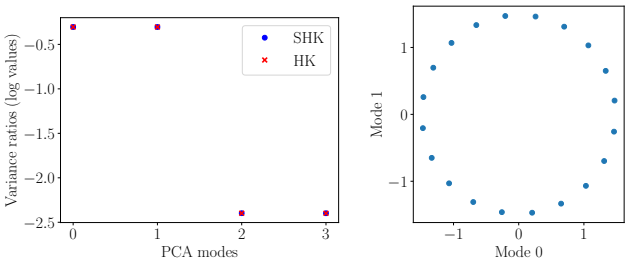


(b) Exponential map along first three PCA modes for SHK.

Figure 8: Visualizing the dominant PCA eigenvectors via the exponential map for HK and SHK on the 1-dimensional disk data.

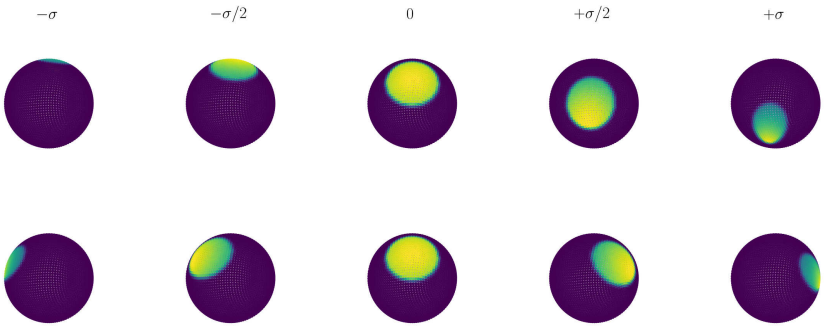


(a) Subset of generated data, reference measure in Black supported on the 'north pole cap'.

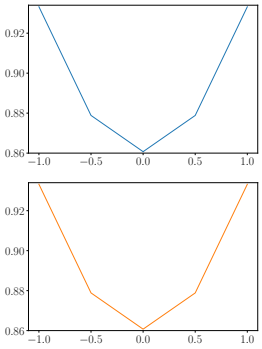


(b) Ratios of explained variance for first PCA modes (log scale), and 2d PCA projection (for SHK embedding, HK embedding yields the same plot).

Figure 9: Principal component analysis for the HK and SHK embeddings of the sphere example.



(a) Exponential map along first two PCA modes for HK.



(b) Mass variations during the shooting for each mode

Figure 10: Visualizing the dominant PCA eigenvectors via the exponential map for HK for the sphere example.