Sliced Hellinger–Kantorovich

Authors TBC

Friday 15th April, 2022

Abstract

Optimal transport distances, such as the Wasserstein distance, are a powerful tool for image and signal analysis, and since their introduction into Generative Adversarial Networks are becoming popular in the machine learning community. In the Wasserstein distance one pays a cost for transporting mass from one signal/image to another that is proportional to the distance between masses. In this setting all mass within the two images must be matched; hence the signals/images are required to be normalised. Recently, the Hellinger–Kantorovich distance was proposed that allows mass creation/destruction and, in particular, introduces a limit on the distance mass is transported. Many of the desirable theoretical properties of the Wasserstein distance, such as geodesics, are also present for the Hellinger–Kantorivich distance and one has the advantage of being able to apply the Hellinger–Kantrovich distance directly to unnormalised signals/images. A disadvantage of optimal transport distances are that they are usually expensive (despite recent advances) to compute in dimensions $d \geq 2$. However, in one dimension computing the Wasserstein distance reduces to solving a sorting problem which can be computed efficiently. Taking advantage of this, the sliced Wasserstein distance was proposed which, for $d \geq 2$, replaces the d-dimensional Wasserstein distance with an average of 1-dimensional Wasserstein distances. We extend this framework to the Hellinger–Kantorovich distance and show that we significantly reduce the computation time.

To do:

- **Write introduction.**
- **Complete Section 2.4.**
- **Relationship between minimisers of the two Kantorovich forms in Section 3.2.**
- **Prove existence of transport maps for the Hellinger–Kantorovich distance between point measures in Section 3.4.**
- **Finish proof of Corollary 3.4.**
- **Finish Section 3.5.**
- **Write Section 4.**
- **Write Section 5.**
- **Write Section 6.**

1 Introduction

1.1 Paper Overview

1.2 Notation

2 The Wasserstein Distance

We follow the historical development of optimal transport, in the context of the Wasserstein distance, by first reviewing the Monge formulation (1781) [9] followed by the Kantorovich formulation (1942) [7]. We then look

at the special case when the probability measures are over the real line. Finally, we include an introduction to the Benamou-Brenier formulation (2000) [3] as this forms the starting point for defining the Hellinger–Kantorovich distance in the following section. Our exposition is brief and we refer to [1, 11–13] for a more in-depth introduction to the theory, or to [10] for computational and applied aspects.

2.1 Monge Formulation

Consider two probability measures μ and ν on a domain $\Omega \subset \mathbb{R}^d$. For our purposes we can consider μ and ν as images or signals and typically the domain will be a subset of $\mathbb{R}^d$. The measures μ and ν are assumed to have densities ρ_μ and ρ_ν, i.e.

$$\mu(A) = \int_A \rho_\mu(x)\,\mathrm{d}x, \quad \nu(B) = \int_B \rho_\nu(y)\,\mathrm{d}y$$

for all open sets A and B. We let $c(x, y)$ be the cost of transporting one unit of mass from x to y. The optimal transport cost measures the total cost of rearranging mass in the distribution μ into the distribution ν. More precisely, we say that T is a transport map between μ and ν if $T_\sharp\mu(A) := \mu(T^{-1}(A)) = \nu(A)$ for all open sets A. We call the set of transport maps $\mathrm{MP}(\mu, \nu)$. Transport maps do not always exist, and a necessary (but not sufficient) condition for the existence of transport maps is $\mu(\Omega) = \nu(\Omega)$ hence we assume that μ and ν are probability measures. If T is smooth and bijective then we can equivalently write

$$\det(DT(x))\rho_\nu(T(x)) = \rho_\mu(x)\ \forall x.$$

The Monge optimal transport problem is

$$\mathrm{M_W}(\mu, \nu) = \inf_{T \in \mathrm{MP}(\mu,\nu)} \int_\Omega c(x, T(x))\rho_\mu(x)\,\mathrm{d}x.$$

Typically one chooses $c(x, y) = |x - y|^p$; in this case the p-Wasserstein distance is defined as $d_{\mathrm{W}^p}(\mu, \nu) = \sqrt[p]{\mathrm{M_W}(\mu, \nu)}$. When $p = 1$ the distance is also known as the earth movers distance, we will mostly consider the case $p = 2$.

2.2 Kantorovich Formulation

In the Monge formulation mass $\mathrm{d}\mu(x) = \rho_\mu(x)\mathrm{d}x$ is transported from x to y via the transport map T. The Kantorovich formulation relaxes this by allowing mass to be distributed to multiple locations. Rather than looking for transport maps we look for transport plans which are measures π on the product space $\Omega \times \Omega$ and $\mathrm{d}\pi(x, y)$ can be understood as the amount of mass that is moved from x to y. Of course the total amount of mass removed from x should still be $\mathrm{d}\mu(x)$ and the amount of mass arriving at y should be $\mathrm{d}\nu(y)$. This introduces the marginal constraints: $P_{X\sharp}\pi = \pi(\cdot \times \Omega) = \mu(\cdot)$ and $P_{Y\sharp}\pi = \pi(\Omega \times \cdot) = \nu$, where $P_X(x, y) = x$ and $P_Y(x, y) = y$. We say π is a transport plan if it is a probability measure on $\Omega \times \Omega$ satisfying the marginal constraints. We call the set of transport plans $\Pi(\mu, \nu)$. The Kantorovich optimal transport problem is

$$\mathrm{K_W}(\mu, \nu) = \min_{\pi \in \Pi(\mu,\nu)} \int_{\Omega\times\Omega} c(x, y)\,\mathrm{d}\pi(x, y).$$

We note that this is a linear optimisation problem with convex constraints. The Monge and Kantorovich forms can be connected as follows: let us assume optimal $T^* \in \mathrm{MP}(\mu, \nu)$ and $\pi^* \in \Pi(\mu, \nu)$ exist, then

$$\pi^* = (\mathrm{Id} \times T^*)_\sharp\mu$$

where $(\mathrm{Id} \times T)_\sharp\mu(A \times B) = \mu(\{x : x \in A \text{ and } T(x) \in B\})$.

Optimal transport plans respect the ordering in the sense of the following theorem.

Theorem 2.1 (Theorem 5.10 [13])**.** *Assume $c : \mathbb{R}^d \times \mathbb{R}^d \to [0, +\infty) \cup \{+\infty\}$ is lower semi-continuous. Let π^* be an optimal transport plan in the Kantorovich formulation of optimal transport, i.e. π^* acheives the infimum in $\mathrm{K_W}(\mu, \nu)$. Then π^* is c-cylically monotone. I.e. for every $n \in \mathbb{N}$, any choice of $(x_i, y_i)_{i=1}^n \subset \mathrm{spt}(\pi^*)$ and every permutation $\kappa : \{1, \dots, n\} \to \{1, \dots, n\}$ we have*

$$\sum_{i=1}^n c(x_i, y_i) \le \sum_{i=1}^n c(x_i, y_{\kappa(i)}).$$

2.3 Dual Formulation

As remarked in the previous section, the Kantorovich formulation of the optimal transport problem is a linear programme with convex constraints. As such it has a dual form which is given by

$$\mathrm{D_W}(\mu,\nu) = \sup\left\{ \int_\Omega \varphi \,\mathrm{d}\mu + \int_\Omega \psi \,\mathrm{d}\nu \,:\, \varphi \in \mathrm{L}^1(\mu), \psi \in \mathrm{L}^1(\nu), \varphi(x) + \psi(y) \le c(x,y) \text{ a.e. } x, y \right\}$$

where $\mathrm{D_W}(\mu,\nu) = \mathrm{K_W}(\mu,\nu)$ holds when $c : \Omega \times \Omega \to [0,+\infty)$ is lower semi-continuous

2.4 The One-Dimensional Wasserstein Distance

to do... 1. montonicity of optimal maps, 2. Wasserstein reduces to a sorting problem, 3. algorithms for sorting and computational complexity, 4. definition of sliced wasserstein.

2.5 Benamou–Brenier Formulation

In the third and final formulation of the Wasserstein distance we think of the measures μ and ν as two fluids; the Wasserstein distance is the minimal amount of kinetic energy required to rearrange the first fluid into the second. This formulation is special to the case $p = 2$. We say $\rho : [0,1] \times \Omega \to \mathbb{R}$ and $v : [0,1] \times \Omega \to \mathbb{R}^d$ satisfies the continuity equation if

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho v) = 0.$$

Note that if (ρ, v) satisfy the continuity equation then $\int_\Omega \rho(t,x)\,\mathrm{d}x$ is constant in t, in particular, mass is neither created nor destroyed. We denote by $\mathrm{CE}(\mu,\nu)$ the set of (ρ, v) satisfying the continuity equation with $\rho(0,\cdot) = \rho_\mu$ and $\rho(1,\cdot) = \rho_\nu$. Now we define the Benamou–Brenier formulation by

$$\mathrm{BB_W}(\mu,\nu) = \min_{(\rho,v)\in\mathrm{CE}(\mu,\nu)} \int_0^1 \int_\Omega |v(t,x)|^2 \rho(t,x)\,\mathrm{d}x\,\mathrm{d}t.$$

When $p = 2$, $c(x,y) = |x-y|^2$, and the densities ρ_μ, ρ_ν exist then

$$\sqrt{\mathrm{BB_W}(\mu,\nu)} = \sqrt{\mathrm{K_W}(\mu,\nu)} = \sqrt{\mathrm{M_W}(\mu,\nu)} = d_{\mathrm{W}^2}(\mu,\nu)$$

are all equivalent formulations of the 2-Wasserstein distance.

3 The Hellinger–Kantorovich Distance

In the previous section we first defined the Wasserstein distance in the Monge formulation and proceeded to derive the Kantorovich and Benamou–Brenier formulation. For the Hellinger–Kantorovich distance we follow the reverse path and start with the Benamou–Brenier formulation then proceed to the Kantorovich and finally Monge formulations. Finally, we consider the one-dimensional special case of the Hellinger–Kantorovich distance.

3.1 Benamou–Brenier Formulation

A natural way to extend to measures with different masses is to allow for creation and destruction of mass in the continuity equation, and penalise any creation and destruction in the objective functional. We say that (ρ, v, ζ) satisfy the continuity equation with source if they satisfy

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho v) = \zeta.$$

Given non-negative measures μ and ν (not necessarily probability measures) with densities ρ_μ and ρ_ν respectively we consider the boundary conditions $\rho(0,\cdot) = \rho_\mu$, $\rho(1,\cdot) = \rho_\nu$. We denote the set of such (ρ, v, ζ) by

$\mathrm{CES}(\mu,\nu)$. There are several ways one could penalise the creation and destruction (which is captured by ζ). One way is to use

$$\mathrm{FR}(\rho,\zeta)=\int_0^1\int_\Omega\left(\frac{\zeta(t,x)}{\rho(t,x)}\right)^2\rho(t,x)\,\mathrm{d}x\,\mathrm{d}t,$$

in particular the objective functional becomes

$$\mathrm{BB}_{\mathrm{HK}_\lambda}(\mu,\nu)=\min_{(\rho,v,\zeta)\in\mathrm{CES}(\mu,\nu)}\int_{[0,1]\times\Omega}\left(\|v\|^2+\lambda\left(\frac{\zeta}{\rho}\right)^2\right)\rho\,\mathrm{d}(t,x)$$

where λ is a parameter that controls the relative importance of transport and pure creation/destruction.

The objective FR is related to the Fisher–Rao metric. Indeed $d_{\mathrm{FR}}(\mu,\nu)=\inf_{(\rho,0,\zeta)\in\mathrm{CES}(\mu,\nu)}\mathrm{FR}(\rho,\zeta)$ is a metric on the space of probability measures. Moreover, it is the unique reparametrisation invariant Riemannian metric on $\mathcal{P}(\Omega)$ [2].

The Hellinger–Kantorovich distance on the space of non-negative measures is defined as

$$d_{\mathrm{HK}_\lambda}(\mu,\nu)=\sqrt{\mathrm{BB}_{\mathrm{HK}_\lambda}(\mu,\nu)}.$$

As $\lambda\to0^+$ the Hellinger–Kantorovich distance converges to the Hellinger distance, and when $\lambda\to\infty$ the Hellinger–Kantorovich distance converges to the Wasserstein distance between μ and ν (assuming that $\mu(\Omega)=\nu(\Omega)$; otherwise it is the Wasserstein distance between the geometric means of μ and ν) [5, Theorem 3.2].

Intuitively one can see that, if comparing two diracs, mass may prefer being destroyed at the source and created at the target rather than transported from source to target (or a combination of destruction/creation and transport) when the transport distance is large. This observation is true and morover the maximum distance is $\sqrt{\lambda}\pi$ [5, Theorem 4.1].

In order to compute the optimal flow (ρ,v,ζ) between arbitrary non-negative measures μ and ν we start by considering two dirac masses $\mu=m_\mu\delta_x$ and $\nu=m_\nu\delta_y$, where $m_\mu,m_\nu\in(0,+\infty)$ and $\|x-y\|<\sqrt{\lambda}\pi$. The motivation is that once we can find the flow between two diracs then we can "decompose" our measure into weighted sums of dirac measures and — after we find how mass is coupled à la Kantorovich or Monge — we "build" flows for more complex measures. The Hellinger–Kantorovich distance between the above diracs is

$$d^2_{\mathrm{HK}_\lambda}(m_\mu\delta_x,m_\nu\delta_y)=4\lambda\left(m_\mu+m_\nu-2\sqrt{m_\mu m_\nu}\cos\left(\frac{\|x-y\|}{2\sqrt{\lambda}}\right)\right)$$

see [5, Corollary 4.1]. In the Wasserstein setting the geodesic moves mass at constant speed along a straight line between x and y; this is only partly true in the Hellinger–Kantorovich setting. If $x(t)=x(t;x,m_\mu,y,m_\nu)$ and $m(t)=m(t;x,m_\mu,y,m_\nu)$ are the position and amount of mass at time t then $\mu_t=m(t)\delta_{x(t)}$ is the constant speed geodesic, in the sense that $d_{\mathrm{HK}_\lambda}(\mu_s,\mu_t)=|t-s|d_{\mathrm{HK}_\lambda}(\mu,\nu)$ for all $s,t\in[0,1]$, where

$$x(t)=x+\frac{2\sqrt{\lambda}(y-x)}{\|y-x\|}\cos^{-1}\left(\frac{(1-t)\sqrt{m_\mu}+t\sqrt{m_\nu}\cos\left(\frac{\|y-x\|}{2\sqrt{\lambda}}\right)}{\sqrt{m(t)}}\right)$$

$$m(t)=(1-t)^2m_\mu+t^2m_\nu+2t(1-t)\sqrt{m_\mu m_\nu}\cos\left(\frac{\|y-x\|}{2\sqrt{\lambda}}\right).$$

Further, $(\rho,v,\zeta)\in\mathrm{CES}(\mu,\nu)$ given by

$$\begin{aligned}\rho(t,\cdot)&=m(t)\delta_{x(t)}\\ v(t,\cdot)&=\dot{x}(t)\delta_{x(t)}\\ \zeta(t,\cdot)&=\dot{m}(t)\delta_{x(t)}\end{aligned}$$

are the minimisers of $\mathrm{BB}_{\mathrm{HK}_\lambda}$ and therefore $d^2_{\mathrm{HK}_\lambda}(\mu,\nu)=\left(\|\dot{x}(t)\|^2+\lambda\left(\frac{\dot{m}(t)}{m(t)}\right)^2\right)m(t)$ for any $t\in[0,1]$ (assuming $m_\mu,m_\nu>0$), see [4, Proposition 3.7].

If $\|x-y\| > \sqrt{\lambda}\pi$ then the geodesic is simply the flow which destroys all mass at x and creates mass at y, in particular,

$$\begin{aligned}\rho(t,\cdot) &= (1-t)^2 I_\mu \delta_x + t^2 I_\nu \delta_y \\ v(t,\cdot) &= 0 \\ \zeta(t,\cdot) &= -2(1-t) I_\mu \delta_x + 2t I_\nu \delta_y.\end{aligned}$$

If $\|x-y\| = \sqrt{\lambda}\pi$ then any convex combination of the above formulations is a geodesic.

3.2 Kantorovich Formulation

There are two Kantorovich-type formulations of the Hellinger–Kantorovich distance. In the first, marginal constraints, rather than being imposed are penalised with respect to the Kullback-Leibler divergence and the cost of transport c_λ takes a different form to the quadratic cost in the Wasserstein case. In particular, we have the following proposition.

Proposition 3.1. *Define*

$$\mathrm{K}^{(\mathrm{soft})}_{\mathrm{HK}_\lambda}(\mu,\nu) = \left(2\sqrt{\lambda}\right)^d \inf\left(\int_{\Omega^2} c_\lambda(x,y)\,\mathrm{d}\pi(x,y) + 4\lambda\mathrm{KL}(P_{1\sharp}\pi|\mu) + 4\lambda\mathrm{KL}(P_{2\sharp}\pi|\nu)\right)$$

where the infimum is taken over positive measures π, KL *is the Kullback–Leibler divergence defined by*

$$\mathrm{KL}(\omega|\mu) = \int \varphi\left(\frac{\rho_\omega(x)}{\rho_\mu(x)}\right)\rho_\mu(x)\,\mathrm{d}x$$

for measures μ,ω with densities ρ_μ,ρ_ω respectively, $\varphi(t) = t\log(t) - t + 1$, and

$$c_\lambda(x,y) = \begin{cases} -8\lambda\log\cos\left(\frac{\|x-y\|}{2\sqrt{\lambda}}\right) & \text{if } \|x-y\| < \sqrt{\lambda}\pi \\ +\infty & \text{else.}\end{cases}$$

Assume... Then, $\mathrm{K}^{(\mathrm{soft})}_{\mathrm{HK}_\lambda}(\mu,\nu) = \mathrm{BB}_{\mathrm{HK}_\lambda}(\mu,\nu)$.

The proposition is proved in Appendix A.

In the form $\mathrm{K}^{(\mathrm{soft})}_{\mathrm{HK}_\lambda}(\mu,\nu)$ the Hellinger–Kantorovich distance can be seen as paying a cost of transport between $P_{1\sharp}\pi$ and $P_{2\sharp}\pi$, where the cost of transporting one unit of mass between x and y is given by $c_\lambda(x,y)$, and of paying a cost due to not matching marginals exactly, which is measured in terms of the Kullback–Leibler divergence.

The second type of Kantorovich formulation can be understood as first coupling between dirac measures and then optimally coupling between diracs. That is, if we know that mass $\mathrm{d}\pi_0(x,y)$ leaves from x to y, and mass $\mathrm{d}\pi_1(x,y)$ arrives at y from x (recalling that since mass can be created and destroyed we may have $\pi_0 \neq \pi_1$), then we can first compute the cost between the dirac at x of mass $\mathrm{d}\pi_0(x,y)$ and the dirac at y of mass $\mathrm{d}\pi_1(x,y)$, one has that this cost is given by [4, Proposition 3.5]

$$\begin{aligned}\mathrm{K}^{(\mathrm{soft})}_{\mathrm{HK}_\lambda}(\mathrm{d}\pi_0(x,y)\delta_x, \mathrm{d}\pi_1(x,y)\delta_y) &= 2\lambda\left(\mathrm{d}\pi_0(x,y) + \mathrm{d}\pi_1(x,y) - 2\sqrt{\mathrm{d}\pi_0(x,y)\mathrm{d}\pi_1(x,y)}\,\overline{\cos}\left(\frac{\|x-y\|}{2\sqrt{\lambda}}\right)\right) \\ &=: \hat{c}_\lambda(\mathrm{d}\pi_0(x,y), x, \mathrm{d}\pi_1(x,y), y)\end{aligned}$$

where $\overline{\cos}(t) = \cos\left(\min\left\{\frac{\pi}{2}, t\right\}\right)$. The amount of mass leaving x should equal $\mathrm{d}\mu(x)$ (the amount of mass at x under the measure μ) and the amount of mass arriving at y should equal $\mathrm{d}\nu(y)$ (the amount of mass at y under the measure ν). Hence, one can write [5, Theorem 5.6]

$$(1)\quad \mathrm{K}^{(\mathrm{hard})}_{\mathrm{HK}_\lambda}(\mu,\nu) = \inf\left\{\int_{\Omega^2}\hat{c}_\lambda\left(\frac{\mathrm{d}\pi_0}{\mathrm{d}\gamma}(x,y), x, \frac{\mathrm{d}\pi_1}{\mathrm{d}\gamma}(x,y), y\right)\mathrm{d}\gamma(x,y) \;:\; P_{1\#}\pi_0 = \mu, P_{2\#}\pi_1 = \nu\right\}$$

where γ is any measure satisfying $\pi_i \ll \gamma$ (and since $\hat{c}_\lambda$ is 1-homogeneous in its first and third arguments the choice of γ does not effect the cost).

Proposition 3.2. *Let π^* be a minimiser of $\mathrm{K}^{(\mathrm{soft})}_{\mathrm{HK}_\lambda}(\mu,\nu)$ and (π_0^*,π_1^*) minimisers of $\mathrm{K}^{(\mathrm{hard})}_{\mathrm{HK}_\lambda}(\mu,\nu)$. Assume... Then [Relationship between minimisers here... (see Prop 3.14 in v3 of LinHK paper)].*

The proof of the proposition is given in Appendix B.

As for the Wasserstein distance, optimal plans in the first Kantorovich form of the Hellinger–Kantorovich distance also satisfy a monotonicity property. If we let π^* be an optimal transport plan in the Hellinger–Kantorovich distance, i.e. π^* achieves the infimum in $\mathrm{K}^{(\mathrm{soft})}_{\mathrm{HK}_\lambda}(\mu,\nu)$. Then π^* is c_λ-cyclically monotone [8, Theorem 6.3]. I.e. for every $n\in\mathbb{N}$, any choice of $(x_i,y_i)_{i=1}^n\subset \mathrm{spt}(\pi^*)$ and every permutation $\kappa:\{1,\dots,n\}\to\{1,\dots,n\}$ we have

$$\sum_{i=1}^n c_\lambda(x_i,y_i)\le\sum_{i=1}^n c_\lambda(x_i,y_{\kappa(i)}).$$

3.3 Dual Formulation

Similarly to Section 2.3 the soft form of the Kantorovich formulation of the Hellinger–Kantorovich distance is a linear programme with convex constraints and therefore admits a dual formulation. In particular, we can define

$$\mathrm{D}_{\mathrm{HK}}(\mu,\nu)=\sup\left\{\int_\Omega\left(1-e^{-\varphi}\right)\,\mathrm{d}\mu+\int_\Omega\left(1-e^{-\psi}\right)\,\mathrm{d}\nu\,:\,\varphi,\psi\in\mathrm{C}^0(\Omega),\varphi(x)+\psi(y)\le c_\lambda(x,y)\,\forall x,y\right\}$$

and we have $\mathrm{D}_{\mathrm{HK}}(\mu,\nu)=\mathrm{K}^{(\mathrm{hard})}_{\mathrm{HK}_\lambda}(\mu,\nu)$ [6, Corollary 5.8].

3.4 Monge Formulation

If μ has a density then, as in the Wasserstein case, the minimiser π^* of the first Kantorovich form of HK is induced by a transport map [8, Theorem 6.6]. This means there exists $T^*:\Omega\to\Omega$ and a measure $\tilde\mu\in\mathcal{M}_+(\Omega)$ such that $\pi^*=(\mathrm{Id}\times T^*)_\sharp\tilde\mu$. Unlike the Wasserstein case it is not true that $\tilde\mu$ is equal to μ, nor is $T^*_\sharp\tilde\mu$ equal to ν.

In the Wasserstein case we also have that this result holds for uniform discrete measures with the same cardinality of the support, i.e. if $\mu=\frac{1}{n}\sum_{i=1}^n\delta_{x_i}$ and $\nu=\frac{1}{n}\sum_{j=1}^n\delta_{y_j}$. The proof is a consequence of more general optimality results regarding the characterisation of solutions to linear optimisation problems with convex constraints as extremal points, see [12, pp. 5-6]. In the Hellinger–Kantorovich case the (soft) Kantorovich formulation no longer fits into this framework, and hence it is not possible to write an analogous theorem for a uniform discrete application of the Hellinger–Kantorovich distance.

When a Monge map exists the minimisers of the second Kantorovich forms can be written:

$$\pi^*_X(x,y)=\begin{cases}\mu^\perp(x) & \text{if } x=y \text{ and } x\in\mathrm{supp}(\mu^\perp)\\ \mu(x) & \text{if } T^*(x)=y \text{ and } x\notin\mathrm{supp}(\mu^\perp)\\ 0 & \text{otherwise}\end{cases}\tag{2}$$

and

$$\pi^*_Y(x,y)=\begin{cases}\nu^\perp(y) & \text{if } x=y \text{ and } x\in\mathrm{supp}(\nu^\perp)\\ \nu(y) & \text{if } T^*(x)=y \text{ and } y\notin\mathrm{supp}(\nu^\perp)\\ 0 & \text{otherwise.}\end{cases}\tag{3}$$

The above characterisation allows us to write an expression for the Hellinger–Kantorovich distance in terms of the cost of destruction/creation and transport which we state in the next proposition for discrete measures.

Proposition 3.3. *Let $\mu=\frac{1}{n}\sum_{i=1}^n\delta_{x_i}$ and $\nu=\frac{1}{n}\sum_{j=1}^n\delta_{y_j}$ where all x_i,y_j are distinct. Let π^* be optimal for $\mathrm{K}^{(\mathrm{soft})}_{\mathrm{HK}_\lambda}(\mu,\nu)$. Assume there exists a map T^* and a measure $\tilde\mu$ satisfying $\pi^*=(\mathrm{Id}\times T^*)_\#\tilde\mu$ and define $\mathcal{I}_{\mathrm{tran}}=\{i\,:\,T^*(x_i)\neq x_i\}$ and $\mathcal{I}_{\mathrm{dest}}=\{i\,:\,T^*(x_i)=x_i\}$. Then,*

$$d_{\mathrm{HK}_\lambda}(\mu,\nu)^2=\sum_{i=1}^n\left(\mathbb{1}_{i\in\mathcal{I}_{\mathrm{tran}}}D(x_i,T^*(x_i))+2\mathbb{1}_{i\in\mathcal{I}_{\mathrm{dest}}}\bar D\right)$$

where

$$D(x,y) = \frac{8\lambda}{n}\left(1 - \overline{\cos}\left(\frac{\|x-y\|}{2\sqrt{\lambda}}\right)\right)$$
$$\bar{D} = \frac{4\lambda}{n}.$$

Proof. We note that $\mathrm{spt}(\pi_X^*) \subseteq \{x_i\}_{i=1}^n \times \left(\{x_i\}_{i=1}^n \cup \{y_j\}_{j=1}^n\right)$ and $\mathrm{spt}(\pi_Y^*) \subseteq \left(\{x_i\}_{i=1}^n \cup \{y_j\}_{j=1}^n\right) \times \{y_j\}_{j=1}^n$. Moreover, by (2) and (3) we have

$$\pi_X^*(x,y)\begin{cases} \frac{1}{n} & \text{if } x \in \mathrm{dom}(T^*) \text{ and } y = T^*(x) \\ \frac{1}{n} & \text{if } x \in \{x_i\}_{i=1}^n \setminus \mathrm{dom}(T^*) \text{ and } y = x \\ 0 & \text{else,} \end{cases}$$
$$\pi_Y^*(x,y)\begin{cases} \frac{1}{n} & \text{if } x \in \mathrm{dom}(T^*) \text{ and } y = T^*(x) \\ \frac{1}{n} & \text{if } y \in (\{y_j\}_{j=1}^n \setminus \mathrm{ran}(T^*) \text{ and } y = x \\ 0 & \text{else.} \end{cases}$$

If $x = x_i$ then $x \in \mathrm{dom}(T^*)$ implies $i \in \mathcal{I}_{\mathrm{tran}}$ and $x \notin \mathrm{dom}(T^*)$ implies $i \in \mathcal{I}_{\mathrm{dest}}$. We also define $\mathcal{I}_{\mathrm{crea}} = \{j : y_j \notin \mathrm{ran}(T^*)\}$. Substituting π_X and π_Y into (1) we have

$$\begin{aligned} d_{\mathrm{HK}_\lambda}(\mu,\nu)^2 &= \sum_{i,j=1}^n \hat{c}_\lambda\left(\pi_X^*(x_i,y_j), x_i, \pi_Y^*(x_i,y_j), y_j\right) + \sum_{i,j=1}^n \hat{c}_\lambda\left(\pi_X^*(x_i,x_j), x_i, \underbrace{\pi_Y^*(x_i,x_j)}_{=0}, x_j\right) \\ &\quad + \sum_{i,j=1}^n \hat{c}_\lambda\left(\underbrace{\pi_X^*(y_i,y_j)}_{=0}, y_i, \pi_Y^*(y_i,y_j), y_j\right) + \sum_{i,j=1}^n \hat{c}_\lambda\left(\underbrace{\pi_X^*(y_i,x_j)}_{=0}, x_i, \underbrace{\pi_Y^*(y_i,x_j)}_{=0}, y_j\right) \\ &= \sum_{i\in\mathcal{I}_{\mathrm{tran}}} \underbrace{\hat{c}_\lambda\left(\frac{1}{n}, x_i, \frac{1}{n}, T^*(x_i)\right)}_{=D(x_i,T^*(x_i))} + \sum_{i\in\mathcal{I}_{\mathrm{dest}}} \underbrace{\hat{c}_\lambda\left(\frac{1}{n}, x_i, 0, x_i\right)}_{=\bar{D}} + \sum_{j\in\mathcal{I}_{\mathrm{crea}}} \underbrace{\hat{c}_\lambda\left(0, y_j, \frac{1}{n}, y_j\right)}_{=\bar{D}} \\ &= \sum_{i=1}^n \left(\mathbb{1}_{i\in\mathcal{I}_{\mathrm{tran}}} D(x_i, T^*(x_i)) + \mathbb{1}_{i\in\mathcal{I}_{\mathrm{dest}}}\bar{D} + \mathbb{1}_{i\in\mathcal{I}_{\mathrm{crea}}}\bar{D}\right). \end{aligned}$$

Since $|\mathcal{I}_{\mathrm{crea}}| = |\mathcal{I}_{\mathrm{dest}}|$ then we have proved the proposition. □

We can understand the previous proposition as follows. Let us denote the mass that is transported from μ by $\tilde{\mu}$. Now, if some mass is transported from x to y, then all mass is transported from x to y. (One can also write this as $\frac{\mathrm{d}\tilde{\mu}}{\mathrm{d}\mu}(x) \in \{0,1\}$ for all x.) Now since both μ and ν are probability measures we know that whatever mass we destroy from μ we must create the same amount of mass in ν. And since the cardinality of the support of μ and ν are equal, and we know there exists a transport map between $\tilde{\mu}$ and $\tilde{\nu}$, then the remaining mass, i.e. $\mu - \tilde{\mu}$ and $\nu - \tilde{\nu}$ must be pure destruction and creation.

3.5 The One-Dimensional Hellinger–Kantorovich Distance

We consider both the Monge form and the dual form of the Hellinger–Kantorovich distance. We later exploit the observations in this section to design our numerical schemes in Section 4.

3.5.1 One-Dimensional Hellinger–Kantorovich in the Monge Formulation

In one-dimension the cyclical monotonicity of optimal plans simplifies (as in the Wasserstein space) to a simple ordering.

Corollary 3.4 (Corollary 6.4 [8])**.** *Let $\Omega = \mathbb{R}$ and π^* be an optimal transport plan for the Hellinger–Kantorovich distance, i.e. π^* achieves the infimum in $\mathrm{K}_{\mathrm{HK}_\lambda}^{(\mathrm{soft})}(\mu,\nu)$. Then for every pair $(x_1,y_1),(x_2,y_2) \in \mathrm{spt}(\pi^*)$ where $x_1 < x_2$ we have that $y_1 \le y_2$.*

For the rest of this section we assume that $\mu = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ and $\nu = \frac{1}{n}\sum_{j=1}^n \delta_{y_j}$. Let $D(x,y)$, when $\|x-y\| \leq \sqrt{\lambda}\pi$, be the cost of "transporting" mass $\frac{1}{n}$ from x to y (that is an initial mass of $\frac{1}{n}$ is transported from x to a final mass of $\frac{1}{n}$ at y). It follows that

$$D(x,y) = \hat{c}_\lambda\left(\frac{1}{n}, x, \frac{1}{n}, y\right) = \frac{8\lambda}{n}\left(1 - \overline{\cos}\left(\frac{\|x-y\|}{2\sqrt{\lambda}}\right)\right).$$

Similarly, we let $\bar{D}$ be the cost of destroying mass $\frac{1}{n}$ (which is independent of the position). We can write

$$\bar{D} = \frac{4\lambda}{n}.$$

By Proposition 3.3 we have $d_{\mathrm{HK}_\lambda}(\mu,\nu) = \sum_{i=1}^n \left(\mathbb{1}_{i\in\mathcal{I}_{\mathrm{tran}}} D(x_i, T^*(x_i)) + 2\mathbb{1}_{i\in\mathcal{I}_{\mathrm{dest}}}\bar{D}\right)$ where T^* is the optimal map, $\mathcal{I}_{\mathrm{tran}}$ is the set of indexes corresponding to where mass is transported from, and $\mathcal{I}_{\mathrm{dest}}$ is the set of indexes corresponding to where mass is destroyed without transport.

Let us suppose that $\{x_i\}_{i=1}^n \subset \mathbb{R}$ and $\{y_j\}_{j=1}^n \subset \mathbb{R}$ are ordered, that is $x_i < x_{i+1}$ and $y_j < y_{j+1}$ for all $i,j = 1,\ldots,n$. For any x_i the transport map T^* will either move mass from x_i to y_j, for $j \in \{1,\ldots,n\}$, or mass is destroyed without transport. Hence $T^*(x_i) \in \{x_i\} \cup \{y_j\}_{j=1}^n$. From the monotonicity of Corollary 3.4 T^* preserves the ordering, i.e. $T^*(x_i) < T^*(x_j)$ for all $i < j$. We immediately can infer a tree-structure to the set of possible optimal transport maps T^*. Let us assume we have already optimally mapped $\{x_1,\ldots,x_k\}$ with cost

$$C_k = \sum_{i=1}^k \left(\mathbb{1}_{i\in\mathcal{I}_{\mathrm{tran}}} D(x_i, T^*(x_i)) + 2\mathbb{1}_{i\in\mathcal{I}_{\mathrm{dest}}}\bar{D}\right)$$

(note that $d_{\mathrm{HK}_\lambda}(\mu,\nu)^2 = C_n$). Now we want to know the admissible choices for $T^*(x_{k+1})$. The first option is that the mass at x_{k+1} is destroyed which will increase the cost by $2\bar{D}$, i.e.

$$C_{k+1} = C_k + 2\bar{D}.$$

The other option is that mass is transported. Letting $I^* : \mathcal{I}_{\mathrm{tran}} \to \{1,\ldots,n\}$ be the index mapping representing the transport, i.e. $I^*(i) = j$ if $T^*(x_i) = y_j$. Then, using the monotonicity of T^*, mass at x_{k+1} is transported to $y_j \in \{y_{I^*(k)+1},\ldots,y_n\}$. In this case the cost is increased by $D(x_k, y_j)$, i.e.

$$C_{k+1} = C_k + D(x_k, y_j).$$

A visualisation of the tree-structure is given in Figure 1. This tree-structure can be exploited to give an efficient algorithm for computing 1D Hellinger–Kantorovich distances.

3.5.2 One-Dimensional Hellinger–Kantorovich in the Dual Formulation

[To do...]

4 Sliced Hellinger–Kantorovich Numerical Schemes

We propose two algorithms based on the monotonicty of transport maps and the dual form.

4.1 Algorithm 1: Exploiting the Monotonicity of Transport Maps

[To do...]

4.2 Algorithm 2: Exploiting the Monotonicity in the Dual Form

[To do...]

5 Numerical Experiments

[To do...]

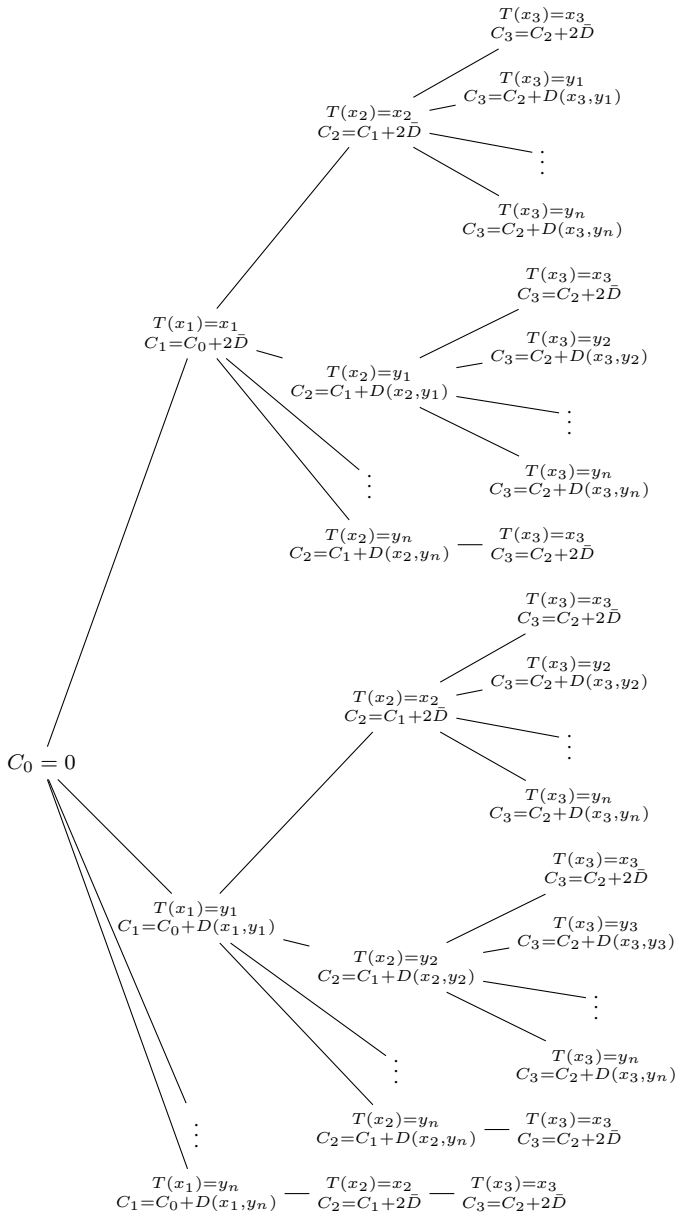

Figure 1: A tree diagram illustrating all possible matchings for the first three points x_1, x_2 and x_3 and how the Hellinger–Kantorovich cost increases.

6 Conclusions

[To do...]

References

[1] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient Flows in Metric Spaces and in the Space of Probability Measures*. Birkhäuser Verlag, 2005.

[2] M. Bauer, M. Bruveris, and P. W. Michor. Uniqueness of the Fisher–Rao metric on the space of smooth densities. *Bulletin of the London Mathematical Society*, 48(3):499–506, 2016.

[3] J.-D. Benamou and Y. Brenier. A computational fluid mechanics solution to the Monge–Kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.

[4] T. Cai, J. Cheng, B. Schmitzer, and M. Thorpe. The linearized Hellinger–Kantorovich distance. *SIAM Journal on Imaging Sciences*, 15(1):45–83, 2022.

[5] L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard. An interpolating distance between optimal transport and Fisher–Rao metrics. *Foundations of Computational Mathematics*, 18(1):1–44, 2018.

[6] L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard. Unbalanced optimal transport: Dynamic and Kantorovich formulations. *Journal of Functional Analysis*, 274(11):3090–3123, 2018.

[7] L. Kantorovich. On the translocation of masses. *Comptes Rendus (Doklady) de I'Académie des Sciences de I'URSS*, 37(7-8):199–201, 1942.

[8] M. Liero, A. Mielke, and G. Savaré. Optimal entropy-transport problems and a new Hellinger–Kantorovich distance between positive measures. *Inventiones mathematicae*, 211(3):969–1117, 2018.

[9] G. Monge. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*, pages 666–704, 1781.

[10] G. Peyré and M. Cuturi. *Computational optimal transport*. Foundations and Trends in Machine Learning. Now Publishers, Inc., 2019.

[11] F. Santambrogio. *Optimal transport for applied mathematicians*. Springer, 2015.

[12] C. Villani. *Topics in Optimal Transportation*. Graduate Studies in Mathematics. American Mathematical Society, 2003.

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

Appendices

A Proof of Proposition 3.1

B Proof of Proposition 3.2