Applications of Optimal Transport

Computational Optimal Transport in Imaging Science
SIAM Conference on Imaging Science

Matthew Thorpe and Yunan Yang

28th May 2024

Disclaimers

1. **Optimal Transport is touching a huge number of applications - impossible to talk about all of them here.**
2. **There is a bias towards applications I've worked on and applications that my collaborators have worked on — but I have tried to be broader than my own interests...**
3. **...which means some of the applications I know very little about!**

Contents

Contents

Computational Methods I

1. Optimal transport is a linear programme, so one can use **linear programming algorithms**, e.g. the simplex method or interior point methods.
2. Angenent, Haker and Tannenbaum's[1] **Flow Minimization** method works directly with the Monge formulation. Starting with an initial transport map (usually the Knoth-Rosenblatt coupling) it is an iterative scheme to update the transport map with one with lower cost (following gradient descent of the Monge cost).
3. 1D optimal transport is very cheap. Meng et al.'s[2] uses iterative 1D projections in their **projection pursuit Monge map** method to find the OT map and cost.

[1] **Angenent, Haker and Tannenbaum**, *Minimizing flows for the Monge–Kantorovich problem*, SIAM journal on mathematical analysis, 35(1):61–97, 2003.

[2] **Meng, Ke, Zhang, Zhang, Zhong and Ma**, *Large-Scale Optimal Transport Map Estimation Using Projection Pursuit*, NeurIPS 32, 2019.

Computational Methods II

4. One of the most popular current approaches, due to Cuturi[3] is to add entropic regularisation to the cost function. The cost can then be reformulated as a Kullback-Leibler divergence which can be solved via the **Sinkhorn Algorithm**.
5. The **back-and-forth method** of Jacobs and Léger[4] alternates between optimising two equivalent dual formulations. If ϕ and ψ are the dual optimal solutions then $\psi = \phi^c$ and $\phi = \psi^c$ leading to a dual problem that can either be written in terms of ϕ or in terms of ψ. Taking alternate H^1 gradients will find the dual optimal pair, from which the transport plans / cost can be recovered.

---

[3] **Cuturi**, *Sinkhorn Distances: Lightspeed Computation of Optimal Transport*, In Advances in Neural Information Processing Systems, pp. 2292–2300, 2013.

[4] **Jacobs and Léger**, *A Fast Approach to Optimal Transport; The Back-And-Forth Method*, Numerische Mathematik, 146(3):513–544, 2020.

Computational Methods III

5. Plenty of others: e.g. **fluid dynamics**[5], **dual gradient descent**[6], **Monge-Ampere solvers**[7] and **semi-discrete approaches**[8].

---

[5] **Benamou and Brenier**, *A Computational Fluid Mechanics Solution to the Monge-Kantorovich Mass Transfer Problem*, Numerische Mathematik, 84(3):375–393, 2000.

[6] **Chartrand, Vixie, Wohlberg and Bollt**, *A Gradient Descent Solution to the Monge-Kantorovich Problem*, Applied Mathematical Sciences, 3(22):1071–1080, 2009.

[7] **Benamou, Froese and Oberman**, *Numerical Solution of the Optimal Transportation Problem Using the Monge-Ampere Equation*, Journal of Computational Physics, 260:107–126, 2014.

[8] **Levy**, *A Numerical Algorithm for L_2 Semi-Discrete Optimal Transport in 3D*, ESAIM Math. Model. Numer. Anal., 49(6):1693–1715, 2015.

Popular Optimal Transport Packages for Python

- **GeomLoss**: Geometric loss Functions between sampled measures, images and volumes. Package authors: Jean Feydy and Pierre Roussillon.
- **OTT-Jax**: optimal transport tools in JAX. Main contributors: Marco Cuturi and Michal Klein.
- **POT**: Python Optimal Transport. Package creators and maintainers: Nicolas Courty, Remi Flamary and Cédric Vincent–Cuaz.
- **PPMP**: Projection pursuit Monge map. Package creators: Cheng Meng, Jingyi Zhang, Mengrui Zhang and Tao Li (built on POT).
- **pysdot**: Python semi-discrete optimal transportation tools. Package contributors: Hugo Leclerc and Quentin Mérigot.

Contents

The Wasserstein Distance

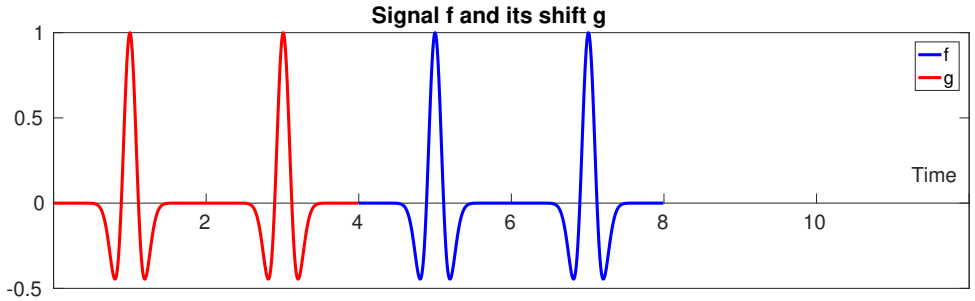
- The Wasserstein distance on the space $\mathcal{P}_p(X)$ is defined by

$$\mathrm{d_{W_p}}(\mu,\nu) = \min_{\pi\in\Pi(\mu,\nu)} \left(\int_{X\times X} |x-y|^p \,\mathrm{d}\pi(x,y)\right)^{\frac{1}{p}}.$$

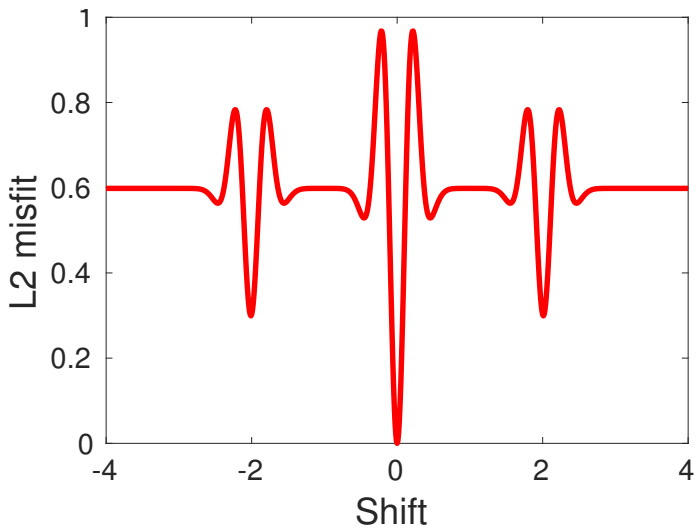
- The Wasserstein distance metrises the weak* convergence (for compact X).
- For $p=1$ we can also represent the Wasserstein distance in the Kantorovich–Rubinstein duality form:

$$\mathrm{d_{W_1}}(\mu,\nu) = \sup\left\{\int_X f \,\mathrm{d}(\mu-\nu) \;:\; f \text{ is 1-Lipschitz}\right\}.$$

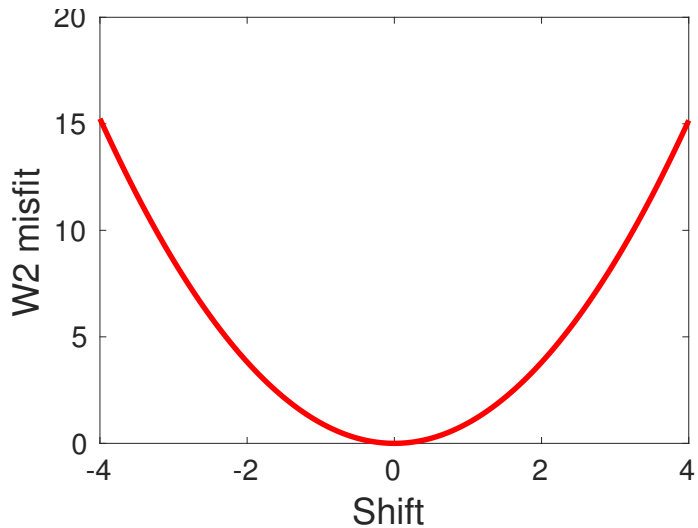
Why the Wasserstein Distance?



(a) Example Signals.



(b) L^2 distance between signals.



(c) W_2 distance between signals.

Source: **Engquist and Yang**, *Seismic Imaging and Optimal Transport*, Communications in Information and Systems, 19(2):95–145, 2019.

Example 1: Seismic Imaging

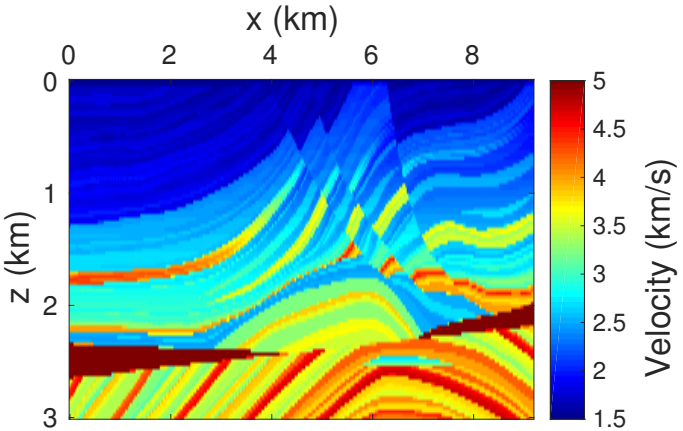
- A simple model for wave propagation is:

$$\begin{cases} \dfrac{1}{v(x)^2}\dfrac{\partial^2 u}{\partial t^2}(x,t) - \Delta u(x,t) &= s(x,t) \\ u(x,0) &= 0 \\ \dfrac{\partial u}{\partial t}(x,0) &= 0 \end{cases}$$

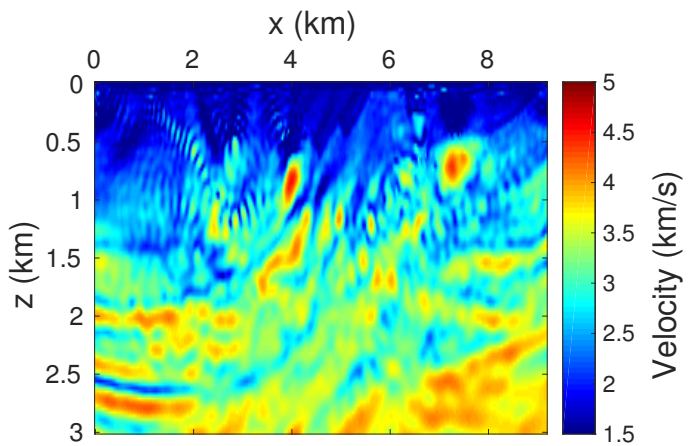
  where u is the wavefield, s is the source and $v(x)$ describes the ground material medium (i.e., the wave speed).
- s is known and u can be observed at locations $\{x_r\}_{r=1}^R$, the **inverse problem** is to find $v(x)$.
- The **forward model** $v \mapsto u(\cdot\,;\, v)$ is highly nonlinear.
- Let $\{g(x_r,\cdot)\}_{r=1}^R$ be observations of the wavefield. One can find v by minimising

$$\mathcal{J}(v) = \sum_{r=1}^{R} d(g(x_r,\cdot)\,,\, u(x_r,\cdot\,;\, v))^2.$$

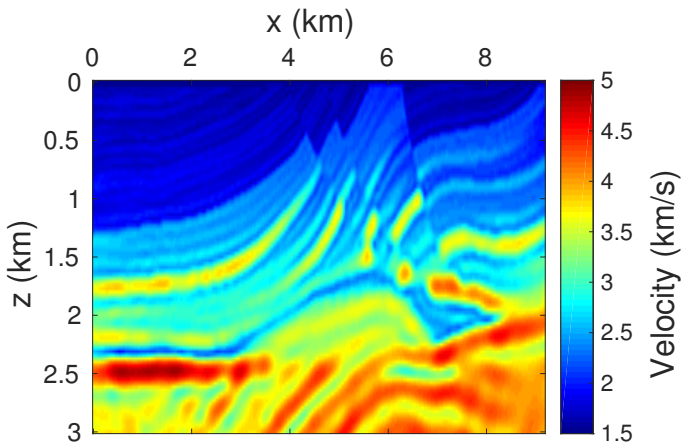
Seismic Imaging with the Wasserstein Distance



(a) True velocity



(b) L^2 cost:
$d(f,g)^2 = \int |f(t) - g(t)|^2 \, \mathrm{d}t$



(c) W_2 cost: $d(f,g)^2 = \inf_{\pi \in \Pi(\tilde{f},\tilde{g})} \int |p-q|^2 \, \mathrm{d}\pi(p,q)$

Source: **Engquist and Yang**, *Seismic Imaging and Optimal Transport*, Communications in Information and Systems, 19(2):95–145, 2019.

Example 2: Generative Adversarial Networks

- GANs were introduced in 2014 by Goodfellow et al.[9] for generative machine learning.
- Core idea: two neural networks compete against each other.
- Generator: creates new images $\mu_G \in \mathcal{P}(\Omega)$.
- Discriminator: tries to determine whether an image is generated or not $D : \Omega \to [0, 1]$.
- Objective functional:

$$\mathcal{L}_{\mathrm{JS}}(\mu_G, D) = \mathbb{E}_{x\sim\mu_{\mathrm{true}}}[\ln D(x)] + \mathbb{E}_{x\sim\mu_G}[\ln(1 - D(x))].$$

- Minimise $\mathcal{L}_{\mathrm{JS}}$ with respect to μ_G and maximise with respect to D.

[9] **Goodfellow, Pouget-Abadie, Mirza, Xu, Warde-Farley, Ozair, Courville and Bengio**, *Generative Adversarial Networks*, arxiv:1406.2661, 2014.

Mode Collapse

- Mode collapse happens when the generator learns too quickly.
- The generator essentially creates images in the set $\operatorname{argmax}_x D(x)$.
- If the discriminator prefers a certain region of Ω then the generator will only sample from this region.
- For example, in MNIST if the discriminator prefers the digit 0 then the generator may only generate 0's.
- Even as training continues the generator can remain stuck in a local minimum and fail to generalise.

Wasserstein Generative Adversarial Networks

- Instead of $\mathcal{L}_{\mathrm{JS}}$ Arjovsky et al.[10] proposed the following objective:

$$\mathcal{L}_{\mathrm{W}}(\mu_G, D) = \mathbb{E}_{x\sim\mu_G}[D(x)] - \mathbb{E}_{x\sim\mu_{\mathrm{true}}}[D(x)].$$

- Maximising over D that are 1-Lipschitz produces the Wasserstein distance:

$$\sup_{\|D\|_{\mathrm{Lip}}\leq 1} \mathcal{L}_{\mathrm{W}}(\mu_G, D) = \mathrm{d}_{\mathrm{W}_1}(\mu_G, \mu_{\mathrm{true}}).$$

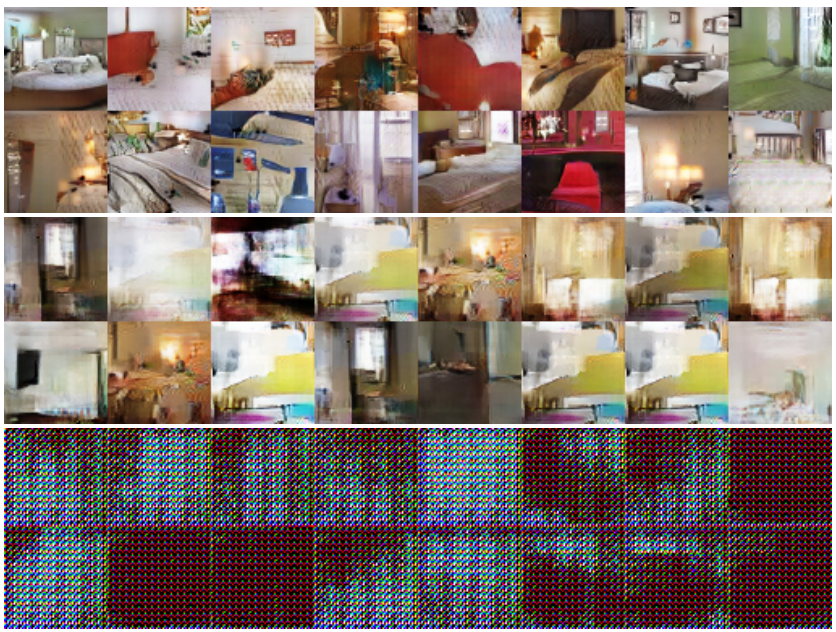
- WGANs are empirically observed as more robust with respect to mode collapse.
- However, WGANs are a poor approximation of the Wasserstein distance, and improving the approximation often leads to worse performance.[11]

[10] **Arjovsky, Chintala and Bottou**, *Wasserstein Generative Adversarial Networks*, International Conference on Machine Learning, 2017.

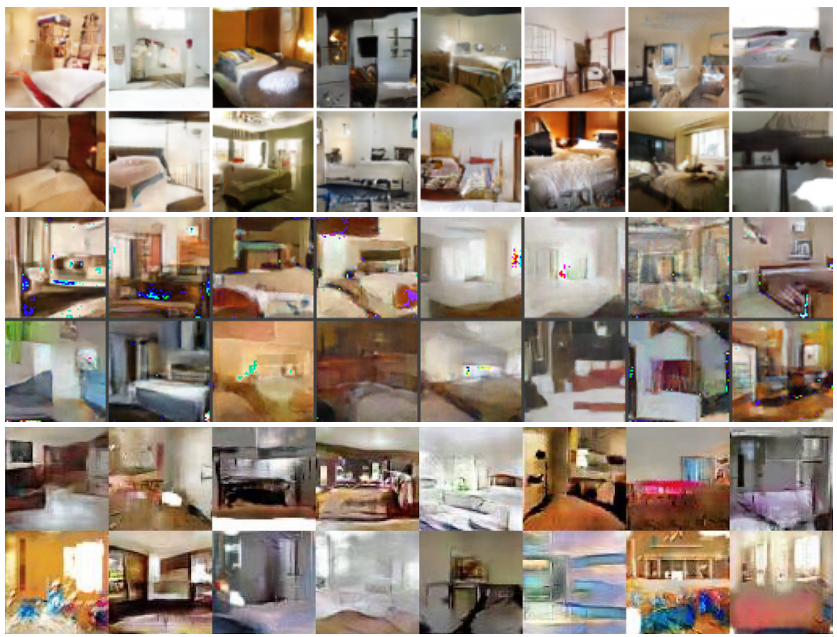
[11] **Stanczuk, Etmann, Kreusser and Schönlieb**, *Wasserstein GANs Work Because They Fail (To Approximate the Wasserstein distance)*, arxiv:2103.01678, 2021.

GANs vs WGANs

Generative Adversarial Network



Wasserstein Generative Adversarial Network



Source: **Arjovsky, Chintala and Bottou**, *Wasserstein Generative Adversarial Networks*, International Conference on Machine Learning, 2017.

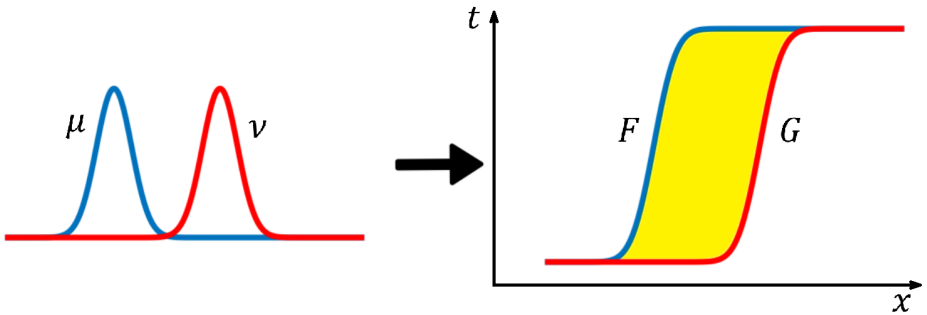
Contents

1D Optimal Transport

- When we have the cumulative distribution functions:

$$\mathrm{d_{W_p}}(\mu,\nu)^p = \int_0^1 |F^{-1}(t) - G^{-1}(t)|^p \,\mathrm{d}t$$

where F, G are the cdf's of μ and ν.



- For uniform empirical distributions computing the distance is equivalent to a sorting problem.

$$\mathrm{d_{W_p}}(\mu,\nu)^p = \frac{1}{n}\sum_{m=1}^{n} |x_{i[m]} - y_{j[m]}|^p$$

where $\mu = \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}$, $\nu = \frac{1}{n}\sum_{j=1}^{n}\delta_{y_j}$ and $i[m]$, $j[m]$ are the sorted indices.

Source: **Kolouri and Rohde**, *Optimal transport a crash course*, IEEE ICIP 2016 Tutorial Slides: Part 1, 2016.

The Sliced Wasserstein Distance

- **Sliced Wasserstein Distance:**
  1. Let $\sigma \in \mathcal{P}(\mathbb{S}^{d-1})$ be a positive measure on the hypersphere $\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d : |x| = 1\}$.
  2. $\langle \theta, \cdot \rangle$ is the projection onto the 1D line spanned by $\theta \in \mathbb{S}^{d-1}$.
  3. $\mathrm{d}_{\mathrm{SW}_p}(\mu, \nu) = \int_{\mathbb{S}^{d-1}} \mathrm{d}_{\mathrm{W}_p}(\langle \theta, \cdot \rangle_{\#}\mu, \langle \theta, \cdot \rangle_{\#}\nu)\, \mathrm{d}\sigma(\theta)$.
- $\mathrm{d}_{\mathrm{SW}_p} : \mathcal{P}_p(\mathbb{R}^d) \times \mathcal{P}_p(\mathbb{R}^d) \to [0, \infty)$ is a metric.
- Approximate computational costs:
  1. Linear programming $O(n^3 \log(n))$,
  2. Auction algorithm $O(n^2 \log(n))$,
  3. Entropy regularised Sinkhorn algorithm $O(n^2)$,
  4. 1D $O(n \log(n))$.

Autoencoders

- An autoencoder aims to generate data by learning a transformation from a samplable (latent) space.
- Let $\mathcal{Z}$ be the latent space and $\mathcal{X}$ be the data space.
- Let $\phi : \mathcal{X} \to \mathcal{Z}$ be the encoder and $\psi : \mathcal{Z} \to \mathcal{X}$ be the decoder.
- Let $\mathbb{P}_{\mathcal{X}}$ be the data distribution on $\mathcal{X}$ and $\mathbb{Q}_{\mathcal{Z}}$ be a given samplable distribution on $\mathcal{Z}$.
- **Properties of a good encoder:**
  1. $\mathbb{Q}_{\mathcal{Z}}$ is easy to sample from;
  2. $\phi_{\#}\mathbb{P}_{\mathcal{X}} \approx \mathbb{Q}_{\mathcal{Z}}$; and
  3. $\psi_{\#}\phi_{\#}\mathbb{P}_{\mathcal{X}} \approx \mathbb{P}_{\mathcal{X}}$.

Example 3: Sliced Wasserstein Autoencoders

- We need a measure for $\phi_{\#}\mathbb{P}_{\mathcal{X}} \approx \mathbb{Q}_{\mathcal{Z}}$.
  1. f-divergences, e.g. Nowozin et al. 2016.[12]
  2. Wasserstein distances, e.g. Tolstikhin et al. 2018.[13]
  3. Sliced Wasserstein distances, e.g. Kolouri et al. 2018.[14]
- Sliced Wasserstein autoencoder:

$$\underset{\phi,\psi}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} |x_i - \psi(\phi(x_i))|^2 + \lambda \mathrm{d}_{\mathrm{SW}_2}(\phi_{\#}\mathbb{P}_{\mathcal{X}}, \mathbb{Q}_{\mathcal{Z}})^2.$$

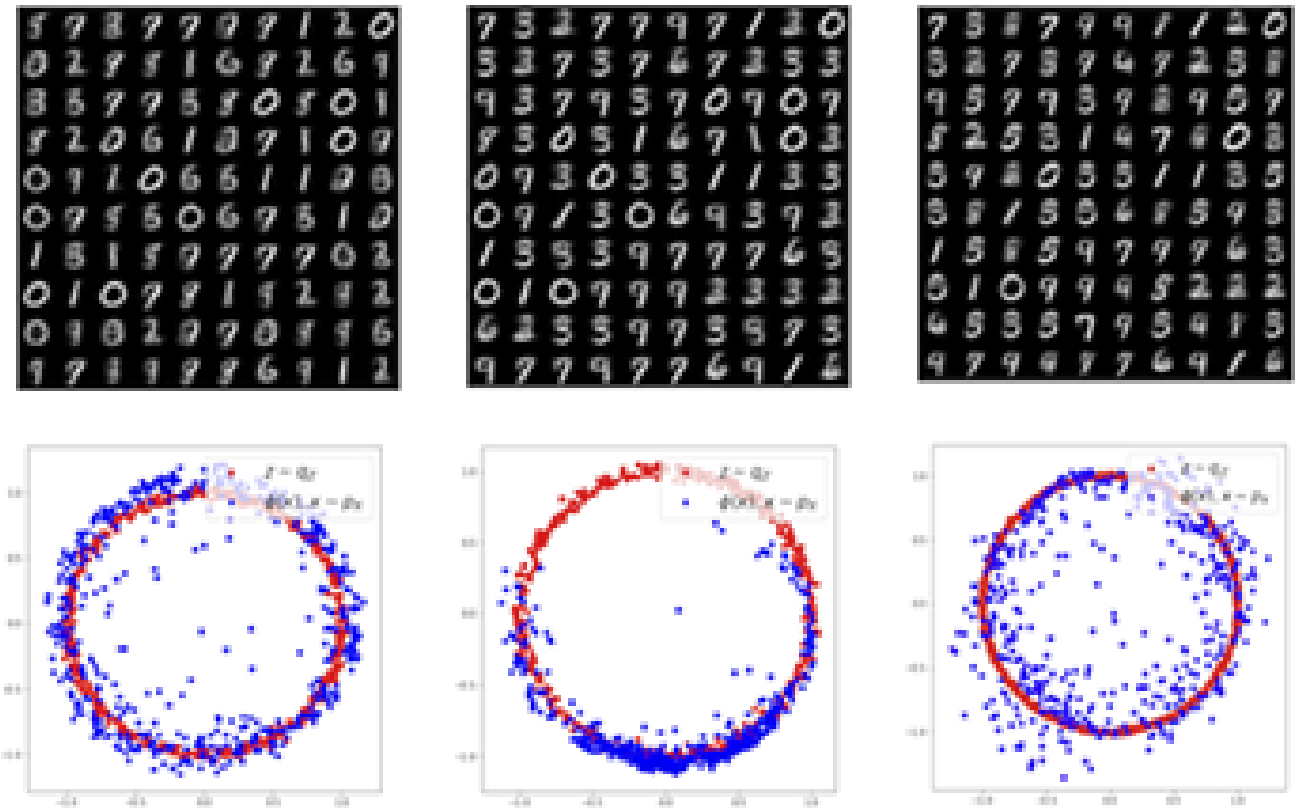
---

[12] **Nowozin, Cseke and Tomioka** *f-GAN: Training generative neural samplers using variational divergence minimization*, NeuRIPS, 2016.

[13] **Tolstikhin, Bousquet, Gelly and Schölkopf**, *Wasserstein Auto-Encoders*, International Conference on Learning Representations, 2018.

[14] **Kolouri, Pope, Martin and Rohde**, *Sliced Wasserstein auto-encoders*, ICLR, 2018.

Representations of MNIST



Left: sliced Wasserstein autoencoder; **centre:** Wasserstein generative adversarial Network; **right:** maximum mean discrepancy autoencoder. **Top:** generated images; **bottom:** latent space samples.

Source: **Kolouri, Pope, Martin and Rohde**, *Sliced Wasserstein auto-encoders*, ICLR, 2018.

Contents

Motivation

The Wasserstein distance is *great* as a distance between signals/images, because...

1. Lagrangian modelling,
2. simple to understand compared to other Lagrangian methods such as large deformation diffeomorphic metric mapping,
3. metric properties (in particular symmetry, triangle inequality,...).
4. geodesics and Riemannian structure,
5. theoretical and characterising properties such as existence of optimal transport maps and optimal transport plans.

But,...

1. it places restrictive conditions on the input, in particular signals have to be probability measures,
2. computationally expensive (despite recent advances),
3. there is a lack of off-the-shelf data analysis tools.

This motivates the **linear Wasserstein distance**.

The Linear Wasserstein Distance

- We recall that the Wasserstein distance can be written

$$\mathrm{d}_{\mathrm{W}_2}(\mu,\nu) = \sqrt{\int |v_0(x)|^2 \,\mathrm{d}\mu(x)}$$

  where $v_0 = v(0,\cdot)$ is the Benamou–Brenier optimal velocity.
- We also have $v_0(x) = T(x) - x$ where T is the optimal Monge map.
- We define the logarithmic map:

$$\mathrm{Log}_{\mathrm{W}_2}(\mu;\mu_i) := v_0.$$

- Wasserstein distance from reference:

$$\mathrm{d}_{\mathrm{W}_2}(\mu,\nu) = \|\mathrm{Log}_{\mathrm{W}_2}(\mu;\nu)\|_{\mathrm{L}^2(\mu)}.$$

- Linear Wasserstein distance:[15]

$$\mathrm{d}_{\mathrm{W}_2,\mathrm{lin},\mu}(\mu_1,\mu_2) = \|\mathrm{Log}_{\mathrm{W}_2}(\mu;\mu_1) - \mathrm{Log}_{\mathrm{W}_2}(\mu;\mu_2)\|_{\mathrm{L}^2(\mu)}.$$

[15] **Wang, Slepčev, Basu, Ozolek and 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.

The Linear Wasserstein Distance Approximation

- We always have

$$\mathrm{d}_{\mathrm{W}_2}(\mu_1, \mu_2) \leq \mathrm{d}_{\mathrm{W}_2,\mathrm{lin},\mu}(\mu_1, \mu_2).$$

- If $\mu_i = S_{i\#}\mu$ where S_i is a shearing and/or translation:[16]

$$\mathrm{d}_{\mathrm{W}_2}(\mu_1, \mu_2) = \mathrm{d}_{\mathrm{W}_2,\mathrm{lin},\mu}(\mu_1, \mu_2).$$

- If μ_i live on a nice submanifold:[17]

$$\mathrm{d}_{\mathrm{W}_2}(\mu_1, \mu_2) = \mathrm{d}_{\mathrm{W}_2,\mathrm{lin},\mu}(\mu_1, \mu_2) + O(\mathrm{d}_{\mathrm{W}_2}(\mu_1, \mu_2)^{\frac{3}{2}}).$$

- In general:[18]

$$\mathrm{d}_{\mathrm{W}_2,\mathrm{lin},\mu}(\mu_1, \mu_2) \leq C\mathrm{d}_{\mathrm{W}_2}(\mu_1, \mu_2)^{\frac{1}{2}}.$$

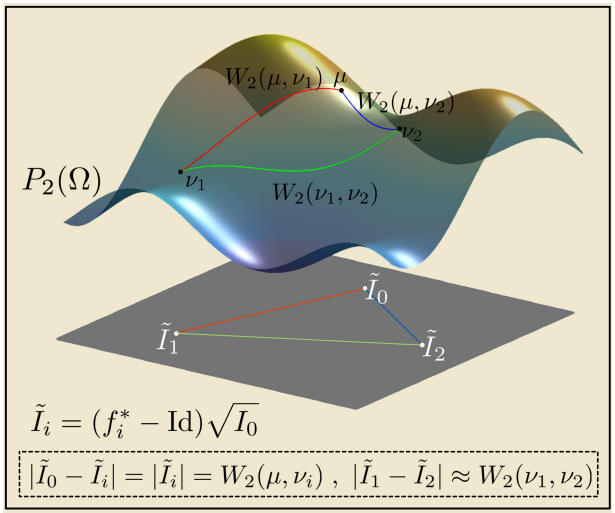


Image Source: **Kolouri, Park, T., Slepčev and Rohde**, *Optimal Mass Transport: Signal Processing and Machine Learning Applications*, IEEE Signal Processing Magazine, 34(4):43–59, 2017.

---

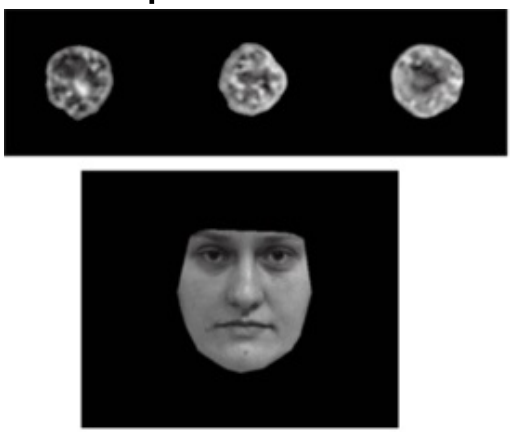
[16] **Khurana, Kannan, Cloninger and Moosmüller**, *Supervised Learning of Sheared Distributions Using Linearized Optimal Transport*, Sampling Theory, Signal Processing, and Data Analysis, 21(1), 2023.

[17] **Hamm, Moosmueller, Schmitzer and T.**, *Manifold Learning in Wasserstein Space*, arxiv:2311.08549, 2023.

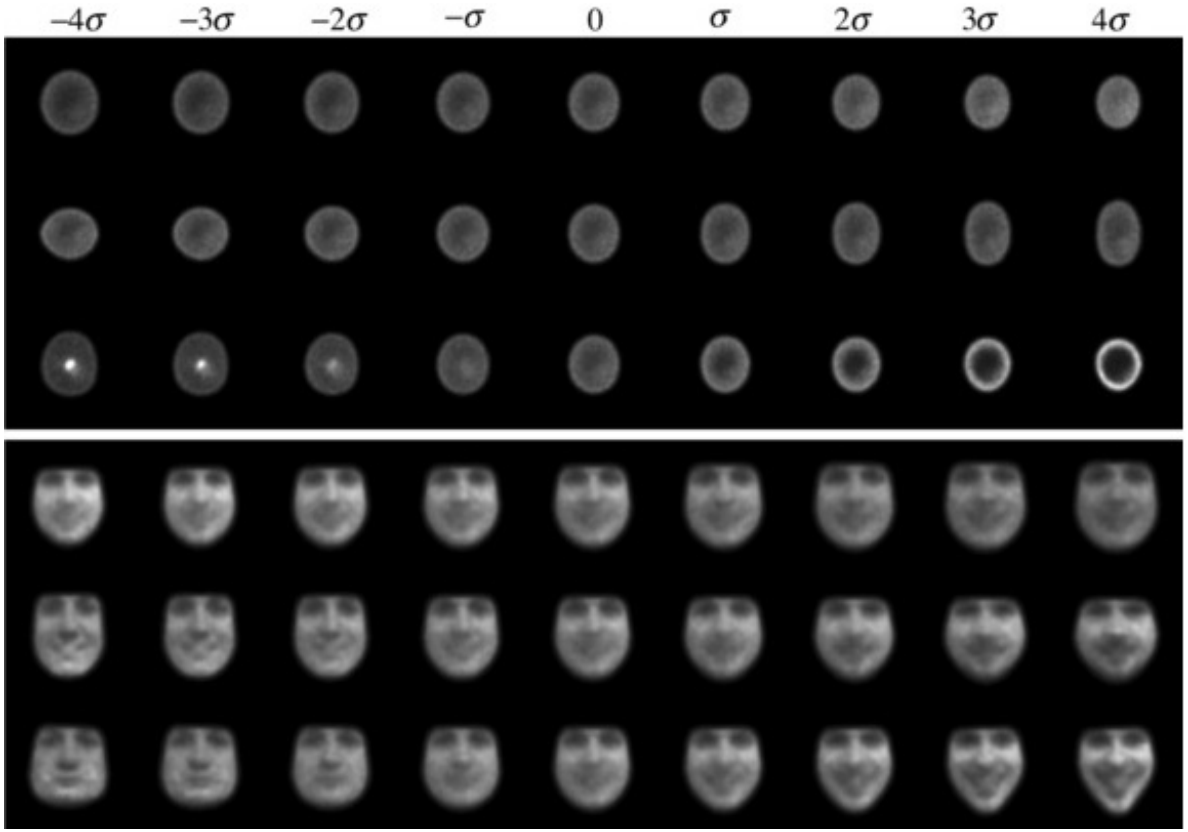
[18] **Gigli**, *On Hölder Continuity-in-Time of the Optimal Transport Map Towards Measures Along a Curve*, Proceedings of the Edinburgh Mathematical Society, 54(2):401–409, 2011.

Example 4: Transport Based Morphometry

Example Data:



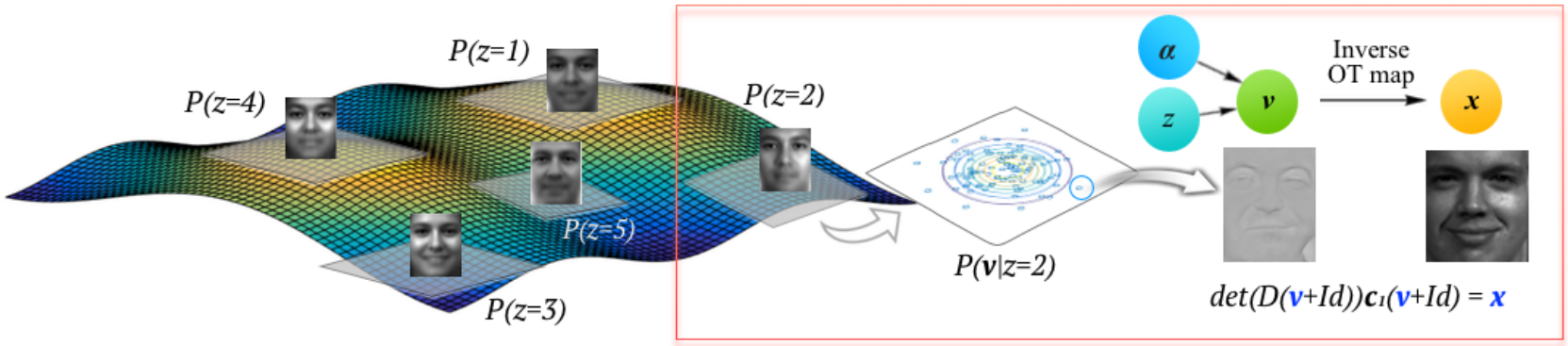
Principle Component Analysis on Linear Embedding:



Source: **Wang, Slepčev, Basu, Ozolek and 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.

Example 5: Data Generating

Pipeline:



Results:

1. Top row, all 19 original images.
2. Second and third rows, generated images.

Source: **Park and T.**, *Representing and Learning High Dimensional Data with the Optimal Transport Map from a Probabilistic Viewpoint*, CVPR, 2018.

Contents

The Hellinger–Kantorovich Distance

1. Recall the continuity equation:
$$(\rho, v) \in \mathcal{CE}(\mu, \nu) \Leftrightarrow \frac{\partial \rho}{\partial t} + \nabla_x \cdot (\rho v) = 0, \rho_0 = \mu, \rho_1 = \nu.$$
2. And the Wasserstein distance
$$\mathrm{d}^2_{\mathrm{W}_2}(\mu, \nu) = \inf_{(\rho,v)\in\mathcal{CE}(\mu,\nu)} \int_0^1 \int_\Omega \|v(t,\cdot)\|^2 \,\mathrm{d}\rho_t \,\mathrm{d}t.$$
3. To allow for mass creation we include a source/sink term:
$$(\rho, v, \zeta) \in \mathcal{CES}(\mu, \nu) \Leftrightarrow \frac{\partial \rho}{\partial t} + \nabla_x \cdot (\rho v) = \zeta, \rho_0 = \mu, \rho_1 = \nu.$$
4. The Hellinger–Kantorovich distance (a.k.a. unbalanced optimal transport and Wasserstein–Fisher–Rao):[19][20][21][22]
$$\mathrm{d}^2_{\mathrm{HK}}(\mu, \nu) := \inf_{(\rho,v,\zeta)\in\mathcal{CES}(\mu,\nu)} \int_0^1 \int_\Omega \left(\|v(t,\cdot)\|^2 + \frac{1}{4}\left(\frac{\mathrm{d}\zeta_t}{\mathrm{d}\rho_t}\right)^2 \right) \mathrm{d}\rho_t \,\mathrm{d}t.$$

[19] **Kondratyev, Monsaingeon and Vorotnikov**, *A New Optimal Transport Distance On The Space Of Finite Radon Measures*, Advances in Differential Equations, 21:1117–1164, 2016.

[20] **Chizat, Peyré, Schmitzer and Vialard**, *An Interpolating Distance Between Optimal Transport and Fisher–Rao Metrics*, Foundations of Computational Mathematics 18:1–44, 2018.

[21] **Chizat, Peyr é, Schmitzer and Vialard**, *Unbalanced Optimal Transport: Dynamic and Kantorovich formulations*, Journal of Functional Analysis, 274:3090–3123, 2018.

[22] **Liero, Mielke and Savaré**, *Optimal Entropy-Transport Problems and a New Hellinger–Kantorovich Distance Between Positive Measures*, Inventiones Mathematicae 211:969–1117, 2018.

The Hellinger–Kantorovich Distance Between Diracs I

1. Let $\mu = m_0 \delta_{x_0}$ and $\nu = m_1 \delta_{x_1}$.
2. If $\|x_0 - x_1\| \leq \frac{\pi}{2}$ then we choose $\rho(t,x) = m(t)\delta_{x(t)}$ for some functions $m : [0,1] \to [0,+\infty)$ and $x : [0,1] \to \Omega$.
3. Further assume that trajectories are straight lines, i.e. there exists $\varphi : [0,1] \to [0,1]$ such that $x(t) = x_0 + (x_1 - x_0)\varphi(t)$.
4. The HK energy is

$$\mathrm{d}^2_{\mathrm{HK}}(\mu,\nu) = \inf_{m,x} \underbrace{\int_0^1 \left(|\dot{x}(t)|^2 + \frac{1}{4} \left(\frac{\dot{m}(t)}{m(t)} \right)^2 \right) m(t)\, \mathrm{d}t}_{=:\mathcal{HK}(m,x)}.$$

The Hellinger–Kantorovich Distance Between Diracs II

5. Taking variations in $\mathcal{HK}$ implies the minimiser satisfies

$$0 = \frac{\mathrm{d}}{\mathrm{d}t}\left(\dot{x}m\right)(t)$$
$$0 = |\dot{x}(t)|^2 - \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\dot{m}}{m}\right)(t) - \frac{1}{4}\frac{\dot{m}(t)^2}{m(t)^2}$$

with $x(0) = x_0$, $x(1) = x_1$, $m(0) = m_0$ and $m(1) = m_1$.

6. The Euler-Lagrange equations are satisfied by

$$m(t) = (1-t)^2 m_0 + t^2 m_1 + 2t(1-t)\sqrt{m_0 m_1}\cos\|x_0 - x_1\|$$
$$\varphi(t) = \frac{1}{\|x_0 - x_1\|}\cos^{-1}\left(\frac{(1-t)\sqrt{m_0} + t\sqrt{m_1}\cos(\|x_0 - x_1\|)}{\sqrt{m(t)}}\right)$$
$$x(t) = x_0 + (x_1 - x_0)\varphi(t).$$

The Hellinger–Kantorovich Distance Between Diracs III

7. The Hellinger–Kantorovich distance between Diracs closer than $\pi/2$ is therefore:

$$\mathrm{d}^2_{\mathrm{HK}}(\mu,\nu) = m_0 + m_1 - 2\sqrt{m_0 m_1}\cos\|x_0 - x_1\|.$$

8. Now if $\|x_0 - x_1\| > \frac{\pi}{2}$ we consider the 'pure creation' and 'pure destruction' interpolation: $\rho(t,x) = m_0(t)\delta_{x_0} + m_1(t)\delta_{x_1}$ where $m_0(0) = m_0$, $m_0(1) = 0$, $m_1(0) = 0$ and $m_1(1) = m_1$.
9. The HK energy is

$$\mathrm{d}^2_{\mathrm{HK}}(\mu,\nu) = \inf_{m_0,m_1} \frac{1}{4}\int_0^1 \left(\frac{\dot{m}_0(t)}{m_0(t)}\right)^2 m_0(t) + \left(\frac{\dot{m}_1(t)}{m_1(t)}\right)^2 m_1(t)\,\mathrm{d}t.$$

10. Optimising implies $m_0(t) = (1-t)^2 m_0$ and $m_1(t) = t^2 m_1$ and

$$\mathrm{d}^2_{\mathrm{HK}}(\mu,\nu) = m_0 + m_1.$$

The Hard Kantorovich Form of the HK Distance

1. Define

$$\hat{c}(x_0, m_0, x_1, m_1) := \mathrm{d}_{\mathrm{HK}}(m_0\delta_{x_0}, m_1\delta_{x_1}) \\ = m_0 + m_1 - 2\sqrt{m_0 m_1}\,\overline{\cos}(\|x_0 - x_1\|)$$

   where $\overline{\cos}(t) = \cos(t \wedge \frac{\pi}{2})$.
2. Let $\pi_0 \in \mathcal{M}_+(\Omega^2)$ be the mass that leaves, i.e. the amount of mass leaving x_0 for x_1 is $\pi_0(x_0, x_1)$.
3. Let $\pi_1 \in \mathcal{M}_+(\Omega^2)$ be the mass that arrives, i.e. the amount of mass arriving at x_1 from x_0 is $\pi_1(x_0, x_1)$.
4. The amount of mass leaving x_0 is given by μ and the amount of mass arriving at x_1 is given by ν, hence $P_{0\#}\pi_0 = \mu$ and $P_{1\#}\pi_1 = \nu$.
5. Hard-marginal Kantorovich form:

$$\mathrm{d}^2_{\mathrm{HK}}(\mu, \nu) = \inf_{\substack{P_{0\#}\pi_0 = \mu \\ P_{1\#}\pi_1 = \nu \\ \pi_0, \pi_1 \ll \lambda}} \int_{\Omega^2} \hat{c}\left(x_0, \frac{\mathrm{d}\pi_0}{\mathrm{d}\lambda}(x_0, x_1), x_1, \frac{\mathrm{d}\pi_1}{\mathrm{d}\lambda}(x_0, x_1)\right) \mathrm{d}\lambda(x_0, x_1).$$

The Soft Kantorovich Form of the HK Distance I

Let $C(x_0, x_1) = \overline{\cos}(\|x_0 - x_1\|)$. We can write

$$
\begin{aligned}
\mathrm{d}^2_{\mathrm{HK}}(\mu,\nu) &= \int_{\Omega^2} \frac{\mathrm{d}\pi_0}{\mathrm{d}\lambda} + \frac{\mathrm{d}\pi_1}{\mathrm{d}\lambda} - 2\sqrt{\frac{\mathrm{d}\pi_0}{\mathrm{d}\lambda}\frac{\mathrm{d}\pi_1}{\mathrm{d}\lambda}}\, C \,\mathrm{d}\lambda \\
&= \int_{\Omega^2} \frac{\mathrm{d}\pi_0}{\mathrm{d}\lambda} + \frac{\mathrm{d}\pi_1}{\mathrm{d}\lambda} \\
&\quad - 2\sqrt{\frac{\mathrm{d}\pi_0}{\mathrm{d}\lambda}\frac{\mathrm{d}\pi_1}{\mathrm{d}\lambda}}\, C \Bigg[1 + \underbrace{\log C - \frac{1}{2}\log\left(\sqrt{\frac{\frac{\mathrm{d}\pi_1}{\mathrm{d}\lambda}}{\frac{\mathrm{d}\pi_0}{\mathrm{d}\lambda}}}\, C\right) - \frac{1}{2}\log\left(\sqrt{\frac{\frac{\mathrm{d}\pi_0}{\mathrm{d}\lambda}}{\frac{\mathrm{d}\pi_1}{\mathrm{d}\lambda}}}\, C\right)}_{=0} \Bigg] \mathrm{d}\lambda \\
&= \int_{\Omega^2} -2\sqrt{\frac{\mathrm{d}\pi_0}{\mathrm{d}\lambda}\frac{\mathrm{d}\pi_1}{\mathrm{d}\lambda}}\, C \log C \,\mathrm{d}\lambda \\
&\quad + \int_{\Omega^2} \frac{\mathrm{d}\pi_0}{\mathrm{d}\lambda}\left(1 - \sqrt{\frac{\frac{\mathrm{d}\pi_1}{\mathrm{d}\lambda}}{\frac{\mathrm{d}\pi_0}{\mathrm{d}\lambda}}}\, C + \sqrt{\frac{\frac{\mathrm{d}\pi_1}{\mathrm{d}\lambda}}{\frac{\mathrm{d}\pi_0}{\mathrm{d}\lambda}}}\, C \log\left(\sqrt{\frac{\frac{\mathrm{d}\pi_1}{\mathrm{d}\lambda}}{\frac{\mathrm{d}\pi_0}{\mathrm{d}\lambda}}}\, C\right)\right) \mathrm{d}\lambda \\
&\quad + \int_{\Omega^2} \frac{\mathrm{d}\pi_1}{\mathrm{d}\lambda}\left(1 - \sqrt{\frac{\frac{\mathrm{d}\pi_0}{\mathrm{d}\lambda}}{\frac{\mathrm{d}\pi_1}{\mathrm{d}\lambda}}}\, C + \sqrt{\frac{\frac{\mathrm{d}\pi_0}{\mathrm{d}\lambda}}{\frac{\mathrm{d}\pi_1}{\mathrm{d}\lambda}}}\, C \log\left(\sqrt{\frac{\frac{\mathrm{d}\pi_0}{\mathrm{d}\lambda}}{\frac{\mathrm{d}\pi_1}{\mathrm{d}\lambda}}}\, C\right)\right) \mathrm{d}\lambda \\
&= \int_{\Omega^2} c \,\mathrm{d}\pi + \int_{\Omega^2} \varphi\left(\sqrt{\frac{\frac{\mathrm{d}\pi_1}{\mathrm{d}\lambda}}{\frac{\mathrm{d}\pi_0}{\mathrm{d}\lambda}}}\, C\right) \mathrm{d}\pi_0 + \int_{\Omega^2} \varphi\left(\sqrt{\frac{\frac{\mathrm{d}\pi_0}{\mathrm{d}\lambda}}{\frac{\mathrm{d}\pi_1}{\mathrm{d}\lambda}}}\, C\right) \mathrm{d}\pi_1
\end{aligned}
$$

where $c(x_0, x_1) = -2\log C(x_0, x_1)$, $\pi = \sqrt{\frac{\mathrm{d}\pi_0}{\mathrm{d}\lambda}\frac{\mathrm{d}\pi_1}{\mathrm{d}\lambda}}\, C\lambda$ and $\varphi(s) = s\log s - s + 1$.

The Soft Kantorovich Form of the HK Distance II

1. Let
$$\mu = u\bar{\mu} + \bar{\mu}^{\perp} \qquad u = \frac{\mathrm{d}\mu}{\mathrm{d}\bar{\mu}}$$
$$\nu = w\bar{\nu} + \bar{\nu}^{\perp} \qquad w = \frac{\mathrm{d}\nu}{\mathrm{d}\bar{\nu}}.$$

2. Then
$$\pi_0 = (u \otimes 1)\pi + (\mathrm{Id} \times \mathrm{Id})_{\#}\bar{\mu}$$
$$\pi_1 = (1 \otimes w)\pi + (\mathrm{Id} \times \mathrm{Id})_{\#}\bar{\nu}.$$

3. We have the densities
$$\frac{\mathrm{d}\pi}{\mathrm{d}\pi_0} = \frac{1}{u \otimes 1} \qquad \frac{\mathrm{d}\pi}{\mathrm{d}\pi_0} = \frac{\frac{\mathrm{d}\pi}{\mathrm{d}\lambda}}{\frac{\mathrm{d}\pi_0}{\mathrm{d}\lambda}} = \frac{\sqrt{\frac{\mathrm{d}\pi_0}{\mathrm{d}\lambda}\frac{\mathrm{d}\pi_1}{\mathrm{d}\lambda}}C}{\frac{\mathrm{d}\pi_0}{\mathrm{d}\lambda}} = \sqrt{\frac{\frac{\mathrm{d}\pi_1}{\mathrm{d}\lambda}}{\frac{\mathrm{d}\pi_0}{\mathrm{d}\lambda}}}C.$$

4. Note also $\frac{\mathrm{d}P_{0\#}\pi}{\mathrm{d}\mu} = \frac{1}{u}$.

The Soft Kantorovich Form of the HK Distance III

5. So

$$\int_{\Omega^2} \varphi\left(\sqrt{\frac{\frac{\mathrm{d}\pi_1}{\mathrm{d}\lambda}}{\frac{\mathrm{d}\pi_0}{\mathrm{d}\lambda}}}\, c\right) \mathrm{d}\pi_0 = \int_{\Omega} \varphi\left(\frac{1}{u(x_0)}\right) \mathrm{d}\pi_0(x_0, x_1)$$
$$= \int_{\Omega} \varphi\left(\frac{1}{u(x_0)}\right) \mathrm{d}\mu(x_0)$$
$$= \int_{\Omega} \varphi\left(\frac{\mathrm{d}P_{0\#}\pi}{\mathrm{d}\mu}\right) \mathrm{d}\mu =: \mathrm{KL}(P_{0\#}\pi|\mu).$$

6. Similarly,

$$\int_{\Omega^2} \varphi\left(\sqrt{\frac{\frac{\mathrm{d}\pi_0}{\mathrm{d}\lambda}}{\frac{\mathrm{d}\pi_1}{\mathrm{d}\lambda}}}\, c\right) \mathrm{d}\pi_1 = \mathrm{KL}(P_{1\#}\pi|\nu).$$

7. So far, for $c(x_0, x_1) = -2 \log \overline{\cos}(\|x_0 - x_1\|)$

$$\mathrm{d}^2_{\mathrm{HK}}(\mu, \nu) = \int_{\Omega^2} c \,\mathrm{d}\pi + \mathrm{KL}(P_{0\#}\pi|\mu) + \mathrm{KL}(P_{1\#}\pi|\nu).$$

8. In fact,

$$\mathrm{d}^2_{\mathrm{HK}}(\mu, \nu) = \inf_{\pi \in \mathcal{M}_+(\Omega^2)} \left(\int_{\Omega^2} c \,\mathrm{d}\pi + \mathrm{KL}(P_{0\#}\pi|\mu) + \mathrm{KL}(P_{1\#}\pi|\nu) \right).$$

Hellinger–Kantorovich Geodesics via Optimal Plans

Let $\mu, \nu \in \mathcal{M}_+(\Omega)$, π^* optimal and T^* be the Monge map $\pi^* = (\mathrm{Id} \times T^*)_{\#}\bar{\mu}$. Let $\bar{\mu} = P_{1\#}\pi^*$, $\bar{\nu} = P_{2\#}\pi^*$ and write

$$\mu = u\bar{\mu} + \mu^{\perp} \qquad\qquad \nu = w\bar{\nu} + \nu^{\perp}.$$

Then a geodesic is given by

$$\begin{aligned}
\bar{\rho}(t,\cdot) &= X\left(t;\cdot,u(\cdot),T^*(\cdot),w\circ T^*(\cdot)\right)_{\#}\left[M\left(t;\cdot,u(\cdot),T^*(\cdot),w\circ T^*(\cdot)\right)\bar{\mu}\right]\\
\rho(t,\cdot) &= \bar{\rho}(t,\cdot) + (1-t)^2\mu^{\perp} + t^2\nu^{\perp}\\
v(t,\cdot) &= \frac{\partial X}{\partial t}\left(t;\cdot,u(\cdot),T^*(\cdot),w\circ T^*(\cdot)\right)\\
\bar{\zeta}(t,\cdot) &= X\left(t;\cdot,u(\cdot),T^*(\cdot),w\circ T^*(\cdot)\right)_{\#}\left[\frac{\partial M}{\partial t}\left(t;\cdot,u(\cdot),T^*(\cdot),w\circ T^*(\cdot)\right)\bar{\mu}\right]\\
\zeta(t,\cdot) &= \bar{\zeta}(t,\cdot) - 2(1-t)\mu^{\perp} + 2t\nu^{\perp}.
\end{aligned}$$

where

$$M(t;x_0,m_0,x_1,m_1) = (1-t)^2 m_0 + t^2 m_1 + 2t(1-t)\sqrt{m_0 m_1}\cos\|x_0 - x_1\|$$

$$\varphi(t) = \frac{1}{\|x_0 - x_1\|}\cos^{-1}\left(\frac{(1-t)\sqrt{m_0} + t\sqrt{m_1}\cos(\|x_0 - x_1\|)}{\sqrt{M(t)}}\right)$$

$$X(t;x_0,m_0,x_1,m_1) = x_0 + (x_1 - x_0)\varphi(t).$$

Time Independent Benamou–Brenier Form

1. Let $\mu, \nu \in \mathcal{M}_+(\Omega)$ and $\pi^* = (\mathrm{Id} \times T^*)_\# \bar{\mu}$ be optimal.
2. Let (ρ, ω, ζ) be the geodesics constructed on the previous slide.
3. Set for $t \in [0, 1)$:
$$\alpha(t, \cdot) = \frac{\mathrm{d}\bar{\zeta}_t}{\mathrm{d}\rho_t} - 2(1-t)\frac{\mathrm{d}\mu^\perp}{\mathrm{d}\rho_t}.$$
4. Then[23]
$$v(0,x) = \begin{cases} \frac{T^*(x)-x}{\|T^*(x)-x\|}\sqrt{\frac{w(T^*(x))}{u(x)}}\sin(\|T^*(x)-x\|) & \bar{\mu}\text{-a.e.}, \\ 0 & \mu^\perp\text{-a.e.}, \end{cases}$$
$$\alpha(0,x) = \begin{cases} 2\left(\sqrt{\frac{w(T^*(x))}{u(x)}}\cos(\|T^*(x)-x\|) - 1\right) & \bar{\mu}\text{-a.e.}, \\ -2 & \mu^\perp\text{-a.e.} \end{cases}$$

and
$$\mathrm{d}^2_{\mathrm{HK}}(\mu, \nu) = \int_\Omega \left(\|v(0,\cdot)\|^2 + \frac{1}{4}(\alpha(0,\cdot))^2\right) \mathrm{d}\mu + \|\nu^\perp\|.$$

[23] **Cai, Cheng, Schmitzer and T.**, *The Linearized Hellinger–Kantorovich Distance*, SIAM Journal on Imaging Sciences, 15(1):45–83, 2022.

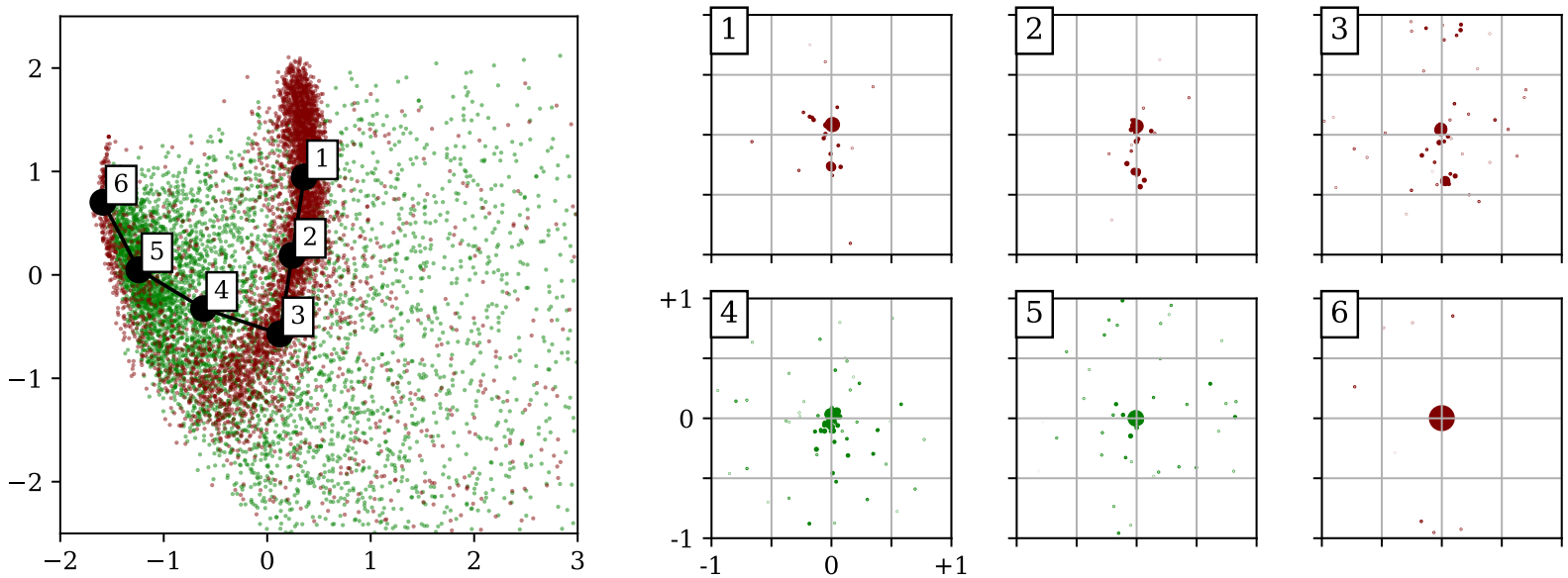
The Linear Hellinger–Kantorovich Distance

1. One can show that $\mu \perp \nu^\perp$.
2. In particular, if $\mathrm{spt}(\mu) = \Omega$ then $\nu^\perp = 0$, and
$$\mathrm{d}^2_{\mathrm{HK}}(\mu, \nu) = \int_\Omega \left(\|v(0,\cdot)\|^2 + \frac{1}{4}(\alpha(0,\cdot))^2 \right) \mathrm{d}\mu.$$
3. 'Logarithmic' map:
$$\mathrm{Log}_{\mathrm{HK}}(\mu; \mu_i) := (v(0,\cdot), \alpha(0,\cdot)).$$
4. Hellinger–Kantorovich distance from reference:
$$\mathrm{d}_{\mathrm{HK}}(\mu, \nu) = \|\mathrm{Log}_{\mathrm{HK}}(\mu; \nu)\|_{\mathrm{L}^2(\mu)}.$$
5. Linear Hellinger–Kantorovich distance:[24]
$$\mathrm{d}_{\mathrm{HK,lin},\mu}(\mu_1, \mu_2) = \|\mathrm{Log}_{\mathrm{HK}}(\mu; \mu_2) - \mathrm{Log}_{\mathrm{HK}}(\mu; \mu_2)\|_{\mathrm{L}^2(\mu)}.$$
6. **Linear Hellinger–Kantorovich Assumption:**
$$\mathrm{d}_{\mathrm{HK}}(\mu_1, \mu_2) \approx \mathrm{d}_{\mathrm{HK,lin},\mu}(\mu_1, \mu_2).$$

[24] **Cai, Cheng, Schmitzer and T.**, *The Linearized Hellinger–Kantorovich Distance*, SIAM Journal on Imaging Sciences, 15(1):45–83, 2022.

Example 6: Particle Decay Classification

Aim: can we classify W boson jets and QCD (quark or gluon) jets?



Source: **Cai, Cheng, Schmitzer and T.**, *The Linearized Hellinger–Kantorovich Distance*, SIAM Journal on Imaging Sciences, 15(1):45–83, 2022.

Contents

The TL^p Distance

1. The Wasserstein distance treats signals/images as probability measures. This can be restrictive. Can we generalise whilst keeping the nice properties of optimal transport?
2. We want to treat the signal/image as a function rather than a measure.
3. The idea is to treat signals as a pair (f, μ) where $f \in \mathrm{L}^p(\mu)$.
4. Mostly we consider when μ is the uniform measure (either continuous or discrete), but one could also trivially adapt in order to weight features of the signal, for example.
5. Note that we can compare signals on different domains.
6. TL^p definition (Monge formulation):[25] (a.k.a. graph space optimal transport[26])

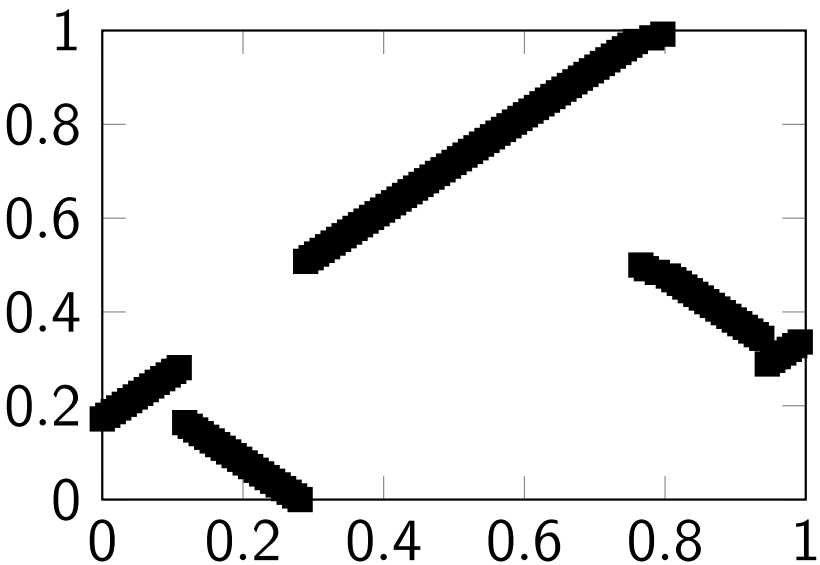
$$\mathrm{d}^p_{\mathrm{TL}^p}((f,\mu),(g,\nu)) = \inf_{T:T_{\#}\mu=\nu} \int_X |x - T(x)|^p + \lambda |f(x) - g(T(x))|^p \,\mathrm{d}\mu(x).$$

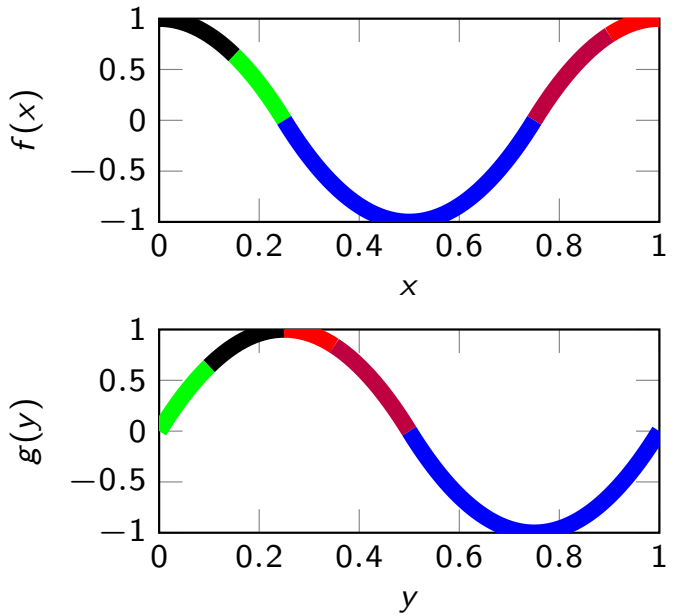
[25] **García Trillos and Slepčev**, *Continuum Limit of Total Variation on Point Clouds*, ARMA, 220:193–241, 2016.

[26] **Métivier, Brossier, Mérigot and Oudet**, *A Graph Space Optimal Transport Distance as a Generalization of L^p Distances: Application to a Seismic Imaging Inverse Problem*, Inverse Problems, 35(8), 2019.

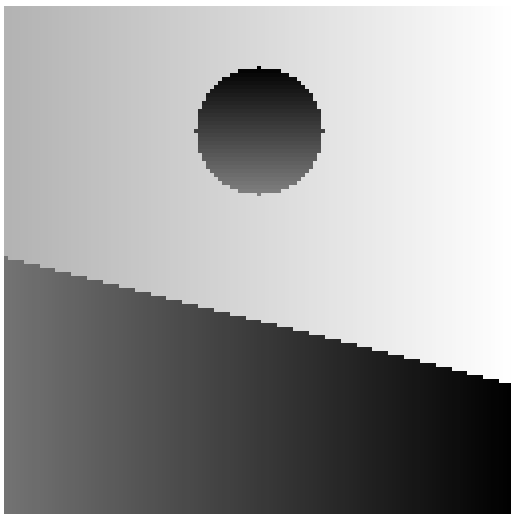
A Simple Example

For example consider the functions $f(x) = \cos(2\pi x)$ and $g(y) = \sin(2\pi y)$ defined on $[0, 1]$ with the uniform measure. The optimal plan using the TL^2 distance is given below.



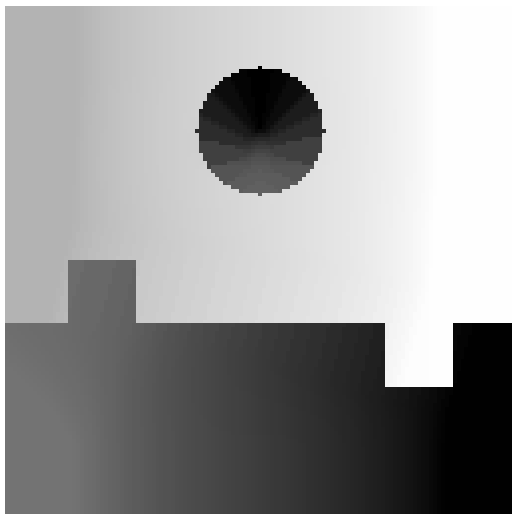


Example 7: Synthetic Colour Transfer



(a) Exemplar image.

(b) Original image to be shaded.



(c) The optimal transport solution.

Source: **T., Park, Kolouri, Rohde and Slepčev**, *A Transportation Lp Distance for Signal Analysis*, Journal of Mathematical Imaging and Vision, 59(2):187–210, 2017.

Example 8: Real World Colour Transfer

(a) Exemplar image.

(b) Original image to be coloured.

(c) The optimal transport solution.

(d) The TL^2 solution.

(e) Reinhard, Ashikhmin, Gooch and Shirley's method.

(f) Pitié and Kokaram's method.

Source: **T., Park, Kolouri, Rohde and Slepčev**, *A Transportation Lp Distance for Signal Analysis*, Journal of Mathematical Imaging and Vision, 59(2):187–210, 2017.

The Linear TL^2 Distance

1. Fix a reference point $(f_0, \mu_0) \in \mathrm{TL}^2$ and let T_i be the TL^2-optimal transport map between (f_0, μ_0) and (f_i, μ_i).
2. I.e. $[T_i]_{\#}\mu_0 = \mu_i$ and

$$\mathrm{d}^2_{\mathrm{TL}^2}((f_i, \mu_i), (f_0, \mu_0)) = \int_X |x - T_i(x)|^2 + \lambda |f_0(x) - f_i(T_i(x))|^2 \, \mathrm{d}\mu_0(x).$$

3. Assume μ_0 has a density ρ_0 then we define

$$P(f_i, \mu_i) = (P_1(f_i, \mu_i), P_2(f_i, \mu_i)) \in \mathrm{L}^2(\mu_0; \mathbb{R}^d) \times \mathrm{L}^2(\mu_0; \mathbb{R})$$

$$[P_1(f_i, \mu_i)](x) = (T_i(x) - x)\sqrt{\rho_0(x)}$$

$$[P_2(f_i, \mu_i)](x) = (f_i(T_i(x)) - f_0(x))\sqrt{\lambda \rho_0(x)}.$$

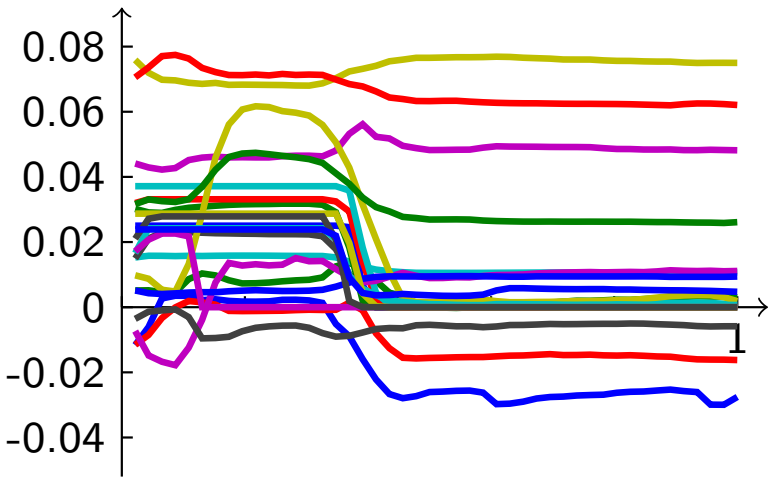
4. Then define

$$\mathrm{d}_{\mathrm{TL}^2, \mathrm{lin}, (f_0, \mu_0)}((f_i, \mu_i), (f_j, \mu_j)) = \|P(f_i, \mu_i) - P(f_j, \mu_j)\|_{\mathrm{L}^2}.$$

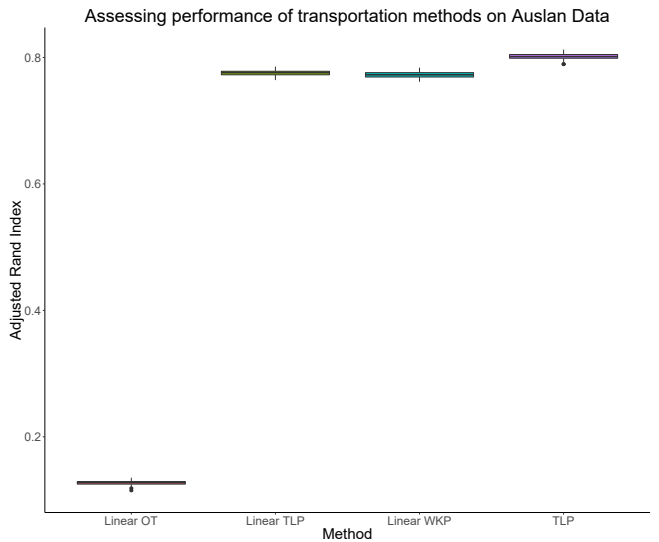
Example 9: AUSLAN Classification

Data: The AUSLAN data set is a set of 95 words 'spoken' by a native AUSLAN (Australian sign language) using 22 sensors on a cyberglove. There are 27 signals in each class, so a total of 2565 signals.

Example data:



Classification results:



Source: **Crook, Cucuringu, Hurst, Schönlieb, T. and Zygalakis**, *A Linear Transportation Lp Distance for Pattern Recognition*, Pattern Recognition 147, 2024.

Take Home Messages

+ Optimal transport distances are often a good model for data and using them in your method can often lead to improved performance.
+ There is a rich theory which includes geodesics and a formal Riemanian structure.
+ Restrictions, such as the data needs to be normalised, are not prohibitive as there are optimal transport distances enjoying many of the same theoretical properties as Wassserstein distances in the unbalanced setting.
– Computation is improving all the time, but is still often much more expensive than Euclidean distances.

References I

1. **Angenent, Haker and Tannenbaum**, *Minimizing flows for the Monge–Kantorovich problem*, SIAM journal on mathematical analysis, 35(1):61–97, 2003.
2. **Arjovsky, Chintala and Bottou**, *Wasserstein Generative Adversarial Networks*, International Conference on Machine Learning, 2017.
3. **Benamou and Brenier**, *A Computational Fluid Mechanics Solution to the Monge-Kantorovich Mass Transfer Problem*, Numerische Mathematik, 84(3):375–393, 2000.
4. **Benamou, Froese and Oberman**, *Numerical Solution of the Optimal Transportation Problem Using the Monge-Ampere Equation*, Journal of Computational Physics, 260:107–126, 2014.
5. **Cai, Cheng, Schmitzer and T.**, *The Linearized Hellinger–Kantorovich Distance*, SIAM Journal on Imaging Sciences, 15(1):45–83, 2022.
6. **Chartrand, Vixie, Wohlberg and Bollt**, *A Gradient Descent Solution to the Monge-Kantorovich Problem*, Applied Mathematical Sciences, 3(22):1071–1080, 2009.
7. **Chizat, Peyré, Schmitzer and Vialard**, *An Interpolating Distance Between Optimal Transport and Fisher–Rao Metrics*, Foundations of Computational Mathematics 18:1–44, 2018.

References II

8. **Chizat, Peyré, Schmitzer and Vialard**, *Unbalanced Optimal Transport: Dynamic and Kantorovich formulations*, Journal of Functional Analysis, 274:3090–3123, 2018.
9. **Crook, Cucuringu, Hurst, Schönlieb, T. and Zygalakis**, *A Linear Transportation Lp Distance for Pattern Recognition*, Pattern Recognition 147, 2024.
10. **Cuturi**, *Sinkhorn Distances: Lightspeed Computation of Optimal Transport*, In Advances in Neural Information Processing Systems, pp. 2292–2300, 2013.
11. **Engquist and Yang**, *Seismic Imaging and Optimal Transport*, Communications in Information and Systems, 19(2):95–145, 2019.
    García Trillos and Slepčev, *Continuum Limit of Total Variation on Point Clouds*, ARMA, 220:193–241, 2016.
12. **Gigli**, *On Hölder Continuity-in-Time of the Optimal Transport Map Towards Measures Along a Curve*, Proceedings of the Edinburgh Mathematical Society, 54(2):401–409, 2011.
13. **Goodfellow, Pouget-Abadie, Mirza, Xu, Warde-Farley, Ozair, Courville and Bengio**, *Generative Adversarial Networks*, arxiv:1406.2661, 2014.
14. **Hamm, Moosmueller, Schmitzer and T.**, *Manifold Learning in Wasserstein Space*, arxiv:2311.08549, 2023.
15. **Jacobs and Léger**, *A Fast Approach to Optimal Transport; The Back-And-Forth Method*, Numerische Mathematik, 146(3):513–544, 2020.

References III

16. **Khurana, Kannan, Cloninger and Moosmüller**, *Supervised Learning of Sheared Distributions Using Linearized Optimal Transport*, Sampling Theory, Signal Processing, and Data Analysis, 21(1), 2023.
17. **Kolouri, Park, T., Slepčev and Rohde**, *Optimal Mass Transport: Signal Processing and Machine Learning Applications*, IEEE Signal Processing Magazine, 34(4):43–59, 2017.
18. **Kolouri, Pope, Martin and Rohde**, *Sliced Wasserstein auto-encoders*, ICLR, 2018.
19. **Kolouri and Rohde**, *Optimal transport a crash course*, IEEE ICIP 2016 Tutorial Slides: Part 1, 2016.
20. **Kondratyev, Monsaingeon and Vorotnikov**, *A New Optimal Transport Distance On The Space Of Finite Radon Measures*, Advances in Differential Equations, 21:1117–1164, 2016.
21. **Levy**, *A Numerical Algorithm for L_2 Semi-Discrete Optimal Transport in 3D*, ESAIM Math. Model. Numer. Anal., 49(6):1693–1715, 2015.
22. **Liero, Mielke and Savaré**, *Optimal Entropy-Transport Problems and a New Hellinger–Kantorovich Distance Between Positive Measures*, Inventiones Mathematicae 211:969–1117, 2018.
23. **Métivier, Brossier, Mérigot and Oudet**, *A Graph Space Optimal Transport Distance as a Generalization of L^p Distances: Application to a Seismic Imaging Inverse Problem*, Inverse Problems, 35(8), 2019.

References IV

24. **Nowozin, Cseke and Tomioka** *f-GAN: Training generative neural samplers using variational divergence minimization*, NeuRIPS, 2016.
25. **Park and T.**, *Representing and Learning High Dimensional Data with the Optimal Transport Map from a Probabilistic Viewpoint*, CVPR, 2018.
26. **Stanczuk, Etmann, Kreusser and Schönlieb**, *Wasserstein GANs Work Because They Fail (To Approximate the Wasserstein distance)*, arxiv:2103.01678, 2021.
27. **T., Park, Kolouri, Rohde and Slepčev**, *A Transportation Lp Distance for Signal Analysis*, Journal of Mathematical Imaging and Vision, 59(2):187–210, 2017.
28. **Tolstikhin, Bousquet, Gelly and Schölkopf**, *Wasserstein Auto-Encoders*, International Conference on Learning Representations, 2018.
29. **Wang, Slepčev, Basu, Ozolek and 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.

Thank you for listening!

In theory, there is no difference between theory and practice. But in practice, there is.

— Yogi Berra