Linearization of Balanced and Unbalanced Optimal Transport

Bernhard Schmitzer

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

July 26, 2021

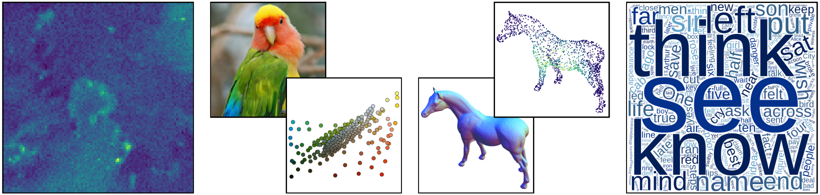
Overview

1. Introduction to optimal transport
2. Linearized optimal transport
3. Unbalanced transport and linearization
4. Conclusion and open questions

Overview

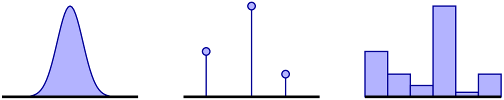
Metric measure spaces for data modelling

Comparing and understanding data



- ‘Are two samples similar?’

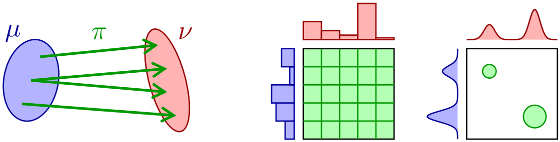
Language: positive Radon measures $\mathcal{M}_+(X)$ on metric space (X, d)



$\mathrm{d}\mu(x) = f(x)\,\mathrm{d}x$ $\quad$ $\mu = \sum_i \mu_i \, \delta_{x_i}$ $\quad$ $\mu \in \mathbb{R}^n_+$

- similarity of samples $\Leftrightarrow$ metric on $\mathcal{M}_+(X)$

Couplings and optimal transport



Couplings

- $\Pi(\mu,\nu) := \{\pi \in \mathcal{M}_+(X \times X) : \mathrm{P}_{1\sharp}\pi = \mu,\ \mathrm{P}_{2\sharp}\pi = \nu\}$
- **marginals**: $\mathrm{P}_{1\sharp}\pi(A) := \pi(A \times X)$, $\mathrm{P}_{2\sharp}\pi(B) := \pi(X \times B)$
- **rearrangement** of mass, generalization of **map**

Optimal transport [Kantorovich, 1942]

$$C(\mu,\nu) := \inf \left\{ \int_{X \times X} c(x,y)\, \mathrm{d}\pi(x,y) \middle| \pi \in \Pi(\mu,\nu) \right\}$$

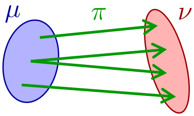
- **cost function** $c : X \times X \to \mathbb{R}$ for moving unit mass from x to y
- **convex problem**: linear program

Wasserstein distance on probability measures $\mathcal{P}(X)$

$$W_p(\mu,\nu) := \left(C(\mu,\nu)\right)^{1/p} \text{ for } c(x,y) := d(x,y)^p, \quad p \in [1,\infty)$$

Wasserstein distances: basic properties

$$W_p(\mu,\nu) := \inf\left\{ \int_{X\times X} d(x,y)^p \, \mathrm{d}\pi(x,y) \,\middle|\, \pi \in \Pi(\mu,\nu) \right\}^{1/p}$$



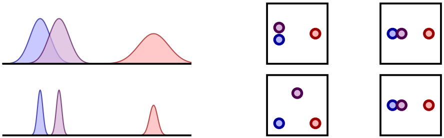
Properties

- ✓ **intuitive:** minimal $\pi \Rightarrow$ optimal assignment
  Thm: $[X = \mathbb{R}^d$, $\mu \ll$ Lebesgue, $c = d^p] \Rightarrow [\pi = (\mathrm{id}, T)_\sharp \mu]$

$$W_2(\mu,\nu) = \left(\int_X \| T(x) - x \|^2 \, \mathrm{d}\mu(x) \right)^{1/2}$$

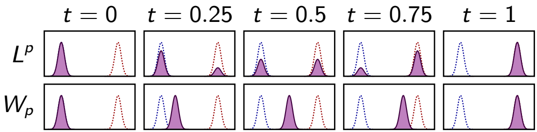
- ✓ metrizes weak$*$ convergence
- ✓ respects (X, d), **robust** to discretization errors, positional noise

$X = \mathbb{R}$, $\mu, \rho, \nu \in \mathcal{P}(X)$ $\quad L_p \quad W_p$



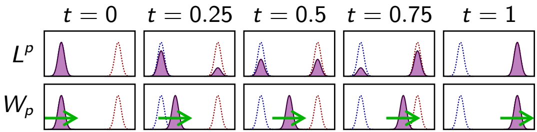
Wasserstein distances: displacement interpolation

- (X, d) length space $\Rightarrow (\mathcal{P}(X), W_p)$ is length space
- $d(x, y)$ is **length of shortest continuous path** between x and y



Wasserstein distances: displacement interpolation

- (X, d) length space $\Rightarrow$ $(\mathcal{P}(X), W_p)$ is length space
- $d(x, y)$ is **length of shortest continuous path** between x and y



Dynamic formulation: Benamou–Brenier formula (on $X = \mathbb{R}^d$)

- (weak) **continuity equation**: mass ρ, velocity field v

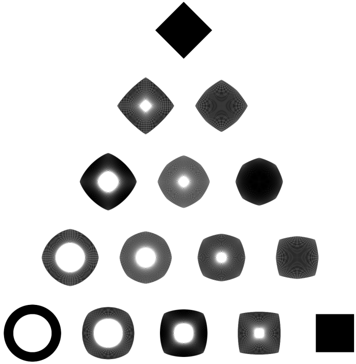
$$\mathcal{CE}(\mu, \nu) := \{(\rho, v) \colon \partial_t \rho + \nabla(v \cdot \rho) = 0,\ \rho_0 = \mu,\ \rho_1 = \nu\}$$

- **least action principle**: minimize Lagrangian / kinetic energy

$$W_2(\mu, \nu)^2 = \inf_{(\rho, v) \in \mathcal{CE}(\mu, \nu)} \int_{[0,1] \times X} \|v_t\|^2 \, \mathrm{d}\rho_t \mathrm{d}t$$

Wasserstein distances: barycenter

[Agueh and Carlier, 2011]



W_2 barycenter: weighted center of mass

- measures $(\mu_i)_{i=1}^n$
- weights $(\lambda_i)_{i=1}^n$, $\lambda_i \geq 0$, $\sum_{i=1}^n \lambda_i = 1$
- **barycenter:** $\text{argmin}_\nu \sum_{i=1}^n \lambda_i \, W_2^2(\nu, \mu_i)$

Overview

1. Introduction to optimal transport
2. Linearized optimal transport
3. Unbalanced transport and linearization
4. Conclusion and open questions

What now?

What we have

- ✓ OT induces meaningful, robust metric on samples
- ✗ numerically more involved, ✓ but good solvers exist
- ✓ can interpolate and average

But...

- ✗ analyzing point clouds in non-linear metric space is tricky
  - ✓ approximate Euclidean embeddings
  - ✗ interpretation not obvious
- ✗ requires computation of all pairwise distances

Wasserstein-2: Local linearization

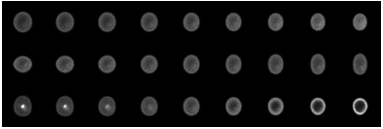
Recall Brenier Theorem

- $[X = \mathbb{R}^d,\ \mu \ll \text{Lebesgue},\ c = d^2] \Rightarrow [\pi = (\text{id}, T)_\sharp \mu]$

$$W_2(\mu, \nu) = \left(\int_X \|T(x) - x\|^2 \, d\mu(x) \right)^{1/2}$$

Idea [Wang et al., 2012]

- set of samples $\{\nu_i\}_{i=1}^N$, **'reference'** measure μ
- represent ν_i by optimal T_i for $W_2(\mu, \nu_i)$, **Lagrangian** representation
- ✓ approximate distance $\text{Lin}W_2(\nu_i, \nu_j) := \|T_i - T_j\|_{L^2(\mu, \mathbb{R}^d)}$
- $\{T_i\}_{i=1}^N$ lie in $L^2(\mu, \mathbb{R}^d) \Rightarrow$ vector space
- ✓ only OT problems $W_2(\mu, \nu_i)$ need to be solved, not all $W_2(\nu_i, \nu_j)$
- ✓ simple post-processing, PCA, classifiers, ...



Wasserstein-2: Local linearization, interpretation

Benamou–Brenier formula (for $X = \mathbb{R}^d$):

$$W_2(\mu, \nu)^2 = \inf_{(\rho, v) \in \mathcal{CE}(\mu, \nu)} \int_{[0,1]\times X} \|v_t\|^2 \, \mathrm{d}\rho_t \mathrm{d}t$$

- $(\mathcal{P}(X = \mathbb{R}^d), W_2)$ 'looks like' **Riemannian manifold** [Otto, 2001] where **tangent space** at ρ_t is $L^2(\rho_t, \mathbb{R}^d)$

Gradient flows via minimizing movements [Ambrosio et al., 2008]

- heat equation as gradient flow of entropy w.r.t. W_2 [Jordan et al., 1998]

Logarithmic and exponential map

- let $\pi = (\mathrm{id}, T)_\sharp \mu$ optimal for $W_2^2(\mu, \nu)$

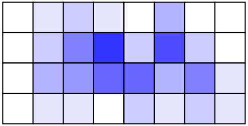
$$\mathrm{Log}_\mu(\nu) = v_0 = T - \mathrm{id}, \qquad \mathrm{Exp}_\mu(v_0) = (\mathrm{id} + v_0)_\sharp \mu$$

Interpretation of [Wang et al., 2012]

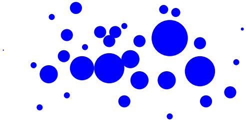
- local approximation of manifold by tangent plane at μ
- simple analogy: sphere
- some insights:
  - μ needs to be chosen carefully, close to $\{\nu_i\}$
    $\Rightarrow$ diameter of samples should not be too high
  - approximation quality depends on curvature of manifold
    [Gigli, 2011; Mérigot et al., 2020; Delalande and Merigot, 2021]

Comparing Eulerian and Lagrangian representation

Eulerian



Lagrangian

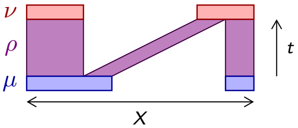


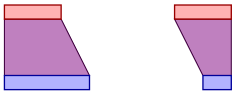
x	y	m
1.1	0.2	0.1
1.9	-0.1	0.2
⋮	⋮	⋮

- better choice depends on problem / context
- Lagrangian representation order invariant
- but consistent order makes comparison easier
- LinOT provides canonical order, ‘know which list items to compare’

Overview

Hellinger–Kantorovich distance: motivation [Kondratyev et al., 2016; Chizat et al., 2018c; Liero et al., 2018]





- **unbalanced continuity equation**: mass ρ, velocity v, **source** α

$$\mathcal{CE}(\mu,\nu) := \{(\rho, v, \alpha)\colon \partial_t \rho + \nabla(v \cdot \rho) = \alpha \cdot \rho,\ \rho_0 = \mu,\ \rho_1 = \nu\}$$

- **unbalanced Benamou–Brenier formula:**

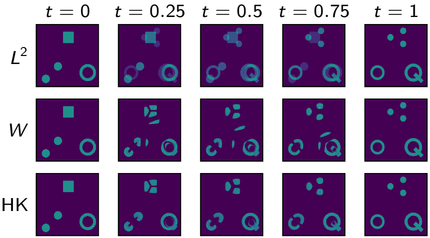
$$\mathrm{HK}(\mu,\nu)^2 := \inf_{(\rho,v,\alpha)\in\mathcal{CE}(\mu,\nu)} \int_{[0,1]\times X} \left[\|v_t\|^2 + \frac{1}{4}\alpha_t^2\right] \mathrm{d}\rho_t \mathrm{d}t$$

- other **unbalanced models:**
  - [Dolbeault et al., 2009]
  - [Piccoli and Rossi, 2016]: TV/L_1-type penalty (cf. [Caffarelli and McCann, 2010]: optimal partial transport)
  - [Maas et al., 2015, 2017]: L_2 and L_2-L_1-type penalty

Hellinger–Kantorovich distance: overview

$$\mathrm{HK}(\mu,\nu)^2 := \inf_{(\rho,v,\alpha)\in\mathcal{CE}(\mu,\nu)} \int_{[0,1]\times X} \left[\|v_t\|^2 + \frac{\kappa}{4}\alpha_t^2\right] \mathrm{d}\rho_t \mathrm{d}t$$

- **Thm:** HK is geodesic distance on **non-negative measures**
  - geodesics well understood [Liero et al., 2018; Chizat et al., 2018a]
  - weak Riemannian structure [Kondratyev et al., 2016; Liero et al., 2016]
- transport up to $\frac{\kappa\pi}{2}$, pure Hellinger after that
  choose κ by physical intuition and cross-validation
- simple **numerical approximation** via **entropic regularization** and **Sinkhorn**-type algorithm [Chizat et al., 2018b]
- **barycenters** [Chung and Phung, 2020; Friesecke et al., 2021]

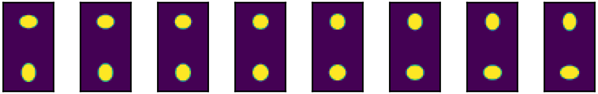


Hellinger–Kantorovich distance: local linearization

with Tianji Cai, Junyi Cheng, Matthew Thorpe [Cai et al., 2021]

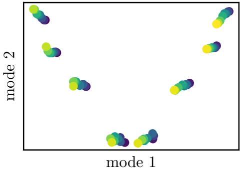
$$\mathrm{HK}(\mu,\nu)^2 := \inf_{(\rho,v,\alpha)\in\mathcal{CE}(\mu,\nu)} \int_{[0,1]\times X} \left[\|v_t\|^2 + \frac{1}{4}\alpha_t^2\right] \mathrm{d}\rho_t \mathrm{d}t$$

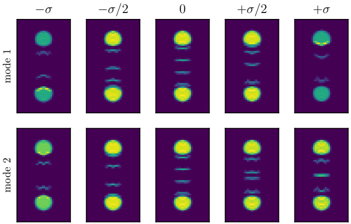
Samples: (varying ellipticities and radii)



$$\mathrm{Log}_\mu^{W_2}(\nu) = v_0 \qquad \mathrm{Log}_\mu^{\mathrm{HK}}(\nu) = (v_0, \alpha_0, \sqrt{\nu^\perp})$$

Principal component analysis in tangent space (W_2):



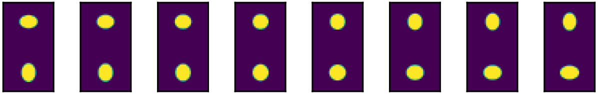


Hellinger–Kantorovich distance: local linearization

with Tianji Cai, Junyi Cheng, Matthew Thorpe [Cai et al., 2021]

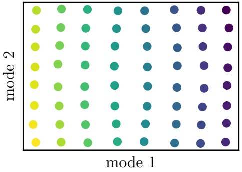
$$\mathrm{HK}(\mu,\nu)^2 := \inf_{(\rho,v,\alpha)\in\mathcal{CE}(\mu,\nu)} \int_{[0,1]\times X} \left[\|v_t\|^2 + \frac{1}{4}\alpha_t^2\right] \mathrm{d}\rho_t \mathrm{d}t$$

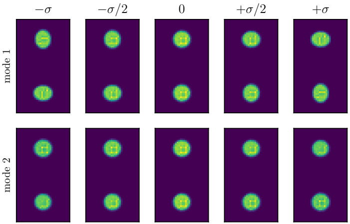
Samples: (varying ellipticities and radii)



$$\mathrm{Log}_\mu^{W_2}(\nu) = v_0 \qquad \mathrm{Log}_\mu^{\mathrm{HK}}(\nu) = (v_0, \alpha_0, \sqrt{\nu^\perp})$$

Principal component analysis in tangent space (HK):



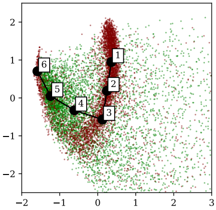


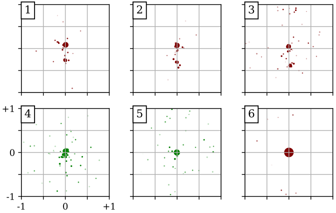
Hellinger–Kantorovich distance: local linearization, cont.

[Cai et al., 2021]

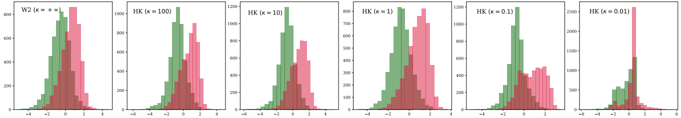
Example application: classification of particle jets

- mass represents energy absorbed in detector plane
- separate weak (red) vs strong (green) decay channels





- LDA: better separation with unbalanced HK metric



Overview

Conclusion

Optimal transport

- ✓ **intuitive, robust, flexible** metric for probability measures
- ✓ **rich geometric structure** (Riemannian flavour...)
- ✓ accessible by **convex optimization**

Local linearization of OT [Wang et al., 2012]

- ✓ **Lagrangian representation**: combine OT metric with **linear structure**
- ✓ **intuitive interpretation** of tangent vectors
- ✓ **useful representation** for subsequent machine learning analysis

Unbalanced transport

- ✓ more **robust to mass fluctuations**
- ✓ carries over to **linearization** [Cai et al., 2021]
- ✓ hyperparameter κ **easy to tune**
- ✓ formulas look scary, but **numerics** almost the same

Example code:
`https://github.com/bernhard-schmitzer/UnbalancedLOT`

Open questions

How well does the linear approximation work?

- tricky question! not my area of expertise, guess: can be arbitrarily bad without regularity assumptions on samples
- some preliminary related results for Wasserstein-2 case exist [Gigli, 2011; Mérigot et al., 2020; Delalande and Merigot, 2021], still open for HK

Riemannian structure of HK metric

- if we have the embedding of two samples, take their average and then apply the exponential map, do we get a valid data point?
- more mathematically: is [range of the logarithmic map] = [domain of the exponential map] convex?
- regularity of logarithmic map?

Beyond simple one-point-linearization

- 'local triangulation' of a sub-manifold?
- barycentric subspace analysis [Pennec, 2018; Bonneel et al., 2016]?

Statistical questions

- how robust is the analysis under sampling of the samples?
- what if samples are themselves only empirical measures?

Even better interpretation of tangent vectors

- relevance for medical diagnosis

References I

M. Agueh and G. Carlier. Barycenters in the Wasserstein space. *SIAM J. Math. Anal.*, 43(2):904–924, 2011. doi: 10.1137/100805741.

L. Ambrosio, N. Gigli, and G. Savaré. *Gradient Flows in Metric Spaces and in the Space of Probability Measures*. Lectures in Mathematics. Birkhäuser Boston, 2nd edition, 2008.

N. Bonneel, G. Peyré, and M. Cuturi. Wasserstein barycentric coordinates: Histogram regression using optimal transport. *ACM Trans. Graph.*, 35(4), 2016. doi: 10.1145/2897824.2925918.

L. A. Caffarelli and R. J. McCann. Free boundaries in optimal transport and Monge-Ampère obstacle problems. *Annals of Math.*, 171(2):673–730, 2010.

T. Cai, J. Cheng, B. Schmitzer, and M. Thorpe. The linearized Hellinger–Kantorovich distance. arXiv:2102.08807, 2021.

L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard. Unbalanced optimal transport: Dynamic and Kantorovich formulations. *J. Funct. Anal.*, 274(11):3090–3123, 2018a. doi: 10.1016/j.jfa.2018.03.008.

L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard. Scaling algorithms for unbalanced optimal transport problems. *Math. Comp.*, 87:2563–2609, 2018b. doi: 10.1090/mcom/3303.

References II

L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard. An interpolating distance between optimal transport and Fisher–Rao metrics. *Found. Comp. Math.*, 18(1): 1–44, 2018c. doi: 10.1007/s10208-016-9331-y.

N.-P. Chung and M.-N. Phung. Barycenters in the Hellinger–Kantorovich space. *Appl Math Optim*, 2020. doi: 10.1007/s00245-020-09695-y.

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

J. Dolbeault, B. Nazaret, and G. Savaré. A new class of transport distances between measures. *Calc. Var. Partial Differential Equations*, 34(2):193–231, 2009. doi: 10.1007/s00526-008-0182-5.

G. Friesecke, D. Matthes, and B. Schmitzer. Barycenters for the Hellinger–Kantorovich distance over Rd. *SIAM J. Math. Anal.*, 53(1):62–110, 2021. doi: 10.1137/20M1315555.

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.

R. Jordan, D. Kinderlehrer, and F. Otto. The variational formulation of the Fokker-Planck equation. *SIAM J. Math. Anal.*, 29(1):1–17, 1998.

L. V. Kantorovich. O peremeshchenii mass. *Doklady Akademii Nauk SSSR*, 37(7–8): 227–230, 1942.

References III

S. Kondratyev, L. Monsaingeon, and D. Vorotnikov. A new optimal transport distance on the space of finite Radon measures. *Adv. Differential Equations*, 21(11-12): 1117–1164, 2016.

M. Liero, A. Mielke, and G. Savaré. Optimal transport in competition with reaction: the Hellinger–Kantorovich distance and geodesic curves. *SIAM J. Math. Anal.*, 48 (4):2869–2911, 2016.

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. doi: 10.1007/s00222-017-0759-8.

J. Maas, M. Rumpf, C. Schönlieb, and S. Simon. A generalized model for optimal transport of images including dissipation and density modulation. *ESAIM Math. Model. Numer. Anal.*, 49(6):1745–1769, 2015.

J. Maas, M. Rumpf, and S. Simon. Transport based image morphing with intensity modulation. In F. Lauze, Y. Dong, and A. B. Dahl, editors, *Scale Space and Variational Methods (SSVM 2017)*, pages 563–577. Springer, 2017. doi: 10.1007/978-3-319-58771-4_45.

Q. Mérigot, A. Delalande, and F. Chazal. Quantitative stability of optimal transport maps and linearization of the 2-Wasserstein space. In *International Conference on Artificial Intelligence and Statistics*, pages 3186–3196, 2020.

References IV

F. Otto. The geometry of dissipative evolution equations: the porous medium equation. *Comm. Partial Differential Equations*, 26(1-2):101–174, 2001. doi: 10.1081/PDE-100002243.

X. Pennec. Barycentric subspace analysis on manifolds. *Ann. Statist.*, 46(6A): 2711–2746, 2018. doi: 10.1214/17-AOS1636.

B. Piccoli and F. Rossi. On properties of the generalized Wasserstein distance. *Arch. Rat. Mech. Analysis*, 222(3):1339–1365, 2016. doi: 10.1007/s00205-016-1026-7.

W. Wang, D. Slepčev, S. Basu, J. A. Ozolek, and G. K. Rohde. A linear optimal transportation framework for quantifying and visualizing variations in sets of images. *Int. J. Comp. Vision*, 101:254–269, 2012.