Leveraging Regularization, Projections and Elliptical Distributions in Optimal Transport

PhD defense

Boris Muzellec (CREST, ENSAE)

Supervisor: Marco Cuturi

October 26th, 2020

Optimal transport, then and today

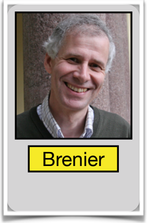
My favorite Panini trading cards:

Monge

Kantorovich

Nobel'75

Dantzig



Brenier

Villani

Fields'10

Figalli

Fields'18

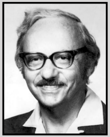
Optimal transport, then and today

My favorite Panini trading cards:

Monge

Kantorovich

Nobel’75


Dantzig

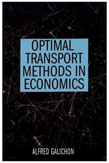
Brenier

Villani

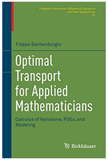
Fields’10

Figalli

Fields’18


OPTIMAL TRANSPORT METHODS IN ECONOMICS

ALFRED GALICHON


Progress in Nonlinear Differential Equations and Their Applications 87

Filippo Santambrogio

Optimal Transport for Applied Mathematicians

Calculus of Variations, PDEs, and Modeling

Birkhäuser

CÉDRIC VILLANI

Volume 338

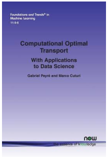
Grundlehren der mathematischen Wissenschaften

A Series of Comprehensive Studies in Mathematics

OPTIMAL TRANSPORT

OLD AND NEW

Springer


Foundations and Trends® in Machine Learning 11:5-6

Computational Optimal Transport

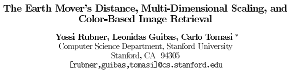
With Applications to Data Science

Gabriel Peyré and Marco Cuturi

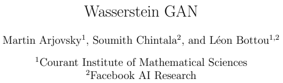
now the essence of knowledge

OT: Everybody does it!

Pure maths, applied maths, economics, biology, ML...

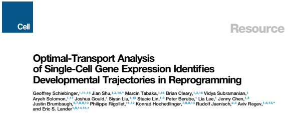

The Earth Mover’s Distance, Multi-Dimensional Scaling, and Color-Based Image Retrieval

Yossi Rubner, Leonidas Guibas, Carlo Tomasi *
Computer Science Department, Stanford University
Stanford, CA 94305
[rubner,guibas,tomasi]@cs.stanford.edu


Wasserstein GAN

Martin Arjovsky[1], Soumith Chintala[2], and Léon Bottou[1,2]

[1]Courant Institute of Mathematical Sciences
[2]Facebook AI Research


Cell

Resource

Optimal-Transport Analysis of Single-Cell Gene Expression Identifies Developmental Trajectories in Reprogramming

Geoffrey Schiebinger,[1,11,16] Jian Shu,[1,2,16,*] Marcin Tabaka,[1,16] Brian Cleary,[1,3,16] Vidya Subramanian,[1] Aryeh Solomon,[1,17] Joshua Gould,[1] Siyan Liu,[1,15] Stacie Lin,[1,6] Peter Berube,[1] Lia Lee,[1] Jenny Chen,[1,4] Justin Brumbaugh,[5,7,8,9,10] Philippe Rigollet,[11,12] Konrad Hochedlinger,[7,8,9,13] Rudolf Jaenisch,[2,3] Aviv Regev,[1,6,13,*] and Eric S. Lander[1,6,14,18,*]

What this thesis is about

OT applications: challenges

Optimal transport is an old mathematical problem, with a rich theory still actively studied and applications that are gaining momentum. But:

- It is hard to compute efficiently;
- It has unfavorable statistical properties.

Workarounds

- Adding regularization to the primal problem:
  - **Entropic** [Cuturi 2013; J. Solomon et al. 2015; Genevay, Cuturi, et al. 2016; Genevay, Peyre, et al. 2018],
  - **Quadratic** [Dessein et al. 2018; Blondel et al. 2018], **Tsallis** [Muzellec, Nock, et al. 2017],
  - **Unbalanced** [Frogner et al. 2015; Chizat et al. 2018; Schiebinger et al. 2019],
- Modeling maps/potentials using Neural Networks:
  - **Potentials:** WGAN [Arjovsky et al. 2017],
  - **Maps:** [Seguy et al. 2018], ICNN [Makkuva et al. 2020].

What this thesis is about

Workarounds (cont'd)

- Fall back on cases with closed forms:
  - **1D setting** (Sliced Wasserstein): [Rabin et al. 2011; Bonneel et al. 2015; Titouan et al. 2019; Kolouri et al. 2019],
  - **Gaussian/Elliptical distributions:** [Heusel et al. 2017; Chen, Georgiou, & Tannenbaum 2019], *this thesis*.

Contributions in this thesis

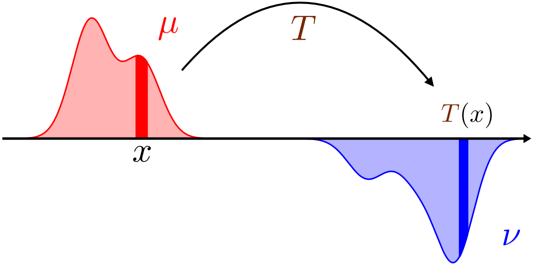
- Leverage the particular case of OT between **Gaussian/Elliptical** distr. to design efficient ML tools [NeurIPS'18];
- Lift transport maps from **low-dimensional projections**, with closed forms for Gaussians [NeurIPS'19];
- Prove closed forms of Gaussian OT with entropic **regularization** and unbalanced relaxation [NeurIPS'20];
- Apply **entropic** OT to missing data imputation [ICML'20].

Monge's problem

Monge's Problem (1781)[1]

How to move earth with minimal effort w.r.t. cost function $c(x,y)$?

$$\inf_{T:\mathbb{R}^d\to\mathbb{R}^d} \int_{\mathbb{R}^d} c(x,T(x))\mathrm{d}\mu(x) \quad \text{s.t.} \quad \forall A \subset \mathbb{R}^d, \nu(A) = \mu(T^{-1}(A)) \quad (\mathcal{M})$$



- Challenge: existence of a solution (ex: Dirac to 2 half-Diracs ✗).

[1] G. Monge. "Mémoire sur la théorie des déblais et des remblais". *Histoire de l'Académie Royale des Sciences de Paris* [1781].

Kantorovich's problem

Kantorovich's Problem (1942)[2]

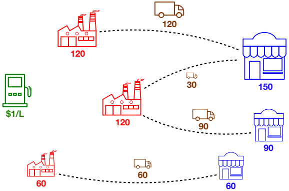
How to transport goods at a minimal cost?

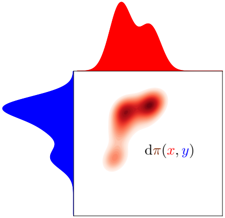
$$\inf_{\pi\in\mathcal{P}(\mathbb{R}^d\times\mathbb{R}^d)} \iint_{\mathbb{R}^d\times\mathbb{R}^d} c(x,y)\mathrm{d}\pi(x,y) \quad \text{s.t.} \quad p_{1\sharp}\pi=\mu, p_{2\sharp}\pi=\nu \qquad (\mathcal{K})$$

Transportation Plans

$$\Pi(\mu,\nu) \overset{\text{def}}{=} \{\pi\in\mathcal{P}(\mathbb{R}^d\times\mathbb{R}^d) : p_{1\sharp}\pi=\mu, p_{2\sharp}\pi=\nu\},\ p_1(x,y)=x, p_2(x,y)=y.$$

- Existence of a solution under very mild conditions.





[2] L. V. Kantorovich. "On the translocation of masses". *Dokl. Akad. Nauk. USSR*. 1942.

Computing OT in practice

Discrete-discrete: LP in $O(n^3 \log n)$, regularized approaches[3] in $O(n^2)$.

Discrete-continuous: density approx on grid[4], stochastic approx[5].

Continuous-continuous: In general, difficult.

- In low dimension, if $c = \| \cdot \|^2$:
  - Benamou-Brenier's[6] dynamic formulation (variational problem),
  - Equivalent to Monge-Ampère PDE (by Brenier's theorem[7]),
- In high dimension:
  - NN parameterization of potentials[8] or maps[9] (very active in ML),
  - Closed forms (this thesis):
    - Project to low dimension: Sliced Wasserstein[10][11],
    - Gaussians[12][13][14], Elliptical distributions[15].

[3] Cuturi 2013; [4] Mérigot 2011; [5] Genevay, Cuturi, et al. 2016; [6] Benamou et al. 2000; [7] Brenier 1987; [8] Arjovsky et al. 2017; [9] Makkuva et al. 2020; [10] Rabin et al. 2011; [11] Bonneel et al. 2015; [12] Dowson et al. 1982; [13] Olkin et al. 1982; [14] Takatsu 2011; [15] Gelbrich 1990.

Seminal closed forms

Regularizing the data to fall back to closed forms:

1-dimensional OT

If $\mu, \nu \in \mathcal{P}(\mathbb{R})$, quantile functions F_μ^{-1}, F_ν^{-1}, $c(x,y) = c(|x-y|)$, c convex,

$$\mathrm{OT}(\mu,\nu) = \int_0^1 c(|F_\mu^{-1}(x) - F_\nu^{-1}(x)|)\mathrm{d}x.$$

(Used in Sliced Wasserstein [Rabin et al. 2011; Bonneel et al. 2015].)

Bures-Wasserstein Geometry

Gaussian and elliptical distributions[4] with $c(x,y) = \|x-y\|^2$.

Let $\alpha = \mathcal{N}(\mathbf{a}, \mathbf{A})$, and $\beta = \mathcal{N}(\mathbf{b}, \mathbf{B})$. Then,

$$\begin{aligned}\mathrm{OT}(\alpha,\beta) &= \|\mathbf{m}_\alpha - \mathbf{m}_\beta\|^2 + \mathfrak{B}^2(\mathbf{A},\mathbf{B}) \\ &= \|\mathbf{m}_\alpha - \mathbf{m}_\beta\|^2 + \mathrm{Tr}\mathbf{A} + \mathrm{Tr}\mathbf{B} - 2\mathrm{Tr}(\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}})^{\frac{1}{2}}.\end{aligned}$$

(E.g. FID [Heusel et al. 2017], GMM-OT [Chen, Georgiou, & Tannenbaum 2019], this thesis.)

[4] M. Gelbrich. "On a formula for the L2 Wasserstein metric between measures on Euclidean and Hilbert spaces". *Mathematische Nachrichten* [1990].

Contributions of this thesis

Algorithms to optimize Bures-Wasserstein large ML applications, leveraging Riemannian structure for projection-free gradients & backprop.

Methods to extract maps and plans defined on the full space from subspace projections, in closed forms for Gaussian measures.

Entropic regularization and unbalanced relaxation of the Bures-Wasserstein geometry in closed form.

Distribution-preserving missing data imputation using entropy-regularized OT (not in this presentation).

Outline

Reminders: The Bures-Wasserstein Geometry on Elliptical Distributions

Learning with BW: Computing and Differentiating BW [NeurIPS'18]

Building OT Plans on Subspace Projections [NeurIPS'19]

Unbalanced Entropic OT for Gaussian Measures [NeurIPS'20]

Elliptical Distributions

Generalization of Gaussians: densities with elliptical level sets.

Definition (Elliptical Distributions[5])

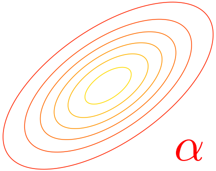
Let $\mathbf{a} \in \mathbb{R}^d$, $\mathbf{A} \in S_+^d$. Let $\lambda_{\mathrm{Im}\mathbf{A}}$ denote the Lebesgue measure over $\mathrm{Im}\mathbf{A}$. An *elliptical distribution* with mean $\mathbf{a}$ and scale parameter $\mathbf{A}$ is a probability measure of the form

$$\mathrm{d}\mu_{g,\mathbf{a},\mathbf{A}}(x) = g\left((x-\mathbf{a})^\top \mathbf{A}^\dagger (x-\mathbf{a})\right) \mathrm{d}\lambda_{\mathrm{Im}\mathbf{A}}(x),$$

with $g : \mathbb{R}^d \to \mathbb{R}_+$ s.t. $\int_{\mathrm{Im}\mathbf{A}} g(\|x\|^2_{\mathbf{A}^\dagger})\mathrm{d}\lambda_{\mathrm{Im}\mathbf{A}}(x) = 1$, and $\mathbf{A}^\dagger$ is the pseudo-inverse of $\mathbf{A}$.

Examples:

- Dirac measures ($\mathbf{A} = 0$),
- Gaussian distributions ($g(\cdot) \propto \exp(-\cdot/2)$),
- Uniform measure on ellipsoids ($g(\cdot) \propto \mathbb{1}_{\cdot \leq 1}$), ...



[5] M. Gelbrich. "On a formula for the L2 Wasserstein metric between measures on Euclidean and Hilbert spaces". *Mathematische Nachrichten* [1990].

The Bures-Wasserstein Geometry

From now on, $c(x,y) = \|x-y\|^2$:

- OT defines the 2-Wasserstein distance: $\mathrm{OT}(\mu,\nu) = W_2^2(\mu,\nu)$.

Theorem (Bures-Wasserstein Distance[6])

Let α and β be two elliptical distributions from the same family. Then,

$$W_2^2(\alpha,\beta) = \|\mathbf{m}_\alpha - \mathbf{m}_\beta\|^2 + \mathfrak{B}^2(\operatorname{Cov}\alpha, \operatorname{Cov}\beta),$$

with $\operatorname{Cov}(\alpha) \overset{\text{def}}{=} \mathbb{E}_{X\sim\alpha}[(X-\mathbf{m}_\alpha)(X-\mathbf{m}_\alpha)^\top] \propto \mathbf{A}$.

Definition (Bures Distance[7][8])

$$\forall \mathbf{A}, \mathbf{B} \in S_+^d, \quad \mathfrak{B}^2(\mathbf{A},\mathbf{B}) \overset{\text{def}}{=} \mathrm{Tr}\mathbf{A} + \mathrm{Tr}\mathbf{B} - 2\mathrm{Tr}(\mathbf{A}^{1/2}\mathbf{B}\mathbf{A}^{1/2})^{1/2}.$$

Defines a Riemannian metric on PSD matrices.

[6] M. Gelbrich. "On a formula for the L2 Wasserstein metric between measures on Euclidean and Hilbert spaces". *Mathematische Nachrichten* [1990].

[7] D. Bures. "An extension of Kakutani's theorem on infinite product measures to the tensor product of semifinite w*-algebras". *Trans. of the Am. Math. Soc.* [1969].

[8] R. Bhatia et al. "On the Bures-Wasserstein distance between positive definite matrices". *Expositiones Mathematicae* [2018].

Elliptical Monge Maps and Geodesics

Proposition (Gelbrich, 1990[9])

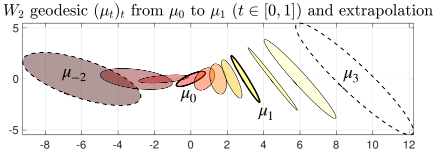
Let $\alpha = \mu_{g,\mathbf{a},\mathbf{A}}$ *and* $\beta = \mu_{g,\mathbf{b},\mathbf{B}}$ *be two elliptical distributions from the same family, s.t.* $\mathrm{Im}\mathbf{B} \subset \mathrm{Im}\mathbf{A}$. *Then,* $T_{\alpha\beta} : x \mapsto \mathbf{T}^{\mathbf{AB}}(x - \mathbf{a}) + \mathbf{b}$ *with*

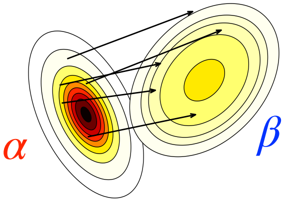
$$\mathbf{T}^{\mathbf{AB}} \overset{\text{def}}{=} \mathbf{A}^{\dagger/2}(\mathbf{A}^{1/2}\mathbf{B}\mathbf{A}^{1/2})^{1/2}\mathbf{A}^{\dagger/2}$$

is the Monge map from α *to* β, *where* $\mathbf{A}^{\dagger/2}$ *is the pseudo-inverse of* $\mathbf{A}^{1/2}$.

Riemannian geodesics[10]

$$\mathbf{C}_{\mathbf{AB}}(t) = [(1-t)\mathbf{I}_\mathrm{d} + t\mathbf{T}^{\mathbf{AB}}]\mathbf{A}[(1-t)\mathbf{I}_\mathrm{d} + t\mathbf{T}^{\mathbf{AB}}]$$





$(T_{\alpha\beta})_\sharp \alpha = \beta$

[9] M. Gelbrich. "On a formula for the L2 Wasserstein metric between measures on Euclidean and Hilbert spaces". *Mathematische Nachrichten* [1990].

[10] A. Takatsu. "Wasserstein geometry of Gaussian measures". *Osaka J. Math.* [2011].

Outline

Reminders: The Bures-Wasserstein Geometry on Elliptical Distributions

Learning with BW: Computing and Differentiating BW [NeurIPS'18]

Building OT Plans on Subspace Projections [NeurIPS'19]

Unbalanced Entropic OT for Gaussian Measures [NeurIPS'20]

Gradient-Based Optimization

- Many ML apps. can be cast as min problems. E.g. $\min_\theta \mathbb{E}_{x\sim\mathcal{P}}[l_\theta(x)]$.
- Classic method: (stochastic) gradient descent $\theta \leftarrow \theta - \eta \nabla_\theta l_\theta(x_i)$.

Bures-Wasserstein as a loss function

To fit models using Bures-Wasserstein *as a loss*, we need to be able to perform gradient steps. E.g. barycenter problem:

$$\text{Minimization problem:} \quad \min_{\alpha=\mathcal{N}(\mathbf{a},\mathbf{A})} \sum_i W_2^2(\alpha, \beta_i),$$

$$\text{Gradient updates:} \quad \mathbf{A} \leftarrow \mathbf{A} - \eta \nabla_{\mathbf{A}} \sum_i W_2^2(\alpha, \beta_i)$$

Can be generalized using chain rule and backprop to $\min_\alpha f(W_2^2(\alpha, \beta))$.

Computing and Differentiating the Bures Metric

$$\mathfrak{B}^2(\mathbf{A}, \mathbf{B}) = \mathrm{Tr}\mathbf{A} + \mathrm{Tr}\mathbf{B} - 2\mathrm{Tr}(\mathbf{A}^{1/2}\mathbf{B}\mathbf{A}^{1/2})^{1/2}$$

Challenge: How to compute and differentiate Bures?

Naive idea: differentiate matrix square roots with SVD. But:

- Expensive, hard to parallelize (in batches) on GPUs?[a]
- Automatic differentiation can be unstable (e.g. with non-distinct singular values, or singular matrices)

[a] Supported from `TensorFlow 1.12.0` (8/11/2018) and `Pytorch 1.2.0` (8/8/2019).

Explicit differentiation amounts to solving the Lyapunov equation:

$$(\mathrm{D}\mathbf{A}^{1/2})[\mathbf{H}]\mathbf{A}^{1/2} + \mathbf{A}^{1/2}(\mathrm{D}\mathbf{A}^{1/2})[\mathbf{H}] = \mathbf{H}.$$

The Monge Map is All You Need

The Bures distance and its gradient can be computed from $\mathbf{T}^{\mathbf{AB}}$.

$$\begin{aligned}\mathfrak{B}^2(\mathbf{A},\mathbf{B}) &= \mathrm{Tr}\mathbf{A} + \mathrm{Tr}\mathbf{B} - 2\mathrm{Tr}(\mathbf{A}^{1/2}\mathbf{B}\mathbf{A}^{1/2})^{1/2} \\ &= \mathrm{Tr}\mathbf{A} + \mathrm{Tr}\mathbf{B} - 2\mathrm{Tr}(\mathbf{A}\mathbf{T}^{\mathbf{AB}}).\end{aligned}$$

Proposition (Bures Gradient)

Let $\mathbf{A},\mathbf{B} \in S_+^d$ *s.t.* $\mathrm{Im}\mathbf{B} \subset \mathrm{Im}\mathbf{A}$. *Then* $\nabla_{\mathbf{A}}\mathfrak{B}^2(\mathbf{A},\mathbf{B}) = \mathbf{I}_\mathrm{d} - \mathbf{T}^{\mathbf{AB}}$*.*

If we need both $\nabla_{\mathbf{A}}\mathfrak{B}^2(\mathbf{A},\mathbf{B})$ and $\nabla_{\mathbf{B}}\mathfrak{B}^2(\mathbf{A},\mathbf{B})$, we need an efficient method to compute

- $\mathbf{T}^{\mathbf{AB}} = \mathbf{A}^{-1/2}(\mathbf{A}^{1/2}\mathbf{B}\mathbf{A}^{1/2})^{1/2}\mathbf{A}^{-1/2}$, and
- $\mathbf{T}^{\mathbf{BA}} = \mathbf{B}^{-1/2}(\mathbf{B}^{1/2}\mathbf{A}\mathbf{B}^{1/2})^{1/2}\mathbf{B}^{-1/2} = (\mathbf{T}^{\mathbf{AB}})^{-1}$.

Elliptical Monge Maps and Matrix Sign Function

Theorem (Higham, Mackey, Tisseur[11])

Let $\mathbf{A}, \mathbf{B} \in S^d_{++}$. *Then*

$$\operatorname{sign}\begin{pmatrix} 0 & \mathbf{B} \\ \mathbf{A} & 0 \end{pmatrix} = \begin{pmatrix} 0 & \mathbf{T}^{\mathbf{AB}} \\ \mathbf{T}^{\mathbf{BA}} & 0 \end{pmatrix},$$

where $\operatorname{sign}(\mathbf{M}) \overset{\text{def}}{=} \mathbf{M}\left(\mathbf{M}^2\right)^{-1/2}$.

$\mathbf{T}^{\mathbf{AB}}$ and $\mathbf{T}^{\mathbf{BA}}$ can be obtained using matrix sign iterations.

[11] N. J. Higham. *Functions of Matrices: Theory and Computation*. SIAM, 2008.

Newton-Schulz iterations

Newton-Schulz iterations

$$\begin{aligned} \mathbf{Y}_{k+1} &= \tfrac{1}{2}\mathbf{Y}_k(3\mathbf{I}_\mathrm{d} - \mathbf{Y}_k\mathbf{Z}_k\mathbf{Y}_k), \quad \mathbf{Y}_0 = \mathbf{B} \\ \mathbf{Z}_{k+1} &= \tfrac{1}{2}\mathbf{Z}_k(3\mathbf{I}_\mathrm{d} - \mathbf{Z}_k\mathbf{Y}_k\mathbf{Z}_k), \quad \mathbf{Z}_0 = \mathbf{A} \end{aligned}$$

Proposition (Higham[12])

If $\|\mathbf{I}_\mathrm{d} - \left(\begin{smallmatrix} 0 & \mathbf{B} \\ \mathbf{A} & 0 \end{smallmatrix}\right)^2\|_{op} < 1$, *then* $\mathbf{Y}_k \to \mathbf{T}^{\mathbf{AB}}$ *and* $\mathbf{Z}_k \to \mathbf{T}^{\mathbf{BA}}$ *quadratically*[a].

[a] i.e. $\exists c > 0, \|\mathbf{Y}_{k+1} - \mathbf{T}^{\mathbf{AB}}\|_\mathrm{op} \le c\|\mathbf{Y}_k - \mathbf{T}^{\mathbf{AB}}\|^2_\mathrm{op}$, and likewise for $\mathbf{Z}_k$.

Why bother?

- Easy to parallelize on GPUs (only requires matrix multiplications)
- Yields both $\mathbf{T}^{\mathbf{AB}}$ and $\mathbf{T}^{\mathbf{BA}}$: we get $\nabla_\mathbf{A}\mathfrak{B}^2(\mathbf{A},\mathbf{B})$ and $\nabla_\mathbf{B}\mathfrak{B}^2(\mathbf{A},\mathbf{B})$

We can now use the Bures-Wasserstein distance in gradient-based optimization.

[12] N. J. Higham. *Functions of Matrices: Theory and Computation*. SIAM, 2008.

Dealing with PSD constraints: avoiding projections

Last issue

- $\mathbf{A} - t\nabla_{\mathbf{A}}\mathfrak{B}^2(\mathbf{A},\mathbf{B})$ is not necessarily PSD.
- Projected gradient descent requires eigen-decomposition.

Classic workaround

- Use a $\mathbf{A} = \Pi(\mathbf{L}_{\mathbf{A}}) \overset{\text{def}}{=} \mathbf{L}_{\mathbf{A}}\mathbf{L}_{\mathbf{A}}{}^T$ parameterization.
- Effect on gradient methods?

Riemannian geodesics at the cost of Euclidean descent[13]

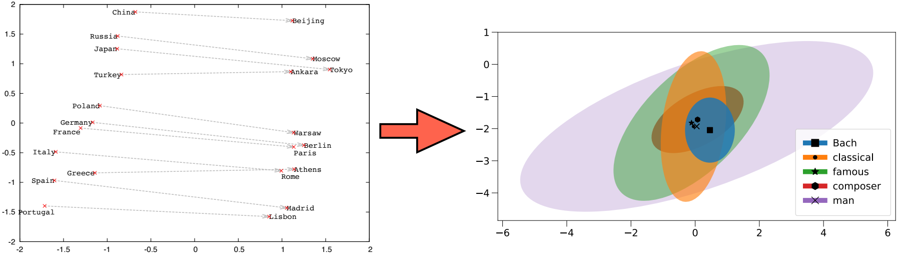
- $\nabla_{\mathbf{L}_{\mathbf{A}}}\frac{1}{2}\mathfrak{B}^2(\mathbf{L}_{\mathbf{A}}\mathbf{L}_{\mathbf{A}}{}^T,\mathbf{B}) = (\mathbf{I}_\mathrm{d} - \mathbf{T}^{\mathbf{AB}})\mathbf{L}_{\mathbf{A}}$
- Riem. geodesics: $\mathbf{C}_{\mathbf{AB}}(t) = [(1-t)\mathbf{I}_\mathrm{d} + t\mathbf{T}^{\mathbf{AB}}]\mathbf{A}[(1-t)\mathbf{I}_\mathrm{d} + t\mathbf{T}^{\mathbf{AB}}]$
- "$\Pi(\cdot)$ makes $\mathfrak{B}^2$ flat" : $\mathbf{L}_{\mathbf{A}} - t\nabla_{\mathbf{L}_{\mathbf{A}}}\frac{1}{2}\mathfrak{B}^2(\mathbf{A},\mathbf{B}) \in \Pi^{-1}\{\mathbf{C}_{\mathbf{AB}}(t)\}$

[13] B. Muzellec & M. Cuturi. “Generalizing Point Embeddings using the Wasserstein Space of Elliptical Distributions”. *NeurIPS*. 2018.

Application: Learning Representations

Problem: find representations for objects x in some space $\mathcal{X}$ (e.g. words, graphs, high-dimensional vectors...)

- **Classic approach:** represent each x as a point $y \in \mathbb{R}^k$[14].
- ***Elliptical embeddings***[15]: represent each x as an ell. distr. α with params $\mathbf{a} \in \mathbb{R}^k$ and $\mathbf{A} \in \mathcal{S}_+^k$. Use Bures-Wasserstein geometry.



[14] Figure from T. Mikolov et al. "Distributed representations of words and phrases and their compositionality". *NeurIPS*. 2013.

[15] B. Muzellec & M. Cuturi. "Generalizing Point Embeddings using the Wasserstein Space of Elliptical Distributions". *NeurIPS*. 2018.

Elliptical Word Embeddings

Advantage: encodes uncertainty, or spread.

Training: $$\min \sum_{w} \sum_{c \in \mathrm{Pos}(w)} \left[M - [\mu_w : \nu_c] + \frac{1}{n} \sum_{c' \in \mathrm{Neg}(w)} [\mu_w : \nu_{c'}] \right]_+$$

ALL MODELS ARE WRONG BUT SOME ARE USEFUL

ALL MODELS ARE WRONG BUT SOME ARE USEFUL

ALL MODELS ARE WRONG BUT SOME ARE USEFUL

Polarization: $[\alpha : \beta] \overset{\mathrm{def}}{=} \langle \mathbf{a}, \mathbf{b} \rangle + \mathrm{Tr}(\mathbf{A}^{1/2}\mathbf{B}\mathbf{A}^{1/2})^{1/2}$

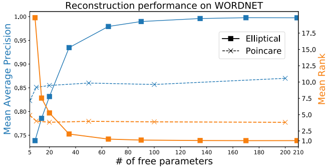
Datasets

`ukWaC` + `WaCkypedia`: 3 billion tokens, 250K unique[16]
`WordNet`: DAG, 80K unique nouns, 740K relationships[17]

Implementation: `cupy` (GPU) + `cython`, on GitHub.

Similarity Benchmark: Spearman Rank Correlation

Dataset	W2G/45/C	Ell/12/CM
SimLex	**25.09**	24.09
WordSim	53.45	**66.02**
WordSim-R	61.70	**71.07**
WordSim-S	48.99	**60.58**
MEN	65.16	**65.58**
MC	59.48	**65.95**
RG	**69.77**	65.58
YP	**37.18**	25.14
MT-287	**61.72**	59.53
MT-771	**57.63**	56.78
RW	**40.14**	29.04



[16] L. Vilnis et al. “Word representations via Gaussian embedding”. *ICLR* [2015].
[17] M. Nickel et al. “Poincaré Embeddings for Learning Hierarchical Representations”. *NeurIPS*. 2017.

Outline

Reminders: The Bures-Wasserstein Geometry on Elliptical Distributions

Learning with BW: Computing and Differentiating BW [NeurIPS'18]

Building OT Plans on Subspace Projections [NeurIPS'19]

Unbalanced Entropic OT for Gaussian Measures [NeurIPS'20]

Projected OT variants

Average on 1D projections[18]

$$\mathrm{SW}(\mu, \nu) \overset{\text{def}}{=} \mathbb{E}_{\theta \sim S^{d-1}}[\mathrm{OT}(\theta^\top{}_\sharp \mu, \theta^\top{}_\sharp \nu)].$$

Projection on adversarially-selected k-D subspace[19]

$$P_k^2(\mu, \nu) \overset{\text{def}}{=} \max_{E:\mathsf{dim}(E)=k} W_2(p_{E\sharp}\mu, p_{E\sharp}\nu).$$

Yields better computational and statistical properties than vanilla OT.

Lifting from subspace to the full space

Projecting first, transporting next is promising, but it restricts OT to happen in the projected space. *Can we recover a plan in the entire space?*

[18] J. Rabin et al. “Wasserstein barycenter and its application to texture mixing”. *SSVM*. 2011.

[19] F.-P. Paty et al. “Subspace Robust Wasserstein Distances”. *ICML*. 2019.

Subspace OT

Definition (Subspace-Optimal Plans, Muzellec & Cuturi, 2019)

- E a k-dimensional subspace of $\mathbb{R}^d$, projection operator p_E,
- $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, $\mu_E := p_{E\sharp}\mu$ (resp. $\nu_E := p_{E\sharp}\nu$),
- S a Monge map from μ_E to ν_E.

E-optimal plans: $\Pi_E(\mu, \nu) \overset{\text{def}}{=} \{\pi \in \Pi(\mu, \nu) : (p_E, p_E)_\sharp \pi = (id_E, S)_\sharp \mu_E\}$.

- Existence guaranteed by the "Gluing Lemma":

Gluing Lemma[20][21]

Given $(X_1, X_2) \sim (\mu_1, \mu_2)$ and $(Y_2, Y_2) \sim (\mu_2, \mu_3)$, there exists (Z_1, Z_2, Z_3) s.t. $(Z_1, Z_2) \sim (\mu_1, \mu_2)$ and $(Z_2, Z_3) \sim (\mu_2, \mu_3)$

What are the degrees of freedom in $\Pi_E(\mu, \nu)$?

[20] I. Berkes et al. “An almost sure invariance principle for the empirical distribution function of mixing random variables”. *Probability Theory and Rel. Fields* [1977].

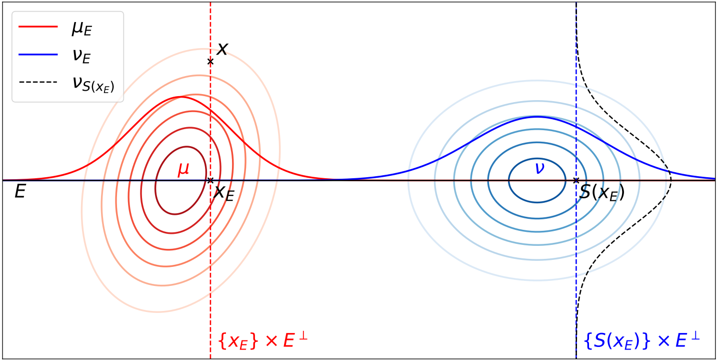
[21] C. Villani. *Optimal transport: old and new*. 2008.

Characterization of Subspace-Optimal Plans

Proposition (Muzellec & Cuturi, 2019)

Let $\pi \in \Pi_E(\mu, \nu)$*. Then (equivalently),*

- π *is supported on* $\mathcal{G}(S) \times E^\perp$*, where* $\mathcal{G}(S) \overset{\text{def}}{=} \{(x_E, S(x_E) : x_E \in E\}$*;*
- π *is characterized by its disintegration on* $\mathcal{G}(S)$*:* $\pi_{x_E, S(x_E)}, x_E \in E$*.*



Monge-Knothe Transport

Definition (Monge-Knothe Transport)

$\forall x_E \in E$, let $\hat{T}(x_E; \cdot) : E^\perp \to E^\perp$ denote the Monge map from μ_{x_E} to $\nu_{S(x_E)}$. The Monge-Knothe transportation map is defined as

$$\begin{aligned} T_{\mathsf{MK}} :& E \oplus E^\perp \to E \oplus E^\perp \\ & (x_E, x_{E^\perp}) \mapsto (S(x_E), \hat{T}(x_E; x_{E^\perp})). \end{aligned}$$

Proposition (Muzellec & Cuturi, 2019)

The Monge-Knothe plan is optimal in $\Pi_E(\mu, \nu)$, *namely*

$$\pi^{MK} \in \underset{\gamma \in \Pi_E(\mu,\nu)}{\arg\min}\, \mathbb{E}_{(X,Y)\sim\gamma}[\|X - Y\|^2].$$

MK transport for Gaussians

Proposition (Muzellec & Cuturi, 2019)

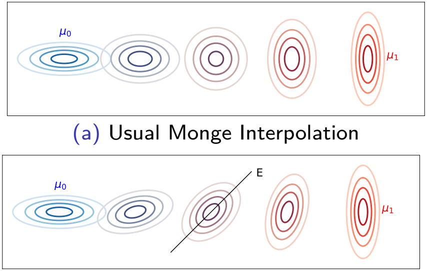
The MK transport map on E between $\alpha = \mathcal{N}(0_d, \mathbf{A})$ and $\beta = \mathcal{N}(0_d, \mathbf{B})$ is linear, and represented by the following matrix:

$$\mathbf{T}_{\mathrm{MK}} = \begin{pmatrix} \mathbf{T}^{\mathbf{A}_E\mathbf{B}_E} & 0_{k\times(d-k)} \\ \left[\mathbf{B}_{EE^\perp}^\top (\mathbf{T}^{\mathbf{A}_E\mathbf{B}_E})^{-1} - \mathbf{T}^{(\mathbf{A}/\mathbf{A}_E)(\mathbf{B}/\mathbf{B}_E)}\mathbf{A}_{EE^\perp}^\top\right](\mathbf{A}_E)^{-1} & \mathbf{T}^{(\mathbf{A}/\mathbf{A}_E)(\mathbf{B}/\mathbf{B}_E)} \end{pmatrix}.$$

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_E & \mathbf{A}_{EE^\perp} \\ \mathbf{A}_{EE^\perp}^T & \mathbf{A}_{E^\perp} \end{pmatrix}, \mathbf{B} = \begin{pmatrix} \mathbf{B}_E & \mathbf{B}_{EE^\perp} \\ \mathbf{B}_{EE^\perp}^T & \mathbf{B}_{E^\perp} \end{pmatrix}$$

Schur complements:

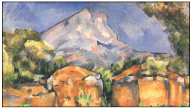
- $\mathbf{A}/\mathbf{A}_E \overset{\text{def}}{=} \mathbf{A}_{E^\perp} - \mathbf{A}_{EE^\perp}^T \mathbf{A}_E^{-1} \mathbf{A}_{EE^\perp}$
- Prop: $\mathbf{A}/\mathbf{A}_E = \mathrm{Cov}_{X\sim\alpha}(X_{E^\perp} | X_E)$

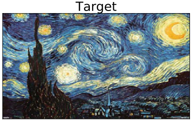


(a) Usual Monge Interpolation

(b) MK Interpolation through E

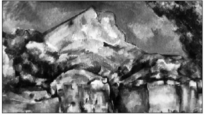
Application: Color transfer


Source


Target

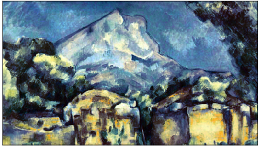
- KMeans quantization in 3D RGB space, 3000 clusters.
- 1D OT in gray space: $O(n \log n)$ time.

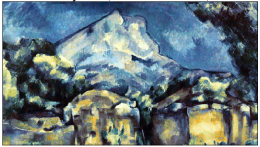
Gray Source

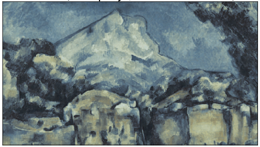

Gray Transfer

Gray Target

- Extend with MK map (small 3D problems): $\sim \times 50$ speedup.


Full OT (runtime 2.83s)


Gray MK (runtime 0.041s)


Sliced (100 projs, runtime 0.069s)

Outline

Reminders: The Bures-Wasserstein Geometry on Elliptical Distributions

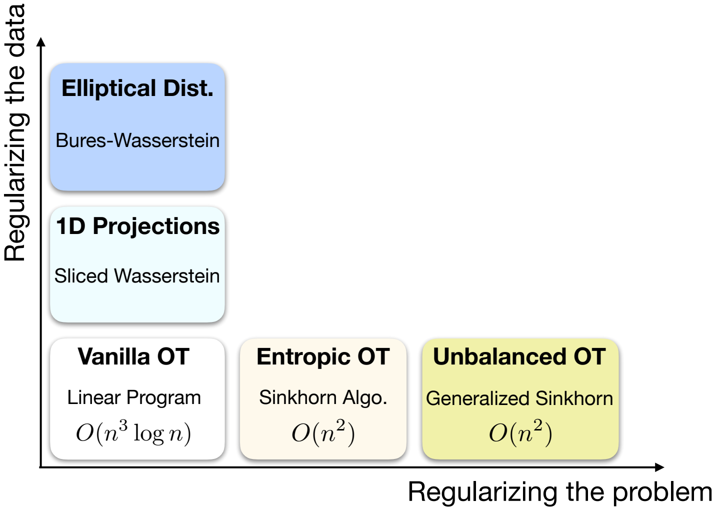
Learning with BW: Computing and Differentiating BW [NeurIPS'18]

Building OT Plans on Subspace Projections [NeurIPS'19]

Unbalanced Entropic OT for Gaussian Measures [NeurIPS'20]

Regularizing OT: A Recap

Real data comes in a discrete form: $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}$. What to do with it?



Regularizing the data

Elliptical Dist.

Bures-Wasserstein

1D Projections

Sliced Wasserstein

Vanilla OT

Linear Program

$O(n^3 \log n)$

Entropic OT

Sinkhorn Algo.

$O(n^2)$

Unbalanced OT

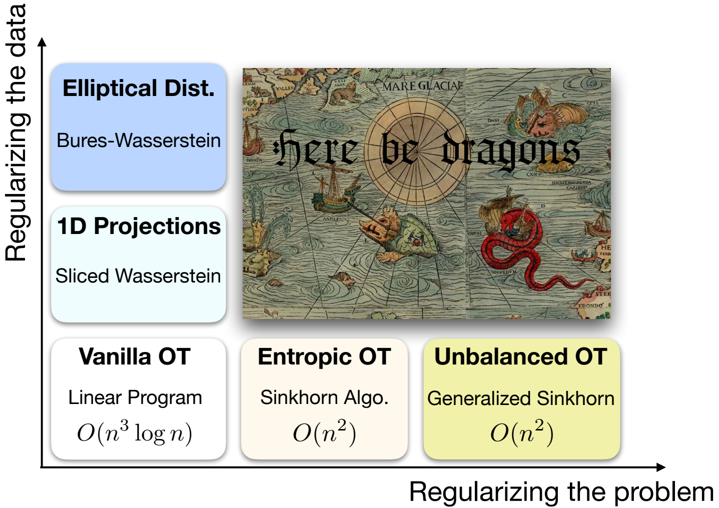
Generalized Sinkhorn

$O(n^2)$

Regularizing the problem

Regularizing OT: A Recap

Real data comes in a discrete form: $\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}$. What to do with it?



Regularizing the data

Elliptical Dist.

Bures-Wasserstein

1D Projections

Sliced Wasserstein

Vanilla OT

Linear Program

$O(n^3 \log n)$

Entropic OT

Sinkhorn Algo.

$O(n^2)$

Unbalanced OT

Generalized Sinkhorn

$O(n^2)$

Regularizing the problem

Entropy-regularized OT: Reminders

Primal Problem

$$\mathrm{OT}_\sigma(\mu,\nu) \overset{\text{def}}{=} \min_{\pi\in\Pi(\mu,\nu)} \iint_{\mathbb{R}^d\times\mathbb{R}^d} \|x-y\|^2 \mathrm{d}\pi(x,y) + 2\sigma^2\, \mathrm{KL}(\pi\|\mu\otimes\nu).$$

Dual Problem

$$\mathrm{OT}_\sigma(\mu,\nu) = \max_{\substack{f\in\mathcal{L}_2(\mu)\\ g\in\mathcal{L}_2(\nu)}} \mathbb{E}_\mu(f) + \mathbb{E}_\nu(g) - 2\sigma^2\Big(\iint_{\mathbb{R}^d\times\mathbb{R}^d} e^{\frac{f(x)+g(y)-\|x-y\|^2}{2\sigma^2}} \mathrm{d}\mu(x)\mathrm{d}\nu(y) - 1\Big).$$

Sinkorn iterations

$$g_{n+1}(y) = -2\sigma^2 \log \int_{\mathbb{R}^d} e^{\frac{-\|x-y\|^2+f_n(x)}{2\sigma^2}} \mathrm{d}\mu(x),$$

$$f_{n+1}(x) = -2\sigma^2 \log \int_{\mathbb{R}^d} e^{\frac{-\|x-y\|^2+g_{n+1}(y)}{2\sigma^2}} \mathrm{d}\nu(y).$$

Primal-Dual Relationship

$$\frac{\mathrm{d}\pi^\star}{\mathrm{d}\mu\mathrm{d}\nu}(x,y) = e^{\frac{f^\star(x)+g^\star(y)-\|x-y\|^2}{2\sigma^2}}$$

$$(f_n, g_n) \xrightarrow[n\to+\infty]{} (f^\star, g^\star)$$

[21] M. Cuturi. "Sinkhorn distances: Lightspeed computation of OT". *NeurIPS*. 2013.

Related Work

- Prior work in economics and control theory [BG16][22]& [CGP16][23]: closed forms for Gaussian Ent-OT.
- Subsequent work [MGM20][24]& [BL20][25]: closed forms for Gaussian Ent-OT & Sinkhorn barycenters restricted to Gaussian measures.

Novel contributions (Janati, Muzellec, et al. 2020)

1. Properties of Entropic Bures: convexity, gradients, minimizers;
2. Sinkhorn barycenters restricted to *sub*-Gaussian measures;
3. Closed forms for *unbalanced* Ent-OT with Gaussians.

[22] R. Bojilov et al. “Matching in closed-form: equilibrium, identification, and comparative statics”. *Economic Theory* [2016].

[23] Y. Chen, T. T. Georgiou, & M. Pavon. “On the relation between optimal transport and Schrödinger bridges: A stochastic control viewpoint”. *Jour. of Optim. Th. and App.* [2016].

[24] A. Mallasto et al. “Entropy-Regularized 2-Wasserstein Distance between Gaussian Measures”. *arXiv preprint* [2020].

[25] E. del Barrio et al. “The statistical effect of entropic regularization in optimal transportation”. *arXiv preprint* [2020].

Entropic OT for Gaussians

Theorem (Janati, Muzellec, et al. 2020)

Let $\mathbf{A}, \mathbf{B} \in S_{++}^d$, $\alpha \sim \mathcal{N}(\mathbf{a}, \mathbf{A})$ *and* $\beta \sim \mathcal{N}(\mathbf{b}, \mathbf{B})$.
Let $\mathbf{D}_\sigma = (4\mathbf{A}^{1/2}\mathbf{B}\mathbf{A}^{1/2} + \sigma^4 \mathbf{I}_\mathrm{d})^{1/2}$. *Then,*

$$\mathrm{OT}_\sigma(\alpha, \beta) = \|\mathbf{a} - \mathbf{b}\|^2 + \mathfrak{B}_\sigma^2(\mathbf{A}, \mathbf{B}), \text{ where}$$

$$\mathfrak{B}_\sigma^2(\mathbf{A}, \mathbf{B}) = \mathrm{Tr}(\mathbf{A} + \mathbf{B} - \mathbf{D}_\sigma) + \sigma^2 \log\det\left(\mathbf{D}_\sigma + \sigma^2 \mathbf{I}_\mathrm{d}\right) + d\sigma^2(1 - \log(2\sigma^2)).$$

Moreover, the entropic optimal transportation plan is also a Gaussian over $\mathbb{R}^d \times \mathbb{R}^d$: $\pi^\star = \mathcal{N}\left(\left(\begin{smallmatrix} \mathbf{a} \\ \mathbf{b} \end{smallmatrix}\right), \left(\begin{smallmatrix} \mathbf{A} & \mathbf{C}_\sigma \\ \mathbf{C}_\sigma^\top & \mathbf{B} \end{smallmatrix}\right)\right)$, *with* $\mathbf{C}_\sigma = \frac{1}{2}\mathbf{A}^{1/2}\mathbf{D}_\sigma\mathbf{A}^{-1/2} - \frac{\sigma^2}{2}\mathbf{I}_\mathrm{d}$

Proposition (Janati, Muzellec, et al. 2020)

- $\nabla_{\mathbf{A}} \mathfrak{B}_\sigma^2(\mathbf{A}, \mathbf{B}) = \mathbf{I}_\mathrm{d} - \mathbf{B}^{1/2}\left((\mathbf{B}^{1/2}\mathbf{A}\mathbf{B}^{1/2} + \frac{\sigma^4}{4}\mathbf{I}_\mathrm{d})^{1/2} + \frac{\sigma^2}{2}\mathbf{I}_\mathrm{d}\right)^{-1}\mathbf{B}^{1/2}$,
- $\mathfrak{B}_\sigma^2(\mathbf{A}, \mathbf{B})$ *is convex in* $\mathbf{A}$ *and in* $\mathbf{B}$, *but not jointly.*

Some elements of the proof

Main points

- The optimal potentials $(f^\star, g^\star)$ contain all the information;
- Sinkhorn iterations preserve quadratic forms $\mathcal{Q}(\mathbf{H}) \overset{\text{def}}{=} x \mapsto x^T \mathbf{H} x$.

Lemma (Janati, Muzellec, et al. 2020)

- *Let* $\mathbf{U}_0 = \mathbf{V}_0 = 0$, $f_0 = \mathcal{Q}(\mathbf{U}_0), g_0 = \mathcal{Q}(\mathbf{V}_0)$
- *Then* $\forall n \geq 0, \frac{f_n}{2\sigma^2} = \mathcal{Q}(\mathbf{U}_n)$ *and* $\frac{g_n}{2\sigma^2} = \mathcal{Q}(\mathbf{V}_n)$, *with*

$$\mathbf{V}_{n+1} = \frac{1}{\sigma^2}((\sigma^2 \mathbf{U}_n + \sigma^2 \mathbf{A}^{-1} + \mathbf{I}_{\mathrm{d}})^{-1} - \mathbf{I}_{\mathrm{d}}),$$
$$\mathbf{U}_{n+1} = \frac{1}{\sigma^2}((\sigma^2 \mathbf{V}_{n+1} + \sigma^2 \mathbf{B}^{-1} + \mathbf{I}_{\mathrm{d}})^{-1} - \mathbf{I}_{\mathrm{d}}).$$

After change of variables: $\mathbf{G}_{n+1} = \sigma^2 \mathbf{B}^{-1} + \mathbf{F}_n^{-1}, \mathbf{F}_{n+1} = \sigma^2 \mathbf{A}^{-1} + \mathbf{G}_{n+1}^{-1}$.

Leads to fixed-point equation $\mathbf{C}_\sigma^2 + \sigma^2 \mathbf{C}_\sigma - \mathbf{A}\mathbf{B} = 0$.

Sinkhorn Debiased Barycenters for Gaussians

Sinkhorn Divergences[26]

$$S_\sigma \overset{\text{def}}{=} (\mu, \nu) \mapsto \mathrm{OT}_\sigma(\mu, \nu) - \tfrac{1}{2}(\mathrm{OT}_\sigma(\mu, \mu) + \mathrm{OT}_\sigma(\nu, \nu))$$

Theorem (Janati, Muzellec, et al. 2020)

- $\mathcal{G} \overset{\text{def}}{=} \{\mu \in \mathcal{P}_2 | \exists q > 0, \mathbb{E}_\mu(e^{q\|X\|^2}) < +\infty\}$ *(sub-Gaussian measures).*
- $\alpha_k \sim \mathcal{N}(\mathbf{a}_k, \mathbf{A}_k), w_k \in \mathbb{R}_+, k = 1, ..., K,$ *s.t.* $\sum_{k=1}^K w_k = 1.$

The debiased barycenter $\beta^{(bar)} \overset{\text{def}}{=} \mathrm{argmin}_{\beta \in \mathcal{G}} \sum_{k=1}^K w_k S_\sigma(\alpha_k, \beta)$

is given by $\beta^{(bar)} = \mathcal{N}(\sum_{k=1}^K w_k \mathbf{a}_k, \mathbf{B})$, *where* $\mathbf{B} \in \mathcal{S}_+^d$ *is a solution of*

$$\sum_{k=1}^K w_k(\mathbf{B}^{1/2}\mathbf{A}_k\mathbf{B}^{1/2} + \tfrac{\sigma^4}{4}\mathbf{I}_{\mathrm{d}})^{1/2} = (\mathbf{B}^2 + \tfrac{\sigma^4}{4}\mathbf{I}_{\mathrm{d}})^{1/2}.$$

Generalizes the Bures barycenter equation[27].

[26] A. Genevay, G. Peyre, et al. "Learning Generative Models with Sinkhorn Divergences". *AISTATS*. 2018.

[27] M. Agueh et al. "Barycenters in the Wasserstein space". *SIAM* [2011].

Entropy-regularized Unbalanced OT

Remove the $\pi \in \Pi(\mu, \nu)$ constraints, replace them with a KL penalty.

Unblanced Ent-OT[28][29]

$$\mathrm{UOT}_{\sigma,\gamma}(\mu,\nu) \overset{\text{def}}{=} \inf_{\pi \in \mathcal{M}_2^+} \Big\{ \iint_{\mathbb{R}^d \times \mathbb{R}^d} \|x-y\|^2 \mathrm{d}\pi(x,y) + 2\sigma^2 \,\mathrm{KL}(\pi \| \mu \otimes \nu) + \gamma \,\mathrm{KL}(\pi_1 \| \mu) + \gamma \,\mathrm{KL}(\pi_2 \| \nu) \Big\}. \qquad (1)$$

Proposition

Assume π^* is a solution of (1). *Then*

$$\mathrm{UOT}_{\sigma,\gamma}(\mu,\nu) = \gamma(m_\mu + m_\nu) + 2\sigma^2 m_\mu m_\nu - 2(\sigma^2+\gamma)\pi^*(\mathbb{R}^d \times \mathbb{R}^d). \qquad (2)$$

[28] C. Frogner et al. "Learning with a Wasserstein Loss". *NeurIPS*. 2015.

[29] L. Chizat. "Unbalanced optimal transport: Models, numerical methods, applications". PhD thesis. 2017.

Entropic Unbalanced OT for Gaussians

Theorem (Janati, Muzellec, et al. 2020)

Let $\alpha = m_\alpha \mathcal{N}(\mathbf{a}, \mathbf{A})$ *and* $\beta = m_\beta \mathcal{N}(\mathbf{b}, \mathbf{B})$ *be two unnormalized Gaussian measures. Then*

- *The unbalanced optimal transport plan, minimizer of* (1), *is an unnormalized Gaussian over* $\mathbb{R}^d \times \mathbb{R}^d$: $\pi^\star = m_{\pi^\star} \mathcal{N}(\mathbf{m}, \mathbf{H})$,
- $\mathrm{UOT}_{\sigma,\gamma}$ *can be derived using* (2) *with* $\pi^\star(\mathbb{R}^d \times \mathbb{R}^d) = m_{\pi^\star}$.

- $\mathbf{m}, \mathbf{H}$ and m_π are in closed form.

Remark

Contrary to balanced (entropic) OT, we cannot consider the centered problem without loss of generality!

Some elements of the proof

Unbalanced Ent-OT: Dual Problem[30]

$$\sup_{\substack{f\in\mathcal{L}_2(\mu)\\ g\in\mathcal{L}_2(\nu)}} \left\{\gamma \int_{\mathbb{R}^d}(1-e^{-\frac{f}{\gamma}})\mathrm{d}\mu + \gamma \int_{\mathbb{R}^d}(1-e^{-\frac{g}{\gamma}})\mathrm{d}\nu - 2\sigma^2 \iint_{\mathbb{R}^d\times\mathbb{R}^d}(e^{\frac{-\|x-y\|^2+f(x)+g(y)}{2\sigma^2}} - 1)\mathrm{d}\mu(x)\mathrm{d}\nu(y)\right\}.$$

Generalized Sinkorn iterations[30]

$$g_{n+1}(y) = -\tau \log \int_{\mathbb{R}^d} e^{\frac{-\|x-y\|^2+f_n(x)}{2\sigma^2}}\mathrm{d}\mu(x),$$

$$f_{n+1}(x) = -\tau \log \int_{\mathbb{R}^d} e^{\frac{-\|x-y\|^2+g_{n+1}(y)}{2\sigma^2}}\mathrm{d}\nu(y).$$

Primal-Dual Relationship

$$\frac{\mathrm{d}\pi^\star}{\mathrm{d}\mu\mathrm{d}\nu}(x,y) = e^{\frac{f^\star(x)+g^\star(y)-\|x-y\|^2}{2\sigma^2}}$$

$$\tau \overset{\mathrm{def}}{=} \frac{\gamma}{\gamma+2\sigma^2}$$

Stable parameterization (Janati, Muzellec, et al. 2020)

$$\frac{f_n(\mathbf{x})}{2\sigma^2} = -\frac{1}{2}(x^\top \mathbf{U}_n\mathbf{x} - 2x^\top \mathbf{u}_n) + \log(m_{un}), \quad \frac{g_n(x)}{2\sigma^2} = -\frac{1}{2}(x^\top \mathbf{V}_n\mathbf{x} - 2x^\top \mathbf{v}_n) + \log(m_{vn}).$$

Solve for $\mathbf{U}^\star, \mathbf{V}^\star, \mathbf{u}^\star, \mathbf{v}^\star, m_u{}^\star$ and $m_v{}^\star$ (fixed-point equations).

[30] L. Chizat. "Unbalanced optimal transport: Models, numerical methods, applications". PhD thesis. 2017.

Numerical Experiments

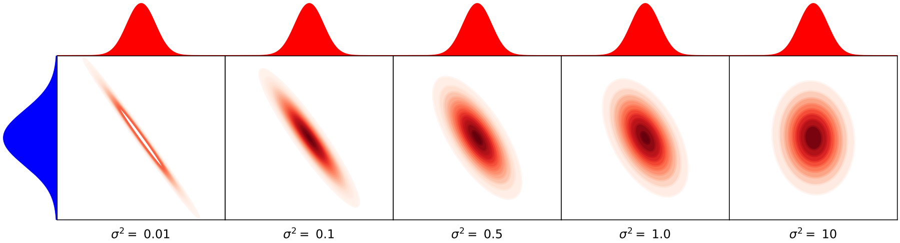


Figure 4: Effect of regularization on balanced transportation plans.

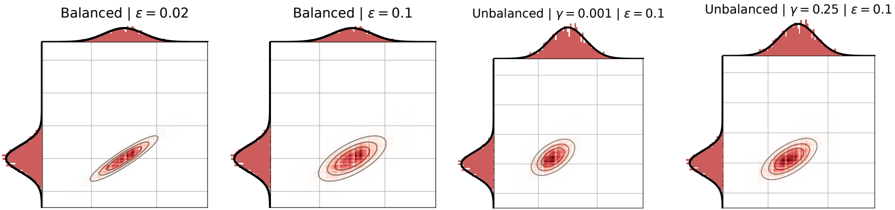


Figure 5: Effect of σ in balanced OT and γ in unbalanced OT. Empirical plans (red) correspond to the expected Gaussian contours depicted in black.

Take-home messages

A common approach is to regularize OT problems. But regularizing *data* to fall back to closed form solutions of OT is also a powerful approach in ML.

The Bures-Wasserstein geometry has all the tools and properties to scale to large ML gradient-based applications.

Maps and plans defined on the full space can be extracted after projecting distributions to lower dimension, in closed forms for Gaussians.

The Bures-Wasserstein geometry can seamlessly incorporate entropic and unbalanced regularization, in closed form.

What I did not talk about

Missing data imputation with entropic OT (Chapter 5)

References I

Agueh, M. & G. Carlier. “Barycenters in the Wasserstein space”. *SIAM* (2011).

Arjovsky, M., S. Chintala, & L. Bottou. “Wasserstein Generative Adversarial Networks”. *ICML*. 2017.

Barrio, E. del & J.-M. Loubes. “The statistical effect of entropic regularization in optimal transportation”. *arXiv preprint* (2020).

Benamou, J.-D. & Y. Brenier. “A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem”. *Numerische Mathematik* (2000).

Berkes, I. & W. Philipp. “An almost sure invariance principle for the empirical distribution function of mixing random variables”. *Probability Theory and Rel. Fields* (1977).

Bhatia, R., T. Jain, & Y. Lim. “On the Bures-Wasserstein distance between positive definite matrices”. *Expositiones Mathematicae* (2018).

References II

Blondel, M., V. Seguy, & A. Rolet. "Smooth and sparse optimal transport". *AISTATS*. 2018.

Bojilov, R. & A. Galichon. "Matching in closed-form: equilibrium, identification, and comparative statics". *Economic Theory* (2016).

Bonneel, N. et al. "Sliced and Radon Wasserstein Barycenters of Measures". *Journal of Mathematical Imaging and Vision* (2015).

Brenier, Y. "Décomposition polaire et réarrangement monotone des champs de vecteurs". *CR Acad. Sci. Paris Sér. I Math.* (1987).

Bures, D. "An extension of Kakutani's theorem on infinite product measures to the tensor product of semifinite w*-algebras". *Trans. of the Am. Math. Soc.* (1969).

Chen, Y., T. T. Georgiou, & A. Tannenbaum. "Optimal Transport for Gaussian Mixture Models". *IEEE Access* (2019).

References III

- Chen, Y., T. T. Georgiou, & M. Pavon. "On the relation between optimal transport and Schrödinger bridges: A stochastic control viewpoint". *Jour. of Optim. Th. and App.* (2016).
- Chizat, L. "Unbalanced optimal transport: Models, numerical methods, applications". PhD thesis. 2017.
- Chizat, L. et al. "Scaling algorithms for unbalanced optimal transport problems". *Mathematics of Computation* (2018).
- Cuturi, M. "Sinkhorn distances: Lightspeed computation of OT". *NeurIPS*. 2013.
- Dessein, A., N. Papadakis, & J.-L. Rouas. "Regularized optimal transport and the rot mover's distance". *The Journal of Machine Learning Research* 19.1 (2018), pp. 590–642.
- Dowson, D. & B. Landau. "The Fréchet distance between multivariate normal distributions". *Journal of multivariate analysis* (1982).

References IV

- Frogner, C. et al. “Learning with a Wasserstein Loss”. *NeurIPS*. 2015.
- Gelbrich, M. “On a formula for the L2 Wasserstein metric between measures on Euclidean and Hilbert spaces”. *Mathematische Nachrichten* (1990).
- Genevay, A., M. Cuturi, et al. “Stochastic optimization for large-scale OT”. *NeurIPS*. 2016.
- Genevay, A., G. Peyre, & M. Cuturi. “Learning Generative Models with Sinkhorn Divergences”. *AISTATS*. 2018.
- Heusel, M. et al. “GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium”. *NeurIPS*. 2017.
- Higham, N. J. *Functions of Matrices: Theory and Computation*. SIAM, 2008.
- Janati, H. et al. “Entropic Optimal Transport between (Unbalanced) Gaussian Measures has a Closed Form”. *NeurIPS* (2020).

References V

- Kantorovich, L. V. "On the translocation of masses". *Dokl. Akad. Nauk. USSR*. 1942.
- Kolouri, S. et al. "Generalized sliced Wasserstein distances". *NeurIPS*. 2019.
- Makkuva, A. V. et al. "Optimal transport mapping via input convex neural networks". *ICML* (2020).
- Mallasto, A., A. Gerolin, & H. Q. Minh. "Entropy-Regularized 2-Wasserstein Distance between Gaussian Measures". *arXiv preprint* (2020).
- Mérigot, Q. "A multiscale approach to optimal transport". *Comp. Grap. Forum*. 2011.
- Mikolov, T. et al. "Distributed representations of words and phrases and their compositionality". *NeurIPS*. 2013.
- Monge, G. "Mémoire sur la théorie des déblais et des remblais". *Histoire de l'Académie Royale des Sciences de Paris* (1781).

References VI

Muzellec, B. & M. Cuturi. “Generalizing Point Embeddings using the Wasserstein Space of Elliptical Distributions”. *NeurIPS*. 2018.

– .“Subspace detours: Building transport plans that are optimal on subspace projections”. *NeurIPS*. 2019.

Muzellec, B., J. Josse, et al. “Missing Data Imputation using Optimal Transport”. *ICML* (2020).

Muzellec, B., R. Nock, et al. “Tsallis regularized optimal transport and ecological inference”. *AAAI*. 2017.

Muzellec, B., K. Sato, et al. “Dimension-free convergence rates for gradient Langevin dynamics in RKHS”. *arXiv preprint* (2020).

Nickel, M. & D. Kiela. “Poincaré Embeddings for Learning Hierarchical Representations”. *NeurIPS*. 2017.

Olkin, I. & F. Pukelsheim. “The distance between two random vectors with given dispersion matrices”. *Linear Algebra and its Applications* (1982).

References VII

- Paty, F.-P. & M. Cuturi. “Subspace Robust Wasserstein Distances”. *ICML*. 2019.
- Rabin, J. et al. “Wasserstein barycenter and its application to texture mixing”. *SSVM*. 2011.
- Schiebinger, G. et al. “Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming”. *Cell* (2019).
- Seguy, V. et al. “Large-Scale Optimal Transport and Mapping Estimation”. *ICLR*. 2018.
- Solomon, J. et al. “Convolutional wasserstein distances: Efficient optimal transportation on geometric domains”. *TOG* (2015).
- Takatsu, A. “Wasserstein geometry of Gaussian measures”. *Osaka J. Math.* (2011).
- Titouan, V. et al. “Sliced Gromov-Wasserstein”. *NeurIPS*. 2019.
- Villani, C. *Optimal transport: old and new*. 2008.

References VIII

Vilnis, L. & A. McCallum. “Word representations via Gaussian embedding”. *ICLR* (2015).