Optimal Transport Metrics

Cambridge MLG Reading Group

Shreyas Padhy
09 February 2022

Computational and Biological Learning
DEPARTMENT OF ENGINEERING

UNIVERSITY OF CAMBRIDGE

Why Optimal Transport?

- The natural geometry for **probability measures** supported on a **metric space**
- **Shortest path principle**
  - OT generalises this: one item -> groups of items
- Borrows key geometric properties of underlying “ground” space on which distributions are defined
  - Euclidean metric -> interpolation, barycenters, etc -> Wasserstein space
- Provides a metric (or discrepancy measure) for probability measures with **non-overlapping support**

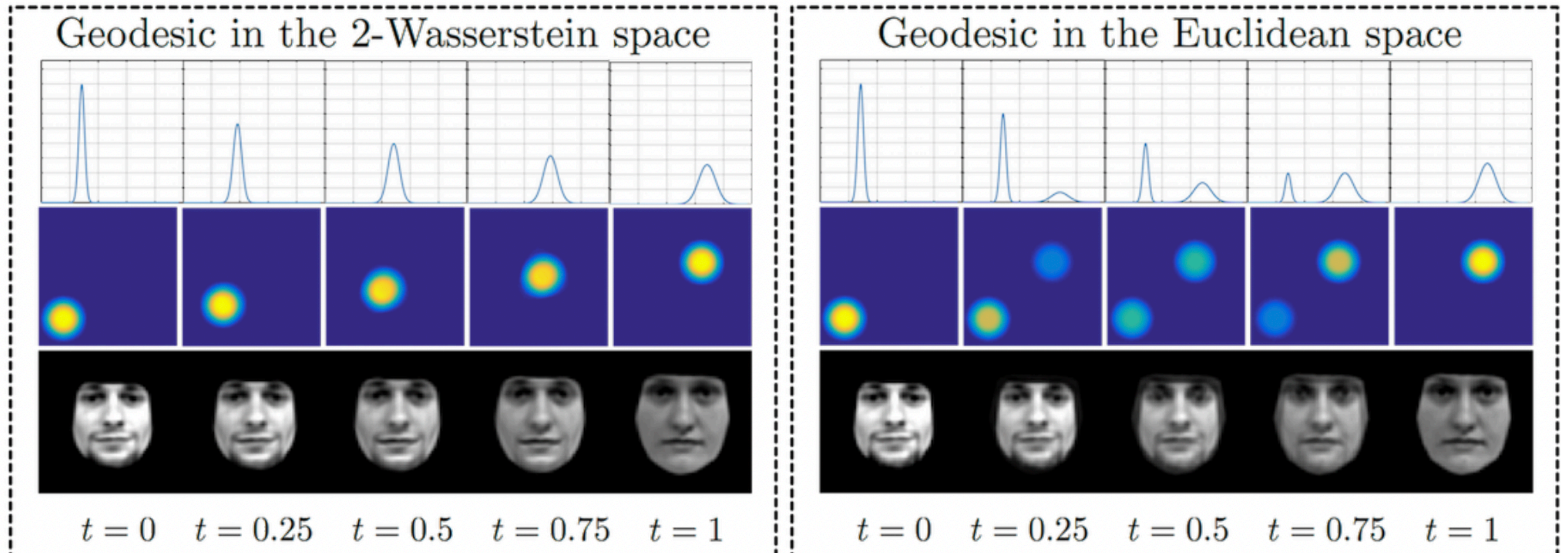


Image credit: [Kolouri et al. 2017]

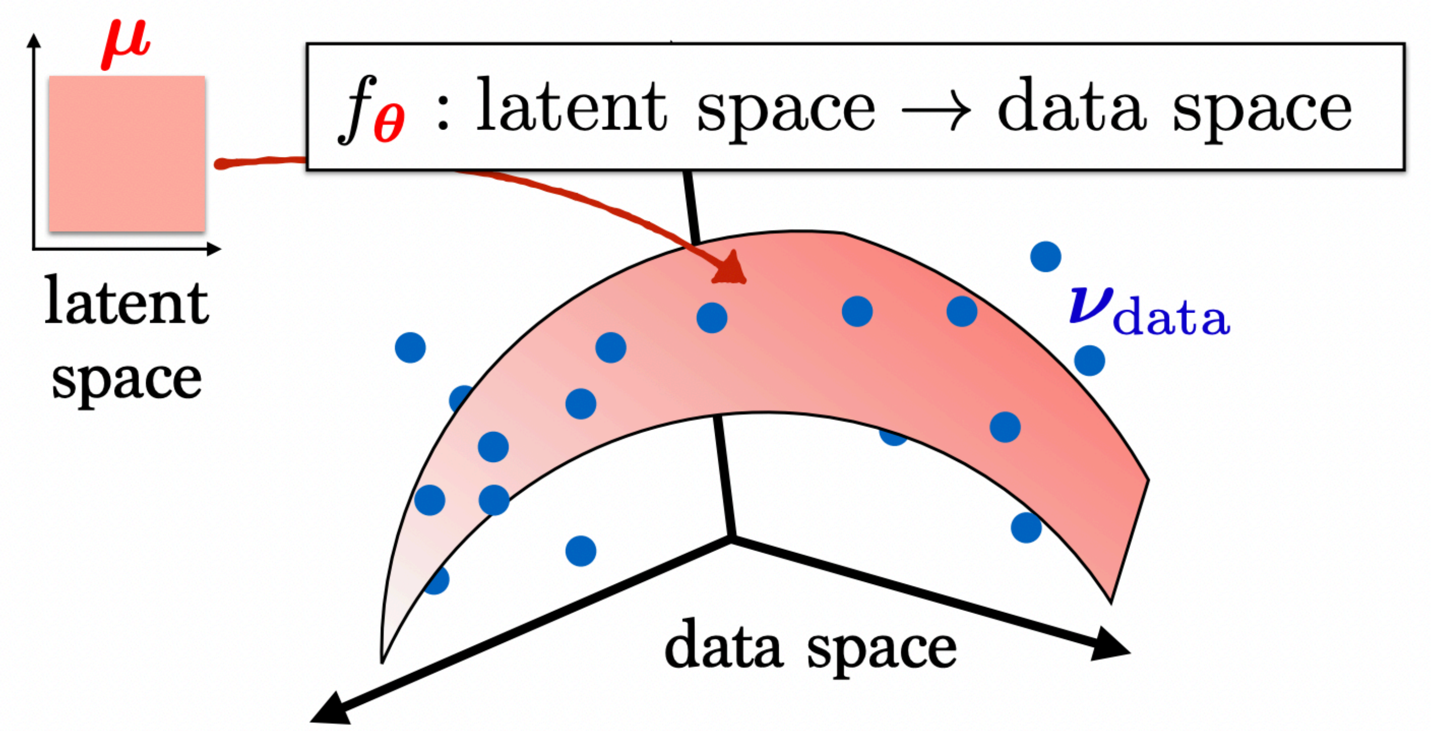


Image credit: Marco Cuturi

In this talk

I. Mathematical Formulation of Optimal Transport Theory

   Wasserstein Distances

   Computational and Statistical Issues

II. Approximate/Regularised OT

   Sliced Wasserstein Distances

   Sinkhorn Divergences

III. Applications of OT in Machine Learning

IV. Extensions of OT

   Unbalanced OT

   OT on separate metrics

Mathematical Preliminaries

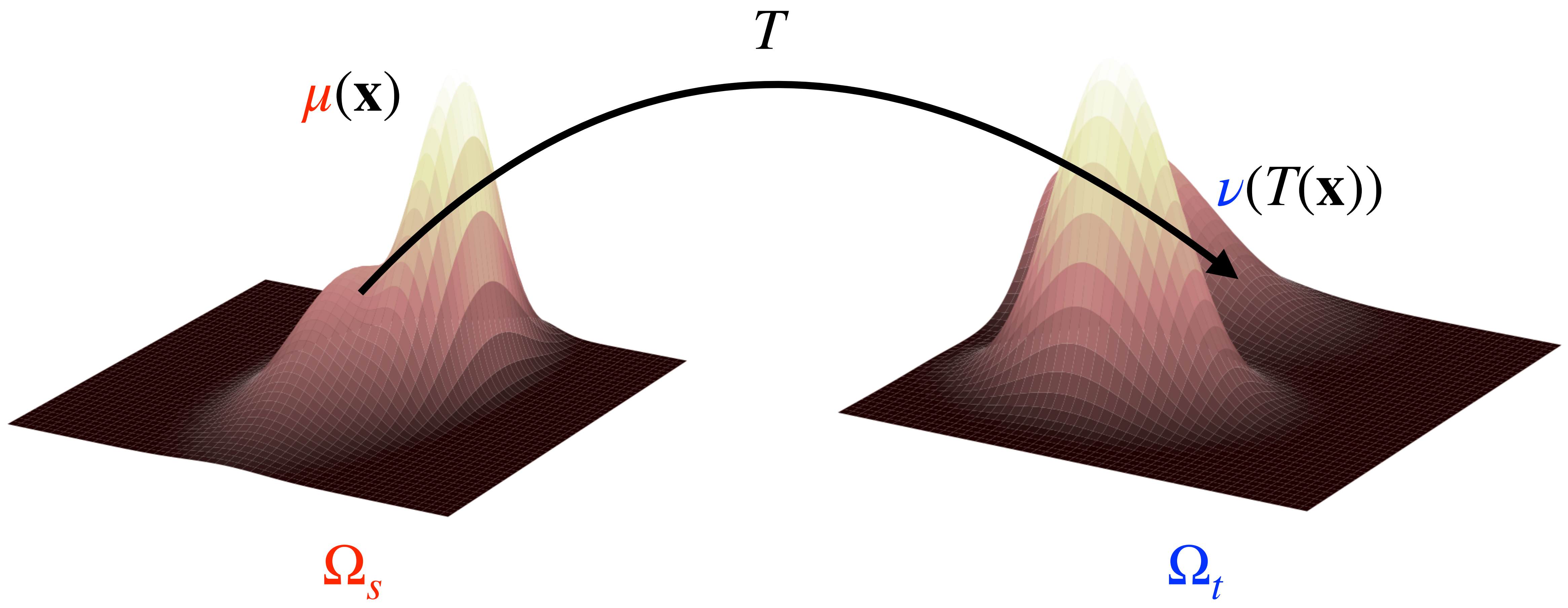
Monge Problem

- [Monge, 1781] How does one move one pile of dirt to another while minimising effort?
- Probability measures $\mu \in P(\Omega_s)$, $\nu \in P(\Omega_t)$, on metric spaces, and a cost function $c : \Omega_s \times \Omega_t \to \mathbb{R}^+$
- Push-forward operator $T\#$ transfers measures from one space Ω_s to another Ω_t

$$\nu(A) = \mu(T^{-1}(A)), \forall \text{Borel subsets } A \in \Omega_t \qquad \text{(conservation of mass)}$$

- The Monge formulation wishes to find a mapping $T : \Omega_s \to \Omega_t$ that minimises

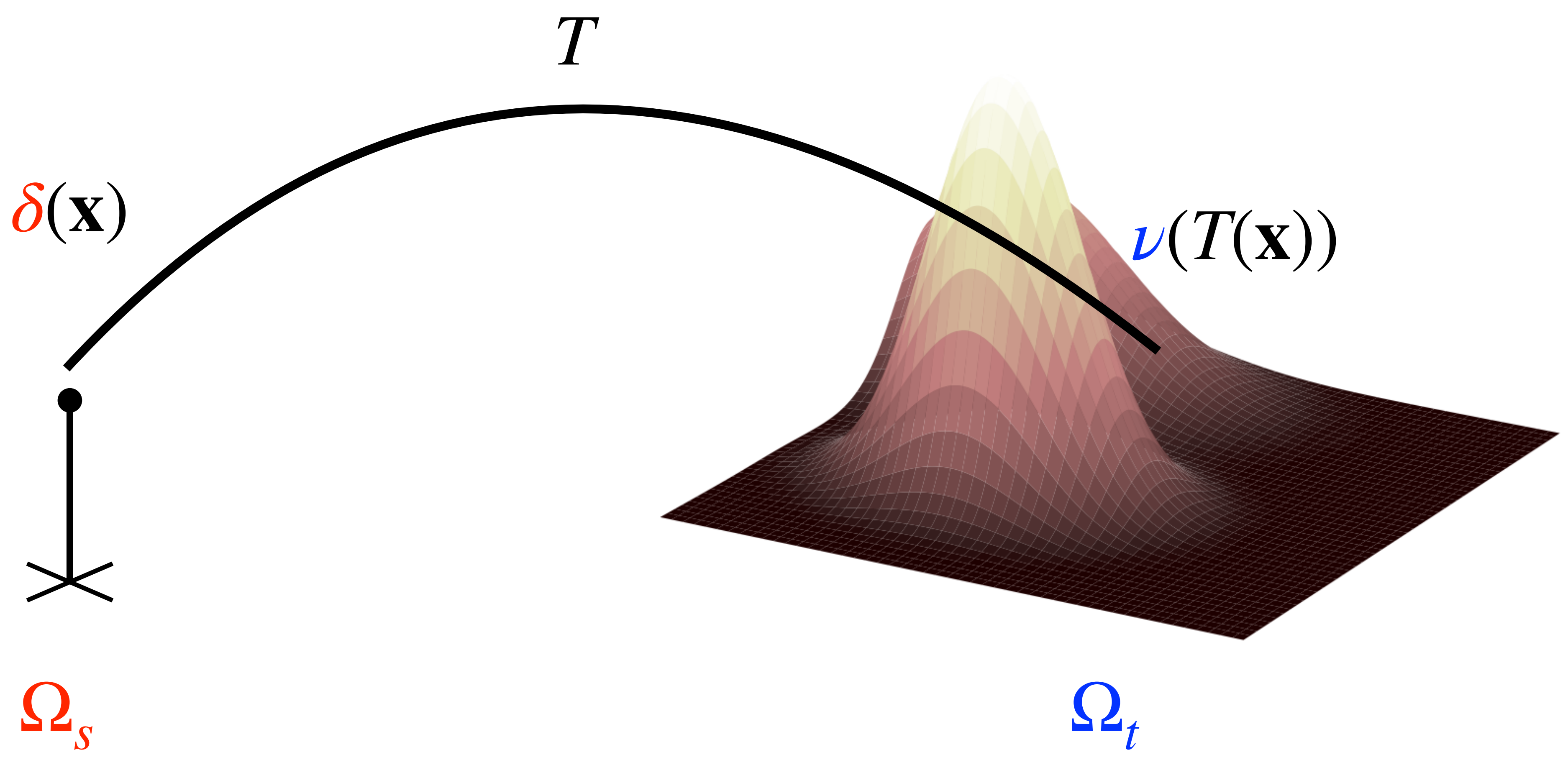
$$\inf_{T\#\mu=\nu} \int_{\Omega_s} c(\mathbf{x}, T(\mathbf{x}))\mu(\mathbf{x})d\mathbf{x}$$



Monge Problem - Issues

$$\inf_{T\#\mu=\nu} \int_{\Omega_s} c(\mathbf{x}, T(\mathbf{x}))\mu(\mathbf{x})d\mathbf{x}$$

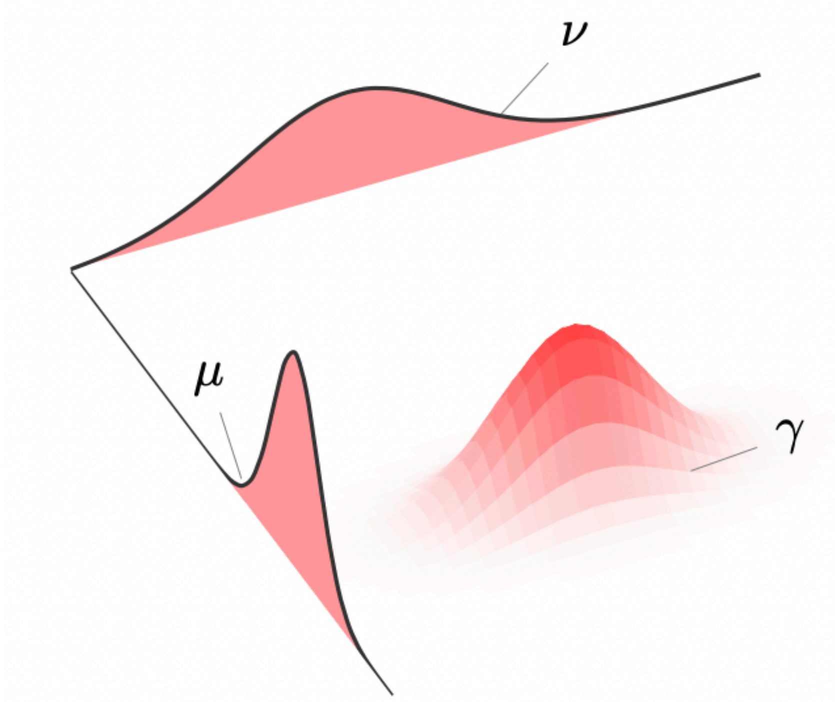
- $T\#\mu = \nu$ is not a convex constraint, *Existence* and *Unicity* of T is not guaranteed
- Can't split mass (one-to-one, but not one-to-many)
- Ex: Can't map Dirac measures δ_x to continuous measures



Kantorovich Relaxation

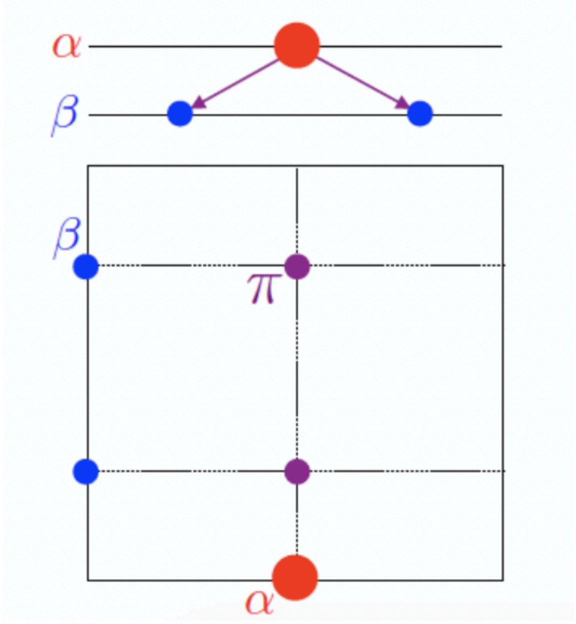
- [Kantorovich, 1942] Relax the requirement of maps T to *probabilistic couplings* $\gamma \in \mathscr{P}(\Omega_s \times \Omega_t)$

$$\gamma \in \mathscr{P} = \left\{ \gamma \geq \mathbf{0}, \int_{\Omega_t} \gamma(\mathbf{x}, \mathbf{y}) \mathrm{d}\mathbf{y} = \mu, \int_{\Omega_s} \gamma(\mathbf{x}, \mathbf{y}) \mathrm{d}\mathbf{x} = \nu \right\}$$



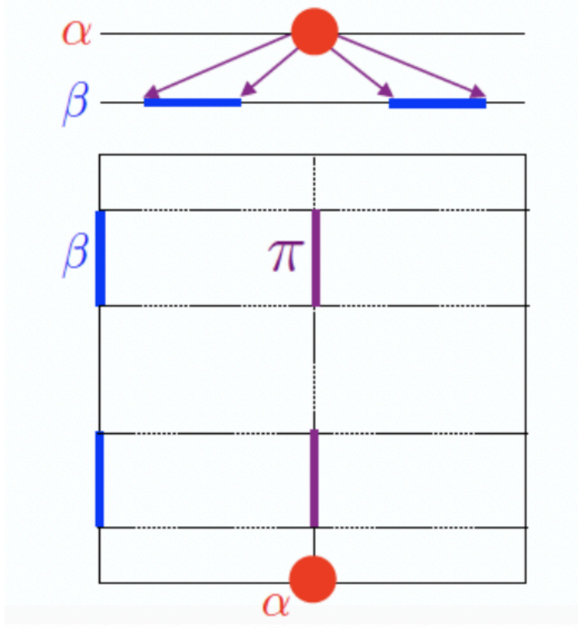
Product Coupling $\gamma = \mu \otimes \nu$

Image credit: Lenaic Chizat



Coupling for Dirac -> Dirac

Image credit: Remi Flamary



Coupling for Dirac -> Continuous

Image credit: Remi Flamary

- Given $\mu \in P(\Omega_s)$, $\nu \in P(\Omega_t)$, on metric spaces, a cost function $c : \Omega_s \times \Omega_t \to \mathbb{R}^+$, find couplings γ that minimise

$$\underset{\gamma}{\operatorname{argmin}} \int_{\Omega_s \times \Omega_t} c(\mathbf{x}, \mathbf{y}) \gamma(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y} \quad \text{s.t.}$$

$$\gamma \in \mathscr{P} = \left\{ \gamma \geq \mathbf{0}, \int_{\Omega_t} \gamma(\mathbf{x}, \mathbf{y}) \mathrm{d}\mathbf{y} = \mu, \int_{\Omega_s} \gamma(\mathbf{x}, \mathbf{y}) \mathrm{d}\mathbf{x} = \nu \right\}$$

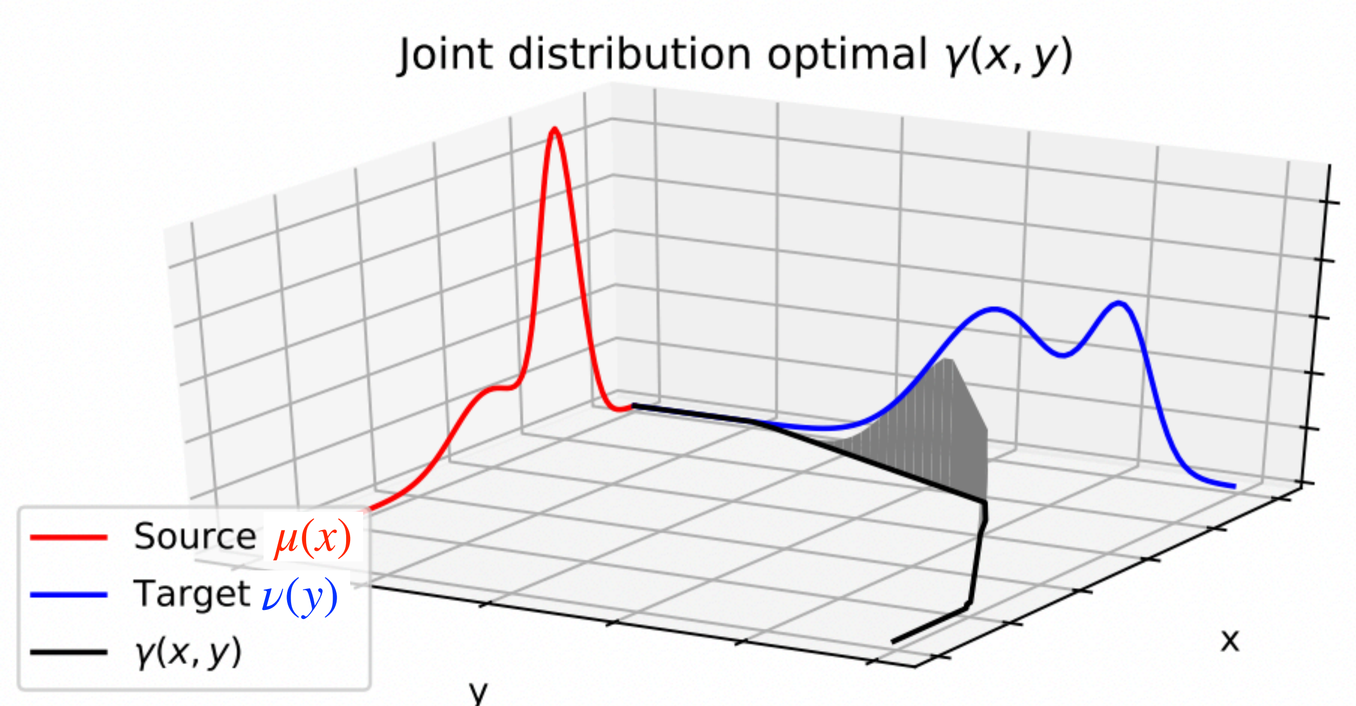
Kantorovich Dual Formulation

- Instead of optimising over all couplings γ that satisfy the constraints, consider two measurable functions $\phi \in L_1(\mu)$, $\psi \in L_1(\nu)$
  - Reminder: A fn $f : \mathscr{X} \to \mathscr{Y}$ is Lipschitz continuous if there exists a real constant $K \geq 0$ s.t

$$d_{\mathscr{Y}}(f(x_1), f(x_2)) \leq K d_{\mathscr{X}}(x_1, x_2)$$

$$\text{Solve} \qquad \max_{\phi,\psi} \left\{ \int \phi d\mu + \int \psi d\nu \quad \text{s.t} \quad \phi(\mathbf{x}) + \psi(\mathbf{y}) \leq c(\mathbf{x}, \mathbf{y}) \right\}$$

- The primal and dual formulations solve exactly the same problem at the equality
  - support of $\gamma(\mathbf{x}, \mathbf{y})$ is where $\phi(\mathbf{x}) + \psi(\mathbf{y}) = c(\mathbf{x}, \mathbf{y})$



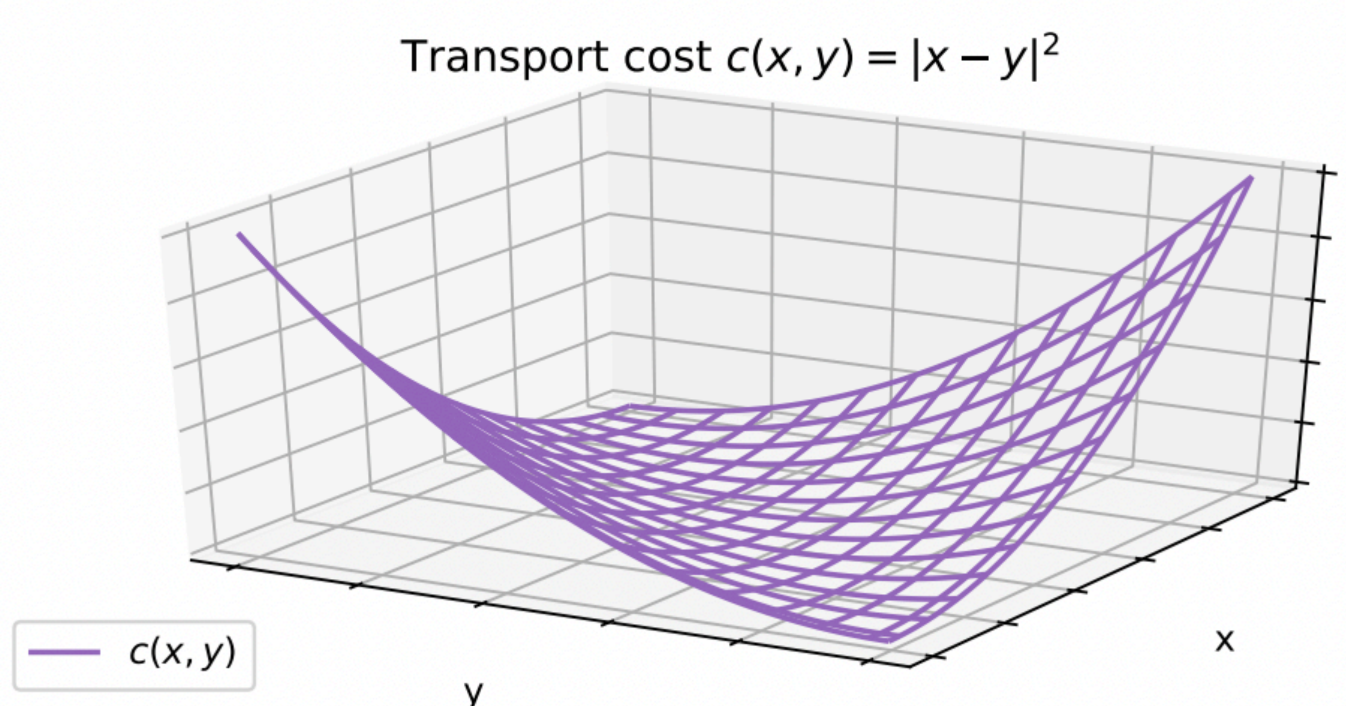


Image credit: Remi Flamary

Semi-dual formulation: c-Conjugates

- Instead of optimising over all possible ϕ, ψ given constraints, can we find the best ψ given a ϕ?
- Given a ϕ, we need that ψ satisfies for all $\mathbf{x}, \mathbf{y}$

$$\phi(\mathbf{x}) + \psi(\mathbf{y}) \leq c(\mathbf{x}, \mathbf{y})$$

$$\psi(\mathbf{y}) \leq c(\mathbf{x}, \mathbf{y}) - \phi(\mathbf{x})$$

$$\psi(\mathbf{y}) \leq \inf_{\mathbf{x}} c(\mathbf{x}, \mathbf{y}) - \phi(\mathbf{x})$$

$$\text{define } \phi^c(\mathbf{y}) = \inf_{\mathbf{x}} c(\mathbf{x}, \mathbf{y}) - \phi(\mathbf{x})$$

- Can simplify to a semi-dual formulation that depends on only one function ϕ through the c-conjugate

$$\max_{\phi, \psi} \left\{ \int \phi \, d\mu + \int \psi \, d\nu \quad \text{s.t} \quad \phi(\mathbf{x}) + \psi(\mathbf{y}) \leq c(\mathbf{x}, \mathbf{y}) \right\} \quad \Longrightarrow \quad \max_{\phi} \left\{ \int \phi \, d\mu + \int \phi^c \, d\nu \right\}$$

Wasserstein Distances

- If $c(x, y) = D^p(x, y)$, a distance-metric, then for measures $\mu, \nu \in P(\Omega)$, the p-Wasserstein Distance is
  - $W_p^p(\mu, \nu) = \left(\inf_{\gamma \in \mathscr{P}} \iint D(x, y)^p \gamma(dx, dy) \right) = \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y}) \sim \gamma} \left[D(x, y)^p \right]$
- In dual formulation
  - $W_p^p(\mu, \nu) = \sup_{\phi \in L_1(\mu), \psi \in L_1(\nu)} \int \phi d\mu + \int \psi d\nu$, where $\phi(x) + \psi(y) \leq D^p(x, y)$
- Special Case of semi-dual formulation - W_1 **Distance**
  - Proposition: if $c = |x - y|$, then $\phi^c = -\phi$ for all ϕ that are 1-Lipschitz.
  - $W_1(\mu, \nu) = \sup_{\phi \text{ is 1-Lipschitz}} \int \phi (d\mu - d\nu)$

Wasserstein Distances are natural metrics

- W-distances encode very different geometries from standard information divergences (KL, Euclidean)
- W-distances borrow key properties from the underlying distance metric and port them into the space of probability distributions
  - Euclidean distance -> interpolation, barycenters, etc

Wasserstein: $W_2^2(\alpha,\beta) \overset{\text{def.}}{=} \sup_{f,g} \{\int f \mathrm{d}\alpha + \int g \mathrm{d}\beta \; ; \; f(x)+g(y) \leqslant \|x-y\|^2\}$

Hellinger: $H^2(\alpha,\beta) \overset{\text{def.}}{=} \int \left(\sqrt{\frac{\mathrm{d}\alpha}{\mathrm{d}x}} - \sqrt{\frac{\mathrm{d}\beta}{\mathrm{d}x}}\right)^2 \mathrm{d}x$

Kullback-Leibler: $\mathrm{KL}(\alpha|\beta) \overset{\text{def.}}{=} \int \log\left(\frac{\mathrm{d}\alpha}{\mathrm{d}\beta}\right)\mathrm{d}\beta$

Burg: $B(\alpha|\beta) \overset{\text{def.}}{=} \mathrm{KL}(\beta|\alpha)$

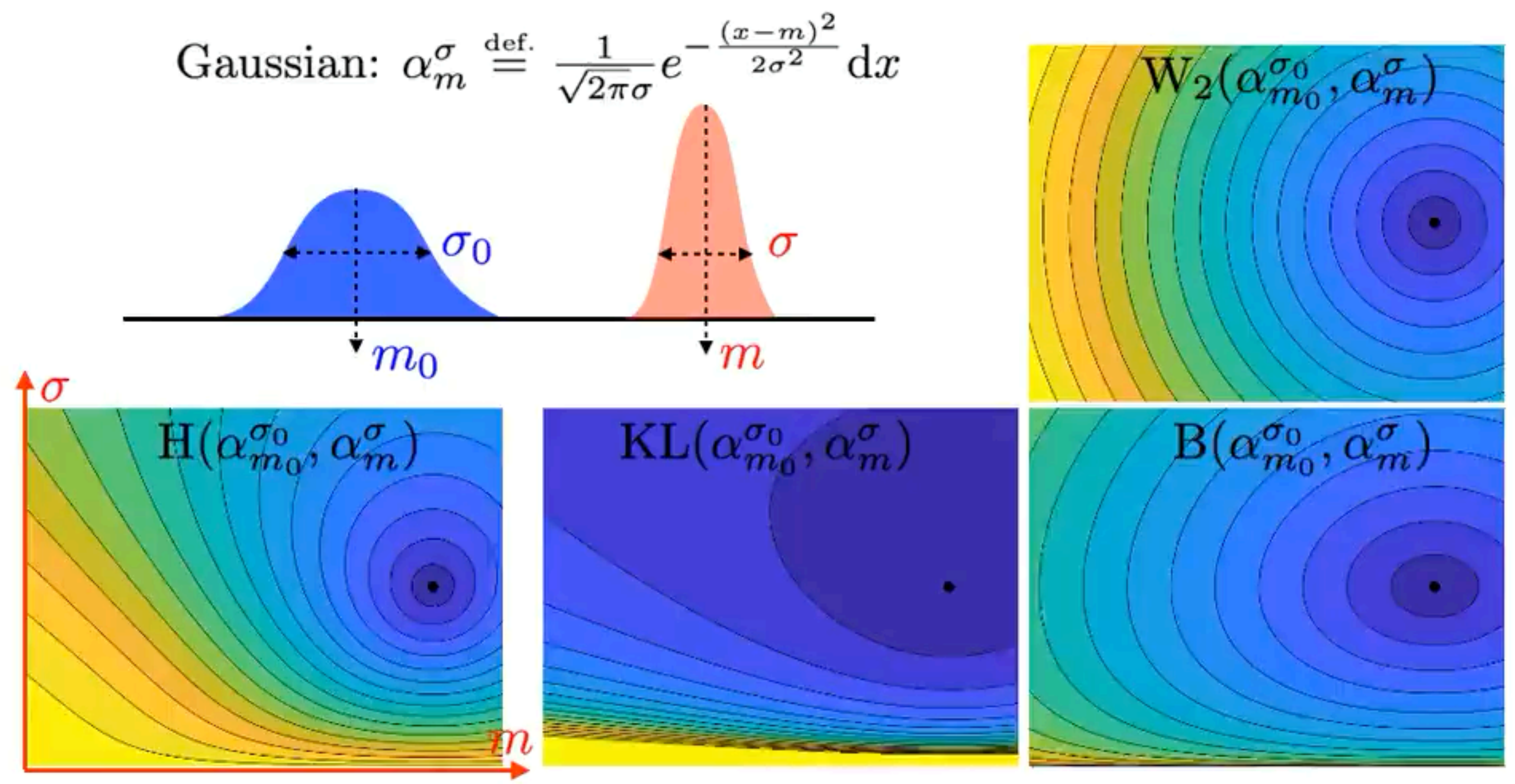


Image credit: Gabriel Peyre

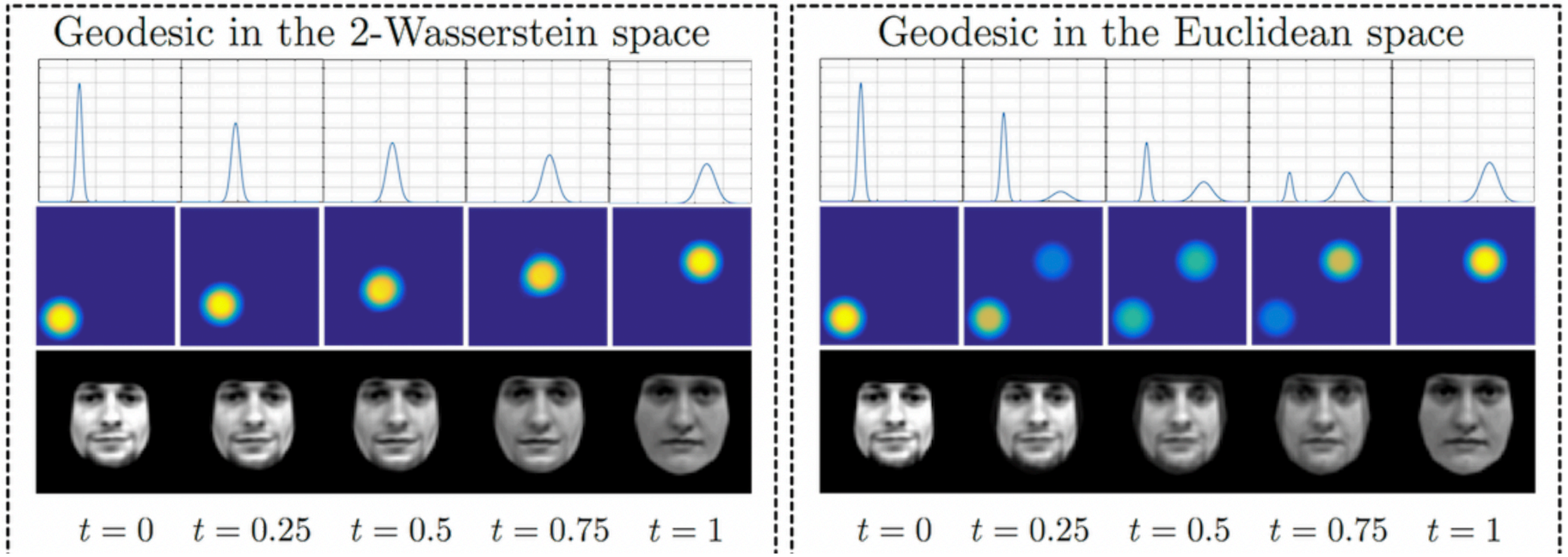


Image credit: [Kolouri et al. 2017]

Wasserstein Distances are natural metrics

- W-distances encode very different geometries from standard information divergences (KL, Euclidean)
- W-distances borrow key properties from the underlying distance metric and port them into the space of probability distributions
  - Euclidean distance -> interpolation, barycenters, convexity

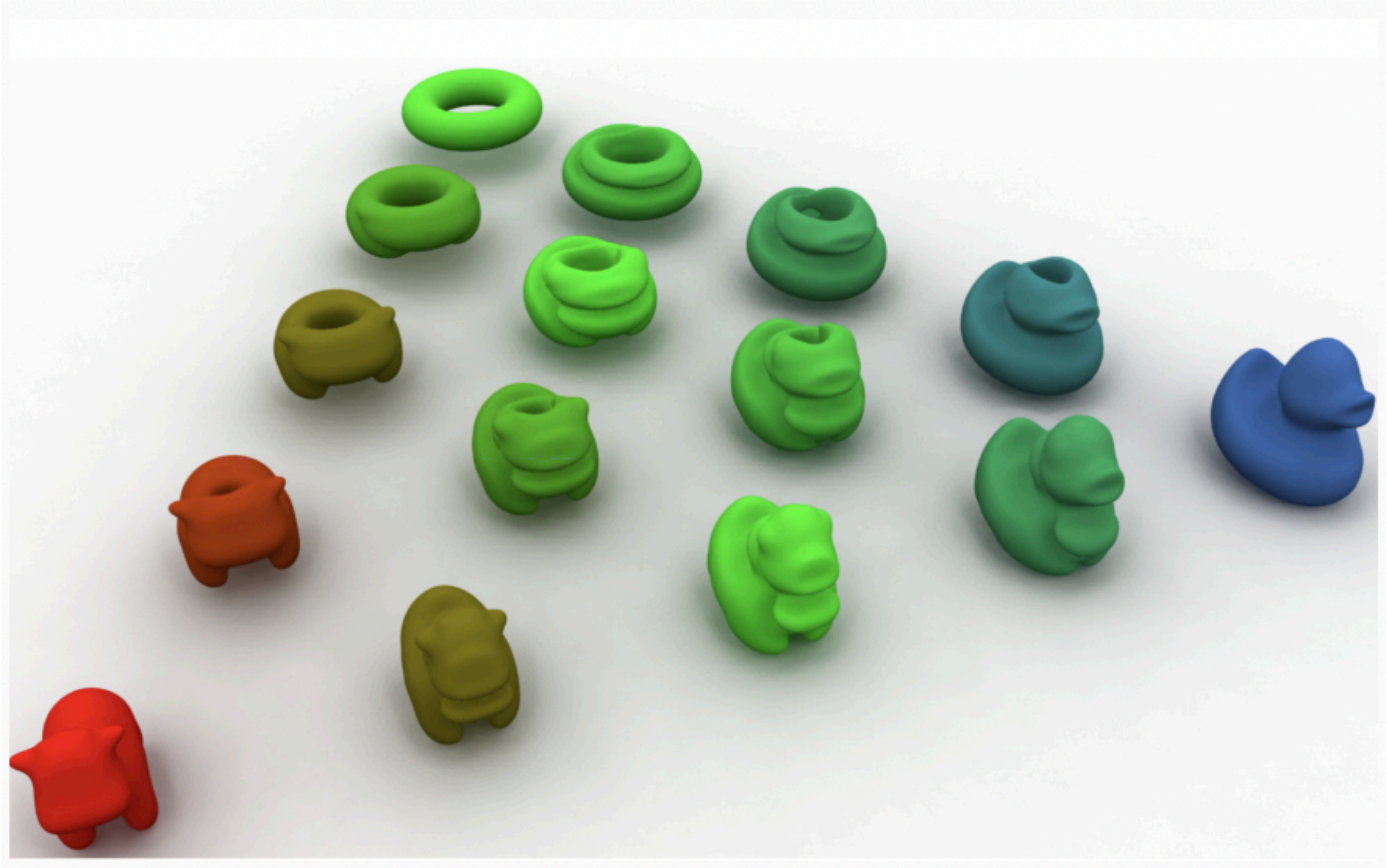


Image credit: [Solomon et al. 2015]

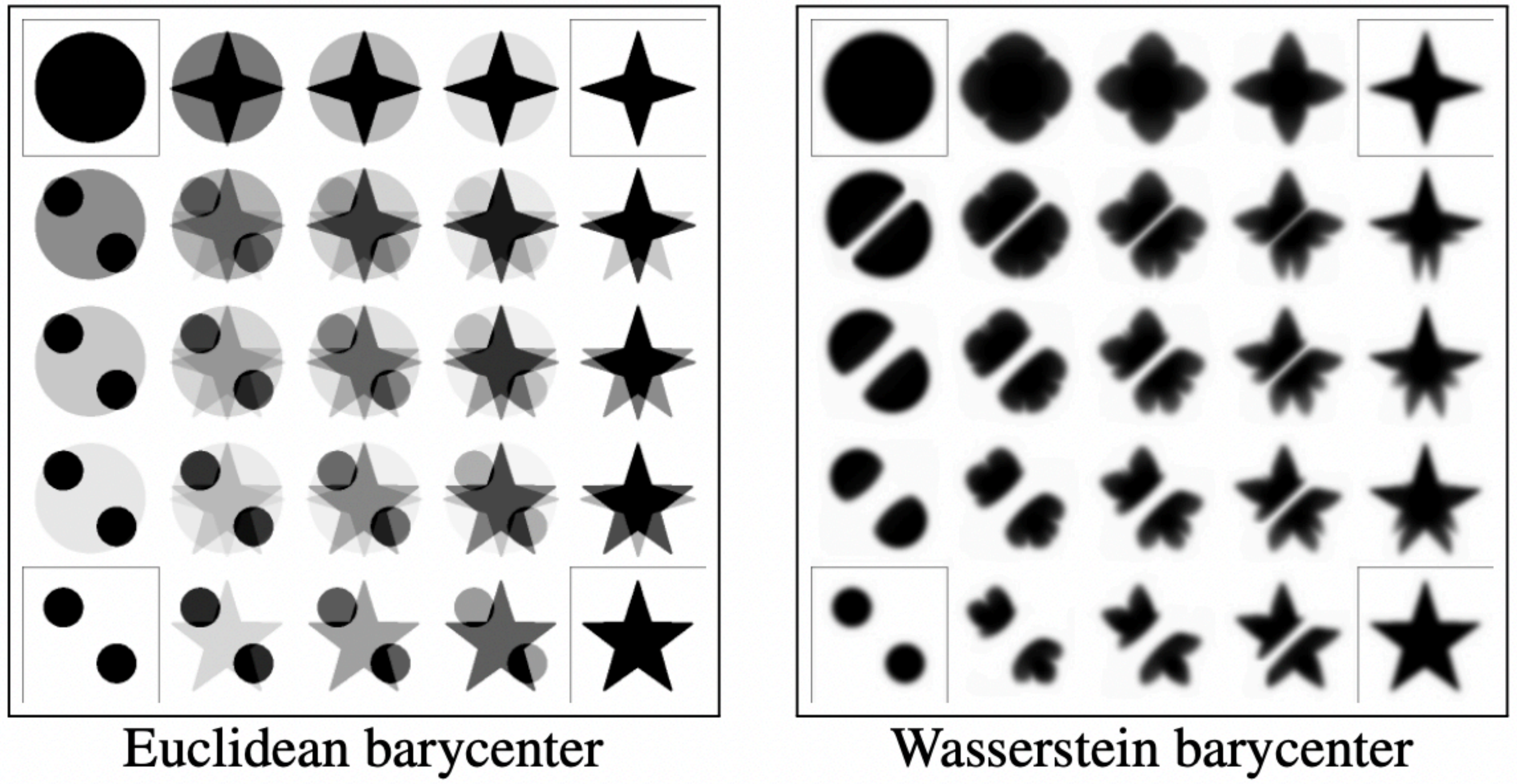


Image credit: [Solomon et al. 2015]

Wasserstein Distances are natural metrics

- W-distances encode very different geometries from standard information divergences (KL, Euclidean)
- W-distances borrow key properties from the underlying distance metric and port them into the space of probability distributions
  - Euclidean distance -> interpolation, barycenters, convexity
- What's the catch?
  - Quite expensive to calculate in practice
  - Not differentiable generally
  - Statistical properties don't scale to high-D distributions

Example - OT for Discrete Distributions

- Consider discrete measures $\mu = \sum_i^n a_i \delta_{\mathbf{x}_i}, \nu = \sum_i^m b_j \delta_{\mathbf{y}_j}$, where $\mathbf{x}_i, \mathbf{y}_j \in \Omega$, and $\sum_i^n a_i = 1, \sum_j^m b_j = 1$
  - Langrangian point clouds ($a_i = \frac{1}{n}, b_j = \frac{1}{m}$), Eulerian Histograms ($\mathbf{x}_i, \mathbf{y}_j$ are points on a grid)

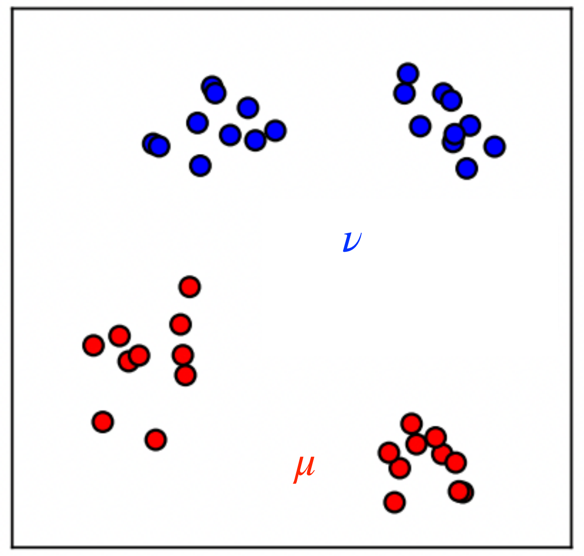


Image credit: Remi Flamary

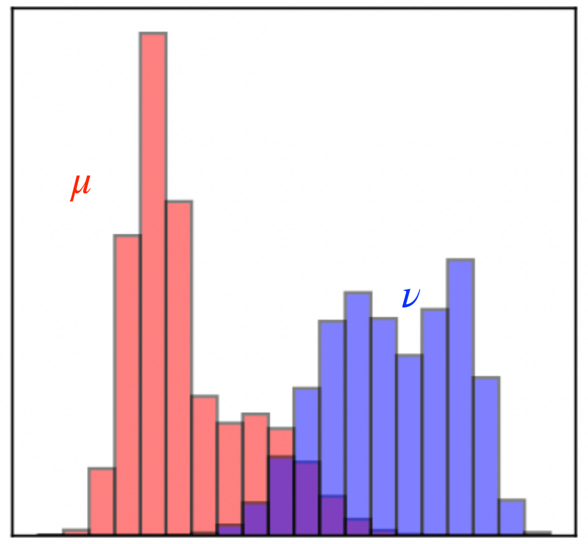


Image credit: Remi Flamary

- Given a cost matrix $\mathbf{C} = c(\mathbf{x}_i, \mathbf{y}_j)$, the optimal coupling between measures is a linear program given by

$$\gamma_0 = \underset{\gamma \in \mathscr{P}}{\operatorname{argmin}} \langle \mathbf{C}, \gamma \rangle_F = \sum_{i,j} \gamma_{i,j} c_{i,j} \text{ where } \mathscr{P} = \left\{ \gamma \in (\mathbb{R}^+)^{n \times m} \,|\, \gamma \mathbf{1}_n = \mathbf{a}, \gamma \mathbf{1}_m = \mathbf{b} \right\}$$

- Alternative dual formulation is given by $n + m$ variables and nm constraints

$$\max_{\boldsymbol{\alpha} \in \mathbb{R}^n, \boldsymbol{\beta} \in \mathbb{R}^m} \quad \boldsymbol{\alpha}^T \mathbf{a} + \boldsymbol{\beta}^T \mathbf{b} \qquad \text{s.t. } \alpha_i + \beta_j \leq c_{i,j} \quad \forall i, j$$

OT for Discrete Distributions - Issues

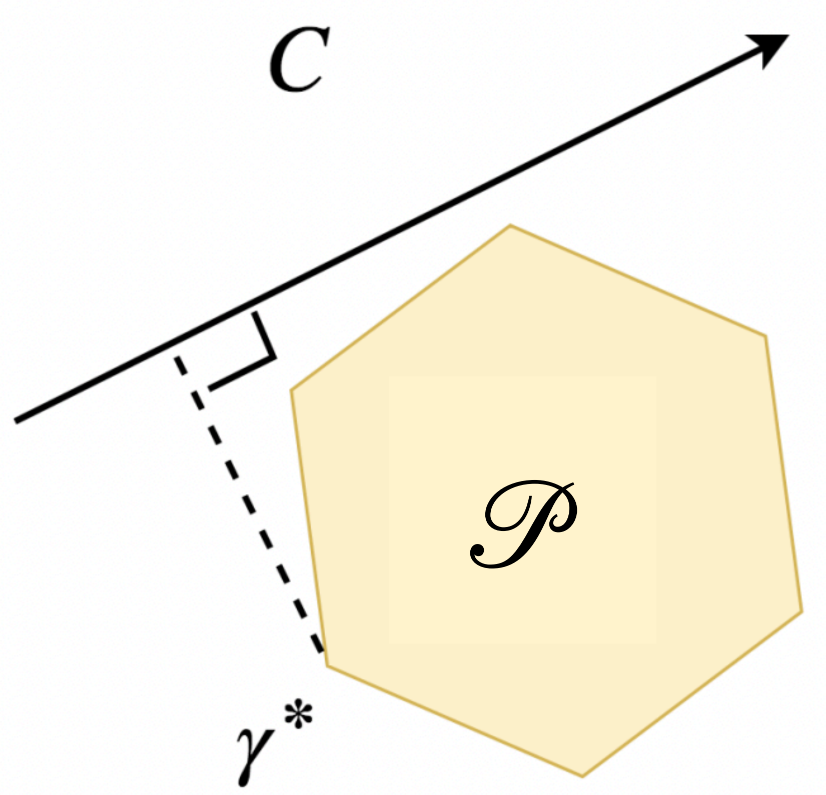


Image credit: Remi Flamary

- Linear Program - no unique solution sometimes, numerical instabilities
  - $W_p^p(\mu, \nu)$ is not differentiable
  - Not parallelisable on GPU hardware
  - Solving a linear problem is $\mathcal{O}((n+m)nm\log(n+m))$
- Assuming we have samples $x_1, \ldots, x_n \sim \mu, y_1, \ldots, y_m \sim \nu$, what are the considerations involved when computing $W_p^p(\hat{\mu}_n, \hat{\nu}_m)$, where $\hat{\mu}_n = \frac{1}{n}\sum_i \delta_{x_i}, \hat{\nu}_m = \frac{1}{m}\sum_j \delta_{y_j}$?
  - Can we bound $\mathbb{E}\left[\left| W_p(\mu, \nu) - W_p\left(\hat{\mu}_n, \hat{\nu}_m\right)\right|\right]$?
  - [Peyre et al., 15] If $\Omega = \mathbb{R}^d, d > 3$ then $\mathbb{E}\left[\left| W_p(\mu, \nu) - W_p\left(\hat{\mu}_n, \hat{\nu}_m\right)\right|\right] = \mathcal{O}(n^{-1/d})$
- What machine learning applications would ideally like
  - Faster, scalable, more stable, differentiable (ideally using autodiff), better statistical convergence

Approximate/Regularised OT

Sliced Wasserstein Distances

- For 1-D distributions $\Omega \in \mathbb{R}$, the W_p Distance is a function of the quantile functions $F_\mu^{-1}(x), F_\nu^{-1}(x)$

$$W_p(\mu, \nu) = \int_0^1 c\left(\left|F_\mu^{-1}(x) - F_\nu^{-1}(x)\right|^p\right) dx$$

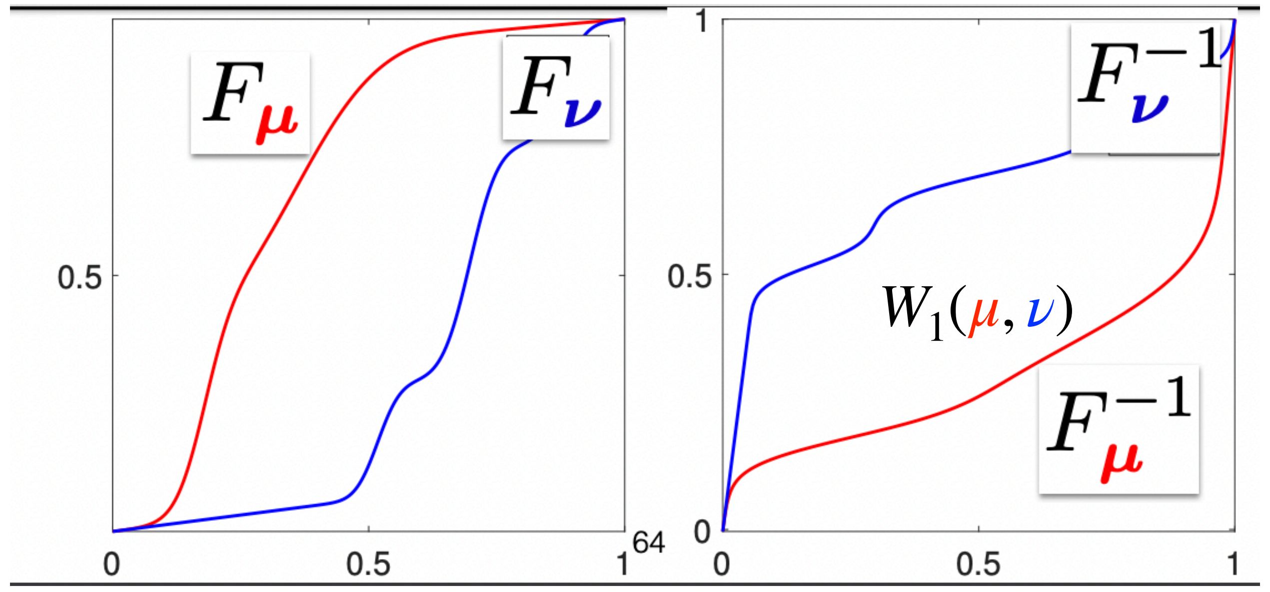


Image credit:Marco Cuturi

- For discrete distributions, very fast $\mathcal{O}(n \log n)$ algorithms exist
- **Idea - Project the high-dimensional distributions into 1 dimension, and calculate 1-D W_p distances**
- [Bonneel et al. 2015, Kolouri et al. 2017] accomplish this using the Radon Transform

$$\mathscr{R}(\mu, \theta) = \int_{\mathbb{S}^{d-1}} \delta(t - x^T\theta)\mu(x)dx, \quad t \in \mathbb{R}, \quad \theta \in \mathbb{S}^{d-1}$$

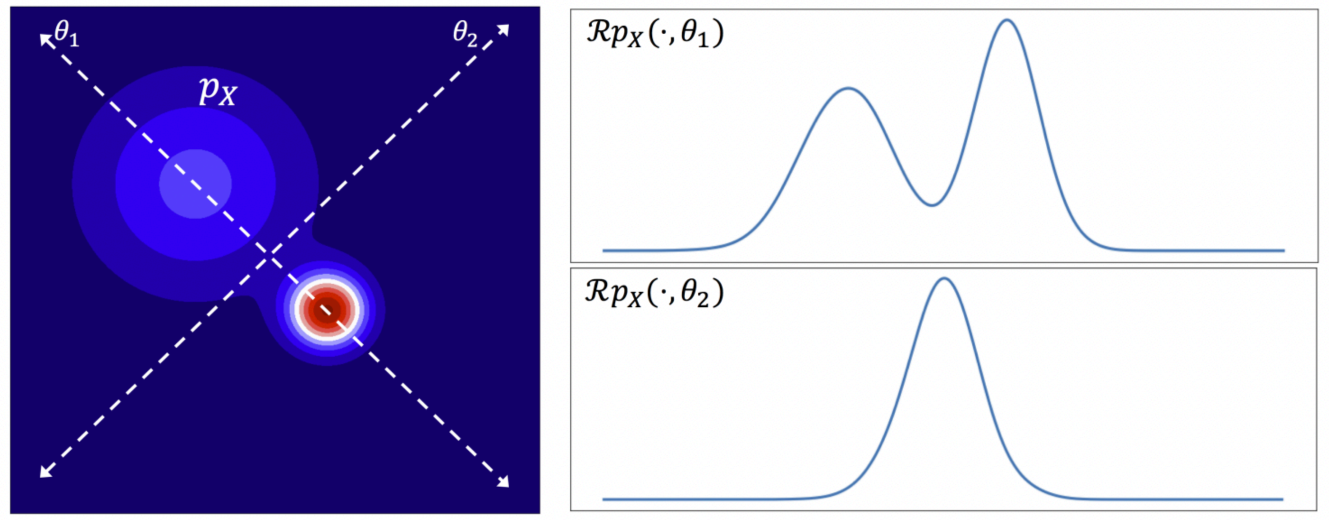


Image credit: [Kolouri et al 2017]

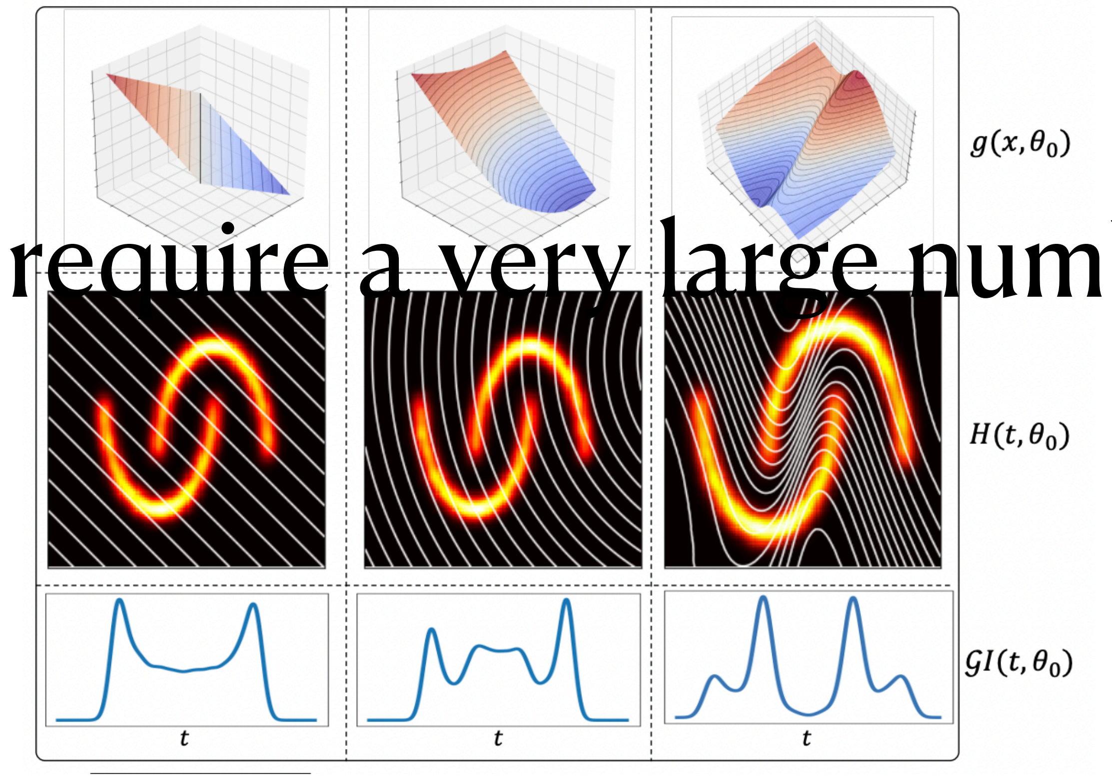
Sliced Wasserstein Distances

- [Bonneel et al. 2015] p-sliced Wasserstein distance

$$pSW_p^p(\mu,\nu) = \int_{\mathbb{S}^{d-1}} W_p^p\left(\mathscr{R}(\mu,\theta), \mathscr{R}(\nu,\theta)\right) d\theta$$

$$pSW_{p,K}^p(\mu,\nu) = \sum_l \frac{1}{K} W_p^p\left(\mathscr{R}(\mu,\theta_l), \mathscr{R}(\nu,\theta_l)\right), \qquad \mathcal{O}(Kn\log n)$$

- [Nadjahi et al, 2020] sliced W-distances are true metrics, topologically equivalent and weaker to W_p
  - Statistical convergence $\sim \mathcal{O}(K^{-1/2}n^{-1/2})$
  - [Kolouri et al, 2020] generalise this distance by formulating generalised Radon transforms onto general hyper-surfaces



- Still not differentiable, in practise can require a very large number of MC estimates if d is large

Regularised Optimal Transport

- **Idea - OT with Regularisation**
  - Option 1: Add priors to the family of couplings to consider
    - Add a regularisation term to the OT formulation, $\gamma_0^\lambda = \underset{\gamma \in \mathcal{P}}{\operatorname{argmin}} \langle \gamma, \mathbf{C} \rangle_F + \lambda R(\gamma)$
      - [Cuturi, 2013] Entropic Regularisation, $R(\gamma) = \sum_{i,j} \gamma_{i,j}(\log \gamma_{i,j} - 1)$
      - [Courty et al., 2016] Group Lasso, $R(\gamma) = \sum_g \sqrt{\sum_{i,j \in \mathcal{G}_g} \gamma_{i,j}^2}$
  - Option 2: Relax the requirement for $W_1(\mu, \nu) = \sup_{\phi \text{ is 1-Lipschitz}} \int \phi(d\mu - d\nu)$
    - [Makkouva et al., 17] Use RELU Networks with bounded weights
    - [Shirdhonkar'08] - Use low-dimensional wavelet decompositions
  - Option 3: Change the cost function in $\underset{\gamma \in \mathcal{P}}{\operatorname{argmin}} \int_{\Omega_s \times \Omega_t} c(\mathbf{x}, \mathbf{y}) \gamma(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y}$
    - [Solomon+, '17] Geodesic Distances on graphs simplify the Linear Program

Entropic Regularised OT

- We have $\gamma_0^\lambda = \underset{\gamma \in \mathcal{P}}{\operatorname{argmin}} \langle \gamma, \mathbf{C} \rangle_F + \lambda \sum_{i,j} \gamma_{i,j} (\log \gamma_{i,j} - 1) = \underset{\gamma \in \mathcal{P}}{\operatorname{argmin}} \langle \gamma, \mathbf{C} \rangle_F - \lambda \mathbb{H}(\gamma)$
- [Wilson, '69] Define a regularised Wasserstein distance, for $\lambda \geq 0$

$$W_\lambda(\mu, \nu) = \min_{\gamma \in \mathcal{P}} \langle \gamma, \mathbf{C} \rangle_F - \lambda \mathbb{H}(\gamma)$$

- If $\lambda \geq 0$, then the linear program becomes a λ-strongly convex optimisation problem
- Fast and scalable, differentiable - **Sinkhorn's Algorithm**
    - $\mathcal{O}(nm)$ complexity in general, $\simeq \mathcal{O}(n \log n)$ on gridded spaces with convolutions [Solomon et al., '15]
- Better statistical convergence properties - **Sinkhorn Divergences**

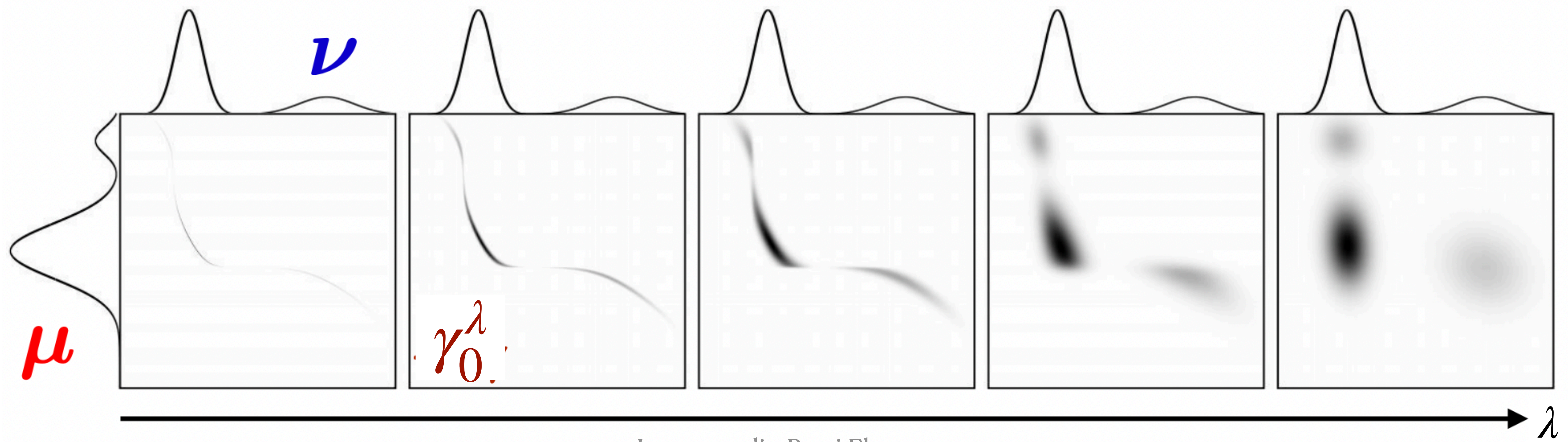


Image credit: Remi Flamary

Sinkhorn's Algorithm - A Fast and Scalable OT Solver

- Proposition: If $\gamma_0^\lambda = \underset{\gamma \in \mathscr{P}}{\operatorname{argmin}} \langle \gamma, \mathbf{C} \rangle_F - \lambda \mathbb{H}(\gamma)$, then there exists $\mathbf{u} \in \mathbb{R}_+^n$, $\mathbf{v} \in \mathbb{R}_+^m$ such that

$$\gamma_0^\lambda = \operatorname{diag}(\mathbf{u})\mathbf{K}\operatorname{diag}(\mathbf{v}), \text{ where } \mathbf{K} = e^{-\mathbf{C}/\lambda}$$

- Write down the Lagrangian to solve the convex optimisation problem

$$L(\gamma, \alpha, \beta) = \sum_{ij} \gamma_{i,j}\mathbf{C}_{i,j} + \lambda\gamma_{i,j}\left(\log \gamma_{i,j} - 1\right) + \alpha^T(\gamma\mathbf{1} - \mathbf{a}) + \beta^T\left(\gamma^T\mathbf{1} - \mathbf{b}\right)$$

$$\partial L / \partial \gamma_{i,j} = \mathbf{C}_{i,j} + \lambda \log \gamma_{i,j} + \alpha_i + \beta_j \Rightarrow 0$$

$$\gamma_{i,j} = e^{\frac{\alpha_i}{\beta}} e^{-\frac{\mathbf{C}_{i,j}}{\lambda}} e^{\frac{\beta_j}{\lambda}} = u_i K_{ij} v_j$$

Ref: Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. Advances in neural information processing systems, 26.

Sinkhorn's Algorithm - A Fast and Scalable OT Solver

- Proposition: If $\gamma_0^\lambda = \underset{\gamma \in \mathscr{P}}{\operatorname{argmin}} \langle \gamma, \mathbf{C} \rangle_F - \lambda \mathbb{H}(\gamma)$, then there exists $\mathbf{u} \in \mathbb{R}_+^n, \mathbf{v} \in \mathbb{R}_+^m$ such that

$$\gamma_0^\lambda = \operatorname{diag}(\mathbf{u}) \mathbf{K} \operatorname{diag}(\mathbf{v}), \text{ where } \mathbf{K} = e^{-\mathbf{C}/\lambda}$$

- To solve, first use the marginalisation constraints

$$\begin{cases} \operatorname{diag}(\mathbf{u}) K \operatorname{diag}(\mathbf{v}) \mathbf{1}_m &= \mathbf{a} \\ \operatorname{diag}(\mathbf{v}) K^T \operatorname{diag}(\mathbf{u}) \mathbf{1}_n &= \mathbf{b} \end{cases}$$

$$\begin{cases} \mathbf{u} \odot K \mathbf{v} &= \mathbf{a} \\ \mathbf{v} \odot K^T \mathbf{u} &= \mathbf{b} \end{cases}$$

- Fixed-point algorithm, repeat until convergence [Sinkhorn, '67]

$$\mathbf{u} \leftarrow \mathbf{a} / \mathbf{K} \mathbf{v} \quad \text{followed by} \quad \mathbf{v} \leftarrow \mathbf{b} / \mathbf{K}^T \mathbf{u}$$

Sinkhorn's Algorithm - A Fast and Scalable OT Solver

- Fixed-point algorithm, repeat until convergence [Sinkhorn, '67]

$$\mathbf{u} \leftarrow \mathbf{a}/\mathbf{K}\mathbf{v} \quad \text{followed by} \quad \mathbf{v} \leftarrow \mathbf{b}/\mathbf{K}^T\mathbf{u}$$

- Define the iterative Wasserstein Distance

$$W_L(\mu, \nu) = \langle \gamma_L, \mathbf{C} \rangle, \quad \text{where } \gamma_L = \text{diag}(\mathbf{u}_L)\mathbf{K}\,\text{diag}(\mathbf{v}_L)$$

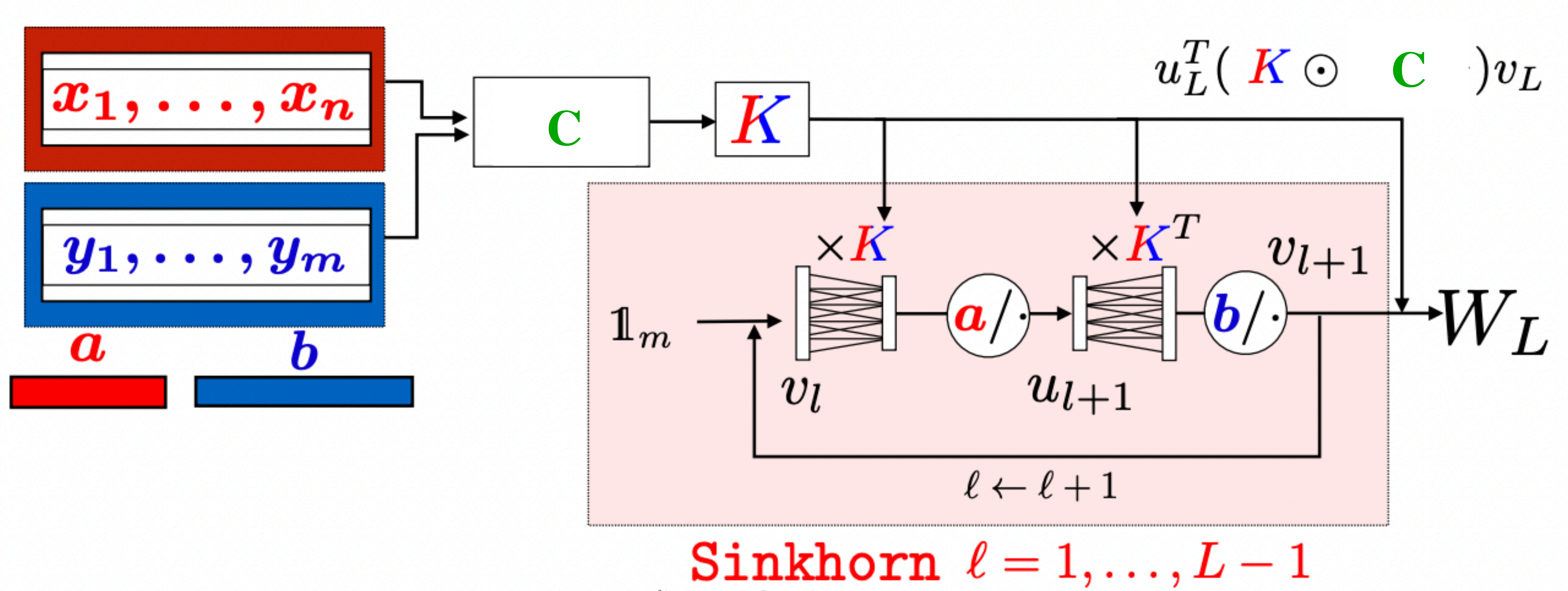


Image credit: Marco Cuturi

- $\dfrac{\partial W_L}{\partial \mathbf{X}}, \dfrac{\partial W_L}{\partial \mathbf{a}}, \dfrac{\partial W_L}{\partial \mathbf{Y}}, \dfrac{\partial W_L}{\partial \mathbf{b}}$ can be computed recursively (and using autodiff)

Sinkhorn's Algorithm - A Fast and Scalable OT Solver

- Computational complexity - $\mathcal{O}((n+m)^2) \times \mathcal{O}(d^2)$
- Linear convergence for **u**, **v** -> Rate bounded by λ

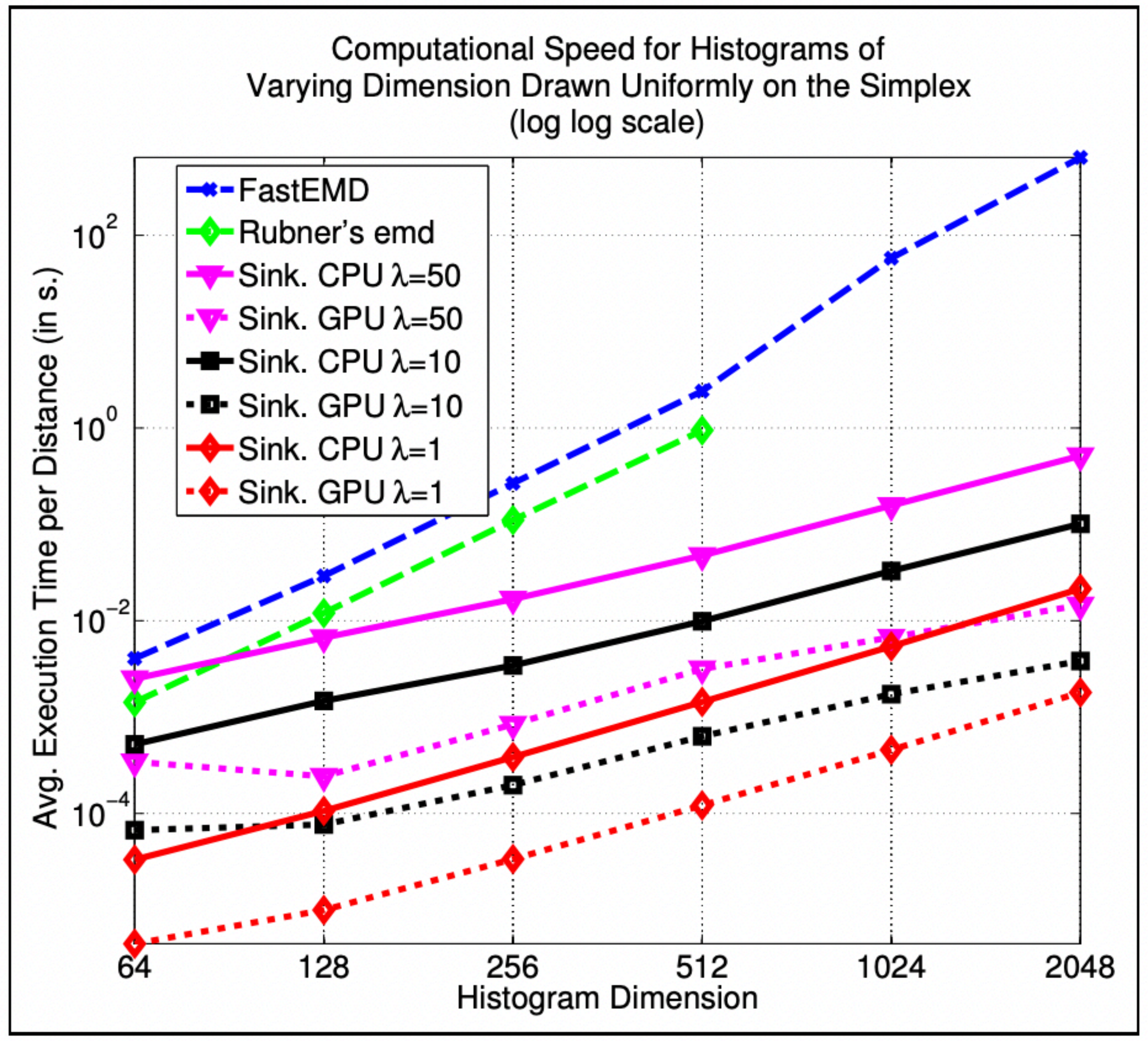


Image credit: [Cuturi et al., 2013]

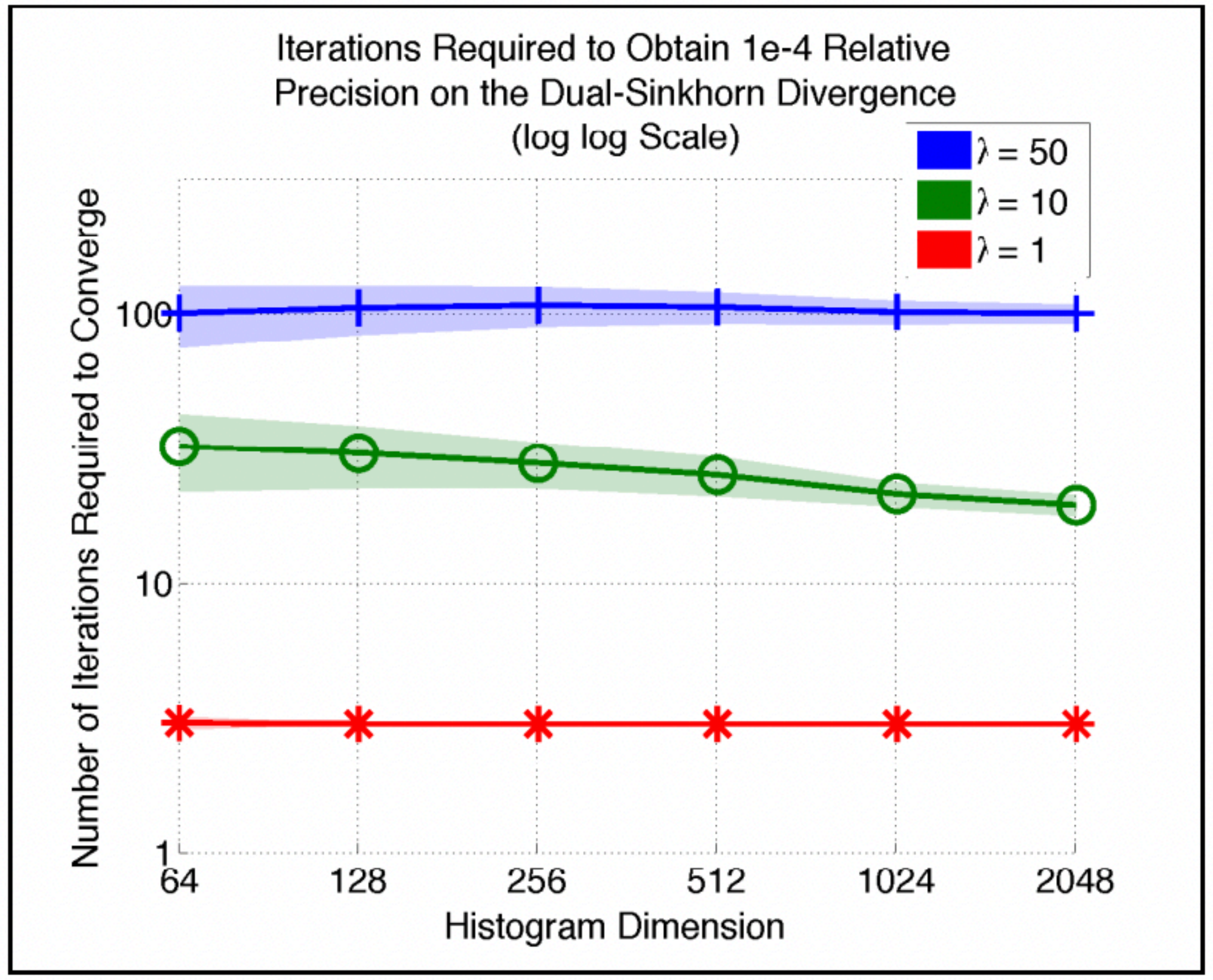


Image credit: [Cuturi et al., 2013]

Sinkhorn's Algorithm as Bregman Projections

- Fixed-point algorithm, repeat until convergence [Sinkhorn, '67]

$$\mathbf{u} \leftarrow \mathbf{a}/\mathbf{K}\mathbf{v} \quad \text{followed by} \quad \mathbf{v} \leftarrow \mathbf{b}/\mathbf{K}^T\mathbf{u}$$

- [Benamou et al., 2015] show that solving entropic regularised OT is the same as Bregman projections

- Proposition: γ_0^λ is the solution of the following Bregman projection

$$\gamma_0^\lambda = \underset{\gamma \in \mathscr{P}}{\operatorname{argmin}}\, \mathrm{KL}(\gamma, \mathbf{K})$$

- Can be generalised to calculate Wasserstein barycenters

$$\min_{\mu} \sum_{i=1}^{N} \lambda_i W_\lambda(\mu, \nu_i) \quad \rightarrow \quad \boldsymbol{\gamma} = [\gamma_1, \ldots, \gamma_N] = \underset{\gamma \in \mathscr{P}_i^K}{\operatorname{argmin}} \sum_{i}^{N} \lambda_i \mathrm{KL}(\gamma_i, \mathbf{K})$$

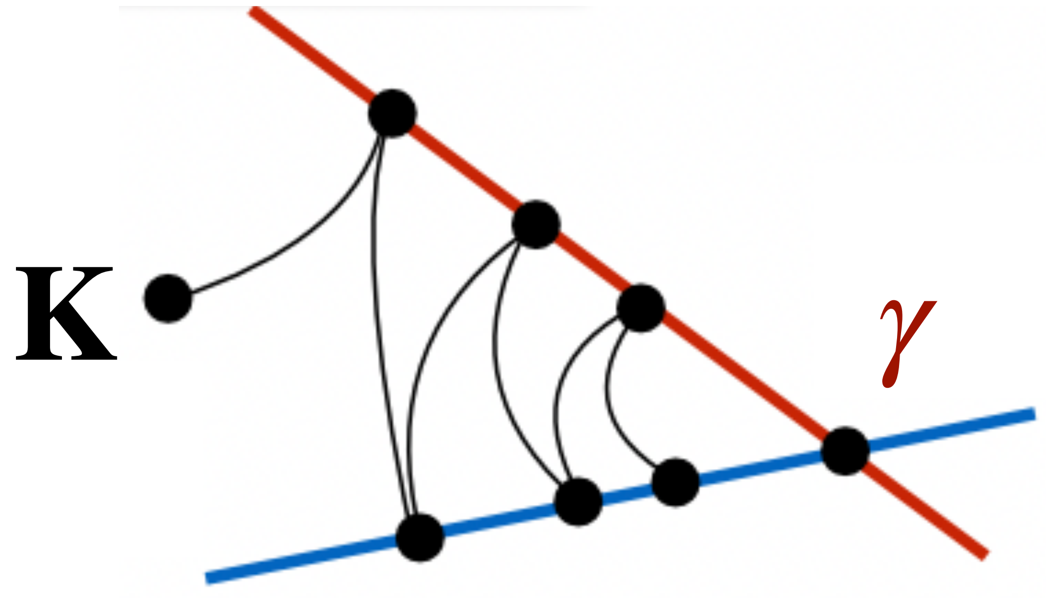


Image credit: Marco

Sinkhorn Divergences

- Given the regularised Wasserstein Distance $W_\lambda(\mu, \nu) = \min_{\gamma \in \mathscr{P}} \langle \gamma, \mathbf{C} \rangle_F - \lambda \mathbb{H}(\gamma)$
  - Issue: $W_\lambda(\mu, \mu) \neq 0$
  - Fix [Ramdas et al., 2017] : $\overline{W}_\lambda(\mu, \nu) = W_\lambda(\mu, \nu) - \dfrac{1}{2} W_\lambda(\mu, \mu) - \dfrac{1}{2} W_\lambda(\nu, \nu)$
    - Sinkhorn Divergences have some nice distance-based and interpolating properties
    - When $\lambda \to 0$, we re-obtain OT
      - $\lim_{\lambda \to 0} \overline{W}_\lambda(\mu, \nu) = W_p^p(\mu, \nu)$
    - When $\lambda \to \infty$, we obtain kernel-based distances (Maximum Mean Discrepancy, Energy Distance)
      - $\lim_{\lambda \to \infty} \overline{W}_\lambda(\mu, \nu) = E(\mu, \nu) - \dfrac{1}{2} E(\mu, \mu) - \dfrac{1}{2} E(\nu, \nu)$, where $E(\mu, \nu) = \langle \mathbf{a}\mathbf{b}^T, \mathbf{C} \rangle$

Sinkhorn Divergences

- Assuming we have samples $x_1, \ldots, x_n \sim \mu, y_1, \ldots, y_m \sim \nu$, what are the considerations involved when computing $W_p^p(\hat{\mu}_n, \hat{\nu}_m)$, where $\hat{\mu}_n = \frac{1}{n}\sum_i \delta_{x_i}, \hat{\nu}_m = \frac{1}{m}\sum_j \delta_{y_j}$?

Computational Costs		**Statistical Convergence**
$(n+m)^2$	$MMD(\mu,\nu) = E(\mu,\nu) - \frac{1}{2}E(\mu,\mu) - \frac{1}{2}E(\nu,\nu)$	$\mathcal{O}(1/\sqrt{n})$
	$\uparrow \lambda \to \infty$	
$\mathcal{O}((n+m)^2)$	$\overline{W}_\lambda(\mu,\nu) = W_\lambda(\mu,\nu) - \frac{1}{2}W_\lambda(\mu,\mu) - \frac{1}{2}W_\lambda(\nu,\nu)$	$\mathcal{O}\left(\frac{1}{\lambda^{d/2}\sqrt{n}}\right)$
	$\downarrow \lambda \to 0$	
$\mathcal{O}((n+m)nm\log nm)$	$W_p^p(\mu,\nu)$	$\mathcal{O}\left(1/n^{1/d}\right)$

Ref: Gretton, Arthur, et al. "A kernel two-sample test." The Journal of Machine Learning Research 13.1 (2012): 723-773,

Applications in Machine Learning

OT for Supervised Learning - Wasserstein Loss

- [Frogner et al 2015] Multiclass classification - learn optimal maps from $\mathscr{X} \in \mathbb{R}^d$ to $\mathscr{Y} = \mathbb{R}_+^K$ through $\mathscr{H} = h_\theta : \mathscr{X} \to \mathscr{Y}$
  - $h_\theta, y \in \Delta^k$ (the K-d simplex), and $\mathbf{C} \in \mathbb{R}_+^{K,K}$ where $\mathbf{C}_{\kappa,\kappa'} = d^p(\kappa, \kappa')$
  - Minimise the entropic regularised Wasserstein Distance $W_p^\lambda(h(\cdot \mid x), y(\cdot))$
  - Ground-truth metric can encode semantic similarity
    - Flickr Creative Commons 100M dataset : $d^p(\kappa, \kappa') = \|\text{word2vec}(\kappa) - \text{word2vec}(\kappa')\|_2^2$
      - Example labels - travel, square, wedding, art, flower, music, nature, ...

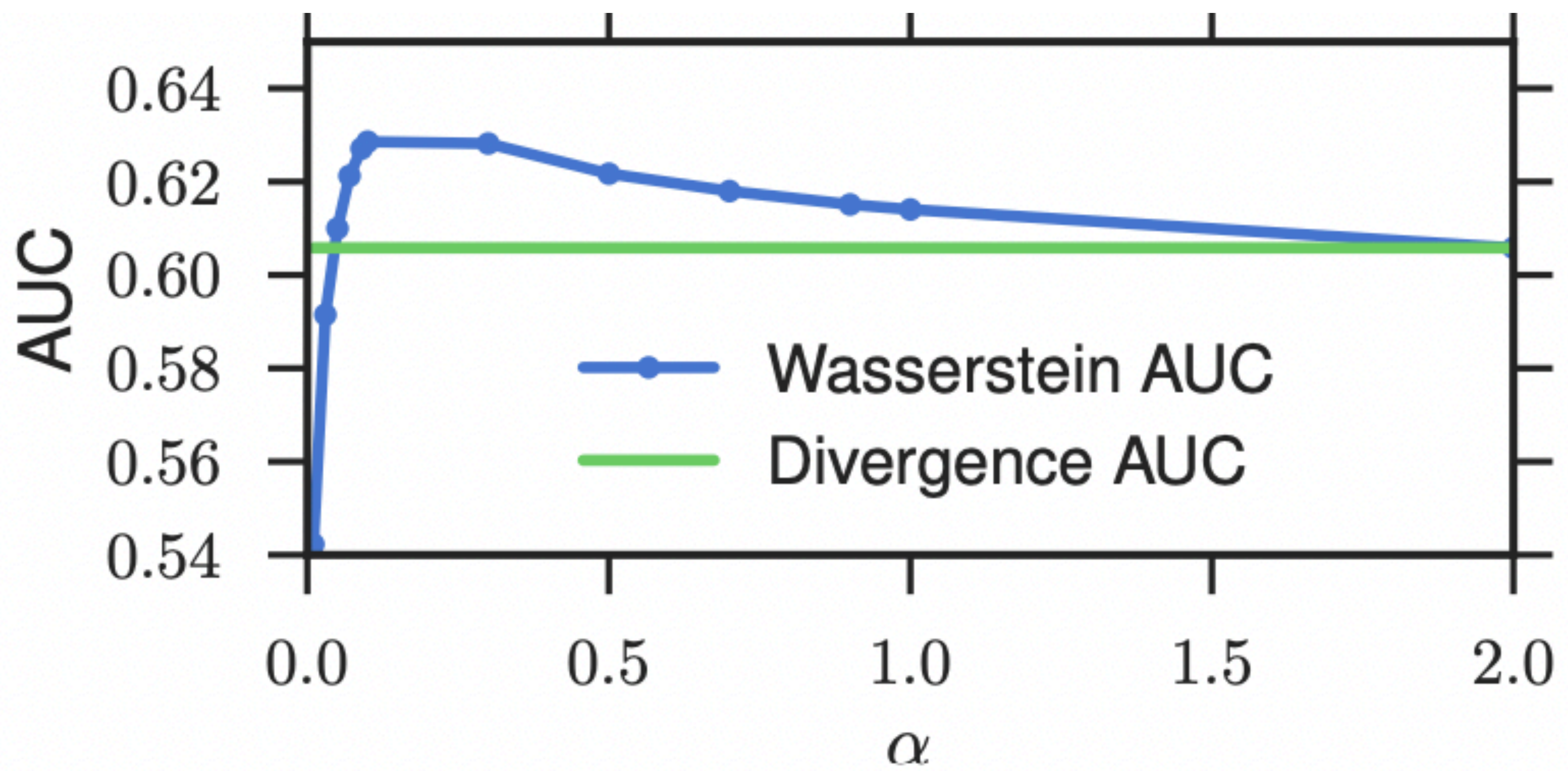


Image credit: [Frogner et al 2015]

Siberian husky Eskimo dog

Image credit: [Frogner et al 2015]

OT for Generative Modelling - WGAN

- Let $\mathbb{P}_r$ denote the real data distribution over a metric space Ω (i.e image space of $[0,1]^{h\times w\times 3}$),
- Let Z be a random variable over a space $\mathcal{Z}$, $g : \mathcal{Z} \times \mathbb{R}^d \to \Omega$ a function parametrised by $\theta \in \mathbb{R}^d$
- Let $\mathbb{P}_\theta$ denote the distribution over $g_\theta(Z)$
- [Arjovsky et al., 2017] trains generative models by minimising the W_1 distance b/w $\mathbb{P}_r$ and $\mathbb{P}_\theta$

$$W_1^1\left(\mathbb{P}_r, \mathbb{P}_\theta\right) = \inf_{\gamma \in \mathcal{P}\left(\mathbb{P}_r, \mathbb{P}_\theta\right)} \mathbb{E}_{(x,y)\sim\gamma}[\|x-y\|]$$

- Using the semi-dual formulation, where f is a 1-Lipschitz function -

$$W_1^1\left(\mathbb{P}_r, \mathbb{P}_\theta\right) = \sup_{\|f\|_L \le 1} \mathbb{E}_{x\sim\mathbb{P}_r}[f(x)] - \mathbb{E}_{x\sim\mathbb{P}_\theta}[f(x)]$$

- If instead we consider K-Lipschitz functions instead, we get

$$\sup_{\|f\|_L \le 1} \mathbb{E}_{x\sim\mathbb{P}_r}[f(x)] - \mathbb{E}_{x\sim\mathbb{P}_\theta}[f(x)] \le \sup_{\|f\|_L \le K} \mathbb{E}_{x\sim\mathbb{P}_r}[f(x)] - \mathbb{E}_{x\sim\mathbb{P}_\theta}[f(x)] = K \,.\, W_1^1\left(\mathbb{P}_r, \mathbb{P}_\theta\right)$$

OT for Generative Modelling - WGAN

- Therefore, for parametrised family of functions $\left\{f_\phi\right\}_{\phi\in\Phi}$ that are all K-Lipschitz, solve instead

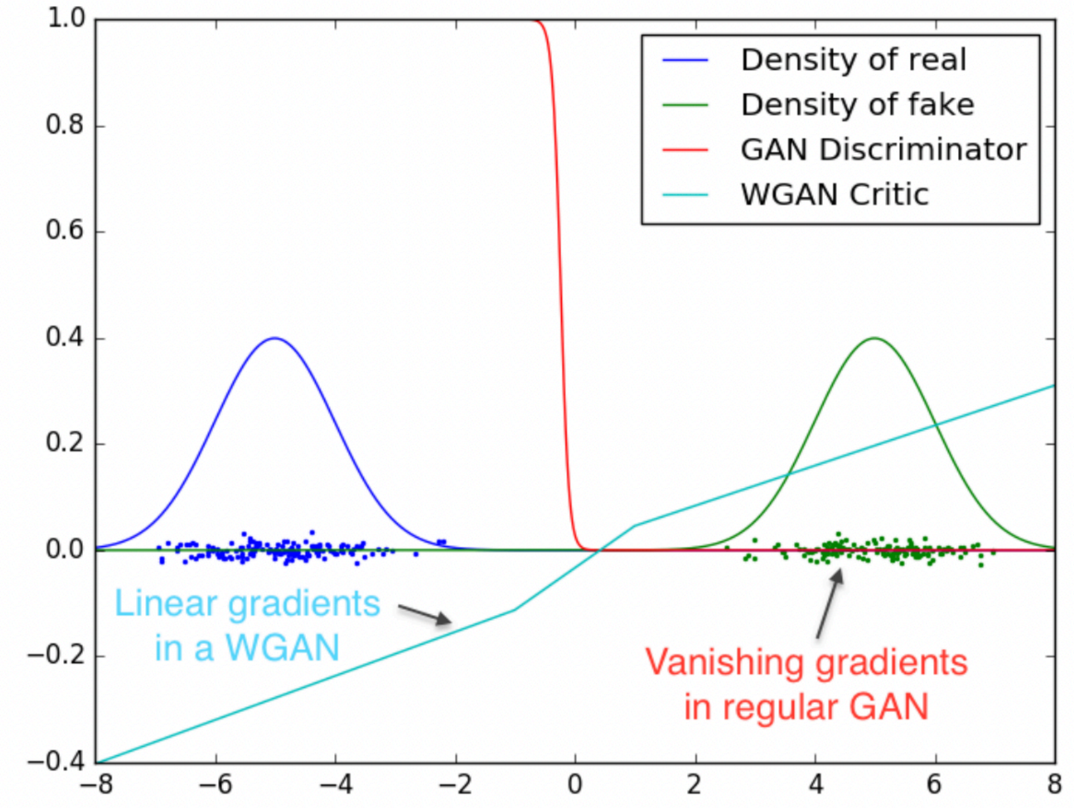
$$W\left(\mathbb{P}_r, \mathbb{P}_\theta\right) = \max_{\phi\in\Phi} \mathbb{E}_{x\sim\mathbb{P}_r}\left[f_\phi(x)\right] - \mathbb{E}_{z\sim p(z)}\left[f_\phi\left(g_\theta(z)\right)\right]$$

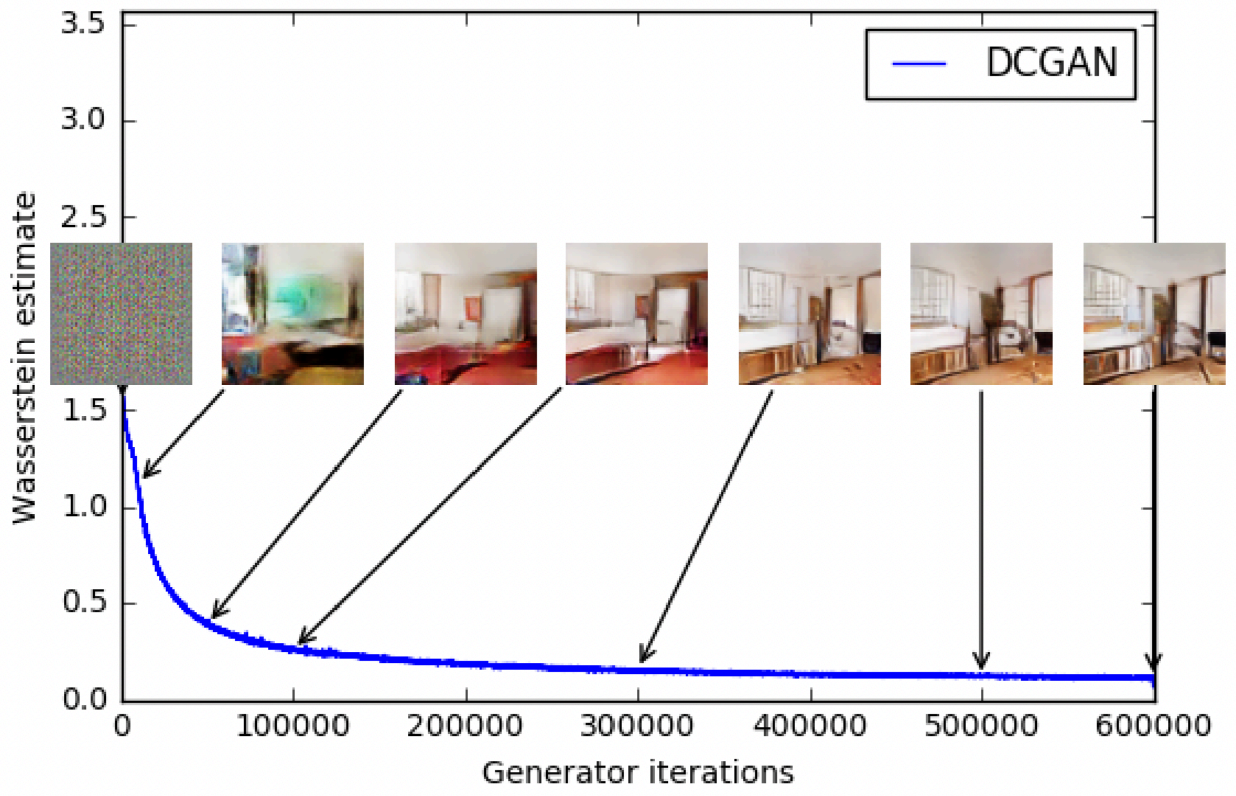
- The paper proves that $W\left(\mathbb{P}_r, \mathbb{P}_\theta\right)$ is the W_1 distance unto a multiplicative factor, and further that

$$\nabla_\theta W\left(\mathbb{P}_r, \mathbb{P}_\theta\right) = -\mathbb{E}_{z\sim p(z)}\left[\nabla_\theta f\left(g_\theta(z)\right)\right]$$

- K-Lipschitz bound is roughly enforced by gradient clipping

$$\phi \leftarrow \mathrm{clip}(\phi, -c, c)$$





OT for Generative Modelling - Extensions

- [Guljarani et al., 2017] Improved WGAN - Replace weight clipping with constraint on gradient norm
  - $W\left(\mathbb{P}_r, \mathbb{P}_\theta\right) = \max_{\phi \in \Phi} \mathbb{E}_{x \sim \mathbb{P}_r}\left[f_\phi(x)\right] - \mathbb{E}_{z \sim p(z)}\left[f_\phi\left(g_\theta(z)\right)\right] + \lambda \mathbb{E}_{x \sim \mathbb{P}_r}\left[\left(\|\nabla f_\phi(\mathbf{x})\|_2 - 1\right)^2\right]$
  - A differentiable function is 1-Lipschitz i.f.f it has gradients with norm at most 1 everywhere

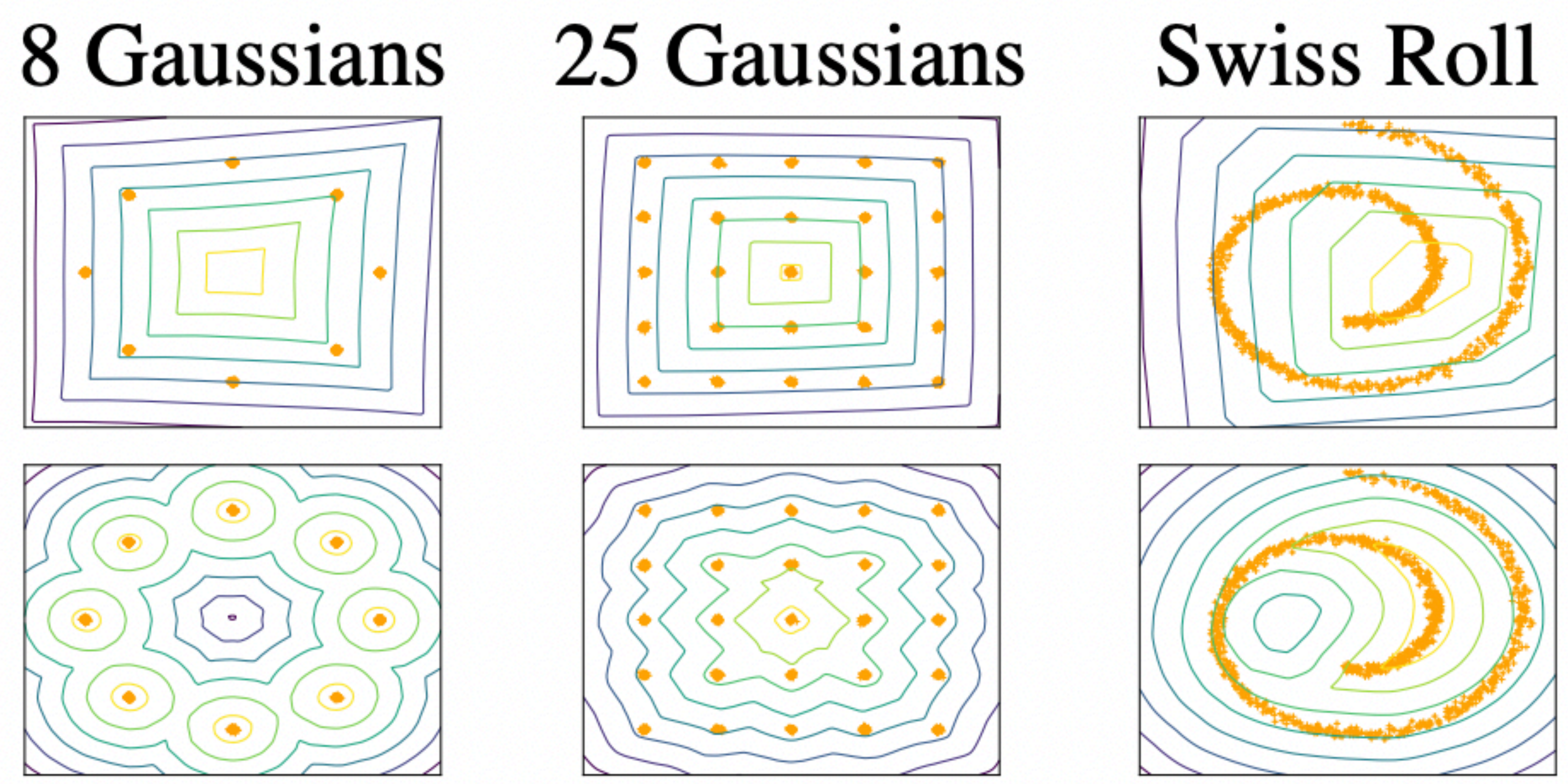


Image credit:[Guljarani et al 2017]

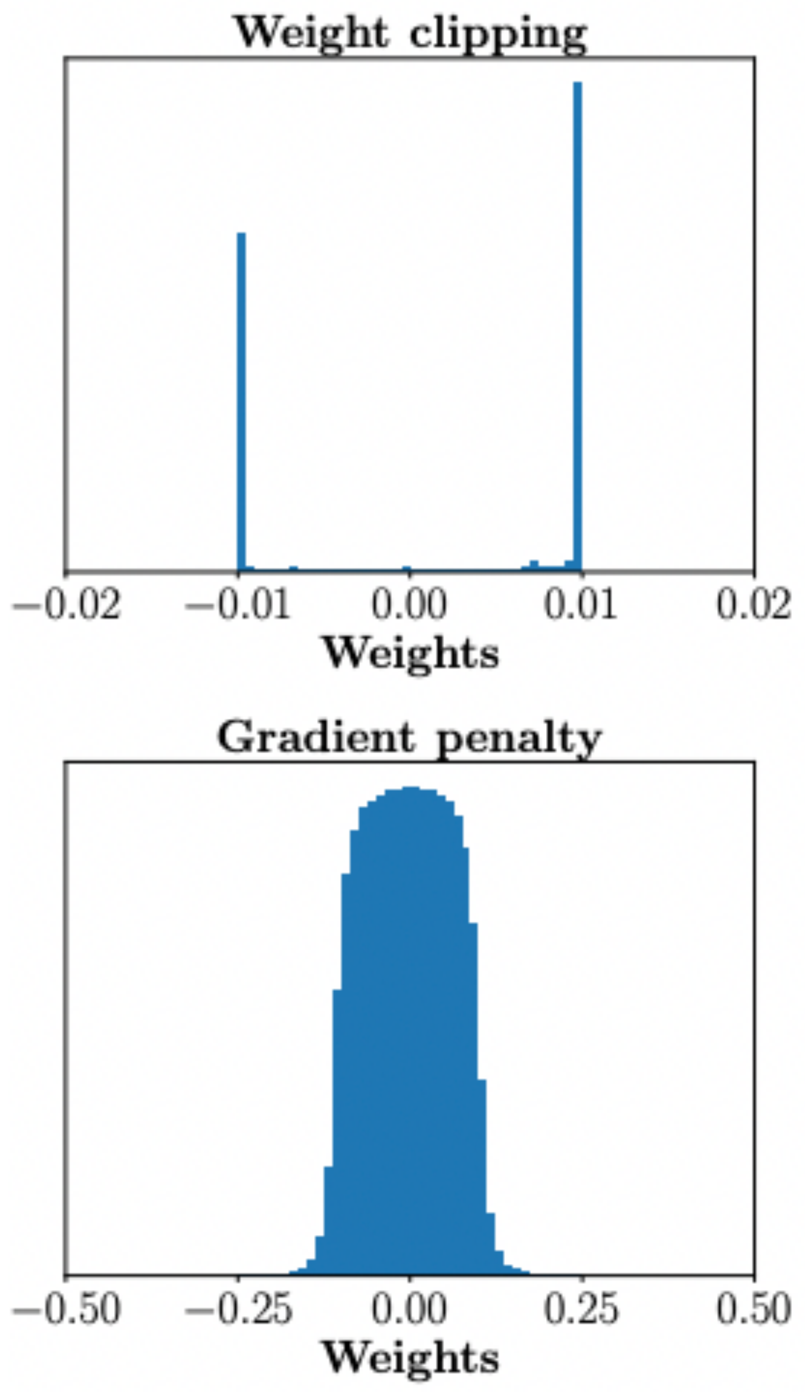


Image credit:[Guljarani et al 2017]

OT for Generative Modelling - Sinkhorn Divergences

- [Genevay et al., 2017] Generative Models with Sinkhorn Divergences
  - Define $\mathbb{P}_r = \frac{1}{N}\sum_{j=1}^{N} \delta_{y_j}$ the empirical data distribution, $\mathbb{P}_\theta = g_\theta(Z)$
  - Generator is trained through $\min_\theta \hat{E}_L(\theta) = \overline{W}_\lambda(\mathbb{P}_r, \mathbb{P}_\theta) \simeq 2W_L(\mathbb{P}_r, \mathbb{P}_\theta) - W_L(\mathbb{P}_r, \mathbb{P}_r) - W_L(\mathbb{P}_\theta, \mathbb{P}_\theta)$
  - Cost function in general is $c_\phi(x, y) = \left\| f_\phi(x) - f_\phi(y) \right\|$ where $f_\phi : \mathcal{X} \to \mathbb{R}^p$
  - $\frac{\partial W_L}{\partial \theta}, \frac{\partial W_L}{\partial \phi}$ can be obtained through autodiff

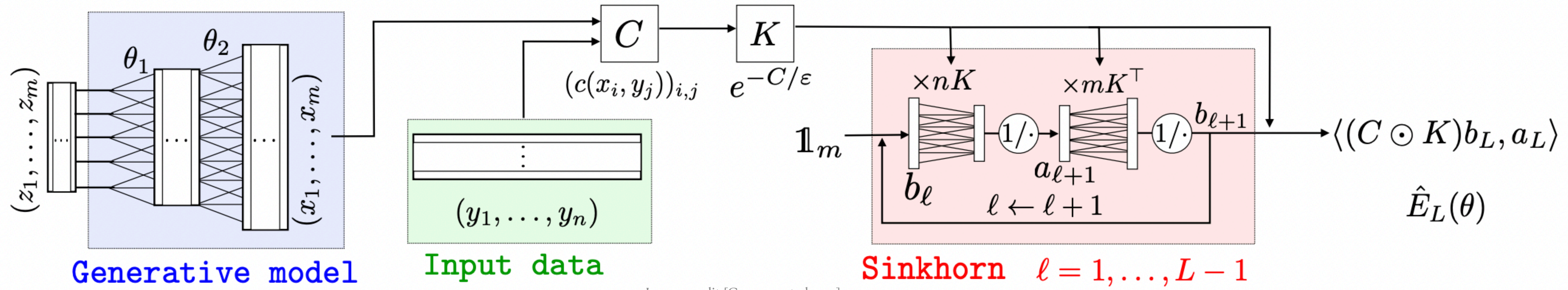


Image credit:[Genevay et al 2017]

Extensions to OT

Unbalanced Optimal Transport

- ($\mu(\Omega_s) = \nu(\Omega_t)$) no longer holds true?
- Modify the OT problem into a variational formulation - adding infinite sources/sinks, mass creation
- [Matthias et al 2016] Given two measures $\mu \in M_+(\Omega_s), \nu \in M_+(\Omega_t)$,
  - Choose $0 < m \leq \min\{\mu(\Omega_s), \nu(\Omega_t)\}$
  - Define $\gamma_t = \int_{\Omega_t} \gamma(\mathbf{x}, \mathbf{y}) d\mathbf{y}, \gamma_s = \int_{\Omega_s} \gamma(\mathbf{x}, \mathbf{y}) d\mathbf{x}$ and solve

$$\min_{\gamma \in \mathscr{M}_+(\Omega_s \times \Omega_t)} \int c(x, y) d\gamma(x, y) \quad \text{subject to } \gamma_t \leq \mu, \gamma_s \leq \nu, \gamma(\Omega_s \times \Omega_t) = m$$

- Generalise the Wasserstein distance to this setting with the **Wasserstein Fisher-Rao distance**

$$\widehat{W}_2^2(\mu, \nu) = \min_{\gamma \in M_+(\Omega_s \times \Omega_t)} KL\left(\gamma_t \mid \mu\right) + KL\left(\gamma_s \mid \nu\right) + \int c_\ell(x, y) d\gamma(x, y)$$

- [Peyre et al., 2017] General algorithm using entropic regularised WFR with Sinkhorn iterations

OT between different metric spaces

- Can you perform OT between two spaces without $c(x, y)$ or when $\dim(\Omega_s) \neq \dim(\Omega_t)$?
- Extending OT metrics to measures with no common ground space
- [Memoli, 2011] proposed Gromov-Wasserstein distance

$$\mathcal{GW}_p(\mu, \nu) = \left(\min_{\gamma \in \mathcal{P}(\mu, \nu)} \mathcal{L}(D_{i,k}, D'_{j,l}) \times \gamma_{i,j} \times \gamma_{k,l} \right)^{\frac{1}{p}}$$

with $D_{i,k} = \| \mathbf{x}_i^s - \mathbf{x}_k^s \|$, $D'_{j,l} = \| \mathbf{x}_j^t - \mathbf{x}_l^t \|$, $\mathcal{L}(D_{i,k}, D'_{j,l})$ is a dissimilarity metric b/w distances

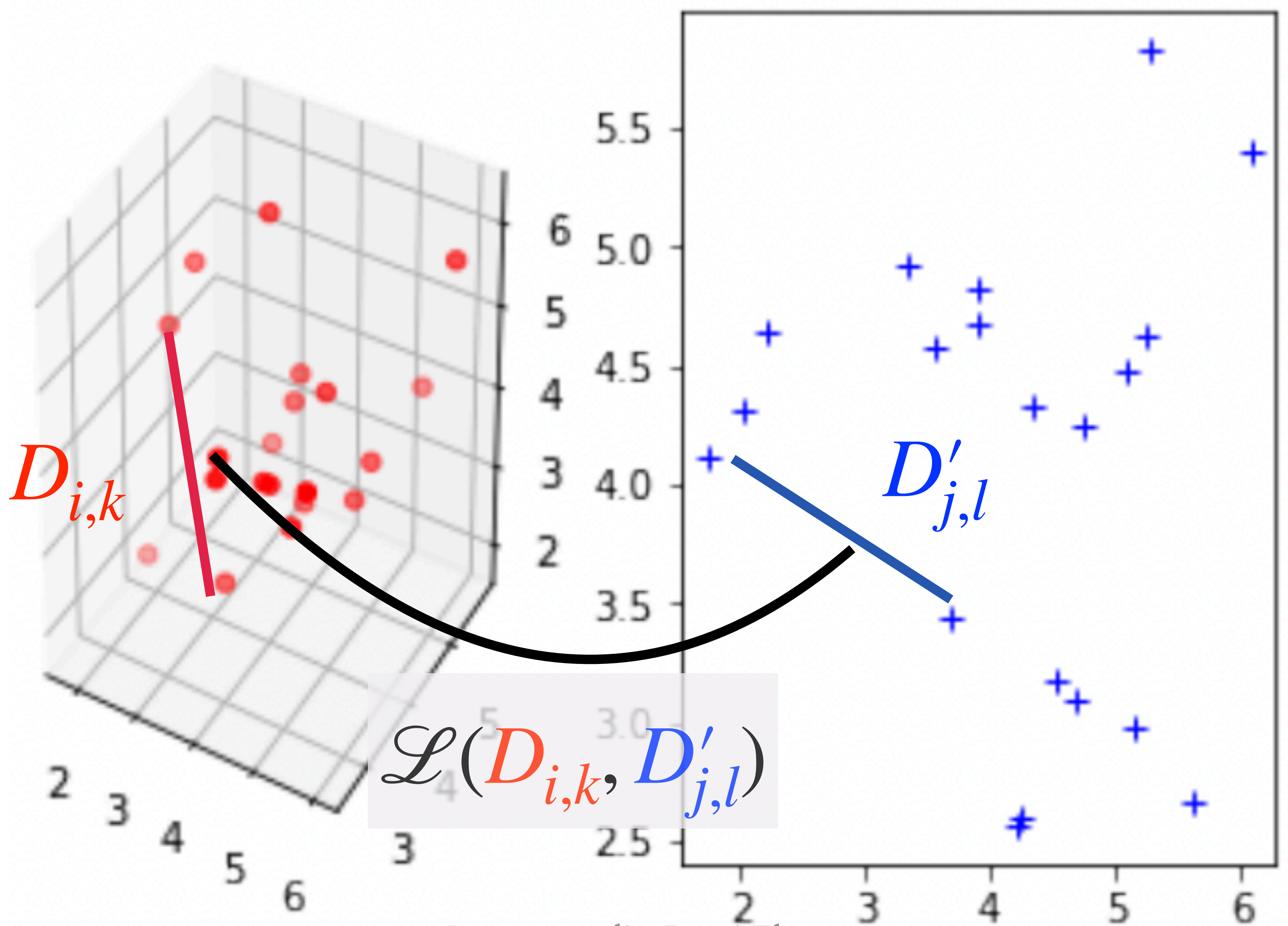


Image credit: Remi Flamary

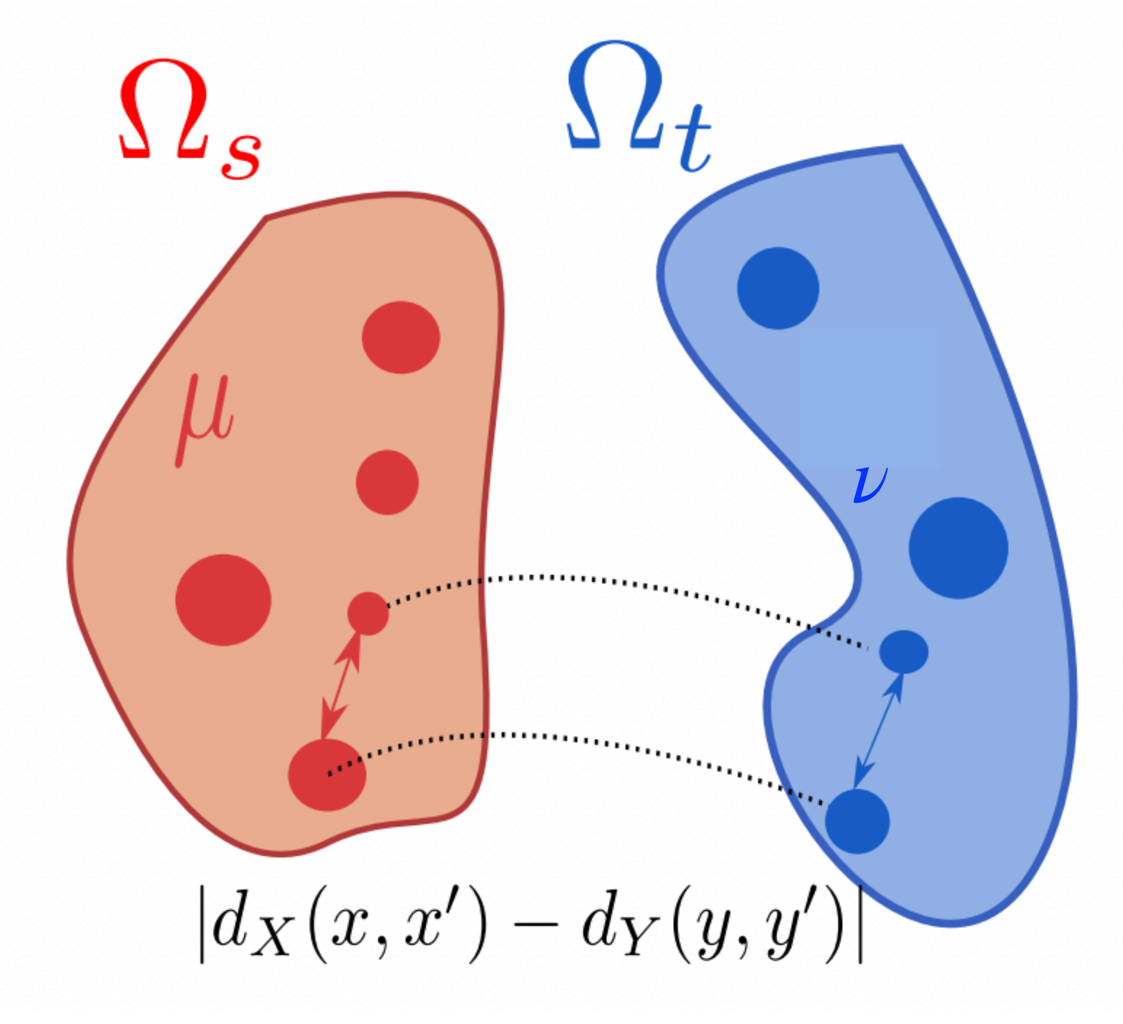


Image credit: Remi Flamary

OT between different metric spaces

- This is a Quadratic Program - Nonconvex, NP-hard
- [Peyre et al., 2016] proposed an entropic regularisation relaxation of this problem

$$\mathscr{GW}_\lambda(\mu,\nu) = \left(\min_{\gamma \in \mathscr{P}(\mu,\nu)} \mathscr{L}(D_{i,k}, D'_{j,l}) \times \gamma_{i,j} \times \gamma_{k,l} \right) - \lambda \mathbb{H}(\gamma)$$

  - This regularised term can be solved using projected gradient descent/Sinkhorn's algorithm

$$\gamma^{k+1} \leftarrow \operatorname*{argmin}_{\gamma^k \in \mathscr{P}} \left\langle \gamma, \mathscr{L}(D_{i,k}, D'_{j,l}) \otimes \gamma^k \right\rangle - \lambda H(\gamma)$$

    - Where $\mathbf{K}' = \mathscr{L}(D_{i,k}, D'_{j,l}) \otimes \gamma^k$, the tensor product where $\mathscr{L}(D_{i,k}, D'_{j,l}) \otimes \gamma^k = \left(\mathscr{L}(D_{i,k}, D'_{j,l}) \gamma_{k,l} \right)_{i,j}$

  - Sinkhorn's algorithm returns a stationary point of the nonconvex optimisation problem

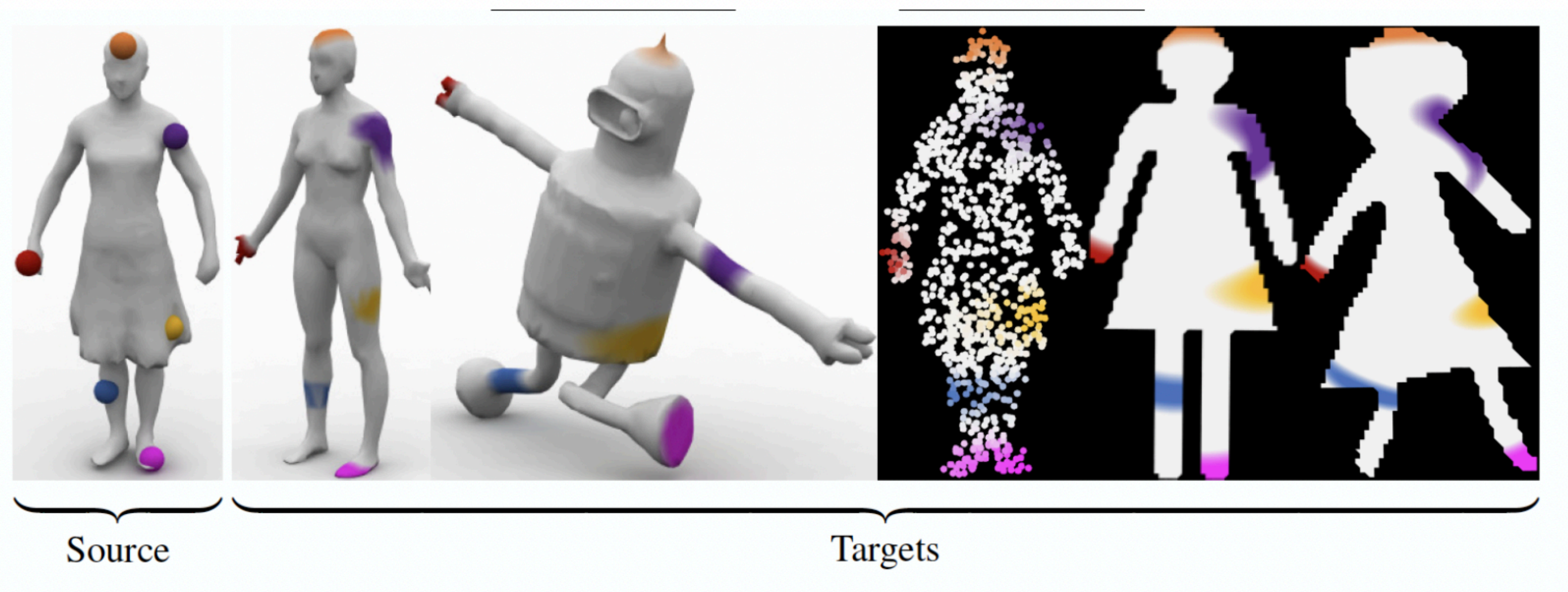


Image credit: [Peyre et al 2016]

Conclusions

- Optimal Transport Theory provides a rigorous and rich mathematical formulation for defining metrics/discrepancy measures between probability measures
- In practise, cheap and efficient approximations have been developed recently
- Applications in generative modelling, supervised learning, computer vision and graphics
- Other cool research to read about
  - [Blanchet et al., 2021] Distributionally Robust Optimisation
  - [Durmus et. Al, 2019] Convergence of Langevin Dynamics Monte Carlo in Wasserstein geometry
  - [Kolouri et al., 2020] Optimal Transport on graphs and arbitrary manifolds through Wasserstein embeddings
  - [Courty et al., 2015] Domain Adaptation with Optimal Transport
  - [Craig et al., 2017] Wasserstein Gradient Flows

References I

- Peyré, Gabriel, and Marco Cuturi. "Computational optimal transport: With applications to data science." Foundations and Trends® in Machine Learning 11.5-6 (2019): 355-607.
- Kolouri, Soheil, et al. "Optimal mass transport: Signal processing and machine-learning applications." IEEE signal processing magazine 34.4 (2017): 43-59.
- Solomon, Justin, et al. "Convolutional wasserstein distances: Efficient optimal transportation on geometric domains." ACM Transactions on Graphics (ToG) 34.4 (2015): 1-11.
- Kolouri, Soheil, et al. "Sliced-wasserstein autoencoder: An embarrassingly simple generative model." arXiv preprint arXiv:1804.01947 (2018).
- Nadjahi, Kimia, et al. "Statistical and topological properties of sliced probability divergences." *Advances in Neural Information Processing Systems* 33 (2020): 20802-20812.
- Kolouri, Soheil, et al. "Generalized sliced wasserstein distances." Advances in Neural Information Processing Systems 32 (2019).
- Peyré, Gabriel, Marco Cuturi, and Justin Solomon. "Gromov-Wasserstein averaging of kernel and distance matrices." International Conference on Machine Learning. PMLR, 2016.

References II

- Makkuva, Ashok, et al. "Optimal transport mapping via input convex neural networks." International Conference on Machine Learning. PMLR, 2020.
- Shirdhonkar, Sameer, and David W. Jacobs. "Approximate earth mover's distance in linear time." 2008 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2008.
- Cuturi, Marco. "Sinkhorn distances: Lightspeed computation of optimal transport." Advances in neural information processing systems 26 (2013).
- Wilson, Alan Geoffrey. "The use of entropy maximising models, in the theory of trip distribution, mode split and route split." Journal of transport economics and policy (1969): 108-126.
- Sinkhorn, Richard. "Diagonal equivalence to matrices with prescribed row and column sums." The American Mathematical Monthly 74.4 (1967): 402-405.
- Ramdas, Aaditya, Nicolás García Trillos, and Marco Cuturi. "On wasserstein two-sample testing and related families of nonparametric tests." Entropy 19.2 (2017): 47.
- Frogner, Charlie, et al. "Learning with a Wasserstein loss." Advances in neural information processing systems 28 (2015).

References III

- Arjovsky, Martin, Soumith Chintala, and Léon Bottou. "Wasserstein generative adversarial networks." International conference on machine learning. PMLR, 2017.
- Gulrajani, Ishaan, et al. "Improved training of wasserstein gans." Advances in neural information processing systems 30 (2017).
- Genevay, Aude, Gabriel Peyré, and Marco Cuturi. "Learning generative models with sinkhorn divergences." International Conference on Artificial Intelligence and Statistics. PMLR, 2018.
- Mémoli, Facundo. "Gromov–Wasserstein distances and the metric approach to object matching." Foundations of computational mathematics 11.4 (2011): 417-487.
- Peyré, Gabriel, Marco Cuturi, and Justin Solomon. "Gromov-Wasserstein averaging of kernel and distance matrices." International Conference on Machine Learning. PMLR, 2016.
- Liero, Matthias, Alexander Mielke, and Giuseppe Savaré. "Optimal transport in competition with reaction: The Hellinger--Kantorovich distance and geodesic curves." SIAM Journal on Mathematical Analysis 48.4 (2016): 2869-2911.
- Chizat, Lenaic, et al. "Unbalanced optimal transport: geometry and Kantorovich formulation." (2015).

References IV

- Blanchet, Jose, Karthyek Murthy, and Fan Zhang. "Optimal Transport-Based Distributionally Robust Optimization: Structural Properties and Iterative Schemes." Mathematics of Operations Research (2021).
- Durmus, Alain, Szymon Majewski, and Błażej Miasojedow. "Analysis of Langevin Monte Carlo via convex optimization." The Journal of Machine Learning Research 20.1 (2019): 2666-2711.
- Kolouri, Soheil, et al. "Wasserstein embedding for graph learning." arXiv preprint arXiv:2006.09430 (2020).
- Courty, Nicolas, et al. "Optimal transport for domain adaptation. CoRR." arXiv preprint arXiv:1507.00504 (2015).
- Craig, Katy. "Nonconvex gradient flow in the Wasserstein metric and applications to constrained nonlocal interactions." Proceedings of the London Mathematical Society 114.1 (2017): 60-102.
- Remi Flamary, Optimal Transport for Machine Learning tutorial [https://remi.flamary.com/cours/otml/OTML_ISBI_2019.pdf]

References V

- Lénaïc Chizat, Tutorial on Optimal Transport with a Machine Learning Touch, IISc Bangalore, 2019, [https://lchizat.github.io/files/presentations/chizat2019IFCAM_OT.pdf]
- Marco Cuturi, MLSS South Africa, A Primer on Optimal Transport, 2019, [https://www.dropbox.com/s/wlxvbxs4r5zbr77/mlss19stellenbosch.pdf?dl=0]
- Gabriel Peyre, Ecole Normale Superieure, Optimal Transport for Machine Learning, [https://www.youtube.com/watch?v=mITml5ZpqM8]