Optimal transport for machine learning

Introduction to optimal transport

Rémi Flamary¹, Nicolas Courty²,

April 8 2019
Tutorial ISBI 2019, Venice, Italy

¹ Université Côte d’Azur, OCA, Lagrange, CNRS
² IRISA, University of Bretagne Sud, France

http://tinyurl.com/otml-isbi
Overview of the tutorial

Part 1: Introduction to optimal transport (≈1:30)
- Optimal transport problem
- Wasserstein distance and geometry
- Computational aspects and regularized OT

Part 2: Learning with optimal transport (≈1:30)
- Learning to map with OT
- Learning from histograms
- Learning from empirical distributions
Table of content (Part 1)

**Optimal transport**

- Monge and Kantorovitch
- OT on discrete distributions
- Wasserstein distances
- Barycenters and geometry of optimal transport

**Computational aspects of optimal transport**

- Special cases
- Regularized optimal transport
- Minimizing the Wasserstein distance

**Gromov-Wasserstein**
Optimal transport
What is optimal transport?

The natural geometry of probability measures

Monge, Kantorovich, Koopmans, Dantzig, Brenier, Otto, McCann, Villani, Figalli

Nobel '75, Fields '10, Fields '18
The origins of optimal transport

Problem [Monge, 1781]

- How to move dirt from one place (déblais) to another (remblais) while minimizing the effort?
- Find a mapping $T$ between the two distributions of mass (transport).
- Optimize with respect to a displacement cost $c(x, y)$ (optimal).
The origins of optimal transport

Problem [Monge, 1781]

- How to move dirt from one place (déblais) to another (remblais) while minimizing the effort?
- Find a mapping $T$ between the two distributions of mass (transport).
- Optimize with respect to a displacement cost $c(x, y)$ (optimal).
Optimal transport (Monge formulation)

- Probability measures $\mu_s$ and $\mu_t$ on and a cost function $c : \Omega_s \times \Omega_t \rightarrow \mathbb{R}^+$.
- The Monge formulation [Monge, 1781] aim at finding a mapping $T : \Omega_s \rightarrow \Omega_t$

$$\inf_{T \# \mu_s = \mu_t} \int_{\Omega_s} c(x, T(x)) \mu_s(x) dx$$

(1)

- Non convex problem because of the constraint $T \# \mu_s = \mu_t$. 
What is $T^#\mu_s = \mu_t$?

**Pushforward operator** $T^#$

- Transfers measures from one space $\Omega_s$ to another space $\Omega_t$

  $$\mu_t(A) = \mu_s(T^{-1}(A)), \quad \forall \text{ Borel subset } A \in \Omega_s$$

- For smooth measures $\mu_s = \rho(x)dx$ and $\mu_t = \eta(x)dx$

  $$T^#\mu_s = \mu_t \equiv \rho(T(x))|\det(\partial T(x))| = \eta(x)$$

  a.k.a. change of variable formula
Properties of mapping $T$

Non-existence / Non-uniqueness

- $T\#\mu_s = \mu_t$ is a non-convex constraint.
- Existence of $T$ is not guaranteed.
- Unicity of $T$ is not guaranteed.
- [Brenier, 1991] proved existence and unicity of the Monge map for $c(x, y) = \|x - y\|^2$ and distributions with densities (i.e. continuous).
Kantorovich relaxation

Leonid Kantorovich (1912–1986), Economy nobelist in 1975

Focus on where the mass goes, allow splitting [Kantorovich, 1942].

Applications mainly for resource allocation problems
The Kantorovich formulation [Kantorovich, 1942] seeks for a probabilistic coupling \( \gamma \in \mathcal{P}(\Omega_s \times \Omega_t) \) between \( \Omega_s \) and \( \Omega_t \):

\[
\gamma_0 = \arg\min_{\gamma} \int_{\Omega_s \times \Omega_t} c(x, y) \gamma(x, y) \, dx \, dy,
\]

\[\text{s.t. } \gamma \in \mathcal{P} = \left\{ \gamma \geq 0, \int_{\Omega_t} \gamma(x, y) \, dy = \mu_s, \int_{\Omega_s} \gamma(x, y) \, dx = \mu_t \right\}\]

- \( \gamma \) is a joint probability measure with marginals \( \mu_s \) and \( \mu_t \).
- Linear Program that always has a solution.
The Kantorovich formulation [Kantorovich, 1942] seeks for a probabilistic coupling $\gamma \in \mathcal{P}(\Omega_s \times \Omega_t)$ between $\Omega_s$ and $\Omega_t$:

$$
\gamma_0 = \arg\min_{\gamma} \int_{\Omega_s \times \Omega_t} c(x, y) \gamma(x, y) \, dx \, dy,
$$

s.t. $\gamma \in \mathcal{P} = \left\{ \gamma \geq 0, \int_{\Omega_t} \gamma(x, y) \, dy = \mu_s, \int_{\Omega_s} \gamma(x, y) \, dx = \mu_t \right\}$

- $\gamma$ is a joint probability measure with marginals $\mu_s$ and $\mu_t$.
- Linear Program that always has a solution.
Couplings for 1D distributions

![Diagrams showing coupling schemes for 1D distributions. Each diagram illustrates different coupling strategies between two distributions α and β, with intermediate points labeled π.]
Optimal transport (Kantorovich dual formulation)

Joint distribution optimal \( \gamma(x, y) \)

Source \( \mu_s(x) \)
Target \( \mu_t(y) \)

Transport cost \( c(x, y) = |x - y|^2 \)

**Dual formulation of the OT linear program**

\[
\max_{\phi, \psi} \left\{ \int \phi \, d\mu_s + \int \psi \, d\mu_t \mid \phi(x) + \psi(y) \leq c(x, y) \right\}
\]  

(3)

- \( \phi \) and \( \psi \) are scalar function also known as Kantorovich potentials.
- Equivalent problem by the Rockafellar-Fenchel theorem.
- Objective value separable wrt \( \mu_s \) and \( \mu_t \).
- Primal-dual relation: the support of \( \gamma(x, y) \) is where \( \phi(x) + \psi(y) = c(x, y) \)
Optimal transport (Kantorovich dual formulation)

Joint distribution optimal $\gamma(x, y)$

Source $\mu_s(x)$
Potential $\phi(x)$
Target $\mu_t(y)$
Potential $\psi(y)$
$\gamma(x, y)$

Transport cost $c(x, y) = |x - y|^2$ and dual constraint

Dual formulation of the OT linear program

$$\max_{\phi, \psi} \left\{ \int \phi d\mu_s + \int \psi d\mu_t \mid \phi(x) + \psi(y) \leq c(x, y) \right\}$$

\begin{itemize}
  \item $\phi$ and $\psi$ are scalar function also known as Kantorovich potentials.
  \item Equivalent problem by the Rockafellar-Fenchel theorem.
  \item Objective value separable wrt $\mu_s$ and $\mu_t$.
  \item Primal-dual relation: the support of $\gamma(x, y)$ is where $\phi(x) + \psi(y) = c(x, y)$
\end{itemize}
The linear dual constraint suggest that there exits an optimal $\psi$ for a given $\phi$.

**c-transform (or c-conjugate)**

\[
\phi^c(y) \overset{\text{def}}{=} H^c(\phi) = \inf_x c(x, y) - \phi(x)
\]  

(4)

Similar a Legendre transform (equal when $c(x, y) = x^\top y$).

**Semi-dual formulation**

\[
\max_\phi \left\{ \int \phi d\mu_s + \int \phi^c d\mu_t \right\}
\]  

(5)

- Depends only on one dual potential through the c-transform.
- Nice reformulation when $H^c$ is easy to compute of close form.
- Special case when $c(x, y) = \|x - y\|$.
Case \( c(x, y) = \|x - y\| \) (a.k.a \( W_1^1 \))

**Case\( c(x, y) = \|x - y\| \)**

- Existence of a solution but not unique.
- For any \( \phi \in \text{Lip}^1 \) (set of 1-Lipschitz functions), we have \( \phi^c(x) = -\phi(x) \).
- The dual OT problem can be reformulated as

\[
\sup_{\phi \in \text{Lip}^1} \int \phi d(\mu_s - \mu_t) = \sup_{\phi \in \text{Lip}^1} \mathbb{E}_{x \sim \mu_s} [\phi(x)] - \mathbb{E}_{y \sim \mu_t} [\phi(y)]
\]  

(6)

- Also known as Kantorovich-Rubinstein duality
- Formulation used for Wasserstein GAN (more details in next part).
Case $c(x, y) = \|x - y\|^2 / 2$ (a.k.a $W_2^2$)

Joint distribution optimal $\gamma(x, y)$

Transport cost $c(x, y) = |x - y|^2$ and dual constraint

Strict equality $c(x, y) = (x) + (y)$

Case $c(x, y) = \|x - y\|^2 / 2$

- When $\mu_s$ and $\mu_t$ are continuous, $T(x)$ the OT mapping exists and is unique.
- More remarkably, it is a gradient of a convex functions $\Phi(x)$

$$T(x) = x - \nabla \phi(x) = \nabla \left( \frac{\|x\|^2}{2} - \phi(x) \right) = \nabla (\Phi(x))$$

- This is also known as Brenier’s Theorem [Brenier, 1991].
Discrete distributions: Empirical vs Histogram

Discrete measure: \[ \mu = \sum_{i=1}^{n} a_i \delta_{x_i}, \quad x_i \in \Omega, \quad \sum_{i=1}^{n} a_i = 1 \]

Lagrangian (point clouds)

- Constant weight: \( a_i = \frac{1}{n} \)
- Quotient space: \( \Omega^n, \Sigma_n \)

Eulerian (histograms)

- Fixed positions \( x_i \) e.g. grid
- Convex polytope \( \Sigma_n \) (simplex): \( \{(a_i)_i \geq 0; \sum_i a_i = 1\} \)
Optimal transport with discrete distributions

**Distributions**

Source $\mu_s$  
Target $\mu_t$

**Matrix $C$**

**OT matrix $\gamma$**

**OT Linear Program**

When $\mu_s = \sum_{i=1}^{n} a_i \delta_{x_i^s}$ and $\mu_t = \sum_{i=1}^{n} b_i \delta_{x_i^t}$

$$\gamma_0 = \arg\min_{\gamma \in \mathcal{P}} \left\{ \langle \gamma, C \rangle_F = \sum_{i,j} \gamma_{i,j} c_{i,j} \right\}$$

where $C$ is a cost matrix with $c_{i,j} = c(x_i^s, x_j^t)$ and the marginals constraints are

$$\mathcal{P} = \left\{ \gamma \in (\mathbb{R}^+)^{n_s \times n_t} \mid \gamma 1_{n_t} = a, \gamma^T 1_{n_s} = b \right\}$$

Linear program with $n_s n_t$ variables and $n_s + n_t$ constraints. Demo
Optimal transport with discrete distributions

OT Linear Program
When \( \mu_s = \sum_{i=1}^{n_s} a_i \delta_{x_i^s} \) and \( \mu_t = \sum_{i=1}^{n_t} b_i \delta_{x_i^t} \)

\[
\gamma_0 = \arg\min_{\gamma \in \mathcal{P}} \left\{ \langle \gamma, C \rangle_F = \sum_{i,j} \gamma_{i,j} c_{i,j} \right\}
\]

where \( C \) is a cost matrix with \( c_{i,j} = c(x_i^s, x_j^t) \) and the marginals constraints are

\[
\mathcal{P} = \left\{ \gamma \in (\mathbb{R}^+)^{n_s \times n_t} \mid \gamma 1_{n_t} = a, \gamma^T 1_{n_s} = b \right\}
\]

Linear program with \( n_s n_t \) variables and \( n_s + n_t \) constraints. Demo.
Optimal transport with discrete distributions

Distributions

Source $\mu_s$
Target $\mu_t$

Matrix $C$

OT matrix with samples

OT Linear Program

When $\mu_s = \sum_{i=1}^{n} a_i \delta_{x_i^s}$ and $\mu_t = \sum_{i=1}^{n} b_i \delta_{x_i^t}$

$$\gamma_0 = \arg\min_{\gamma \in \mathcal{P}} \left\{ \langle \gamma, C \rangle_F = \sum_{i,j} \gamma_{i,j} c_{i,j} \right\}$$

where $C$ is a cost matrix with $c_{i,j} = c(x_i^s, x_j^t)$ and the marginals constraints are

$$\mathcal{P} = \left\{ \gamma \in (\mathbb{R}^+)^{n_s \times n_t} | \gamma 1_{n_t} = a, \gamma^T 1_{n_s} = b \right\}$$

Linear program with $n_sn_t$ variables and $n_s + n_t$ constraints. Demo
Optimal transport with discrete distributions

- $\mathcal{P}$ is the Birkhoff polytope (for uniform weights).
- No unique solution in some cases, numerical instabilities
- OT loss not differentiable!
OT Dual for discrete distributions

Discrete OT dual formulation

\[
\begin{align*}
\max_{\alpha \in \mathbb{R}^{n_s}, \beta \in \mathbb{R}^{n_t}} & \quad \alpha^T a + \beta^T b \\
\text{s.t.} & \quad \alpha_i + \beta_j \leq c_{i,j} \quad \forall i, j
\end{align*}
\] 

(8) \hspace{1cm} (9)

- With \( \mu_s = \sum_{i=1}^{n} a_i \delta_{x_i}^s \) and \( \mu_t = \sum_{i=1}^{n} b_i \delta_{x_i}^t \)
- Linear program with \( n_s + n_t \) variables and \( n_s n_t \) constraints.
- Solved with Network Flow solver of complexity \( O(n^3 \log(n)) \) with \( n = \max(n_s, n_t) \).
Matching words embedding

Word mover’s distance [Kusner et al., 2015]

- Words embedded in a high-dimensional space with neural networks.
- Matching two documents is an OT problem, with the cost being the $l_2$ distance in the embedded space.
- Small value of the objective means similar documents.
- OT matrix provide interpretability (word correspondence).
Wasserstein distance

\[ W_p^p(\mu_s, \mu_t) = \min_{\gamma \in \mathcal{P}} \int_{\Omega_s \times \Omega_t} \|x - y\|^p \gamma(x, y) dx dy = \mathbb{E}_{(x, y) \sim \gamma} [\|x - y\|^p] \quad (10) \]

In this case we have \( c(x, y) = \|x - y\|^p \)

- A.K.A. Earth Mover’s Distance (\( W_1^1 \)) [Rubner et al., 2000].
- Do not need the distribution to have overlapping support.
- Works for continuous and discrete distributions (histograms, empirical).
Earth Mover’s Distance (EMD)

EMD for image retrieval [Rubner et al., 2000]

- Represent images as histograms.
- Color histogram measure de color proportion
- Histogram of gradient encode texture.
- FastEMD [Pele and Werman, 2009] is a fast approximation.
Wasserstein barycenter

Barycenters [Agueh and Carlier, 2011]

\[ \bar{\mu} = \arg \min_{\mu} \sum_{i}^{n} \lambda_i W_p^{\infty}(\mu_i, \mu) \]

- \( \lambda_i > 0 \) and \( \sum_{i}^{n} \lambda_i = 1 \).
- Uniform barycenter has \( \lambda_i = \frac{1}{n}, \forall i \).
- Interpolation with \( n=2 \) and \( \lambda = [1 - t, t] \) with \( 0 \leq t \leq 1 \) [McCann, 1997].
- Regularized barycenters using Bregman projections [Benamou et al., 2015].
- The cost and regularization impacts the interpolation trajectory.
Wasserstein barycenter

Barycenters [Agueh and Carlier, 2011]

\[ \tilde{\mu} = \arg \min_{\mu} \sum_{i}^{n} \lambda_{i} W_{p}^{\mu}(\mu_{i}, \mu) \]

- \( \lambda_{i} > 0 \) and \( \sum_{i}^{n} \lambda_{i} = 1 \).
- Uniform barycenter has \( \lambda_{i} = \frac{1}{n} \), \( \forall i \).
- Interpolation with \( n=2 \) and \( \lambda = [1 - t, t] \) with \( 0 \leq t \leq 1 \) [McCann, 1997].
- Regularized barycenters using Bregman projections [Benamou et al., 2015].
- The cost and regularization impacts the interpolation trajectory.
The space of probability distribution equipped with the Wasserstein metric \( (\mathcal{P}_p(X), W^2_2(X)) \) defines a geodesic space with a Riemannian structure \cite{Santambrogio,2014}.

- Geodesics are shortest curves on \( \mathcal{P}_p(X) \) that link two distributions.

Illustration from \cite{Kolouri,2017} and maze example from \cite{Papadakis,2014}.
Wasserstein space

- The space of probability distribution equipped with the Wasserstein metric $(\mathcal{P}_p(X), W_2^2(X))$ defines a geodesic space with a Riemannian structure [Santambrogio, 2014].
- Geodesics are shortest curves on $\mathcal{P}_p(X)$ that link two distributions.
- Cost between two pixels is the shortest path in the maze (Riemannian metric).

Illustration from [Kolouri et al., 2017] and maze example from [Papadakis et al., 2014]
3D Wasserstein barycenter

Shape interpolation [Solomon et al., 2015]
Wasserstein averaging of fMRI

OT averaging of neurological data [Gramfort et al., 2015]

- Average fMRI activation maps on voxels or cortical surface (natural metric).
- Classical average across subjects and gaussian blur loose information.
- OT averaging recover central activation areas with better precision.
- Can encode both geometrical (3D position) or anatomical connectivity information.
- Extension using OT-Lp seems more robust to noise [Wang et al., 2018].
Optimal transport

Monge and Kantorovitch

OT on discrete distributions

Wasserstein distances

Barycenters and geometry of optimal transport

Computational aspects of optimal transport

Special cases

Regularized optimal transport

Minimizing the Wasserstein distance

Gromov-Wasserstein
Special case: OT in 1D

- When $c(x, y)$ is a strictly convex and increasing function of $|x - y|$.
- If $x_1 < x_2$ and $y_1 < y_2$, we have $c(x_1, y_1) + c(x_2, y_2) < c(x_1, y_2) + c(x_2, y_1)$
- The OT plan respects the ordering of the elements.
- Solution is given by the monotone rearrangement of $\mu_1$ onto $\mu_2$.
- Simple algorithm for discrete distribution by sorting $O(N \log N)$.
Special case: OT in 1D

- When \( c(x, y) \) is a strictly convex and increasing function of \(|x - y|\).
- If \( x_1 < x_2 \) and \( y_1 < y_2 \), we have \( c(x_1, y_1) + c(x_2, y_2) < c(x_1, y_2) + c(x_2, y_1) \).
- The OT plan respects the ordering of the elements.
- Solution is given by the monotone rearrangement of \( \mu_1 \) onto \( \mu_2 \).
- Simple algorithm for discrete distribution by sorting \( O(N \log N) \).
Special case: OT in 1D

Illustration with cumulative distributions

- $F_\mu$ cumulative distribution function of $\mu$: $F_\mu(t) = \mu(-\infty, t]$.
- $F_\mu^{-1}(q), q \in [0, 1]$ is the quantile function: $F_\mu^{-1}(q) = \inf\{x \in \mathbb{R} : F_\mu(x) \geq q\}$.
- The value of the $W_1$ Wasserstein distance

$$W_1(\mu_s, \mu_t) = \int_0^1 c(F_{\mu_s}^{-1}(q), F_{\mu_t}^{-1}(q))dq$$

- Very fast $O(n \log(n))$ computation on discrete distributions.
Illustration with cumulative distributions

- $F_\mu$ cumulative distribution function of $\mu$: $F_\mu(t) = \mu(-\infty, t]$.
- $F_\mu^{-1}(q), \ q \in [0, 1]$ is the quantile function: $F_\mu^{-1}(q) = \inf\{x \in \mathbb{R} : F_\mu(x) \geq q\}$.
- The value of the $W_1$ Wasserstein distance
  
  $$W_1(\mu_s, \mu_t) = \int_0^1 c(F_\mu^{-1}(q), F_\mu^{-1}(q))dq$$

- Very fast $O(n \log(n))$ computation on discrete distributions.
Sliced Radon Wasserstein

\[ pSW^p_\mu (\mu_s, \mu_t) = \int_{S^{d-1}} W^p_\mu (R(\mu_s, \theta), R(\mu_t, \theta)) d\theta \]

where \( R \) is the Radon transform \( R(\mu, \theta) = \int_{S^{d-1}} \mu(x) \delta(t - \theta^\top x) dx \) \( \forall \theta \in S^{d-1} \)

- Can be approximated by discrete sampling of the directions \( \theta \).
- Fast 1D wasserstein on 1D projections when \( d > 1 \), fast distance and barycenter computation.
- p-sliced Wasserstein distance used for kernel learning between distributions [Kolouri et al., 2016].
Wasserstein between Gaussian distributions

- $\mu_s \sim \mathcal{N}(m_1, \Sigma_1)$ and $\mu_t \sim \mathcal{N}(m_2, \Sigma_2)$
- Wasserstein distance with $c(x, y) = \|x - y\|_2^2$ reduces to:

$$W_2^2(\mu_s, \mu_t) = \|m_1 - m_2\|_2^2 + B(\Sigma_1, \Sigma_2)^2$$

where $B(\cdot, \cdot)$ is the so-called Bures metric:

$$B(\Sigma_1, \Sigma_2)^2 = \text{trace}(\Sigma_1 + \Sigma_2 - 2(\Sigma_1^{1/2} \Sigma_2 \Sigma_1^{1/2})^{1/2})$$
OT mapping between Gaussian distributions

- $\mu_s \sim \mathcal{N}(m_1, \Sigma_1)$ and $\mu_t \sim \mathcal{N}(m_2, \Sigma_2)$
- The optimal map $T$ for $c(x, y) = \|x - y\|_2^2$ is given by

$$T(x) = m_2 + A(x - m_1)$$

with

$$A = \Sigma_1^{-1/2} \left( \Sigma_1^{1/2} \Sigma_2 \Sigma_1^{1/2} \right)^{1/2} \Sigma_1^{-1/2}$$
Regularized optimal transport

\[ \gamma_0^\lambda = \arg\min_{\gamma \in P} \langle \gamma, C \rangle_F + \lambda \Omega(\gamma), \quad (11) \]

**Regularization term** \( \Omega(\gamma) \)

- Entropic regularization [Cuturi, 2013].
- Group Lasso [Courty et al., 2016a].
- KL, Itakura Saito, \( \beta \)-divergences, [Dessein et al., 2016].

**Why regularize?**

- Smooth the “distance” estimation:
  \[ W_\lambda(\mu_s, \mu_t) = \langle \gamma_0^\lambda, C \rangle_F \]
- Encode prior knowledge on the data.
- Better posed problem (convex, stability).
- Fast algorithms to solve the OT problem.
Entropic regularized optimal transport

Distributions
Source $\mu_s$
Target $\mu_t$

Reg. OT matrix with $\lambda=1e-3$

Reg. OT matrix with $\lambda=1e-2$

Entropic regularization [Cuturi, 2013]

$$\gamma^\lambda_0 = \arg\min_{\gamma \in \mathcal{P}} \langle \gamma, C \rangle_F + \lambda \sum_{i,j} \gamma(i,j)(\log \gamma(i,j) - 1)$$

- Regularization with the negative entropy of $\gamma$.
- Loose sparsity, gains stability.
- Strictly convex optimization problem.
- Loss and OT matrix are differentiable.
Entropic regularized optimal transport distributions

\[ \gamma_0^\lambda = \arg\min_{\gamma \in \mathcal{P}} \langle \gamma, C \rangle_F + \lambda \sum_{i,j} \gamma(i, j)(\log \gamma(i, j) - 1) \]

- Regularization with the negative entropy of \( \gamma \).
- Looses sparsity, gains stability.
- Strictly convex optimization problem.
- Loss and OT matrix are differentiable.
Solving the entropy regularized problem

Lagrangian of the optimization problem

\[ L(\gamma, \alpha, \beta) = \sum_{ij} \gamma_{ij} C_{ij} + \lambda \gamma_{ij} (\log \gamma_{ij} - 1) + \alpha^T (\gamma 1_{nt} - a) + \beta^T (\gamma^T 1_{ns} - b) \]

\[ \frac{\partial L(\gamma, \alpha, \beta)}{\partial \gamma_{ij}} = C_{ij} + \lambda \log \gamma_{ij} + \alpha_i + \beta_j \]

\[ \frac{\partial L(\gamma, \alpha, \beta)}{\partial \gamma_{ij}} = 0 \implies \gamma_{ij} = \exp\left(\frac{\alpha_i}{\lambda}\right) \exp\left(-\frac{C_{ij}}{\lambda}\right) \exp\left(\frac{\beta_j}{\lambda}\right) \]

Entropy-regularized transport

The solution of entropy regularized optimal transport problem is of the form

\[ \gamma^\lambda = \text{diag}(u) \exp(-C/\lambda) \text{diag}(v) \]

- Through the **Sinkhorn theorem** \( \text{diag}(u) \) and \( \text{diag}(v) \) exist and are unique.
- Relation with dual variables: \( u_i = \exp(\alpha_i/\lambda), \quad v_j = \exp(\beta_j/\lambda) \).
- Can be solved by the **Sinkhorn-Knopp** algorithm.
Algorithm 1 Sinkhorn-Knopp Algorithm (SK).

Require: $a, b, C, \lambda$

$u^{(0)} = 1, K = \exp(-C/\lambda)$

for $i$ in $1, \ldots, n_{it}$ do

$v^{(i)} = b \odot K^T u^{(i-1)}$ // Update right scaling

$u^{(i)} = a \odot Kv^{(i)}$ // Update left scaling

end for

return $T = \text{diag}(u^{(n_{it})})K\text{diag}(v^{(n_{it})})$

- The algorithm performs alternatively a scaling along the rows and columns of $K = \exp(-C/\lambda)$ to match the desired marginals.
- Complexity $O(kn^2)$, where $k$ iterations are required to reach convergence.
- Fast implementation in parallel, GPU friendly.
- Convolutive/Heat structure for $K$ [Solomon et al., 2015]
Dual formulation of entropic OT

Primal formulation of entropic OT

$$\min_{\gamma \in \mathcal{P}} \langle \gamma, C \rangle_F + \lambda \sum_{i,j} \gamma_{i,j} (\log \gamma_{i,j} - 1)$$

Dual formulation of entropic OT

$$\max_{\alpha, \beta} \alpha^T a + \beta^T b - \frac{1}{\lambda} \exp \left( \frac{\alpha}{\lambda} \right)^T K \exp \left( \frac{\beta}{\lambda} \right) \quad \text{with} \quad K = \exp \left( -\frac{C}{\lambda} \right) \quad (12)$$

- Sinkhorn algorithm is a gradient ascent on the dual variables.
- Dual problem is unconstrained: stochastic gradient descent (SGD) [Genevay et al., 2016, Seguy et al., 2017] or L-BFGS [Blondel et al., 2017].
- Semi-dual : closed form for \( \beta \) for a fixed \( \alpha \) (sumlogexp) leads to fast SAG algorithm [Genevay et al., 2016].
Solving entropic OT with Bregman Projections

**Kullback Leibler (KL) divergence**

\[
\text{KL}(\gamma, \rho) = \sum_{ij} \gamma_{ij} \log \frac{\gamma_{ij}}{\rho_{ij}} = \langle \gamma, \log \frac{\gamma}{\rho} \rangle_F,
\]

where \( \gamma \) and \( \rho \) are discrete distributions with the same support.

**OT as a Bregman projection [Benamou et al., 2015]**

\( \gamma^* \) is the solution of the following Bregman projection

\[
\gamma^* = \arg\min_{\gamma \in \mathcal{P}} \text{KL}(\gamma, K), \quad \text{where } K = \exp\left(\frac{C}{\lambda}\right)
\]  

\( (13) \)

- Sinkhorn is an iterative projection scheme, with alternative projections on marginal constraints.
- Generalizes to Barycenter computation [Benamou et al., 2015].
- Also generalizes to other regularization but less efficient (Dykstra’s Projection algorithm [Dessein et al., 2016]).
Sinkhorn divergence

Sinkhorn loss

\[
W_\lambda(\mu_s, \mu_t) = \min_{\gamma \in \mathcal{P}} \langle \gamma, C \rangle_F + \lambda \sum_{i,j} \gamma(i,j) \log \gamma(i,j)
\]

- Entropic term has smoothing effect.
- Not a divergence ($W_\lambda(\mu, \mu) > 0$ for $\lambda > 0$).

OT loss (aka Sharp Sinkhorn [Luise et al., 2018])

\[
OT_\lambda(\mu_s, \mu_t) = \langle \gamma_0^\lambda, C \rangle_F
\]

- $\gamma_0^\lambda$ is the solution of entropic OT above.
- Not a divergence ($OT_\lambda(\mu, \mu) > 0$ for $\lambda > 0$).

Sinkhorn divergence [Genevay et al., 2017]

\[
SD_\lambda(\mu_s, \mu_t) = W_\lambda(\mu_s, \mu_t) - \frac{1}{2} W_\lambda(\mu_s, \mu_s) - \frac{1}{2} W_\lambda(\mu_t, \mu_t)
\]

- True divergence ($SD_\lambda(\mu, \mu) = 0$).
- Better statistical properties as Wasserstein distance [Genevay et al., 2018].
Regularized OT (general case)

\[
\gamma_0^\lambda = \arg\min_{\gamma \in \mathcal{P}} \langle \gamma, \mathbf{C} \rangle_F + \lambda \Omega(\gamma),
\]

- **Group lasso** [Courty et al., 2016b]

\[
\Omega(\gamma) = \sum_g \sqrt{\sum_{i,j \in \mathcal{G}_g} \gamma_{i,j}^2}
\]

Promotes group sparsity (also submodular reg. [Alvarez-Melis et al., 2017])

- **Frobenius norm** [Blondel et al., 2017]

\[
\Omega(\gamma) = \sum_{i,j} \gamma_{i,j}^2
\]

Strongly convex regularization that keeps some sparsity in the solution.

- [Dessein et al., 2016]: KL, Itakura Saito, \(\beta\)-divergences.

Solved with Alternative optimization techniques when projection is efficient.
Minimizing the Wasserstein distance

Let \( \mu_s = \sum_{i=1}^{n} a_i \delta_{x_i^s} \). We seek the minimal Wasserstein estimator:

\[
\min_{\mu_s} W(\mu_s, \mu_t)
\]

In practice for a discrete distribution \( \mu_s \) there are two ways of doping this:

- **Case 1:** For a fixed support \( X_s = \{x_i^s\} \) find the optimal weights \( a \) (Eulerian).
- **Case 2:** For fixed weights \( a \) find the optimal support \( X_s = \{x_i^s\} \) (Lagrangian).
Gradient with respect to weights $\alpha$

$$W(\mu_s, \mu_t) = \max_{\alpha \in \mathbb{R}^{n_s}, \beta \in \mathbb{R}^{n_t}, \alpha_i + \beta_j \leq c(x_i^s, x_j^t)} \alpha^T \alpha + \beta^T \beta$$  \tag{14}$$

- $W(\mu_s, \mu_t)$ is convex wrt. $\alpha$
- Dual solution $\alpha^*$ is a sub-gradient: $\alpha^* \in \partial_\alpha W(\mu_s, \mu_t)$
- Entropy regularized: $W(\mu_s, \mu_t)$ is smooth, convex and $\nabla_\alpha W(\mu_s, \mu_t) = \lambda \log u$.
- OT loss: $\nabla_\alpha OT(\mu_s, \mu_t)$ computed using the implicit function theorem [Luise et al., 2018].
Case 2: fixed probability masses $a$

Gradient and update respect to weights $X_s = \{x^s_i\}$ for $c(x, y) = \|x - y\|^2$

$$W_2^2(\mu_s, \mu_t) = \min_{\gamma \in \mathcal{P}} \sum_{i,j} \gamma_{i,j} \|x^s_i - x^t_j\|^2$$  \hspace{1cm} (15)

- Gradient: $\nabla_{x^s_i} W_2^2(\mu_s, \mu_t) = 2x^s_i - 2 \frac{1}{a_i} \sum_j \gamma_{i,j} x^t_j$
- $W_2^2(\mu_s, \mu_t)$ decreases if $X_s \leftarrow \text{diag}(a^{-1})\gamma^* X_t$
- Expression above called barycentric interpolation [Ferradans et al., 2014].
Case 2: fixed probability masses \( a \)

**Distributions**

**Grad. wrt \( x_\mathbf{s}^i \) of \( W(\mu_s, \mu_t) \)**

**Update \( x_\mathbf{s}^i \) for fixed \( y \)**

Gradient and update respect to weights \( \mathbf{X}_s = \{ \mathbf{x}_\mathbf{s}^i \} \) for \( c(\mathbf{x}, \mathbf{y}) = \| \mathbf{x} - \mathbf{y} \|^2 \)

\[
W_2^2(\mu_s, \mu_t) = \min_{\gamma \in \mathcal{P}} \sum_{i,j} \gamma_{i,j} \| \mathbf{x}_i^s - \mathbf{x}_j^t \|^2
\]  

- Gradient: \( \nabla_{\mathbf{x}_\mathbf{s}^i} W_2^2(\mu_s, \mu_t) = 2\mathbf{x}_i^s - 2 \frac{1}{a_i} \sum_j \gamma_{i,j} \mathbf{x}_j^t \)
- \( W_2^2(\mu_s, \mu_t) \) decreases if \( \mathbf{X}_s \leftarrow \text{diag}(a^{-1}) \gamma^* \mathbf{X}_t \)
- Expression above called barycentric interpolation [Ferradans et al., 2014].
General case for entropic OT: autodifferentiation

Sinkhorn Autodiff [Genevay et al., 2017]

- Computing gradients through implicit function theorem can be costly [Luise et al., 2018].
- Each iteration of the Sinkhorn algorithm is differentiable.
- Modern neural network toolboxes can perform autodiff (Pytorch, Tensorflow).
- Fast but needs log-stabilization for numerical stability.
Outline

Optimal transport
- Monge and Kantorovitch
- OT on discrete distributions
- Wasserstein distances
- Barycenters and geometry of optimal transport

Computational aspects of optimal transport
- Special cases
- Regularized optimal transport
- Minimizing the Wasserstein distance

Gromov-Wasserstein
Can you transport between different spaces?

- $\Omega_s$ : source space, $\Omega_t$ : target space.
- Both domains/spaces do not share the same variables.
- There is no $c(x, y)$ between the two domains.
- They are related (observe similar objects) but not registered.
- Example: multi-modality with observations on different objects.
Gromov-Wasserstein distance

Inspired from Gabriel Peyré

GW for discrete distributions [Memoli, 2011]

\[ \mathcal{GW}_p(\mu_s, \mu_t) = \left( \min_{\gamma \in \Pi(\mu_s, \mu_t)} \sum_{i,j,k,l} |D_{i,k} - D'_{j,l}|^p \gamma_{i,j} \gamma_{k,l} \right)^{\frac{1}{p}} \]

with \( \mu_s = \sum_i a_i \delta_{x_i^s} \) and \( \mu_t = \sum_j b_j \delta_{x_j^t} \) and \( D_{i,k} = \|x_i^s - x_k^s\|, D'_{j,l} = \|x_j^t - x_l^t\| \)

- Distance over measures with no common ground space.
- Works well on graphs and structured data.
- Invariant to rotations and translation in either spaces.
Solving the Gromov Wasserstein optimization problem

\[ GW_p^p(\mu_s, \mu_t) = \min_{\gamma \in \Pi(\mu_s, \mu_t)} \sum_{i,j,k,l} |D_{i,k} - D'_{j,l}|^p \gamma_{i,j} \gamma_{k,l} \]

with \( \mu_s = \sum_i a_i \delta_{x_i^s} \) and \( \mu_t = \sum_j b_j \delta_{x_j^t} \) and \( D_{i,k} = \|x_i^s - x_k^s\|, D'_{j,l} = \|x_j^t - x_l^t\| \)

**Optimization problem**

- Quadratic Program (Wasserstein is a linear program).
- Nonconvex, NP-hard, related to Quadratic Assignment Problem (QAP).

**Optimization algorithm**

- Large problem and non convexity forbid standard QP solvers.
- Local solution can be obtained with conditional gradient (each iteration is an OT problems).
- Using entropy regularization leads to efficient projected gradient (each iteration is a sinkhorn) [Peyré et al., 2016].
Applications of GW [Solomon et al., 2016]

Shape matching between 3D and 2D objects

Multidimensional scaling (MDS) of shape collection
Summary for Part 1

Optimal transport

- Theoretically grounded ways of comparing probability distributions.
- Non-parametric comparison (between empirical distributions)
- Ground metric encode the geometry of the space (barycenters, geodesic).
- Two aspects: mapping vs coupling.

Optimization

- Solving OT is a linear program.
- Regularization (entropic) leads to faster algorithms.
- Minimization of Wasserstein distance can be done.

Next step: how to use it in machine learning?
**Barycenters in the wasserstein space.**  

**Structured optimal transport.**  

**Iterative Bregman projections for regularized transportation problems.**  
*SISC.*
**Smooth and sparse optimal transport.**

**Sliced and radon Wasserstein barycenters of measures.**

**Polar factorization and monotone rearrangement of vector-valued functions.**
*Communications on pure and applied mathematics*, 44(4):375–417.


**Regularized optimal transport and the rot mover’s distance.** 

**Regularized discrete optimal transport.** 
*SIAM Journal on Imaging Sciences, 7*(3).

**Sample complexity of sinkhorn divergences.** 

**Stochastic optimization for large-scale optimal transport.** 
In *NIPS*, pages 3432–3440.
**Sinkhorn-autodiff: Tractable wasserstein learning of generative models.**

**Fast optimal transport averaging of neuroimaging data.**

Kantorovich, L. (1942). 
**On the translocation of masses.**


Kolouri, S., Zou, Y., and Rohde, G. K. (2016). **Sliced wasserstein kernels for probability distributions.**


Monge, G. (1781).

Mémoire sur la théorie des déblais et des remblais.
De l’Imprimerie Royale.


Optimal Transport with Proximal Splitting.


Fast and robust earth mover’s distances.

In *2009 IEEE 12th International Conference on Computer Vision*, pages 460–467. IEEE.

**Gromov-Wasserstein Averaging of Kernel and Distance Matrices.**


**The earth mover’s distance as a metric for image retrieval.**


**Introduction to optimal transport theory.**
*Notes*.


**Large-scale optimal transport and mapping estimation.**

Convolutional wasserstein distances: Efficient optimal transportation on geometric domains.


Entropic metric alignment for correspondence problems.


Population averaging of neuroimaging data using $l_p$ distance-based optimal transport.
In 2018 International Workshop on Pattern Recognition in Neuroimaging (PRNI), pages 1–4. IEEE.