Overview

• Measures and Histograms
• From Monge to Kantorovitch Formulations
• Entropic Regularization and Sinkhorn
• Barycenters
• Unbalanced OT and Gradient Flows
• Minimum Kantorovitch Estimators
• Gromov-Wasserstein
Comparing Measures and Spaces

- Probability distributions and histograms
  → images, vision, graphics and machine learning, . . .
Comparing Measures and Spaces

- **Probability distributions and histograms** → images, vision, graphics and machine learning.

- **Optimal transport** → takes into account a metric $d$. 

\[ L^2 \text{ mean} \quad \text{Optimal transport mean} \]
Probability Measures

Positive Radon measure $\mu$ on a set $X$.

$$d\mu(x) = m(x)dx$$

Measure of sets $A \subset X$: $\mu(A) = \int_A d\mu(x) \geq 0$
Probability Measures

Positive Radon measure $\mu$ on a set $X$.

$$d\mu(x) = m(x)dx$$

$$\mu = \sum_i \mu_i \delta_{x_i}$$

Integration against continuous functions:

$$\int_X g(x)d\mu(x) \geq 0$$

Measure of sets $A \subset X$: $\mu(A) = \int_A d\mu(x) \geq 0$

Integration against continuous functions:

$$\int_X g(x)d\mu(x) = \int_X m(x)dx$$

$$\mu = \sum_i \mu_i \delta_{x_i}$$

$$\int_X gd\mu = \sum_i \mu_i g(x_i)$$

Measure of sets $A \subset X$: $\mu(A) = \int_A d\mu(x) \geq 0$

Integration against continuous functions:

$$\int_X g(x)d\mu(x) = \int_X m(x)dx$$

$$\mu = \sum_i \mu_i \delta_{x_i}$$

$$\int_X gd\mu = \sum_i \mu_i g(x_i)$$
Probability Measures

Positive Radon measure $\mu$ on a set $X$.

$\mu = \sum_i \mu_i \delta_{x_i}$

Measure of sets $A \subset X$: $\mu(A) = \int_A d\mu(x) \geq 0$

Integration against continuous functions: $\int_X g(x) d\mu(x) \geq 0$

$\mu = \sum_i \mu_i \delta_{x_i}$

Probability (normalized) measure: $\mu(X) = \int_X d\mu(x) = 1$
Random vectors

\[ \mathbb{P}(X \in A) \]

Weak* convergence:

\[ \forall \text{ set } A \quad \mathbb{P}(X_n \in A) \xrightarrow{n \to +\infty} \mathbb{P}(X \in A) \]

Radon measures

\[ \int_A d\mu(x) \]

Convergence in law:

\[ \forall \text{ continuous function } f \quad \int f d\mu_n \xrightarrow{n \to +\infty} \int f d\mu \]
Random vectors

\[ P(X \in A) \]

Weak* convergence:

\[ \forall \text{ set } A \quad P(X_n \in A) \xrightarrow{n \to +\infty} P(X \in A) \]

Radon measures

\[ \int_A d\mu(x) \]

Convergence in law:

\[ \forall \text{ continuous function } f \quad \int f d\mu_n \xrightarrow{n \to +\infty} \int f d\mu \]

Weak convergence:
Discretization: Histogram vs. Empirical

Discrete measure: \( \mu = \sum_{i=1}^{N} \mu_i \delta_{x_i} \) \( x_i \in X \), \( \sum_i \mu_i = 1 \)

**Lagrangian (point clouds)**
Constant weights \( \mu_i = \frac{1}{N} \)

**Eulerian (histograms)**
Fixed positions \( x_i \) (e.g. grid)

Quotient space: \( X^N / \Sigma_N \)

Convex polytope (simplex): \( \{(\mu_i)_i \geq 0 ; \sum_i \mu_i = 1\} \)
Push Forward

Radon measures \((\mu, \nu)\) on \((X, Y)\).

Transfer of measure by \(f : X \rightarrow Y\): push forward.

\[
\nu = f_\# \mu \text{ defined by: } \quad \nu(A) \overset{\text{def.}}{=} \mu(f^{-1}(A))
\]

\[
\iff \int_Y g(y) d\nu(y) \overset{\text{def.}}{=} \int_X g(f(x)) d\mu(x)
\]
Push Forward

Radon measures \((\mu, \nu)\) on \((X, Y)\).

Transfer of measure by \(f : X \to Y\): \textit{push forward}. 

\[ \nu = f^*\mu \] defined by:

\[ \nu(A) \overset{\text{def.}}{=} \mu(f^{-1}(A)) \]

\[ \iff \int_Y g(y) d\nu(y) \overset{\text{def.}}{=} \int_X g(f(x)) d\mu(x) \]

Smooth densities: \(d\mu = \rho(x)dx\), \(d\nu = \xi(x)dx\)

\[ f^*\mu = \nu \iff \rho(f(x)) | \det(\partial f(x)) | = \xi(x) \]
Push-forward vs. Pull-back

Measures: push-forward

\[ f \colon X \to Y \]

\[ f_\# : \mathcal{M}(X) \to \mathcal{M}(Y) \]

\[ \mu = \sum_i \delta_{x_i} \]

\[ f_\# \mu \overset{\text{def.}}{=} \sum_i \delta_{f(x_i)} \]

Functions: pull-back

\[ f \colon Y \to X \]

\[ f^\# : \mathcal{C}(Y) \to \mathcal{C}(X) \]

\[ f^\# \varphi \overset{\text{def.}}{=} \varphi \circ f \]

Remark: \( f^\# \) and \( f_\# \) are adjoints

\[ \int_Y \varphi d(f_\# \mu) = \int_X (f^\# \varphi) d\mu \]
Convergence of Random Variables

In mean

$$\lim_{n \to +\infty} \mathbb{E}(|X_n - X|^p) = 0$$

Almost sure

$$\mathbb{P}(\lim_{n \to +\infty} X_n = X) = 1$$

In probability

$$\forall \varepsilon > 0, \mathbb{P}(|X_n - X| > \varepsilon) \xrightarrow{n \to +\infty} 0$$

In law

$$\mathbb{P}(X_n \in A) \xrightarrow{n \to +\infty} \mathbb{P}(X \in A)$$

(the $X_n$ can be defined on different spaces)
Overview

- Measures and Histograms
- **From Monge to Kantorovitch Formulations**
- Entropic Regularization and Sinkhorn
- Barycenters
- Unbalanced OT and Gradient Flows
- Minimum Kantorovitch Estimators
- Gromov-Wasserstein
MÉMOIRE
SUR LA
THÉORIE DES DÉBLAIS
ET DES REMBLAIS.
Par M. MONGE.

Lorsqu'on doit transporter des terres d'un lieu dans un autre, on a coutume de donner le nom de Déblais au volume des terres que l'on doit transporter, & le nom de Remblais à l'espace qu'elles doivent occuper après le transport.

Le prix du transport d'une molécule étant, toutes choses d'ailleurs égales, proportionnel à son poids & à l'espace qu'on lui fait parcourir, & par conséquent le prix du transport total devant être proportionnel à la somme des produits des molécules multipliées chacune par l'espace parcouru, il s'ensuit que le déblai & le remblai étant donnés de figure & de position, il n'est pas indifférent que telle molécule du déblai soit transportée dans tel ou tel autre endroit du remblai, mais qu'il y a une certaine disposition à faire des molécules du premier dans le second, d'après laquelle la somme de ces produits fera la moindre possible, & le prix du transport total fera un minimum.
Monge Transport

\[
\min_{\nu = f_#\mu} \int_X c(x, f(x)) d\mu(x)
\]
Monge Transport

\[ \min_{\nu = f \# \mu} \int_X c(x, f(x)) \, d\mu(x) \]

**Theorem:** [Brenier] for \( c(x, y) = \|x - y\|^2 \), \((\mu, \nu)\) with density, there exists a unique optimal \( f \). One has \( f = \nabla \psi \) where \( \psi \) is the unique convex function such that \((\nabla \psi) \# \mu = \nu\)
Monge Transport

\[
\min_{\nu = f_\# \mu} \int_X c(x, f(x)) d\mu(x)
\]

**Theorem:** [Brenier] for \( c(x, y) = \|x - y\|^2 \), \((\mu, \nu)\) with density, there exists a unique optimal \( f \). One has \( f = \nabla \psi \) where \( \psi \) is the unique convex function such that \((\nabla \psi)_\# \mu = \nu\).

**Monge-Ampère equation:** \( \rho(\nabla \psi) \det(\partial^2 \psi) = \xi \)
Monge Transport

\[ \min_{\nu=f_\# \mu} \int_X c(x, f(x)) \, d\mu(x) \]

**Theorem:** [Brenier] for \( c(x, y) = \|x - y\|^2 \), \((\mu, \nu)\) with density, there exists a unique optimal \( f \). One has \( f = \nabla \psi \) where \( \psi \) is the unique convex function such that \( (\nabla \psi)_\# \mu = \nu \)

**Monge-Ampère equation:** \( \rho(\nabla \psi) \det(\partial^2 \psi) = \xi \)

**Non-uniqueness / non-existence:**

\[ f \quad f' \]
\[ \nu \quad \nu \]
\[ \mu \quad \mu \]
\[ ? \quad \nu \]

\[ f \quad \nu \quad \nu \]
\[ f \quad \nu \quad \nu \]
\[ f \quad \nu \quad \nu \]
\[ f \quad \nu \quad \nu \]
Leonid Kantorovich (1912-1986)

Леонид Витальевич Канторович

ON THE TRANSLOCATION OF MASSES
L. V. Kantorovich

The original paper was published in Dokl. Akad. Nauk SSSR, 37, No. 7-8, 227-229 (1942).

We assume that $R$ is a compact metric space, though some of the definitions and results given below can be formulated for more general spaces.

Let $\Phi(e)$ be a mass distribution, i.e., a set function such that: (1) it is defined for Borel sets, (2) it is nonnegative, $\Phi(e) \geq 0$, (3) it is absolutely additive if $e = e_1 + e_2 + \cdots + e_n$ ($e_i \neq e_j$, $i \neq j$), then $\Phi(e) = \Phi(e_1) + \Phi(e_2) + \cdots$. Let $\Phi_1(e')$ be another mass distribution such that $\Phi(e') = \Phi(e)$. By definition, a translocation of masses is a function $\Psi(e, e')$ defined for pairs of Borel sets $e, e' \in R$ such that (1) it is nonnegative and absolutely additive with respect to each of its arguments, (2) $\Psi(e, e) = \Phi(e)$, $\Psi(e', e') = \Phi(e')$.$^*$

Let $\rho(x, y)$ be a known continuous nonnegative function representing the work required to move a unit mass from $x$ to $y$.

We define the work required for the translocation of two given mass distributions as

$$W(\Phi, \Phi') = \int_{R} \rho(x, y)|\Phi(dx) - \Phi'(dy)|,$$

where $e_i$ are disjoint and $\sum_i e_i = R$, $e'_i$ are disjoint and $\sum_i e'_i = R$, $x_i \in e_i$, $x'_i \in e'_i$, and $\lambda$ is the largest of the numbers $\lambda_1 = \lambda_1(e_i) = \min(x_i, x'_i)$ and $\lambda_2 = \lambda_2(e'_i) = \max(x_i, x'_i)$. Clearly, this integral does exist.

We call the quantity

$$W(\Phi, \Phi') - \rho W(\Phi, \Phi'),$$

the minimal translocation work. Since the set of all functions $\Psi(e)$ is compact, there exists a function $\Phi_0$ realizing this minimum, so that

$$W(\Phi, \Phi_0) = W(\Phi_0, \Phi_0).$$
Before Kantorovitch

Figure 1: Figure from Tolstoi [1930] to illustrate a negative cycle

Optimal Transport was formulated in 1930 by A.N. Tolstoi, 12 years before Kantorovich. He even solved a "large scale" 10×68 instance!
Kantorovitch’s Formulation

Input distributions
\[ \mu = \sum_i \mu_i \delta_{x_i} \]
\[ \nu = \sum_j \nu_j \delta_{y_j} \]

Points \((x_i)_i, (y_j)_j\)

Weights \(\mu_i \geq 0, \nu_j \geq 0\).
\[ \sum_{i=1}^{N_1} \mu_i = \sum_{j=1}^{N_2} \nu_j = 1 \]

Def. Couplings
\[ C_{\mu, \nu} \overset{\text{def.}}{=} \left\{ T \in \mathbb{R}_{+}^{N_1 \times N_2} ; T1_{N_1} = \mu, T^\top 1_{N_2} = \nu \right\} \]
Kantorovitch’s Formulation

Input distributions

\[
\begin{align*}
\mu &= \sum_i \mu_i \delta_{x_i} \\
\nu &= \sum_j \nu_j \delta_{y_j}
\end{align*}
\]

Points \((x_i)_i, (y_j)_j\)

Weights \(\mu_i \geq 0, \nu_j \geq 0\).

\[
\sum_{i=1}^{N_1} \mu_i = \sum_{j=1}^{N_2} \nu_j = 1
\]

\[
d_{i,j} = d(x_i, y_j)
\]

Def. Couplings

\[
C_{\mu,\nu}^{\text{def.}} = \left\{ T \in \mathbb{R}^{N_1 \times N_2}_+ ; T1_{N_1} = \mu, T^\top 1_{N_2} = \nu \right\}
\]

Def. Wasserstein Distance / EMD

\[
W_p^p(\mu, \nu) \overset{\text{def.}}{=} \min \left\{ \sum_{i,j} T_{i,j} d_{i,j}^p ; T \in C_{\mu,\nu} \right\}
\]

[Kantorovich 1942]

\(\rightarrow W_p\) is a distance over Radon probability measures.
OT Between General Measures

**Couplings:** \( \Pi(\mu, \nu) \overset{\text{def.}}{=} \{\pi \in \mathcal{M}_+(X \times X) \ ; \ P_1\#\pi = \mu, P_2\#\pi = \nu\} \)

**Marginals:** \( P_1\#\pi(S) \overset{\text{def.}}{=} \pi(S, X) \quad P_2\#\pi(S) \overset{\text{def.}}{=} \pi(X, S) \)

**Optimal transport:** [Kantorovitch 1942]

\[
W_p^p(\mu, \nu) \overset{\text{def.}}{=} \min_{\pi} \left\{ \langle d^p, \pi \rangle = \int_{X \times X} d(x, y)^p \, d\pi(x, y) \ ; \ \pi \in \Pi(\mu, \nu) \right\}
\]
Couplings: the 3 Settings

Discrete

Semi-discrete

Continuous
Couplings

\[ \alpha \quad \beta \]

\[ \beta \quad \pi \quad \beta \]

\[ \beta \quad \pi \quad \beta \]

\[ \beta \quad \pi \quad \beta \]

\[ \alpha \quad \beta \]

\[ \beta \quad \pi \quad \beta \]

\[ \beta \quad \pi \quad \beta \]
1-D Optimal Transport

**Remark.** If $\Omega = \mathbb{R}$, $c(x, y) = c(|x - y|)$, $c$ convex, $F^{-1}_\mu, F^{-1}_\nu$ quantile functions,

\[
W(\mu, \nu) = \int_0^1 c(|F^{-1}_\mu(x) - F^{-1}_\nu(x)|)dx
\]
Remark. If $\Omega = \mathbb{R}^d$, $c(x, y) = \|x - y\|^2$, and
$\mu = \mathcal{N}(m_\mu, \Sigma_\mu)$, $\nu = \mathcal{N}(m_\nu, \Sigma_\nu)$ then
\[ W_2^2(\mu, \nu) = \|m_\mu - m_\nu\|^2 + B(\Sigma_\mu, \Sigma_\nu)^2 \]
where $B$ is the Bures metric
\[ B(\Sigma_\mu, \Sigma_\nu)^2 = \text{trace}(\Sigma_\mu + \Sigma_\nu - 2(\Sigma_{1/2}^\mu \Sigma_\nu \Sigma_{1/2}^\mu)^{1/2}). \]

The map $T : x \mapsto m_\nu + A(x - m_\mu)$ is optimal,
where $A = \Sigma^{-1/2}_\mu \left( \Sigma_{1/2}^\mu \Sigma_\nu \Sigma_{1/2}^\mu \right)^{1/2} \Sigma^{-1/2}_\mu$. 

\[ T : x \mapsto m_\nu + A(x - m_\mu) \]
Remark 2.11 (Distance between Gaussians). If $\alpha = \mathcal{N}(m_\alpha, C_\alpha)$ and $\beta = \mathcal{N}(m_\beta, C_\beta)$, then one can show that

$$\mathcal{W}_2^2(\alpha, \beta) = \|m_\alpha - m_\beta\|^2 + \mathcal{B}(C_\alpha, C_\beta)^2$$

(2.19)

where $\mathcal{B}$ is the so-called Bures metric

$$\mathcal{B}(C_\alpha, C_\beta)^2 \overset{\text{def.}}{=} \text{tr} \left( C_\alpha + C_\alpha - 2(C_\alpha^{1/2}C_\beta C_\alpha^{1/2})^{1/2} \right)$$

(2.20)
\[ W_1(a, b) = \min_{s \in \mathbb{R}_+^\mathcal{E}} \left\{ \sum_{(i, j) \in \mathcal{E}} w_{i, j} s_{i, j} : \text{div}(s) = a - b \right\} \]
Metrics on the Space of Measures

$$d\mu(x) = \rho(x)dx$$
$$d\tilde{\mu}(x) = \tilde{\rho}(x)dx$$

Bins-to-bins metrics:

Kullback-Leibler divergence:

$$D_{KL}(\mu, \tilde{\mu}) = \int \rho(x) \log \frac{\rho(x)}{\tilde{\rho}(x)} dx$$

Hellinger distance:

$$D_H(\mu, \tilde{\mu})^2 = \int \left( \sqrt{\rho(x)} - \sqrt{\tilde{\rho}(x)} \right)^2 dx$$
Bins-to-bins metrics:

Kullback-Leibler divergence:

\[ D_{KL}(\mu, \tilde{\mu}) = \int \rho(x) \log \frac{\rho(x)}{\tilde{\rho}(x)} \, dx \]

Hellinger distance:

\[ D_H(\mu, \tilde{\mu})^2 = \int \left( \sqrt{\rho(x)} - \sqrt{\tilde{\rho}(x)} \right)^2 \, dx \]

Effect of translation:

\[ D(\mu, \mu_\delta) \approx \text{cst} \quad \text{and} \quad W_2(\mu, \mu_\delta) = \delta \]
**Csiszár Divergence vs Dual Norms**

\[ D_\varphi(\alpha|\beta) \overset{\text{def.}}{=} \int_X \varphi \left( \frac{d\alpha}{d\beta} \right) d\beta \]

\[ \|\alpha - \beta\|_B \overset{\text{def.}}{=} \max_{f \in B} \int_X f(x) (d\alpha(x) - d\beta(x)) \]

**Csiszár divergences:**

- Weak topology
  - KL, TV, \( \chi^2 \), Hellinger, ...

**Dual norms:**

- Strong topology
  - \( W_1 \), flat, RKHS*, energy dist, ...

**Strong topology**

**Weak topology**
Csiszár divergences, a unifying way to define losses between arbitrary positive measures (discrete & densities). https://en.wikipedia.org/wiki/F-divergence

\[ D_\varphi(\alpha|\beta) \overset{\text{def.}}{=} \int X \varphi \left( \frac{d\alpha}{d\beta} \right) d\beta + \varphi'_\infty \alpha'(X) \]

\[ \varphi'_\infty = \lim_{x \to +\infty} \varphi(x)/x \in \mathbb{R} \cup \{\infty\} \]

Csiszár divergences, a unifying way to define losses between arbitrary positive measures (discrete & densities). https://en.wikipedia.org/wiki/F-divergence
Dual norms: (aka Integral Probability Metrics)

\[ \|\alpha - \beta\|_B^{\text{def.}} = \max \left\{ \int_X f(x)(d\alpha(x) - d\beta(x)) \mid f \in B \right\} \]

Wasserstein 1: \( B = \{ f ; \|\nabla f\|_\infty \leq 1 \} \).

Flat norm: \( B = \{ f ; \|f\|_\infty \leq 1, \|\nabla f\|_\infty \leq 1 \} \).

RKHS: \( B = \{ f ; \|f\|_k^2 \leq 1 \} \).

\[ \|\alpha - \beta\|_B^2 = \int k(x, x')d\alpha(x)d\alpha(x') + \int k(x, x')d\beta(y)d\beta(y') - 2 \int k(x, y)d\alpha(x)d\beta(y) \]

Energy distance: \( k(x, y) = -\|x - y\| \)

Gaussian: \( k(x, y) = e^{-\|x - y\|^2 / 2\sigma^2} \)
RKHS Norms aka Maximum Mean Discrepency

Figure 8.4: Top row: display of $\psi$ such that $\|\alpha - \beta\|_k = \|\psi \ast (\alpha - \beta)\|_{L^2(\mathbb{R}^2)}$, formally defined over Fourier as $\hat{\psi}(\omega) = \sqrt{\hat{\varphi}(\omega)}$ where $k^\ast(x, x') = \varphi(x - x')$. Bottom row: display of $\psi \ast (\alpha - \beta)$. $(G, \sigma)$ stands for Gaussian kernel of variance $\sigma^2$ and ED for Energy Distance kernel (in which case $\psi(x) = 1/\sqrt{\|x\|}$).
The Earth Mover’s Distance

\[ \text{dist}(I_1, I_2) = W_1(\mu, \nu) \]

[Rubner’98]
The Word Mover’s Distance

\[ \text{dist}(D_1, D_2) = W_2(\mu, \nu) \]

[\text{Kusner’15}]
Overview

- Measures and Histograms
- From Monge to Kantorovitch Formulations
- **Linear Programming and Semi-discrete**
- Entropic Regularization and Sinkhorn
- Barycenters
- Unbalanced OT and Gradient Flows
- Minimum Kantorovitch Estimators
- Gromov-Wasserstein
Linear programming:

$$\mu = \sum_{i=1}^{N_1} p_i \delta x_i, \quad \nu = \sum_{j=1}^{N_2} p_j \delta y_i$$
## Algorithms

### Linear programming:

\[
\mu = \sum_{i=1}^{N_1} p_i \delta x_i, \quad \nu = \sum_{j=1}^{N_2} p_j \delta y_i
\]

### Hungarian/Auction:

\[
\mu = \frac{1}{N} \sum_{i=1}^{N} \delta x_i, \quad \nu = \frac{1}{N} \sum_{j=1}^{N} \delta y_j
\]

\[
T_{i,j} = \begin{cases} 
1/N & \text{if } j = \sigma(i), \\
0 & \text{otherwise.}
\end{cases}
\]
Linear programming:

\[ \mu = \sum_{i=1}^{N_1} p_i \delta x_i, \ \nu = \sum_{j=1}^{N_2} p_j \delta y_i \]

Hungarian/Auction:

\[ \mu = \frac{1}{N} \sum_{i=1}^{N} \delta x_i, \ \nu = \frac{1}{N} \sum_{j=1}^{N} \delta y_j \sim O(N^3) \]

1-D case, \( d = | \cdot |^p, p \geq 1 \).

→ sorting, \( O(N \log(N)) \) operations.

1/\(N\) if \( j = \sigma(i) \),
0 otherwise.
Hungarian/Auction: $\mu = \sum_{i=1}^{N_1} p_i \delta x_i, \nu = \sum_{j=1}^{N_2} p_j \delta y_i$

Linear programming: $\mu = \sum_{i=1}^{N_1} p_i \delta x_i, \nu = \sum_{j=1}^{N_2} p_j \delta y_i$

Hungarian/Auction: $\sim O(N^3)$

$\mu = \frac{1}{N} \sum_{i=1}^{N} \delta x_i, \nu = \frac{1}{N} \sum_{j=1}^{N} \delta y_j$

1-D case, $d = | \cdot |^p, p \geq 1$. Sorting, $O(N \log(N))$ operations.

Monge-Ampère/Benamou-Brenier, $d = \| \cdot \|_2^2$. 

$T_{i,j} = \begin{cases} 1/N & \text{if } j = \sigma(i), \\ 0 & \text{otherwise.} \end{cases}$

$\sigma$
**Algorithms**

<table>
<thead>
<tr>
<th>Linear programming:</th>
</tr>
</thead>
<tbody>
<tr>
<td>[ \mu = \sum_{i=1}^{N_1} p_i \delta x_i, \nu = \sum_{j=1}^{N_2} p_j \delta y_i ]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Hungarian/Auction:</th>
</tr>
</thead>
<tbody>
<tr>
<td>[ \mu = \frac{1}{N} \sum_{i=1}^{N} \delta x_i, \nu = \frac{1}{N} \sum_{j=1}^{N} \delta y_j \sim O(N^3) ]</td>
</tr>
</tbody>
</table>

| 1-D case, \( d = | \cdot |^p, p \geq 1 \). |
| Sorting, \( O(N \log(N)) \) operations. |

| Monge-Ampère/Benamou-Brenier, \( d = \| \cdot \|_2^2 \). |

| Semi-discrete: Laguerre cells, \( d = \| \cdot \|_2^2 \). |
| [Merigot 2013] |

\[ T_{i,j} = \begin{cases} 
1/N & \text{if } j = \sigma(i), \\
0 & \text{otherwise.} 
\end{cases} \]
Algorithms

Linear programming:

\[ \mu = \sum_{i=1}^{N_1} p_i \delta x_i, \nu = \sum_{j=1}^{N_2} p_j \delta y_i \]

Hungarian/Auction:

\[ \mu = \frac{1}{N} \sum_{i=1}^{N} \delta x_i, \nu = \frac{1}{N} \sum_{j=1}^{N} \delta y_j \sim O(N^3) \]

1-D case, \( d = | \cdot |^p, p \geq 1. \)
→ sorting, \( O(N \log(N)) \) operations.

Monge-Ampère/Benamou-Brenier, \( d = \| \cdot \|_2 \).

Semi-discrete: Laguerre cells, \( d = \| \cdot \|_2^2 \).
[Merigot 2013]

\[ d = \| \cdot \|, p = 1 : W_1(\mu, \nu) = \min_{\text{div}(\nu) = \mu - \nu} \int \|u(x)\| dx \rightarrow \text{max-flow}. \]
Algorithms

**Linear programming:**
\[
\mu = \sum_{i=1}^{N_1} p_i \delta x_i, \quad \nu = \sum_{j=1}^{N_2} p_j \delta y_i
\]

**Hungarian/Auction:**
\[
\mu = \frac{1}{N} \sum_{i=1}^{N} \delta x_i, \quad \nu = \frac{1}{N} \sum_{j=1}^{N} \delta y_j
\]

1-D case, \( d = | \cdot |^p, p \geq 1 \). → sorting, \( O(N \log(N)) \) operations.

**Monge-Ampère/Benamou-Brenier,** \( d = \| \cdot \|_2^2 \).

**Semi-discrete: Laguerre cells,** \( d = \| \cdot \|_2^2 \).

\[ d = \| \cdot \|, p = 1 : W_1(\mu, \nu) = \min_{\text{div}(\nu) = \mu - \nu} \int \|u(x)\|dx \rightarrow \text{max-flow}. \]

Need for fast approximate algorithms for generic \( c \).
∀ y ∈ Y, \( f^c(y) \overset{\text{def.}}{=} \inf_{x \in X} c(x, y) - f(x) \),

∀ x ∈ X, \( \bar{g}^c(x) \overset{\text{def.}}{=} \inf_{y \in Y} c(x, y) - g(y) \),

\( \mathcal{L}_c(\alpha, \beta) = \max_{f \in \mathcal{C}(X)} \int_X f(x) d\alpha(x) + \int_Y f^c(y) d\beta(y) \),

= \max_{g \in \mathcal{C}(Y)} \int_X \bar{g}^c(x) d\alpha(x) + \int_Y g(y) d\beta(y) \).
Semi-discrete Descent Algorithm

Figure 5.2: Iterations of the semi-discrete OT algorithm minimizing (5.8) (here a simple gradient descent is used). The support \( (y_j) \) of the discrete measure — is indicated by the red points, while the continuous measure is the uniform measure on a square. The blue cells display the Laguerre partition \((Lg(\cdot))\) where \(g(\cdot)\) is the discrete dual potential computed at iteration \(\ell\).

Use of a Newton solver which is applied to sampling in computer graphics is proposed in [De Goes et al., 2012], see also [Lévy, 2015] for applications to 3-D volume and surface processing. An important area of application of semi-discrete method is for the resolution of incompressible fluid dynamics (Euler’s equations) using Lagrangian methods [De Goes et al. 2015, Gallouët and Mérigot 2017]. The semi-discrete OT solver enforces incompressibility at each iteration by imposing that the (possibly weighted) points cloud approximates a uniform inside the domain. The convergence (with linear rate) of damped Newton iterations is proved in [Mirebeau 2015] for the Monge-Ampère equation, and is refined in [Kitagawa et al. 2016] for optimal transport. Semi-discrete OT finds important applications to illumination design, see [Mérigot et al. 2017].

5.3 Entropic Semi-discrete Formulation

The dual of the entropic regularized problem between arbitrary measures (4.9) is

\[
\mathcal{L}_c(\mu, \nu) \triangleq \max_{f, g} \mathcal{C}(\mu) \times \mathcal{C}(\nu) \quad \text{subject to} \quad \int f(x) \, d\mu(x) = \int g(y) \, d\nu(y).
\]

This is a smooth unconstrained optimization problem.
Semi-discrete Stochastic Descent

Stochastic gradient descent for the semi-discrete Optimal Transport, illustration of convergence and corresponding Laguerre cells.

https://arxiv.org/abs/1605.08527
Entropic Regularization

**Entropy:** \( H(T) \overset{\text{def.}}{=} - \sum_{i,j=1}^{N} T_{i,j} (\log(T_{i,j}) - 1) \)

**Def. Regularized OT:** [Cuturi NIPS’13]

\[
\min_T \left\{ \sum_{i,j} d_{i,j}^p T_{i,j} - \varepsilon H(T) \mid T \in \mathcal{C}_{\mu,\nu} \right\}
\]
**Entropy:** \[ H(T) \overset{\text{def.}}{=} - \sum_{i,j=1}^{N} T_{i,j} (\log(T_{i,j}) - 1) \]

**Def. Regularized OT:** [Cuturi NIPS’13]
\[
\min_T \left\{ \sum_{i,j} d_{i,j}^p T_{i,j} - \varepsilon H(T) ; T \in \mathcal{C}_{\mu, \nu} \right\}
\]

**Regularization impact on solution:**
Overview

- Measures and Histograms
- From Monge to Kantorovich Formulations
- Entropic Regularization and Sinkhorn
- Barycenters
- Unbalanced OT and Gradient Flows
- Minimum Kantorovich Estimators
- Gromov-Wasserstein
Sinkhorn’s Algorithm

\[
\min_T \left\{ \sum_{i,j} d_{i,j}^p T_{i,j} + \varepsilon T_{i,j} \log(T_{i,j}) \ ; \ T \in \mathcal{C}_{\mu,\nu} \right\} \quad (\star)
\]

**Prop.** One has \( T = \text{diag}(a)K\text{diag}(b) \), where \( K = e^{-\frac{d^p}{\varepsilon}} \).
Sinkhorn’s Algorithm

\[
\min_T \left\{ \sum_{i,j} d_{i,j}^p T_{i,j} + \varepsilon T_{i,j} \log(T_{i,j}) ; \ T \in \mathcal{C}_{\mu,\nu} \right\} \quad (\star)
\]

Prop. One has \( T = \text{diag}(a)K \text{diag}(b) \), where \( K = e^{-\frac{d^p}{\varepsilon}} \).

Row constraint: \( T\mathbf{1}_{N_2} = \mu \iff a \odot (Kb) = \mu \)

Col. constraint: \( T^\top \mathbf{1}_{N_2} = \nu \iff b \odot (K^\top a) = \nu \)

Sinkhorn iterations: \( a \leftarrow \frac{\mu}{Kb} \) and \( b \leftarrow \frac{\nu}{K^\top a} \)
Sinkhorn’s Algorithm

\[
\min_T \left\{ \sum_{i,j} d_{i,j}^p T_{i,j} + \varepsilon T_{i,j} \log(T_{i,j}) \mid T \in \mathcal{C}_{\mu,\nu} \right\} \quad (\ast)
\]

**Prop.** One has \( T = \text{diag}(a)K \text{diag}(b) \), where \( K = e^{-\frac{d^p}{\varepsilon}} \).

Row constraint: \( T1_{N_2} = \mu \iff a \odot (Kn) = \mu \)

Col. constraint: \( T^\top 1_{N_2} = \nu \iff b \odot (K^\top a) = \nu \)

**Sinkhorn iterations:** \( a \leftarrow \frac{\mu}{Kn} \) and \( b \leftarrow \frac{\nu}{K^\top a} \)

Only matrix/vector multiplications. \( \rightarrow \) Parallelizable.

\( \rightarrow \) Streams well on GPU.

\( \rightarrow \) convolutive/heat structure for \( K \)
Sinkhorn’s Algorithm

\[
\min_T \left\{ \sum_{i,j} d_{i,j}^p T_{i,j} + \varepsilon T_{i,j} \log(T_{i,j}) ; \ T \in \mathcal{C}_{\mu,\nu} \right\} \quad (\star)
\]

Prop. One has \( T = \text{diag}(a)K\ \text{diag}(b) \), where \( K = e^{-\frac{d^p}{\varepsilon}} \).

Row constraint: \( T1_{N_2} = \mu \iff a \odot (Kb) = \mu \)

Col. constraint: \( T^\top 1_{N_2} = \nu \iff b \odot (K^\top a) = \nu \)

Sinkhorn iterations: \( a \leftarrow \frac{\mu}{Kb} \) and \( b \leftarrow \frac{\nu}{K^\top a} \)

Only matrix/vector multiplications. \( \rightarrow \) Parallelizable.

\( \rightarrow \) Streams well on GPU.

\( \rightarrow \) convolutive/heat structure for \( K \)

Prop. \( (\star) \iff \min_T \{ \text{KL}(T|K) ; \ T \in \mathcal{C}_{\mu,\nu} \} \)

Sinkhorn \( \iff \) iterative projections.
Optimal Transport on Surfaces

Triangulated mesh $M$. Geodesic distance $d_M$. 
Triangulated mesh $M$. Geodesic distance $d_M$.

Ground cost: $c(x, y) = d_M(x, y)\alpha$. 

$d(x_i, \cdot)$ Level sets
Optimal Transport on Surfaces

Triangulated mesh $M$. Geodesic distance $d_M$.

Ground cost: $c(x, y) = d_M(x, y)^\alpha$.

Computing $c$ (Fast-Marching): $N^2 \log(N) \rightarrow$ too costly.
Heat equation on $M$: $\partial_t u_t(x, \cdot) = \Delta_M u_t(x, \cdot)$, $u_{t=0}(x, \cdot) = \delta_x$
Heat equation on $M$: $\partial_t u_t(x, \cdot) = \Delta_M u_t(x, \cdot)$, $u_{t=0}(x, \cdot) = \delta_x$

**Theorem:** [Varadhan] $-\varepsilon \log(u_\varepsilon) \xrightarrow{\varepsilon \to 0} d^2_M$
Entropic Transport on Surfaces

Heat equation on $M$: $\partial_t u_t(x, \cdot) = \Delta_M u_t(x, \cdot)$, $u_{t=0}(x, \cdot) = \delta_x$

Theorem: [Varadhan] $-\varepsilon \log(u_\varepsilon) \xrightarrow{\varepsilon \to 0} d^2_M$

Sinkhorn kernel: $K \overset{\text{def.}}{=} e^{-\frac{d_M^2}{\varepsilon}} \approx u_\varepsilon \approx (\text{Id} - \frac{\varepsilon}{\ell} \Delta_M)^{-\ell}$
Ground cost $c = d_M$: geodesic on cortical surface $M$. 

$L^2$ barycenter

$W_2^2$ barycenter
\[ \pi_{\varepsilon} \overset{\text{def.}}{=} \arg\min_{\pi} \left\{ \langle d^p, \pi \rangle + \varepsilon \KL(\pi|\pi_0) ; \; \pi \in \Pi(\mu, \nu) \right\} \]

**Schrödinger's problem:** \[ \pi_{\varepsilon} = \arg\min_{\pi \in \Pi(\mu, \nu)} \KL(\pi|K) \]

\[ K(x, y) \overset{\text{def.}}{=} e^{-\frac{d_p(x, y)}{\varepsilon}} \pi_0(x, y) \]

Landmark computational paper: [Cuturi 2013].
Regularization for General Measures

\[ \pi_\varepsilon \overset{\text{def.}}{=} \arg\min_{\pi} \{ \langle d^p, \pi \rangle + \varepsilon \text{KL}(\pi|\pi_0) ; \pi \in \Pi(\mu, \nu) \} \]

**Schrödinger's problem:** \[ \pi_\varepsilon = \arg\min_{\pi \in \Pi(\mu, \nu)} \text{KL}(\pi|K) \]

\[ K(x, y) \overset{\text{def.}}{=} e^{-\frac{d^p(x,y)}{\varepsilon}} \pi_0(x, y) \]

Landmark computational paper: [Cuturi 2013].

**Proposition:** [Carlier, Duval, Peyré, Schmitzer 2015]

\[ \pi_\varepsilon \overset{\varepsilon \to 0}{\longrightarrow} \arg\min_{\pi \in \Pi(\mu, \nu)} \langle d^p, \pi \rangle \quad \pi_\varepsilon \overset{\varepsilon \to +\infty}{\longrightarrow} \mu(x)\nu(y) \]
Back to Sinkhorn’s Algorithm

Optimal transport problem:

\[ f_1 = \nu_\mu \quad \rightarrow \quad \text{Prox}_{f_1/\varepsilon}^{\text{KL}}(\tilde{\mu}) = \mu \]

\[ f_2 = \nu_\nu \quad \rightarrow \quad \text{Prox}_{f_2/\varepsilon}^{\text{KL}}(\tilde{\nu}) = \nu \]
Back to Sinkhorn’s Algorithm

Optimal transport problem: \[ f_1 = \nu_{\mu} \quad \longrightarrow \quad \text{Prox}_{f_1/\varepsilon}^{KL}(\tilde{\mu}) = \mu \]
\[ f_2 = \nu_{\nu} \quad \longrightarrow \quad \text{Prox}_{f_2/\varepsilon}^{KL}(\tilde{\nu}) = \nu \]

Sinkhorn/IPFP algorithm: [Sinkhorn 1967][Deming, Stephan 1940]
\[ a^{(\ell+1)} \overset{\text{def.}}{=} \frac{\mu}{Kb^{(\ell)}} \quad \text{and} \quad b^{(\ell+1)} \overset{\text{def.}}{=} \frac{\nu}{K^*a^{(\ell+1)}} \]
Back to Sinkhorn’s Algorithm

Optimal transport problem:

\[ f_1 = \nu_\mu \quad \longrightarrow \quad \text{Prox}_{f_1/\varepsilon}^{\text{KL}}(\tilde{\mu}) = \mu \]

\[ f_2 = \nu_\nu \quad \longrightarrow \quad \text{Prox}_{f_2/\varepsilon}^{\text{KL}}(\tilde{\nu}) = \nu \]

Sinkhorn/IPFP algorithm: [Sinkhorn 1967][Deming, Stephan 1940]

\[
a(\ell+1) \overset{\text{def.}}{=} \frac{\mu}{Kb(\ell)} \quad \text{and} \quad b(\ell+1) \overset{\text{def.}}{=} \frac{\nu}{K^*a(\ell+1)}
\]

Proposition: \( \| \log(\pi^{(\ell)}) - \log(\pi^*) \|_\infty = O(1 - \delta)^\ell, \delta \sim \kappa_c^{-1/\varepsilon} \)

\( \pi^{(\ell)} \overset{\text{def.}}{=} \text{diag}(a^{(\ell)})K\text{diag}(b^{(\ell)}) \)

Local rate: [Knight 2008]
5.3. Entropic Semi-discrete Formulation

Figure 5.3: Top: examples of entropic semi-discrete $\bar{c}$-transforms $g_{\bar{c}}$, for ground cost $c(x, y) = |x \neq y|$ for varying $\varepsilon$ (see colorbar). The red points are at locations $(y_j, \neq g_j)$. Bottom: examples of entropic semi-discrete $\bar{c}$-transforms $g_{\bar{c}}$, for ground cost $c(x, y) = \sqrt{x \neq y}$ for varying $\varepsilon$. The black curves are the level sets of the function $g_{\bar{c}}$, while the colors indicate the smoothed indicator function of the Laguerre cells $\mathbb{K}_j$. The red points are at locations $y_j \in R^2$, and their size is proportional to $g_j$.
Overview

- Measures and Histograms
- From Monge to Kantorovitch Formulations
- Entropic Regularization and Sinkhorn
- **Barycenters**
- Unbalanced OT and Gradient Flows
- Minimum Kantorovitch Estimators
- Gromov-Wasserstein
Barycenters of measures \((\mu_k)_k:\ \sum_k \lambda_k = 1\)

\[\mu^* \in \arg\min_{\mu} \sum_k \lambda_k W_2^2(\mu_k, \mu)\]
Wasserstein Barycenters

Barycenters of measures \((\mu_k)_k\): \[ \sum_k \lambda_k = 1 \]
\[ \mu^* \in \arg\min_{\mu} \sum_k \lambda_k W_2^2(\mu_k, \mu) \]

Generalizes Euclidean barycenter:
If \(\mu_k = \delta_{x_k}\) then \(\mu^* = \delta_{\sum_k \lambda_k x_k}\)
Barycenters of measures \((\mu_k)_k\): \[ \sum_k \lambda_k = 1 \]

\[ \mu^* \in \text{argmin} \sum_k \lambda_k W_2^2(\mu_k, \mu) \]

Generalizes Euclidean barycenter:

If \(\mu_k = \delta_{x_k}\) then \(\mu^* = \delta \sum_k \lambda_k x_k\)
Wasserstein Barycenters

Barycenters of measures \((\mu_k)_k\): 
\[
\sum_k \lambda_k = 1 \\
\mu^* \in \text{argmin}_\mu \sum_k \lambda_k W_2^2(\mu_k, \mu)
\]

Generalizes Euclidean barycenter:

If \(\mu_k = \delta_{x_k}\) then \(\mu^* = \delta \sum_k \lambda_k x_k\)

Mc Cann’s displacement interpolation.
Wasserstein Barycenters

Barycenters of measures \((\mu_k)_k\): \[\sum_k \lambda_k = 1\]
\[\mu^* \in \arg\min_{\mu} \sum_k \lambda_k W_2^2(\mu_k, \mu)\]

Generalizes Euclidean barycenter:
If \(\mu_k = \delta_{x_k}\) then \(\mu^* = \delta\sum_k \lambda_k x_k\)

Mc Cann’s displacement interpolation.
Wasserstein Barycenters

Barycenters of measures \((\mu_k)_k\): \[\sum_k \lambda_k = 1\]
\[\mu^* \in \arg\min_{\mu} \sum_k \lambda_k W_2^2(\mu_k, \mu)\]

Generalizes Euclidean barycenter:
If \(\mu_k = \delta_{x_k}\) then \(\mu^* = \delta \sum_k \lambda_k x_k\)

Mc Cann’s displacement interpolation.

Theorem: [Agueh, Carlier, 2010]
(for \(c(x, y) = \|x - y\|^2\))
if \(\mu_1\) does not vanish on small sets, \(\mu^*\) exists and is unique.
Displacement Interpolation
Figure 7.2: Comparison of displacement interpolation (7.8) of discrete measures. Top: point clouds (empirical measures \( \alpha_0 \), \( \alpha_1/5 \), \( \alpha_2/5 \), \( \alpha_3/5 \), \( \alpha_4/5 \), \( \alpha_1 \)) with the same number of points. Bottom: same but with weight. For \( 0 < t < 1 \), the top example corresponds to an empirical measure interpolation \( t \mapsto \tilde{P}_t \) with \( N \) points, while the bottom one defines a measure supported on \( 2N \neq 1 \) points.

In the case that there is only a coupling \( f \) (not necessarily supported on a Monge map), one can compute this interpolant as

\[
\tilde{P}_t(x, y) = \frac{1}{N} \sum_{i,j} P_{i,j} \left( x_i + ty_j \right).
\]

For instance, in the discrete setup (2.3), denoting \( P \) a solution to (2.11), an interpolation is defined as

\[
\tilde{P}_t(x, y) = \frac{1}{N} \sum_{i,j} P_{i,j} \left( x_i + ty_j \right).
\]

Such an interpolation is typically supported on \( n + m \neq 1 \) points, which is the maximum number of nonzero elements of \( P \). Figure 7.2 shows two examples of such displacement interpolation of discrete measures. This construction can be generalized to geodesic spaces \( X \) by replacing \( P_t \) by the interpolation along geodesic paths.

McCann interpolation finds many applications, for instance color, shape and illumination interpolations in computer graphics [Boineau et al., 2011].
Wasserstein Barycenters

\[ \lambda \in \Sigma_3 \]

Wasserstein mean

\[ L_2 \text{ mean} \]
Wasserstein Barycenters

\[ \lambda \in \Sigma_3 \]

Wasserstein mean

\[ L_2 \text{ mean} \]
Regularized Barycenters

\[
\min_{(\pi_k)_k, \mu} \left\{ \sum_k \lambda_k \left( \langle c, \pi_k \rangle + \varepsilon \text{KL}(\pi_k \| \pi_{0,k}) \right) ; \forall k, \pi_k \in \Pi(\mu_k, \mu) \right\}
\]
Regularized Barycenters

\[
\min_{(\pi_k)_k, \mu} \left\{ \sum_k \lambda_k \left( \langle c, \pi_k \rangle + \varepsilon \text{KL}(\pi_k | \pi_{0,k}) \right) ; \forall k, \pi_k \in \Pi(\mu_k, \mu) \right\}
\]

→ Need to fix a discretization grid for \( \mu \), i.e. choose \((\pi_{0,k})_k\)
Regularized Barycenters

\[
\min_{(\pi_k)_k, \mu} \left\{ \sum_k \lambda_k \left( \langle c, \pi_k \rangle + \varepsilon \text{KL}(\pi_k | \pi_{0,k}) \right) ; \forall k, \pi_k \in \Pi(\mu_k, \mu) \right\}
\]

→ Need to fix a discretization grid for \( \mu \), i.e. choose \( (\pi_{0,k})_k \)

→ Sinkhorn-like algorithm [Benamou, Carlier, Cuturi, Nenna, Peyré, 2015].
Regularized Barycenters

\[
\min_{(\pi_k)_k, \mu} \left\{ \sum_k \lambda_k \left( \langle c, \pi_k \rangle + \varepsilon \KL(\pi_k | \pi_{0,k}) \right) ; \forall k, \pi_k \in \Pi(\mu_k, \mu) \right\}
\]

→ Need to fix a discretization grid for \( \mu \), i.e. choose \((\pi_{0,k})_k\)

→ Sinkhorn-like algorithm [Benamou, Carlier, Cuturi, Nenna, Peyré, 2015].

[Solomon et al, SIGGRAPH 2015]
Barycenters of 2D Shapes
Barycenters of 3D Shapes
Barycenter on a Surface
Barycenter on a Surface

\[ \mu_1, \mu_2, \mu_3, \mu_4, \mu_5, \mu_6 \]
Barycenter on a Surface

\[ \lambda = \frac{1}{6} (\mu_1, \ldots, \mu_6) \]
Color Transfer

Input images: \((f, g)\) (chrominance components)

Input measures: \(\mu(A) = \mathcal{U}(f^{-1}(A)), \nu(A) = \mathcal{U}(g^{-1}(A))\)
Color Transfer

Input images: \((f, g)\) (chrominance components)

Input measures: \(\mu(A) = \mathcal{U}(f^{-1}(A)), \nu(A) = \mathcal{U}(g^{-1}(A))\)
Color Transfer

Input images: \((f, g)\) (chrominance components)

Input measures: \(\mu(A) = \mathcal{U}(f^{-1}(A)), \nu(A) = \mathcal{U}(g^{-1}(A))\)

\[ f \xrightarrow{T_\gamma} T_\gamma \circ f \]

\[ \mu \xrightarrow{} \nu \]

\[ \tilde{T}_\gamma \circ g \xleftarrow{} g \]
Topic Models

[Rolet’16]
Overview

- Measures and Histograms
- From Monge to Kantorovitch Formulations
- Entropic Regularization and Sinkhorn
- Barycenters
- **Unbalanced OT and Gradient Flows**
- Minimum Kantorovitch Estimators
- Gromov-Wasserstein
Unbalanced Transport

\((\xi, \mu) \in \mathcal{M}_+(X)^2\), \(\text{KL}(\xi|\mu) \overset{\text{def.}}{=} \int_X \log \left( \frac{d\xi}{d\mu} \right) d\mu + \int_X (d\mu - d\xi)\)

\[
WF_c(\mu, \nu) \overset{\text{def.}}{=} \min_{\pi} \langle c, \pi \rangle + \lambda \text{KL}(P_1 \# \pi | \mu) + \lambda \text{KL}(P_2 \# \pi | \nu)
\]

[Liere, Mielke, Savaré 2015]
Unbalanced Transport

\((\xi, \mu) \in \mathcal{M}_+(X)^2, \quad \text{KL}(\xi|\mu) \overset{\text{def.}}{=} \int_X \log \left( \frac{d\xi}{d\mu} \right) d\mu + \int_X (d\mu - d\xi)\)

\[
WF_c(\mu, \nu) \overset{\text{def.}}{=} \min_{\pi} \langle c, \pi \rangle + \lambda \text{KL}(P_1#\pi|\mu) + \lambda \text{KL}(P_2#\pi|\nu)
\]

[Liere, Mielke, Savaré 2015]

**Proposition:** If \(c(x, y) = -\log(\cos(\min(d(x, y), \frac{\pi}{2})))\)
then \(WF_c^{1/2}\) is a distance on \(\mathcal{M}_+(X)\).
[Liere, Mielke, Savaré 2015] [Chizat, Schmitzer, Peyré, Vialard 2015]
Unbalanced Transport

\((\xi, \mu) \in \mathcal{M}_+(X)^2\), \(\text{KL}(\xi|\mu) \overset{\text{def.}}{=} \int_X \log \left( \frac{d\xi}{d\mu} \right) d\mu + \int_X (d\mu - d\xi)\)

\(WF_c(\mu, \nu) \overset{\text{def.}}{=} \min_{\pi} \langle c, \pi \rangle + \lambda \text{KL}(P_1#\pi|\mu) + \lambda \text{KL}(P_2#\pi|\nu)\)  

[Liégeois, Mielke, Savaré 2015]

**Proposition:** If \(c(x, y) = -\log(\cos(\min(d(x, y), \frac{\pi}{2}))\) 
then \(WF_{c}^{1/2}\) is a distance on \(\mathcal{M}_+(X)\).  
[Liégeois, Mielke, Savaré 2015] [Chizat, Schmitzer, Peyré, Vialard 2015]

→ “Dynamic” Benamou-Brenier formulation.  
[Liégeois, Mielke, Savaré 2015][Kondratyev, Monsaingeon, Vorotnikov, 2015]  
[Chizat, Schmitzer, P, Vialard 2015]

Balanced OT

Unbalanced OT
Metric space \((\mathcal{X}, d)\), minimize \(F(x)\) on \(\mathcal{X}\).

Implicit Euler step:

\[
x_{k+1} \overset{\text{def.}}{=} \arg\min_{x \in \mathcal{X}} d(x_k, x)^2 + \tau F(x)
\]

\[F(x) = \|x\|^2 \text{ on } (\mathcal{X} = \mathbb{R}^2, \| \cdot \|_p)
\]

\[
\{x ; d(x_k, x) \sim \tau\}
\]
Implicit vs. Explicit Stepping

Metric space \((\mathcal{X}, d)\), minimize \(F(x)\) on \(\mathcal{X}\).

**Explicit**

\[ x_{k+1} = \arg\min_{x \in \mathcal{X}} d(x_k, x)^2 + \tau \langle \nabla F(x_k), x \rangle \]

**Implicit**

\[ x_{k+1} = \arg\min_{x \in \mathcal{X}} d(x_k, x)^2 + \tau F(x) \]

\[ F(x) = \| x \|^2 \text{ on } (\mathcal{X} = \mathbb{R}^2, \| \cdot \|_p) \]
Implicit Euler step: 

\[ \mu_{t+1} = \text{Prox}_{\tau f} W(\mu_t) \overset{\text{def.}}{=} \arg\min_{\mu \in \mathcal{M}_+(X)} W^2_2(\mu_t, \mu) + \tau f(\mu) \]
Implicit Euler step: \[ \mu_{t+1} = \text{Prox}_{\tau f}^W (\mu_t) \overset{\text{def.}}{=} \arg\min_{\mu \in \mathcal{M}_+(X)} W_2^2 (\mu_t, \mu) + \tau f(\mu) \]

Formal limit $\tau \to 0$: \[ \partial_t \mu = \text{div} (\mu \nabla (f'(\mu))) \]
Implicit Euler step: \[ \mu_{t+1} = \text{Prox}_{\tau f}^W(\mu_t) \overset{\text{def.}}{=} \arg\min_{\mu \in \mathcal{M}_+(X)} W_2^2(\mu_t, \mu) + \tau f(\mu) \]

Formal limit $\tau \to 0$: \[ \partial_t \mu = \text{div} (\mu \nabla (f'(\mu))) \]

\[ f(\mu) = \int \log(\frac{d\mu}{dx}) d\mu \rightarrow \partial_t \mu = \Delta \mu \quad \text{(heat diffusion)} \]
Wasserstein Gradient Flows

Implicit Euler step: \[ \mu_{t+1} = \text{Prox}_{\tau f}^W(\mu_t) \overset{\text{def.}}= \arg\min_{\mu \in \mathcal{M}_+(X)} W_2^2(\mu_t, \mu) + \tau f(\mu) \]

Formal limit $\tau \to 0$: $\partial_t \mu = \text{div} (\mu \nabla (f'(\mu)))$

- $f(\mu) = \int \log\left(\frac{d\mu}{dx}\right) d\mu$ \quad \rightarrow \quad $\partial_t \mu = \Delta \mu$ (heat diffusion)
- $f(\mu) = \int w d\mu$ \quad \rightarrow \quad $\partial_t \mu = \text{div}(\mu \nabla w)$ (advection)
Wasserstein Gradient Flows

Implicit Euler step: \[
\mu_{t+1} = \text{Prox}^{W}_{\tau f}(\mu_t) \overset{\text{def.}}{=} \arg\min_{\mu \in \mathcal{M}_+(X)} W_2^2(\mu_t, \mu) + \tau f(\mu)
\]

Formal limit $\tau \to 0$: \[
\partial_t \mu = \text{div}(\mu \nabla (f'(\mu)))
\]

\[f(\mu) = \int \log\left(\frac{d\mu}{dx}\right) d\mu \quad \Rightarrow \quad \partial_t \mu = \Delta \mu \quad \text{(heat diffusion)}
\]

\[f(\mu) = \int w d\mu \quad \Rightarrow \quad \partial_t \mu = \text{div}(\mu \nabla w) \quad \text{(advection)}
\]

\[f(\mu) = \frac{1}{m-1} \int \left(\frac{d\mu}{dx}\right)^{m-1} d\mu \quad \Rightarrow \quad \partial_t \mu = \Delta \mu^m \quad \text{(non-linear diffusion)}
\]
Eulerian vs. Lagrangian Discretization
Lagrangian Discretization of Entropy

\[ \hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} \delta x_i \]

\[ \hat{H}(\hat{\mu}_n) \overset{\text{def.}}{=} \sum_i \log(\min_{j \neq i} \| x_i - x_j \|) \]

\[ H(\mu) \overset{\text{def.}}{=} - \int \log(\frac{d\mu}{dx}(x))d\mu(x) \]
Lagrangian Discretization of Gradient Flows

\[ \min_{\rho} E(\rho) \overset{\text{def.}}{=} \int V(x) \rho(x) \, dx + \int \rho(x) \log(\rho(x)) \, dx \]

Wasserstein flow of \( E \):

\[ \frac{d\rho_t}{dt} = \Delta \rho_t + \nabla (V \rho_t) \]
Primal: $\min_{\pi} \langle d^p, \pi \rangle + f_1(P_1 \# \pi) + f_2(P_2 \# \pi) + \varepsilon KL(\pi|\pi_0)$
Generalized Entropic Regularization

Primal: \( \min_{\pi} \langle d^p, \pi \rangle + f_1(P_1 \# \pi) + f_2(P_2 \# \pi) + \varepsilon \text{KL}(\pi | \pi_0) \)

Dual: \( \max_{u,v} - f_1^*(u) - f_2^*(v) - \varepsilon \langle e^{-\frac{u}{\varepsilon}}, K e^{-\frac{v}{\varepsilon}} \rangle \)

\( \pi(x, y) = a(x) K(x, y) b(y) \)

\( (a, b) \overset{\text{def.}}{=} (e^{-\frac{u}{\varepsilon}}, e^{-\frac{v}{\varepsilon}}) \)
**Generalized Entropic Regularization**

**Primal:**
\[
\min_\pi \langle d^p, \pi \rangle + f_1(P_1\|\pi) + f_2(P_2\|\pi) + \varepsilon \text{KL}(\pi|\pi_0)
\]

**Dual:**
\[
\max_{u,v} - f_1^*(u) - f_2^*(v) - \varepsilon \langle e^{-u/\varepsilon}, Ke^{-v/\varepsilon} \rangle
\]

\[
\pi(x,y) = a(x)K(x,y)b(y) \quad (a,b) \overset{\text{def.}}{=} (e^{-u/\varepsilon}, e^{-v/\varepsilon})
\]

**Block coordinates relaxation:**
\[
\max_u - f_1^*(u) - \varepsilon \langle e^{-u/\varepsilon}, Ke^{-v/\varepsilon} \rangle \quad (I_u)
\]
\[
\max_v - f_2^*(v) - \varepsilon \langle e^{-v/\varepsilon}, K^*e^{-u/\varepsilon} \rangle \quad (I_v)
\]
**Generalized Entropic Regularization**

### Primal:
\[
\min_{\pi} \langle d^p, \pi \rangle + f_1(P_1 \# \pi) + f_2(P_2 \# \pi) + \varepsilon \text{KL}(\pi | \pi_0)
\]

### Dual:
\[
\max_{u,v} - f_1^*(u) - f_2^*(v) - \varepsilon \langle e^{-\frac{u}{\varepsilon}}, Ke^{-\frac{v}{\varepsilon}} \rangle
\]

\[
\pi(x, y) = a(x) K(x, y) b(y) \quad (a, b) \overset{\text{def.}}{=} (e^{-\frac{u}{\varepsilon}}, e^{-\frac{v}{\varepsilon}})
\]

### Block coordinates relaxation:
\[
\begin{align*}
\max_u &- f_1^*(u) - \varepsilon \langle e^{-\frac{u}{\varepsilon}}, Ke^{-\frac{v}{\varepsilon}} \rangle \quad (I_u) \\
\max_v &- f_2^*(v) - \varepsilon \langle e^{-\frac{v}{\varepsilon}}, K^*e^{-\frac{u}{\varepsilon}} \rangle \quad (I_v)
\end{align*}
\]

### Proposition:
the solutions of $(I_u)$ and $(I_v)$ read:
\[
\begin{align*}
a &= \frac{\text{Prox}^\text{KL}_{f_1/\varepsilon}(Kb)}{Kb} \\
b &= \frac{\text{Prox}^\text{KL}_{f_2/\varepsilon}(K^*a)}{K^*a}
\end{align*}
\]

\[
\text{Prox}^\text{KL}_{f_1/\varepsilon}(\mu) \overset{\text{def.}}{=} \arg\min_v f_1(v) + \varepsilon \text{KL}(v | \mu)
\]
**Generalized Entropic Regularization**

**Primal:**
\[
\min_{\pi} \langle d^p, \pi \rangle + f_1(P_1 \# \pi) + f_2(P_2 \# \pi) + \varepsilon \text{KL}(\pi | \pi_0)
\]

**Dual:**
\[
\max_{u,v} - f_1^*(u) - f_2^*(v) - \varepsilon \langle e^{-\frac{u}{\varepsilon}}, Ke^{-\frac{v}{\varepsilon}} \rangle
\]

\[
\pi(x, y) = a(x)K(x, y)b(y) \quad (a, b) \overset{\text{def.}}{=} (e^{-\frac{u}{\varepsilon}}, e^{-\frac{v}{\varepsilon}})
\]

**Block coordinates relaxation:**
\[
\max_u - f_1^*(u) - \varepsilon \langle e^{-\frac{u}{\varepsilon}}, Ke^{-\frac{v}{\varepsilon}} \rangle \quad (\mathcal{I}_u)
\]
\[
\max_v - f_2^*(v) - \varepsilon \langle e^{-\frac{v}{\varepsilon}}, K^*e^{-\frac{u}{\varepsilon}} \rangle \quad (\mathcal{I}_v)
\]

**Proposition:**
the solutions of \((\mathcal{I}_u)\) and \((\mathcal{I}_v)\) read:
\[
a = \frac{\text{Prox}_{f_1/\varepsilon}(Kb)}{Kb}
\]
\[
b = \frac{\text{Prox}_{f_2/\varepsilon}(K^*a)}{K^*a}
\]

\[
\text{Prox}_{f_1/\varepsilon}(\mu) \overset{\text{def.}}{=} \arg\min_{\nu} f_1(\nu) + \varepsilon \text{KL}(\nu | \mu)
\]

→ Only matrix-vector multiplications. → Highly parallelizable.
→ On regular grids: only convolutions! Linear time iterations.
Gradient Flows: Crowd Motion

\[ \mu_{t+1} \overset{\text{def.}}{=} \arg\min_{\mu} W_\alpha(\mu_t, \mu) + \tau f(\mu) \]

Congestion-inducing function:
\[ f(\mu) = \iota_{[0,\kappa]}(\mu) + \langle w, \mu \rangle \]

[Maury, Roudneff-Chupin, Santambrogio 2010]
Gradient Flows: Crowd Motion

\[ \mu_{t+1} \overset{\text{def.}}{=} \operatorname{argmin}_\mu W^\alpha(\mu_t, \mu) + \tau f(\mu) \]

Congestion-inducing function:
\[ f(\mu) = \nu_{[0,\kappa]}(\mu) + \langle w, \mu \rangle \]
[Maury, Roudneff-Chupin, Santambrogio 2010]

Proposition: \( \operatorname{Prox}_{\frac{1}{\epsilon}} f(\mu) = \min(e^{-\epsilon w} \mu, \kappa) \)
Gradient Flows: Crowd Motion

\[ \mu_{t+1} \overset{\text{def.}}{=} \arg\min_{\mu} W_\alpha(\mu_t, \mu) + \tau f(\mu) \]

**Congestion-inducing function:**

\[ f(\mu) = \nu_{[0,\kappa]}(\mu) + \langle w, \mu \rangle \]

[Maury, Roudneff-Chupin, Santambrogio 2010]

**Proposition:** \( \text{Prox}_{\frac{1}{\varepsilon}} f(\mu) = \min(e^{-\varepsilon w} \mu, \kappa) \)

\[ \kappa = \|\mu_{t=0}\|_\infty \]

\[ \kappa = 2\|\mu_{t=0}\|_\infty \]

\[ \kappa = 4\|\mu_{t=0}\|_\infty \]
Crowd Motion on a Surface

\[ X = \text{triangulated mesh}. \]

\[ \kappa = \| \mu_{t=0} \|_\infty \]

\[ \kappa = 6 \| \mu_{t=0} \|_\infty \]

Potential \( \cos(w) \)
Crowd Motion on a Surface

\[ X = \text{triangulated mesh.} \]

\[ \kappa = \| \mu_{t=0} \|_{\infty} \]

\[ \kappa = 6 \| \mu_{t=0} \|_{\infty} \]

Potential \( \cos(w) \)
Gradient Flows: Crowd Motion with Obstacles

\[ X = \text{sub-domain of } \mathbb{R}^2. \]
Gradient Flows: Crowd Motion with Obstacles

\[ X = \text{sub-domain of } \mathbb{R}^2. \]
\[(\mu_{1,t+1}, \mu_{2,t+1}) \overset{\text{def.}}{=} \arg\min_{(\mu_1, \mu_2)} W_\alpha^\alpha(\mu_1, t, \mu_1) + W_\alpha^\alpha(\mu_2, t, \mu_2) + \tau f(\mu_1, \mu_2)\]
Multiple-Density Gradient Flows

$$(\mu_1,t+1, \mu_2,t+1) \overset{\text{def.}}{=} \arg\min_{(\mu_1, \mu_2)} W_\alpha^\mu(\mu_1,t, \mu_1) + W_\alpha^\mu(\mu_2,t, \mu_2) + \tau f(\mu_1, \mu_2)$$

Wasserstein attraction:

$$f(\mu_1, \mu_2) = W_\alpha^\mu(\mu_1, \mu_2) + h_1(\mu_1) + h_2(\mu_2)$$

$$\nabla w$$

$$h_i(\mu) = \langle w, \mu \rangle$$
Multiple-Density Gradient Flows

$$(\mu_{1,t+1}, \mu_{2,t+1}) \overset{\text{def.}}{=} \arg\min_{(\mu_1, \mu_2)} W_\alpha^{\mu_1}(\mu_{1,t}, \mu_1) + W_\alpha^{\mu_2}(\mu_{2,t}, \mu_2) + \tau f(\mu_1, \mu_2)$$

Wasserstein attraction:

$$f(\mu_1, \mu_2) = W_\alpha^{\mu_1}(\mu_1, \mu_2) + h_1(\mu_1) + h_2(\mu_2)$$

$$h_i(\mu) = \langle w, \mu \rangle$$  

$$h_i(\mu) = \iota_{[0,\kappa]}(\mu).$$
Overview

- Measures and Histograms
- From Monge to Kantorovich Formulations
- Entropic Regularization and Sinkhorn
- Barycenters
- Unbalanced OT and Gradient Flows
- Minimum Kantorovich Estimators
- Gromov-Wasserstein
Discriminative vs Generative Models

Generative
\[ g_\theta \]

Discriminative
\[ d_\xi \]

Low dimension

High dimension

\[ X \]

\[ Z \]
**Discriminative vs Generative Models**

**Generative**
- Model: $g_\theta$
- Mapping: $g_\theta \rightarrow \mathbf{x}$
- Low dimension
- Examples: Images

**Discriminative**
- Model: $d_\xi$
- Mapping: $d_\xi \rightarrow \mathbf{z}$
- High dimension
- Examples: Features

**Supervised: classification, $z =$ class probability**

→ Learn $d_\xi$ from labeled data $(\mathbf{x}_i, z_i)_i$.

Note: The diagram illustrates the difference between discriminative and generative models. Discriminative models learn a mapping from the input space to the feature space, while generative models learn a mapping from the feature space to the input space. The feature space is typically lower-dimensional than the input space, which can lead to more efficient and interpretable models.
**Discriminative vs Generative Models**

---

**Low dimension**

**Discriminative**

\[ d_\xi \]

**High dimension**

**Generative**

\[ g_\theta \]

---

**Supervised:** classification, \( z = \text{class probability} \)

\[ \rightarrow \text{Learn } d_\xi \text{ from labeled data } (x_i, z_i)_i. \]

**Un-supervised:** Compression: \( z = d_\xi(x) \) is a representation.

Generation: \( x = g_\theta(z) \) is a synthesis.

\[ \rightarrow \text{Learn } (g_\theta, d_\xi) \text{ from data } (x_i)_i. \]
Discriminative vs Generative Models

**Supervised:** classification, \( z = \text{class probability} \)

→ Learn \( d_\xi \) from labeled data \((x_i, z_i)_i\).

**Un-supervised:**

Compression: \( z = d_\xi(x) \) is a representation.

Generation: \( x = \theta_g(z) \) is a synthesis.

→ Learn \((\theta_g, d_\xi)\) from data \((x_i)_i\).

Density fitting

\( \theta_g(\{z_i\}_i) \approx \{x_i\}_i \)

Auto-encoders

\( \theta_g(d_\xi(x_i)) \approx x_i \)
Discriminative vs Generative Models

\[ \begin{align*}
\text{Low dimension} & \quad \text{High dimension} \\
\text{Generative} & \quad g_{\theta} \quad \text{Discriminative} & \quad d_{\xi}
\end{align*} \]

**Supervised:** classification, \( z = \text{class probability} \)

- Learn \( d_{\xi} \) from labeled data \((x_i, z_i)_i\).

**Un-supervised:** Compression: \( z = d_{\xi}(x) \) is a representation.

- Generation: \( x = g_{\theta}(z) \) is a synthesis.

- Learn \((g_{\theta}, d_{\xi})\) from data \((x_i)_i\).

**Density fitting**

\[ g_{\theta}(\{z_i\}_i) \approx \{x_i\}_i \]

**Optimal transport**

map \( d_{\xi} \)

**Auto-encoders**

\[ g_{\theta}(d_{\xi}(x_i)) \approx x_i \]
Density Fitting and Generative Models

Observations: $\nu = \frac{1}{n} \sum_{i=1}^{n} \delta x_i$

Parametric model: $\theta \mapsto \mu_\theta$
**Density Fitting and Generative Models**

Observations: \( \nu = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i} \)

Parametric model: \( \theta \mapsto \mu_\theta \)

Density fitting: \( d\mu_\theta(y) = f_\theta(y)dy \)

\[
\min_{\theta} \hat{\text{KL}}(\nu|\mu_\theta) \overset{\text{def.}}{=} - \sum_{j} \log(f_\theta(y_j))
\]

Maximum likelihood (MLE)
Density Fitting and Generative Models

Observations: \( \nu = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i} \)

Parametric model: \( \theta \mapsto \mu_{\theta} \)

\[
\text{Density fitting:} \quad \text{d} \mu_{\theta}(y) = f_{\theta}(y) \text{d}y
\]

\[
\min_{\theta} \hat{KL}(\nu|\mu_{\theta}) \overset{\text{def.}}{=} - \sum_{j} \log(f_{\theta}(y_j))
\]

Maximum likelihood (MLE)

Generative model fit: \( \mu_{\theta} = g_{\theta}, \# \zeta \)

\[
\hat{KL}(\mu_{\theta}|\nu) = +\infty
\]

\( \rightarrow \) MLE undefined.

\( \rightarrow \) Need a weaker metric.
Loss Functions for Measures

Density fitting: \[ \min_{\theta} D(\mu_\theta, \nu) \]

\[ \nu = \frac{1}{P} \sum_j y_j \]
\[ \mu = \frac{1}{N} \sum_i x_i \]

Optimal Transport Distances

\[ W(\mu, \nu)^p \overset{\text{def.}}{=} \min_{T \in \mathcal{C}_{\mu, \nu}} \sum_{i,j} T_{i,j} \| x_i - y_j \|^p \]
Loss Functions for Measures

Density fitting: \( \min_\theta D(\mu_\theta, \nu) \)

\[ \nu = \frac{1}{P} \sum_j \delta_{y_j} \]

\[ \mu = \frac{1}{N} \sum_i \delta_{x_i} \]

Optimal Transport Distances

\[ W(\mu, \nu)^p \overset{\text{def.}}{=} \min_{T \in \mathcal{C}_{\mu, \nu}} \sum_{i,j} T_{i,j} \|x_i - y_j\|^p \]

Maximum Mean Discrepancy (MMD)

\[ \|\mu - \nu\|^2_k = \frac{1}{N^2} \sum_{i,i'} k(x_i, x_{i'}) + \frac{1}{P^2} \sum_{j,j'} k(y_j, y_{j'}) - \frac{2}{NP} \sum_{i,j} k(x_i, y_j) \]

Gaussian: \( k(x, y) = e^{-\frac{\|x - y\|^2}{2\sigma^2}} \)

Energy distance: \( k(x, y) = -\|x - y\|^2 \).
Loss Functions for Measures

Density fitting: $\min_{\theta} D(\mu_\theta, \nu)$

\[ \nu = \frac{1}{P} \sum_j \delta_{y_j} \]

\[ \mu = \frac{1}{N} \sum_i \delta_{x_i} \]

Optimal Transport Distances

\[ W(\mu, \nu)^p \overset{\text{def.}}{=} \min_{T \in C_{\mu, \nu}} \sum_{i,j} T_{i,j} \| x_i - y_j \|^p \]

Maximum Mean Discrepancy (MMD)

\[ \| \mu - \nu \|^2_k \overset{\text{def.}}{=} \frac{1}{N^2} \sum_{i,i'} k(x_i, x_{i'}) + \frac{1}{P^2} \sum_{j,j'} k(y_j, y_{j'}) - \frac{2}{NP} \sum_{i,j} k(x_i, y_j) \]

Gaussian: $k(x, y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}}$. Energy distance: $k(x, y) = -\|x-y\|^2$.

Sinkhorn divergences [Cuturi 13]

\[ W_\varepsilon(\mu, \nu)^p \overset{\text{def.}}{=} \sum_{i,j} T_{i,j}^\varepsilon \| x_i - y_j \|^p \]

\[ \tilde{W}_\varepsilon(\mu, \nu)^p \overset{\text{def.}}{=} W_\varepsilon(\mu, \nu)^p - \frac{1}{2} W_\varepsilon(\mu, \mu)^p - \frac{1}{2} W_\varepsilon(\nu, \nu)^p \]
Loss Functions for Measures

Optimal Transport Distances

$$W(\mu, \nu)^p \overset{\text{def.}}{=} \min_{T \in C_{\mu, \nu}} \sum_{i,j} T_{i,j} \|x_i - y_j\|^p$$

Maximum Mean Discrepancy (MMD)

$$\|\mu - \nu\|^2_k \overset{\text{def.}}{=} \frac{1}{N^2} \sum_{i,i'} k(x_i, x_{i'}) + \frac{1}{P^2} \sum_{j,j'} k(y_j, y_{j'}) - \frac{2}{NP} \sum_{i,j} k(x_i, y_j)$$

Gaussian: $$k(x, y) = e^{-\frac{\|x - y\|^2}{2\sigma^2}}.$$ Energy distance: $$k(x, y) = -\|x - y\|^2.$$

Sinkhorn divergences [Cuturi 13]

$$W_\varepsilon(\mu, \nu)^p \overset{\text{def.}}{=} \sum_{i,j} T_{i,j}^\varepsilon \|x_i - y_j\|^p$$

$$\bar{W}_\varepsilon(\mu, \nu)^p \overset{\text{def.}}{=} W_\varepsilon(\mu, \nu)^p - \frac{1}{2} W_\varepsilon(\mu, \mu)^p - \frac{1}{2} W_\varepsilon(\nu, \nu)^p$$

Theorem: [Ramdas, G.Trillos, Cuturi 17] $$\bar{W}_\varepsilon(\mu, \nu)^p \overset{\varepsilon \to 0}{\to} W(\mu, \nu)^p$$

for $$k(x, y) = -\|x - y\|^p$$

Density fitting: $$\min_\theta D(\mu_\theta, \nu)$$

$$\nu = \frac{1}{P} \sum_j \delta_{y_j}$$

$$\mu = \frac{1}{N} \sum_i \delta_{x_i}$$
Loss Functions for Measures

Density fitting: \( \min_{\theta} D(\mu_\theta, \nu) \)

\[
\nu = \frac{1}{P} \sum_j \delta_{y_j}
\]

\[
\mu = \frac{1}{N} \sum_i \delta_{x_i}
\]

Optimal Transport Distances

\[
W(\mu, \nu)^p \overset{\text{def.}}{=} \min_{T \in \mathcal{C}_{\mu, \nu}} \sum_{i,j} T_{i,j} \|x_i - y_j\|^p
\]

Maximal Mean Discrepancy (MMD)

\[
\|\mu - \nu\|_k^2 \overset{\text{def.}}{=} \frac{1}{N^2} \sum_{i,i'} k(x_i, x_{i'}) + \frac{1}{P^2} \sum_{j,j'} k(y_j, y_{j'}) - \frac{2}{NP} \sum_{i,j} k(x_i, y_j)
\]

Gaussian: \( k(x, y) = e^{-\frac{\|x - y\|^2}{2\sigma^2}} \), Energy distance: \( k(x, y) = -\|x - y\|^2 \).

Sinkhorn divergences [Cuturi 13]

\[
W_\varepsilon(\mu, \nu)^p \overset{\text{def.}}{=} \sum_{i,j} T_{i,j}^\varepsilon \|x_i - y_j\|^p
\]

\[
\bar{W}_\varepsilon(\mu, \nu)^p \overset{\text{def.}}{=} W_\varepsilon(\mu, \nu)^p - \frac{1}{2} W_\varepsilon(\mu, \mu)^p - \frac{1}{2} W_\varepsilon(\nu, \nu)^p
\]

Theorem: [Ramdas, G.Trillos, Cuturi 17]

\[
\bar{W}_\varepsilon(\mu, \nu)^p \xrightarrow{\varepsilon \to 0} W(\mu, \nu)^p \quad \xrightarrow{\varepsilon \to +\infty} \|\mu - \nu\|_k^2
\]

for \( k(x, y) = -\|x - y\|^p \)

Best of both worlds:

\( \rightarrow \) cross-validate \( \varepsilon \)

- Scale free (no \( \sigma \), no heavy tail kernel).
- Non-Euclidean, arbitrary ground distance.
- Less biased gradient.
- No curse of dimension (low sample complexity).
Deep Discriminative vs Generative Models

Deep networks:

\[
d_\xi(x) = \rho(\xi_K(\ldots \rho(\xi_2(\rho(\xi_1(x))\ldots)
\]

\[
g_\theta(z) = \rho(\theta_K(\ldots \rho(\theta_2(\rho(\theta_1(z))\ldots)
\]
Deep Discriminative vs Generative Models

Deep networks:

\[ d_\xi(x) = \rho(\xi_K(\ldots \rho(\xi_2(\rho(\xi_1(x)) \ldots) \right) \]
\[ g_\theta(z) = \rho(\theta_K(\ldots \rho(\theta_2(\rho(\theta_1(z)) \ldots) \right) \]

Discriminative

Generative
Examples of Image Generation

[Credit ArXiv:1511.06434]
Overview

• Measures and Histograms
• From Monge to Kantorovitch Formulations
• Entropic Regularization and Sinkhorn
• Barycenters
• Unbalanced OT and Gradient Flows
• Minimum Kantorovitch Estimators
• Gromov-Wasserstein
Gromov-Wasserstein

Inputs: \{(similarity/kernel matrix, histogram)\}

\[
\begin{align*}
(d, \mu) & \quad \mu = \sum_i \mu_i \delta_{x_i} \quad d_{i,i'} = d(x_i, x_{i'}) \\
(\bar{d}, \nu) & \quad \nu = \sum_j \nu_j \delta_{y_j} \quad \bar{d}_{j,j'} = \bar{d}(y_j, y_{j'})
\end{align*}
\]
Gromov-Wasserstein

**Inputs:** \{(similarity/kernel matrix, histogram)\}

\[
(d, \mu) \quad \mu = \sum_i \mu_i \delta_{x_i} \quad d_{i,i'} = d(x_i, x_{i'})
\]

\[
(d, \nu) \quad \nu = \sum_j \nu_j \delta_{y_j} \quad \bar{d}_{j,j'} = \bar{d}(y_j, y_{j'})
\]

**Def. Gromov-Wasserstein distance:**

\[
GW_p(d, \mu, \bar{d}, \nu) \overset{\text{def.}}{=} \min_{T \in C_{\mu, \nu}} \mathcal{E}_p^{d, \bar{d}}(T)
\]

\[
\mathcal{E}_p^{d, \bar{d}}(T) \overset{\text{def.}}{=} \sum_{i,i',j,j'} |d_{i,i'} - \bar{d}_{j,j'}|^p T_{i,j} T_{i',j'}
\]
Gromov-Wasserstein

Inputs: \{(similarity/kernel matrix, histogram)\}

\[(d, \mu) \quad \mu = \sum_i \mu_i \delta_{x_i} \quad d_{i,i'} = d(x_i, x_{i'})\]

\[(\bar{d}, \nu) \quad \nu = \sum_j \nu_j \delta_{y_j} \quad \bar{d}_{j,j'} = \bar{d}(y_j, y_{j'})\]

Def. Gromov-Wasserstein distance:

\[GW^p_d,\mu,\bar{d},\nu \overset{\text{def.}}{=} \min_{T \in C_{\mu,\nu}} \mathcal{E}^p_{d,\bar{d}}(T)\]

\[\mathcal{E}^p_{d,\bar{d}}(T) \overset{\text{def.}}{=} \sum_{i,i',j,j'} |d_{i,i'} - \bar{d}_{j,j'}|^p T_{i,j} T_{i',j'}\]

Computation of GW is a QAP:

→ NP-hard in general.
→ need for a fast approximate solver.
Gromov-Wasserstein as a Metric

\[ \mu = \sum_i \mu_i \delta_{x_i} \in \mathcal{M}_+(X) \quad d_{i,i'} = d(x_i, x_{i'}) \]

\[ \nu = \sum_j \nu_j \delta_{y_j} \in \mathcal{M}_+(Y) \quad \bar{d}_{j,j'} = \bar{d}(y_j, y_{j'}) \]

**Def.** Metric-measured spaces \((X, \mu, d) \in \mathbb{M}: \quad \mu \in \mathcal{M}_+(X) \) and \(d\) is a distance on \(X\)
Gromov-Wasserstein as a Metric

$$\mu = \sum_i \mu_i \delta_{x_i} \in \mathcal{M}^1_+(X)$$
$$d_{i,i'} = d(x_i, x_{i'})$$

$$\nu = \sum_j \nu_j \delta_{y_j} \in \mathcal{M}^1_+(Y)$$
$$\bar{d}_{j,j'} = \bar{d}(y_j, y_{j'})$$

**Def.** Metric-measured spaces $$(X, \mu, d) \in \mathcal{M}$$:
$$\mu \in \mathcal{M}^1_+(X)$$ and $d$ is a distance on $X$

**Def.** Isometries on $\mathcal{M}$: $$(\mu, d) \sim (\nu, \bar{d})$$
$$\iff \exists f : X \to Y, \left\{ \begin{array}{l}
f_{\#} \mu = \nu, \\
d(x, x') = \bar{d}(f(x), f(x')). \end{array} \right.$$
Gromov-Wasserstein as a Metric

\[ \mu = \sum_i \mu_i \delta_{x_i} \in \mathcal{M}_+^1(X) \quad d_{i,i'} = d(x_i, x_{i'}) \]
\[ \nu = \sum_j \nu_j \delta_{y_j} \in \mathcal{M}_+^1(Y) \quad \bar{d}_{j,j'} = \bar{d}(y_j, y_{j'}) \]

**Def.** Metric-measured spaces \((X, \mu, d) \in \mathbb{M} : \mu \in \mathcal{M}_+^1(X) \) and \(d\) is a distance on \(X\)

**Def.** Isometries on \(\mathbb{M} : (\mu, d) \sim (\nu, \bar{d})\)
\[ \Longleftrightarrow \exists f : X \to Y, \left\{ \begin{array}{l}
    f_# \mu = \nu, \\
    d(x, x') = \bar{d}(f(x), f(x')).
\end{array} \right. \]

**Prop.** GW defines a distance on \(\mathbb{M}/\sim\), [Memoli 2011]

\(\longrightarrow\) “bending-invariant” objects recognition.
Metric-measure spaces $(X, Y): (d_X, \mu), (d_Y, \nu)$
For Arbitrary Spaces

**Metric-measure spaces** \((X, Y)\): \((d_X, \mu), (d_Y, \nu)\)

**Def. Gromov-Wasserstein distance:**

\[
GW^2_2(d_X, \mu, d_Y, \nu) \overset{\text{def.}}{=} \min_{\pi \in \Pi(\mu, \nu)} \int_{X^2 \times Y^2} |d_X(x, x') - d_Y(y, y')|^2 d\pi(x, y) d\pi(x', y')
\]

[Sturm 2012] [Memoli 2011]
For Arbitrary Spaces

Metric-measure spaces \((X, Y)\): \((d_X, \mu), (d_Y, \nu)\)

**Def.** Gromov-Wasserstein distance:

\[
GW^2_2(d_X, \mu, d_Y, \nu) \overset{\text{def.}}{=} \min_{\pi \in \Pi(\mu, \nu)} \int_{X^2 \times Y^2} |d_X(x, x') - d_Y(y, y')|^2 d\pi(x, y) d\pi(x', y')
\]

[Sturm 2012] [Memoli 2011]

**Prop.** GW is a distance on mm-spaces/isometries.

→ “bending-invariant” objects recognition.
For Arbitrary Spaces

**Metric-measure spaces** \((X, Y)\): \((d_X, \mu), (d_Y, \nu)\)

**Def.** Gromov-Wasserstein distance:

\[
\text{GW}_2^2(d_X, \mu, d_Y, \nu) \overset{\text{def.}}{=} \min_{\pi \in \Pi(\mu, \nu)} \int_{X^2 \times Y^2} |d_X(x, x') - d_Y(y, y')|^2 d\pi(x, y) d\pi(x', y')
\]

[Sturm 2012] [Memoli 2011]

**Prop.** GW is a distance on mm-spaces/isometries.

→ “bending-invariant” objects recognition.
→ QAP: NP-hard in general.
→ need for a fast approximate solver.
Entropic Gromov Wasserstein

**Def.** *Entropic Gromov-Wasserstein*

\[ GW^p_{p,\varepsilon}(d, \mu, \bar{d}, \nu) \overset{\text{def.}}{=} \min_{T \in C_{\mu,\nu}} E^p_{d,\bar{d}}(T) - \varepsilon H(T) \]
Entropic Gromov Wasserstein

**Def.** Entropic Gromov-Wasserstein

\[ \text{GW}^p_{p, \varepsilon}(d, \mu, \bar{d}, \nu) \overset{\text{def.}}{=} \min_{T \in \mathcal{C}_{\mu, \nu}} \mathcal{E}^p_{d, \bar{d}}(T) - \varepsilon H(T) \]

**Def.** Projected mirror descent:

\[ T \leftarrow \text{Proj}^{\text{KL}}_{\mathcal{C}_{\mu, \nu}} \left( T \odot e^{-\tau(-\nabla \mathcal{E}^p_{d, \bar{d}}(T) - \varepsilon \nabla H(T))} \right) \]

where \( \text{Proj}^{\text{KL}}_{\mathcal{C}_{\mu, \nu}}(K) \overset{\text{def.}}{=} \text{argmin}_T \{ \text{KL}(T|K) ; T \in \mathcal{C}_{\mu, \nu} \} \)

**Prop.** for \( \tau = 1/\varepsilon \), the iteration reads

\[ T \leftarrow \text{Sinkhorn}(\mu, \nu, -d \times T \times \bar{d}) \]

**Prop.** \( T \) converges to a stationary point.

**func** \( T = \text{GW}(C, \bar{C}, p, q) \)

**initialization:**

\[ T \leftarrow \mu \nu^\top \]

**repeat:**

\[ D \leftarrow -d \times T \times \bar{d} \]

\[ T \leftarrow \text{Sinkhorn}(\mu, \nu, D) \]

**until convergence.**

**return** \( T \)
Applications of GW: Shapes Analysis

Use $T$ to define registration between:

Shape $\leftrightarrow$ Shape

Colors distribution $\leftrightarrow$ Shape
Applications of GW: Shapes Analysis

Use $T$ to define registration between:

\[
\text{Shapes} (X_s)_s \rightarrow \text{Geodesic distances} d_s = (D_{X_s}(x_i, x_i'))_{i,i'} \rightarrow \text{GW distances} (GW_\varepsilon(d_s, d_{s'}))_{s,s'} \rightarrow \text{MDS Visualization}
\]
Applications of GW: Quantum Chemistry

*Input*: Molecules with positions and charges \( \mu = \sum_i \mu_i \delta_{x_i} \).

*Regression problem*: approximate ground state energy \( \mu \mapsto f(\mu) \).

\( f \) by solving DFT approximation is too costly.
Applications of GW: Quantum Chemistry

**Input:** Molecules with positions and charges \( \mu = \sum_i \mu_i \delta_{x_i} \).

**Regression problem:** approximate ground state energy \( \mu \mapsto f(\mu) \). \( \rightarrow f \) by solving DFT approximation is too costly.

**Coulomb matrices** \( d = d(\mu) \):

\[
d_{i,i'} \overset{\text{def.}}{=} \begin{cases} \frac{\mu_i \mu_{i'}}{\|x_i - x_{i'}\|} & \text{for } (i \neq i') \\ \frac{1}{2} \mu_i^2.4 & \text{for } (i = i') \end{cases}
\]

[Rupp et al 2012]
Applications of GW: Quantum Chemistry

Input: Molecules with positions and charges \( \mu = \sum_i \mu_i \delta_{x_i} \).

Regression problem: approximate ground state energy \( \mu \mapsto f(\mu) \).

→ \( f \) by solving DFT approximation is too costly.

\textbf{Coulomb matrices} \( d = d(\mu) \):

\[
d_{i,i'} \overset{\text{def.}}{=} \begin{cases} 
\frac{\mu_i \mu_{i'}}{\|x_i - x_{i'}\|} & \text{for } (i \neq i') \\
\frac{1}{2} \mu_i^2 & \text{for } (i = i').
\end{cases}
\]

Learning: \((\mu_s, f(\mu_s))_s \to \text{approximation } \tilde{f}.

\textbf{GW-interpolation:} \( \tilde{f}(\mu) = f(\mu_{s^*}) \)

\( s^* = \arg\min_s \text{GW}(d(\mu), d(\mu_s)) \)

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>( |f - \tilde{f}|_1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( k )-nearest neighbors</td>
<td>71.54</td>
</tr>
<tr>
<td>Linear regression</td>
<td>20.72</td>
</tr>
<tr>
<td>Gaussian kernel ridge regression</td>
<td>8.57</td>
</tr>
<tr>
<td>Laplacian kernel ridge regression (8)</td>
<td>3.07</td>
</tr>
<tr>
<td>Multilayer Neural Network (1000)</td>
<td>3.51</td>
</tr>
<tr>
<td>GW ( 3 )-nearest neighbors</td>
<td>10.83</td>
</tr>
</tbody>
</table>
Gromov-Wasserstein Geodesics

**Def.**  \(Gromov-Wasserstein Geodesic\)

\[
(\mu_t, d_t) \in \arg \min_{(\mu, d) \in \mathcal{X}} (1 - t)GW^2_2(\mu_0, d_0, \mu, d) + tGW^2_2(\mu_1, d_1, \mu, d)
\]
**Def.**  \( Gromov\text{-}Wasserstein Geodesic \)

\[
(\mu_t, d_t) \in \arg\min_{(\mu,d)\in\mathcal{X}} (1 - t)GW_2^2(\mu_0, d_0, \mu, d) + tGW_2^2(\mu_1, d_1, \mu, d)
\]

**Optimal coupling** \( T^* \):

\[
GW_2^2(d_0, \mu_0, d_1, \mu_1) \overset{\text{def.}}{=} \mathcal{E}^2_{d_0,d_1}(T^*)
\]

**Prop.** One can define \((\mu_t, d_t)\) on \(X \times Y\) as

\[
\mu_t = \sum_{i,j} T_{i,j}^* \delta_{x_i,y_j}
\]

\[
d_t((x, y), (x', y')) = (1 - t)d_0(x, x') + td_1(y, y')
\]

[Sturm 2012]
Gromov-Wasserstein Geodesics

**Def.** *Gromov-Wasserstein Geodesic*

\[(\mu_t, d_t) \in \arg\min_{(\mu, d) \in X} (1 - t)GW^2_2(\mu_0, d_0, \mu, d) + tGW^2_2(\mu_1, d_1, \mu, d)\]

Optimal coupling \(T^*:\)

\[GW^2_2(d_0, \mu_0, d_1, \mu_1) \overset{\text{def.}}{=} E^2_{d_0, d_1}(T^*)\]

**Prop.** One can define \((\mu_t, d_t)\) on \(X \times Y\) as

\[
\mu_t = \sum_{i,j} T_{i,j}^* \delta_{x_i, y_j}
\]

\[d_t((x, y), (x', y')) = (1 - t)d_0(x, x') + td_1(y, y')\]

[Sturm 2012]

\(\rightarrow X \times Y\) is not practical for most applications.

(need to fix the size of the geodesic embedding space)

\(\rightarrow\) Extension to more than 2 input spaces?
Gromov-Wasserstein Barycenters

Input: Measures \((\mu_s)_s\), matrices \((d_s)_s\)
Weights \(\lambda\), size \(N\), \(\mu \in \mathbb{R}^N_+\) probability vector

Def. GW Barycenters

$$\min_{d \in \mathbb{R}^{N \times N}} \sum_s \lambda_s GW^2_{2,\varepsilon}(d_s, \mu_s, d, \mu)$$
Gromov-Wasserstein Barycenters

**Input:** Measures $(\mu_s)_s$, matrices $(d_s)_s$
Weights $\lambda$, size $N$, $\mu \in \mathbb{R}_+^N$ probability vector

**Def. GW Barycenters**

\[
\min_{d \in \mathbb{R}^{N \times N}} \sum_s \lambda_s \text{GW}^2_{2,\varepsilon}(d_s, \mu_s, d, \mu)
\]

\[
\min_{d, (T_s)_s} \left\{ \sum_s \lambda_s \left( \mathcal{E}_{d,d_s}^2(T_s) - \varepsilon H(T_s) \right) ; \forall s, T_s \in \mathcal{C}_{\mu,\mu_s} \right\}
\]

**Alternating minimization:**

```
func C = GW-bary(d_s, \mu_s, \mu)_s

initialization: C \leftarrow C_0
repeat:
    for s = 1 to S do
        T_s \leftarrow \text{GW}(d, \mu, d_s, \mu_s)
    d \leftarrow \frac{1}{\mu \mu^\top} \sum \lambda_s T_s^\top d_s T_s
until convergence.
return C
```
Conclusion: Toward High-dimensional OT

Optimal transport framework

Applications

Application to Color Transfer

Source image ($X$)

Style image ($Y$)

Sliced Wasserstein projection of $X$ to style image color statistics

Source image after color transfer

J. Rabin

Wasserstein Regularization