Optimal Transport

http://bicmr.pku.edu.cn/~wenzw/bigdata2023.html

Acknowledgement: this slides is based on Prof. Gabriel Peyré's lecture notes

Outline

A Geometric Motivation

Setting: Probability measures $\mathcal{P}(\mathcal{X})$ on a metric space $(\mathcal{X}, \mathrm{dist})$.

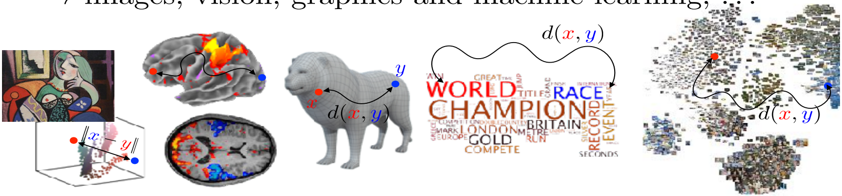
distance between μ and ν:

- $\mu = \delta_{x_1}$ and $\nu = \delta_{y_1}$
  $\mathrm{dist}(\mu, \nu) = \mathrm{dist}(x_1, y_1)$
- $\mu = \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$ and $\nu = \frac{1}{n} \sum_{i=1}^n \delta_{y_i}$
  $\mathrm{dist}(\mu, \nu) = \frac{1}{n^2} \sum_{ij} \mathrm{dist}(x_i, y_j)$? or
  $\mathrm{dist}(\mu, \nu) = \min_{\sigma \text{ permutation}} \frac{1}{n} \sum_i \mathrm{dist}(x_i, y_{\sigma(i)})$
- What if $\mu, \nu \in \mathcal{P}(\mathcal{X})$?

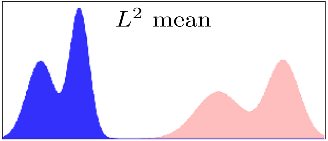
Goal: Build a metric on $\mathcal{P}(\mathcal{X})$ with the geometry of $(\mathcal{X}, \mathrm{dist})$.

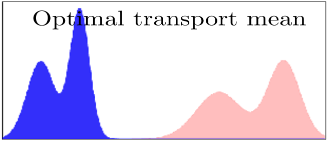
Comparing Measures

→ images, vision, graphics and machine learning, ...

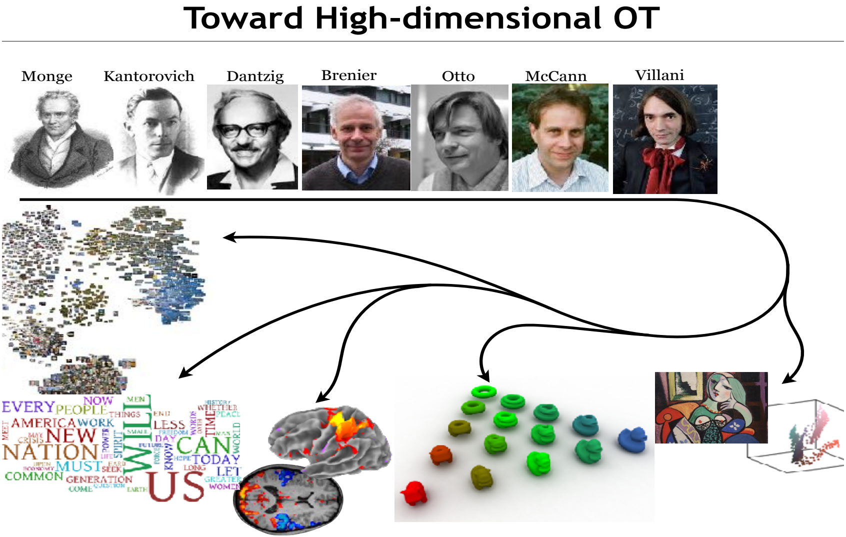


- *Optimal transport*
  → takes into account a metric d.





Applications: toward high-dimensional OT



Kantorovitch's Formulation

Discrete Optimal Transport

Input two discrete probability measures

$$\alpha = \sum_{i=1}^{m} a_i \delta_{x_i}, \quad \beta = \sum_{j=1}^{n} b_j \delta_{y_j}. \tag{1}$$

- $X = \{x_i\}_i$, $Y = \{x_j\}_j$: are given points clouds, x_i, y_i are vectors.
- a_i, b_j : positive weights, $\sum_{i=1}^{m} a_i = \sum_{j=1}^{n} b_j = 1$.
- C_{ij}: costs, $C_{ij} = c(x_i, y_j) \geq 0$.

Couplings

$$\mathbf{U}(\alpha, \beta) \stackrel{\text{def}}{=} \{\Pi \in \mathbb{R}_+^{m \times n}; \Pi \mathbf{1}_n = a, \Pi^\top \mathbf{1}_m = b\} \tag{2}$$

is called the set of couplings with respect to α and β.

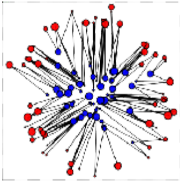
Kantorovitch's Formulation

Discrete Optimal Transport

In the optimal transport, we want to compute the following quantity [Kantorovich 1942]

Optimal transport distance

$$\mathcal{L}(\alpha, \beta, C) \stackrel{\text{def}}{=} \min \left\{ \sum_{i,j} C_{i,j} \Pi_{i,j}; \Pi \in \mathbf{U}(a, b) \right\}. \tag{3}$$



Push Forward

- Radon measures (α, β) on $(\mathcal{X}, \mathcal{Y})$.
- Transfer of measure by $T : \mathcal{X} \to \mathcal{Y}$: push forward.
- The measure $T_{\#}\alpha$ on $\mathcal{Y}$ is defined by

$$T_{\#}\alpha(Y) = \alpha(T^{-1}(Y)), \quad \text{for all measurable } Y \in \mathcal{Y}. \tag{4}$$

Equivalently,

$$\int_{\mathcal{Y}} g(y) \mathrm{d}T_{\#}\alpha(y) \stackrel{\text{def}}{=} \int_{\mathcal{X}} g(T(x)) \mathrm{d}\alpha(x). \tag{5}$$

- Discrete measures: $T_{\#}\alpha = \sum_i \alpha_i \delta_{T(x_i)}$
- Smooth densities: $\mathrm{d}\alpha = \rho(x)\mathrm{d}x$, $\mathrm{d}\beta = \xi(x)\mathrm{d}x$.

$$T_{\#}\alpha = \beta \iff \rho(T(x))|\mathbf{det}(\partial T(x))| = \xi(x). \tag{6}$$

Monge problem

- Monge problem seeks for a map that associates to each point x_i a single point y_j, and which must push the mass of α toward the mass of β, namely:

$$\forall j, \quad b_j = \sum_{i:T(x_i)=y_j} a_i$$

- Discrete case:

$$\min_T \quad \sum_i c(x_i, T(x_i)), \quad \text{s.t.} \quad T_\# \alpha = \beta$$

- Arbitrary measures:

$$\min_T \quad \int_{\mathcal{X}} c(x, T(x)) d\alpha(x), \quad \text{s.t.} \quad T_\# \alpha = \beta$$

Couplings between General Measures

Projectors:

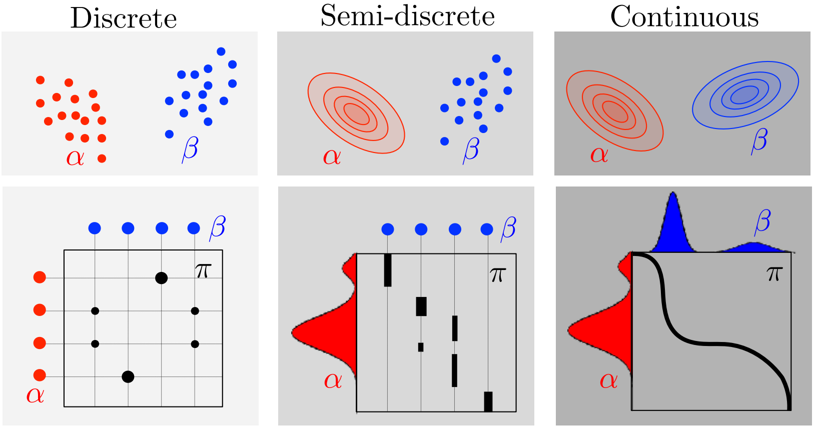
$$\begin{aligned} P_{\mathcal{X}} &: (x, y) \in \mathcal{X} \times \mathcal{Y} \to x \in \mathcal{X}, \\ P_{\mathcal{Y}} &: (x, y) \in \mathcal{X} \times \mathcal{Y} \to y \in \mathcal{Y}. \end{aligned} \tag{7}$$

Couplings between General Measures

$$\mathcal{U}(\alpha, \beta) \stackrel{\text{def}}{=} \{\pi \in \mathcal{M}_+(\mathcal{X} \times \mathcal{Y}); P_{\mathcal{X}\#}\pi = \alpha, P_{\mathcal{Y}\#}\pi = \beta\}. \tag{8}$$

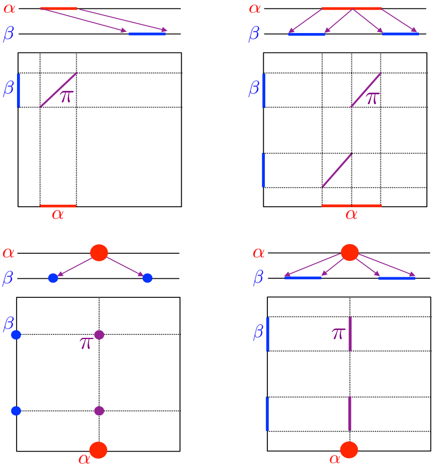
is called the set of couplings with respect to α and β.

Couplings: the 3 Settings



More Examples

Examples of Couplings



Kantorovitch Problem for General Measures

Optimal transport distance between General Measures

$$\mathcal{L}(\alpha, \beta, c) \stackrel{\text{def}}{=} \min_{\pi \in \mathcal{U}(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y) \mathrm{d}\pi(x, y). \tag{9}$$

Probability interpretation:

$$\min_{(X,Y)} \{\mathbb{E}_{(X,Y)}(c(X, Y)), X \sim \alpha, Y \sim \beta\}. \tag{10}$$

Wasserstein Distance

Metric Space $\mathcal{X} = \mathcal{Y}$.
Distance $d(x, y)$ (nonegative, symmetric, identity, triangle inequality).
Cost $c(x, y) = d(x, y)^p$, $p \geq 1$.

Wasserstein Distance

$$\mathcal{W}_p(\alpha, \beta) \stackrel{\text{def}}{=} \mathcal{L}(\alpha, \beta, d^p)^{1/p}. \tag{11}$$

Theorem

$\mathcal{W}_p$ is a distance, and

$$\mathcal{W}_p(\alpha_n, \alpha) \to 0 \iff \alpha_n \stackrel{\text{weak}}{\to} \alpha. \tag{12}$$

Example

$$\mathcal{W}_p(\delta_x, \delta_y) = d(x, y). \tag{13}$$

Dual form

Dual problem (discrete case)

$$\begin{aligned} \max_{w \in \mathbb{R}^m, r \in \mathbb{R}^n} \quad & w^\top a + r^\top b, \\ \text{s.t.} \quad & w_i + r_j \leq C_{ij}, \quad \forall (i,j) \end{aligned} \tag{14}$$

Relation between any primal and dual solutions:

$$P_{ij} > 0 \Rightarrow w_i + r_j = C_{ij}.$$

Wasserstein barycenter

- Define $C \stackrel{\text{def}}{=} M_{XY}$, where $(M_{XY})_{ij} = d(x_i, y_i)^p$. The Wasserstein distance as

$$\mathcal{L}(a, b, C) \stackrel{\text{def}}{=} \min \left\{ \sum_{i,j} C_{i,j} \Pi_{i,j}; \Pi \in \mathbf{U}(a, b) \right\}. \tag{15}$$

- Given a set of point clouds and their corresponding probability vector $\{(Y^i, b^i)\}$, $i = 1, \ldots, N$.
- Find a support $X = \{x_i\}$ with a probability vector a such that (X, a) is the optimal solution of the following problem

$$\min_{X,a} \sum_{k=1}^{N} \lambda_k \mathcal{L}(a, b^k, M_{XY^k}),$$

where $\sum_k \lambda_k = 1$ and $\lambda_k \geq 0$.

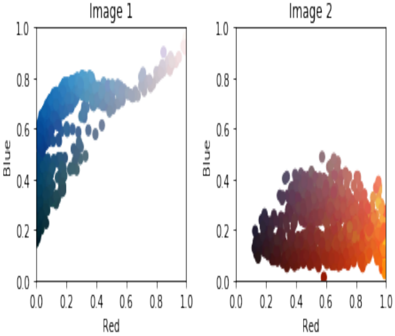
Outline

Applications: image color adaptation

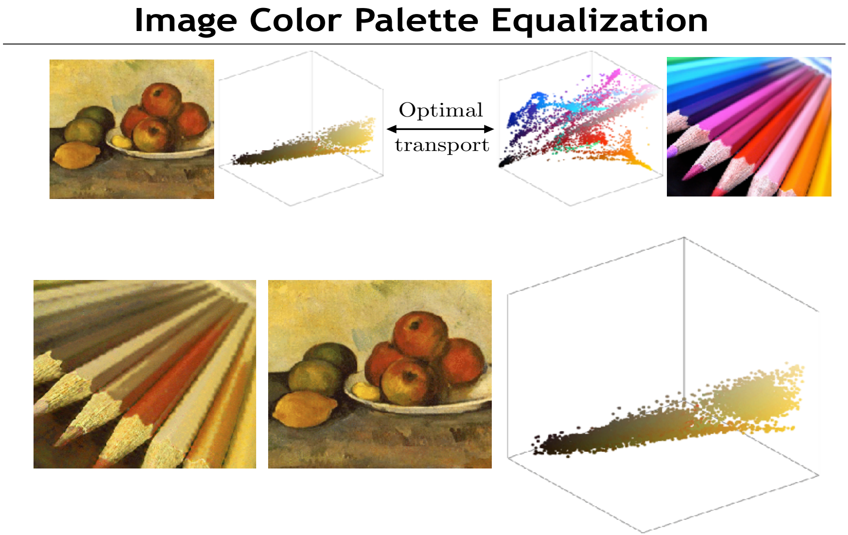
Example: https://pythonot.github.io/auto_examples/domain-adaptation/plot_otda_color_images.html

```
Given color image stored in the RGB format: I1, I2
# Converts an image to matrix (one pixel per line)
X1 = im2mat(I1), X2 = im2mat(I2)
# Take samples
Xs = X1[idx1, :], Xt = X2[idx2, :]
# Scatter plot of colors
pl.scatter(Xs[:, 0], Xs[:, 2], c=Xs)
# Sinkhorn Transport
ot_sinkhorn = ot.da.SinkhornTransport(reg_e=1e-1)
ot_sinkhorn.fit(Xs=Xs, Xt=Xt)
# prediction between images
transp_Xs_sinkhorn = ot_sinkhorn.transform(Xs=X1)
transp_Xt_sinkhorn = ot_sinkhorn.inverse_transform(Xt=X2)
```

Applications: image color adaptation

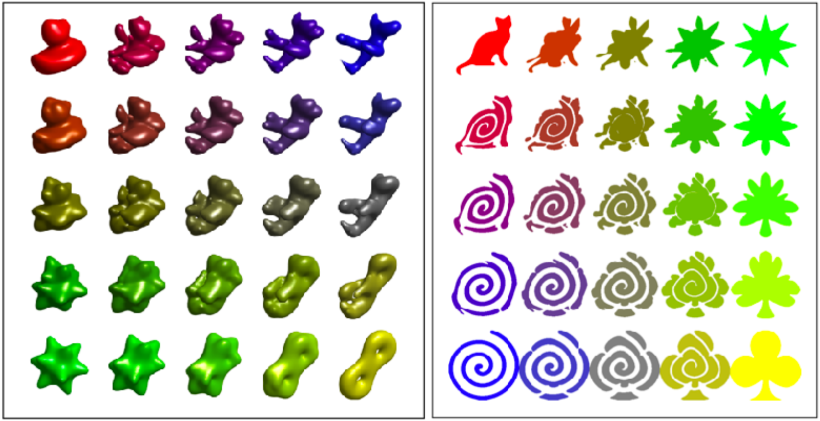


Applications: image color palette equalization



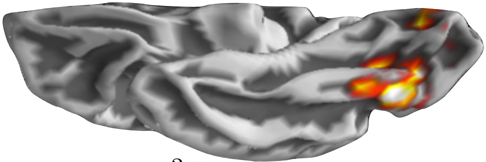
Applications: shape interpolation

Shape Interpolation

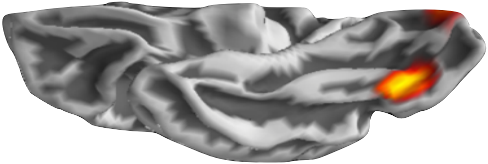


MRI Data Procesing [with A. Gramfort]

Ground cost $c = d_M$: geodesic on cortical surface M.



L^2 barycenter

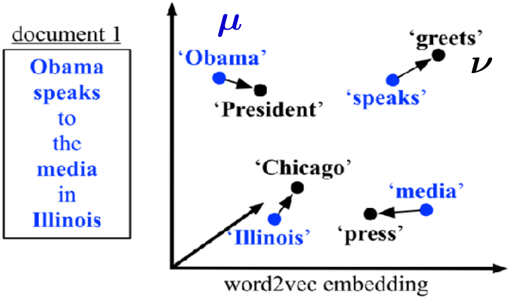


W_2^2 barycenter

Applications: word mover's distance

normalized bag-of-words (nBOW), word travel cost (word2vec distance), document distance $T_{ij}c(i,j)$, transportation problem

Bag of Words



[Kusner'15] $\mathrm{dist}(D_1, D_2) = W_2(\boldsymbol{\mu}, \boldsymbol{\nu})$

Applications: word mover's distance

$$\min_{\Pi \geq 0} \sum_{ij} \Pi_{ij} c_{ij}$$

$$\text{s.t.} \sum_{j=1}^{n} \Pi_{ij} = d_i$$

$$\sum_{i=1}^{n} \Pi_{ij} = d'_j$$

- x_i: word2vec embedding
- $c_{ij} = \|x_i - x_j\|_2$
- if word i appears w_i times in the document, we denote $d_i = \frac{w_i}{\sum w_j}$

Distributional Robust Optimization (DRO)

- stochastic optimization:

$$\inf_{\beta \in B} E_{P^*}[\ell(\beta^\top X)],$$

  where B is a convex set, ℓ is a loss function, $E_{P^*}[\cdot]$ represents the expectation operator associated to the probability model P^*, which describes the random element X.

- The DRO model:

$$\inf_{\beta \in B} \sup_{P \in \mathcal{U}_\delta(P_0)} E_P[\ell(\beta^\top X)],$$

  where $\mathcal{U}_\delta(P_0)$ is a so-called distributional uncertainty region "centered" around some benchmark model, P_0, which may be data-driven (for example, an empirical distribution) and δ >0 parameterizes the sizeof the distributional uncertainty.

- Wasserstein distance: $\mathcal{U}_\delta(P_0) = \{P \mid \mathcal{W}(P, P_0) \leq \delta\}$.

Outline

Discrete OT Review

Given an integer $n \geqslant 1$, we write Σ_n for the discrete probability simplex

$$\Sigma_n \overset{\text{def}}{=} \left\{ a \in \mathbb{R}_n^+ ; \sum_{i=1}^{n} a_i = 1. \right\} \tag{16}$$

Given $a \in \Sigma_m$, $b \in \Sigma_n$, the Optimal Transport problem is to compute

$$L(a, b, C) \overset{\text{def}}{=} \min\{ \sum_{i,j} C_{i,j} \mathbf{P}_{i,j}; \ \ \text{s.t.} \ \ \mathbf{P} \in \mathbf{U}(a, b) \}. \tag{17}$$

Where $\mathbf{U}(a, b)$ is the set of couplings between a and b.

Entropy

The discrete entropy of a positive matrix $\mathbf{P}$ ($\sum_{ij} \mathbf{P}_{ij} = 1$) is defined as

$$H(\mathbf{P}) \stackrel{\text{def}}{=} -\sum_{i,j} \mathbf{P}_{i,j}(\log(\mathbf{P}_{i,j}) - 1). \tag{18}$$

For a positive vector $u \in \Sigma_n$, the entropy is defined analogously:

$$H(\mathbf{u}) \stackrel{\text{def}}{=} -\sum_{i} \mathbf{u}_i(\log(\mathbf{u}_i) - 1). \tag{19}$$

For two positive vector $u, v \in \Sigma_n$, the Kullback-Leibler divergence (or, KL divergence) is defined to be

$$\mathbf{KL}(u\|v) = -\sum_{i=1}^{n} u_i \log(\frac{v_i}{u_i}). \tag{20}$$

The KL divergence is always non-negative: $\mathbf{KL}(u\|v) \geq 0$ (Jensen's inequality: $E[f(g(X))] \geq f(E[g(X)])$).

Entropic regularization

- Given $a \in \Sigma_m$, $b \in \Sigma_n$ and cost matrix $\mathbf{C} \in \mathbb{R}_+^{m \times n}$. The entropic regularization of the transportation problem reads

$$L^{\varepsilon}(a, b, \mathbf{C}) = \min_{\mathbf{P} \in \mathbf{U}(a,b)} \langle \mathbf{P}, \mathbf{C} \rangle - \varepsilon H(\mathbf{P}). \tag{21}$$

- The case $\varepsilon = 0$ corresponds to the classic (linear) optimal transport problem.
- For $\varepsilon > 0$, problem (21) has an ε-strongly convex objective and therefore admits a unique optimal solution $\mathbf{P}_{\varepsilon}^{\star}$.

- This is not (necessarily) true for $\varepsilon = 0$. But we have the following proposition.

Entropic regularization

Proposition

When $\varepsilon \to 0$, the unique solution P_ε of (21) converges to the optimal solution with maximal entropy within the set of all optimal solutions of the unregularized transportation problem, namely,

$$\mathbf{P}_\varepsilon \overset{\varepsilon \to 0}{\to} \text{argmax}_{\mathbf{P}}\{H(\mathbf{P}); \mathbf{P} \in U(a,b), \langle \mathbf{P}, \mathbf{C} \rangle = L^0(a,b,\mathbf{C})\} \tag{22}$$

The above proposition motivates us to solve the problems in (21) sequentially and then take $\epsilon \to 0$.

Entropic regularization

Proof

We consider a sequence $(\varepsilon_\ell)_\ell$ such that $\varepsilon_\ell \to 0$ and $\varepsilon_\ell > 0$. We denote $\mathbf{P}_\ell = \mathbf{P}^\star_{\varepsilon_\ell}$. Since $\mathbf{U}(a,b)$ is bounded, we can extract a sequence (that we do not relabel for the sake of simplicity) such that $\mathbf{P}_\ell \to \mathbf{P}^\star$. Since $\mathbf{U}(a,b)$ is closed, $\mathbf{P}^\star \in \mathbf{U}(a,b)$. We consider any $\mathbf{P}$ such that $\langle \mathbf{C}, \mathbf{P} \rangle = L^0(a,b,\mathbf{C})$. By optimality of $\mathbf{P}$ and $\mathbf{P}_\ell$ for their respective optimization problems (for $\varepsilon = 0$ and $\varepsilon = \varepsilon_\ell$), one has

$$0 \leq \langle \mathbf{C}, \mathbf{P}_\ell \rangle - \langle \mathbf{C}, \mathbf{P} \rangle \leq \varepsilon_\ell (H(\mathbf{P}_\ell) - H(\mathbf{P})). \tag{23}$$

Since H is continuous, taking the limit $\ell \to +\infty$ in this expression shows that $\langle C, \mathbf{P}^\star \rangle = \langle C, \mathbf{P} \rangle$. Furthermore, dividing by ε_ℓ and taking the limit shows that $H(\mathbf{P}) \leqslant H(\mathbf{P}^\star)$. Now the result follows from the strictly convexity of $-H$.

Entropic regularization

By the concavity of entropy, for $\alpha > 0$, we introduce the convex set

$$\begin{aligned} \mathbf{U}_\alpha(a,b) &\overset{\text{def}}{=} \{\mathbf{P} \in \mathbf{U}(a,b) | \mathbf{KL}(\mathbf{P} \| ab^\top) \le \alpha\} \\ &= \{\mathbf{P} \in \mathbf{U}(a,b) | H(\mathbf{P}) \ge H(a) + H(b) - 1 - \alpha\}. \end{aligned} \tag{24}$$

Definition: Sinkhorn Distance

$$d_{\mathbf{C},\alpha}(a,b) \overset{\text{def}}{=} \min_{\mathbf{P} \in \mathbf{U}_\alpha(a,b)} \langle \mathbf{C}, \mathbf{P} \rangle. \tag{25}$$

Proposition

For $\alpha \ge 0$, $d_{\mathbf{C},\alpha}(a,b)$ is symmetric and satisfies all triangle inequalities. Moreover, $\mathbf{1}_{a \neq b} d_{\mathbf{C},\alpha}(a,b)$ satisfies all three distance axioms.

Entropic regularization

Proposition

For α large enough, the Sinkhorn distance $d_{C,\alpha}$ is the transport distance $d_{\mathbf{C}}$.

Proof.

Note that for any $\mathbf{P} \in U(a, b)$, we have

$$H(\mathbf{P}) \geq \frac{1}{2}(H(a) + H(b)), \tag{26}$$

so for $\alpha \geq \frac{1}{2}(H(a) + H(b)) - 1$, we have

$$U_\alpha(a, b) = U(a, b).$$

Outline

Sinkhorn's algorithm

For solving (21), consider its Lagrangian dual function

$$\mathcal{L}_{\mathbf{C}}^{\varepsilon}(\mathbf{P}, w, r) = \langle \mathbf{C}, \mathbf{P} \rangle - \varepsilon H(\mathbf{P}) + w^{\top}(\mathbf{P}\mathbf{1}_n - a) + r^{\top}(\mathbf{P}^{\top}\mathbf{1}_m - b). \tag{27}$$

Now let $\partial \mathcal{L}_{\mathbf{C}}^{\varepsilon} / \partial \mathbf{P}_{ij} = 0$, i.e.,

$$\mathbf{P}_{ij} = e^{-\frac{c_{ij} + w_i + r_j}{\varepsilon}}, \tag{28}$$

so we can write

$$\mathbf{P}_{\varepsilon} = \textbf{diag}(e^{-\frac{w}{\varepsilon}}) e^{-\frac{\mathbf{C}}{\varepsilon}} \textbf{diag}(e^{-\frac{r}{\varepsilon}}). \tag{29}$$

Note that

$$\mathbf{P}_{\varepsilon}\mathbf{1}_n = a, \quad \mathbf{P}_{\varepsilon}^{\top}\mathbf{1}_m = b, \tag{30}$$

we can then use Sinkhorn's algorithm to find $\mathbf{P}_{\varepsilon}$!

Let $u = e^{-\frac{w}{\varepsilon}}$, $v = e^{-\frac{r}{\varepsilon}}$ and $\mathbf{K} = e^{-\mathbf{C}/\varepsilon}$. We again state the KKT system of (21):

$$
\begin{aligned}
\mathbf{P}_\varepsilon &= \textbf{diag}(u)\mathbf{K}\textbf{diag}(v), \\
a &= \textbf{diag}(u)\mathbf{K}v, \\
b &= \textbf{diag}(v)\mathbf{K}^\top u.
\end{aligned} \tag{31}
$$

Then the Sinkhorn's algorithm amounts to alternating updates in the form of

$$
\begin{aligned}
u^{(k+1)} &= \textbf{diag}(\mathbf{K}v^{(k)})^{-1}a, \\
v^{(k+1)} &= \textbf{diag}(\mathbf{K}^\top u^{(k+1)})^{-1}b.
\end{aligned} \tag{32}
$$

Sinkhorn's algorithm

1. Compute $\mathbf{K} = e^{-\frac{\mathbf{C}}{\varepsilon}}$.
2. Compute $\hat{\mathbf{K}} = \mathbf{diag}(a^{-1})\mathbf{K}$.
3. Initial scale factor $u \in \mathbb{R}^m$.
4. Iteratively update u:

$$u = 1./(\hat{\mathbf{K}}(b./(\mathbf{K}^\top u))),$$

   until reaches certain stopping criterion.
5. Compute

$$v = b./(\mathbf{K}^\top u),$$

   and eventually

$$\mathbf{P}_\varepsilon = \mathbf{diag}(u)\mathbf{K}\mathbf{diag}(v).$$

Outline

Sinkhorn-Newton method

The dual problem of (21) is

$$\begin{aligned} \min_{w,r} \quad & \langle a, w\rangle + \langle b, r\rangle + \varepsilon\langle e^{-\frac{w}{\varepsilon}}, \mathbf{K}e^{-\frac{r}{\varepsilon}}\rangle, \\ \text{s.t.} \quad & \textbf{diag}(e^{-\frac{w}{\varepsilon}})\mathbf{K}e^{-\frac{r}{\varepsilon}} = a, \\ & \textbf{diag}(e^{-\frac{r}{\varepsilon}})\mathbf{K}^\top e^{-\frac{w}{\varepsilon}} = b. \end{aligned} \tag{33}$$

with w, r being the dual variables.

Sinkhorn-Newton method

Let

$$F(w,r) = \begin{pmatrix} \textbf{diag}(e^{-\frac{w}{\varepsilon}})\mathbf{K}e^{-\frac{r}{\varepsilon}} - a \\ \textbf{diag}(e^{-\frac{r}{\varepsilon}})\mathbf{K}^\top e^{-\frac{w}{\varepsilon}} - b \end{pmatrix}. \tag{34}$$

We want to find w, r such that $F(w,r) = 0$ so that

$$\mathbf{P}_\varepsilon = \textbf{diag}(e^{-\frac{w}{\varepsilon}})e^{-\frac{\mathbf{C}}{\varepsilon}}\textbf{diag}(e^{-\frac{r}{\varepsilon}}). \tag{35}$$

The Newton iteration is given by

$$\begin{pmatrix} w^{(k+1)} \\ r^{(k+1)} \end{pmatrix} = \begin{pmatrix} w^{(k)} \\ r^{(k)} \end{pmatrix} - J_F^{-1}(w^{(k)}, r^{(k)})F(w^{(k)}, r^{(k)}), \tag{36}$$

where

$$J_F = \frac{1}{\varepsilon}\begin{pmatrix} \textbf{diag}(\mathbf{P}\mathbf{1}_n) & \mathbf{P} \\ \mathbf{P}^\top & \text{diag}(\mathbf{P}^\top\mathbf{1}_m) \end{pmatrix}. \tag{37}$$

Sinkhorn-Newton method: Convergence

Proposition

For $w \in \mathbb{R}^m$ and $r \in \mathbb{R}^n$, the Jacobian matrix $J_F(w,r)$ is symmetric positive semidefinite, and its kernel is given by

$$\textbf{ker}(J_F(w,r)) = \textbf{span}\left\{ \begin{pmatrix} \mathbf{1}_m \\ -\mathbf{1}_n \end{pmatrix} \right\}. \tag{38}$$

Proof

J_F is clearly symmetric. For arbitrary $\gamma \in \mathbb{R}^m$ and $\phi \in \mathbb{R}^n$, one has

$$\begin{pmatrix} \gamma^\top & \phi^\top \end{pmatrix} J_F \begin{pmatrix} \gamma \\ \phi \end{pmatrix} = \frac{1}{\varepsilon} \sum_{ij} \mathbf{P}_{ij}(\gamma_i + \phi_j)^2 \geq 0,$$

which holds with equality if and only if $\gamma_i + \phi_j = 0$ for all i, j, leading us to (38).

Sinkhorn-Newton method: Convergence

Lemma

Let $F : D \to \mathbb{R}^n$ be a continuously differentiable mapping with $D \subset \mathbb{R}^n$ open and convex. Suppose that $F(x)$ is invertible for each $x \in D$. Assume that the following affine covariant Lipschitz condition holds

$$\|F'(x)^{-1}(F'(y) - F'(x))(y - x)\| \leq \omega\|y - x\|^2 \tag{39}$$

for $x, y \in D$. Let $F(x) = 0$ have a solution x^*. For the initial guess $x^{(0)}$ assume that $B(x^*, \|x^{(0)} - x^*\|) \subset D$ and that

$$\omega\|x^{(0)} - x^*\| < 2.$$

Then the ordinary Newton iterates remain in the open ball $B(x^*, \|x^{(0)} - x^*\|)$ and converge to x^* at an estimated quadratic rate

$$\|x^{(k+1)} - x^*\| \leq \frac{\omega}{2}\|x^{(k)} - x^*\|^2. \tag{40}$$

Moreover, the solution x^* is unique in the open ball $B(x^*, 2/\omega)$.

Sinkhorn-Newton method: Convergence

Proof

Denote $e^{(k)} = x^{(k)} - x^*$. Let us prove the lemma by induction:

$$\begin{aligned}
\|e^{(k+1)}\| &= \|x^{(k)} - (F'(x^{(k)}))^{-1}(F(x^{(k)} - F(x^*)) - x^*\| \\
&= \|e^{(k)} - (F'(x^{(k)}))^{-1}(F(x^{(k)} - F(x^*))\| \\
&= \|(F'(x^{(k)}))^{-1}((F(x^*) - F(x^{(k)})) + F'(x^{(k)})e^{(k)})\| \\
&= \|(F'(x^{(k)}))^{-1}\int_{s=0}^{-1}(F'(x^{(k)} + se^{(k)}) - F'(x^{(k)}))e^{(k)}\,\mathrm{d}s\| \\
&\leq \omega\|\int_{s=0}^{-1} s\,\mathrm{d}s\|e^{(k)}\|^2 = \frac{\omega}{2}\|e^{(k)}\|^2 < \|e^{(k)}\|.
\end{aligned} \tag{41}$$

Also

$$\omega\|e^{(k+1)}\| \leq \omega\|e^{(k)}\| < 2. \tag{42}$$

For the uniqueness part, let $x^{(0)} = x^{**} \neq x^*$ be a different solution,then $x^{(1)} = x^{**}$, then consider (40) when $k = 0$.

Sinkhorn-Newton method: Convergence

Proposition

For any $k \in \mathbb{N}$ with $P_{\varepsilon,ij}^{(k)} > 0$, the affine covariante Lipschitz condition holds in the ℓ_∞-norm for

$$\omega \le (e^{\frac{1}{\varepsilon}} - 1)\left(1 + 2e^{\frac{1}{\varepsilon}} \frac{\max\{\|\mathbf{P}_\varepsilon^{(k)} \mathbf{1}_n\|_\infty, \|(\mathbf{P}_\varepsilon^{(k)})^\top \mathbf{1}_m\|_\infty\}}{\min_{ij} \mathbf{P}_{\varepsilon,ij}^{(k)}}\right) \tag{43}$$

when $\|y - x\|_\infty \le 1$.

The proof for this proposition is tedious and therefore we refer the interested readers to the paper [**?**].

Relationship with Sinkhorn's algorithm

Let $u = e^{-\frac{w}{\varepsilon}}$, $v = e^{-\frac{r}{\varepsilon}}$ and $\mathbf{K} = e^{-\mathbf{C}/\varepsilon}$. We again state the KKT system of (21):

$$\begin{aligned} \mathbf{P}_\varepsilon &= \mathbf{diag}(u)\mathbf{K}\mathbf{diag}(v), \\ a &= \mathbf{diag}(u)\mathbf{K}v, \\ b &= \mathbf{diag}(v)\mathbf{K}^\top u. \end{aligned} \tag{44}$$

Then the Sinkhorn's algorithm amounts to alternating updates in the form of

$$\begin{aligned} u^{(k+1)} &= \mathbf{diag}(\mathbf{K}v^{(k)})^{-1}a, \\ v^{(k+1)} &= \mathbf{diag}(\mathbf{K}^\top u^{(k+1)})^{-1}b. \end{aligned} \tag{45}$$

Relationship with Sinkhorn's algorithm

Define

$$G(u,v) = \begin{pmatrix} \textbf{diag}(u)\mathbf{K}v - a \\ \textbf{diag}(v)\mathbf{K}^\top u - b \end{pmatrix}. \tag{46}$$

Process analogously to the Sinkhorn-Newton method we just discussed, note that

$$J_G(u,v) = \begin{pmatrix} \textbf{diag}(\mathbf{K}v) & \textbf{diag}(u)\mathbf{K} \\ \textbf{diag}(v)\mathbf{K}^\top & \textbf{diag}(\mathbf{K}^\top u) \end{pmatrix}. \tag{47}$$

If we neglect the off-diagonal blocks above, i.e.,

$$\hat{J}_G(u,v) = \begin{pmatrix} \textbf{diag}(\mathbf{K}v) & \mathbf{0} \\ \mathbf{0} & \textbf{diag}(\mathbf{K}^\top u) \end{pmatrix}, \tag{48}$$

and perform the Newton iteration

$$\begin{pmatrix} u^{(k+1)} \\ v^{(k+1)} \end{pmatrix} = \begin{pmatrix} u^{(k)} \\ v^{(k)} \end{pmatrix} - \hat{J}_G^{-1}(u^{(k)}, v^{(k)}) G(u^{(k)}, v^{(k)}), \tag{49}$$

Relationship with Sinkhorn's algorithm

We get

$$u^{(k+1)} = \textbf{diag}(\mathbf{K} v^{(k)})^{-1} a, \\ v^{(k+1)} = \textbf{diag}(\mathbf{K}^\top u^{(k)})^{-1} b. \tag{50}$$

So the Sinkhorn's algorithm simply approximates one Newton step by neglecting the off-diagonal blocks and replacing $u^{(k)}$ by $u^{(k+1)}$.

Outline

Wasserstein barycenter

- Define $C \overset{\text{def}}{=} M_{XY}$, where $(M_{XY})_{ij} = d(x_i, y_i)^p$. The Wasserstein distance as

$$\mathcal{L}(a, b, C) \overset{\text{def}}{=} \min \left\{ \sum_{i,j} C_{i,j} \Pi_{i,j}; \Pi \in \mathbf{U}(a, b) \right\}. \tag{51}$$

- Given a set of point clouds and their corresponding probability vector $\{(Y^i, b^i)\}$, $i = 1, \dots, N$.
- Find a support $X = \{x_i\}$ with a probability vector a such that (X, a) is the optimal solution of the following problem

$$\min_{X,a} \ \psi(a, X) = \sum_{k=1}^{N} \lambda_k \mathcal{L}(a, b^k, M_{XY^k}), \text{ s.t. } \sum_i a_i = 1, a \geq 0.$$

where $\sum_k \lambda_k = 1$ and $\lambda_k \geq 0$.

Differentiability of $\mathcal{L}(a, b, C)$ w.r.t. a

- The primal problem:

$$\mathcal{L}(a, b, C) \overset{\text{def}}{=} \min_{\Pi} \sum_{i,j} C_{i,j} \Pi_{i,j} \quad \text{s.t.} \quad \Pi \mathbf{1}_n = a, \Pi^\top \mathbf{1}_m = b, \Pi \geq 0.$$

- Let u^* is the optimal dual vector of the dual problem:

$$\max_{u \in \mathbb{R}^m, v \in \mathbb{R}^n} \quad u^\top a + v^\top b, \text{ s.t.} \quad u_i + v_j \leq C_{ij}, \quad \forall (i, j)$$

- Suppose $\mathcal{L}(a, b, C)$ is finite, the strong duality holds. Then u^* is a subgradient of $\mathcal{L}(a, b, C)$ w.r.t. a.

Subgradient of optimal value function

define $h(u, v)$ as the optimal value of convex problem

$$\begin{array}{rl} \min & f_0(x) \\ \text{s.t.} & f_i(x) \leq u_i,\ i = 1, \cdots, m \\ & Ax = b + v \end{array}$$

(functions f_i are convex; optimization variable is x)

weak result: suppose $h(\hat{u}, \hat{v})$ is finite, strong duality holds with the dual

$$\begin{array}{rl} \max & \inf_x \left(f_0(x) + \sum_i \lambda_i (f_i(x) - \hat{u}_i) + \nu^\top (Ax - b - \hat{v}) \right) \\ \text{s.t.} & \lambda \geq 0 \end{array}$$

if $\hat{\lambda}, \hat{\nu}$ are optimal dual variables (for r.h.s. $\hat{u}, \hat{v}$) then $(\hat{\lambda}, \hat{\nu}) \in \partial h(\hat{u}, \hat{v})$

proof : by weak duality for problem with r.h.s. u, v

$$
\begin{aligned}
h(u, v) &\geq \inf_x \left(f_0(x) + \sum_i \hat{\lambda}_i (f_i(x - u_i) + \hat{\nu}^\top (Ax - b - v) \right) \\
&= \inf_x \left(f_0(x) + \sum_i \hat{\lambda}_i (f_i(x - \hat{u}_i) + \hat{\nu}^\top (Ax - b - \hat{v}) \right) \\
&- \hat{\lambda}^\top (u - \hat{u}) - \hat{\nu}^\top (v - \hat{v}) \\
&= h(\hat{u}, \hat{v}) - \hat{\lambda}^\top (u - \hat{u}) - \hat{\nu}^\top (v - \hat{v})
\end{aligned}
$$

minimizing $\psi(a, X)$ w.r.t a

For a fixed X, consider the problem

$$\min_{a} \quad \psi(a, X) = \sum_{k=1}^{N} \lambda_k \mathcal{L}(a, b^k, M_{XY^k}), \text{ s.t. } \sum a_i = 1, a \geq 0$$

- Let u^k be the optimal dual variable of $\mathcal{L}(a, b^k, M_{XY^k})$ w.r.t. a. Then

$$g = \sum_{k=1}^{N} \lambda_k u^k \in \partial_a \psi(a, X)$$

- Let $h(a) = \sum_{i=1}^{m} a_i \log a_i$. The associated Bregman divergence is

$$D_h(y, x) = h(y) - h(x) - \nabla h(x)^T (y - x)$$

- The mirror descent method is

$$a^{j+1} = \underset{\sum a_i = 1, a \geq 0}{\operatorname{argmin}} \left\{ g^T (a - a^j) + t D_h(a, a^j) \right\}$$

Minimizing $\psi(a, X)$ w.r.t. X

Denote $X = [x_1, \ldots, x_m]$ and $Y = [y_1, \ldots, y_n]$.

- Consider $(M_{XY})_{ij} = \|x_i - y_i\|_2^2$. Let $\mathrm{x} = \mathrm{diag}(X^\top X)$ and $\mathrm{y} = \mathrm{diag}(Y^\top Y)$. Then we have:

$$M_{XY} = \mathrm{x}1_n^\top + 1_m^\top \mathrm{y} - 2X^\top Y \in \mathbb{R}^{m \times n}$$

- Let Π be the optimal matrix corresponding to a

$$\begin{aligned}
\mathcal{L}(a, b, M_{XY}) &= \langle \Pi, M_{XY} \rangle \\
&= \left\langle \Pi, \mathrm{x}1_n^\top + 1_m^\top \mathrm{y} - 2X^\top Y \right\rangle \\
&= \langle \mathrm{x}, \Pi 1_n \rangle + \left\langle \mathrm{y}, \Pi^\top 1_m \right\rangle - 2 \left\langle \Pi, X^\top Y \right\rangle \\
&= \mathrm{x}^\top a + \mathrm{y}^\top b - 2 \left\langle \Pi, X^\top Y \right\rangle \\
&= \|X \mathrm{diag}(a^{1/2}) - Y \Pi^\top \mathrm{diag}(a^{-1/2})\|_F^2 + \mathrm{const}.
\end{aligned}$$

Minimizing $\psi(a, X)$ w.r.t. X

For a fixed a, consider the problem

$$\min_X \quad \psi(a, X) = \sum_{k=1}^{N} \lambda_k \mathcal{L}(a, b^k, M_{XY^k}).$$

Then, it is equivalent to

$$\min_X \quad \sum_{k=1}^{N} \lambda_k \left(\mathrm{x}^\top a - 2 \left\langle \Pi^k, X^\top Y^k \right\rangle \right)$$

$$\min_X \quad \mathrm{x}^\top a - 2 \left\langle \sum_{k=1}^{N} \lambda_k \Pi^k, X^\top Y^k \right\rangle$$

$$\min_X \quad \|X \mathrm{diag}(a^{1/2}) - \sum_{k=1}^{N} \lambda_k Y^k (\Pi^k)^\top \mathrm{diag}(a^{-1/2})\|_F^2$$

The optimal solution is:

$$X = \sum_{k=1}^{N} \lambda_k Y^k (\Pi^k)^\top \mathrm{diag}(a^{-1})$$