Optimal Transport

http://bicmr.pku.edu.cn/~wenzw/bigdata2022.html

Acknowledgement: this slides is based on Prof. Gabriel Peyré’s lecture notes
Applications: comparing measures

Comparing Measures

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

- Optimal transport
  → takes into account a metric $d$.
Applications: toward high-dimensional OT

Toward High-dimensional OT

Monge  Kantorovich  Dantzig  Brenier  Otto  McCann  Villani
1. Problem Formulation
2. Applications
3. Entropic Regularization
4. Sinkhorn’s Algorithm
5. Sinkhorn-Newton method
6. Wasserstein barycenter
7. A Multiscale Semismooth Newton Method
A Geometric Motivation

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

distance between $\mu$ and $\nu$:

- $\mu = \delta_{x_1}$ and $\nu = \delta_{y_1}$
  - $\text{dist}(\mu, \nu) = \text{dist}(x_1, y_1)$

- $\mu = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}$ and $\nu = \frac{1}{n} \sum_{i=1}^{n} \delta_{y_i}$
  - $\text{dist}(\mu, \nu) = \frac{1}{n^2} \sum_{ij} \text{dist}(x_i, y_j)$? or
  - $\text{dist}(\mu, \nu) = \min_{\sigma \text{ permutation}} \frac{1}{n} \sum_{i} \text{dist}(x_i, y_{\sigma(i)})$

- What if $\mu, \nu \in \mathcal{P}(\mathcal{X})$?

Goal: Build a metric on $\mathcal{P}(\mathcal{X})$ with the geometry of $(\mathcal{X}, \text{dist})$. 
Kantorovitch’s Formulation

Discrete Optimal Transport
Input two discrete probability measures

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

(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

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

(2)

is called the set of couplings with respect to \( \alpha \) and \( \beta \).
Kantorovitch’s Formulation

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

Optimal transport distance

\[
\mathcal{L}(\alpha, \beta, C) \overset{\text{def}}{=} \min \left\{ \sum_{i,j} C_{i,j} \Pi_{i,j}; \Pi \in \mathbf{U}(a, b) \right\}.
\]
Push Forward

- Radon measures \((\alpha, \beta)\) 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,
  \[
  \int_{\mathcal{Y}} g(y) dT_#\alpha(y) \overset{\text{def}}{=} \int_{\mathcal{X}} g(T(x)) d\alpha(x). \tag{5}
  \]
- Discrete measures: \(T_#\alpha = \sum_i \alpha_i \delta_{T(x_i)}\)
- Smooth densities: \(d\alpha = \rho(x) dx, \quad d\beta = \xi(x) dx\).
  \[
  T_#\alpha = \beta \iff \rho(T(x)) |\text{det}(\partial T(x))| = \xi(x). \tag{6}
  \]
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 $\alpha$ toward the mass of $\beta$, namely:

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

- Discrete case:

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

- Arbitrary measures:

$$\min_T \int_X c(x, T(x)) d\alpha(x), \quad \text{s.t.} \quad T \# \alpha = \beta$$
Couplings between General Measures

Projectors:

\[ P_X : (x, y) \in X \times Y \rightarrow x \in X, \]
\[ P_Y : (x, y) \in X \times Y \rightarrow y \in Y. \]  \hspace{1cm} (7)

\[ U(\alpha, \beta) \overset{\text{def}}{=} \{ \pi \in M_+(X \times Y); P_X \# \pi = \alpha, P_Y \# \pi = \beta \}. \]  \hspace{1cm} (8)

is called the set of couplings with respect to \( \alpha \) and \( \beta \).
Cases of Couplings

Couplings: the 3 Settings

Discrete

Semi-discrete

Continuous
More Examples

Examples of Couplings
Kantorovich Problem for General Measures

Optimal transport distance between General Measures

\[ \mathcal{L}(\alpha, \beta, c) \overset{\text{def}}{=} \min_{\pi \in \mathcal{U}(\alpha, \beta)} \int_{X \times Y} c(x, y) d\pi(x, y). \]  \hspace{1cm} (9)

Probability interpretation:

\[ \min_{(X,Y)} \{ \mathbb{E}_{(X,Y)}(c(X, Y)), X \sim \alpha, Y \sim \beta \}. \]  \hspace{1cm} (10)
Wasserstein Distance

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

Wasserstein Distance

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

Theorem

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

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

Example

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

Dual problem (discrete case)

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

Relation between any primal and dual solutions:

\[ P_{ij} > 0 \Rightarrow u_i + v_j = C_{ij}. \]
Wasserstein barycenter

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

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

(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 $\lambda_k \geq 0$. 
提纲

1. Problem Formulation
2. Applications
3. Entropic Regularization
4. Sinkhorn’s Algorithm
5. Sinkhorn-Newton method
6. Wasserstein barycenter
7. A Multiscale Semismooth Newton Method
Applications: image color adaptation

Example: https://github.com/rflamary/POT/blob/master/notebooks/plot_otda_color_images.ipynb

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

Image Color Palette Equalization

Optimal transport

[Images of color palettes and transport diagrams]
Applications: shape interpolation

Shape Interpolation
Applications: MRI Data Processing

MRI Data Processing [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 $T_{ijc}(i, j)$, transportation problem

Bag of Words

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

[Kusner’15]
Applications: word mover’s distance

\[
\begin{align*}
\min_{\Pi \geq 0} & \quad \sum_{ij} \Pi_{ij} c_{ij} \\
s.t. & \quad \sum_{j=1}^{n} \Pi_{ij} = d_i \\
& \quad \sum_{i=1}^{n} \Pi_{ij} = d'_j
\end{align*}
\]

- \(x_i\): 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, \(\ell\) 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 \(\delta > 0\) parameterizes the size of the distributional uncertainty.

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

1. Problem Formulation
2. Applications
3. Entropic Regularization
4. Sinkhorn’s Algorithm
5. Sinkhorn-Newton method
6. Wasserstein barycenter
7. A Multiscale Semismooth Newton Method
Given an integer $n \geq 1$, we write $\Sigma_n$ for the discrete probability simplex

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

(16)

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

$$
L(a, b, C) \overset{\text{def}}{=} \min \left\{ \sum_{i,j} C_{i,j} P_{i,j} ; \text{s.t. } P \in U(a, b) \right\}.
$$

(17)

Where $U(a, b)$ is the set of couplings between $a$ and $b$. 
Entropy

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

\[
H(P) \overset{\text{def}}{=} - \sum_{i,j} P_{i,j} (\log(P_{i,j}) - 1).
\]  

(18)

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

\[
H(u) \overset{\text{def}}{=} - \sum_i u_i (\log(u_i) - 1).
\]

(19)

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

\[
\text{KL}(u\|v) = - \sum_{i=1}^n u_i \log\left(\frac{v_i}{u_i}\right).
\]

(20)

The KL divergence is always non-negative: \( \text{KL}(u\|v) \geq 0 \) (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 $C \in \mathbb{R}_+^{m \times n}$. The entropic regularization of the transportation problem reads

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

- The case $\varepsilon = 0$ corresponds to the classic (linear) optimal transport problem.
- For $\varepsilon > 0$, problem (21) has an $\varepsilon$-strongly convex objective and therefore admits a unique optimal solution $P^\star_\varepsilon$.
- This is not (necessarily) true for $\varepsilon = 0$. But we have the following proposition.
Proposition

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

\[
P_{\varepsilon} \xrightarrow{\varepsilon \to 0} \text{argmax}_P \{ H(P); P \in U(a, b), \langle P, C \rangle = L^0(a, b, C) \} \tag{22}
\]

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

**Proof**

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

\[
0 \leq \langle C, P_\ell \rangle - \langle C, P \rangle \leq \varepsilon_\ell (H(P_\ell) - H(P)).
\]  

(23)

Since \(H\) is continuous, taking the limit \(\ell \to +\infty\) in this expression shows that \(\langle C, P^* \rangle = \langle C, P \rangle\). Furthermore, dividing by \(\varepsilon_\ell\) and taking the limit shows that \(H(P) \leq H(P^*)\). 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

$$U_\alpha(a, b) \overset{\text{def}}{=} \{ P \in U(a, b) | KL(P \parallel ab^\top) \leq \alpha \}$$

$$= \{ P \in U(a, b) | H(P) \geq H(a) + H(b) - 1 - \alpha \}.$$  \hspace{1cm} (24)

**Definition: Sinkhorn Distance**

$$d_{C,\alpha}(a, b) \overset{\text{def}}{=} \min_{P \in U_\alpha(a, b)} \langle C, P \rangle.$$  \hspace{1cm} (25)

**Proposition**

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

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

Proof.

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

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

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

$$U_\alpha(a, b) = U(a, b).$$
1 Problem Formulation
2 Applications
3 Entropic Regularization
4 Sinkhorn’s Algorithm
5 Sinkhorn-Newton method
6 Wasserstein barycenter
7 A Multiscale Semismooth Newton Method
Sinkhorn’s algorithm

For solving (21), consider its Lagrangian dual function

\[ \mathcal{L}_C^\varepsilon(P, \alpha, \beta) = \langle C, P \rangle - \varepsilon H(P) + \alpha^\top (P1_n - a) + \beta^\top (P^\top 1_m - b). \]  

(27)

Now let \( \partial \mathcal{L}_C^\varepsilon / \partial p_{ij} = 0 \), i.e.,

\[ p_{ij} = e^{-\frac{c_{ij} + \alpha_i + \beta_j}{\varepsilon}}, \] 

(28)

so we can write

\[ P_\varepsilon = \text{diag}(e^{-\frac{\alpha}{\varepsilon}}) e^{-\frac{C}{\varepsilon}} \text{diag}(e^{-\frac{\beta}{\varepsilon}}). \] 

(29)

Note that

\[ P_\varepsilon 1_n = a, \quad P_\varepsilon^\top 1_m = b, \] 

(30)

we can then use Sinkhorn’s algorithm to find \( P_\varepsilon \)!
Sinkhorn’s algorithm

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

\[
P_{\varepsilon} = \text{diag}(u)K\text{diag}(v),
\]
\[
a = \text{diag}(u)Kv,
\]
\[
b = \text{diag}(v)K^\top u. \tag{31}
\]

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

\[
u^{(k+1)} = \text{diag}(Kv^{(k)})^{-1}a,
\]
\[
u^{(k+1)} = \text{diag}(K^\top u^{(k+1)})^{-1}b. \tag{32}
\]
Sinkhorn’s algorithm

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

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

   until reaches certain stopping criterion.
5. Compute

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

and eventually

$$P_\epsilon = \text{diag}(u)K\text{diag}(v).$$
Problem Formulation

Applications

Entropic Regularization

Sinkhorn’s Algorithm

Sinkhorn-Newton method

Wasserstein barycenter

A Multiscale Semismooth Newton Method
The dual problem of (21) is

$$\min_{\alpha, \beta} \quad \langle a, \alpha \rangle + \langle b, \beta \rangle + \varepsilon \langle e^{-\frac{\alpha}{\varepsilon}}, K e^{-\frac{\beta}{\varepsilon}} \rangle,$$

s.t. \quad \text{diag}(e^{-\frac{\alpha}{\varepsilon}})K e^{-\frac{\beta}{\varepsilon}} = a,
\quad \text{diag}(e^{-\frac{\beta}{\varepsilon}})K^{\top} e^{-\frac{\alpha}{\varepsilon}} = b. \quad \text{(33)}$$

with $\alpha, \beta$ being the dual variables. 
[1] proposes using Newton method to solve this system.
Sinkhorn-Newton method

Let

\[ F(\alpha, \beta) = \begin{pmatrix} \text{diag}(e^{-\frac{\alpha}{\varepsilon}})Ke^{-\frac{\beta}{\varepsilon}} - a \\ \text{diag}(e^{-\frac{\beta}{\varepsilon}})K^\top e^{-\frac{\alpha}{\varepsilon}} - b \end{pmatrix}. \] (34)

We want to find \( \alpha, \beta \) such that \( F(\alpha, \beta) = 0 \) so that

\[ P_\varepsilon = \text{diag}(e^{-\frac{\alpha}{\varepsilon}})e^{-\frac{c}{\varepsilon}} \text{diag}(e^{-\frac{\beta}{\varepsilon}}). \] (35)

The Newton iteration is given by

\[ \begin{pmatrix} \alpha^{(k+1)} \\ \beta^{(k+1)} \end{pmatrix} = \begin{pmatrix} \alpha^{(k)} \\ \beta^{(k)} \end{pmatrix} - J_F^{-1}(\alpha^{(k)}, \beta^{(k)})F(\alpha^{(k)}, \beta^{(k)}), \] (36)

where

\[ J_F = \frac{1}{\varepsilon} \begin{pmatrix} \text{diag}(P1_n) & P \\ P^\top & \text{diag}(P^\top 1_m) \end{pmatrix}. \] (37)
Sinkhorn-Newton method: Convergence

Proposition

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

$$\ker(J_F(\alpha, \beta)) = \text{span}\left\{ \begin{pmatrix} 1_m \\ -1_n \end{pmatrix} \right\}.$$  \hfill (38)

Proof

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

$$(\gamma^T \quad \phi^T) J_F \begin{pmatrix} \gamma \\ \phi \end{pmatrix} = \frac{1}{\varepsilon} \sum_{ij} P_{ij} (\gamma_i + \phi_j)^2 \geq 0,$$

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

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

(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

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

(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:

\[ \|e^{(k+1)}\| = \|x^{(k)} - (F'(x^{(k)}))^{-1}(F(x^{(k)}) - F(x^*)) - x^*\| \]
\[ = \|e^{(k)} - (F'(x^{(k)}))^{-1}(F(x^{(k)}) - F(x^*))\| \]
\[ = \| (F'(x^{(k)}))^{-1}((F(x^*) - F(x^{(k)})) + F'(x^{(k)})e^{(k)})\| \]
\[ = \| (F'(x^{(k)}))^{-1} \int_{s=0}^{-1} (F'(x^{(k)}) + se^{(k)}) - F'(x^{(k)}))e^{(k)} \, ds\| \]
\[ \leq \omega \int_{s=0}^{-1} s \, ds \|e^{(k)}\|^2 = \frac{\omega}{2} \|e^{(k)}\|^2 < \|e^{(k)}\|. \] (41)

Also

\[ \omega \|e^{(k+1)}\| \leq \omega \|e^{(k)}\| < 2. \] (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 \( \ell_\infty \)-norm for

\[
\omega \leq (e^{\frac{1}{\varepsilon}} - 1) \left( 1 + 2e^{\frac{1}{\varepsilon}} \frac{\max\{\|P_{\varepsilon}^{(k)} 1_n\|_\infty, \|(P_{\varepsilon}^{(k)})^\top 1_m\|_\infty\}}{\min_{ij} P_{\varepsilon,ij}} \right)
\]

when \( \|y - x\|_\infty \leq 1 \).

The proof for this proposition is tedious and therefore we refer the interested readers to the paper [1].
Relationship with Sinkhorn’s algorithm

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

\[
P_\varepsilon = \text{diag}(u) K \text{diag}(v),
\]
\[
a = \text{diag}(u) K v,
\]
\[
b = \text{diag}(v) K^\top u.
\]

(44)

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

\[
u^{(k+1)} = \text{diag}(Kv^{(k)})^{-1} a,
\]
\[
u^{(k+1)} = \text{diag}(K^\top u^{(k+1)})^{-1} b.
\]

(45)
Relationship with Sinkhorn’s algorithm

Define

\[ G(u, v) = \begin{pmatrix} \text{diag}(u)Kv - a \\ \text{diag}(v)K^T u - b \end{pmatrix}. \] (46)

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

\[ J_G(u, v) = \begin{pmatrix} \text{diag}(Kv) & \text{diag}(u)K \\ \text{diag}(v)K^T & \text{diag}(K^T u) \end{pmatrix}. \] (47)

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

\[ \hat{J}_G(u, v) = \begin{pmatrix} \text{diag}(Kv) & 0 \\ 0 & \text{diag}(K^T u) \end{pmatrix}, \] (48)

and perform the Newton iteration

\[ \begin{pmatrix} u^{(k+1)} \\ v^{(k+1)} \end{pmatrix} = \begin{pmatrix} u^{(k)} \\ v^{(k)} \end{pmatrix} - \hat{J}_G^{-1}(u^{(k)}, v^{(k)})G(u^{(k)}, v^{(k)}), \] (49)
Relationship with Sinkhorn’s algorithm

We get

\[
\begin{align*}
u^{(k+1)} &= \text{diag}(Kv^{(k)})^{-1}a, \\
v^{(k+1)} &= \text{diag}(K^\top u^{(k)})^{-1}b. 
\end{align*}
\]

(50)

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

1. Problem Formulation
2. Applications
3. Entropic Regularization
4. Sinkhorn’s Algorithm
5. Sinkhorn-Newton method
6. Wasserstein barycenter
7. A Multiscale Semismooth Newton Method
Wasserstein barycenter

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

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

(51)

- 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

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

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:

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

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

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

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

\[
\begin{align*}
\min & \quad f_0(x) \\
\text{s.t.} & \quad f_i(x) \leq u_i, \ i = 1, \ldots, m \\
& \quad Ax = b + v
\end{align*}
\]

(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

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

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. \( u, v \)

\[
h(u, v) \geq \inf_x \left( f_0(x) + \sum_i \hat{\lambda}_i (f_i(x - u_i) + \hat{\nu}^\top (Ax - b - v)) \right)
\]

\[
= \inf_x \left( f_0(x) + \sum_i \hat{\lambda}_i (f_i(x - \hat{u}_i) + \hat{\nu}^\top (Ax - b - \hat{v})) \right)
\]

\[
- \hat{\lambda}^\top (u - \hat{u}) - \hat{\nu}^\top (v - \hat{v})
\]

\[
= h(\hat{u}, \hat{v}) - \hat{\lambda}^\top (u - \hat{u}) - \hat{\nu}^\top (v - \hat{v})
\]
minimizing $\psi(a, X)$ w.r.t. $a$

For a fixed $X$, consider the problem

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

- Let $u^k$ be the optimal dual variable of $\mathcal{L}(a, b^k, M_{XYk})$ 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

$$a^{i+1} = \arg\min_{\sum a_i = 1, a \geq 0} \left\{ g^T (a - a^i) + \frac{1}{\alpha} D_h(a, a^i) \right\}$$
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 $x = \text{diag}(X^\top X)$ and $y = \text{diag}(Y^\top Y)$. Then we have:

\[
M_{XY} = x1_n^\top + 1_m^\top y - 2X^\top Y \in \mathbb{R}^{m \times n}
\]

- Let $\Pi$ be the optimal matrix corresponding to $a$

\[
\begin{align*}
\mathcal{L}(a, b, M_{XY}) &= \langle \Pi, M_{XY} \rangle \\
&= \langle \Pi, x1_n^\top + 1_m^\top y - 2X^\top Y \rangle \\
&= \langle x, \Pi 1_n \rangle + \langle y, \Pi^\top 1_m \rangle - 2 \langle \Pi, X^\top Y \rangle \\
&= x^\top a + y^\top b - 2 \langle \Pi, X^\top Y \rangle \\
&= \|X\text{diag}(a^{1/2}) - Y\Pi^\top \text{diag}(a^{-1/2})\|_F^2 + \text{const.}
\end{align*}
\]
Minimizing $\psi(a, X)$ w.r.t. $X$

For a fixed $a$, consider the problem

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

Then, it is equivalent to

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

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

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

The optimal solution is:

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

\[
\min_{X \in \mathbb{R}^{m \times n}} \langle C, X \rangle,
\]

subject to

\[
su\text{bject to } \sum_{j=1}^{n} X_{ij} = u_i, \quad 1 \leq i \leq m,
\]

\[
\sum_{i=1}^{m} X_{ij} = v_j, \quad 1 \leq j \leq n,
\]

\[
X_{ij} \geq 0, \quad 1 \leq i \leq m, 1 \leq j \leq n,
\]

where \( C \in \mathbb{R}_{+}^{m \times n} \) is the given cost matrix, and \( u \in \mathbb{R}_{+}^{m}, v \in \mathbb{R}_{+}^{n} \) satisfy \( \sum_{i=1}^{m} u_i = \sum_{j=1}^{n} v_j = 1 \).
The problem in (52) can be rewritten as an LP:

\[
\begin{align*}
\min_{x \in \mathbb{R}^{mn}} & \quad c^T x, \\
\text{subject to} & \quad Ax = b, \\
& \quad x \geq 0,
\end{align*}
\]

where

\[
A = \begin{bmatrix}
1^T_n \\
1^T_n \\
\vdots \\
I_n & I_n & \cdots & I_n
\end{bmatrix} \in \mathbb{R}^{(m+n) \times mn}, \quad b = \begin{bmatrix} u \\ v \end{bmatrix} \in \mathbb{R}^{m+n},
\]

\(c \in \mathbb{R}^{mn}\) is the vectorization of \(C\) in (52): \(c_{(i-1)n+j} = C_{ij}\), and \(1_n\) denotes the \(n \times 1\) all-one vector.
Semismooth Newton Method

Let

\[ h(x) = \delta_{\{Ax=b\}}, \quad g(x) = c^T x + \delta_{\{x \geq 0\}}, \]  

where \( \delta_S \) is the indicator function of the set \( S \), i.e.,

\[ \delta_S(x) = \begin{cases} 
0, & x \in S, \\
\infty, & x \notin S. 
\end{cases} \]

Problem in (53) becomes

\[ \min_{x \in \mathbb{R}^{mn}} h(x) + g(x). \]  

Note that both \( h \) and \( g \) are convex.
Semismooth Newton Method

The DRS method gives us the following iterative process

\[
\begin{align*}
    x^{(k)} &= \text{prox}_{tg}(z) = (z^{(k-1)} - tc)_+, \\
    y^{(k)} &= \text{prox}_{th}(2x^{(k)} - z^{(k-1)}) \\
    &= P_{\{Ax=b\}}(2x^{(k)} - z^{(k-1)}), \\
    z^{(k)} &= z^{(k-1)} + y^{(k)} - x^{(k)}.
\end{align*}
\] (58)

\(P_S\) is the projector of set \(S\).

**Lemma**

If \(b \in \mathcal{R}(A)\), then

\[
P_{\{Ax=b\}}(x) = x - A^\dagger(Ax - b).
\] (59)
Substituting $z - tc$ by $z$, we have

\[
\begin{align*}
    x^{(k)} &= (z^{(k-1)})_+ , \\
    y^{(k)} &= \mathcal{P}_{\{Ax=b\}}(2x^{(k)} - z^{(k-1)} - tc) , \\
    z^{(k)} &= z^{(k-1)} + y^{(k)} - x^{(k)} .
\end{align*}
\]

(60)

We could write the above iterative process as a fixed-point iteration

\[
    z^{(k)} = \mathcal{T}_{\text{DRS}}(z^{(k-1)}) = z^{(k-1)} - F(z^{(k-1)}) ,
\]

(61)

where

\[
    F(z) = A^\dagger A(2z_+ - z - tc) - (z_+ - z - tc) - A^\dagger b .
\]

(62)
Semismooth Newton Method

\[ F(z) = A^\dagger A(2z_+ - z - tc) - (z_+ - z - tc) - A^\dagger b. \]

We want to find \( z^* \) such that \( F(z^*) = 0 \). The generalized Jacobian of \( F \) at \( z \) is

\[ J(z) = A^\dagger A(2D - I) + (I - D) \in \partial_C F(z), \tag{63} \]

where \( D \overset{\text{def}}{=} D(z) \in \mathbb{R}^{mn \times mn} \) is a diagonal matrix with

\[
D_{ii} = \begin{cases} 
1, & z_i > 0, \\
0, & z_i \leq 0, 
\end{cases} \quad 1 \leq i \leq mn. \tag{64}
\]

**Lemma**

\( J \) is positive-semidefinite.

**Proposition**

For any \( \mu > 0 \), the condition number \( \kappa(J + \mu I) \leq \sqrt{\frac{1+\mu}{\mu}} \).
At the $k$-th iteration, we want to find the update step $d^{(k)}$ such that

\[(J^{(k)} + \mu^{(k)} I) d^{(k)} = -F(z^{(k)}) \overset{\text{def}}{=} -F^{(k)}, \quad u^{(k)} = z^{(k)} + d^{(k)}.\] (65)

- $\mu^{(k)} > 0$ is the regularizer, $J^{(k)} + \mu^{(k)} I$ is positive definite.
- $\mu^{(k)}$ should scales along with $\|F^{(k)}\|_2$. 
Define the ratio

\[ \rho^{(k)} = \frac{-\langle F(u^{(k)}), d^{(k)} \rangle}{\|d^{(k)}\|_2^2}. \]  

(66)

If \( \rho^{(k)} \) is large enough, it is usually a signal of a good Newton step, therefore we decrease \( \mu^{(k)} \) and do the update

\[ z^{(k+1)} = u^{(k)}. \]  

(67)

Otherwise, we increase \( \mu^{(k)} \) and let \( z^{(k+1)} = z^{(k)} \).
Algorithm 1: Semismooth Newton method

1. Input: $\mu^{(1)}$, $\eta > 0$.
2. while not converged do
   3. Compute $d^{(k)}$ by solving the Newton system in (65) inexactley.
   4. $u^{(k)} = z^{(k)} + d^{(k)}$.
   5. Compute the ratio $\rho^{(k)}$ as in (66).
   6. If $\rho^{(k)} \geq \eta$, then update $z^{(k+1)} = u^{(k)}$. Otherwise $z^{(k+1)} = z^{(k)}$.
   7. Update $\mu^{(k)}$.
   8. Update $\eta$ according to the value of $\|F(z^{(k+1)})\|_2$. 
At $z^{(k)}$, from (63) we have

$$J + \mu I = (\mu + 1)I - D - A^\dagger A(I - 2D)$$

$$= H - A^\dagger A(I - 2D),$$

(68)

where $H = (\mu + 1)I - D$.

If $A$ has linearly independent rows (discussed later), then

$$A^\dagger = A^\top (AA^\top)^{-1},$$

$$J + \mu I = H - A^\top (AA^\top)^{-1}A(I - 2D).$$

(69)
By Sherman-Morrison-Woodbury formula, we could verify

\[(J + \mu I)^{-1} = H^{-1} + H^{-1} A^\top (AKA^\top)^{-1} A(I - 2D)H^{-1}.\]  

(70)

- \(K = (I - (I - 2D)H^{-1})\).
- The diagonal element of \(K\) is either \(v_1 = \frac{\mu}{\mu + 1} > 0\) or \(v_2 = \frac{\mu + 1}{\mu} > 0\).
- \(H, D\) and \(K\) are all diagonal mappings.
- The only obstacle is solving the positive-semidefinite linear system \(AKA^\top\). This could be done by the method of Conjugate Gradient.
Sparsity of OT

From the theory of general LP we know that the support of the optimal solution of (53) contains at most $m + n$ elements though the overall number of variables is $mn$, i.e., the solution is supposed to be very sparse. This simple fact motivates us to restrict the set of variables to a small subset $\mathcal{N} \subset S \times D$ and fix other variables as zeros. Here

$$S = \{s_i, 1 \leq i \leq m\},$$
$$D = \{d_j, 1 \leq j \leq n\},$$

are the sets of origins and destinations, respectively.
Sparsity of OT

Given $\mathcal{N} = \{(s_{ik}, d_{jk}), 1 \leq k \leq |\mathcal{N}|\}$, we solve the following LP

$$\min_{x \in \mathbb{R}^{|\mathcal{N}|}} c^T_N x, \tag{71}$$

subject to

$$A_N x = b,$$

$$x \geq 0.$$

- $c_N \in \mathbb{R}^{|\mathcal{N}|}$ is the cost function corresponds to the subset $\mathcal{N}$, i.e., the $k$-th element of $c_N$ is

$$c_{(i_k-1)n+j_k} = C_{i_kj_k}.$$

- $A_N = (a_{ik}) \in \mathbb{R}^{(m+n) \times |\mathcal{N}|}$ and

$$a_{ik} = \begin{cases} 1, & 1 \leq i \leq m, i = i_k, \\ 0, & 1 \leq i \leq m, i \neq i_k, \\ 1, & m < i \leq m + n, i - m = j_k, \\ 0, & m < i \leq m + n, i - m \neq j_k, \end{cases} \quad 1 \leq k \leq |\mathcal{N}|. \tag{72}$$
Remove redundant rows from $A_N$

- Each column of $A_N$ has exactly two ones (each variable is involved in exactly two constraints).
- The first $m$ rows of $A_N$ are linearly independent.
- We sequentially label the $i$-th row ($m + 1 \leq i \leq m + n$) as either redundant or irredundant.
- Define $\mathcal{J} = \{ j | a_{ij} = 1 \}$. In order for the $i$-th row to be linearly expressed by the previous rows of $A_N$, for each $j \in \mathcal{J}$ there is exactly one $i_j$ such that
  \begin{equation}
  i_j \leq m < i, a_{ij} = 1, \tag{73}
  \end{equation}
  and the $i_j$-th row of $A_N$ must be included the linear expression.
- Now update $\mathcal{J}$ by
  \begin{equation}
  \mathcal{J} \leftarrow (\bigcup_{j \in \mathcal{J}} \{ t | a_{ij} = 1 \}) \setminus \mathcal{J}. \tag{74}
  \end{equation}
  If $\mathcal{J}$ is empty, the $i$-th row is redundant. Otherwise keep updating $\mathcal{J}$ like above.
Sparsity of OT

Assume the size of $A_N$ after reduction is $(m + r(n)) \times |N|$. 

$$(J + \mu I)^{-1} = H^{-1} + H^{-1} A_N^\top (A_N K A_N^\top)^{-1} A_N (I - 2D) H^{-1}. \tag{75}$$

where $K = (I - (I - 2D) H^{-1})$. We can compute 

$$A_N^\top K A_N = \begin{bmatrix} \text{diag}(r_{\hat{K}}) & \hat{K} \\ \hat{K}^\top & \text{diag}(c_{\hat{K}}) \end{bmatrix}, \tag{76}$$

where $r_{\hat{K}} \in \mathbb{R}^m$ is the row sums of $\hat{K}$ and $c_{\hat{K}} \in \mathbb{R}^n$ is the column sums of $\hat{K}$, and $\hat{K} \in \mathbb{R}^{m \times r(n)}$ is defined as follows. First we set $\hat{K} \in \mathbb{R}^{m \times n}$ by 

$$\hat{k}_{ikj} = \begin{cases} \nu_1, z_k \leq 0, \\ \nu_2, z_k > 0, \end{cases} \quad 1 \leq k \leq |N|, \tag{77}$$

while filling all the other positions with 0. Here $(s_{ik}, d_{jk})$ is the $k$-th element of $N$ and $\nu_1 = \mu / (\mu + 1)$, $\nu_2 = \nu_1^{-1}$. Then reduce $\hat{K}$ from $n$ columns to $r(n)$ columns according to the result of Algorithm ??.
How to choose $\mathcal{N}$?

**Proposition [2]**

Given $\mathcal{N}$, suppose $X$ is an optimal solution of the local OT problem restricted to $\mathcal{N}$ ($\text{spt } X \subset \mathcal{N}$). If for every $(s_i, d_j) \in (S \times D) \setminus \mathcal{N}$, there exists some $(s_{i_0}, d_{j_0}) \in \text{spt } X$ with $(s_i, d_j) \in \mathcal{N}$ such that

$$c_{ij} + c_{i_0j_0} > c_{ij_0} + c_{i_0j},$$

(78)

then $X$ is the global optimizer of problem (52).

**Proposition [2]**

Given an initial feasible subset $\mathcal{N}^{(1)}$. If at the $i$-th iteration we obtain the local optimizer $x^{(i)}$, and we construct the new subset $\mathcal{N}^{(i+1)}$ such that it satisfies the condition mentioned above (with respect to $x^{(i)}$), then the algorithm terminates after a finite number of iterations and returns a global primal optimizer.
Hierarchical partition

For a discrete set $S$, consider its hierarchical partition, i.e. an ordered tuple $(S_0, \ldots, S_{K_d})$. Here

$$S_0 = \{\{s\} : s \in S\},$$

is the trivial partition into singletons and each subsequent level is generated by merging cells from the previous level.

Hierarchical measures

For probability measure $u \in \mathcal{P}(S)$, its multiscale measure approximation is the tuple $(u_0, \ldots, u_{K_d})$ of probability measures $u_k \in \mathcal{P}(S_k)$ defined by $u_k(\hat{S}) = u(\bigcup_{\hat{s} \in \hat{S}} x)$ for all subsets $\hat{S} \subset S_k$ and $k = 0, \ldots, K_d$. 

Multiscale Structure
Multiscale Structure

Now suppose the source set and destination set in the optimal transport problem is $S$ and $D$, also the source measure $u \in \mathcal{P}(S)$ and destination measure $v \in \mathcal{P}(D)$ are given. Then follow the process mentioned above we can construct two hierarchical $2^n$-trees $(S_0, \ldots, S_{K_d})$ and $(D_0, \ldots, D_{K_d})$ and correspondingly two series of probability measures $(u_0, \ldots, u_{K_d})$ and $(v_0, \ldots, v_{K_d})$. We can also generate a series of hierarchical costs generated by the original cost function.

**hierarchical cost function**

For a cost function $C : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$, let

$$C_k : S_k \times D_k \rightarrow \mathbb{R}, \quad C_k(s, d) = C(\text{center}(s), \text{center}(d)), \quad 1 \leq k \leq K_d.$$ (79)
Multiscale Structure

We are now ready to describe a hierarchical series of OT problems: for $1 \leq k \leq K_d$, the $k$-th OT problem is generated by the source set $S_k$, the destination set $D_k$, the source measure $u_k$, the destination measure $v_k$ and the cost function $C_k$:

Multiscale OT

\[
\begin{align*}
\min_{X \in \mathbb{R}^{|S_k| \times |D_k|}} & \quad \langle C_k, X \rangle, \\
\text{subject to} & \quad \sum_{j=1}^{|D_k|} X_{ij} = u_{k,i}, \quad 1 \leq i \leq |S_k|, \\
& \quad \sum_{i=1}^{|S_k|} X_{ij} = v_{k,j}, \quad 1 \leq j \leq |D_k|, \\
& \quad X_{ij} \geq 0, \quad 1 \leq i \leq |S_k|, 1 \leq j \leq |D_k|. 
\end{align*}
\]

(80)

We also consider restricting these OT problems to some corresponding subset $N_k \subset S_k \times D_k$. 
Multiscale Semismooth Newton method for OT

Algorithm 2: Multiscale Semismooth Newton Method

1. Solve the dense OT problem at the $K_d$-th level and get a solution $x$.

2. for $k = K_d - 1, K_d - 2, \ldots$ do
   
   3. $\mathcal{N}^{(1)} \leftarrow \bigcup_{(s,d) \in \text{spt}(x)} (\text{children}(s) \times \text{children}(d))$.
   
   4. while $\langle c_{\mathcal{N}(i)}, x^{(i)} \rangle \neq \langle c_{\mathcal{N}(i-1)}, x^{(i-1)} \rangle$ do
      
      5. Reduce $A$ and $b$ w.r.t. the neighborhood $\mathcal{N}_i$.
      
      6. while not converged do
         
         7. Compute $F, D, H = (\mu + 1)I - D$, and $M = A_{\mathcal{N}(i)}KA_{\mathcal{N}(i)}^T$.
         
         8. $d \leftarrow -A_{\mathcal{N}(i)}(I - 2D)H^{-1}F$.
         
         9. Use PCG to compute $d \leftarrow M^{-1}d$ inexacty.
         
         10. $d \leftarrow H^{-1}A_{\mathcal{N}(i)}^Td - H^{-1}F$, $u \leftarrow z + d$.
         
         11. Compute the ratio $\rho$ as in (66).
         
         12. If $\rho \geq \eta$, then update $z = u$.
         
         13. Update $\mu$ and $\eta$.
         
         14. $x^{(i)} = \max(z, 0)$.
         
         15. Construct a new neighborhood $\mathcal{N}^{(i+1)}$ from $x^{(i)}$.
         
         16. $i \leftarrow i + 1$.
         
         17. $x \leftarrow x^{(i)}$. 

18. end while

19. end while

20. end for
This is a set of $256 \times 256$ miscellaneous images collected by the research group CVG $^1$ from University of Granada. Here we test our algorithm on the g256_002 and the g256_006 image sets, which consists of 12 and 9 grayscale images with $256 \times 256$ pixels, respectively. For each of the image sets, we fix one image as the source distribution and set all the others as the target distribution. This results in a total of 19 OT instances, each of them could be formulated as an LP with 4,294,967,296 nonnegative variables and 131,072 equality constraints. Note that we normalize the sum of all pixels to eliminate the possible imbalance between different images.

$^1$http://decsai.ugr.es/cvg/introduccion.php
g256 dataset

Figure: Images in g256_002 (size: 256 × 256).
g256 dataset: $\ell_2$ squared cost

<table>
<thead>
<tr>
<th>Image Pair</th>
<th>MSSN</th>
<th>MCPPLEX</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TIME/SSN</td>
<td>gap/pinf/dinf</td>
</tr>
<tr>
<td>indor1</td>
<td>77.15/2165</td>
<td>1.1e-7/5.6e-9/4.2e-7</td>
</tr>
<tr>
<td>indor1</td>
<td>60.50/4432</td>
<td>5.7e-9/5.5e-9/9.3e-7</td>
</tr>
<tr>
<td>indor1</td>
<td>71.88/2404</td>
<td>3.9e-8/5.1e-9/3.8e-7</td>
</tr>
<tr>
<td>indor1</td>
<td>84.21/2405</td>
<td>6.6e-8/4.8e-9/9.2e-8</td>
</tr>
<tr>
<td>indor1</td>
<td>46.06/1609</td>
<td>8.4e-8/5.5e-9/1.2e-7</td>
</tr>
<tr>
<td>indor1</td>
<td>84.71/2140</td>
<td>2.5e-9/5.5e-9/5.0e-7</td>
</tr>
<tr>
<td>indor1</td>
<td>71.56/2824</td>
<td>3.8e-8/5.7e-9/2.9e-7</td>
</tr>
<tr>
<td>indor1</td>
<td>52.86/2395</td>
<td>2.3e-8/5.4e-9/3.5e-7</td>
</tr>
<tr>
<td>indor1</td>
<td>47.58/1651</td>
<td>1.3e-9/4.4e-9/2.2e-7</td>
</tr>
<tr>
<td>indor1</td>
<td>76.78/1645</td>
<td>7.5e-8/5.2e-9/2.4e-7</td>
</tr>
<tr>
<td>indor1</td>
<td>90.35/4802</td>
<td>4.2e-8/5.2e-9/1.6e-8</td>
</tr>
</tbody>
</table>

**Table:** OT distance (Euclidean squared distance) on g256_002.

https://github.com/rflamary/POT
