Optimal Transport

http://bicmr.pku.edu.cn/~wenzw/bigdata2019.html

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

1. Problem Formulation
2. Applications
3. Entropic Regularization
4. Sinkhorn’s Algorithm
5. Sinkhorn-Newton method
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_i\}_{i}, \{y_j\}_{j}: points clouds.
- \(a_i, b_j\): 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

\[ U(a, b) \overset{\text{def}}{=} \{ P \in \mathbb{R}_{+}^{m \times n}; P1_n = a, P^\top 1_m = b \} \]  

(2)

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

Discrete Optimal Transport

In the Optimal transport, we want to compute the following quantity

\[ L_C(a, b) \overset{\text{def}}{=} \min \left\{ \sum_{i,j} C_{i,j} \mathbf{P}_{i,j} ; \mathbf{P} \in \mathbf{U}(a, b) \right\} . \]  

(3)
Push Forward

- Radon measures \((\alpha, \beta)\) on \((\mathcal{X}, \mathcal{Y})\).
- Transfer of measure by \(T : \mathcal{X} \rightarrow \mathcal{Y}\): push forward.
- The measure \(T\#\alpha\) on \(\mathcal{Y}\) is defined by
  \[
  T\#\alpha(Y) = \alpha(T^{-1}(Y)), \quad \text{for all measurable } Y \in \mathcal{Y}.
  \]

  Equivalently,
  \[
  \int_{\mathcal{Y}} g(y) dT\#\alpha(y) \overset{\text{def}}{=} \int_{\mathcal{X}} g(T(x)) d\alpha(x).
  \]

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

  \[
  T\#\alpha = \beta \iff \rho(T(x)) |\text{det}(\partial T(x))| = \xi(x).
  \]
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:

$$\min_T \int_{\chi} c(x, T(x))d\alpha(x), \quad \text{s.t.} \quad T\#\alpha = \beta$$
Projectors:

\[ P_X : (x, y) \in X \times Y \rightarrow x \in X, \]
\[ P_Y : (x, y) \in X \times Y \rightarrow y \in Y. \]  

(7)

\[ \mathcal{U}(\alpha, \beta) \overset{\text{def}}{=} \{ \pi \in \mathcal{M}_+(X \times Y); P_X \# \pi = \alpha, P_Y \# \pi = \beta \}. \]

(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
Kantorovitch Problem for General Measures

Optimal transport distance between General Measures

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

(9)

Probability interpretation:

\[ \min_{(X, Y)} \{ \mathbb{E}_{(X,Y)}(c(X, Y)), X \sim \alpha, Y \sim \beta \}. \]

(10)
Wasserstein Distance

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

Wasserstein Distance

$$W_p(\alpha, \beta) \overset{\text{def}}{=} \mathcal{L}_{dp}(\alpha, \beta)^{1/p}. \quad (11)$$

Theorem

$W_p$ is a distance, and

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

Example

$$W_p(\delta_x, \delta_y) = d(x, y). \quad (13)$$
Dual form

Dual problem (discrete case)

\[
\begin{align*}
\max_{f \in \mathbb{R}^m, g \in \mathbb{R}^n} & \quad f^\top a + g^\top b, \\
\text{subject to} & \quad f_i + g_j \leq C_{ij}, \quad \forall(i,j)
\end{align*}
\]  

(14)

Relation between any primal and dual solutions:

\[ P_{ij} > 0 \Rightarrow f_i + g_j = C_{ij}. \]
Outline

1. Problem Formulation
2. Applications
3. Entropic Regularization
4. Sinkhorn’s Algorithm
5. Sinkhorn-Newton method
Comparing Measures

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

• Optimal transport
  → takes into account a metric $d$. 

$L^2$ mean

Optimal transport mean
Applications: toward high-dimensional OT

Toward High-dimensional OT

Monge  Kantorovich  Dantzig  Brenier  Otto  McCann  Villani
Applications: Grayscale image equalization

Grayscale Image Equalization

$t = 0 \quad t = 0.25 \quad t = 0.5 \quad t = .75 \quad t = 1$
Applications: image color palette equalization

Image Color Palette Equalization

Optimal transport
Applications: shape interpolation

Shape Interpolation
MRI Data Processing [with A. Gramfort]

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

$L^2$ barycenter

$W_2^2$ barycenter
Applications

Gradient Flows Simulation

https://www.youtube.com/watch?v=tDQw21ntR64
Tim Whittaker (New Zealand)
Applications

**OT Loss for Diffeomorphic Registration**

*Joint work with J. Feydy, B. Charier, F-X. Vialard.*

Shape registration: \[
\min_{\varphi \text{ diffeo}} D(\varphi(\mu), \nu) + R(\varphi)
\]

Hilbertian loss (MMD/RKHS):
\[
D(\mu, \nu) = \|k_\sigma * (\mu - \nu)\|_2^2
\]

Sinkhorn divergence:
\[
D(\mu, \nu) = \tilde{W}_\epsilon(\mu, \nu)
\]

→ Do not use OT for registration . . . but as a loss.
→ Sinkhorn’s iterates “propagate” a small bandwidth kernel.
→ Automatic differentiation: game changer for advanced loss and models.
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: topic models


Topic Models
Applications

Shapes Analysis with Gromov-Wasserstein

Use $T$ to define registration between:

- **Shape**
- **Shape**
- **Colors distribution**
- **Shape**

**Shapes** $(X_s)_s$ → **Geodesic distances** $d_s = (D_{X_s}(x_i, x_{i'}))_{i,i'}$ → **GW distances** $(GW_{\varepsilon}(d_s, d_{s'}))_{s,s'}$ → **MDS Visualization**

MDS in 2-D

MDS in 3-D

Source

Targets
Outline

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

$$\Sigma_n \overset{\text{def}}{=} \left\{ a \in \mathbb{R}_{n}^+ ; \sum_{i=1}^{n} a_i = 1 \right\}$$ (15)

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

$$L_C(a, b) \overset{\text{def}}{=} \min \left\{ \sum_{i,j} C_{i,j} P_{i,j} ; P \in U(a, b) \right\}.$$ (16)

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

$$H(P) \overset{\text{def}}{=} - \sum_{i,j} P_{i,j} (\log(P_{i,j}) - 1).$$  \hspace{1cm} (17)

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

$$H(u) \overset{\text{def}}{=} - \sum_i u_i (\log(u_i) - 1).$$ \hspace{1cm} (18)

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

$$\text{KL}(u \parallel v) = - \sum_{i=1}^{n} u_i \log\left(\frac{v_i}{u_i}\right).$$ \hspace{1cm} (19)

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

$$L^\varepsilon_C(a, b) = \min_{P \in \mathcal{U}(a, b)} \langle P, C \rangle - \varepsilon H(P).$$ (20)

The case $\varepsilon = 0$ corresponds to the classic (linear) optimal transport problem.

For $\varepsilon > 0$, problem (20) has an $\varepsilon$-strongly convex objective and therefore admits a unique optimal solution $P^*_\varepsilon$.

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

**Proposition**

When $\varepsilon \to 0$, the unique solution $P_\varepsilon$ of (20) converges to the optimal solution with maximal entropy within the set of all optimal solutions of the unregularized transportation problem, namely,

$$P_\varepsilon \xrightarrow{\varepsilon \to 0} \arg\max_P \{ H(P); P \in U(a, b), \langle P, C \rangle = L_C^0(a, b) \}$$

The above proposition motivates us to solve the problems in (20) sequentially and then take $\varepsilon \to 0$. 
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 \(\langle C, P \rangle = L_0^C(a, b)\). By optimality of \(P\) and \(P_\ell\) for their respective optimization problems (for \(\varepsilon = 0\) and \(\varepsilon = \varepsilon_\ell\)), one has

\[
0 \leq \langle C, P_\ell \rangle - \langle C, P \rangle \leq \varepsilon_\ell (H(P_\ell) - H(P)). \tag{22}
\]

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

$$U_\alpha(a, b) \overset{\text{def}}{=} \{ P \in U(a, b) | \KL(P \| ab^\top) \leq \alpha \} \quad (23)$$

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

**Definition: Sinkhorn Distance**

$$d_{C,\alpha}(a, b) \overset{\text{def}}{=} \min_{P \in U_\alpha(a, b)} \langle C, P \rangle. \quad (24)$$

**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

$$H(P) \geq \frac{1}{2}(H(a) + H(b)), \quad (25)$$

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

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

1. Problem Formulation
2. Applications
3. Entropic Regularization
4. Sinkhorn’s Algorithm
5. Sinkhorn-Newton method
Sinkhorn’s algorithm

The Sinkhorn’s algorithm exploits the following result, which was first introduced in [2].

Proposition

Let $a_1, a_2, \ldots, a_m, b_1, \ldots, b_n$ be fixed positive numbers. Then, corresponding to each positive $m \times n$ matrix $A$, there is a unique matrix $D_1AD_2$ with row sums $\mu a_1, \mu a_2, \ldots, \mu a_m$ and column sums $b_1, b_2, \ldots b_n$, where $\mu = \sum_j b_j / \sum_i a_i$, $D_1$ and $D_2$ are respectively $m \times m$ and $n \times n$ with positive diagonals and are themselves unique up to a scalar multiple.

The iterative process of alternately scaling the rows and columns of $A$ to have row and column sums respectively $a_i$ and $b_j$ can be used to find $D_1AD_2$. The subsequence from the iteration in which column sums are scaled converges to $D_1AD_2$ while the subsequence in which the row sums are scaled converges to $(1/\mu)D_1AD_2$. 

Sinkhorn’s algorithm

For solving (20), consider its Lagrangian dual function

\[
\mathcal{L}_C^\varepsilon(P, \alpha, \beta) = \langle C, P \rangle - \varepsilon H(P) + \alpha^\top (P1_n - a) + \beta^\top (P^\top 1_m - b). \tag{26}
\]

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

\[
p_{ij} = e^{-\frac{c_{ij} + \alpha_i + \beta_j}{\varepsilon}}, \tag{27}
\]

so we can write

\[
P_\varepsilon = \text{diag}(e^{-\frac{\alpha}{\varepsilon}}) e^{-\frac{C}{\varepsilon}} \text{diag}(e^{-\frac{\beta}{\varepsilon}}). \tag{28}
\]

Note that

\[
P_\varepsilon 1_n = a, \quad P_\varepsilon^\top 1_m = b, \tag{29}
\]

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

Let \( u = \text{diag}(e^{-\alpha \varepsilon}) \), \( v = \text{diag}(e^{-\beta \varepsilon}) \) and \( K = e^{-C/\varepsilon} \). We again state the KKT system of (20):

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

(30)

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

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

(31)
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 \( u \):
   \[
   u = 1. / (\hat{K}(b. / (K^\top u)))
   \]
   until reaches certain stopping criterion.
5. Compute
   \[
   v = b. / (K^\top u),
   \]
   and eventually
   \[
   P_\varepsilon = \text{diag}(u)K\text{diag}(v).\]
Sinkhorn’s algorithm: Parallelism

As can be seen right above, Sinkhorn’s algorithm can be vectorized and generalized to \( N \) target histograms \( b^{(1)}, \ldots, b^{(N)} \). When \( N > 1 \), the computations for \( N \) target histograms can be simultaneously carried out by updating a single matrix of scaling factors \( u \in \mathbb{R}^{m \times N} \) instead of updating a scaling vector \( u \in \mathbb{R}^{d} \).

This important observation makes the execution of Sinkhorn’s algorithm particularly suited to GPU platforms.
Parallelized Sinkhorn’s algorithm

1. \( \mathbf{B} = [b^{(1)}, b^{(2)}, \ldots, b^{(N)}] \).

2. Compute \( \mathbf{K} = e^{-\frac{C}{\epsilon}} \).

3. Compute \( \hat{\mathbf{K}} = \text{diag}(a^{-1}) \mathbf{K} \).

4. Initial scale factor \( \mathbf{u} \in \mathbb{R}^{m \times N} \).

5. Iteratively update \( \mathbf{u} \):

\[
\mathbf{u} = 1./ (\hat{\mathbf{K}} (\mathbf{B}./(\mathbf{K}^\top \mathbf{u}))),
\]

until reaches certain stopping criterion.

6. Compute

\[
\mathbf{v} = \mathbf{B}./(\mathbf{K}^\top \mathbf{u}),
\]

and eventually

\[
\mathbf{P}_{\epsilon,i} = \text{diag}(\mathbf{u}(::i)) \mathbf{K} \text{diag}(\mathbf{v}(::i)), \quad i = 1, \ldots N.
\]
Outline

1. Problem Formulation
2. Applications
3. Entropic Regularization
4. Sinkhorn’s Algorithm
5. Sinkhorn-Newton method
The dual problem of (20) is

$$\min_{\alpha, \beta} \langle a, \alpha \rangle + \langle b, \beta \rangle + \epsilon \langle e^{-\frac{\alpha}{\epsilon}}, Ke^{-\frac{\beta}{\epsilon}} \rangle,$$

s.t. \[ \text{diag}(e^{-\frac{\alpha}{\epsilon}})Ke^{-\frac{\beta}{\epsilon}} = a, \]

\[ \text{diag}(e^{-\frac{\beta}{\epsilon}})K^\top e^{-\frac{\alpha}{\epsilon}} = b. \]

(32)

with $\alpha, \beta$ being the dual variables.

[1] proposes using Newton method to solve this system.
Sinkhorn-Newton method

Let

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

(33)

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

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

(34)

The Newton iteration is given by

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

(35)

where

\[
J_F = \frac{1}{\varepsilon}\begin{pmatrix}
\text{diag}(P1_n) & P \\
P^\top & \text{diag}(P^\top 1_m)
\end{pmatrix}.
\]  

(36)
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

$$\ker(J_F(\alpha, \beta)) = \text{span}\left\{ \begin{pmatrix} 1_m \\ -1_n \end{pmatrix} \right\}. \quad (37)$$

Proof

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

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

which holds with equality if and only if $\gamma_i + \phi_j = 0$ for all $i, j$, leading us to (37).
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

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

(38)

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

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

(39)

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:

\[
\|e^{(k+1)}\| = \|x^{(k)} - (F'(x^{(k)}))^{-1}(F(x^{(k)}) - F(x^*)) - x^*\|
\]

\[
= \|e^{(k)} - (F'(x^{(k)}))^{-1}(F(x^{(k)}) - F(x^*))\|
\]

\[
= \|(F'(x^{(k)}))^{-1}((F(x^*) - F(x^{(k)})) + F'(x^{(k)})e^{(k)})\|
\]

\[
= \|(F'(x^{(k)}))^{-1} \int_{s=0}^{-1} (F'(x^{(k)} + se^{(k)}) - F'(x^{(k)}))e^{(k)} ds\|
\]

\[
\leq \omega \| \int_{s=0}^{-1} s ds \| \|e^{(k)}\|^2 = \frac{\omega}{2} \|e^{(k)}\|^2 < \|e^{(k)}\|.
\]

Also

\[
\omega \|e^{(k+1)}\| \leq \omega \|e^{(k)}\| < 2.
\]

For the uniqueness part, let $x^{(0)} = x^{**} \neq x^*$ be a different solution, then $x^{(1)} = x^{**}$, then consider (39) when $k = 0$. 

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

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

(42)

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

The proof for this proposition is tedious and therefore we refer the interested readers to the paper [1].
Let $u = \text{diag}(e^{-\frac{\alpha}{\varepsilon}})$, $v = \text{diag}(e^{-\frac{\beta}{\varepsilon}})$ and $K = e^{-C/\varepsilon}$. We again state the KKT system of (20):

$$P_\varepsilon = \text{diag}(u) K \text{diag}(v),$$

$$a = \text{diag}(u) K v,$$

$$b = \text{diag}(v) K^\top u.$$ (43)

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

$$u^{(k+1)} = \text{diag}(K v^{(k)})^{-1} a,$$

$$v^{(k+1)} = \text{diag}(K^\top u^{(k+1)})^{-1} b.$$ (44)
Relationship with Sinkhorn’s algorithm

Define

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

(45)

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

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

(46)

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

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

(47)

and perform the Newton iteration

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

(48)
Relationship with Sinkhorn’s algorithm

We get

\[ u^{(k+1)} = \text{diag}(K v^{(k)})^{-1} a, \]
\[ v^{(k+1)} = \text{diag}(K^\top u^{(k)})^{-1} b. \] (49)

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