Optimal Transportation

2019 Winter Camp at PHBS

Xianhua Peng
Associate Professor
PHBS, Peking University

January 2019

Outline

1. Optimal Transportation
2. Sinkhorn distance
3. Computing Sinkhorn Distances

Outline

1 Optimal Transportation

2 Sinkhorn distance

3 Computing Sinkhorn Distances

Frobenius dot-product of two matrices

$$\langle A, B \rangle := \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} B_{ij}$$

Transportation polytope

- For two histograms r and c in the simplex

$$\Sigma_d := \{x \in \mathbb{R}_+^d : x^T 1_d = 1\},$$

we write $U(r, c)$ for the transportation polytope of r and c, namely the polyhedral set of $d \times d$ matrices:

$$U(r, c) := \{P \in \mathbb{R}_+^{d \times d} \mid P 1_d = r, \; P^T 1_d = c\},$$

where 1_d is the d dimensional vector of ones.

- $U(r, c)$ contains all nonnegative $d \times d$ matrices with row and column sums r and c respectively.

Transportation polytope

- $U(r, c)$ has a probabilistic interpretation: for X and Y two random variables taking values in $\{1, \cdots, d\}$, each with marginal distribution r and c respectively, the set $U(r, c)$ contains all possible *joint probabilities* of (X, Y).
- Indeed, any matrix $P \in U(r, c)$ can be identified with a joint probability distribution for (X, Y) such that $p(X = i, Y = j) = p_{ij}$. Such joint probabilities are also known as *contingency tables*.

Entropy

The entropy h and the Kullback-Leibler divergences of these tables and their marginals are

$$r \in \Sigma_d,\ h(r) := -\sum_{i=1}^{d} r_i \log r_i,$$

$$P \in U(r,\ c), h(P) := -\sum_{i,j=1}^{d} p_{ij} \log p_{ij}$$

$$P, Q \in U(r,\ c), KL(P\|Q) = \sum_{ij} p_{ij} \log \frac{p_{ij}}{q_{ij}}.$$

Optimal Transportation

- Given a $d \times d$ cost matrix M, the cost of mapping r to c using a transportation matrix (or joint probability) P can be quantified as

$$\langle P, M \rangle.$$

- The following problem:

$$d_M(r,\ c) := \min_{P \in U(r,c)} \langle P,\ M \rangle$$

  is called an *optimal transportation* problem between r and c given cost M.
- An optimal table $P^\star$ for this problem can be obtained by solving a linear programming problem

Optimal Transportation

- The optimum of this problem, $d_M(r, c)$, is a distance whenever the matrix M is itself a metric matrix, namely whenever M belongs to the cone of distance matrices

$$\mathrm{M} = \{M \in \mathbb{R}_+^{d\times d} \ : \ \forall i \leq d,\ m_{ii} = 0; \forall i, j, k \leq d,\ m_{ij} \ \leq m_{ik} + m_{kj}\}.$$

- For a general matrix M, the worst case complexity of computing that optimum with any of the algorithms known so far scales in $O(d^3 \log d)$

Outline

1. Optimal Transportation
2. Sinkhorn distance
3. Computing Sinkhorn Distances

- The optimum P^* of classical optimal transportation distances is achieved on vertices of $U(r, c)$, that is $d \times d$ matrices with only up to $2d - 1$ non-zero elements
- The optimal P^* can be interpreted as quasi-deterministic joint probabilities, since if $p_{ij} > 0$, then very few values $p_{ij'}$ will have a non-zero probability.
- The optimal P^* is an “extreme" joint distribution
- Sinkhorn distance provides a more smooth version of distance between marginals r and c.

An information theoretic inequality

A basic information theoretic inequality: for $\forall r,\ c \in \Sigma_d,\ \forall P \in U(r,\ c)$,

$$h(P)\ \leq h(r) + h(c),$$

where the equality holds if $P = rc^T$, known as the independence table.

A subset $U_\alpha(r, c)$ of $U(r, c)$

- Define the convex set $U_\alpha(r, c) \subset U(r, c)$:

$$
\begin{aligned}
U_\alpha(r, c) := & \{P \in U(r, c) \mid KL(P\|rc^T) \leq \alpha\} \\
& \{P \in U(r, c) \mid h(r) + h(c) - h(P) \leq \alpha\} \\
& \{P \in U(r, c) \mid h(P) \ \geq h(r) + h(c) - \alpha\}
\end{aligned}
$$

- $KL(P\|rc^T) = h(r) + h(c) - h(P)$ is also the mutual information $I(X\|Y)$ of two random variables (X, Y) should they follow the joint probability P
- $U_\alpha(r, c)$ is the set of tables P in $U(r, c)$ which have *sufficient* entropy with respect to $h(r) + h(c)$, or joint probabilities which display a small enough *mutual information*.

Sinkhorn Distances

Definition

For given $\alpha > 0$, the Sinkhorn distances $d_{M,\alpha}(r,\ c)$ is defined as

$$\begin{aligned} d_{M,\alpha}(r,\ c) &:= \min_{P \in U_\alpha(r,c)} \langle P, M \rangle, \\ &= \min_{P \in U(r,\ c), h(P)\ \geq h(r)+h(c)-\alpha} \langle P, M \rangle. \end{aligned}$$

- The entropic constraint “smooths" the optimal P^*
- The entropic constraint provides a regularization of the optimal P^*

Sinkhorn Distances

- *For α large enough, the Sinkhorn distance $d_{M,\alpha}$ is the transportation distance d_M.*
- The function $(r, c) \mapsto 1_{r \neq c} d_{M,\alpha}(r, c)$ *satisfies all three distance axioms*.

Outline

1. Optimal Transportation
2. Sinkhorn distance
3. Computing Sinkhorn Distances

Alternative formulation of the Sinkhorn distance by Lagrange multiplier

- Recall that

$$d_{M,\alpha}(r,\ c) = \min_{P \in U(r,\ c), h(P)\ \geq h(r)+h(c)-\alpha} \langle P, M \rangle.$$

- Introducing a Lagrange multiplier λ for the entropic constraint $h(P)\ \geq h(r) + h(c) - \alpha$, and consider the problem

$$P^{\lambda} = \arg \min_{P \in U(r,c)} \langle P, M \rangle - \frac{1}{\lambda} h(P).$$

- Then, define

$$d_M^{\lambda}(r,\ c) := \langle P^{\lambda}, M \rangle$$

Alternative formulation of the Sinkhorn distance by Lagrange multiplier

- By duality theory, for every pair $(r,\ c)$, each α corresponds to an $\lambda \in [0,\ \infty]$ such that

$$d_{M,\alpha}(r,c) = d_M^\lambda(r,\ c).$$

- d_M^λ is called the dual-Sinkhorn divergence and can be computed at a much cheaper cost than the optimal transportation distance.

Solve for dual-Sinkhorn divergence

The problem:

$$P^{\lambda} = \arg \min_{P \in U(r,c)} \langle P, M \rangle - \frac{1}{\lambda} h(P).$$

- When $\lambda > 0$, the solution P^{λ} is unique by strict convexity of minus the entropy.
- Form the Lagrangian $\mathcal{L}(P,\ \alpha,\ \beta)$, where α, β are Lagrangian multiplier for the two equality constraints in $U(r,\ c)$:

$$\mathcal{L}(P,\ \alpha,\ \beta) = \sum_{ij} \left(p_{ij} m_{ij} + \frac{1}{\lambda} p_{ij} \log p_{ij} \right) + \alpha^T (P \mathbb{1}_d - r) + \beta^T (P^T \mathbb{1}_d - c)$$

Solve for dual-Sinkhorn divergence

- By the first order condition of optimality

$$\frac{\partial \mathcal{L}}{\partial p_{ij}^{\lambda}} = 0, \forall i, \forall j,$$

it follows that

$$\begin{aligned} p_{ij}^{\lambda} &= e^{-\frac{1}{2}-\lambda\alpha_i} e^{-\lambda m_{ij}} e^{-\frac{1}{2}-\lambda\beta_j}, \forall i, \forall j. \\ P^{\lambda} 1_d &= r \\ (P^{\lambda})^T 1_d &= c \end{aligned}$$

- By Sinkhorn and Knopp's theorem, P^{λ} is the *only matrix* with row-sum r and column-sum c of the form

$$\exists u, v > 0_d : P^{\lambda} = \mathrm{diag}(u) e^{-\lambda M} \mathrm{diag}(v).$$

Compute dual-Sinkhorn divergence

By Sinkhorn and Knopp's theorem, P^λ can be computed by the following iterative algorithm:

1. given: M, r, c, and λ, initialize $P^\lambda = e^{-\lambda M}$
2. repeat
   - scale the rows of P^λ such that the row sums match r
   - scale the columns of P^λ such that the column sums match c

   until convergence