大数据算法-2025 春

# Lecture 16: 最优传输

2025.5.13

*Lecturer:* 丁虎 *Scribe:* 王运韬

最优传输 (Optimal Transport, OT) 为比较概率分布提供了一个强大的几何工具，在数据科学领域有着广泛的应用。然而，对于大型数据集而言，精确 OT 的计算通常过于缓慢。本笔记介绍了 Sinkhorn 距离，它源于 OT 问题的熵正则化版本。熵正则化平滑了问题，使得通过 Sinkhorn 算法可以实现更快速的迭代求解。我们将探讨其公式、算法，并讨论为什么 Sinkhorn 距离已成为在实际大规模数据科学任务中应用 OT 原理的基石。

## 1 最优传输 (OT) 简介

### 1.1 问题背景：搬土堆

想象一下，你有两堆土，代表两个概率分布。最优传输，在其最初由 Monge 提出的形式中，旨在找到将第一堆土以最有效的方式移动以匹配第二堆土的形状，同时最小化所做的总功（例如，质量 × 距离）。

在数据科学中，这些“土堆”可以是直方图、词嵌入、图像像素强度或数据的任何其他离散表示。移动“土”的“成本”可以通过距离度量来定义（例如，特征向量之间的欧几里得距离）。

### 1.2 数学公式 (Kantorovich 松弛)

令 $r \in \mathbb{R}_+^n$ 和 $c \in \mathbb{R}_+^m$ 为两个离散概率分布（直方图），即 $\sum_{i=1}^n r_i = \sum_{j=1}^m c_j = 1$。（原论文也考虑了非归一化的正向量，这是一个轻微的推广。）令 $C \in \mathbb{R}_+^{n \times m}$ 为**成本矩阵**，其中 $C_{ij}$ 是将单位质量从分布 $r$ 的第 $i$ 个箱格运输到分布 $c$ 的第 $j$ 个箱格的成本。

目标是找到一个**传输计划**（或耦合）$P \in \mathbb{R}_+^{n \times m}$，它指定了从箱格 $r_i$ 到箱格 $c_j$ 流动多

少质量 $P_{ij}$。该计划必须满足边际约束：

$$\sum_{j=1}^{m} P_{ij} = r_i \quad \forall i = 1, \dots, n \quad (P\mathbf{1}_m = r) \tag{1}$$

$$\sum_{i=1}^{n} P_{ij} = c_j \quad \forall j = 1, \dots, m \quad (P^T\mathbf{1}_n = c) \tag{2}$$

所有此类有效传输计划的集合表示为 $U(r, c)$。

**最优传输距离**（或推土机距离、Wasserstein 距离）是最小的总成本：

$$L_C(r, c) = \min_{P \in U(r,c)} \sum_{i=1}^{n} \sum_{j=1}^{m} P_{ij} C_{ij} = \min_{P \in U(r,c)} \langle P, C \rangle_F$$

其中 $\langle \cdot, \cdot \rangle_F$ 是 Frobenius 内积。

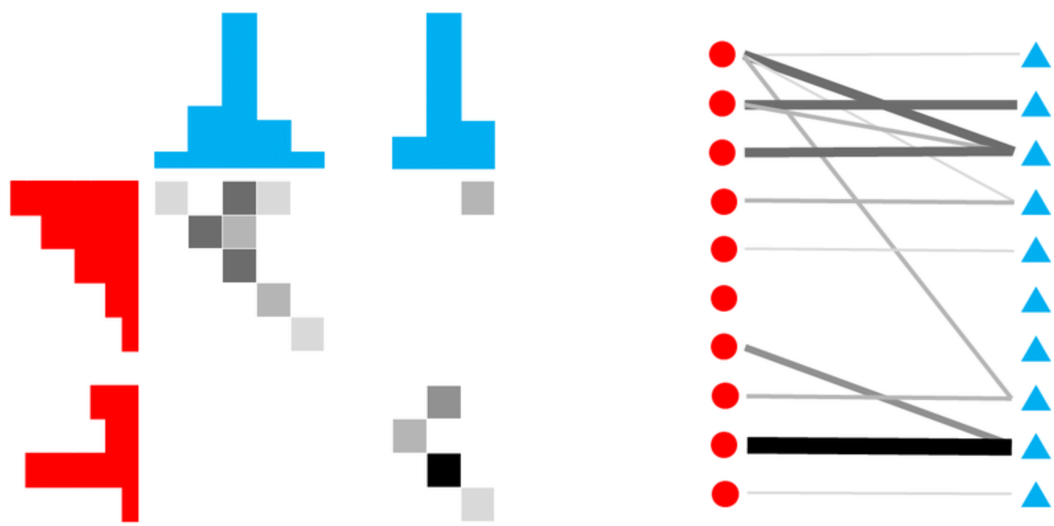

Figure 1: 最优传输概念示意图。左侧矩阵中的项对应两个分布的支撑点对应的 coupling 权重，颜色越深代表数值越大。右图是直观的匹配。

## 2 挑战：计算成本

如上定义的 OT 问题是一个线性规划问题。对于具有 $n$ 和 $m$ 个箱格的分布，精确求解器复杂度约为 $O((n+m)^3 \log(n+m))$，如果 $n \approx m \approx N$，则为 $O(N^3 \log N)$。这对于大的

$N$（例如，$N > 1000$）来说计算成本过高，而这在许多数据科学应用中很常见（例如，比较高分辨率图像或大型词汇表）。

# 3 熵正则化：一条更平滑的路径

为了克服计算障碍，Cuturi (2013) 推广了在 OT 问题中添加**熵正则化**项的思想。

## 3.1 添加熵

传输计划 $P$ 的熵定义为：

$$H(P) = -\sum_{i,j} P_{ij}(\log P_{ij} - 1)$$

最大化熵鼓励传输计划 $P$ 更“平滑”或更分散，避免过于稀疏的解。

## 3.2 正则化 OT 问题

熵正则化 OT 问题为：

$$L_C^\lambda(r,c) = \min_{P \in U(r,c)} \left( \sum_{i,j} P_{ij} C_{ij} - \lambda H(P) \right)$$

这里，$\lambda > 0$ 是**正则化参数**。

- 当 $\lambda \to 0$ 时，问题接近原始（未正则化）的 OT 问题。
- 当 $\lambda \to \infty$ 时，熵项占主导地位，$P_{ij}$ 趋向于 $r_i c_j$（边际的乘积，忽略成本）。
- 对于有限的 $\lambda > 0$，我们得到一个权衡：一个计划 $P^\lambda$ 平衡了最小化真实传输成本和保持足够的熵。

关键的洞察是，这个正则化问题是严格凸的，并且可以更有效地求解。

## 3.3 关于解空间的性质

在讨论 Sinkhorn 算法之前，我们先引入两个与熵约束相关的引理。这里我们定义香农熵 (Shannon entropy) 为 $h(X) = -\sum_k X_k \log X_k$。注意这与上文定义的 $H(P)$ 略有不同：$H(P) = h(P) + \sum_{i,j} P_{ij}$。如果 $P$ 是一个概率分布 ($\sum P_{ij} = 1$)，则 $H(P) = h(P) + 1$。

**Lemma 3.1.** 令 $U_\alpha(r,c) := \{P \in U(r,c) | KL(P||rc^T) \le \alpha\}$。则 $U_\alpha(r,c) = \{P \in U(r,c) | h(P) \ge h(r) + h(c) - \alpha\}$，并且 $U_\alpha(r,c) \subset U(r,c)$。

*Proof.* KL 散度定义为 $KL(P||Q) = \sum_{i,j} P_{ij} \log \frac{P_{ij}}{Q_{ij}}$。令 $Q_{ij} = r_i c_j$。则 $rc^T$ 表示 $r$ 和 $c$ 的独立耦合。

$$\begin{aligned}
KL(P||rc^T) &= \sum_{i,j} P_{ij} \log \frac{P_{ij}}{r_i c_j} \\
&= \sum_{i,j} P_{ij} \log P_{ij} - \sum_{i,j} P_{ij} \log r_i - \sum_{i,j} P_{ij} \log c_j \\
&= -h(P) - \sum_i \left( \sum_j P_{ij} \right) \log r_i - \sum_j \left( \sum_i P_{ij} \right) \log c_j
\end{aligned}$$

由于 $P \in U(r,c)$，我们有 $\sum_j P_{ij} = r_i$ 和 $\sum_i P_{ij} = c_j$。代入上式：

$$\begin{aligned}
KL(P||rc^T) &= -h(P) - \sum_i r_i \log r_i - \sum_j c_j \log c_j \\
&= -h(P) + h(r) + h(c)
\end{aligned}$$

因此，条件 $KL(P||rc^T) \le \alpha$ 等价于 $-h(P) + h(r) + h(c) \le \alpha$，即 $h(P) \ge h(r) + h(c) - \alpha$。$U_\alpha(r,c) \subset U(r,c)$ 的关系由 $U_\alpha(r,c)$ 的定义直接得出，因为它是在 $U(r,c)$ 的基础上增加了一个额外的约束。 □

**Lemma 3.2.** 集合 $U_\alpha(r,c)$ 是凸集。

*Proof.* 首先，集合 $U(r,c) = \{P \in \mathbb{R}_+^{n\times m} | P1_m = r, P^T 1_n = c\}$ 是凸集。这是因为它是由一系列线性等式约束 ($P1_m = r, P^T 1_n = c$) 和非负约束 ($P_{ij} \ge 0$) 定义的，这些约束共同形成一个多面体，因此是凸的。

其次，函数 $f(P) = KL(P||rc^T)$ 对于固定的 $rc^T$ 是关于 $P$ 的凸函数。因此，集合 $\{P \in \mathbb{R}_+^{n\times m} | KL(P||rc^T) \le \alpha\}$ 是一个凸函数的子水平集 (sublevel set)，所以它也是凸集。

集合 $U_\alpha(r,c)$ 是两个凸集 $U(r,c)$ 和 $\{P | KL(P||rc^T) \le \alpha\}$ 的交集。两个凸集的交集仍然是凸集。因此，$U_\alpha(r,c)$ 是凸集。 □

**Theorem 3.3.** 对任意 $\alpha \ge 0$, $M \in \mathcal{M}$, $d_{M,\alpha}$ 都对称且满足三角不等式。

**Lemma 3.4** (熵约束下的粘合引理)**.** 设 $\alpha \ge 0$，$x, y, z$ 为 $\Sigma_d$ 中的三个元素。设 $P \in U_\alpha(x,y)$ 和 $Q \in U_\alpha(y,z)$ 分别为 $(x,y)$ 和 $(y,z)$ 对应传输多面体中满足熵约束的两个联合概率分布。设 $S$ 为 $d \times d$ 矩阵，其第 $(i,k)$ 个元素定义为 $s_{ik} = \sum_j \frac{p_{ij} q_{jk}}{y_j}$。则 $S \in U_\alpha(x,z)$。

证明见 Cuturi 原文。

*Proof of Theorem 3.3.* $d_{M,\alpha}$ 的对称性源于 $M$ 的对称性. 设 $x, y, z \in \Sigma_d$. 设 $P \in U_\alpha(x, y)$ 和 $Q \in U_\alpha(y, z)$ 为分别计算 $d_{M,\alpha}(x, y)$ 和 $d_{M,\alpha}(y, z)$ 得到的最优解. 使用 Lemma **??**得到的 $S$ of $U_\alpha(x, z)$, 我们有如下的不等式:

$$\begin{aligned}
d_{M,\alpha}(x,z) &= \min_{P\in U_\alpha(x,z)} \langle X, M\rangle \le \langle S, M\rangle = \sum_{ik} m_{ik} \sum_j \frac{p_{ij}q_{jk}}{y_j} \\
&\le \sum_{ijk} (m_{ij} + m_{jk}) \frac{p_{ij}q_{jk}}{y_j} = \sum_{ijk} m_{ij}\frac{p_{ij}q_{jk}}{y_j} + m_{jk}\frac{p_{ij}q_{jk}}{y_j} \\
&= \sum_{ij} m_{ij}p_{ij} \sum_k \frac{q_{jk}}{y_j} + \sum_{jk} m_{jk}q_{jk} \sum_i \frac{p_{ij}}{y_j} \\
&= \sum_{ij} m_{ij}p_{ij} + \sum_{jk} m_{jk}q_{jk} = d_{M,\alpha}(x,y) + d_{M,\alpha}(y,z). \blacksquare
\end{aligned}$$

# 4 Sinkhorn 算法

熵正则化 OT 问题的解 $P^\lambda$ 具有特定结构:

$$P^\lambda_{ij} = u_i K_{ij} v_j$$

其中:

- $K$ 是一个 Gibbs 核矩阵，其中 $K_{ij} = e^{-C_{ij}/\lambda}$。
- $u \in \mathbb{R}^n_+$ 和 $v \in \mathbb{R}^m_+$ 是缩放向量（对偶变量）。

这些缩放向量 $u$ 和 $v$ 可以通过一个称为 **Sinkhorn 算法**（或 Sinkhorn-Knopp 算法，最初为矩阵缩放开发）的简单迭代过程找到。

## 4.1 迭代缩放

给定 $r, c, C$ 和 $\lambda$:

1. 计算 $K_{ij} = e^{-C_{ij}/\lambda}$。
2. 初始化 $v^{(0)} = 1_m$（长度为 $m$ 的全 1 向量）。

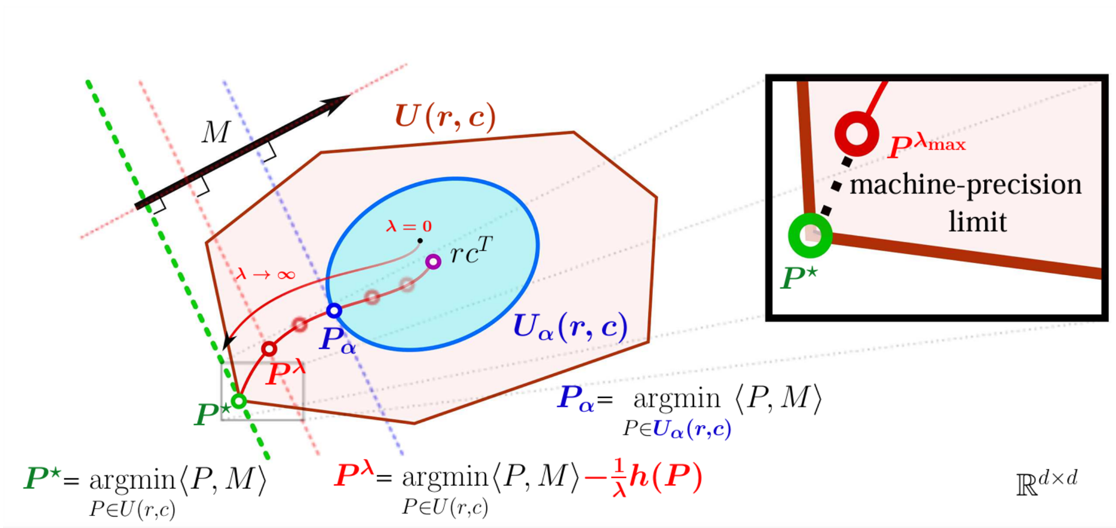


Figure 2: Sinkhorn 算法的迭代过程。可以看到，对于固定的 $\lambda$,$U_\alpha(r,c)$ 平滑地逼近原线性规划问题的凸多面体定义域。随着 $\lambda$ 变大，Sinkhorn 算法求得的解越来越趋近于最优传输问题的精确解。

3. 对于 $l = 0, 1, 2, \ldots$ 直到收敛：

$$u^{(l+1)} = r./(Kv^{(l)}) \quad \text{(逐元素除法)} \tag{3}$$

$$v^{(l+1)} = c./(K^T u^{(l+1)}) \quad \text{(逐元素除法)} \tag{4}$$

（这里，./ 表示逐元素除法。为了数值稳定性，通常会向分母添加小常数，或在对数空间中进行计算。）

一旦 $u$ 和 $v$ 收敛（例如，到 $u^*$ 和 $v^*$），最优正则化传输计划为 $P^\lambda = diag(u^*)Kdiag(v^*)$。

## 4.2 计算优势

Sinkhorn 算法的每次迭代都涉及矩阵-向量乘法（$Kv$ 和 $K^T u$），其成本为 $O(nm)$。该算法通常收敛非常快（几何级数收敛）。与精确 OT 求解器的 $O(N^3 \log N)$ 复杂度相比，这是一个巨大的速度提升，使其对于数千甚至更大的 $N$ 都是可行的。

# 5 Sinkhorn 距离

从熵正则化问题的解中获得的 $L_C^\lambda(r,c)$ 的传输成本部分 $\langle P^\lambda, C\rangle_F$ 通常被称为 **Sinkhorn 距离**。

$$d_{C,\lambda}(r,c) = \sum_{i,j} P_{ij}^\lambda C_{ij}$$

值得注意的是：

- Sinkhorn 距离是真实 OT 距离 $L_C(r,c)$ 的一个近似。正则化目标函数值 $L_C^\lambda(r,c)$ 是真实 OT 成本 $L_C(r,c)$ 的一个下界（如果 $\lambda H(P)$ 如目标函数中那样被减去）。更准确地说，$L_C^\lambda(r,c)$ 是正则化目标的值，而 Sinkhorn 距离本身是 $\langle P^\lambda, C\rangle_F$。对于 $\lambda > 0$，这个值通常大于 $L_C(r,c)$。
- 近似的质量取决于 $\lambda$。较小的 $\lambda$ 给出更接近的近似，但可能需要更多迭代并且可能存在数值不稳定性。较大的 $\lambda$ 导致更快的收敛和更稳定的计算，但是一个更粗糙的近似（一个更“模糊”的传输计划）。