AE 8803: Optimal Transport Theory and Applications
Spring 2023

Lecturer: Yongxin Chen

Lecture: 16
Scribe: Bipin Koirala

Date: 03/02/2023

---

1 Sinkhorn Algorithm

The Sinkhorn algorithm, also known as the Sinkhorn-Knopp algorithm or entropy-scaling method, is an efficient iterative method used for solving the entropy-regularized optimal transport problem. While not designed specifically for the Schrödinger Bridge Problem (SBP), it can be applied to the entropy-regularized version of SBP, as the regularized problem is a special case of the entropy-regularized optimal transport problem.

In a discrete setting the problem formulation of SBP is:

$$\min_{\pi \in \prod(\mu,\nu)} \sum_{ij} c_{ij}\ \pi_{ij} - \varepsilon H(\pi)$$

where μ and ν are the initial and final probability distributions, respectively, $\pi(x, y)$ is the joint probability distribution of the initial and final states, $c(x, y)$ is the cost function, $H(\pi)$ is the entropy of π, and ε is the regularization parameter.

We introduce the Lagrangian to solve the above linear programming.

$$\mathcal{L}(\pi, \alpha, \beta) = \sum_{ij} c_{ij}\ \pi_{ij} + \varepsilon \sum_{ij} \pi_{ij} \log \pi_{ij} + \sum_{i} \alpha_i \Big(\sum_{j} \pi_{ij} - \mu_i \Big) + \sum_{j} \beta_j \Big(\sum_{i} \pi_{ij} - \nu_j \Big)$$

where α, β are the Lagrange multipliers associated with the marginal μ and ν respectively.

Differentiating the Lagrangian w.r.t. π_{ij} and setting it to zero gives the following expression;

$$\pi_{ij} = \exp\Big(-1 - \frac{c_{ij}}{\varepsilon} - \frac{\alpha_i}{\varepsilon} - \frac{\beta_j}{\varepsilon} \Big)$$

Denote

$$\mathrm{u} = \exp\Big(-\frac{1}{2} - \frac{\alpha}{\varepsilon} \Big) \in \mathbb{R}_+^m$$
$$\mathrm{v} = \exp\Big(-\frac{1}{2} - \frac{\beta}{\varepsilon} \Big) \in \mathbb{R}_+^n$$
$$\mathrm{k}_{ij} = \exp\Big(-\frac{c_{ij}}{\varepsilon} \Big)$$

Sinkhorn algorithm iteratively updates the positive scaling vectors $\mathbf{u}$, $\mathbf{v}$, such that $\pi = (\mathrm{diag\ u})\ (K)\ (\mathrm{diag\ v}) \in \prod(\mu, \nu)$. It finds an optimal coupling matrix π^* such that the marginals match the given probability distributions μ and ν. The algorithm proceeds as follows:

1. Initialize $\mathbf{u}^{(0)}$ and $\mathbf{v}^{(0)}$ as vectors of ones

2. For $k = 0, 1, 2, \ldots$ until convergence:
   a. Update the scaling vector $\mathbf{u}$: $\mathbf{u}^{(k+1)} = \frac{\mu}{K^T \mathbf{v}^{(k)}}$
   b. Update the scaling vector $\mathbf{v}$: $\mathbf{v}^{(k+1)} = \frac{\nu}{K \mathbf{u}^{(k)}}$

3. Compute the optimal coupling matrix $\pi^* = \text{diag}(\mathbf{u}^{(k)})\ K\ \text{diag}(\mathbf{v}^{(k)})$

The algorithm iteratively normalizes the rows and columns of the matrix product $\text{diag}(\mathbf{u})\ K\ \text{diag}(\mathbf{v})$ to satisfy the marginal constraints for μ and ν. The convergence criterion for the algorithm can be based on a stopping tolerance for the change in $\mathbf{u}$ and $\mathbf{v}$ between iterations or a maximum number of iterations.

The Sinkhorn algorithm's efficiency stems from the fact that it leverages the structure of the problem and only requires matrix-vector multiplications, which can be performed quickly for sparse cost matrices. The entropy regularization in the Schrödinger Bridge Problem allows the use of the Sinkhorn algorithm, enabling the efficient computation of approximate solutions to the optimal transport problem.

Furthermore, it is worth noting that the variables $(\mathbf{u}, \mathbf{v})$ must satisfy the following two equations corresponding to the conservation of mass in the joint distribution $\prod(\mu, \nu)$.

$$\text{diag}(\mathbf{u}) K \text{diag}(\mathbf{v}) \mathbf{1}_m = \mu \qquad \text{and} \qquad \text{diag}(\mathbf{v}) K^T \text{diag}(\mathbf{u}) \mathbf{1}_n = \nu$$

1.1 History Behind Sinkhorn Algorithm

The Sinkhorn Algorithm is named after Richard Sinkhorn, who, along with Paul Knopp, independently studied the algorithm in the context of linear algebra in 1964. The Sinkhorn Algorithm is an iterative method used to compute a doubly stochastic matrix (a matrix with non-negative elements where the sum of each row and each column is equal to 1) from an arbitrary non-negative matrix.

The paper "On a least square adjustment of sample frequency table when the expected marginal totals are known" dates back to 1932, which is before the Sinkhorn Algorithm was introduced. However, it is worth noting that the problem studied in the paper is related to adjusting the observed frequency distributions in a contingency table when the expected marginal totals are known. Specifically, the paper discusses a problem where there are two variables - weights and wealth - both with 1 million data points. Instead of evaluating all 1 million samples, the authors sampled 1000 individuals and analyzed their joint distribution, R. The goal is to find an approximation, π, which is close to R, and belongs to the set of joint probability distributions with given marginals μ (wealth) and ν (weights). This is exactly a Schrödinger Bridge Problem (SBP). Therfore, this problem is, in some sense, a precursor to the problem that the Sinkhorn Algorithm addresses since both involve adjusting the elements of a matrix to fit certain criteria.

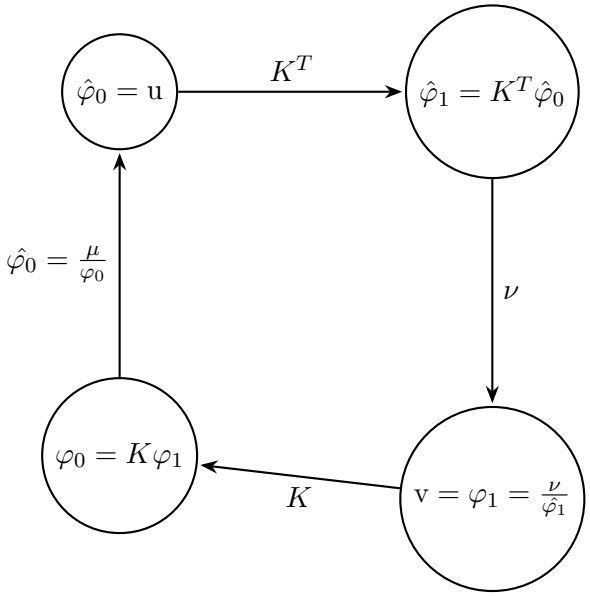
In 1964, Sinkhorn and Knopp independently studied the algorithm in the context of linear algebra, focusing on finding a doubly stochastic matrix. Sinkhorn's theorem states the following:

Theorem 1.1 *If $A \in \mathbb{R}^{n \times n}$ is a square matrix with positive entries, $\exists$ diagonal matrices D_1, D_2 s.t. $D_1 A D_2$ is doubly stochastic i.e. $\frac{1}{n} D_1 A D_2 \in \prod \left(\frac{1}{n}\mathbf{1}, \frac{1}{n}\mathbf{1}\right)$*

The Sinkhorn Algorithm provides an iterative method to find these diagonal matrices.

2 On Convergence of Sinkhorn Algorithm

Below is a schematic diagram for each iteration of the algorithm.



Recall, from SBP in a continuous-time setting, we have the following:

$$\rho_0 = \hat{\varphi}(0,\cdot)\ \varphi(0,\cdot) \qquad \partial_t\hat{\varphi} = \frac{1}{2}\Delta\hat{\varphi}$$
$$\rho_1 = \hat{\varphi}(1,\cdot)\ \varphi(1,\cdot) \qquad \partial_t\varphi = -\frac{1}{2}\Delta\varphi$$

It is clear that discrete setting $\hat{\varphi}_1 = K^T\hat{\varphi}_0$ corresponds to $\partial_t\hat{\varphi} = \frac{1}{2}\Delta\hat{\varphi}$, and $\varphi_0 = K\varphi_1$ corresponds to $\partial_t\varphi = -\frac{1}{2}\Delta\varphi$. This means we aim to find $\hat{\varphi}$ and φ so that $\hat{\varphi}$ solves the forward heat equation $\partial_t\hat{\varphi} = \frac{1}{2}\Delta\hat{\varphi}$ and φ solves the backward heat equation $\partial_t\varphi = -\frac{1}{2}\Delta\varphi$. In this sense, $\hat{\varphi}$ and φ have to match each other in the marginals.

With a slight abuse of notation, in regards to the figure above, the update in $\mathbf{u}$ can be written as $\mathbf{u}^{k+1} = \mathcal{D}_\mu \circ K \circ \mathcal{D}_\nu \circ K^T(\mathbf{u}^k)$. In order to show that the Sinkhorn algorithm converges, we need to show that this expression is a contractive map.

To show the contractiveness property of this map $\mathbf{u}^{k+1}$, we resort to Hilbert metric:

$$d_H(x,y) = \log\frac{\max_i(x_i/y_i)}{\min_i(x_i/y_i)} \geq 1 \quad ;\forall\ x,y \in \mathbb{R}^n_+$$

It turns out that $\mathcal{D}_\mu$ and $\mathcal{D}_\nu$ are isometric and K and K^T are strictly contractive w.r.t. the Hilbert metric. For a contractive ratio κ we have;

$$\kappa(K^T) = \kappa(K) = \tanh\left(\frac{1}{2}\log\frac{\max K_{ij}}{\min K_{ij}}\right) < 1$$

where $K_{ij} = \exp(-\frac{C_{ij}}{\varepsilon})$ and $\kappa(\mathcal{D}_\mu \circ K \circ \mathcal{D}_\nu \circ K^T) < 1$. Furthermore, when $\varepsilon \to 0;\ \kappa(K) \to 1$ and the mapping has a slow convergence rate.

2.1 An Alternative Way to View Sinkhorn Algorithm and Quantify Its Convergence

With a slight modification to the entropy regularized optimal transport (OT) problem we can write:

$$\min_{\pi \in \prod(\mu,\nu)} \langle c, \pi \rangle + \varepsilon \langle \pi, \log \pi - \log \mu - \log \nu - 1 \rangle$$

which is the same as SBP. The dual problem of this optimization problem is;

$$\min_{\alpha,\beta} \psi(\alpha, \beta) = \varepsilon \sum K_{ij} e^{\frac{\alpha_i}{\varepsilon} + \frac{\beta_j}{\varepsilon}} \mu_i \nu_j - \mathbf{u}^T \alpha - \mathbf{v}^T \beta$$

The optimal solution to this problem is: $\pi_{ij} = K_{ij} e^{\frac{\alpha_i}{\varepsilon} + \frac{\beta_j}{\varepsilon}} \mu_i \nu_j$. Let, $u = e^{\alpha/\varepsilon}$ and $v = e^{\beta/\varepsilon}$. With this notation, the dual problem becomes;

$$\min_{u,v} \psi(u, v) = \varepsilon \sum K_{ij} u_i v_j \mu_i \nu_j - \varepsilon \mu^T \log u - \varepsilon \nu^T \log v$$

which is an unconstrained optimization problem and can be solved using block coordinate descent as follows:

$$u^{t+1} = \arg\min_u \psi(u, v^t) = \frac{\mu}{\sum_j K_{ij} v_j^t \mu_i \nu_j}; \quad (t - \text{even})$$
$$v^{t+1} = \arg\min_v \psi(u^t, v) = \frac{\nu}{\sum_i K_{ij} u_i^t \mu_i \nu_j}; \quad (t - \text{odd})$$

In this regards, Sinkhorn algorithm is the block coordinate descent for the dual problem.

2.2 Convergence Results

Lemma 2.2.1

$$\max_i \alpha_i^t - \min_i \alpha_i^t \leq ||C||_\infty;$$
$$\max_i \beta_i^t - \min_i \beta_i^t \leq ||C||_\infty;$$
$$\max \alpha^* - \min \alpha^* \leq ||C||_\infty;$$
$$\max \beta^* - \min \beta^* \leq ||C||_\infty; \quad \forall\ C \geq 0$$

Lemma 2.2.2

$$0 \leq \psi(\alpha, \beta) - \psi(\alpha^*, \beta^*) \leq ||C||_\infty \Big(||P_1(\pi(\alpha, \beta) - \mu)||_1 + ||P_2(\pi(\alpha, \beta) - \nu)||_1 \Big)$$

where; $\pi_{ij} = K_{ij} \exp\left(\frac{\alpha_i}{\varepsilon} + \frac{\beta_i}{\varepsilon}\right) \mu_i \nu_j$

Lemma 2.2.2 suggests that the difference between the value attained by the Sinkhorn algorithm and the optimal value is non-negative and this quantity is upper bounded by a quantity that depends on the infinity norm of the cost matrix C and the sum of the L_1 norms of the differences between the marginal distributions of the joint distribution $\pi(\alpha, \beta)$ and the target marginal distributions μ and ν.

Denote; $\widetilde{\psi}(\alpha,\beta) = \psi(\alpha,\beta) - \psi(\alpha^*,\beta^*) \geq 0$. Now;

$$\begin{aligned}
\widetilde{\psi}(\alpha^t,\beta^t) - \widetilde{\psi}(\alpha^{t+1},\beta^{t+1}) &= \begin{cases} \langle u\ , \alpha^{t+1} - \alpha^t \rangle = \varepsilon H\Big(\mu | P_1(\pi^t)\Big) & \text{when } t \text{ is odd} \\ \langle v\ , \beta^{t+1} - \beta^t \rangle = \varepsilon H\Big(\nu | P_2(\pi^t)\Big) & \text{when } t \text{ is even} \end{cases} \\
&\geq \begin{cases} \varepsilon/2 ||P_1(\pi^t) - \mu||_1^2 \\ \varepsilon/2 ||P_2(\pi^t) - \nu||_1^2 \quad \text{where } \pi^t = \pi(\alpha^t,\beta^t) \end{cases} \\
&= \frac{\varepsilon}{2}\Big(||P_1(\pi^t) - \mu||_1 + ||P_2(\pi^t) - \nu||_1\Big)^2 \\
&\geq \frac{\varepsilon}{2||C||_\infty^2}\ \widetilde{\psi}(\alpha^t,\beta^t)^2
\end{aligned}$$

This result implies that the Sinkhorn algorithm converges faster when the computed joint distribution $\pi(\alpha,\beta)$ at each iteration is closer to the set of joint probability distributions with given marginals μ and ν. The rate of convergence depends on the step size ε and the infinity norm of the cost matrix C ($||C||_\infty$).