**Math 612: Single Cell Analysis** **2019W Term 1**

Lecture 13: October 17

*Lecturer: Geoffrey Schiebinger* *Scribe: Delbert Yip*

## 13.1 Sinkhorn's Algorithm for Entropic Optimal Transport

**Theorem 13.1** *The solution to entropic optimal transport,*

$$\begin{aligned} \min_{\pi} \quad & \sum_{ij} \pi_{ij} c_{ij} + \epsilon \sum_{ij} \log \pi_{ij} \\ s.t. \quad & \pi \mathbb{1} = a \\ & \pi^T \mathbb{1} = b \end{aligned}$$

*is unique and has the form:*

$$\pi^*_{ij} = u_i K_{ij} v_j \tag{13.1}$$

*for scaling vectors* $u \in \mathbb{R}^n, v \in \mathbb{R}^m$*, where* $K$ *is a matrix with elements* $K_{ij} = \exp\left(-\frac{c_{ij}}{\epsilon}\right)$*. This means:*

$$\pi^* = Diag(u) K Diag(v). \tag{13.2}$$

**Proof:** The Lagrangian is the function:

$$\begin{aligned} L(\pi, \nu^{(1)}, \nu^{(2)}) = & \sum_{ij} \pi_{ij} c_{ij} + \epsilon \sum_{ij} \pi_{ij} \log \pi_{ij} \\ & - {\nu^{(1)}}^T (\pi \mathbb{1} - a) - {\nu^{(2)}}^T (\pi^T \mathbb{1} - b) - \epsilon (\sum_{ij} \pi_{ij} - 1). \end{aligned} \tag{13.3}$$

At optimality,

$$\frac{\partial L}{\partial \pi_{ij}} = 0 \tag{13.4}$$

$$= c_{ij} + \epsilon (\log \pi_{ij} + 1) - \nu^{(1)}_i - \nu^{(2)}_j - \epsilon$$

$$= c_{ij} + \epsilon \log \pi_{ij} - \nu^{(1)}_i - \nu^{(2)}_j \tag{13.5}$$

$$\implies \log \pi^*_{ij} = \frac{-c_{ij} + \nu^{(1)}_i + \nu^{(2)}_j}{\epsilon} \tag{13.6}$$

$$\begin{aligned} \pi^*_{ij} &= \exp\left( \frac{-c_{ij} + \nu^{(1)}_i + \nu^{(2)}_j}{\epsilon} \right) \\ &= \exp\left( \frac{\nu^{(1)}_i}{\epsilon} \right) \exp\left( -\frac{c_{ij}}{\epsilon} \right) \exp\left( \frac{\nu^{(2)}_j}{\epsilon} \right). \end{aligned} \tag{13.7}$$

■

### 13.1.1 Scaling Algorithm

Now, we solve for the vectors

$$u = \exp\left(\frac{\nu_i^{(1)}}{\epsilon}\right)$$
$$v = \exp\left(\frac{\nu_j^{(2)}}{\epsilon}\right)$$

$$\text{s.t.} \quad [\text{Diag}(u)K\text{Diag}(v)]\,\mathbb{1} = a = \pi\mathbb{1}$$
$$\left[\text{Diag}(v)K^T\text{Diag}(u)\right]\mathbb{1} = b = \pi^T\mathbb{1}$$

Note that the product of a diagonal matrix $D$ and $\mathbb{1}$ yields a column vector containing the diagonal elements of $D$. For example,

$$\text{Diag}(u) = \mathbb{1} = u = (u_1, u_2, \dots, u_n)^T$$

Thus, we get the following system of equations, with two equations and two unknowns:

$$\text{Diag}(u)Kv = a \tag{13.8}$$
$$\text{Diag}(v)K^Tu = b. \tag{13.9}$$

To solve this system,

1. Initialize $v^{(0)} = (1, 1, \dots, 1)^T$.

2. Update $u^{(1)}$ so that equation (13.8) holds.

3. Update $v^{(1)}$ so that equation (13.9) holds.

At the $(l+1)^{\text{th}}$ iteration, we would have used $v^{(l)}$ to update $u^{(l)}$, giving us $u^{(l+1)}$, which would then be used to update $v^{(l)}$, yielding $v^{(l+1)}$. This is shown below:

$$u^{(l+1)} = \frac{a}{Kv^{(l)}} \quad \rightarrow \quad v^{(l+1)} = \frac{b}{K^Tu^{(l+1)}}. \tag{13.10}$$

For the $n^{\text{th}}$ iteration, we would have:

$$\text{Diag}(u^{(n)}) \exp\left(-\frac{c_{ij}}{\epsilon}\right) \text{Diag}(v^{(n)}). \tag{13.11}$$

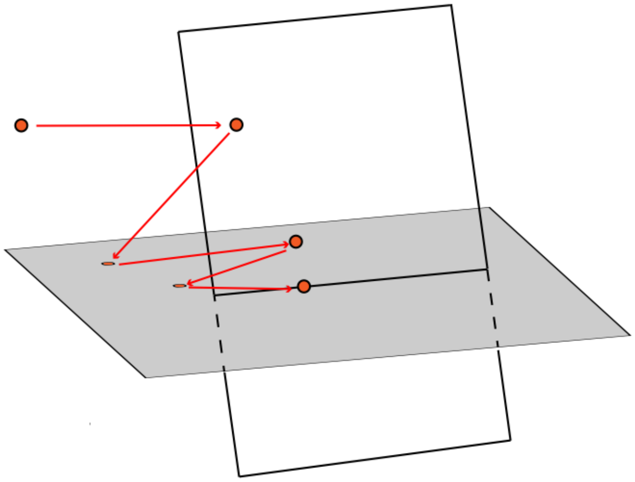

Figure 13.1: The Sinkhorn algorithm.
Starting from some initial values for $u$ and $v$, we alternatingly project between the gray rectangle, representing the space of all matrices with row sums equal to $a$, and the white rectangle, representing the space of all matrices with column sums equal to $b$. The algorithm eventually converges at the intersection of the rectangles, representing the set of matrices with row sums equal to $a$ and column sums equal to $b$. This would be the optimal solution.

**Theorem 13.2** *Let $C$ and $D$ be closed convex sets. Let $P_D$ and $P_C$ denote projections onto $D$ and $C$, respectively. Then, alternatingly projecting onto $C$ and $D$ will converge to a point at the intersection $C \cap D$.*

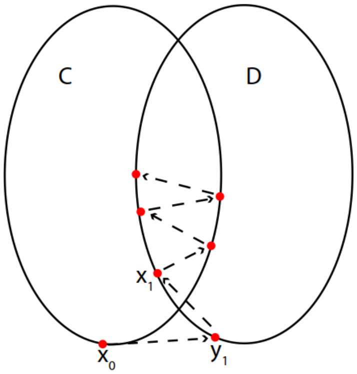

Figure 13.2: Converging in the intersection of two convex sets.
Starting from some $x_0 \in C$, we project onto $D$ to get $y_1 = P_D(x_0)$, and vice versa to get $x_1 = P_C(y_1)$.
At the $k^{\text{th}}$ iteration, $y_k = P_D(x_K)$, and $x_{k+1} = P_C(y_k)$.

**Proof:** Let $\overline{x}$ be any point in the intersection $C \cap D$. We claim each projection brings us closer to $\overline{x}$. To show this,

$$\begin{aligned}\|x_k - \overline{x}\|^2 &= \|(x_k - y_k) + (y_k - \overline{x})\|^2 \\ &= \|x_k - y_k\|^2 + \|y_k - \overline{x}\|^2 + 2(x_k - y_k)^T (y_k - \overline{x}). \end{aligned} \tag{13.12}$$

The third term in equation (13.12) is greater than or equal to zero, because $x_k - y_k$ and $y_k - \overline{x}$ are in the same direction (Figure 13.3).

Since C and D are closed, $\|x_k - \overline{x}\|^2$ will converge to the limit of both C and D, and will therefore end up in the intersection of C and D. ■

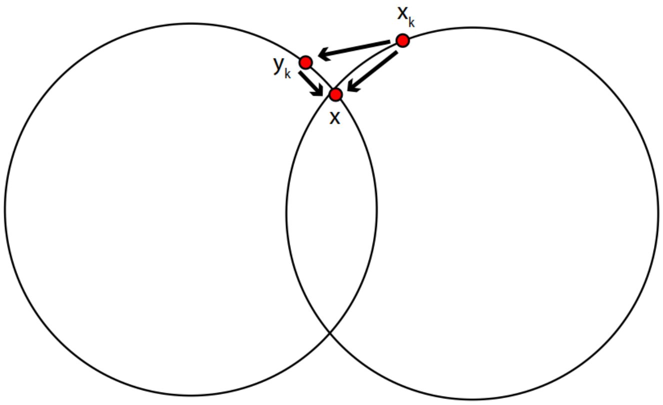

Figure 13.3: Projections between $x_i \in C$ and $y_i \in D$ eventually reach the intersection $C \cap D$.

### 13.1.2 Development as a Curve

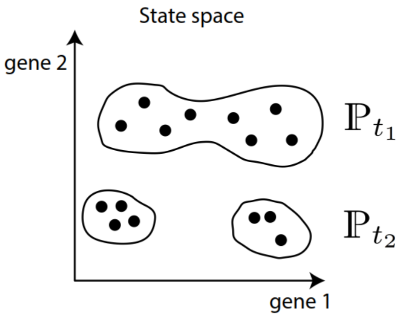

Figure 13.4: Cells as points in state space.

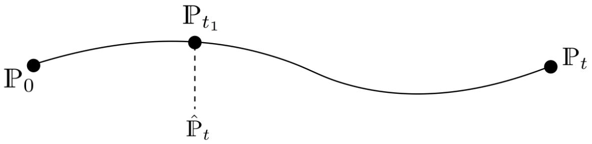

Figure 13.5: Distributions as points.

In state space, a cell is a point $x$, whereas an organism is a distribution $\mathbb{P}$. Over time, there is a stochastic process $\mathbb{P}_t$.

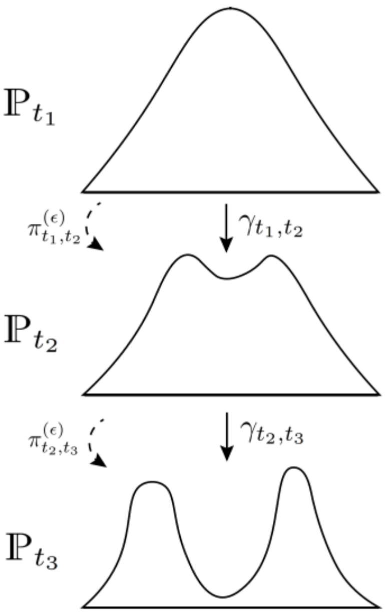

Figure 13.6: Couplings and distributions across time in state space.
From the distribution $\mathbb{P}_{t_1}$, we can sample cells $x_1, x_2, \ldots, x_{n_1} \sim \mathbb{P}_{t_1}$, yielding an empirical distribution $\hat{\mathbb{P}}_{t_1} = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$. We can do the same at times $t_2$ and $t_3$ . We are interested in inferring the couplings $\gamma_{t_k,t_{k+1}}$, which specify ancestor-descendant relationships. We can perform an approximation using entropy-regularized optimal transport, which gives couplings $\pi^{(\epsilon)}_{t_k,t_{k+1}}$.

In the next part of the course, we will learn to think about developmental curves with the optimal transport metric (Figure ). The space of distributions is known as ”Wasserstein space.”

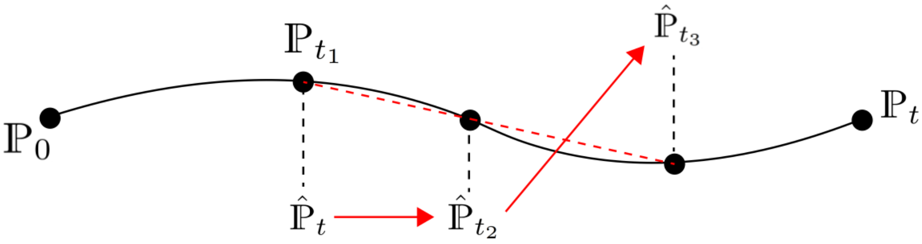

Figure 13.7: Developmental process in Wasserstein space.
Optimal transport gives linear trajectories (red). The dashed line represents the trajectory given the actual distributions $\mathbb{P}_t$, whereas the red arrows show the trajectories between empirical distributions $\hat{\mathbb{P}}_t$.

To infer the coupling $\gamma_{k,k+1}$, we assume that, on short time scales, $\gamma_{t_1,t_2}$ can be approximated by $\pi^{(\epsilon)}_{t_1,t_2}$.

$$\gamma_{t_1,t_2} \sim \pi^{(2)}_{t_1,t_2} \leftarrow \min_{\pi} \sum_{ij} \pi_{ij} c_{ij} + \epsilon \sum_{ij} \pi_{ij} \log \pi_{ij}$$

$$\begin{aligned} \min_{\pi} & \int \pi(x,y)c(x,y) + \epsilon \int \pi(x,y) \log \pi(x,y) \\ \text{s.t.} & \int_x \pi(x,y) = \mathbb{P}_{t_2(y)} \\ & \int_y \pi(x,y) = \mathbb{P}_{t_1}(x) g(x)^{t_2-t_1} . \hat{\mathbb{P}}_{t_2}, \hat{\mathbb{P}}_{t_3} \end{aligned} \tag{13.13}$$

The above applies when the full distributions $\mathbb{P}_t$ are available. In reality, we have empirical distributions $\hat{\mathbb{P}}_t$, and the integrals in equation (13.13) and its constraints would be replaced by sums.

To summarize, the three reasons for applying entropy-regularized optimal transport are:

1. Biologically, we can avoid limit points, or extremities, of the solution space that would otherwise be prescribed by a linear program.

2. Computationally, there is an efficient algorithm that finds the dual vectors by projecting alternatingly between closed, convex sets.

3. Statistically, we can avoid overfitting to a single sample's data, which may allow us to better approximate the true behaviour of the system.