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

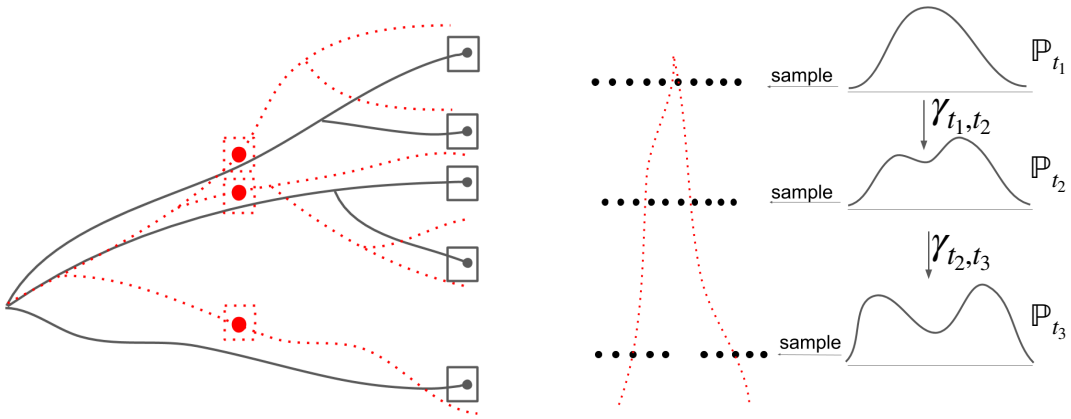
Lecture 9: October 3

Lecturer: Geoffrey Schiebinger *Scribe: Qiong Zhang*

Inferring developmental couplings with growth

Definition 9.1 (Developmental coupling) *γ_{t_1,t_2} is the joint distribution of a random cell $X_{t_2} \sim \mathbb{P}_{t_2}$ at time t_2 and its ancestor $A_{t_1} = Ancest(X_{t_2}) \sim \mathbb{Q}_{t_1} \propto \mathbb{P}_{t_1} g^{t_2 - t_1}$ where g is called the growth rate.*

Intuition for the growth function g: We model the growth rate according to a *birth-death process* with some birth rate $\beta(x)$ and death rate $\delta(x)$. At a "birth event", the mass in a small volume dx increases by 1; at a "death event", path stops and mass decreases. That is, $g(x) = \exp\{\beta(x) - \delta(x)\}$.



(a) Realizations of cell growth in the embedded space (b) Developmental coupling and Waddington's landscape

Figure 9.1: Illustration of developmental process.

In the actual inference, we sample from the developmental process $\mathbb{P}_t$ independently at time $t_1, t_2, \cdots$. For example, Figure 9.1a shows two realizations of cell growth represented by the red dotted line and black solid line in the embedded space. The red points represent for the random cell samples at time t_1 and black points represent for the random cell samples at time t_2. Our goal is to infer the coupling γ_{t_1,t_2} and the transition kernel $\gamma_{t_2|t_1}$ where

$$\gamma_{t_2|t_1}(\cdot|a) \sim \{X_{t_2}|A_{t_1} = a\}$$

and

$$\gamma_{t_1,t_2}(da, dx) = \mathbb{Q}_{t_1}(da)\gamma_{t_2|t_1}(dx|da).$$

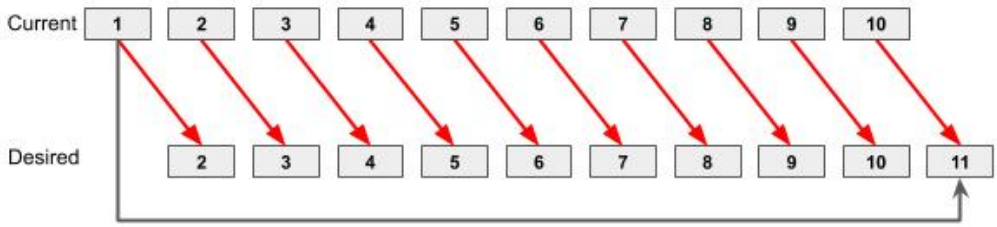


Figure 9.2: Two strategies in red and black respectively for moving the boxes from current to desired location.

Figure 9.1b illustrate the connection of developmental coupling and the recovery of Waddington's landscape. We sample from population $\mathbb{P}_t$ at time $t_1, t_2, t_3, \cdots$. The estimated developmental coupling based on these samples are given in red dotted line. We hope the red dotted line can recovery the Waddington's landscape. Traditional approaches for the recovery of Waddington's landscape includes Monocle, URD, AGA, slingshot, etc which involves non-convex optimization. The non-convex optimization problems are difficult in general as they may have multiple locally optimal points. We instead consider the recovery of Waddington's landscape with convex optimization.

Inferring developmental couplings with convex optimization

The inference of developmental couplings in this course is mainly based on the optimal transportation theory. Optimal transport (OT) has a long history starting from 1781 when the French mathematician and physicist Gaspard Monge (1746-1818) formulated the original problem. He considered the problem of how to move a pile of sand to fill up a hole with minimal cost. The assignment of the sand from the original location to the final location is called the "transport plan" or "transportation plan". The terminology "optimal transport" comes from the fact that we are looking for the transportation plan with minimum cost. The modern definition of optimal transport is the reformulation of Monge's problem by the Russian economist Leonid Kantorovich in 1947.

Definition 9.2 (Optimal transport) *Let $\mathbb{P}$ and $\mathbb{Q}$ be two probability measures defined on space $\mathcal{X}$, let $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+ \cup \{+\infty\}$ be a cost function. The optimal transport problem look for a coupling π^* so that*

$$\pi^* = \arg\min_{\pi} \mathbb{E}_\pi\{c(X, Y)\}$$

under constraints

$$(X, Y) \sim \pi,\ X \sim \mathbb{P},\ Y \sim \mathbb{Q}.$$

Example 9.3 (Optimal transportation plan and cost function) *Let us consider the following example where ten identical boxes are at current location $1, 2, \cdots, 10$ needs to be moved to desired location at $2, 3, \cdots, 11$. There are two proposed strategies as illustrated in Figure 9.2 to move these boxes: the strategy in black that keeps box at location 2 to 10 and move box at location 1 to location 11; the strategy in red moves all the boxes to the location on its right. The question is which strategy is better? Indeed, the optimal strategy depends on the choice of the function. Let $c(x, y)$ be the cost for moving unit mass from x to y.*

- *If $c(x, y) = |x - y|^2$, the total cost for red strategy is 10×1^2 while the total cost for the black strategy is 1×10^2, therefore, red strategy is better.*

- *If $c(x,y) = |x-y|$, the total cost for both strategy is 10, then these two strategies are equally good.*
- *If $c(x,y) = \sqrt{|x-y|}$, then the black strategy is better.*

Example 9.4 (Deliver goods) *There are 3 bakeries at position x_1, x_2, x_3 and 3 restaurants at position y_1, y_2, y_3. The 3 bakeries respectively produce $b_1 = 10$, $b_2 = 5$, and $b_3 = 3$ loaves. The 3 restaurants respectively consumes $r_1 = 6$, $r_2 = 6$, and $r_3 = 6$ loaves.*

***Question**: If it costs $c(x,y)$ to move a loaf from bakery at location x to a restaurant at location y, what is the optimal way to transport the goods?*

In this case, the transport plan is a matrix $\pi = (\pi_{ij}) \in \mathbb{R}^{3\times 3}_{\geq 0}$ with π_{ij} represents the number of loaves that is transported from bakery at location x_i to restaurant at location y_j. Let us assume that the loaves are allowed to break into pieces. Then the optimal transportation plan π^* is

$$\pi^* = \arg\min_{\pi} \sum_{i=1}^{3}\sum_{j=1}^{3} \pi_{ij} c(x_i, y_j)$$

subject to

$$\begin{cases} \sum_{j=}^{3} \pi_{ij} = b_i, & \forall i = 1,2,3 \\ \sum_{i=}^{3} \pi_{ij} = r_j, & \forall j = 1,2,3 \end{cases}$$

The optimization problem above is called a "linear program".

The estimation of developmental coupling for developmental process can also be formulated as a optimal transportation problem. We will connect $\hat{\mathbb{Q}}_{t_1} \propto \hat{\mathbb{P}}_{t_1} g^{t_2-t_1}$ where $\hat{\mathbb{Q}}_{t_1} = N_1^{-1} \sum_{i=1}^{N_1} \frac{\delta_{x_i} g^{t_2-t_1}(x_i)}{N_1^{-1}\sum_i g^{t_2-t_1}(x_i)}$ to $\hat{\mathbb{P}}_{t_2} = N_2^{-1}\sum_i \delta_{y_i}$ with an optimal transport problem:

$$\hat{\pi} = \arg\min_{\pi} \mathbb{E}_{\pi}\{c(X,Y)\}$$

such that $(X,Y) \sim \pi$, $X \sim \hat{\mathbb{Q}}_{t_1}$, and $Y \sim \hat{\mathbb{P}}_{t_2}$. Then $\hat{\pi}$ will be an estimate for γ_{t_1,t_2}.