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

Lecture 12: October 15

Lecturer: Geoffrey Schiebinger *Scribe: Delbert Yip*

12.1 Entropy Regularized Optimal Transport

Definition 12.1 *The Shannon entropy of a distribution on a finite space* $\left(\sum_{i=1}^{k} p_i = 1\right)$ *is:*

$$-\sum_{i=1} p_i \log p_i. \tag{12.1}$$

For a coupling π, the entropy is:

$$\begin{aligned} \min_{\pi} \quad & \sum_{ij} \pi_{ij} c_{ij} \\ \text{s.t.} \quad & \boldsymbol{\pi} \mathbb{1} = r, \\ & \boldsymbol{\pi}^{\boldsymbol{T}} \mathbb{1} = c. \end{aligned} \tag{12.2}$$

In entropy-regularized optimal transport (EROT), the entropy is:

$$\begin{aligned} \min_{\pi} \quad & \sum_{ij} \pi_{ij} c_{ij} + \epsilon \sum_{ij} \pi_{ij} \log \pi_{ij} \\ \text{s.t.} \quad & \boldsymbol{\pi} \mathbb{1} = r, \\ & \boldsymbol{\pi}^{\boldsymbol{T}} \mathbb{1} = c. \end{aligned} \tag{12.3}$$

Where, $\epsilon \geq 0$ is a regularization parameter.

For a developmental stochastic process in which we sample two distributions at timepoints t_1 and t_2, we infer the coupling $\gamma_{t_1 t_2}$ with optimal transport, giving $\pi_{t_1 t_2}$ with constraints:

$$\begin{cases} \sum_{y_i} \pi(x, y_i) = \hat{\mathbb{P}}_{t_1}(x) g(x)^{t_2 - t_1} \\ \sum_{x_i} \pi(x_i, y) = \hat{\mathbb{P}}_{t_2}(y). \end{cases}$$

Where, c_{ij} is a cost function.

12.1.1 Entropy

Definition 12.2 ***Entropy***
For a "thermodynamic system," entropy is:

$$\log\left(\textit{number of microstates for a given macrostate}\right). \tag{12.4}$$

12.1.2 Gas in a box

Consider gas in a box that is discretized into k volume elements (i.e. small cubes). For molecules $x_1 \dots x_n$, the microstate would correspond to the box each molecule is in, whereas the macrostate would be the number of molecules in each box.

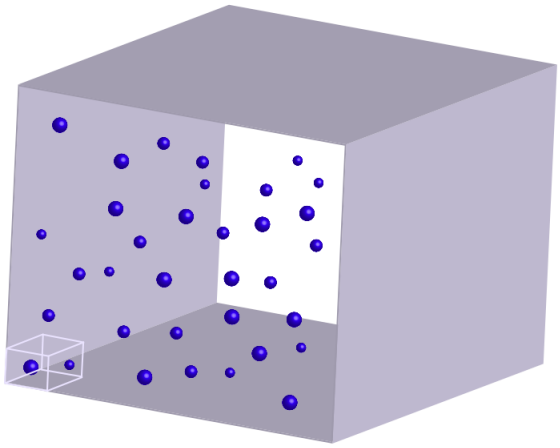

Figure 12.1: Gas particles enclosed in a box. An example of a discretized volume element is shown in the lower-left corner (light grey).

For N_i molecules in a box, there are $\frac{N!}{N_1!N_2!\dots N_k!}$ ways N_i molecules can find themselves in this box if there are k boxes. From this, we can derive an expression for entropy:

$$
\begin{aligned}
\log \frac{N!}{N_1!N_2!\dots N_k!} &= \log N! - \sum_{i=1}^{k} \log N_i! \\
&\approx (N \log N - N) - \left(\sum_{i=1}^{k} N_i \log N_i - \sum_{i=1}^{k} N_i \right) \quad \text{by Sterling's approximation.} \\
&\approx N \log N - \sum_{i=1}^{k} N_i \log N_i \\
&\approx N \log N - \sum_{i=1}^{k} (Np_i) \log(Np_i) \quad \text{where we substitute } N_i = Np_i. \\
&\approx N \log N - N \sum_{i=1}^{k} p_i (\log N + \log p_i) \\
&\approx N \log N - N \log N \sum_{i=1}^{k} p_i - N \sum_{i=1}^{k} p_i \log p_i \\
&\approx -N \sum_{i=1}^{k} p_i \log p_i.
\end{aligned}
$$

12.1.3 Relative Entropy and the Second Law

Definition 12.3 ***Relative entropy.***
Let $\mathbb{P}$ and $\mathbb{Q}$ be distributions on a finite state space (i.e. $\sum_{i=1}^{k} p_i = 1, \sum_{i=1}^{k} q_i = 1$). Then, the relative entropy is:

$$D(\mathbb{P}\|\mathbb{Q}) = \sum_{i=1}^{k} p_i \log \frac{p_i}{q_i}. \tag{12.5}$$

Fact: $D(\mathbb{P}\|\mathbb{Q}) \geq 0$ with equality if and only if $p = q$.

Theorem 12.4 ***The second law of thermodynamics.***
Consider a Markov chain on a finite state space. For example, the number of molecules in k volume elements of a box. The states are the number of molecules in each box: $N_1, N_2, \ldots N_k$. Let μ_0 and μ_0' be initial distributions. Then, we have that $D(\mu_0\|\mu_0')$ decreases monotically with N.

Proof:
Conditional relative entropy is defined as:

$$D(p(y \mid x)\|q(y \mid x)) = \sum_{x} p(x) \sum_{y} p(y \mid x) \log \frac{p(y \mid x)}{q(y \mid x)}. \tag{12.6}$$

Fact: (Chain rule)

$$D(p(x,y)\|q(x,y)) = D(p(x)\|q(x)) + D(p(y \mid x)\|q(y \mid x)). \tag{12.7}$$

We will show that $D(\mu_{n+1}\|\mu_{n+1}')$ is monotically decreasing.

Let $p(x_n, x_{n+1})$ denote the joint distribution for μ_n, μ_{n+1}. If the transition kernel is γ, we have:

$$p(x_n, x_{n+1}) = \mu_n(x_n)\gamma(x_{n+1} \mid x_n) \tag{12.8}$$
$$q(x_n, x_{n+1}) = \mu_n'(x_n)\gamma(x_{n+1} \mid x_n). \tag{12.9}$$

By the chain rule, we have:

$$D(p(x_n, x_{n+1})\|q(x_n, x_{n+1})) = D(p(x_n)\|q(x_n)) + D(p(x_{n+1})\|q(x_{n+1} \mid x_n)) \tag{12.10}$$
$$= D(p(x_{n+1})\|q(x_{n+1})) + D(p(x_n \mid x_{n+1})\|q(x_n \mid x_{n+1})). \tag{12.11}$$

In equation (12.11), note the following:

- $p(x_{n+1}) = \mu_{n+1}$
- $q(x_{n+1}) = \mu_{n+1}'$
- $p(x_n \mid x_{n+1}) \geq 0$
- $q(x_n \mid x_{n+1}) \geq 0$

Therefore, we have:

$$D(\mu_n\|\mu_n') = D(\mu_{n+1}\|\mu_{n+1}') + \text{something positive}. \tag{12.12}$$

Since the right-hand side of equation (12.12) is greater than its left-hand side, this means that the state distributions μ_n and μ_n' will get closer together with time. ■

Corollary 12.5 *If μ' is the uniform distirbution and is stationary for the Markov process, then the divergence between μ_n and μ'_n is always decreasing:*

$$D(\mu_n \| \mu'_0) = \sum_{i=1} p_i \log p_i. \tag{12.13}$$

The right-hand side of equation equation (12.13) *represents negative entropy. Entropy increases if we have μ'_0 as the uniform distribution and if μ'_0 is also the stationary distribution.*

Example: Low versus high entropy

1. Low entropy:
   Let $p_1 = 1$, and $p_i = 0$ for all $i \neq 1, i \leq k$.
   This situation is deterministic. For the entropy, we have:

$$-\sum_{i=1}^{k} p_i \log p_i = 0.$$

2. High entropy:
   Let $p_i = \frac{1}{k}$ for all $i \in [1, k]$. Then, the entropy is:

$$-\sum_{i=1}^{k} \frac{1}{k} \log \frac{1}{k} = -\log \frac{1}{k} = \log k.$$

   As $k \to \infty$, so does the entropy.

12.1.4 Developmental Stochastic Processes

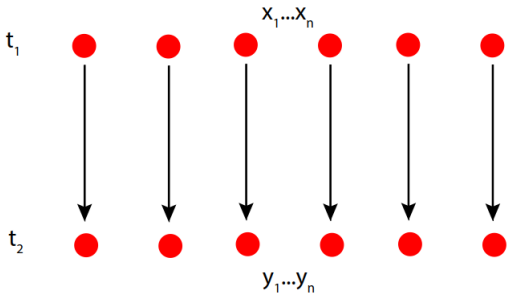

Figure 12.2: Cells x_i at time t_1 give rise to a descendant distribution y_i at time t_2.

A developmental stochastic process is depicted in Figure 12.2, where a population of ancestral cells x_i, observed at time t_1, give rise to descendants y_i, sampled at a later time t_2. For a given cell x_i, the entropy of the descendant distribution specifies how 'fated' the cell is. For instance, if a cell is equally fated for all lineages - that is, it could become any y_i - this is a case of high entropy.

12.1.4.1 Optimal transport with or without entropy regularization

Without entropy regulation, optimal transport is a linear program. The solution to a linear program is attained at an extreme point of the constraint set (circled areas in Figure 12.3). For couplings, the extreme points have zero entropy. When we apply entropy, we have level sets in a third dimension, which ensures that we avoid the extreme points (Figure 12.4).

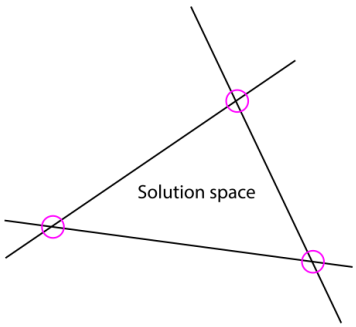

Figure 12.3: Optimal transport without entropy.

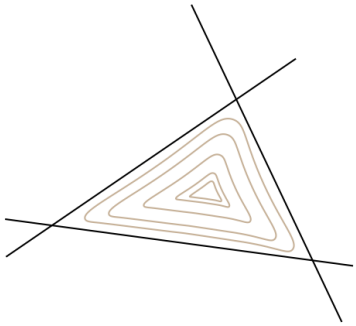

Figure 12.4: Optimal transport with entropy regularization.