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

Lecture 5: September 19

Lecturer: Geoffrey Schiebinger *Scribe: Naba Mukhtar*

5.1 Introduction

This is the first lecture on probability. Our goal is to understand mathematically how a sample of cells is representative of the whole population.

Definition 5.1 *A* ***measure*** *assigns a mass to subsets of a space* $\boldsymbol{X}$*. For a subset* $A \subset \boldsymbol{X}$*,* $\mu(A)$ *is the measure of* A*; where* μ *satisfies all of the following properties:*

1. μ *is non-negative* ($\mu(A) \geq 0, \forall A \in \boldsymbol{X}$)*;*

2. μ *is additive* ($\mu(A \cup B) = \mu(A) + \mu(B), \forall A, B \in \boldsymbol{X}$)*;*

3. $\mu(\emptyset) = 0$*.*

Definition 5.2 *A* ***probability measure****, or distribution,* $\mathbb{P}$*, assigns mass 1 to the whole space; that is,* $\mathbb{P}(\boldsymbol{X}) = 1$*.*

A probability measure $\mathbb{P}$ tells us $Prob\{X \in A\} = \mathbb{P}(A)$.

Definition 5.3 *A* ***random variable****,* X*, is a random element of* $\boldsymbol{X}$*.*

Example 5.4 *Coin flip:* $H =$ *heads,* $T =$ *tails*

$\boldsymbol{X} = \{H, T\}, \mathbb{P}(H) = \frac{1}{2}, \mathbb{P}(T) = \frac{1}{2}$

Random variable: flip (F)

$Prob\{F = H\} = \mathbb{P}(H) = \frac{1}{2}$.

Example 5.5 *A Gaussian (Normal) random variable has distribution* $\mathbb{P}$ *with density* $p(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}$ *on* $\boldsymbol{X} = \mathbb{R}$*.*

$Prob\{X \in A\} = \int_A p(x)dx = \mathbb{P}(A)$.

Example 5.6 *A pair of random variables (X, Y) is a random variable.*

Definition 5.7 *A distribution for a pair of random variables is called a* ***joint distribution****.*

Definition 5.8 *A pair of random variables are* ***independent*** *if the realization of one does not affect the realization of the other; that is,* $p(x, y) = p(x)p(y)$*.*

Put another way, $Prob\{X \in A \text{ and } Y \in B\} = Prob\{X \in A\}Prob\{Y \in B\}$*.*

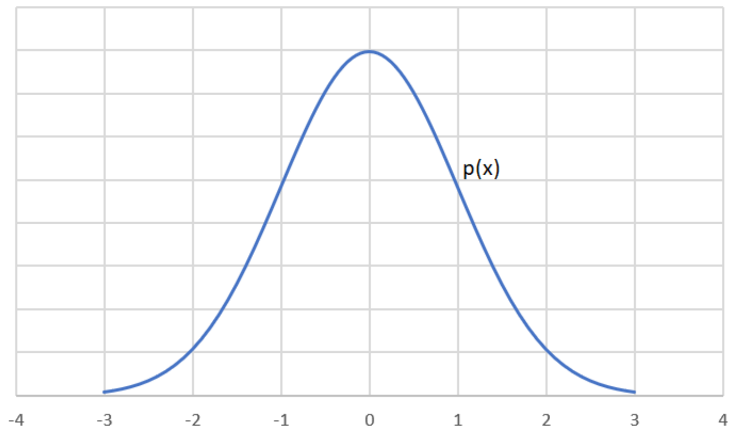

Figure 5.1: Density function *p(x)* for Gaussian distribution as in Example 5.5

5.2 Application to Cell Sampling

Consider a population of cells represented by $\mathbb{P}$, a probability distribution on the space of *all cell states*. For now assume RNA completely encodes cell state. This implies $\mathbb{P}$ is a probability distribution on the gene expression space, $\mathbf{X}$.

Sample $X_1, X_2, ..., X_n \sim \mathbb{P}$, where $\mathbb{P}$ is on the order of $\mathbb{R}^{20,000}$. Then for each integer $i = 1, 2, ..., 20000$, we have

$$X_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{i20,000} \end{bmatrix}, \; X_i \in \mathbb{R}^{20,000}, \; Prob\{X_i \in A\} = \mathbb{P}(A), \tag{5.1}$$

where A is a set of expression vectors.

Example 5.9 *Suppose you have a plate with 50% cell type 1 and 50% cell type 2. Then X_1 is like the coin-flip example (Example 5.4).*

Example 5.10 *Dirac δ-function is a probability measure with no density. δ_x assigns mass 1 to any set containing the point x. δ_x is a "point mass".*

Definition 5.11 *The* ***empirical measure*** *for samples $X_1, ..., X_n$ is:*

$\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$.

$\mathbb{P}_n$ is a probability measure.

Proof: $\mathbb{P}_n(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} 1 = \frac{1}{n}n = 1$. ∎

How similar is $\mathbb{P}_n$ to $\mathbb{P}$?

Definition 5.12 *The* ***expected value*** *of a random variable with distribution* $\mathbb{P}$ *is:*

$\mathbb{E}[X] = \int x d\mathbb{P}(x) = \int x p(x) dx,$

where the second equality applies if and only if $\mathbb{P}$ *has density.*

Fact 1: $\int f(y)\delta_x(y)dy \, (= \int f(y) d\delta_x(y)) = f(x).$

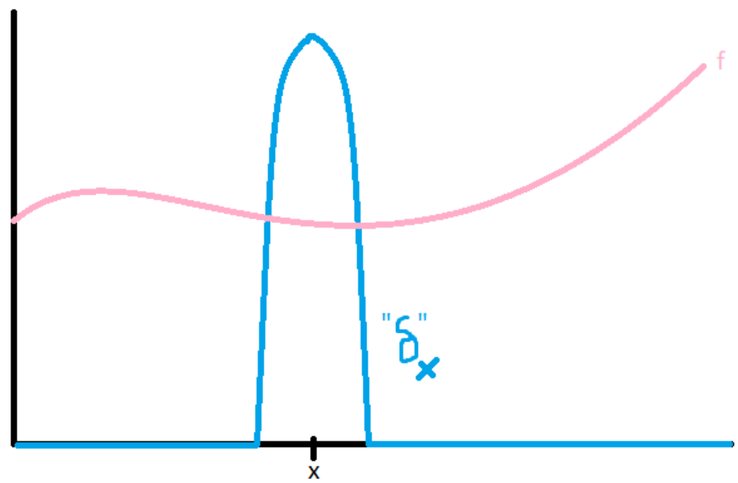

Figure 5.2: Illustration of **Fact 1**. f is an arbitrary function, "δ_x" is a function whose integral over the real line is 1. As the peak of "δ_x" gets narrower, we see that "δ_x" $\to \delta_x$, and $\int f(y)$"δ_x"$(y)dy \to f(x)$.

Definition 5.13 *The* ***mean*** *of* $\mathbb{P}_n$ *is* $\frac{1}{n}\sum_{i=1}^n X_i$.

Fact: $\int x d\mathbb{P}_n(x) = \int x d(\frac{1}{n}\sum_{i=1}^n \delta_{X_i})$.

The Law of Large Numbers says that $\int x d\mathbb{P}_n(x) \to \int x d\mathbb{P}$ as $n \to \infty$ and also $\int f(x) d\mathbb{P}_n(x) \to \int f(x) d\mathbb{P}(x)$.

5.3 Dimensionality Reduction

We can linearize the problem using Principal Component Analysis (PCA). Given samples $X_1, ..., X_n \sim \mathbb{P}$, PCA identifies the linear subspace of $\mathbf{X} = \mathbb{R}^{20,000}$, where the data is spread out.

We construct a gene expression matrix (this is an $n \times 20000$ random matrix):

$\chi = [X_1\, X_2 \dots X_n]$ (note that each X_i is a 20,000-component vector).

Then we define the (random) matrix $S^{(n)} = \frac{1}{n}\chi\chi^T$. That is,

$$S^{(n)} = [X_1\, X_2 \dots X_n] \begin{bmatrix} X_1^T \\ X_2^T \\ \vdots \\ X_n^T \end{bmatrix} = \frac{1}{n}\sum_{i=1}^n X_i X_i^T = \int x x^T d\mathbb{P}_n(x) = \int x x^T d(\frac{1}{n}\sum_{i=1}^n \delta_{X_i}). \tag{5.2}$$

Note that, since χ is an $n \times 20000$ matrix, then χ^T is a $20000 \times n$ matrix, so $S^{(n)}$ is a 20000×20000 matrix. Also, as $n \to \infty$, $S^{(n)} \to S$, where we define

$S = \int xx^T d\mathbb{P}(x)$.

We compute the eigenvectors of $S^{(n)}$:

$S^{(n)}v = \lambda v$, so $S^{(n)} = \sum_{i=1}^{d} \lambda_i v_i v_i^T$,

where $\lambda_1 > \lambda_2 > ... > \lambda_d > 0$.

Since working in higher dimensions requires more computations, we project the data $X_1, ..., X_n$ onto

$span\{v_1, ..., v_k\}$ for some $k < d$ and define

$\widehat{S^{(n)}} = \sum_{i=1}^{k} \lambda_i v_i v_i^T$.

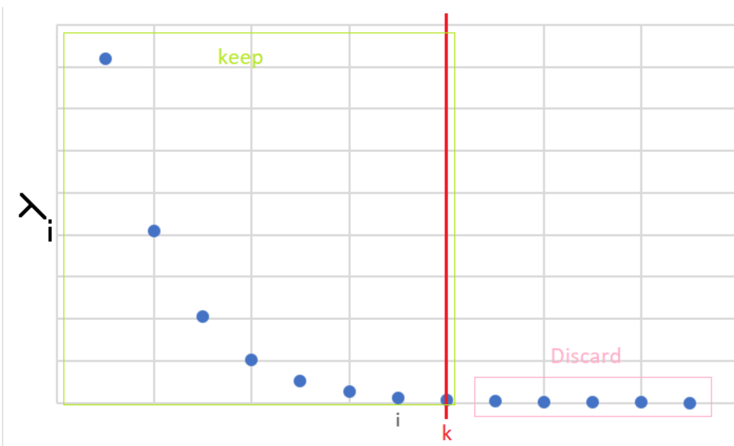

Figure 5.3: There are $d = 13$ eigenvalues and we keep the first $k = 8$ of them, projecting the data onto $span\{v_1, ..., v_8\}$. Note that each $\lambda_i > \lambda_{i+1} > 0$.

Example 5.14 *Suppose $k = 2$. Then each X_i is mapped to $(\lambda_1 v_1^T X_i, \lambda_2 v_2^T X_i)$.*