CSE 312 – Section 9

Spring 2026

Review of Main Concepts

Law of Total Expectation

- **Conditional Expectation**: Let X and Y be **discrete** random variables. Then, the conditional expectation of X given $Y = y$ is

$$\mathbb{E}[X|Y=y] = \sum_x x \cdot p_{X|Y}(x|y) = \sum_x x \cdot \mathbb{P}(X=x|Y=y).$$

  Note that linearity of expectation still applies to conditional expectation:

$$\mathbb{E}[X+Y|A] = \mathbb{E}[X|A] + \mathbb{E}[Y|A].$$

- **Discrete Law of Total Expectation (event version):** Let $A_1 \dots, A_n$ be a partition of the sample space. Then

$$\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X|A_i]\,\mathbb{P}(A_i).$$

- **Discrete Law of Total Expectation (r.v. version):** Let X and Y be two random variables. Then

$$\mathbb{E}[X] = \sum_{y \in \Omega_Y} \mathbb{E}[X|Y=y] \cdot \mathbb{P}(Y=y).$$

- **Continuous Law of Total Expectation:**

$$\mathbb{E}[X] = \int_{y \in \Omega_Y} \mathbb{E}[X|Y=y]\, f_Y(y)\, \mathrm{d}y$$

- **Expected value of X conditioned on r.v. Y:** Suppose that Y is a random variable that takes values $y_1, \dots, y_k$. Then $\mathbb{E}[X|Y]$ is the following random variable

$$\mathbb{E}[X|Y] = \begin{cases} \mathbb{E}[X|Y=y_1] & \text{with probability } \mathbb{P}(Y=y_1) \\ \mathbb{E}[X|Y=y_2] & \text{with probability } \mathbb{P}(Y=y_2) \\ \dots & \\ \mathbb{E}[X|Y=y_k] & \text{with probability } \mathbb{P}(Y=y_k) \end{cases}$$

- **Law of total expectation (rewritten):** Given the above definition, we can write

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]] = \sum_{i=1}^{k} \mathbb{E}[X|Y=y_i] \cdot \mathbb{P}(Y=y_i).$$

- **Covariance:** We may not get to this in class, but it's important to know about. To find out more, check out section 5.4 in the Tsun book. And now the definition: For any two random variables X, Y the *covariance* is defined as

$$\operatorname{Cov}(X,Y) = \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right].$$

  It can also be shown that

$$\operatorname{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y].$$

- **Conditional distributions:** We are not explicitly covering this topic in class, but it is **highly** recommended that you study it. Much of the above can be more appropriately rewritten in terms of conditional distributions. See Tsun, Section 5.3.

	Discrete	**Continuous**
Conditional PMF/PDF	$p_{X\|Y}(x\|y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}$	$f_{X\|Y}(x\|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$
Conditional Expectation	$\mathbb{E}[X\|Y=y] = \sum_x x p_{X\|Y}(x\|y)$	$\mathbb{E}[X\|Y=y] = \int_{-\infty}^{\infty} x f_{X\|Y}(x\|y)\,\mathrm{d}x$

Maximum Likelihood Estimation

We assume access to samples $X_1, X_2, \ldots, X_n$ from a distribution (e.g. Bernoulli, Geometric, Normal, etc) with unknown parameters. Our goal is to estimate this parameters. When the parameter is unknown we denote it by θ.

- **Realization/Sample**: A realization/sample x of a random variable X is the value that is actually observed.
- **Likelihood**: Let $X_1 = x_1, \ldots X_n = x_n$ be iid samples of a random variable with probability mass function $p_X(\mathrm{x}\,;\theta)$ (if X discrete) or density $f_X(\mathrm{x}\,;\theta)$ (if X continuous), where θ is a parameter (or a vector of parameters). We define the likelihood function to be the probability of seeing the data (in the discrete case). If X is discrete:

$$L(X_1 = x_1, \ldots X_n = x_n\,;\,\theta) = \prod_{i=1}^{n} p_X(x_i\,;\,\theta)$$

  If X is continuous:

$$L(X_1 = x_1, \ldots X_n = x_n\,;\,\theta) = \prod_{i=1}^{n} f_X(x_i\,;\,\theta)$$

- **Log-Likelihood**: We define the log-likelihood as the natural logarithm of the likelihood function. Since the logarithm is a strictly increasing function, the value of θ that maximizes the likelihood will be exactly the same as the value that maximizes the log-likelihood. If X is discrete:

$$\ln L(x_1, \ldots, x_n\,;\,\theta) = \sum_{i=1}^{n} \ln p_X(x_i\,;\,\theta)$$

  If X is continuous:

$$\ln L(x_1, \ldots, x_n\,;\,\theta) = \sum_{i=1}^{n} \ln f_X(x_i\,;\,\theta)$$

- **Maximum Likelihood Estimator (MLE)**: We denote the MLE of θ as $\hat{\theta}_{\text{MLE}}$ or simply $\hat{\theta}$, the parameter (or vector of parameters) that maximizes the likelihood function (probability of seeing the data) or equivalently, maximizes the log-likelihood. For a fixed set of realizations/samples $x_1, \dots, x_n$

$$\hat{\theta}_{\text{MLE}}(x_1, \dots, x_n) = \arg\max_{\theta} L(x_1, \dots, x_n\,;\,\theta) = \arg\max_{\theta} \ln L(x_1, \dots, x_n\,;\,\theta)$$

  Viewing the estimator $\hat{\theta}_{\text{MLE}}$ as a function of the random variables $X_1, \dots, X_n$, the estimator $\hat{\theta}_{\text{MLE}}(X_1, \dots, X_n)$ is itself a random variable.

- **Bias**: The bias of an estimator $\hat{\theta}$ for a true parameter θ is defined as

$$\text{Bias}(\hat{\theta}, \theta) = \mathbb{E}[\hat{\theta}(X_1, \dots, X_n)] - \theta.$$

  Here, $\hat{\theta}$ is treated as a *random variable*, computed from i.i.d. random variables $X_1, \dots . X_n$.

- **Unbiased Estimator**: An estimator $\hat{\theta}$ of θ is unbiased iff $\text{Bias}(\hat{\theta}, \theta) = 0$, or equivalently $\mathbb{E}[\hat{\theta}(X_1, \dots, X_n)] = \theta$.

- **Steps to find the maximum likelihood estimator $\hat{\theta}$:**

  a) Find the likelihood and log-likelihood of the data.

  b) Take the derivative of the log-likelihood

  c) Set it to 0 to find a candidate for the MLE, $\hat{\theta}$. (note: at this step, we change from the θ to the $\hat{\theta}$ because in this step we are solving for the *maximum* likelihood estimator for θ)

  d) Take the second derivative and show that $\hat{\theta}$ indeed is a maximizer, that $\frac{\partial^2 L}{\partial \theta^2} < 0$ at $\hat{\theta}$. Also ensure that it is the global maximizer: check points of non-differentiability and boundary values.

  Similar steps apply when the distribution has multiple parameters, e.g. a normal distribution with unknown mean and variance. In this case, maximizing the likelihood or log-likelihood involves maximizing a function of multiple variables (the different parameters). See, for example, problem 11 on this worksheet.

Tail bounds ++

- **Union Bound:** The union bound is a simple application of inclusion/exclusion and says that for any two events A and B, it holds that

$$\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B).$$

  The union bound is used to bound probabilities when, e.g. $\mathbb{P}(A \cap B)$ is difficult to compute.

  We may or may not get to cover the following tail bounds.

- **Markov's Inequality**: Let X be a non-negative random variable, and $\alpha > 0$. Then,

$$\mathbb{P}\left(X \geq \alpha\right) \leq \frac{\mathbb{E}\left[X\right]}{\alpha}$$

- **Chebyshev's Inequality**: Suppose Y is a random variable with $\mathbb{E}[Y] = \mu$ and $\text{Var}(Y) = \sigma^2$. Then, for any $\alpha > 0$,
$$\mathbb{P}(|Y - \mu| \geq \alpha) \leq \frac{\sigma^2}{\alpha^2}$$
- **(Multiplicative) Chernoff Bound**: Let $X_1, X_2, ..., X_n$ be *independent* Bernoulli random variables. Let $X = \Sigma_{i=1}^{n} X_i$, and $\mu = \mathbb{E}[X]$. Then, for any $0 \leq \delta \leq 1$,
  - $\mathbb{P}(X \geq (1+\delta)\mu) \leq e^{-\frac{\delta^2 \mu}{3}}$
  - $\mathbb{P}(X \leq (1-\delta)\mu) \leq e^{-\frac{\delta^2 \mu}{2}}$

Section plan

- Content review I
- Problem 7
- Problem 8
- Problem 11
- If time, Problem 12

1 Content Review I

a) **Law of Total Expectation.** Let X and Y be discrete random variables. Which of the following represents the Law of Total Expectation for X?

- ○ $\mathbb{E}[X] = \sum_y \mathbb{E}[X \mid Y = y] p_X(x)$
- ○ $\mathbb{E}[X] = \sum_y \mathbb{E}[X \mid Y = y] p_Y(y)$
- ○ $\mathbb{E}[X] = \sum_x \mathbb{E}[X \mid Y = y] p_Y(y)$
- ○ $\mathbb{E}[X] = \sum_y \mathbb{E}[X] p_{X|Y}(x|y)$

b) **Linearity of Conditional Expectation.** True or False: Linearity of expectation does **not** apply to conditional expectations. That is, $\mathbb{E}[X + Y \mid A]$ is generally NOT equal to $\mathbb{E}[X \mid A] + \mathbb{E}[Y \mid A]$.

- ○ True
- ○ False

c) Suppose X and Y are random variables and A is an event. Given that $\mathbb{E}[X|A] = 4$ and $\mathbb{E}[Y|A] = 10$, what is $\mathbb{E}\left[2X + \frac{Y}{2} \,\middle|\, A\right]$?

- ○ 14
- ○ 18
- ○ 9
- ○ 13

d) **True or False**: We get a different answer if we maximize log-likelihood rather than likelihood, but because it is close enough and easier to compute, we use it for our estimate of θ.

e) **True or False**: [Notation question] $\hat{\theta}$ is the true parameter and θ is our estimate.

f) **True or False**: An estimator is unbiased if $\text{Bias}(\hat{\theta}, \theta) = \mathbb{E}[\hat{\theta}(X_1, \ldots, X_n)] - \theta = 0$ or equivalently $\mathbb{E}[\hat{\theta}(X_1, \ldots, X_n)] = \theta$

g) You flip a coin 10 times and observe HHHTHHTHHH (8 heads, 2 tails). What is the MLE of θ, where θ is the true probability of seeing tails?

- ○ $\hat{\theta} = .2$
- ○ $\hat{\theta} = .25$
- ○ $\hat{\theta} = .8$
- ○ $\hat{\theta} = .3$

h) **True or False**: The Union Bound always gives a result in $[0, 1]$.

2 Content Review II (material not properly covered in class)

a) Suppose C and D are discrete random variables. Then $\mathbb{E}[C|D = d] =$

- ○ $\sum_d d p_{D|C}(d|c)$
- ○ $\sum_c c p_{C|D}(c|d)$
- ○ $\int_{-\infty}^{\infty} c f_{C|D} \, \mathrm{d}x$
- ○ $\frac{\mathbb{E}[C]}{\mathbb{E}[D]}$

b) True or false: Markov's Inequality always gives a non-negative result.

c) True or false: Chebyshev's Inequality can best be described as giving an upper bound on the distribution's right tail.

3 Trapped Miner

A miner is trapped in a mine containing 3 doors.

- D_1: The 1^{st} door leads to a tunnel that will take him to safety after 3 hours.
- D_2: The 2^{nd} door leads to a tunnel that returns him to the mine after 5 hours.
- D_3: The 3^{rd} door leads to a tunnel that returns him to the mine after a number of hours that is Binomial with parameters $(12, \frac{1}{3})$.

At all times, he is equally likely to choose any one of the doors. What is the expected number of hours for this miner to reach safety?

4 The Compound Server

A web server receives requests according to a Poisson process with a rate of λ requests per minute. Each incoming request is independently classified as a "database query" with probability p, or a "static asset request" with probability $1 - p$. Let N be the total number of requests in a minute, and X be the number of database queries in a minute.

a) Derive the marginal distribution of X. Also, write down the marginal distribution of Y, defined as the number of static asset requests in a minute. (You only need to prove your answer for X.)

b) Show that X and Y are independent.

c) Given that exactly k database queries were received in a particular minute, what is the expected *total* number of requests $\mathbb{E}[N \mid X = k]$ received during that minute?

5 Lemonade Stand

Suppose I run a lemonade stand, which costs me \$100 a day to operate. I sell a drink of lemonade for \$20. Every person who walks by my stand either buys a drink or doesn't (no one buys more than one). If it is raining, n_1 people walk by my stand, and each buys a drink independently with probability p_1. If it isn't raining, n_2 people walk by my stand, and each buys a drink independently with probability p_2. It rains each day with probability p_3, independently of every other day. Let X be my profit over the next week. In terms of n_1, n_2, p_1, p_2 and p_3, what is $\mathbb{E}[X]$?

6 The Dice and Coin Cascade

Alice repeatedly rolls a fair six-sided die until she rolls a 6. Let N be the total number of rolls she makes (including the final roll of 6). Once Alice is finished, Bob flips a fair coin exactly N times. Let H be the total number of Heads Bob flips.

a) Compute the exact probability that Bob flips zero Heads, $\mathbb{P}(H = 0)$.

b) Compute the expected value $\mathbb{E}[H]$ and variance $\text{Var}(H)$.

7 Nested Uniforms

Let $X \sim \mathsf{Unif}(0, 1)$. We draw a random variable Y such that, conditioned on $X = x$, $Y \sim \mathsf{Unif}(0, x)$. What is the expected value of Y, $\mathbb{E}[Y]$?

8 A Red Poisson

Suppose that $x_1, \ldots, x_n$ are i.i.d. samples from a Poisson(θ) random variable, where θ is unknown. Find the MLE for θ.

9 Independent Shreds, You Say?

You are given 100 independent samples $X_1 = x_1, X_2 = x_2, \ldots, X_{100} = x_{100}$ from Bernoulli(θ), where θ is unknown. (Each sample is either a 0 or a 1). These 100 samples sum to 30. We saw in class that the maximum likelihood estimate for θ is

$$\hat{\theta}(x_1, \ldots, x_{100}) = \frac{1}{100} \sum_{i=1}^{100} x_i = \frac{30}{100}.$$

Is $\hat{\theta}(X_1, \ldots, X_{100}) = \frac{\sum_{i=1}^{100} X_i}{100}$ an unbiased estimator of θ?

10 Laplace MLE

Suppose $x_1, \ldots, x_{2n}$ are iid realizations from the Laplace density (double exponential density): for $x \in \mathbb{R}$,

$$f_X(x \mid \theta) = \frac{1}{2} e^{-|x-\theta|}$$

Find the MLE for θ. For this problem, you need not verify that the MLE is indeed a maximizer. You may find the **sign** function useful:

$$\operatorname{sgn}(x) = \begin{cases} +1, & x \geq 0 \\ -1, & x < 0 \end{cases}$$

11 Bird Watching

You are an ornithologist studying a rare species of birds in a nature reserve. Over a period of 50 days, you record the number of sightings of this bird (you see $x_1, x_2, ..., x_{50}$ birds on each day). Your research has shown that the number of sightings of this species depends on the average number of monkeys in the reserve, θ_1, and the average number of eagles in the reserve, θ_2. After years of studying this rare species in other environments, you've found that the number of birds observed on a particular day follows the following distribution:

$$p_X(k) = \frac{1}{k!}(\theta_1^k \cdot e^{-\theta_1} \cdot \theta_2^k \cdot e^{-3\theta_2})$$

Find the MLE for θ_1 and θ_2 (i.e., find $\hat{\theta_1}$ and $\hat{\theta_2}$).

a) What is the likelihood function?

b) What is the log-likelihood function?

c) We want to find values of θ_1 and θ_2 that maximize the likelihood function. To do this, we will take the partial derivative with respect to each of these parameters and solve for the values that make them both zero. First, take the partial derivative of the likelihood function with respect to θ_1.

d) Now, take the partial derivative with respect to θ_2.

e) Set both these partial derivatives to 0, and solve for $\hat{\theta_1}$ and $\hat{\theta_2}$.

Remark: Similar to the single-variable case, we treat the points where both partial derivatives are 0 (also known as the critical point) directly as the maximum of the function. This is just a convenient simplification for 312. For general functions, we need to do the second derivative test to confirm whether a critical point is indeed a local maximum/minimum. For multi-variable functions, the second derivative test is much more complicated - it is not enough to take the partial second derivatives, you would have to check the Hessian matrix.

12 A biased estimator

In class, we showed that the maximum likelihood estimate of the variance θ_2 of a normal distribution (when both the true mean μ and true variance σ^2 are unknown) is what's called the *population variance*. That is

$$\hat{\theta}_2 = \left(\frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\theta}_1)^2 \right)$$

where $\hat{\theta}_1 = \frac{1}{n} \sum_{i=1}^{n} x_i$ is the MLE of the mean. Is $\hat{\theta}_2$ unbiased?

13 It Means Nothing

Suppose $x_1, x_2, \ldots, x_n$ are samples from a normal distribution whose mean is known to be μ but the variance is unknown. How does the maximum likelihood estimator for the variance differ from the maximum likelihood estimator when both mean and variance are unknown? Which, if any, is unbiased?

14 Maximum Likelihood Estimators

a) Let $x_1, \ldots, x_n$ i.i.d. samples from a random variable that follows a Rademacher distribution with unknown parameter $p \in [0, 1]$, i.e., a distribution from the family

$$\mathbb{P}[x; p] = \begin{cases} p & \text{if } x = +1 \\ 1 - p & \text{if } x = -1 \end{cases},$$

What is the maximum likelihood estimator for p? Write this as a function of the x_i's and n.

b) Let $x_1, \ldots, x_n$ be i.i.d samples that follow a $\mathsf{Two}(\theta)$ distribution with unknown parameter $\theta \in [0, 1]$, where the probabilities from the family are given by

$$\mathbb{P}[x; \theta] = \begin{cases} (1-\theta)^2 & x = 0 \\ 2\theta(1-\theta) & x = 1 \\ \theta^2 & x = 2 \end{cases}$$

Suppose that in the sample there are n_0 0's, n_1 1's, and n_2 2's. What is the maximum likelihood estimator for θ in terms of n, n_0, n_1, n_2?

c) Let $x_1, \ldots, x_n$ be i.i.d. samples from a random variable that follows a so-called Borel distribution with unknown parameter θ, i.e., a distribution from the family

$$\mathbb{P}[k; \theta] = \frac{e^{-\theta k} (\theta k)^{k-1}}{k!},$$

where $0 < \theta \leq 1$ is a real number, and $k \geq 1$ is an integer. What is the maximum likelihood estimator for θ?

d) If the samples from the Borel distribution are 5, 7, 10, 2, 7, 5, 12, 13, 11, what is the maximum likelihood estimator for θ? Give an exact answer as a simplified fraction.

The remaining problems cover material we may not get to this quarter.

15 Tail bounds

Suppose $X \sim \mathsf{Binomial}(6, 0.4)$. We will bound $\mathbb{P}(X \geq 4)$ using the tail bounds we've learned, and compare this to the true result.

a) Give an upper bound for this probability using Markov's inequality. Why can we use Markov's inequality?

b) Give an upper bound for this probability using Chebyshev's inequality. You may have to rearrange algebraically and it may result in a weaker bound.

c) Give an upper bound for this probability using the Chernoff bound.

d) Give the exact probability.

16 Exponential Tail Bounds

Let $X \sim \text{Exp}(\lambda)$ and $k > 1/\lambda$.

a) Use Markov's inequality to bound $\mathbb{P}(X \geq k)$.

b) Use Markov's inequality to bound $\mathbb{P}(X < k)$.

c) Use Chebyshev's inequality to bound $\mathbb{P}(X \geq k)$.

d) What is the exact formula for $\mathbb{P}(X \geq k)$?

e) For $\lambda k \geq 3$, how do the bounds given in parts (a), (c), and (d) compare?

17 Robbie's Late!

Suppose the probability Robbie is late to teaching lecture on a given day is at most 0.01. Do not make any independence assumptions.

a) Use a Union Bound to bound the probability that Robbie is late at least once over a 30-lecture quarter.

b) Use a Union Bound to bound the probability that Robbie is **never** late over a 30-lecture quarter.

c) Use a Union Bound to bound the probability that Robbie is late at least once over a 120-lecture quarter.