

ECE368 Cheatsheet

Hanhee Lee

January 22, 2025

Contents

| | | |
|----------|--|-----------|
| 1 | L1: Probability Review | 2 |
| 1.1 | Sample Space | 2 |
| 1.2 | Probability Definitions | 2 |
| 1.3 | Axioms of Probability | 3 |
| 1.4 | Conditional Probability | 3 |
| 1.4.1 | Consequences of Conditional Probability | 3 |
| 1.4.2 | Independence | 3 |
| 1.4.3 | Importance of Labelling | 3 |
| 2 | L2: Probability Review | 4 |
| 2.1 | Total Probability | 4 |
| 2.2 | Bayes' Rule | 4 |
| 2.2.1 | Posteriori Probability, Priori Probability (Prior), Likelihood | 4 |
| 2.2.2 | Interpretation of Bayes' Rule | 4 |
| 2.3 | Random Variables | 5 |
| 2.4 | Distribution of RV | 5 |
| 2.4.1 | Cumulative Distribution Function (CDF) of RV | 5 |
| 2.4.2 | Discrete RV Probability Mass Function (PMF) | 5 |
| 2.4.3 | Continuous RV Probability Density Function (PDF) | 5 |
| 2.4.4 | Conditional PMF/PDF | 6 |
| 2.5 | Expected Values | 6 |
| 3 | L3: Probability Review | 8 |
| 3.1 | 2 RVs | 8 |
| 3.2 | Joint PMF/PDF | 8 |
| 3.3 | Expectations | 8 |
| 3.3.1 | Correlation | 8 |
| 3.3.2 | Covariance | 9 |
| 3.3.3 | Correlation Coefficient | 9 |
| 3.4 | Marginal PMF/PDF | 9 |
| 3.5 | Conditional PMF/PDF | 9 |
| 3.6 | Bayes' Rule | 10 |
| 3.7 | Independent vs. Uncorrelated vs. Orthogonal | 10 |
| 3.8 | Conditional Expectation | 10 |
| 4 | L4: Estimation of Sample Mean | 13 |
| 4.1 | Parameter Estimation: | 13 |
| 4.2 | Estimator: | 13 |
| 4.2.1 | Estimation Error: | 13 |
| 4.2.2 | Unbiased | 13 |
| 4.2.3 | Consistent | 13 |
| 4.3 | Sample Mean & Law of Large Numbers | 14 |
| 4.3.1 | Digression for Sum of RVs (not necessarily independent or identically distributed) | 14 |
| 4.3.2 | Unbiased (i.i.d.) | 14 |
| 4.3.3 | Consistent (i.i.d.) | 15 |

| | | |
|----------|--|-----------|
| 4.3.4 | Weak Law of Large Numbers | 15 |
| 4.3.5 | Confidence Interval: Finding n | 16 |
| 5 | L5: Sample Mean and Maximum Likelihood Estimation | 17 |
| 5.1 | Maximum Likelihood Estimation | 17 |
| 5.1.1 | Log-Likelihood | 17 |
| 6 | L6: Maximum Likelihood and Laplace | 20 |
| 6.1 | MLE for Categorical Random Variables | 20 |
| 6.2 | MLE for Gaussian Random Variables | 21 |
| 6.3 | Will the Sun Rise Tomorrow? (Laplace's Problem) | 21 |
| 6.3.1 | Frequentist Approach | 22 |
| 6.3.2 | Bayesian Approach | 22 |

Summary: On second thoughts, the lecture notes he posts are good, so I think I'll just do the cheatsheet.

W1 (LG-IPPR 1.1, 1.2; Murphy 2.1 – 2.3)

1 L1: Probability Review

Summary:

FAQ:

- How to study? Practice, practice.
- What textbooks? Use 2024 version of Murphy, Leon Garcia as main reference, Bishop, 4th textbook is intro.
- How is HW graded? Effort, and tutorials are used to explain soln.

1.1 Sample Space

Motivation: If you have 4 sheep and a flea, the probability that starting from sheep 1, the flea will jump to sheep 4 in 10 steps is 0.2.

- Ambiguous as there are 2 different interpretations for the sample space (i.e. space of probability is not clear):
 - Set of sheep
 - Set of number of steps

1.2 Probability Definitions

Definition:

- **Random Experiment:** An outcome (realization) for each run.
- **Sample Space Ω :** Set of all possible outcomes.
- **Events:** (measurable) subsets of Ω .
- **Probability of Event A :** $P[A] \equiv P[\text{'outcome is in } A\text{'}]$.

Example: Roll Fair Die

- $\Omega = \{1, 2, 3, 4, 5, 6\}$.
- $P[\text{'even number'}] = \frac{1}{2}$.

1.3 Axioms of Probability

Definition:

1. $P[A] \geq 0$ for all $A \in \Omega$.
2. $P[\Omega] = 1$.
3. If $A \cap B = \emptyset$, then $P[A \cup B] = P[A] + P[B]$ for all $A, B \in \Omega$.



Figure 1: 3rd Axiom

1.4 Conditional Probability

Definition:

$$P[A|B] = \frac{P[A \cap B]}{P[B]} \quad (1)$$

- $|\cdot$: Given event (data/obs.).

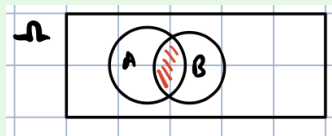


Figure 2: Conditional Probability

Notes:

- Changing sample space to B .
- Conditional probability satisfy the 3 axioms (i.e. are probabilities), can be viewed as probability measure on new sample space B .

1.4.1 Consequences of Conditional Probability

Definition:

$$P[A \cap B] = P[A|B]P[B] = P[B|A]P[A] \quad (2)$$

1.4.2 Independence

Definition: A and B are independent iff

$$P[A \cap B] = P[A]P[B] \iff P[A|B] = P[A] \iff P[B|A] = P[B] \quad (3)$$

1.4.3 Importance of Labelling

Example: Toss 2 Fair Coins

1. **Given:** Given that one of the coins is heads, what is the probability that the other coin is tails?
2. **Wrong Solution:** $\frac{1}{2}$ since $\{HH, HT, TH, TT\}$, so $P[T|H] = \frac{1}{2}$, which assumes that the coins are distinguishable (i.e. coin #1 is heads)
3. **Correct Solution:** $\frac{2}{3}$ since $\{HH, HT, TH\}$ as we didn't specify which coin was heads, so $P[T|H] = \frac{2}{3}$, which assumes that the coins are indistinguishable.

2 L2: Probability Review

2.1 Total Probability

Definition: If H_1, \dots, H_n form a partition of Ω , then

$$P[A] = \sum_{i=1}^n P[A|H_i]P[H_i] \quad (4)$$

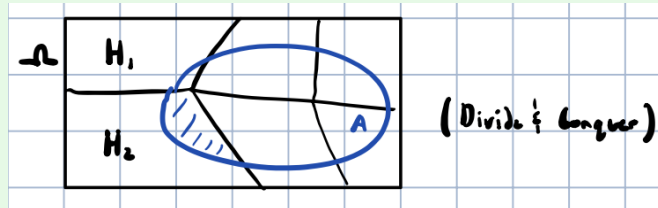


Figure 3: Total Probability

2.2 Bayes' Rule

Definition:

$$P[H_k|A] = \frac{P[H_k \cap A]}{P[A]} = \frac{P[A|H_k]P[H_k]}{\sum_{i=1}^n P[A|H_i]P[H_i]} \quad (5)$$

2.2.1 Posteriori Probability, Priori Probability (Prior), Likelihood

Definition:

- **Posteriori:** $P[H_k|A]$.
- **Priori:** $P[H_k]$.
- **Likelihood:** $P[A|H_k]$.

Example: Suppose a lie detector is 95% accurate, i.e. $P[\text{'out=truth'}|\text{'in=truth'}] = 0.95$ and $P[\text{'out=lie'}|\text{'in=lie'}] = 0.95$. It says that Mr. Ernst is lying. What is the probability Mr. Ernst is actually lying.

- **Observation:** $A = \text{'out=lie'}$.
- **Hypothesis:** $H_0 = \text{'in=lie'}$ and $H_1 = \text{'in=truth'}$.
- **Solution:**
$$P[H_0|A] = \frac{P[A|H_0]P[H_0]}{P[A|H_0]P[H_0] + P[A|H_1]P[H_1]} = \frac{0.95 \times P[H_0]}{0.95 \times P[H_0] + 0.05 \times (1 - P[H_0])}.$$
- $H_0 = 0.01$: i.e. 1% of the population are liars, then
$$P[H_0|A] = \frac{0.95 \times 0.01}{0.95 \times 0.01 + 0.05 \times 0.99} = 0.16.$$

Warning: Need to know priori probability.

2.2.2 Interpretation of Bayes' Rule

Notes: Taking one component of the total probability and normalizing it by the sum of all components.

2.3 Random Variables

Motivation: Coin Toss Mapping of each outcome to a real number

- $w \in \Omega$ is the outcome of a coin toss, and X is the RV, so $H \rightarrow 0$ and $T \rightarrow 1$.



Figure 4: Random Variables

- Mapping is deterministic function. RV is not random or variable.

Definition: Mapping from Ω to \mathbb{R} .

2.4 Distribution of RV

2.4.1 Cumulative Distribution Function (CDF) of RV

Definition:

$$F_X(x) \equiv P[X \leq x] \quad (6)$$

2.4.2 Discrete RV Probability Mass Function (PMF)

Definition:

$$P_X(x_j) \equiv P[X = x_j] \quad j = 1, 2, 3, \dots \quad (7)$$

Example: Binomial RV w/ (n, p)

$$P_X(k) = \binom{n}{k} p^k (1-p)^{n-k} \quad (8)$$

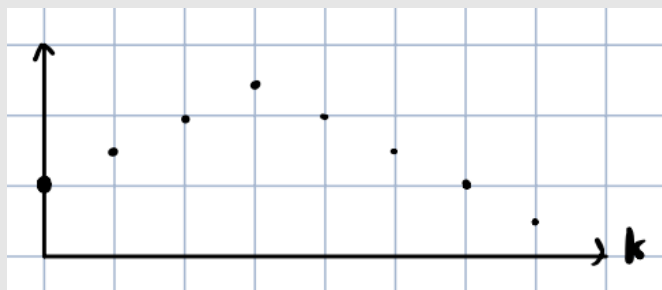


Figure 5: Binomial RV

2.4.3 Continuous RV Probability Density Function (PDF)

Definition:

$$f_X(x) \equiv \frac{d}{dx} F_X(x) \quad (9)$$

$$P[x < X < x + dx] = f_X(x)dx \quad (10)$$

Example: Gaussian RV w/ (μ, σ^2)

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (11)$$

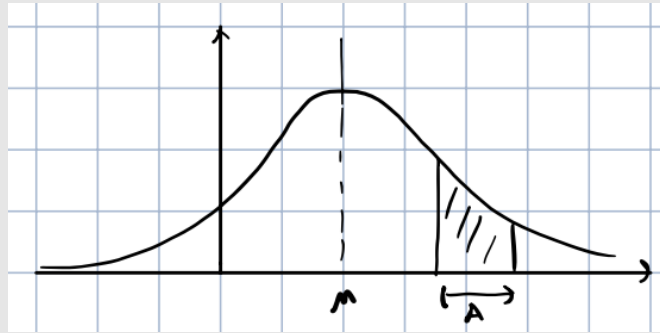


Figure 6: Gaussian RV

- $P[X \in A] = \int_A f_X(x)dx.$

Notes: Discrete RV has pdf w/ δ functions.

2.4.4 Conditional PMF/PDF

Definition:

$$P_X(x|A) \quad (12)$$

$$f_X(x|A) \quad (13)$$

Example: Continuous

$$f(x|X > a) = \begin{cases} \frac{f_X(x)}{P[X > a]} & \text{if } x > a \\ 0 & \text{otherwise} \end{cases} \quad (14)$$

Example: Geometric RV Geometric RV X w/ success probability p

$$P_X(k) = (1-p)^{k-1}p \quad (15)$$

- **Memoryless Property:** $P_X[k|X > m] = \frac{p(1-p)^{k-1}}{(1-p)^m} = p(1-p)^{k-m-1}.$
 - So it only cares about the additional trials (i.e. same as resetting after m trials).

2.5 Expected Values

Definition:

$$E[X] = \int_{-\infty}^{\infty} x f_X(x) dx \stackrel{\text{If int. values}}{=} \sum_{k=-\infty}^{\infty} k f_X(k) \quad (16)$$

$$E[h(X)] = \int_{-\infty}^{\infty} h(x) f_X(x) dx \stackrel{\text{If int. values}}{=} \sum_{k=-\infty}^{\infty} h(k) f_X(k) \quad (17)$$

$$\text{Var}[X] = E[(X - E[X])^2] = E[X^2] - E[X]^2 \quad (18)$$

$$E[X|A] = \int_{-\infty}^{\infty} x f_X(x|A) dx \quad (19)$$

Example: Lottery Ticket (Geometric RV)

1. **Given:** Buying one lottery ticket per week
 - Each ticket has $10^{-7} = p$ chance of winning the jackpot.
 - X = '# of weeks to win jackpot'.
2. **Problem:** What is the expected number of weeks to win the jackpot?
3. **Solution:** $E[X] = \sum_{k=1}^{\infty} k(1-p)^{k-1}p = \dots = \frac{1}{p} = 10^7$ weeks.
4. **Extension (Memoryless Property):** If I have already played for 999999 weeks, what is the expected number of weeks to win the jackpot? $E[X - 999999 | X > 999999] = E[X] = 10^7$ weeks.

3 L3: Probability Review

3.1 2 RVs

Notes: RVs are neither random nor a variable.

$$\underline{Z} = (X, Y)$$

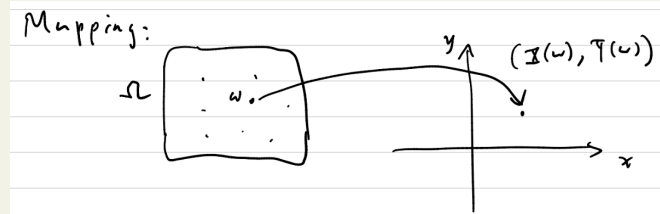


Figure 7: Mapping of RVs

3.2 Joint PMF/PDF

Definition:

$$P_{X,Y}(x, y) = P[X = x, Y = y] \quad (20)$$

$$f_{X,Y}(x, y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x, y) \quad (21)$$

$$P[(X, Y) \in A] = \int \int_{(x,y) \in A} f_{X,Y}(x, y) dx dy \quad (22)$$

Example: Jointly Gaussian RVs X and Y with $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$

$$f_{X,Y}(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp \left\{ -\frac{1}{2(1-\rho^2)} \left[\left(\frac{x-\mu_1}{\sigma_1} \right)^2 - 2\rho \left(\frac{x-\mu_1}{\sigma_1} \right) \left(\frac{y-\mu_2}{\sigma_2} \right) + \left(\frac{y-\mu_2}{\sigma_2} \right)^2 \right] \right\}$$

3.3 Expectations

Definition:

$$E[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f_{X,Y}(x, y) dx dy$$

Notes:

- $g(X, Y)$ is also an RV, but inside the integral or sum, you use x and y as dummy variables to vary through the values of the RVs.

3.3.1 Correlation

Definition:

$$E[XY] \quad (23)$$

3.3.2 Covariance

Definition:

$$\text{Cov}[X, Y] = E[(X - \mu_X)(Y - \mu_Y)] = E[XY] - \mu_X \mu_Y = E[XY] - E[X]E[Y] \quad (24)$$

Notes:

- Mean shifted to 0.

3.3.3 Correlation Coefficient

Definition:

$$\rho_{X,Y} = E \left[\left(\frac{X - \mu_X}{\sigma_X} \right) \left(\frac{Y - \mu_Y}{\sigma_Y} \right) \right] = \frac{\text{Cov}[X, Y]}{\sigma_X \sigma_Y} \quad (25)$$

- $|\rho_{X,Y}| \leq 1$

Notes:

- Mean shifted to 0 and normalized by the standard deviation.

3.4 Marginal PMF/PDF

Definition:

$$P_X(x) = \sum_{j=1}^{\infty} P_{X,Y}(x, y_j), \quad P_Y(y) = \sum_{j=1}^{\infty} P_{X,Y}(x_j, y) \quad (26)$$

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dy, \quad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dx \quad (27)$$

Notes:

- Total probability theorem is being used here.

Example: Jointly Gaussian X and Y :

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f_{X,Y}(x, y) dy \\ &= \dots \quad (\text{completing the square}) \\ &= \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}, \quad \text{marginally Gaussian} \end{aligned}$$

- Gaussian RVs has a property that the PDF of a single variable is equal to the marginal Gaussian of two variables.

3.5 Conditional PMF/PDF

Definition:

$$P_{X|Y}(x|y) \triangleq P[X = x|Y = y] = \frac{P_{X,Y}(x, y)}{P_Y(y)} \quad (28)$$

$$f_{X|Y}(x|y) \triangleq \frac{f_{X,Y}(x, y)}{f_Y(y)} \quad (29)$$

3.6 Bayes' Rule

Definition:

$$P_{Y|X}(x|y) = \frac{P_{X,Y}(x,y)}{P_X(x)} = \frac{P_{X|Y}(x|y)P_Y(y)}{\sum_{j=1}^{\infty} P_{X,Y}(x,y_j)P_Y(y_j)} \quad (30)$$

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{f_{X|Y}(x|y)f_Y(y)}{\int_{-\infty}^{\infty} f_{X|Y}(x|y')f_Y(y') dy'} \quad (31)$$

3.7 Independent vs. Uncorrelated vs. Orthogonal

Definition:

1. Independent:

$$f_{X|Y}(x|y) = f_X(x) \quad \forall y \Leftrightarrow f_{X,Y}(x,y) = f_X(x)f_Y(y) \quad (32)$$

2. Uncorrelated:

$$\text{Cov}[X,Y] = 0 \quad \Leftrightarrow \quad \rho_{X,Y} = 0 \quad (33)$$

3. Orthogonal:

$$E[XY] = 0 \quad (34)$$

Theorem: If independent, then uncorrelated.

Derivation:

$$\begin{aligned} \text{Independent} &\implies E[XY] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy f_{X,Y}(x,y) dx dy \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy f_X(x) f_Y(y) dx dy \\ &= \left(\int_{-\infty}^{\infty} x f_X(x) dx \right) \left(\int_{-\infty}^{\infty} y f_Y(y) dy \right) \\ &\implies E[XY] = E[X]E[Y] \\ &\implies \text{Cov}[X,Y] = 0, \quad \text{uncorrelated} \\ &\neq \text{in general.} \end{aligned}$$

Example: Jointly Gaussian RVs X and Y : If uncorrelated, i.e. $\rho_{X,Y} = 0$, then X and Y are independent.

$$\begin{aligned} f_{X,Y}(x,y) &= \frac{1}{2\pi\sigma_1\sigma_2} \exp \left\{ -\frac{1}{2} \left[\left(\frac{x-\mu_1}{\sigma_1} \right)^2 + \left(\frac{y-\mu_2}{\sigma_2} \right)^2 \right] \right\} \\ &= \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} \cdot \frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{(y-\mu_2)^2}{2\sigma_2^2}} \\ &= f_X(x)f_Y(y) \quad \text{independent} \end{aligned}$$

3.8 Conditional Expectation

Definition:

$$E[Y] = E[E[Y|X]] \quad (35)$$

$$E[h(Y)] = E[E[h(Y)|X]] \quad (36)$$

Notes:

- $E[E[Y|X]]$ is w.r.t. X .
- $E[Y|X]$ is w.r.t. Y .

Derivation:

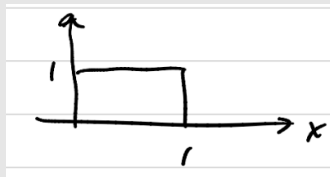
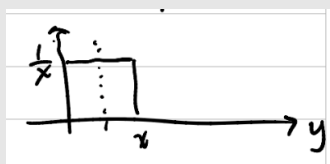
$$\begin{aligned}
 E[Y] &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f_{X,Y}(x,y) dx dy \\
 &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f_{Y|X}(y|x) f_X(x) dx dy \\
 &= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} y f_{Y|X}(y|x) dy \right) f_X(x) dx \\
 &= \int_{-\infty}^{\infty} E[Y|X=x] f_X(x) dx \quad (\text{using the total probability theorem}) \\
 &= \int_{-\infty}^{\infty} g(x) f_X(x) dx \\
 &= E[g(X)] \\
 &= E[E[Y|X]].
 \end{aligned}$$

Example:

1. **Given:** An unknown voltage. $X \sim \text{Uniform}(0, 1)$. Measurement from a (bad) voltmeter: $Y \sim \text{Uniform}(0, X)$.

$$\begin{aligned}
 f_X(x) &= \begin{cases} 1, & 0 < x < 1 \\ 0, & \text{otherwise} \end{cases} \\
 f_{Y|X}(y|x) &= \begin{cases} \frac{1}{x}, & 0 < y < x \\ 0, & \text{otherwise} \end{cases}
 \end{aligned}$$

- **Note:** Area under PDF is 1.

Figure 8: Uniform Distribution of X Figure 9: Uniform Distribution of Y

2. Expected Value (Average Reading of Bad Voltmeter):

$$\begin{aligned}
 E[Y] &= E[E[Y|X]] \\
 &= E\left[\frac{X}{2}\right] \quad \text{Since in the middle of 0 and x} \\
 &= \frac{1}{2} \cdot E[X] \\
 &= \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4} \quad \text{Since } E[X] \text{ (i.e. mean) is 0.5}
 \end{aligned}$$

3. The Long Way:

$$\begin{aligned}
 f_Y(y) &= \int_{-\infty}^{\infty} f_{Y|X}(y|x) f_X(x) dx \\
 &= \int_y^1 f_{Y|X}(y|x) f_X(x) dx \\
 &= \int_y^1 \frac{1}{x} \cdot 1 dx \\
 &= -\ln y. \\
 E[Y] &= \int_0^1 y \cdot (-\ln y) dy = \dots = \frac{1}{4}
 \end{aligned}$$

4. **Question:** Suppose $Y = \frac{1}{8}$. What is "best" given X ? This will be the question for the rest of the course.

4 L4: Estimation of Sample Mean

4.1 Parameter Estimation:

Motivation: The readout of a sensor is $X = \theta + N$ volts

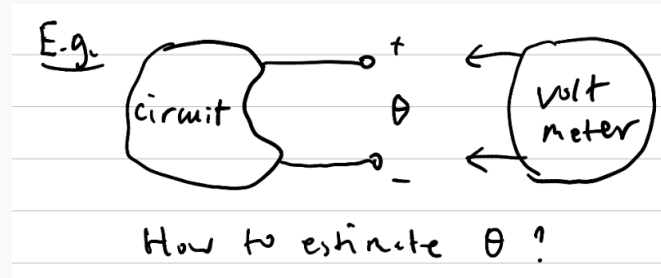


Figure 10:

- There is some noise N in the sensor, so we want to estimate the true value of θ (unknown parameter to be estimated)
 - e.g. Mean and/or variance of X .

4.2 Estimator:

Definition: Perform n independent and identically distributed (i.i.d.) measurements/observations of X : X_1, X_2, \dots, X_n .

$$\hat{\Theta} = \hat{\Theta}(\underline{X}) = g(X_1, X_2, \dots, X_n) \quad (37)$$

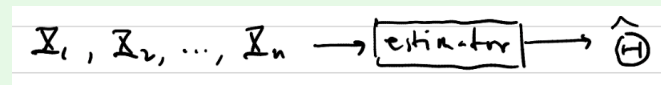


Figure 11:

4.2.1 Estimation Error:

Definition:

$$\hat{\Theta}(\underline{X}) - \theta \quad (38)$$

4.2.2 Unbiased

Definition: The estimator $\hat{\Theta}$ is unbiased if

$$\mathbb{E}[\hat{\Theta}(\underline{X})] = \theta \quad (39)$$

- **Asymptotically Unbiased:** $\lim_{n \rightarrow \infty} \mathbb{E}[\hat{\Theta}(\underline{X})] = \theta$ (big data)

4.2.3 Consistent

Definition: The estimator $\hat{\Theta}$ is consistent if $\hat{\Theta}(\underline{X}) \rightarrow \theta$ as $n \rightarrow \infty$, in probability, i.e., $\forall \epsilon > 0$,

$$\lim_{n \rightarrow \infty} P(|\hat{\Theta}(\underline{X}) - \theta| < \epsilon) \rightarrow 1 \quad (40)$$

as $n \rightarrow \infty$.

4.3 Sample Mean & Law of Large Numbers

Definition: Given a sequence of i.i.d. random variables (RVs), X_1, X_2, \dots, X_n , w/ unknown mean μ , estimate μ . Let $S_n = X_1 + X_2 + \dots + X_n$. The *sample mean* is

$$M_n = \frac{1}{n} S_n$$

- How good is M_n as an estimator of μ ?
 - Use unbiased and consistent to evaluate M_n .

Example: Previous voltage measurement, e.g.,

$$X_i = \mu + N_i$$

where μ is the true value and N_i is the noise.

If we assume N_i are i.i.d. with zero mean,

$$E[X_i] = E[\mu + N_i] = E[\mu] + E[N_i] = \mu + 0 = \mu, \quad \forall i$$

4.3.1 Digression for Sum of RVs (not necessarily independent or identically distributed)

Derivation:

$$\begin{aligned} E[S_n] &= E[X_1 + \dots + X_n] \\ &= E[X_1] + \dots + E[X_n] \end{aligned}$$

Derivation:

$$\begin{aligned} \text{Var}[S_n] &= E[(S_n - E[S_n])^2] \\ &= E\left[\left(\sum_{i=1}^n X_i - E[X_i]\right)^2\right] \\ &= E\left[\sum_{i=1}^n \sum_{j=1}^n (X_i - E[X_i])(X_j - E[X_j])\right] \\ &= \sum_{i=1}^n \sum_{j=1}^n E[(X_i - E[X_i])(X_j - E[X_j])] \\ &= \sum_{i=1}^n \sum_{j=1}^n \text{Cov}[X_i, X_j] \\ &= \sum_{i=1}^n \text{Var}[X_i] + \sum_{i \neq j} \sum_{j=1}^n \text{Cov}[X_i, X_j] \end{aligned}$$

4.3.2 Unbiased (i.i.d.)

Derivation:

$$\begin{aligned}
E[M_n] &= E\left[\frac{1}{n}S_n\right] \\
&= \frac{1}{n}(E[X_1] + \dots + E[X_n]) \\
&= \frac{1}{n}(n\mu) \quad \text{since } X_i \text{ are i.i.d. so same expectation} \\
&= \mu \Rightarrow \text{Unbiased!}
\end{aligned}$$

4.3.3 Consistent (i.i.d.)**Derivation:**

$$\begin{aligned}
\text{Var}[M_n] &= \text{Var}\left[\frac{1}{n}S_n\right] \\
&= \frac{1}{n^2}\text{Var}[S_n] \quad \text{taking out constant requires squaring} \\
&= \frac{1}{n^2}\left(\sum_{i=1}^n \text{Var}[X_i] + \sum_{i \neq j} \text{Cov}[X_i, X_j]\right) \\
&= \frac{1}{n^2}(n\sigma^2) \quad \sigma^2 \triangleq \text{Var}[X_i] \text{ and } X_i \text{ are i.i.d. so covariance is 0} \\
&= \frac{\sigma^2}{n} \rightarrow 0 \text{ as } n \rightarrow \infty.
\end{aligned}$$

- This means that there is no variance in the sample mean as n approaches infinity, so it converges to the true mean.

Recall the Chebyshev Inequality:

$$P[|X - E[X]| \geq \epsilon] \leq \frac{\text{Var}[X]}{\epsilon^2}, \quad \forall \epsilon > 0.$$

Substitute in M_n :

$$\begin{aligned}
P[|M_n - E[M_n]| \geq \epsilon] &\leq \frac{\text{Var}[M_n]}{\epsilon^2} \\
P[|M_n - \mu| \geq \epsilon] &\leq \frac{\sigma^2}{n\epsilon^2} \\
\Rightarrow P[|M_n - \mu| < \epsilon] &\geq 1 - \frac{\sigma^2}{n\epsilon^2} \rightarrow 1 \text{ as } n \rightarrow \infty \text{ then it is consistent}
\end{aligned}$$

Warning: Cov = 0 because independence implies uncorrelated.

4.3.4 Weak Law of Large Numbers

Definition: Even if σ is infinite, then $\forall \epsilon > 0$,

$$\lim_{n \rightarrow \infty} P[|M_n - \mu| < \epsilon] = 1$$

4.3.5 Confidence Interval: Finding n

Example: Measure an unknown voltage θ for n times and obtain independent measurements:

$$X_i = \theta + N_i,$$

where N_i are i.i.d. random variables with mean 0 and variance 1.

- We want to determine how many measurements n are sufficient so that

$$P(|M_n - \theta| < 0.1) \geq 0.95,$$

where 0.1 is the desired precision and 0.95 is the confidence level.

- The sample mean is given by:

$$M_n = \frac{1}{n} \sum_{i=1}^n X_i = \theta + \frac{1}{n} \sum_{i=1}^n N_i.$$

- The variance of X_i is:

$$\sigma^2 = \text{Var}[X_i] = \text{Var}[\theta + N_i] = \text{Var}[N_i] = 1.$$

– $\text{Var}[aX + b] = a^2 \text{Var}[X]$, where $a = 1$ and $b = \theta$.

- Using Chebyshev's inequality:

$$1 - \frac{\sigma^2}{n\epsilon^2} \geq 0.95,$$

where $\epsilon = 0.1$ (precision).

- Solving for n :

$$\begin{aligned} 1 - \frac{1}{n(0.1)^2} &\geq 0.95, \\ \frac{1}{n(0.1)^2} &\leq 0.05, \\ n &\geq 2000. \end{aligned}$$

Thus, at least 2000 measurements are needed to achieve the desired precision and confidence level.

5 L5: Sample Mean and Maximum Likelihood Estimation

FAQ:

- Why can we say that it is consistent for the last example?

5.1 Maximum Likelihood Estimation

Motivation: Choose parameter θ that is most likely to generate the observation x_1, x_2, \dots, x_n .

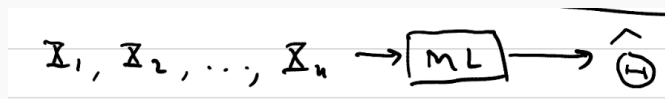


Figure 12:

Definition:

$$\hat{\Theta} = \arg \max_{\theta} P_{\underline{X}}(\underline{x}|\theta), \text{ discrete } X. \quad (41)$$

$$\hat{\Theta} = \arg \max_{\theta} f_{\underline{X}}(\underline{x}|\theta), \text{ continuous } X. \quad (42)$$

5.1.1 Log-Likelihood

Definition:

$$\hat{\theta} = \arg \max_{\theta} \sum_{i=1}^n \log P_X(x_i|\theta) \quad (43)$$

$$\hat{\theta} = \arg \max_{\theta} \sum_{i=1}^n \log f_X(x_i|\theta). \quad (44)$$

Warning: Can only go from argmax of fcn to argmax of log fcn, if it is i.i.d.

Derivation:

$$\begin{aligned}
 \text{i.i.d. } X_1, X_2, \dots, X_n &\implies \\
 p_{\underline{X}}(\underline{x}|\theta) &= \prod_{i=1}^n p_{X_i}(x_i|\theta) \\
 &= \prod_{i=1}^n p_X(x_i|\theta) \quad \text{drop the i due to i.i.d. assumption} \\
 \log p_{\underline{X}}(\underline{x}|\theta) &= \sum_{i=1}^n \log p_X(x_i|\theta).
 \end{aligned}$$

Example:

1. Model and Observations:

- Assume a biased coin with probability θ of showing heads. Find ML estimator for θ .
- Toss the coin n times and obtain Bernoulli random variables X_1, \dots, X_n such that:

$$\text{"heads"} \rightarrow 1, \quad \text{"tails"} \rightarrow 0.$$

- Total number of heads is:

$$k = \sum_{i=1}^n X_i.$$

For example:

$$\underline{x} = (1, 0, 0, 0, 1, 1, 0, 1, 1, 1), \quad k = 6.$$

- Probability of observations x_1, \dots, x_n corresponding to parameter θ is:

$$p_{\underline{X}}(\underline{x}|\theta) = \theta^k (1 - \theta)^{n-k}.$$

- It is sufficient to know only k .
- Note: Don't need the $\binom{n}{k}$ term because we are given the specific sequence of heads and tails.

2. Log-Likelihood and Maximization:

- The log-likelihood function is:

$$\log p_{\underline{X}}(\underline{x}|\theta) = k \log(\theta) + (n - k) \log(1 - \theta).$$

- To maximize the log-likelihood over θ , set:

$$\begin{aligned} 0 &= \frac{\partial}{\partial \theta} \log p_{\underline{X}}(\underline{x}|\theta), \\ 0 &= \frac{k}{\theta} - \frac{n - k}{1 - \theta}, \\ \theta &= \frac{k}{n}. \end{aligned}$$

- Thus, the Maximum Likelihood Estimator (MLE) is:

$$\hat{\Theta} = \frac{k}{n}, \quad \text{where } k = \sum_{i=1}^n X_i.$$

This corresponds to the observed frequency of heads, which is intuitive b/c the more heads we see, the more likely the coin is biased towards heads.

3. Examples:

- For $\underline{x} = (1, 0, 0, 0, 1, 1, 0, 1, 1, 1)$:

$$\begin{aligned} p_X(\underline{x}|\theta) &= \theta^6 (1 - \theta)^4, \\ \hat{\theta} &= \frac{6}{10} = 0.6. \end{aligned}$$

- For $\underline{x} = (0, 1, 1, 1, 0, 0, 1, 0, 1, 0)$:

$$\begin{aligned} p_X(\underline{x}|\theta) &= \theta^5 (1 - \theta)^5, \\ \hat{\theta} &= \frac{5}{10} = 0.5. \end{aligned}$$

Notes:

1. k is a sufficient statistic for this Maximum Likelihood (ML) estimator.

2. The expectation of the estimator $\hat{\theta}$ is:

$$\begin{aligned} E[\hat{\Theta}] &= E\left[\frac{1}{n} \sum_{i=1}^n X_i\right] \\ &= \frac{1}{n} \sum_{i=1}^n E[X_i] \\ &= \frac{1}{n}(n\theta) \\ &= \theta \quad (\text{Unbiased}). \end{aligned}$$

$$- E[X_i] = (1)\theta + (0)(1 - \theta) = \theta$$

3. In fact, $\hat{\Theta} = \frac{1}{n} \sum_{i=1}^n X_i$ is the sample mean, and θ is the true mean. Therefore, $\hat{\theta} \rightarrow \theta$ in probability, which implies that $\hat{\theta}$ is *consistent*.

6 L6: Maximum Likelihood and Laplace

6.1 MLE for Categorical Random Variables

Example:

1. We say that $X \sim \text{Cat}(\underline{\theta})$ if

$$P[X = m] = \theta_m, \quad m = 1, 2, \dots, M.$$

- Going from 2 to M categories is a generalization of the Bernoulli distribution.

The parameter $\underline{\theta}$ is a vector:

$$\underline{\theta} = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_M \end{bmatrix},$$

such that $\theta_m \geq 0$ and $\sum_{m=1}^M \theta_m = 1$.

2. Given n i.i.d. observations X_1, \dots, X_n , we aim to find the maximum likelihood estimator (MLE) of $\underline{\theta}$.
3. Define n_m as the number of observations that equal m :

$$n_m = \sum_{i=1}^n 1(x_i = m),$$

where $1(x_i = m)$ is the indicator function. Note that $\sum_{m=1}^M n_m = n$.

4. The likelihood function is:

$$p_{\underline{X}}(\underline{x} | \underline{\theta}) = \prod_{m=1}^M \theta_m^{n_m}.$$

- Similar to the Bernoulli distribution, but with M categories.

Taking the log, we get:

$$\log p_{\underline{X}}(\underline{x} | \underline{\theta}) = \sum_{m=1}^M n_m \log \theta_m.$$

5. To find the optimal $\underline{\theta}$, we minimize the negative log-likelihood:

$$\min_{\underline{\theta}} - \sum_{m=1}^M n_m \log \theta_m,$$

subject to the constraints $\theta_m \geq 0$ for $1 \leq m \leq M$ and $\sum_{m=1}^M \theta_m = 1$.

6. Solving this optimization problem, the MLE is:

$$\hat{\theta}_m = \frac{N_m}{n} = \frac{\sum_{i=1}^n 1(X_i = m)}{n}, \quad \hat{\underline{\theta}} = \begin{bmatrix} \frac{N_1}{n} \\ \vdots \\ \frac{N_M}{n} \end{bmatrix}.$$

6.2 MLE for Gaussian Random Variables

Example:

1. Given n i.i.d. observations X_1, \dots, X_n of a Gaussian random variable with parameters (μ, σ^2) , we aim to find the maximum likelihood estimators (MLEs) of μ and σ^2 .

$$f_{\underline{X}}(\underline{x}|\mu, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right),$$

$$\log f_{\underline{X}}(\underline{x}|\mu, \sigma^2) = \sum_{i=1}^n \left(-\frac{(x_i - \mu)^2}{2\sigma^2} - \log \sigma - \log \sqrt{2\pi}\right).$$

2. To find μ , take the derivative of the log-likelihood with respect to μ and set it to zero:

$$0 = \frac{\partial}{\partial \mu} \sum_{i=1}^n \left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right),$$

$$0 = \frac{1}{n} \sum_{i=1}^n x_i - \mu,$$

$$\mu = \frac{1}{n} \sum_{i=1}^n x_i.$$

3. To find σ^2 , take the derivative of the log-likelihood with respect to σ^2 and set it to zero:

$$0 = \frac{\partial}{\partial \sigma^2} \sum_{i=1}^n \left(-\frac{(x_i - \mu)^2}{2\sigma^2} - \frac{1}{2} \log \sigma^2\right),$$

$$0 = -\frac{1}{2} \sum_{i=1}^n \left(\frac{(x_i - \mu)^2}{\sigma^4}\right) + \frac{1}{2\sigma^2},$$

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2.$$

4. Thus, the MLEs are:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n X_i, \quad (\text{sample mean})$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \hat{\mu})^2. \quad (\text{sample variance})$$

- Note: The sample variance is biased, so we often use $\frac{1}{n-1}$ instead of $\frac{1}{n}$ to make it unbiased.
- Note: The sample mean is unbiased.

6.3 Will the Sun Rise Tomorrow? (Laplace's Problem)

Example:

- Observation: The Sun has risen for n consecutive days. Estimate the probability that it will rise tomorrow.
- Model: Assume n i.i.d. Bernoulli random variables X_1, \dots, X_n with $P[X_i = 1] = \theta$.

6.3.1 Frequentist Approach

Example:

1. The Maximum Likelihood Estimator (MLE) is:

$$\hat{\theta} = \frac{K}{n} = \frac{\sum_{i=1}^n X_i}{n}.$$

2. If $K = n$ (i.e., the Sun has risen every day so far), then:

$$\hat{\theta} = \frac{n}{n} = 1.$$

3. Conclusion: The Sun will rise tomorrow with probability 1, regardless of what n is, based on the Frequentist approach.
 - This doesn't make sense b/c if $n = 1$ then we are assuming 100% it will rise based on one observation.

6.3.2 Bayesian Approach

Example:

1. Assume that θ is not fixed but drawn from a uniform distribution in $[0, 1]$. This means that the probability of the Sun rising is based on a uniform distribution.
2. We want to find the probability that the sun will rise tomorrow given that it has risen for n consecutive days:

$$P[X_{n+1} = 1 | X_1 = 1, \dots, X_n = 1].$$

Using Bayes' Theorem:

$$P[X_{n+1} = 1 | X_1 = 1, \dots, X_n = 1] = \frac{P[X_1 = 1, \dots, X_{n+1} = 1]}{P[X_1 = 1, \dots, X_n = 1]}.$$

3. Compute $P[X_1 = 1, \dots, X_n = 1]$:

$$P[X_1 = 1, \dots, X_n = 1] = \int_0^1 P[X_1 = 1, \dots, X_n = 1 | \Theta = \theta] f_{\Theta}(\theta) d\theta.$$

- The joint probability is calculated by integrating the product of the likelihood and the prior (i.e. marginalizing over θ).
- The likelihood becomes θ^n because the observations are i.i.d.
- The prior is uniform, so $f_{\Theta}(\theta) = 1$.

Since $f_{\Theta}(\theta) = 1$ (uniform prior) and $P[X_1 = 1, \dots, X_n = 1 | \Theta = \theta] = \theta^n$, we have:

$$P[X_1 = 1, \dots, X_n = 1] = \int_0^1 \theta^n d\theta = \frac{1}{n+1}.$$

4. Compute $P[X_1 = 1, \dots, X_{n+1} = 1]$ similarly:

$$P[X_1 = 1, \dots, X_{n+1} = 1] = \int_0^1 \theta^{n+1} d\theta = \frac{1}{n+2}.$$

5. Combine results:

$$P[X_{n+1} = 1 | X_1 = 1, \dots, X_n = 1] = \frac{\frac{1}{n+2}}{\frac{1}{n+1}} = \frac{n+1}{n+2}.$$

6. Conclusion: As n increases, the probability approaches 1, providing more certainty with more data.