EECS 126 Probability and Random Processes — University of California, Berkeley: Fall 2018
Kannan Ramchandran and Abhay Parekh — December 12, 2018

Final

Last Name	First Name	SID

- You have 5 minutes to read the exam and 175 minutes to complete this exam.
- The maximum you can score is 134, but 100 points is considered perfect.
- The exam is not open book, but you are allowed to consult the cheat sheet that we provide. No calculators or phones. No form of collaboration between the students is allowed. If you are caught cheating, you may fail the course and face disciplinary consequences.
- **Show all work to get any partial credit.**
- Take into account the points that may be earned for each problem when splitting your time between the problems.

Problem	points earned	out of
Problem 1		45
Problem 2		20
Problem 3		10
Problem 4		12
Problem 5		7
Problem 6		20
Problem 7		20
Total		100 (+34)

Problem 1: Answer these questions briefly but clearly. [45]

1. **Maximum Variance [5]**
   Let X be a random variable that takes value between 0 and c, where c is positive-valued (i.e. $\mathbb{P}(0 \leq X \leq c) = 1$). What is the maximum value of the variance of X? Provide an example which achieves this bound. You do not need to prove that your bound (if correct) is tight.

   The bound is achieved by $X = \begin{cases} c, \text{w.p. } 1/2 \\ 0, \text{w.p. } 1/2 \end{cases}$

   The value of the variance in this case is $c^2/4$

   Proof: Let $\mu = E[X]$. Then $var(X) = E[X^2] - \mu^2 = \sum x^2 p(x) - \mu^2 \leq \sum c \cdot xp(x) - \mu^2 = c\mu - \mu^2 = \mu(c - \mu)$. This is maximized by $\mu = c/2$, giving an upper bound of $c^2/4$.

2. **Min and Max of Uniform Distribution [5]** Let X and Y be independent random variables distributed as Uniform$[0, 1]$. Let $U = \min\{X, Y\}$ and $V = \max\{X, Y\}$. Find $\text{cov}(U, V)$.

   $cov(U,V) = E[UV] - E[U]E[V] = E[XY] - 1/3 \cdot 2/3 = E[X]E[Y] - 2/9 = 1/4 - 2/9 = 1/36$

3. **Correlation Coefficients [5]**
   Let X, Y, Z be jointly Gaussian zero–mean random variables such that X is conditionally independent of Z given Y. Given that the correlation coefficients of (X, Y) and (Y, Z) are ρ_1 and ρ_2, find the correlation coefficient of (X, Z). *Hint: The answer is in a fairly simple form; use the law of iterated expectation.*

   $$\mathbb{E}[XZ] = \mathbb{E}[\mathbb{E}[XZ|Y]] = \mathbb{E}[\mathbb{E}[X|Y]\mathbb{E}[Z|Y]] = \mathbb{E}[L[X|Y]L[Z|Y]]$$
   $$= \mathbb{E}[\frac{cov(X,Y)}{\sigma_Y^2}Y \cdot \frac{cov(Y,Z)}{\sigma_Y^2}Y] = \frac{cov(X,Y)}{\sigma_Y^2}\frac{cov(Y,Z)}{\sigma_Y^2}\mathbb{E}[Y^2] = \frac{cov(X,Y)cov(Y,Z)}{\sigma_Y^2}$$
   $$\rho = \frac{cov(X,Z)}{\sigma_X\sigma_Z} = \frac{cov(X,Y)}{\sigma_X\sigma_Y}\frac{cov(Y,Z)}{\sigma_Y\sigma_Z} = \rho_1\rho_2$$

   (Note: This result is true even when X,Y,Z are non-zero mean. You would have to handle a few more expectations.)

4. **Short Questions (Justify, no points for only answer.) [6]**

(a) True or False? For Zero Mean Jointly Gaussian RVs X, Y and Z, if $L(X|Y,Z) = L(X|Y) + L(X|Z)$, then Y and Z are independent RVs.

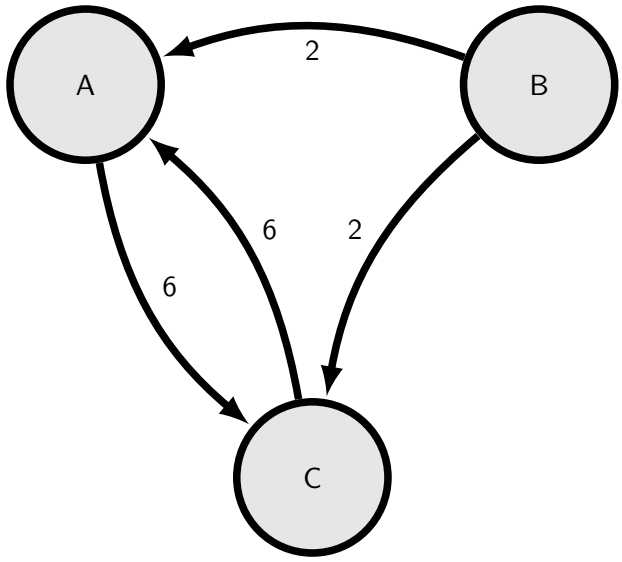
False, when Y and Z are correlated but X is independent of Y and Z. In this case, $L[X|Y,Z] = 0 = L[X|Y] + L[X|Z]$

Full credit was given for answering True, with the reasoning that Y and Z are orthogonal as the innovation is 0, implying they are uncorrelated and hence independent

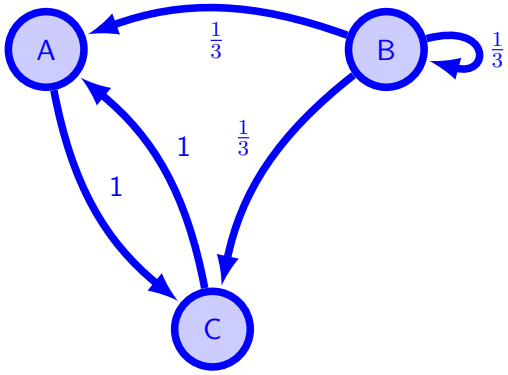
(b) If X is a Poisson Process of rate λ and has N arrivals in $(0,T)$, what is the joint distribution of the first N arrival times?

$f(t_1,\ldots,t_N|N(T)=N) = \frac{N!}{T^N}\mathbf{1}\{0 \le t_1 \le \cdots \le t_N \le T\}$

5. **CTMC [7]**
Consider the CTMC shown below. Write out the transition matrix, find the stationary distribution and then find the corresponding DTMC which has the same stationary distribution as this chain.



The rate matrix for this CTMC is $Q = \begin{bmatrix} -6 & 0 & 6 \\ 2 & -4 & 2 \\ 6 & 0 & -6 \end{bmatrix}$
For the stationary distribution, we solve $\pi Q = 0$ and $\sum_i \pi(i) = 1$ which gives us $\pi = [1/2 \quad 0 \quad 1/2]$ (this can also be seen from the symmetry of states A and C and the transience of state B).



Multiple answers. The DTMC above has the same stationary distribution.

6. **Deterministic Poisson Splitting [5]**

Customers arrive at a store in a Poisson process, $N(t), (t > 0)$, with rate λ. There are two queues, Q_1 and Q_2. Instead of random assignment to the queues, the first customer is deterministically assigned to Q_1, the next is assigned to Q_2, and so on; that is, the customers are assigned alternately to the two queues. Are the arrival processes to the individual queues Poisson? (If yes, provide the rate of the process. If no, show why not.)

The arrival process to an individual queue is not a Poisson process. Since every other arrival comes to a queue, the interarrival times for an individual queue is the sum of two $Exp(\lambda)$ random variables, i.e. it is Erlang of order 2.

7. **Petersburg Revisited [7]**

Recall the St. Petersburg "paradox" example from lecture. Formally, let X be a random variable representing the payoff from a random game such that for $i = 1, ..., P(X = 2^i) = \frac{1}{2^i}$. In lecture, we showed that $\mathbb{E}[X]$ is infinite, but this does not seem to be a reasonable way to model the "fair price" of the game. Here, we explore a different approach. Now, let X_k be i.i.d. realizations of this game at time step k. At each time step, according to some fixed $c \in \mathbb{R}$, define

$$S_n = \prod_{i=1}^{n} \frac{X_i}{c}$$

(a) What is the distribution of $\log_2 X_i$? (Specify parameters, if any. No justification needed.)

$P(\log_2 X_i = x) = P(X_i = 2^x) = \frac{1}{2^x}$. So $\log_2 X_i \sim Geom(\frac{1}{2})$

(b) Show that $\mathbb{E}[\log_2(X_i)] = 2, \forall i$. If needed, use without proof that $\sum_{i=1}^{\infty} \frac{i}{2^i} = 2$.

The expectation of a geometric r.v. is $1/p$. This follows immediately.

(c) Show that $\lim_{n\to\infty} \log_2(S_n)$ is either $-\infty$ or ∞, w.p. 1 according to if $c < c^*$ or $c > c^*$ for some fixed c^*. What is this value c^*?

$\lim_{n\to\infty} \log_2(S_n) = \lim_{n\to\infty} \sum_1^n \log_2(X_i) - \log_2 c = \lim_{n\to\infty} n(\frac{\sum X_i}{n} - \log_2 c)$.

Using the Strong Law of Large Numbers, the above limit is 0, with probability 1 when $\log_2 c = E[\log_2 X] = 2 \implies c^* = 4$.

We then have $\lim_{n\to\infty} \log_2(S_n) = \begin{cases} \infty, c < 4 \\ 0, c = 4 \\ -\infty, c > 4 \end{cases}$

8. **Independent Sum Entropy [5]**

Let $X_1, ..., X_n$ be independent random variables. Let $Y = \sum_{i=1}^{n} X_i$.

(a) Argue that $\forall i$, $H(Y) \geq H(X_i)$.

Intuitive argument: Let $Z = X_1 + \cdots + X_{i-1} + X_{i+1} + \cdots + X_n$. $Y = X_i + Z$ and since X_i and Z are independent, we can treat Z as noise. Independent noise increases the uncertainty (the number of values and how the probability is spread over these), which entropy is a measure of.

Proof:

$$\begin{aligned} H(X_1 + X_2 + ... + X_n) &\geq H(X_1 + X_2 + ... + X_n | X_1, ...X_{i-1}, X_{i+1}...X_n) \\ &= H(X_i | X_1, ..., X_{i-1}, X_{i+1}..., X_n) \\ &= H(X_i) \end{aligned}$$

Step 1 is true because the conditional entropy of a random variable conditioned on others is no more than the entropy of the random variable.
Step 2 is true because conditioned on $X_1, ..., X_{i-1}, X_{i+1}..., X_n$, $X_1 + X_2 + ... + X_n = c + X_i$ and adding a constant does not change entropy.
Step 3 is because X_i is independent of all the others.

Note: $H(Y) = H(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} H(X_i)$ is not true. The sum identity is for joint entropy, i.e. $H(X_1, \ldots, X_n) = \sum_{i=1}^{n} H(X_i)$.

Independence is very important. Suppose

$$X_1, X_2 = \begin{cases} (1,0) & \text{w.p. } \frac{1}{2} \\ (0,1) & \text{w.p. } \frac{1}{2} \end{cases}$$

X_1, X_2 are individually $B(\frac{1}{2})$, so $H(X_1) = 1$. But $X_1 + X_2$ is 1 w.p. 1, so it's entropy is 0.

(b) Give an example where $H(Y) = \sum_{i=1}^{n} H(X_i)$.

An example of equality is when all X_is are constants. In this case both sides are 0.

Another example is when

$$X_i = \begin{cases} 0 & \text{w.p. } \frac{1}{2} \\ 2^{i-1} & \text{w.p. } \frac{1}{2} \end{cases}$$

Each individual X_i has entropy 1. Furthermore $\sum X_i$ generates all 2^n bitstrings, and since they're uniformly probable, it's entropy is n.

Problem 2: Random Graphs and Markov Chains [20]

Assume $N \geq 3$ is a fixed positive integer and G_0 is a graph on N vertices $\{1, 2, \ldots, N\}$ with no edges (empty graph). At each time step $n \geq 1$, starting from the graph G_{n-1}, we pick a pair (i, j), $1 \leq i < j \leq N$ uniformly at random among the $\binom{N}{2}$ such pairs. Then, with probability $1-p$, we do nothing, and with probability p, we alter the edge between vertices i and j; that is, if there is an edge between i and j, we remove it and if there is not an edge, we place an edge. Here, $p \in (0, 1)$ is fixed. Let G_n be the resulting graph. We continue this process inductively, i.e. we generate G_1 from G_0, then G_2 from G_1, and so on.

1. Argue that $(G_n : n \geq 0)$ is a Markov Chain. What is the state space? What is the size of this state space? What are the transition probabilities?

   Note that G_{n+1} is constructed only based on G_n. Hence, conditioned on G_n, G_{n+1} is independent from $G_0, \ldots, G_{n-1}$. This means that $(G_n : n \geq 0)$ is a Markov Chain. The state space is the set of all simple graphs on n vertices, which has size $2^{\binom{N}{2}}$. If $P(G, G') = \mathbb{P}(G_{n+1} = G' | G_n = G)$ denotes the transition probability, we have $P(G, G) = 1 - p$ (since we do nothing with probability $1 - p$), and $P(G, G') = p/\binom{N}{2}$ if G' differs from G in only one edge.

2. Classify this chain in terms of its periodicity, reducibility, and whether it is transient, positive recurrent or null recurrent (Justify your answers for full credit).

   Since there is a loop at each state with probability $1 - p$ and $p \in (0, 1)$, the chain is aperiodic. Moreover, we can reach any graph G' starting with any graph G with nonzero probability. The reason is that we can simply go over the edges in G one by one and alter them if necessary so that we modify G into G'. Therefore, the chain is irreducible and all the states are recurrent.

3. What are the stationary distribution(s) of this Markov Chain? [Hint: Show that the Erdős–Rényi distribution $\mathcal{G}(N, q)$ with an appropriate value of q is the stationary distribution.]

Note that since the chain is irreducible, it has a unique stationary distribution. We claim that the Erdős–Rényi distribution $\mathcal{G}(N, q)$ with an appropriate value of q is the stationary distribution. Under this distribution, we have

$$\pi(G) = q^{n_G}(1-q)^{M-n_G},$$

where n_G is the number of edges in G and M is a shorthand for $\binom{N}{2}$. We should pick q to guarantee that $\pi(G) = \sum_{G'} P(G', G)\pi(G')$. As we have seen before, $P(G', G) \neq 0$ only when either $G = G'$, in which case $P(G', G) = 1-p$, or G' differs from G in only one edge, in which case $P(G', G) = p/M$. If G' is such that it differs from G in one of the n_G edges of G, we have

$$\pi(G') = q^{n_G-1}(1-q)^{M-n_G+1} = \frac{1-q}{q}\pi(G).$$

On the other hand, if G' differs from G in one of the $M - n_G$ edges that are not present in G, we have

$$\pi(G') = q^{n_G+1}(1-q)^{M-n_G-1} = \frac{q}{1-q}\pi(G).$$

Therefore, we must have

$$\pi(G) = (1-p)\pi(G) + n_G\frac{p}{M}\frac{1-q}{q}\pi(G) + (M-n_G)\frac{p}{M}\frac{q}{1-q}\pi(G).$$

We see that this identity is satisfied with $q = 1/2$. Hence, the Erdős–Rényi distribution $\mathcal{G}(N, 1/2)$ is the unique stationary distribution of this chain.

4. Assume that $N = 3$. Let T be the first time such that G_T is the complete graph (the graph on 3 vertices with all the 3 possible edges present). Find $\mathbb{E}(T)$.

Let a_G be the expected number of iterations it takes to reach the complete graph starting from G. We write first step equations for a_G. For a graph G, let n_G denote the number of edges in G. Moreover, let A_G denote the set of graphs which differ from G in only one edge. Then, the first step equations are:

$$a_G = 1 + (1-p)a_G + \sum_{G' \in A_G} \frac{p}{M} a_{G'},$$

and $a_H = 0$, where H denotes the complete graph. Note that there are two types of graphs in A_G: those with $n_G - 1$ many edges and those with $n_G + 1$ many edges. Due to the symmetry in choosing the edge to alter at each step, a_G only depends on n_G. Hence, let b_k for $0 \leq k \leq M = 3$ denote a_G for graphs G with $n_G = k$. We know that $b_3 = 0$ and we are interested in b_0. Rewriting the first step equations in terms of b_k, we get

$$\begin{aligned} b_0 &= 1 + (1-p)b_0 + pb_1 \\ b_k &= 1 + (1-p)b_k + \frac{kp}{3}b_{k-1} + \left(1 - \frac{k}{3}\right)pb_{k+1} \qquad \text{for } 1 \leq k \leq 2 \end{aligned}$$

From the first identity

$$b_0 = \frac{1}{p} + b_1.$$

For $k = 1$, $b_1 = 1 + (1-p)b_1 + \frac{p}{3}b_0 + \frac{2p}{3}b_2$. Substituting $b_0 = b_1 + 1/p$ and simplifying, we get

$$b_1 = b_2 + \frac{2}{p}.$$

For $k = 2$, we get $b_2 = 1 + (1-p)b_2 + \frac{2p}{3}b_1$. Substituting $b_1 = b_2 + 2/p$ and solving for b_2, we get $b_2 = 7/p$. Using this, we get $b_1 = 9/p$ and

$$b_0 = \frac{1}{p} + b_1 = \frac{10}{p}.$$

Thereby, $\mathbb{E}(T) = 10/p$.

Problem 3: MAP with Gaussians [10]

A disease has 2 strains, 0 and 1, which occur with prior probability p_0 and $p_1 = 1-p_0$ respectively. *For both parts of this problem, you are allowed to leave your answer in terms of $\Phi(x)$, the CDF of the standard normal distribution.*

1. A noisy test is developed to find which strain is present for patients with the disease. Let $X \in \{0, 1\}$ be the random variable which denotes the strain. The output of the test is a random variable Y_1, such that $Y_1 = 5 - 4X + Z_1$, where $Z_1 \sim \mathcal{N}(0, \sigma^2)$ and is independent of the strain X. Give a MAP decision rule to output $\hat{X}$, your best guess for X, given Y_1, and compute $\mathbb{P}(\hat{X} \neq 0 | X = 0)$.

   This is a case of binary detection with additive gaussian noise. The MAP rule is,

   $$\hat{X}_{MAP} = 1 \text{ if } \frac{P(Y_1|X=1)P(X=1)}{P(Y_1|X=0)P(X=0)}$$

   otherwise, $\hat{X}_{MAP} = 0$. Since $P(X = 0) = p_0$ and $P(X = 1) = p_1$, the above mentioned rule can be simplified as the following,

   $$\hat{X}_{MAP} = 1 \text{ if } y_1 \leq 3 - \frac{\sigma^2}{4}\log(p_0/p_1) \tag{1}$$

   otherwise $\hat{X}_{MAP} = 0$. Let y^* be the threshold on y_1, we have

   $$y^* = 3 - \frac{\sigma^2}{4}\log(p_0/p_1)$$

   Then,

   $$P(\hat{X}_{MLE} = 1|X = 0) = P(y_1 < y^*|X = 0) = 1 - Q(\frac{y^* - 5}{\sigma})$$

   $$P(\hat{X}_{MLE} = 0|X = 1) = P(y_1 \geq y^*|X = 1) = Q(\frac{y^* - 1}{\sigma})$$

2. A medical researcher proposes a new measurement procedure: he observes Y_1 as done previously, and in addition, "creates" a new measurement, $Y_2 = Y_1 + Z_2$. Assume $Z_2 \sim \mathcal{N}(0, \sigma^2)$ is independent of X and Z_1. Now, find the MAP rule in terms of the joint observation (y_1, y_2) and compute $\mathbb{P}(\hat{X} \neq 0|X = 0)$.

   Note that Y_2 is simlpy Y_1 plus noise and the noise component is independent of X and Y_1. Thus Y_2 conditioned on Y_1 and X is simply $\mathcal{N}(Y_1, \sigma^2)$ which does not depend on X. Thus Y_1 is sufficient and Y_2 is irrelevant. So the decision rule stays the same, so are the error probabilities.

Problem 4: Hypothesis Testing with Gaussians [12]
For this problem also, you may leave your answer in terms of the Guassian CDF $\Phi(x)$.

1. We are told that a random variable X is either $\mathcal{N}(0,1)$ (null hypothesis) or $\mathcal{N}(10,1)$ (alternate hypothesis). We want the probability of false alarm to be no more than 2.5%. What is the Neyman-Pearson optimal test? What is the probability of correct detection for this threshold?

   Let H be the random variable of the hypothesis which can take the values H_1, the alternate hypothesis, and H_0, the null hypothesis. We set up the likelihood ratio to determine how to set our threshold:

   $$L(x) = \frac{f_{X|H_1}(x|H=H_1)}{f_{X|H_0}(x|H=H_0)} = \frac{\frac{1}{\sqrt{2\pi}}e^{-(x-10)^2/2}}{\frac{1}{\sqrt{2\pi}}e^{-x^2/2}} = e^{10x-50}$$

   We see that the likelihood is monotonically increasing in x, so we will look for a threshold τ such that $\hat{H} = H_1$ if $x \geq \tau$. For this we use the threshold on the probability of false alarm.

   $$P(\hat{H} = H_1|H = H_0) = 0.025 \implies P(X \geq \tau|H = H_0) = 0.025$$

   Conditioned on $H = H_0$ we know $X \sim \mathcal{N}(0,1)$ so to achieve the bound we set $\tau = 1.96$.

   To find the probability of correct detection we just plug in our value of τ. Let $Z \sim \mathcal{N}(0,1)$.

   $$\begin{aligned} P(\hat{H} = H_1|H = H_1) &= P(X \geq 1.96|H = H_1) \\ &= P(X \geq 1.96|X \sim N(10,1)) \\ &= P(Z \geq 1.96 - 10) \\ &= \Phi(8.04) \end{aligned}$$

2. Now, we are told that a random variable Y is either $\mathcal{N}(0,1)$ (null hypothesis) or $\mathcal{N}(0,2)$ (alternate hypothesis). We want the probability of false alarm to be no more than 5%. What is the Neyman-Pearson optimal test? What is the probability of correct detection for this threshold?

   Let H be the random variable of the hypothesis which can take the values H_1, the alternate hypothesis, and H_0, the null hypothesis. We set up the likelihood ratio to determine how to set our threshold:

   $$L(y) = \frac{f_{Y|H_1}(y|H=H_1)}{f_{Y|H_0}(y|H=H_0} = \frac{\frac{1}{\sqrt{2\pi 2}}e^{-y^2/(2*2)}}{\frac{1}{\sqrt{2\pi}}e^{-y^2/2}} = \frac{1}{\sqrt{2}}e^{\frac{y^2}{4}}$$

   We see that the likelihood is monotonically increasing in $|y|$, so we will look for a threshold τ such that $\hat{H} = H_1$ if $|y| \geq \tau$. For this we use the threshold on the probability of false alarm.

   $$P(\hat{H} = H_1|H = H_0) = 0.05 \implies P(|X| \geq \tau|H = H_0) = 0.05$$

   Conditioned on $H = H_0$ we know $Y \sim \mathcal{N}(0,1)$ so to achieve the bound we set $\tau = 1.96$.

To find the probability of correct detection we just plug in our value of τ. Let $Z \sim \mathcal{N}(0,1)$.

$$\begin{aligned} P(\hat{H} = H_1 | H = H_1) &= P(|Y| \geq 1.96 | H = H_1) \\ &= P(|Y| \geq 1.96 | Y \sim N(0,2)) \\ &= P(|Z| \geq \frac{1.96}{\sqrt{2}}) \\ &= 2\Phi(-\frac{1.96}{\sqrt{2}}) \end{aligned}$$

Problem 5: Gaussian Product CLT [7]

Let $X_1, ..., X_n \overset{i.i.d.}{\sim} \text{Lognormal}(\mu, \sigma)$. Let $Y_k := (\Pi_{i=1}^{k} X_i)^{1/k}$. *Recall: if X is log-normally distributed, then $\ln(X)$ is $\mathcal{N}(\mu, \sigma)$.*

1. Find $\mathbb{E}[\ln(Y_k)]$.

   $\mathbb{E}[\ln(Y_k)] = \mathbb{E}[\frac{1}{k}\sum_{i=1}^{k} \ln X_i] = \frac{k}{k}\mu = \mu$

2. Find a lower bound on n such that $\mathbb{P}(|\ln(Y_n) - \mathbb{E}[\ln(Y_n)]| > 0.01) < 0.05$. You may leave your answer in terms of $\Phi(x)$, the normal CDF.

   $var(\ln(Y_n)) = var(\frac{1}{n}\sum \ln X_i) = \frac{\sigma^2}{n}$.

$$\begin{aligned}
\mathbb{P}(|\ln(Y_n) - \mathbb{E}[\ln(Y_n)]| > 0.01) &= \mathbb{P}(|\mathcal{N}(0, \frac{\sigma^2}{n})| > 0.01) \\
&= \mathbb{P}(|\mathcal{N}(0,1)| > 0.01\frac{\sqrt{n}}{\sigma}) < 0.05 \\
&\implies 2(1 - \Phi(0.01\frac{\sqrt{n}}{\sigma})) < 0.05 \\
&\implies 0.01\frac{\sqrt{n}}{\sigma} > 1.96 \\
&\implies n > 196^2\sigma^2
\end{aligned}$$

Problem 6: LLSE and Kalman Filter [20]

Consider a sensor network comprising n sensors that take noisy measurements of a temperature variable X as follows: $Y_i = X + W_i$, where $X \sim \mathcal{N}(0, 10)$ and W_i's are i.i.d. $\mathcal{N}(0, 1)$ that model the noise in the system.

1. Let $\hat{X}_{LLSE} = \alpha_1 Y_1 + \alpha_2 Y_2 + \ldots + \alpha_n Y_n$. Find α_i for $i = \{1, \ldots, n\}$. *(Hint: Do it for $n = 2$ first and then generalize).*

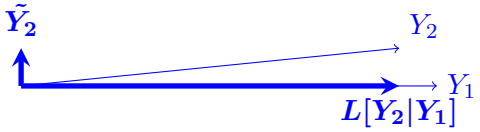
   Solution 1:
   By symmetry, each α_i must be equal ($\alpha_1 = \alpha_2 = \ldots \alpha_n = \alpha$). So $\hat{X}_{LLSE} = \alpha(nX + \sum_i W_i)$. To calculate α, note that $X - \hat{X}_{LLSE}$ must be orthogonal to each Y_i. Choosing Y_1 we have,

$$\begin{aligned}
0 &= E[(X - \alpha(nX + \sum_i W_i))(X + W_1)] \\
&= E[X^2] - \alpha n E[X^2] - \alpha \cancelto{0}{E[(\sum W_i)X]} \\
&+ \cancelto{0}{E[XW_1]} - \alpha n \cancelto{0}{E[XW_1]} - \alpha E[(\sum W_i)W_1] \\
&= E[X^2] - \alpha n E[X^2] - \alpha E[W_1^2] \\
&= 10 - (10n + 1)\alpha \\
\implies \alpha &= \frac{10}{10n + 1}
\end{aligned}$$

   Solution 2:
   If you calculated $\hat{X}_{1|1}$ and $\hat{X}_{2|2}$ below you may have guessed the form. This can then be shown by induction using the Kalman filter equations.

2. Suppose $n = 2$. I want to form $L[X|Y_1, Y_2]$ in an online fashion by first considering Y_1 and then Y_2 as follows:
$$L[X|Y_1, Y_2] = L[X|Y_1] + L[X|\tilde{Y}_2].$$
   What is $\tilde{Y}_2$? Draw a geometric picture relating Y_1, Y_2 and $\tilde{Y}_2$.

   The innovation $\tilde{Y}_2 = Y_2 - L[Y_2|Y_1] = Y_2 - \frac{cov(X+W_2, X+W_1)}{var(X+W_1)} Y_1 = Y_2 - \frac{10}{11} Y_1$



   $\cos\theta = \rho(Y_2, Y_1) = 10/11$

3. Now I want to estimate X recursively by taking the measurements $Y_1, Y_2, \ldots, Y_n$ in an online fashion using a Kalman Filter based approach. Note that the state-space equations degenerate to:
$$X_n = X_{n-1},$$

$$Y_n = X_n + W_n$$

We will use the usual notation seen in lecture. $\hat{X}_{n|n}$ is the best estimate of X_n given $Y_1, Y_2, \ldots, Y_n$. $\hat{X}_{n|n-1}$ is the best estimate of X_n given $Y_1, Y_2, ...Y_{n-1}$ and $\sigma^2_{n|n} = E((X_n - \hat{X}_{n|n})^2)$, etc.

Suppose I initialize $\hat{X}_{1|0} = 0$ and $\sigma^2_{1|0} = 10$ (i.e, variance of X) in the Kalman equations:

$$\hat{X}_{n|n} = \hat{X}_{n|n-1} + k_n(Y_n - \hat{X}_{n|n-1})$$

$$k_n = \frac{\sigma^2_{n|n-1}}{\sigma^2_{n|n-1} + \sigma^2_w}$$

$$\sigma^2_{n|n} = \sigma^2_{n|n-1}(1 - k_n)$$

(a) What are $\hat{X}_{1|1}, \sigma^2_{1|1}, \hat{X}_{2|2}$, and $\sigma^2_{2|2}$?

Using a), $X_{1|1} = \frac{10}{11}Y_1$ and $X_{2|2} = \frac{10}{21}(Y_1 + Y_2)$.

$$\sigma^2_{n|n} = E[(X - \alpha(nX + \sum_i W_i))^2] = E[(X(1 - \alpha n) - \alpha \sum_i W_i)^2]$$

$$= (1 - \alpha n)^2 var(X) + \alpha^2 var(\sum W_i) - 2cov(X, \sum_i W_i)^{0}$$

$$= n\alpha^2 + 10(1 - n\alpha)^2 = \frac{100n}{(10n+1)^2} + 10\frac{1}{(10n+1)^2} = \frac{10}{10n+1}$$

So $\sigma^2_{1|1} = \frac{10}{11}$ and $\sigma^2_{2|2} = \frac{10}{21}$

(b) In the limit as $n \to \infty$, what are k_n and $\sigma^2_{n|n}$?

Since $X_n = X_{n-1}$, $\sigma^2_{n|n-1} = \sigma^2_{n-1|n-1}$. So $k_n = \frac{\frac{10}{10(n-1)+1}}{\frac{10}{10(n-1)+1}+1} = \frac{10}{10n+1}$. We can see that in the limit as $n \to \infty$, both k_n and $\sigma_{n|n}$ go to 0

Problem 7: HMMs and EM [20]

1. There are two identical-looking coins A and B whose biases (probability of Heads) are $\theta_A = 0.4$ and $\theta_B = 0.8$ respectively. Let X_k be the coin at time step k. A Markov Chain with transition probabilities given below describes the coin-picking process: $P(X_{k+1} = A|X_k = B) = 0.2$, $P(X_{k+1} = B|X_k = A) = 0.3$ for $k = \{0, 1. \ldots, \}$. Now, let the initial state X_0 be A. At each time step, we observe the result of flipping the current coin (without knowing which coin it was). The observed sequence of tosses is H,T,T.

   (a) What is the most likely sequence of coin labels picked?

   (A,A,A). We are given that the state is A at time step 0, so the first coin in the sequence is trivially A. Both the transition probability and the tails probability are individually maximized when the next state is A for the subsequent states in the sequence, so the second and third coins are also A

   (b) What is the most likely coin label corresponding to the *second* toss? Is it consistent with the answer in (a)? Does it need to be? Explain.

   The most likely coin label corresponding to the second toss is the coin with the higher probability of tails, i.e, A. It is consistent with a). However, this does not need to be consistent with a) as we are calculating the maximum likelihood estimate of a single toss and not the maximum likelihood estimate of the full sequence of tosses.

2. Now suppose that you do not know the true biases of the two coins and want to estimate them. At each time step, you pick one of the two coins equally at random and toss it once and observe whether it is Heads or Tails. You then replace the coin and repeat the experiment 5 times. Suppose you observe H, T, T, H, H.

   (a) Using the Hard EM algorithm with initial guess $\theta_A = 0.4, \theta_B = 0.8$, what will be your converged estimates of the biases of the coins?

   First iteration - since we flip only one coin each experiment, the coin assignment is just B if you see a head and A if T. The MLE estimates based on just one flip are 1 and 0 for $\hat{\theta_B}^{(1)}$ and $\hat{\theta_A}^{(1)}$ respectively. We see that convergence has been reached since on the next iteration all heads are once again assigned to B and tails to A giving the same estimates of 0 and 1 for $\hat{\theta_A}^{(2)}$ and $\hat{\theta_B}^{(2)}$. The estimates with Hard E.M. are thus A yields tails w.p. 1 and B yields heads w.p. 1.

   (b) Now you use the Soft EM algorithm with the same initial guesses. What will be the estimates for θ_A, θ_B after one iteration?

   E-step:
   First we calculate $P(A|H)$ and $P(A|T)$.

$$P(A|H) = \frac{P(H|A)P(A)}{P(H|A)P(A) + P(H|B)P(B} = \frac{0.4}{0.4 + 0.8} = 1/3$$

   Similarly, we get $3/4$ for $P(A|T)$, $2/3$ for $P(B|H)$ and $1/4$ for $P(B|T)$. Weighting the observations with these probabilities, we have $3 \cdot 1/3 = 1H$ and $2 \cdot 3/4 = 1.5T$s for A and $3 \cdot 2/3 = 2H$s and $2 \cdot 1/4 = 1/2$T for B.

   M-step:
   Finding the MLE, we have the following estimates after one iteration -

$$\hat{\theta_A}^{(1)} = \frac{1}{1 + 1.5} = 0.4$$

$$\hat{\theta_B}^{(1)} = \frac{2}{2 + \frac{1}{2}} = 0.8$$