Quiz Section 9 – Solutions

Review

1) 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).

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} \mathcal{L}(x_1, \ldots, x_n \mid \theta) = \arg\max_{\theta} \ln \mathcal{L}(x_1, \ldots, x_n \mid \theta)$$

2) An estimator $\hat{\theta}$ for a parameter θ of a probability distribution is **unbiased** iff $\mathbb{E}[\hat{\theta}(X_1, \ldots, X_n)] = \theta$

3) A **discrete-time stochastic process (DTSP)** is a sequence of random variables $X^{(0)}, X^{(1)}, X^{(2)}, \ldots$, where $X^{(t)}$ is the value at time t. For example, the temperature in Seattle or stock price of TESLA each day, or which node you are at after each time step on a random walk on a graph.

4) Markov Chain is a DTSP, with the additional following three properties:

(a) ...has a finite (or countably infinite) **state space** $\mathcal{S} = \{s_1, \ldots, s_n\}$ which it bounces between, so each $X^{(t)} \in \mathcal{S}$.

(b) ...satisfies the **Markov property**. A DTSP satisfies the Markov property if the future is (conditionally) independent of the past given the present. Mathematically, it means,

$$\mathbb{P}\left(X^{(t+1)} = x_{t+1} \mid X^{(0)} = x_0, X^{(1)} = x_1, \ldots, X^{(t-1)} = x_{t-1}, X^{(t)} = x_t\right) = \mathbb{P}\left(X^{(t+1)} = x_{t+1} \mid X^{(t)} = x_t\right).$$

(c) ...has **fixed transition probabilities**. Meaning, if we are at some state s_i, we transition to another state s_j with probability *independent* of the current time. Due to this property and the previous, the transitions are governed by n^2 probabilities: the probability of transitioning from one of n current states to one of n next states. These are stored in a square $n \times n$ **transition probability matrix (TPM)** $\mathbf{M}$, where $M_{ij} = \mathbb{P}\left(X^{(t+1)} = s_j \mid X^{(t)} = s_i\right)$ is the probability of transitioning from state s_i to state s_j for any/every value of t.

5) A **stationary distribution** of a Markov chain is a probability distribution on states that is unchanged by taking one step of the Markov chain.

Task 1 – Mystery Dish!

A fancy new restaurant has opened up that features only 4 dishes. The unique feature of dining here is that they will serve you any of the four dishes randomly according to the following probability distribution: give dish A with probability 0.5, dish B with probability θ, dish C with probability 2θ, and dish D with probability $0.5 - 3\theta$. Each diner is served a dish independently. Let x_A be the number of people who received dish A, x_B the number of people who received dish B, etc, where $x_A + x_B + x_C + x_D = n$. Find the MLE for θ, $\hat{\theta}$.

The data tells us, for each diner in the restaurant, what their dish was. We begin by computing the likelihood of seeing the given data given our parameter θ. Because each diner is assigned a dish independently, the likelihood is equal to the product over diners of the chance they got the particular dish they got, which gives us:

$$\mathcal{L}(x \mid \theta) = 0.5^{x_A} \theta^{x_B} (2\theta)^{x_C} (0.5 - 3\theta)^{x_D}$$

From there, we just use the MLE process to get the log-likelihood, take the first derivative, set it equal to 0, and solve for $\hat{\theta}$.

$$\ln \mathcal{L}(x \mid \theta) = x_A \ln(0.5) + x_B \ln(\theta) + x_C \ln(2\theta) + x_D \ln(0.5 - 3\theta)$$

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \ln \mathcal{L}(x \mid \theta) = \frac{x_B}{\theta} + \frac{x_C}{\theta} - \frac{3x_D}{0.5 - 3\theta}$$

$$\frac{x_B}{\hat{\theta}} + \frac{x_C}{\hat{\theta}} - \frac{3x_D}{0.5 - 3\hat{\theta}} = 0$$

Solving yields $\hat{\theta} = \frac{x_B + x_C}{6(x_B + x_C + x_D)}$.

Task 2 – 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. In other words, they follow the distributions $\mathbb{P}(k; \theta) = \theta^k e^{-\theta}/k!$, where $k \in \mathbb{N}$ and $\theta > 0$ is a positive real number.

Find the MLE of θ.

We follow the recipe given in class:

$$\begin{aligned}
\mathcal{L}(x_1, \ldots, x_n \mid \theta) &= \prod_{i=1}^{n} e^{-\theta} \frac{\theta^{x_i}}{x_i!} \\
\ln \mathcal{L}(x_1, \ldots, x_n \mid \theta) &= \sum_{i=1}^{n} \left[-\theta - \ln(x_i!) + x_i \ln(\theta)\right] \\
\frac{\mathrm{d}}{\mathrm{d}\theta} \ln \mathcal{L}(x_1, \ldots, x_n \mid \theta) &= \sum_{i=1}^{n} \left[-1 + \frac{x_i}{\theta}\right] \\
-n + \frac{\Sigma_{i=1}^{n} x_i}{\hat{\theta}} &= 0 \\
\hat{\theta} &= \frac{\Sigma_{i=1}^{n} x_i}{n}
\end{aligned}$$

Task 3 – 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?

Let $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$. Then

$$\mathbb{E}\left[\hat{\theta}_2\right] = \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})^2\right] = \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^{n} (X_i^2 - 2X_i\overline{X} + \overline{X}^2)\right]$$

which by linearity of expectation (and distributing the sum) is

$$\begin{aligned}
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[X_i^2\right] - \mathbb{E}\left[\frac{2}{n}\overline{X}\sum_{i=1}^{n}X_i\right] + \mathbb{E}\left[\overline{X}^2\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[X_i^2\right] - 2\mathbb{E}\left[\overline{X}^2\right] + \mathbb{E}\left[\overline{X}^2\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[X_i^2\right] - \mathbb{E}\left[\overline{X}^2\right] . \qquad (**)
\end{aligned}$$

We know that for any random variable Y, since $\mathrm{Var}\,(Y) = \mathbb{E}\left[Y^2\right] - (\mathbb{E}\left[Y\right])^2$ it holds that

$$\mathbb{E}\left[Y^2\right] = \mathrm{Var}\,(Y) + (\mathbb{E}\left[Y\right])^2.$$

Also, we have $\mathbb{E}\left[X_i\right] = \mu,\ \mathrm{Var}\,(X_i) = \sigma^2\ \forall i$ and $\mathbb{E}\left[\overline{X}\right] = \mu,\ \mathrm{Var}\left(\overline{X}\right) = \frac{\sigma^2}{n}$. Combining these facts, we get

$$\mathbb{E}\left[X_i^2\right] = \sigma^2 + \mu^2 \quad \forall i \quad \text{and} \quad \mathbb{E}\left[\overline{X}^2\right] = \frac{\sigma^2}{n} + \mu^2.$$

Substituting these equations into (**) we get

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2)\right] &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[X_i^2\right] - \mathbb{E}\left[\overline{X}^2\right] = \sigma^2 + \mu^2 - \left(\frac{\sigma^2}{n} + \mu^2\right) \\
&= \left(1 - \frac{1}{n}\right)\sigma^2.
\end{aligned}$$

Thus $\hat{\theta}_2$ is not unbiased.

Task 4 – Weather Forecast

A weather forecaster predicts sun with probability θ_1, clouds with probability $\theta_2 - \theta_1$, rain with probability $\frac{1}{2}$ and snow with probability $\frac{1}{2} - \theta_2$. This year, there have been 55 sunny days, 100 cloudy days, 160 rainy days and 50 snowy days. What is the maximum likelihood estimator for θ_1 and θ_2?

We want to find the likelihood of the data samples given the parameter θ. To do this, we take the following product over all the data points.

$$\mathcal{L}(x_1, ..., x_{365} \mid \theta_1, \theta_2) = \theta_1^{55}(\theta_2 - \theta_1)^{100}\left(\frac{1}{2}\right)^{160}\left(\frac{1}{2} - \theta_2\right)^{50}$$

Then, we use this to determine the log likelihood.

$$\begin{aligned}
\ln\mathcal{L}(x_1, ..., x_{365} \mid \theta_1, \theta_2) &= \ln\theta_1^{55}(\theta_2 - \theta_1)^{100}\left(\frac{1}{2}\right)^{160}\left(\frac{1}{2} - \theta_2\right)^{50} \\
&= \ln\theta_1^{55} + \ln(\theta_2 - \theta_1)^{100} + \ln\left(\frac{1}{2}\right)^{160} + \ln\left(\frac{1}{2} - \theta_2\right)^{50} \\
&= 55\ln\theta_1 + 100\ln(\theta_2 - \theta_1) + 160\ln\left(\frac{1}{2}\right) + 50\ln\left(\frac{1}{2} - \theta_2\right)
\end{aligned}$$

Then, we take the derivative of the log likelihood with respect to θ_1.

$$\frac{\partial}{\partial \theta_1} \ln \mathcal{L}(x_1, ..., x_{365} \mid \theta_1, \theta_2) = \frac{55}{\theta_1} - \frac{100}{\theta_2 - \theta_1}$$

Setting this equal to 0, we solve for $\hat{\theta_1}$:

$$\frac{55}{\hat{\theta_1}} - \frac{100}{\hat{\theta_2} - \hat{\theta_1}} = 0$$

$$55(\hat{\theta_2} - \hat{\theta_1}) - 100\ \hat{\theta_1} = 0$$

$$55\ \hat{\theta_2} = 155\ \hat{\theta_1}$$

$$\hat{\theta_1} = \frac{11}{31}\ \hat{\theta_2}$$

Then, we take the derivative of the log likelihood with respect to θ_2.

$$\frac{\partial}{\partial \theta_2} \ln \mathcal{L}(x_1, ..., x_{365} \mid \theta_1, \theta_2) = \frac{100}{\theta_2 - \theta_1} - \frac{50}{\frac{1}{2} - \theta_2}$$

Setting this equal to 0, we solve for $\hat{\theta_2}$:

$$\frac{100}{\hat{\theta_2} - \hat{\theta_1}} - \frac{50}{\frac{1}{2} - \hat{\theta_2}} = 0$$

$$100\left(\frac{1}{2} - \hat{\theta_2}\right) - 50\ (\hat{\theta_2} - \hat{\theta_1}) = 0$$

$$50 - 150\ \hat{\theta_2} + 50\ \hat{\theta_1} = 0$$

$$\hat{\theta_2} = \frac{\hat{\theta_1} + 1}{3}$$

We can now solve the simultaneous equations we have for θ_1 and θ_2 to obtain the maximum likelihood estimators for each parameter.

$$\hat{\theta_2} = \frac{\hat{\theta_1} + 1}{3}$$

Plugging in the equation for θ_1, we find

$$\hat{\theta_2} = \frac{\frac{11}{31}\ \hat{\theta_2} + 1}{3}$$

$$3\ \hat{\theta_2} = \frac{11}{31}\ \hat{\theta_2} + 1$$

$$93\ \hat{\theta_2} = 11\ \hat{\theta_2} + 31$$

$$\hat{\theta_2} = \frac{31}{82}$$

Plugging in the value for θ_2 into the equation for θ_1,

$$\hat{\theta_1} = \frac{11}{31} \cdot \frac{31}{82} = \frac{11}{82}$$

To confirm that this is in fact a maximum, we could do a second derivative test. We won't ask you do this for this multivariate case, but it would still be good to check!

Task 5 – Faulty Machines

You are trying to use a machine that only works on some days. If on a given day, the machine is working it will break down the next day with probability $0 < b < 1$, and works on the next day with probability $1 - b$. If it is not working on a given day, it will work on the next day with probability $0 < r < 1$ and not work the next day with probability $1 - r$.

a) In this problem we will formulate this process as a Markov chain. First, let $X^{(t)}$ be a variable that denotes the state of the machine at time t. Then, define a state space $\mathcal{S}$ that includes all the possible states that the machine can be in. Lastly, for all $A, B \in \mathcal{S}$ find $\mathbb{P}(X^{(t+1)} = A \mid X^{(t)} = B)$ (A and B can be the same state).

Formally, a Markov chain is defined by a state space $\mathcal{S}$ and a transition probability matrix. The two possible states of the machine are "working" and "broken". So, $\mathcal{S} = \{W, B\}$. Let X_t be the state of the process at time t. Then we can define the following transition probabilities:

$$\mathbb{P}(X^{(t+1)} = W \mid X^{(t)} = W) = 1 - b \qquad \mathbb{P}(X^{(t+1)} = B \mid X^{(t)} = W) = b$$

$$\mathbb{P}(X^{(t+1)} = W \mid X^{(t)} = B) = r \qquad \mathbb{P}(X^{t+1} = B \mid X^{(t)} = B) = 1 - r$$

We can also represent the transition probabilities with the following matrix:

$$M = \begin{bmatrix} 1-b & b \\ r & 1-r \end{bmatrix}$$

where the ij-th entry is probability that the machine is in the j-th state at time $t+1$ given it was in state i at time t. (Here state 1 is working and state 2 is broken.)

b) Suppose that on day 1, the machine is working. What is the probability that it is working on day 3?

We are trying to find $\mathbb{P}(X^{(3)} = W \mid X^{(1)} = W)$. From the law of total probability, and then plugging in the values from our transition matrix:

$$\begin{aligned}
&\mathbb{P}(X^{(3)} = W \mid X^{(1)} = W) \\
&= \sum_{i \in \mathcal{S}} \mathbb{P}(X^{(3)} = W \mid X^{(1)} = W, X^{(2)} = i) \cdot \mathbb{P}(X^{(2)} = i \mid X^{(1)} = W) \\
&= \mathbb{P}(X^{(3)} = W \mid X^{(2)} = W) \cdot \mathbb{P}(X^{(2)} = W \mid X^{(1)} = W) \\
&+ \mathbb{P}(X^{(3)} = W \mid X^{(2)} = B) \cdot \mathbb{P}(X^{(2)} = B \mid X^{(1)} = W) \\
&= \mathbb{P}(X^{(3)} = W \mid X^{(2)} = W) \cdot (1-b) + \mathbb{P}(X^{(3)} = W \mid X^{(2)} = B) \cdot b \\
&= (1-b)(1-b) + rb \\
&= (1-b)^2 + rb
\end{aligned}$$

Alternative solution using matrix operations: Let $q^{(t)}$ be the probability vector at time t associated with this Markov chain. The assumption that the machine is working on day 1 is the same as saying that the probability vector $q^{(1)} = [1, 0]$. Then

$$q^{(2)} = q^{(1)} \cdot M = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} 1-b & b \\ r & 1-r \end{bmatrix} = \begin{bmatrix} 1-b & b \end{bmatrix}.$$

The probability we want to compute is the 1st entry of

$$q^{(3)} = q^{(2)} \cdot M = \begin{bmatrix} 1-b & b \end{bmatrix} \begin{bmatrix} 1-b & b \\ r & 1-r \end{bmatrix}$$

which equals $(1-b) \cdot (1-b) + b \cdot r = (1-b)^2 + br$.

c) As $n \to \infty$, what does the probability that the machine is working on day n converge to? To get the answer, solve for the *stationary distribution*.

The stationary distribution is the row vector $\pi = [\pi_W \ \pi_B]$ such that $\pi P = \pi$. The entries in the vector π_W and π_B can be interpreted as the probabilities that the machine works or is broken converge to. As such, $\pi_W + \pi_B = 1$. Additionally, multiplying the stationary distribution by the TPM gives us the following two equations (one per column of M):

$$\pi_W = \pi_W(1-b) + \pi_B\, r \qquad \pi_B = \pi_W b + \pi_B(1-r)$$

Solving these 3 equations for π_W and π_B gives us the following solutions for the stationary distribution:

$$\pi_W = \frac{r}{b+r} \qquad \pi_B = \frac{b}{b+r}$$

So, as $n \to \infty$ the probability that the machine works on day n is $\pi_W = \frac{r}{b+r}$

Task 6 – Another Markov Chain

Suppose that the following is the transition probability matrix for a 4 state Markov chain (states 1,2,3,4).

$$M = \begin{bmatrix} 0 & 1/2 & 1/2 & 0 \\ 1/3 & 0 & 0 & 2/3 \\ 1/3 & 1/3 & 0 & 1/3 \\ 1/5 & 2/5 & 2/5 & 0 \end{bmatrix}$$

a) What is the probability that $X^{(2)} = 4$ given that $X^{(0)} = 4$?

Let's denote the state space $\mathcal{S} = \{1, 2, 3, 4\}$. Using the law of total probability we can determine that

$$\begin{aligned} \mathbb{P}(X^{(2)} = 4 \mid X^{(0)} = 4) &= \sum_{i \in \mathcal{S}} \mathbb{P}(X^{(2)} = 4 \mid X^{(0)} = 4, X^{(1)} = i)\, \mathbb{P}(X^{(1)} = i \mid X^{(0)} = 4) \\ &= \sum_{i \in \mathcal{S}} \mathbb{P}(X^{(2)} = 4 \mid X^{(1)} = i)\, \mathbb{P}(X^{(1)} = i \mid X^{(0)} = 4) \\ &= 0 + \frac{2}{5} \cdot \frac{2}{3} + \frac{2}{5} \cdot \frac{1}{3} + 0 \\ &= \frac{2}{5} \end{aligned}$$

Alternative solution using matrix operations: Let $q^{(t)}$ be the probability vector at time t associated with this Markov chain. The statement that $X^{(0)} = 4$ is equivalent to saying that the probability vector $q^{(0)} = [0, 0, 0, 1]$. Therefore

$$q^{(1)} = q^{(0)} \cdot M = \begin{bmatrix} 0 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 0 & 1/2 & 1/2 & 0 \\ 1/3 & 0 & 0 & 2/3 \\ 1/3 & 1/3 & 0 & 1/3 \\ 1/5 & 2/5 & 2/5 & 0 \end{bmatrix} = \begin{bmatrix} 1/5 & 2/5 & 2/5 & 0 \end{bmatrix}.$$

What we want is the 4-th entry of

$$q^{(2)} = q^{(1)} \cdot M = \begin{bmatrix} 1/5 & 2/5 & 2/5 & 0 \end{bmatrix} \cdot \begin{bmatrix} 0 & 1/2 & 1/2 & 0 \\ 1/3 & 0 & 0 & 2/3 \\ 1/3 & 1/3 & 0 & 1/3 \\ 1/5 & 2/5 & 2/5 & 0 \end{bmatrix}$$

This is $0 + \frac{2}{5} \cdot \frac{2}{3} + \frac{2}{5} \cdot \frac{1}{3} + 0 = 2/5$.

b) Write down the system of equations that the stationary distribution must satisfy and solve them.

The stationary distribution is the row vector $\pi = [\pi_1\ \pi_2\ \pi_3\ \pi_4]$ such that $\pi P = \pi$. We know that $\pi_1 + \pi_2 + \pi_3 + \pi_4 = 1$. Additionally, multiplying the stationary distribution by the TPM gives us the following equations:

$$\begin{aligned}
\pi_1 &= \frac{1}{3}\pi_2 + \frac{1}{3}\pi_3 + \frac{1}{5}\pi_4 \\
\pi_2 &= \frac{1}{2}\pi_1 + \frac{1}{3}\pi_3 + \frac{2}{5}\pi_4 \\
\pi_3 &= \frac{1}{2}\pi_1 + \frac{2}{5}\pi_4 \\
\pi_4 &= \frac{2}{3}\pi_2 + \frac{1}{3}\pi_3
\end{aligned}$$

Solving these 5 equations for each π_i gives us the following solutions for the stationary distribution:

$$\pi_1 = \frac{46}{206} \qquad \pi_2 = \frac{60}{206} \qquad \pi_3 = \frac{45}{206} \qquad \pi_4 = \frac{55}{206}$$

Task 7 – Three Tails

You flip a fair coin until you see three tails in a row. Model this as a Markov chain with the following states:

- S: start state, which we are only in before flipping any coins.
- H: We see a heads, which means no streak of tails currently exists.
- T: We've seen exactly one tail in a row so far.
- TT: We've seen exactly two tails in a row so far.
- TTT: We've accomplished our goal of seeing three tails in a row, stop flipping, and stay there.

a) Write down the transition probability matrix.

$$M = \begin{bmatrix} 0 & 1/2 & 1/2 & 0 & 0 \\ 0 & 1/2 & 1/2 & 0 & 0 \\ 0 & 1/2 & 0 & 1/2 & 0 \\ 0 & 1/2 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

b) Write down the system of equations whose variables are $D(s)$ for each state $s \in \{S, H, T, TT, TTT\}$, where $D(s)$ is the expected number of steps until state TTT is reached starting from state s. Solve this system of equations to find $D(S)$.

Using the law of total expectation and the transition probability matrix above we can set up and solve the following system of equations:

$$\begin{aligned}
D(TTT) &= 0 \\
D(TT) &= 1 + \frac{1}{2}D(H) + \frac{1}{2}D(TTT) = \frac{1}{2}D(H) + 1 \\
D(T) &= 1 + \frac{1}{2}D(H) + \frac{1}{2}D(TT) = \frac{3}{4}D(H) + \frac{3}{2} \\
D(H) &= 1 + \frac{1}{2}D(H) + \frac{1}{2}D(T) = \frac{7}{8}D(H) + \frac{7}{4} \\
D(S) &= 1 + \frac{1}{2}D(H) + \frac{1}{2}D(T) = \frac{7}{8}D(H) + \frac{7}{4}
\end{aligned}$$

Solving for $D(H)$ gives us that $D(H) = 14$, which allows as to solve for the rest of the expected number of steps, $D(TT) = 8$, $D(T) = 12$, $D(S) = 14$. So, we expect to flip 14 coins before we flip three tails in a row.

c) Write down the system of equations whose variables are $\gamma(s)$ for each state $s \in \{S, H, T, TT, TTT\}$, where $\gamma(s)$ is the expected number of heads seen before state TTT is reached. Solve this system to find $\gamma(S)$, the expected number of heads seen overall until getting three tails in a row.

Like in the previous part we can use the LoTE and the Transition Probability Matrix to set up and solve the following system of equations. We get one equation for each column of M:

$$\begin{aligned}
\gamma(TTT) &= 0 \\
\gamma(TT) &= 0.5\gamma(H) + 0.5\gamma(TTT) = 0.5\gamma(H) \\
\gamma(T) &= 0.5\gamma(H) + 0.5\gamma(TT) = 0.75\gamma(H) \\
\gamma(H) &= 1 + 0.5\gamma(H) + 0.5\gamma(T) = 0.875\gamma(H) + 1 \\
\gamma(S) &= 0.5\gamma(H) + 0.5\gamma(T) = 0.875\gamma(H)
\end{aligned}$$

Solving for $\gamma(H)$ gives us $\gamma(H) = 8$. This allows us to solve for the other expected values which are $\gamma(TT) = 4$, $\gamma(T) = 6$, $\gamma(S) = 7$. So, we expect to see 7 heads before we flip three tails in a row.