Section 8: Solutions

Review of Main Concepts

- **Realization/Sample**: A realization/sample x of a random variable X is the value that is actually observed.
- **Likelihood**: Let $x_1, \dots x_n$ be iid realizations from 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.

  If X is discrete:
  $$L\left(x_1, \dots, x_n \mid \theta\right) = \prod_{i=1}^{n} p_X\left(x_i \mid \theta\right)$$
  If X is continuous:
  $$L\left(x_1, \dots, x_n \mid \theta\right) = \prod_{i=1}^{n} f_X\left(x_i \mid \theta\right)$$
- **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).
  $$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} L\left(x_1, \dots, x_n \mid \theta\right) = \arg\max_{\theta} \ln L\left(x_1, \dots, x_n \mid \theta\right)$$
- **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:
  $$\ln L\left(x_1, \dots, x_n \mid \theta\right) = \sum_{i=1}^{n} \ln p_X\left(x_i \mid \theta\right)$$
  If X is continuous:
  $$\ln L\left(x_1, \dots, x_n \mid \theta\right) = \sum_{i=1}^{n} \ln f_X\left(x_i \mid \theta\right)$$
- **Bias**: The bias of an estimator $\hat{\theta}$ for a true parameter θ is defined as $\text{Bias}\left(\hat{\theta}, \theta\right) = \mathbb{E}[\hat{\theta}] - \theta$. An estimator $\hat{\theta}$ of θ is unbiased iff $\text{Bias}\left(\hat{\theta}, \theta\right) = 0$, or equivalently $\mathbb{E}[\hat{\theta}] = \theta$.
- **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 and set it to 0 to find a candidate for the MLE, $\hat{\theta}$.
  (c) 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.

1. Mystery Dish!

A fancy new restaurant has opened up which 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, $\hat{\theta}$. **Solution:**

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:

$$L(x|\theta) = \binom{n}{x_A}\binom{n-x_A}{x_B}\binom{n-x_A-x_B}{x_C}0.5^{x_A}\theta^{x_B}(2\theta)^{x_C}(0.5-3\theta)^{x_D}$$

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 $\hat{\theta}$.

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

$$\frac{\partial}{\partial\theta}\ln L(x|\theta) = \frac{x_B}{\theta} + \frac{x_C}{\theta} - \frac{3x_D}{0.5-3\theta}$$

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

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

2. A Red Poisson

Suppose that $x_1,\dots,x_n$ are i.i.d. samples from a Poisson(θ) random variable, where θ is unknown. Find the MLE of θ. **Solution:**

Because each Poisson RV is i.i.d., the likelihood of seeing that data is just the PMF of the Poisson distribution multiplied together for every x_i. From there, take the log-likelihood, then the first derivative, set it equal to 0 and solve for for $\hat{\theta}$.

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

3. Independent Shreds, You Say?

You are given 100 independent samples $x_1, x_2, \dots, x_{100}$ from Bernoulli(θ), where θ is unknown. (Each sample is either a 0 or a 1). These 100 samples sum to 30. You would like to estimate the distribution's parameter θ. Give all answers to 3 significant digits.

(a) What is the maximum likelihood estimator $\hat{\theta}$ of θ? **Solution:**

Note that $\Sigma_{i\in[n]}x_i = 30$, as given in the problem spec. Therefore, there are 30 **1**s and 70 **0**s. (Note that they come in some specific order.) Therefore, we can setup L as follows, because there is a θ chance of getting a 1, and a $(1-\theta)$ chance of getting a 0 and they are each i.i.d. From there, take the log-likelihood, then the first derivative, set it equal to 0 and solve for for $\hat{\theta}$.

$$\begin{aligned}
L\left(x_1, \ldots, x_n \mid \theta\right) &= (1-\theta)^{70} \theta^{30} \\
\ln L\left(x_1, \ldots, x_n \mid \theta\right) &= 70 \ln (1-\theta)+30 \ln \theta \\
\frac{\partial}{\partial \theta} \ln L\left(x_1, \ldots, x_n \mid \theta\right) &= -\frac{70}{1-\theta}+\frac{30}{\theta} \\
-\frac{70}{1-\hat{\theta}}+\frac{30}{\hat{\theta}} &= 0 \\
\frac{30}{\hat{\theta}} &= \frac{70}{1-\hat{\theta}} \\
30-30 \hat{\theta} &= 70 \hat{\theta} \\
\hat{\theta} &= \frac{30}{100}
\end{aligned}$$

(b) Is $\hat{\theta}$ an unbiased estimator of θ? **Solution:**

An estimator is unbiased if the expectation of the estimator is equal to the original parameter, i.e.: $E[\hat{\theta}]=\theta$. Setting up the expectation of our estimator and plugging it in for the generic case, we get the following, which we can then reduce with linearity of expectation:

$$\begin{aligned}
\mathbb{E}[\hat{\theta}] &= \mathbb{E}\left[\frac{1}{100} \sum_{i=1}^{100} X_i\right] \\
&= \frac{1}{100} \sum_{i=1}^{100} \mathbb{E}\left[X_i\right] \\
&= \frac{1}{100} \cdot 100 \theta=\theta .
\end{aligned}$$

so it is unbiased.

4. Y Me?

Let $y_1, y_2, \ldots y_n$ be i.i.d. samples of a random variable with density function

$$f_Y(y \mid \theta)=\frac{1}{2 \theta} \exp \left(-\frac{|y|}{\theta}\right)$$

.

Find the MLE for θ in terms of $|y_i|$ and n. **Solution:**

Since the samples are i.i.d., the likelihood of seeing n samples of them is just their PDFs multiplied together.

From there, take the log-likelihood, then the first derivative, set it equal to 0 and solve for for $\hat{\theta}$.

$$\begin{aligned}
L\left(y_1, \ldots, y_n \mid \theta\right) &= \prod_{i=1}^{n} \frac{1}{2\theta} \exp\left(-\frac{|y_i|}{\theta}\right) \\
\ln L\left(y_1, \ldots, y_n \mid \theta\right) &= \sum_{i=1}^{n} \left[-\ln 2 - \ln\theta - \frac{|y_i|}{\theta}\right] \\
\frac{\partial}{\partial\theta} \ln L\left(y_1, \ldots, y_n \mid \theta\right) &= \sum_{i=1}^{n} \left[-\frac{1}{\theta} + \frac{|y_i|}{\theta^2}\right] \\
\sum_{i=1}^{n} \left[-\frac{1}{\hat{\theta}} + \frac{|y_i|}{\hat{\theta}^2}\right] &= 0 \\
-\frac{n}{\hat{\theta}} + \frac{\Sigma_{i=1}^{n}|y_i|}{\hat{\theta}^2} &= 0 \\
\hat{\theta} = \frac{\Sigma_{i=1}^{n}|y_i|}{n}
\end{aligned}$$

5. 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?

Solution:

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

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

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

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

We know that for any random variable Y, since $Var(Y) = E(Y^2) - (E(Y))^2$ it holds that

$$E(Y^2) = Var(Y) + (E(Y))^2.$$

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

$$E(X_i^2) = \sigma^2 + \mu^2\ \ \forall i \quad \text{and} \quad E(\overline{X}^2) = \frac{\sigma^2}{n} + \mu^2.$$

Substituting these equations into (**) we get

$$E\left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2)\right)=\frac{1}{n}\sum_{i=1}^{n}E(X_i^2)-E(\overline{X}^2)=\sigma^2+\mu^2-\left(\frac{\sigma^2}{n}+\mu^2\right)$$
$$=\left(1-\frac{1}{n}\right)\sigma^2.$$

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

6. 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).

Solution:

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? **Solution:**

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:

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

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*. **Solution:**

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

7. 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 $X^{(0)} = 4$? **Solution:**

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. **Solution:**

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:

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

8. 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. **Solution:**

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

(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 $D(S)$. **Solution:**

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. **Solution:**

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.