HW3 Solutions

1. (20 pts.) Packets Over the Internet
n packets are sent over the Internet (n even). Let $X_i = 1$ if the i^{th} packet got lost and $X_i = 0$ otherwise. Consider the following probability models for the packet loss process:

(i) Each packet is routed over a different path and is lost independently with probability p.

(ii) All n packets are routed along the same path, and with probability p, one of the links along the path fails and all n packets are lost. Otherwise all packets are received.

(iii) The n packets are divided into 2 groups of $n/2$ packets, and each group is routed along a different path and lost with probability p. Losses of different groups are independent events.

In each of the three models:

(a) **(6 pts.)** Compute $\mathbf{P}(X_i = 0)$ for all i

Answer:

(i) $\mathbf{P}(X_i = 0) = 1 - p \ \forall i$
(ii) $\mathbf{P}(X_i = 0) = 1 - p \ \forall i$
(iii) $\mathbf{P}(X_i = 0) = 1 - p \ \forall i$

(b) **(14 pts.)** Determine whether X_i and X_j are independent for all $i \neq j$

Answer:

(i) By the problem statement, the random variables **are independent**
(ii) Based on the protocol, we know that if one packet is lost, then they are all lost so

$$\mathbf{P}(X_1 = 1) = p \neq \mathbf{P}(X_1 = 1 | X_2 = 1) = 1$$

hence the random variables are **not independent**.
(iii) Consider two packets i and j which are routed along the same path. The corresponding random variables X_i and X_j are **not independent** by the same logic as for protocol (ii). However, the random variables corresponding to two packets in *different* groups **are independent** by the definition of protocol (iii).

2. (20 pts.) Conditional Independence and Medical Diagnosis
In class we have defined the notion of independence of events: two events A and B are independent if $\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B)$. Now we want to define a notion of *conditional independence* of events A and B given a third event C.

(a) **(2 pts.)** Can you supply a reasonable definition for this notion?

Answer: Since the conditional probability $\mathbf{P}(\cdot|C)$ is also a valid probability assignment, to obtain the conditional version of $\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B)$, we can replace every $\mathbf{P}(\cdot)$ by $\mathbf{P}(\cdot|C)$ to obtain

$$\mathbf{P}(A \cap B|C) = \mathbf{P}(A|C)\mathbf{P}(B|C).$$

Another way to interpret it is that A, B independent means that whether B is given or not would not change the probability of A (i.e., $\mathbf{P}(A|B) = \mathbf{P}(A)$). Similarly, A, B independent given C means that given C occurs, whether we are further given B or not would not change the probability of A (i.e., $\mathbf{P}(A|B, C) = \mathbf{P}(A|C)$). Expanding this equation would give the same answer if $\mathbf{P}(B|C) > 0$.

(b) **(3 pts.)** If A and B are independent given C, show how to simplify the probability of A given B and C, i.e. $\mathbf{P}(A|B, C)$.

Answer:

$$\begin{aligned}\mathbf{P}(A|B,C) &= \frac{\mathbf{P}(A,B,C)}{\mathbf{P}(B,C)} \\ &= \frac{\mathbf{P}(A,B|C)\mathbf{P}(C)}{\mathbf{P}(B|C)\mathbf{P}(C)} \\ &= \frac{\mathbf{P}(A|C)\mathbf{P}(B|C)\mathbf{P}(C)}{\mathbf{P}(B|C)\mathbf{P}(C)} \\ &= \mathbf{P}(A|C).\end{aligned}$$

(c) **(3 pts.)** Can you supply a definition of two random variables X and Y being conditionally independent given a third random variable Z?

Answer: X and Y conditionally independent given Z if

$$\mathbf{P}(X = a, Y = b|Z = c) = \mathbf{P}(X = a|Z = c)\mathbf{P}(Y = b|Z = c)$$

for all a, b, c.

(d) **(3 pts.)** From the examples covered in lectures, give an example of three such random variables.

Answer: In the biased coin example in lecture 4 (or HW2 Q1) where we have two coins with probabilities p and q of obtaining Heads, and we randomly choose one of them (denote the choice by X) and flip it twice to obtain Y_1, Y_2, then Y_1, Y_2 are conditionally independent given X. Note that Y_1, Y_2 are not independent if $p \neq q$.

(e) **(4 pts.)** There is a disease which affects 2% of the population. A medical test is available which gives a false positive rate of 5% and a mis-detection rate of 3%. Compute the false discovery rate and the false omission rate of using this test.

Answer: Define the random variables H and T as in the lecture notes ($H = 0$ if the patient is healthy, $H = 1$ if affected, $T = 0$ if the test is negative, $T = 1$ if positive). $\mathbf{P}(H = 1) = 0.02$. Misdetection rate is $\mathbf{P}(T = 0|H = 1) = 0.03$. False positive rate is $\mathbf{P}(T = 1|H = 0) = 0.05$.

To compute the false discovery rate, by Bayes rule,

$$\begin{aligned}\mathbf{P}(H=0|T=1) &= \frac{\mathbf{P}(H=0)\mathbf{P}(T=1|H=0)}{\mathbf{P}(H=0)\mathbf{P}(T=1|H=0)+\mathbf{P}(H=1)\mathbf{P}(T=1|H=1)}\\ &= \frac{(1-0.02)0.05}{(1-0.02)0.05+0.02(1-0.03)}\\ &\approx 0.716.\end{aligned}$$

To compute the false omission rate, by Bayes rule,

$$\begin{aligned}\mathbf{P}(H=1|T=0) &= \frac{\mathbf{P}(H=1)\mathbf{P}(T=0|H=1)}{\mathbf{P}(H=1)\mathbf{P}(T=0|H=1)+\mathbf{P}(H=0)\mathbf{P}(T=0|H=0)}\\ &= \frac{0.02\cdot 0.03}{0.02\cdot 0.03+(1-0.02)(1-0.05)}\\ &\approx 0.000644.\end{aligned}$$

(f) **(5 pts.)** The doctors complain that the false discovery rate for this test is too high. The medical test company responds with a new test that has a false positive rate of 6% and a mis-detection rate of 4%. Although this new test has worse false positive rate and mis-detection rate compared to the old test, the company claims that when used *in conjunction* with the old test, will give a lower false discovery rate because the results of the two tests are conditionally independent given the disease state of the patient. More specifically, the company recommends the doctors diagnose a patient to have a disease if and only if *both* the old and the new tests are positive.
Do you agree that the company's claim? Justify your answer.

Answer: Let the result of the old test be T_1, and that of the new test be T_2. To compute the false discovery rate of the combined test,

$$\begin{aligned}&\mathbf{P}(H=0|T_1=1,T_2=1)\\ &= \frac{\mathbf{P}(H=0)\mathbf{P}(T_1=1,T_2=1|H=0)}{\mathbf{P}(H=0)\mathbf{P}(T_1=1,T_2=1|H=0)+\mathbf{P}(H=1)\mathbf{P}(T_1=1,T_2=1|H=1)}\\ &= \frac{\mathbf{P}(H=0)\mathbf{P}(T_1=1|H=0)\mathbf{P}(T_2=1|H=0)}{\mathbf{P}(H=0)\mathbf{P}(T_1=1|H=0)\mathbf{P}(T_2=1|H=0)+\mathbf{P}(H=1)\mathbf{P}(T_1=1|H=1)\mathbf{P}(T_2=1|H=1)}\\ &= \frac{(1-0.02)0.05\cdot 0.06}{(1-0.02)0.05\cdot 0.06+0.02(1-0.03)(1-0.04)}\\ &\approx 0.136.\end{aligned}$$

The false discovery rate is indeed lower.

3. (20 pts.) B and B
In class, we said we can define two sets of random variables for the balls and bins problem:

1. X_i = index of the bin where the i^{th} ball lands, $i = 1, \ldots, m$
2. Y_i = number of balls in bin i, $i = 1, \ldots, n$

(a) **(4 pts.)** Do the X_i's contain the same information about the system as the Y_i's? In other words, can one compute the Y_i's given the X_i's and vice versa? If not, which set of random variables contain more information, and give an example of an event that can be expressed in terms of one set of random

variables but not the other.

Answer: No they do not contain the same information. The X_i's contain more information than the Y_i's. By knowing where each ball is located we can can calculate the number of balls in each bin (i.e. we can get Y_i from knowing all the X_i's).

One example of an event that can be expressed in terms of the first set of random variables but not he second is, the event that the first ball lands in the second bin. This can be expressed as $X_1 = 2$ but it cannot be expressed in terms of the random variables Y_i.

(b) **(16 pts.)** Now consider a probability model where the X_i's are mutually independent. In class, we give some intuition why the Y_i's are not mutually independent. Here you will verify that this is indeed the case.

(i) **(5 pts.)** Compute $\mathbf{P}(Y_i = 0), i = 1, \ldots, n$

Answer: We can rewrite $\mathbf{P}(Y_i = 0)$ as $\mathbf{P}(X_1 \neq i, X_2 \neq i, \ldots, X_m \neq i)$. Now using the independence of the X_i's we have

$$\mathbf{P}(Y_i = 0) = \mathbf{P}(X_1 \neq i)\mathbf{P}(X_2 \neq i) \cdots \mathbf{P}(X_m \neq i)$$

Assuming each bin is equally likely $\mathbf{P}(X_i \neq i) = 1 - 1/n$ so we conclude

$$\mathbf{P}(Y_i = 0) = (1 - 1/n)^m$$

(ii) **(6 pts.)** Compute $\mathbf{P}(Y_i = 0, Y_j = 0)$ for $i \neq j$ (You may want to separate out the two cases when $n = 2$ and $n > 2$)

Answer:
$\underline{n = 2}$:
Since we only have two bins we only need to calculate $\mathbf{P}(Y_1 = 0, Y_2 = 0)$. From the problem statement we know that there is at least 1 ball and it can only go into one of the two bins so we conclude that the event that neither bin has a ball is impossible. Specifically,

$$\mathbf{P}(Y_1 = 0, Y_2 = 0) = 0$$

$\underline{n > 2}$:
We can break the probability down using conditional probabilities as follows:

$$\mathbf{P}(Y_i = 0, Y_j = 0) = \mathbf{P}(Y_i = 0 | Y_j = 0)\mathbf{P}(Y_j = 0)$$

From the previous part we know $\mathbf{P}(Y_j = 0) = (1 - 1/n)^m$. To calculate the conditional probability we exploit the fact that knowing bucket j has no balls implies the balls *must* lie in the other $n - 1$ buckets. So

$$\mathbf{P}(Y_i = 0 | Y_j = 0) = \left(1 - \frac{1}{n-1}\right)^m$$

Multiplying the probabilities gives us

$$\mathbf{P}(Y_i = 0, Y_j = 0) = \left(1 - \frac{1}{n-1}\right)^m (1 - 1/n)^m$$

(iii) **(2 pts.)** Are Y_i and Y_j independent?

Answer: No the random variables are not independent since

$$\mathbf{P}(Y_i = 0|Y_j = 0) = \left(1 - \frac{1}{n-1}\right)^m \neq \left(1 - \frac{1}{n}\right)^m = \mathbf{P}(Y_i = 0) \quad i \neq j$$

(iv) **(3 pts.)** What happens when n is very large? Can you give some intuition for your answer?

Answer: When n is very large $\mathbf{P}(Y_i = 0|Y_j = 0) \approx \mathbf{P}(Y_i = 0)$. Intuitively, the more bins you have, the less information knowing one of the bins is empty gives you about the others.

4. (20 pts.) Random Variables

(a) **(4 pts.)** Define the basic random variables and give the sample space and assign probabilities to the outcomes.

Answer: We have 4 random variables X_i which is 1 if the i^{th} coint toss is heads and 0 otherwise. Our sample space is composed of all possible outcomes of the coin flips.

$$\Omega = \{TTTT, HTTT, THTT, \ldots, HHHT, HHHH\}$$

As seen in lecture, the probability of getting $HTTT$ and the probability of getting $TTHT$ are exactly the same: $p(1-p)^3 = (1-p)^2 \cdot p(1-p)$ and in general, the probability of getting a sequence with r heads out of the 4 tosses is $p^r(1-p)^{4-r}$.

(b) **(4 pts.)** Let X be the total number of Heads in the four flips. Draw a Venn diagram showing the five events $X = i, i = 0, 1, 2, 3, 4$ as well as the sample space and the outcomes. Is X a random variable?

Answer: Indeed, X is a random variable. It takes every sample point in our sample space, and assigns it a real value. Our sample space and Venn diagram are shown in figure 1.

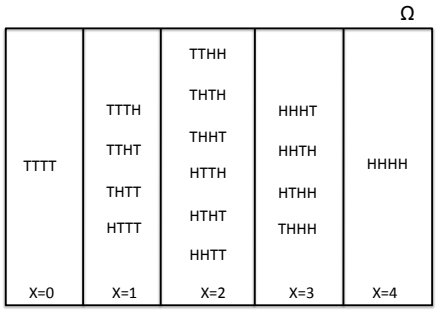

Figure 1: The sample space for X.

(c) **(4 pts.)** Are the events $X = 1$ and $X = 2$ disjoint? Are they independent? What about the events $X = 1$ and $X \leq 2$?

Answer: The events $X = 1$ and $X = 2$ are indeed disjoint (it is impossible that there will be a total of both 1 heads and 2 heads in on outcome) and therefore $\mathbf{P}(X = 1 \cap X = 2) = \mathbf{P}[\emptyset] = 0$. These events however are not independent since $\mathbf{P}(X = 1) \cdot \mathbf{P}(X = 2) \neq 0 = \mathbf{P}(X = 1 \cap X = 2)$.

The events $X = 1$ and $X \leq 2$ are not disjoint. The event $X = 1$ is a subset of the event $X \leq 2$, therefore their intersection is not empty and is the event $X = 1$. They are not independent since $\mathbf{P}(X = 1) \cdot \mathbf{P}(X = 2) \neq \mathbf{P}(X = 1) = \mathbf{P}(X = 1 \cap X \leq 2)$.

(d) **(4 pts.)** Let Y be the first flip when a Heads appears and $Y = 0$ if there is no Heads in the four flips. Draw a Venn diagram showing the five events $Y = i, i = 0, 1, 2, 3, 4$ as well as the sample space and the outcomes.

Answer: Recall that a random variable on a sample space is a function that assigns to each sample point in the sample space a real number. The function Y is not defined for the event in which no heads come out at all, which is a valid point in our sample space. Our sample space and Venn diagram are shown in figure 2.

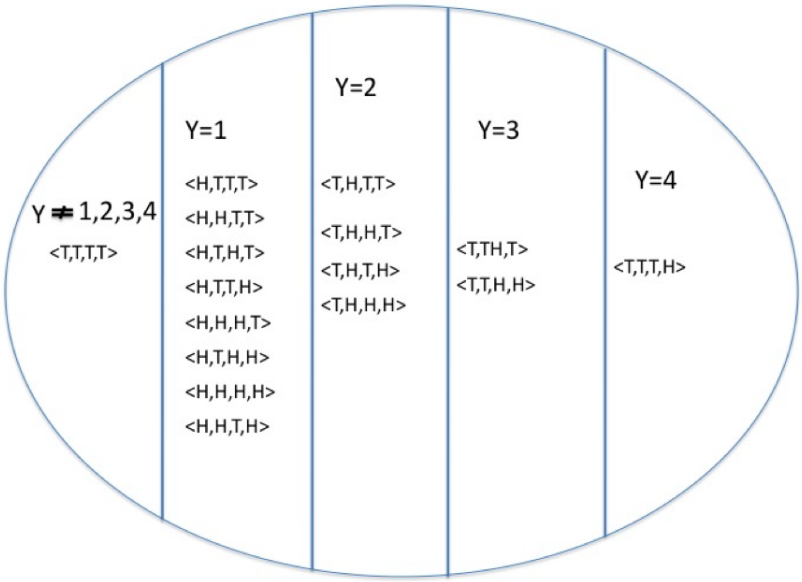

Figure 2: The sample space for Y.

(e) **(4 pts.)** Are the events $X = 3$ and $Y = 4$ disjoint? Are they independent? What about the events $X = 2$ and $Y = 2$?

Answer: The events $X = 3$ and $Y = 4$ are indeed disjoint, since if heads only appears in the last flip, the total number of heads in all four flips is exactly one. Since both events $Y = 4$ and $X = 3$ have some positive probability (i.e. not 0) and $\mathbf{P}(Y = 4)\mathbf{P}(Y = 3) \neq 0 = \mathbf{P}(X = 3 \cap Y = 4)$, and thus these events are not independent

The events $X = 2$ and $Y = 2$ are not disjoint. This is because their are outcomes that are both in $X = 2$ and in $Y = 2$, that is the set $(X = 2) \cap (Y = 2)$ is not empty. For example, the outcome $THTH$ respects both properties of the having it land on heads the first time on the second flip, and the total number heads being 2. These events are not independent. We have that $\mathbf{P}(X = 2) = \binom{4}{2}p^2(1-p)^2$ and $\mathbf{P}(Y = 2) = (1-p)p$, and therefore $\mathbf{P}(X = 2) \cdot \mathbf{P}(Y = 2) = \binom{4}{2}p^3(1-p)^3$. On the other hand, the intersection of these events, $(X = 2) \cap (Y = 2)$ is $\{THTH, THHT\}$, each with probability $p^2(1-p)^2$, and therefore $\mathbf{P}(X = 2 \cap Y = 2) = 2 \cdot p^2(1-p)^2$. In general, $\binom{4}{2}p^3(1-p)^3 \neq p^2(1-p)^2$.

5. (20 pts.) DNA Sequencing In high throughput sequencing technologies, a DNA of length G symbols s_1, s_2,..., s_G (each symbol one of the 4 possible nucleotides A,G,C,T) is sequenced by randomly sampling

short subsequences called reads from it. For simplicity, we will assume that the genome is circular. See figure 1. Each read R_i, of length L symbols, is uniformly sampled from the genome, and the locations of different reads are mutually independent. We sample N such reads in order to reconstruct the underlying genome. You can assume that G is significantly larger than L.

(a) **(12 pts.)** Compute the probability that position i on the genome is not covered by any read.

Answer: Let $E_i^{(R_j)}$ be the event that position i on the genome is **not** covered by read R_j. We want to calculate $\mathbf{P}\left(E_i^{(R_1)} \cap E_i^{(R_2)} \cap \cdots \cap E_i^{(R_N)}\right)$. But since the location of each read is independent we can just multiply the probabilities of each event. Specifically,

$$\begin{aligned}\mathbf{P}\left(E_i^{(R_1)} \cap E_i^{(R_2)} \cap \cdots \cap E_i^{(R_N)}\right) &= \prod_{j=1}^{N} \mathbf{P}\left(E_i^{(R_j)}\right) \\ &= \prod_{j=1}^{N} (1 - L/G) \\ &= (1 - L/G)^N\end{aligned}$$

(b) **(4 pts.)** Give a non-trivial upper bound on the probability that at least one position on the genome is not covered by any reads.

Answer: Let A_i be the event that position i is not covered (i.e. the event for which we calculated the probability in part A), then

$$\begin{aligned}\mathbf{P}(\text{at least one pos. not covered}) &= \mathbf{P}(\cup_{i=1}^{N} A_i) \\ &\leq \sum_{i=1}^{N} \mathbf{P}(A_i) \\ &= G(1 - L/G)^N\end{aligned}$$

Where we have used the union bound to get the upper bound.

(c) **(4 pts.)** Suppose $G = 3 \times 10^9$ and $L = 100$. How many reads do we need to sample to guarantee that the probability in part (b) is at most 1%?

Answer: Plugging in our values and solving for N we have

$$3 \times 10^9 \left(1 - \frac{100}{3 \times 10^9}\right)^N \leq 0.01$$

$$N \log\left(1 - \frac{100}{3 \times 10^9}\right) \leq \log\left(\frac{0.01}{3 \times 10^9}\right)$$

$$N \geq \frac{-26.42705}{-3.333 \times 10^{-8}}$$

$$N \geq 7.93 \times 10^8$$

Where $\log(\cdot)$ is understood to be the natural logarithm (not that it particularly matters)