Problem Set 10 Solutions

Problems:

Problem 1 Recall the envelope guessing game demonstrated in lecture. (You do go to lecture, right?) You are shown two envelopes, each lying face down on a table. There is a number from $\{0, \ldots, n\}$ written inside each. You pick an envelope and get to see the number in that envelope. You can then either choose to stay with that number, or switch to the number written in the other envelope. You *win* if the number you end up with is the larger of the two (that is, if either you originally picked the larger and stayed, or if you originally picked the smaller and switched).

A randomized strategy for this game is a function that, for each number i that may be revealed in the envelope you look at, assigns a probability p_i of switching. For example, a strategy where $p_i = \frac{1}{2}$ for all i effectively ignores the number revealed and chooses randomly. A strategy where $p_i = 1$ for $i < \frac{n}{2}$ and $p_i = 0$ for $i \geq \frac{n}{2}$ would switch if the number revealed is in the lower half of the range and stay if it is in the upper half of the range.

(a) Prove that the strategy described in class for winning with probability better than 1/2 is optimal if the envelopes are filled with x and $x+1$, where x is chosen uniformly from $\{0, \ldots, n-1\}$. In other words, since we have shown that the strategy given in lecture wins with probability at least $\frac{1}{2} + \frac{1}{2n}$, show that the expected value of wins per guess with *any* randomized strategy (any set of $\{p_0, p_1, \ldots p_n\}$ is *at most* $\frac{1}{2} + \frac{1}{2n}$.

Solution.

Let p_i denote the probability that you switch if you see the number i. This is an arbitrary choice, so it covers every strategy (a strategy is such a set of p_i). Then, the probability that you win if the envelopes contain the pair $(i, i+1)$ is the probability that you will see the smaller number and switch or you will see the bigger number and stay. Since the probability of picking either envelope (and thus either number) is 1/2, the probability that you win is

$$\frac{1}{2}p_i + \frac{1}{2}(1 - p_{i+1})$$

Then, since the pair in the envelopes can range from $(0, 1)$ to $(n-1, n)$, the total probability that you win is:

$$\sum_{i=0}^{n-1} \frac{1}{n}\left(\frac{1}{2}p_i + \frac{1}{2}(1 - p_{i+1})\right) = \frac{1}{2} + \frac{1}{2n}\sum_{i=0}^{n-1}(p_i - p_{i+1}) = \frac{1}{2} + \frac{1}{2n}(p_0 - p_n) \leq \frac{1}{2} + \frac{1}{2n}$$

(b) *Optional question:* Can you think of another strategy that is also optimal as described above? Is this new strategy 'really' optimal?

Solution.

This can almost be read directly off of the probability formula. The strategy is to switch if 0 is revealed, stay if n is revealed, and act arbitrarily otherwise (always switch, always stay, switch half the time, etc). This is reasonable, since those two choices are guaranteed to lead to a win.

No, it's not truly an optimal strategy – there is no guarantee that the distribution of the envelopes even includes the endpoints, in the general case. It relies on the envelopes being filled with the given distribution.

Problem 2 A new fault-tolerant computer network contains 100,000 computer nodes, and is designed to function adequately when no more than 5000 computers are down. Suppose that the probability that any particular computer is down is 4%; moreover, these probabilities are independent for different computers.

(a) Estimate the probability that exactly 5000 computers are down.

Solution.

$$f_{100000,0.04}(5000) = 1.7264 \cdot 10^{-55}$$

(b) Estimate the probability that the network can function adequately.

Solution.

Here, we must estimate the probability that more than 95000 computers are fault free. As a rough but high approximation, we can notice that the expected number of computers to be down is 4000. If we assume that all numbers of down computers from 5001 to 100,000 are less likely to occur than 5000 computers being down, we can upper bound the probability that more than 5000 computers are down by

$$95000 \cdot \Pr(95000 \text{ computers working}) = 95000 \cdot 1.7264 \cdot 10^{-55} = 1.640 \cdot 10^{-50}$$

which should suggest that the probability of a functioning network is extremely high, greater than $1 - 1.640 \cdot 10^{-50}$.

For a more precise bound, we can find a lower bound on the probability that the network can function adequately:

$$\Pr(> 95000 \text{ computers working}) = 1 - \Pr(\leq 95000 \text{ computers working})$$

An upper bound for $\Pr(\leq 95000$ computers working) is given by:

$$\begin{aligned} F_{100000,0.96}(0.95 \cdot 100000) &\leq \frac{1-\alpha}{1-\frac{\alpha}{p}} f_{n,p}(\alpha \cdot n) \\ &= \frac{1-\alpha}{1-\frac{\alpha}{p}} f_{n,(1-p)}((1-\alpha) \cdot n) \\ &\approx 4.8 \cdot f_{n,(1-p)}((1-\alpha) \cdot n) \end{aligned}$$

Plugging in for $f_{n,(1-p)}((1-\alpha) \cdot n)$ from part a, we get:

$$F_{100000,0.96}(0.95 \cdot 100000) \leq 8.2866 \cdot 10^{-55}$$

Thus, it follows that:

$$\Pr(> 95000 \text{ computers working}) \approx 1$$

Problem 3 The Gallup polling service is conducting a poll to see what percentage of Americans believe that 6-year-olds should be given the legal right to decide where they should live. They would like to be accurate (after all, this is a very serious issue) to within a 2% margin of error, with at least 98% probability. Estimate how many people they would need to poll to be able to make this claim of reliability. (As in lecture, assume that they are polling with replacement, so they may poll the same person twice).

Solution.

We can follow the form used in lecture, using the bounds shown there. (Note that this does not tell us the *minimum* number of people we need to poll to be sure enough, just a sufficient number). Using the general formula

$$\delta \leq 2F_{n,\frac{1}{2}}((\frac{1}{2}-\epsilon)n)$$

we have $\delta = 0.02$, $\epsilon = 0.02$, so

$$\begin{aligned} \delta &\leq 2F_{n,\frac{1}{2}}(\frac{12}{25}n) \\ &\leq 2 \cdot \frac{1-0.48}{1-\frac{0.48}{0.5}} \cdot f_{n,\frac{1}{2}}(0.48n) \\ &\leq 26 \cdot \left(\frac{2^{\left(0.48lg\left(\frac{0.5}{0.48}\right)+0.52lg\left(\frac{0.5}{0.52}\right)\right)n}}{\sqrt{2\pi 0.48 \cdot 0.52n}} \right) \\ &\leq 26 \cdot \left(\frac{2^{-0.00115446 \cdot n}}{1.25231108\sqrt{n}} \right) \end{aligned}$$

The easiest way to solve this is just plugging in each potential n into the final equation until we find where it must be that $\delta \leq 0.02$. It turns out that it is sufficient to survey approximately 3569 people.

Problem 4 Let X_1, X_2, and X_3 be three mutually independent random variables, each having the uniform distribution, $\Pr(X_i = k)$ equal to $1/3$ for each of $k = 1$, $k = 2$, and $k = 3$. Let M be another random variable giving the maximum of these three random variables.

(a) What is the distribution of M?

Solution.

This can be easily hashed out by counting the possible outcomes:

$$\begin{array}{lll} M & \text{is} & 1 \text{ with probability } \frac{1}{27} \\ & & 2 \text{ with probability } \frac{7}{27} \\ & & 3 \text{ with probability } \frac{19}{27} \end{array}$$

(b) What is the expected value of M?

Solution.

Directly from the above, it is:

$$\frac{1}{27}(1) + \frac{7}{27}(2) + \frac{19}{27}(3) = \frac{8}{3}$$

(c) Generalize to n variables with possible values $1, 2, \ldots n$.

Solution.

We can solve this by examining the probabilities as follows:

$$\Pr(M \leq k) = \Pr(\text{no } X_i \text{ is greater than } k) = \left(\frac{k}{n}\right)^n$$

However,

$$\begin{aligned} \Pr(M = k) &= \Pr(M \leq k) - \Pr(M \leq k - 1) \\ &= \left(\frac{k}{n}\right)^n - \left(\frac{k-1}{n}\right)^n \end{aligned}$$

And the expectation of M becomes:

$$\begin{aligned}
\sum_{k=1}^{n}\left[k\cdot\left(\left(\frac{k}{n}\right)^{n}-\left(\frac{k-1}{n}\right)^{n}\right)\right] &= \sum_{k=1}^{n}\left[k\cdot\left(\frac{k}{n}\right)^{n}\right]-\sum_{k=1}^{n}\left[k\cdot\left(\frac{k-1}{n}\right)^{n}\right] \\
&= \sum_{k=1}^{n}\left[k\cdot\left(\frac{k}{n}\right)^{n}\right]-\sum_{k=0}^{n-1}\left[(k+1)\cdot\left(\frac{k}{n}\right)^{n}\right] \\
&= \sum_{k=1}^{n}\left[k\cdot\left(\frac{k}{n}\right)^{n}\right]-\sum_{k=0}^{n-1}\left[k\cdot\left(\frac{k}{n}\right)^{n}\right]-\sum_{k=0}^{n-1}\left(\frac{k}{n}\right)^{n} \\
&= n-\sum_{k=0}^{n-1}\left(\frac{k}{n}\right)^{n}
\end{aligned}$$

Problem 5 Suppose a length n string of 0's and 1's is chosen uniformly randomly (e.g., by n independent tosses of fair coin) and then wrapped around to form a circle.

(a) What is the maximum number of occurrences of the string 0^k (k 0's) in this circle? What is the expected number of occurrences of 0^k? (Note: A string of $k+1$ 0's counts as *two* strings of length k – starting in the first position of the string and starting in the second position. Also remember that there is no particular start or end of the circle.)

Solution.

The maximum number of 0^k substrings is n – in a string of all 0's, a new such substring occurs at every position. The expected number can be found by simply adding up probabilities of the separate events that the string occurs starting at position i, which should be $\frac{n}{2^k}$. While the events are not independent, we are 'overcounting' in exactly the correct way. This concept is referred to, of course, as *linearity of expectation.*

To be more rigorous, define random variables X_i for $i = 1, 2, \ldots n$, where $X_i = 1$ if a string of k 0's starts at position i and 0 if it does not. Then, we are looking for $E[X_1 + X_2 + \ldots + X_n]$, which by linearity of expectation must be $E[X_1] + E[X_2] + \ldots + E[X_n]$.

This can be envisioned as a counting problem instead. There are 2^n possible strings. Each position will begin a substring of 0^k for 2^{n-k} different strings. Thus, the total number of such substrings divided by the number of possible strings is $\frac{n2^{n-k}}{2^n}$, which simplifies to $\frac{n}{2^k}$.

(b) Suppose k is even. What is the expected number of occurrences of $0^{\frac{k}{2}}1^{\frac{k}{2}}$ (total length k)?

Solution.

This is the same as in the previous part, since the reasoning does not change (perhaps surprisingly).

(c) Now fix $n = 8$, $k = 4$. Is the probability that 0000 appears in the circle equal to the probability that 0011 appears?

Solution.

This can be reasoned out by counting the strings in each case.

For 0000, there is one string that contains 8 copies, 00000000. The rest of the strings containing 0000 can be ordered by the number of copies of 0000 that they contain; each one can also have the long string of 0's begin at any of the 8 positions:

1 string	00000000	(8 copies)
8 strings	00000001	(4 copies)
8 strings	00000011	(3 copies)
16 strings	000001x1	(2 copies)
32 strings	00001xx1	(1 copy)

Thus, in total, 65 of the 256 possible strings contain a copy of 0000. (Note that there are 128 copies in total; this agrees with the expectation of $\frac{n}{2^k}$ multiplied by the number of possible strings, 2^n). So the probability of a random string containing a copy of 0000 is

$$\frac{65}{256}$$

For 0011, we get the following:

4 strings	00110011	(2 copies)
120 strings	0011$xxxx$	(1 copy)

The number of strings with one copy is a bit complex; the 0011 sequence can start in any of 8 places, and the other four characters can be any of the 16 possibilities except 0011. Altogether, 124 of the 256 possible strings contain a copy of 0000. (Note that there are again 128 copies in total). The probability of a random string containing a copy of 0011 is

$$\frac{124}{256}$$

Thus, the probability is not the same; it is significantly more likely that a substring of 0011 will appear.

(d) Justify your answer for the previous part. If your answers in the previous parts seem to be "contradictory", explain.

Solution.

Although the expectations are the same, the extra overlapping possibilities for 0000 make the conditional probability of a second occurrence, conditioned on a first occurrence, higher than it is for 0011. Since it is more likely that there will be a second occurrence, the expectation must be composed of less situations in which the target string occurs at all.

Problem 6 Suppose that you choose a listing (permutation) of the numbers $1, 2, \ldots, n$ uniformly at random. What is the expected number of entries that are greater than all entries that follow them? For example, in the listing 4, 2, 5, 1, 3, the numbers 3 and 5 are greater than all following entries. (Hint: Try using indicator variables).

Solution.

Let X_i be an indicator for the event that the i-th entry is greater than the following $n-i$ entries. Since the indicator random variables are 1 when the event occurs and 0 when it does not, $\mathrm{E}[X_i] = \Pr(X_i = 1)$, as we have seen. By linearity of expectation,

$$\begin{aligned}\mathrm{E}[X_1 + X_2 + \ldots + X_n] &= \sum_{i=1}^{n} \mathrm{E}[X_i] \\ &= \sum_{i=1}^{n} \Pr(X_i = 1)\end{aligned}$$

All that remains is to compute $\Pr(X_i = 1)$. Note that no matter what the last $n-i+1$ numbers are, each of the last $n-i+1$ entries is equally likely to be the largest. Therefore, $\Pr(X_i = 1) = \frac{1}{n-i+1}$. Substituting into the equation above gives:

$$\begin{aligned}\mathrm{E}[\#\text{ new maxima}] &= \sum_{i=1}^{n} \Pr(X_i = 1) \\ &= \sum_{i=1}^{n} \frac{1}{n-i+1} \\ &= \sum_{i=1}^{n} \frac{1}{i} \\ &= H_n\end{aligned}$$

Problem 7 Every day in Boston, a rainstorm occurs with probability 1/3 (independently of any other day). Each day it rains, the radar system at the Nashua airport breaks with probability 1/25. The system never breaks on fair days.

(a) Assuming that the radar was working yesterday (but may break today, if it happens to rain), how many days can we expect the Nashua radar system to work correctly before it next breaks?

Solution.

This can be done with a simple recurrence:

$$\begin{aligned}w &= \frac{2}{3}(1+w) + \frac{1}{3}(\frac{24}{25}(1+w)) \\ w &= \frac{2}{3} + \frac{2}{3}w + \frac{8}{25} + \frac{8}{25}w \\ \frac{1}{75}w &= \frac{74}{75} \\ w &= 74\end{aligned}$$

Notice that in this case, we could have tried to predict the expected time the system will work as simply $\frac{1}{p}$, where p is the probability of failure on any given day, because it never breaks on non-rainy days. However, this would technically be off by 1, since it counts the day that the system breaks as well.

(b) *Optional question:* Suppose that the first time the radar system breaks (on a rainy day with probability 1/25), it is only patched-up with a quick hack to keep it working. (As a computer science student, you should be able to relate to this). After being patched-up, it will fail on future rainy days with probability 1/10, after which the system will need major repairs. Calculate the expected number of days until the radar system needs *major* repairs. **Solution.**

$$\begin{aligned}
w &= \frac{2}{3}(1+w) + \frac{1}{3}(\frac{24}{25}(1+w) + \frac{1}{25}(1+v)) \\
v &= \frac{2}{3}(1+v) + \frac{1}{3}(\frac{9}{10}(1+v)) \\
v &= \frac{2}{3} + \frac{2}{3}v + \frac{3}{10} + \frac{3}{10}v \\
\frac{1}{30}v &= \frac{29}{30} \\
v &= 29 \\
w &= \frac{2}{3}(1+w) + \frac{1}{3}(\frac{24}{25}(1+w) + \frac{1}{25}(30)) \\
w &= \frac{2}{3} + \frac{2}{3}w + \frac{8}{25} + \frac{8}{25}w + \frac{30}{75} \\
\frac{1}{75}w &= \frac{104}{75} \\
w &= 104
\end{aligned}$$

Although you could have also seen these as independent events that occur sequentially. The first breakage is expected on day $1/\frac{1}{75}$, and the second is expected on day $1/\frac{1}{30}$. Thus, we expect their sum to be $75 + 30 = 105$, by linearity of expectation. Depending on how you interpret the day on which it needs the repairs, this gives either 104 or 105 days.

Problem 8 Suppose that R is a random variable. Let E be an event. The *conditional expectation* of R given that event E occurs is denoted $\mathrm{E}[R \mid E]$ and is defined to be

$$\mathrm{E}[R \mid E] = \sum_{x \in \mathrm{range}(R)} x \cdot \Pr(R = x \mid E)$$

(a) Suppose I flip a fair coin 4 times. Compute the expected number of heads, given that all 4 values are not the same.

Solution.

Let the random variable R be the number of heads, and let E be the event that not all flips are the same.

$$\begin{aligned}
\mathrm{E}[R \mid E] &= \sum_{x=1}^{4} \Pr(R = x \mid E)(x) \\
&= 0(0) + \frac{4}{14}(1) + \frac{6}{14}(2) + \frac{4}{14}(3) + 0(4) \\
&= 2
\end{aligned}$$

(b) Compute the expected value of the number rolled on a fair, 6-sided die, given that the outcome is prime.

Solution.

Define R as the number rolled and E as the event that a prime number is rolled.

$$\begin{aligned}
\mathrm{E}[R \mid E] &= \sum_{x=1}^{6} x \cdot \Pr(R = x \mid E) \\
&= 0 \cdot (1) + \frac{1}{3} \cdot (2) + \frac{1}{3} \cdot (3) + 0 \cdot (4) + \frac{1}{3} \cdot (5) + 0 \cdot (6) \\
&= \frac{10}{3}
\end{aligned}$$

(c) Prove the following identity, assuming that all expectations exist.

$$\mathrm{E}[R] = \Pr(E) \cdot \mathrm{E}[R \mid E] + \Pr(\overline{E}) \cdot \mathrm{E}\left[R \mid \overline{E}\right]$$

Solution.

We transform the right side into the left side. All summations are over $x \in \mathrm{Range}(R)$.

$$\begin{aligned}
&\Pr(E) \cdot \mathrm{E}[R \mid E] + \Pr(\overline{E}) \cdot \mathrm{E}\left[R \mid \overline{E}\right] \\
&= \Pr(E) \cdot \left(\sum_x x \cdot \Pr(R = x \mid E)\right) + \Pr(\overline{E}) \cdot \left(\sum_x x \cdot \Pr(R = x \mid \overline{E})\right) \\
&= \Pr(E) \cdot \left(\sum_x x \cdot \frac{\Pr(R = x \cap E)}{\Pr(E)}\right) + \Pr(\overline{E}) \cdot \left(\sum_x x \cdot \frac{\Pr(R = x \cap \overline{E})}{\Pr(\overline{E})}\right) \\
&= \sum_x x \cdot \left(\Pr(R = x \cap E) + \Pr(R = x \cap \overline{E})\right) \\
&= \sum_x x \cdot \left(\Pr((R = x \cap E) \cup (R = x \cap \overline{E}))\right) \\
&= \sum_x x \cdot \Pr(R = x) \\
&= \mathrm{E}[R]
\end{aligned}$$

The first step uses the definition of conditional expectation. The second step uses the definition of conditional probability. The third step is simplification. The fourth step uses the fact that the probability of a disjoint union of events is equal to the sum of the event probabilities. The fifth step is simplification, and the final step uses the definition of expected value.

(d) *Optional problem:* Suppose you are playing poker. You have the following chart of payoffs:

One Pair	:	\$10
Two Pair	:	\$15
Three of a Kind	:	\$30
Full House	:	\$60
All else	:	\$0

Calculate the expected amount of money you will make on a hand, given that the first two cards you are dealt are both Jacks.

Solution.

Note that $\binom{50}{3} = 19600$, the total number of possible hands given that the first two are Jacks. We can compute the probabilities by counting:

For Two Pair, we have to choose the rank of the second pair, out of the 12 choices (besides Jacks). We then must choose 2 of the 4 suits. The last card can be any of the remaining 44 that don't match the pairs.

For Four of a Kind, we can only choose the last card out of the remaining 48.

For Three of a Kind, we can choose either of the remaining two Jacks and are then free to choose the last two cards out of any of the 48 non-Jacks.

For a Full House, we can choose any of the 12 non-Jack ranks for the triplet and any 3 of the 4 cards in that rank.

Two Pair, finally, can either be calculated as the remaining hands, since it is the 'default', or as choosing 3 of the 12 non-Jack suits and 1 of the 4 cards in each.

(There are many ways to calculate any of these figures, depending on how you look at them.)

$$\begin{array}{rcllcl}
\text{Two Pair} & : & 12 \cdot \binom{4}{2} \cdot 44 & = 3168 & \longrightarrow & \frac{3168}{19600} \\
\text{Four of a Kind} & : & 48 & & & \\
\text{Three of a Kind} & : & 2 \cdot \binom{48}{2} + 48 & = 2304 & \longrightarrow & \frac{2304}{19600} \\
\text{Full House} & : & 12 \cdot \binom{4}{3} & = 48 & \longrightarrow & \frac{48}{19600} \\
\text{One Pair} & : & \binom{50}{3} - 3168 - 2304 - 48 & = 14080 & \longrightarrow & \frac{14080}{19600}
\end{array}$$

The problem changes slightly if you did not count Four of a Kind as a valid Three Pair, but by the payoffs given that is the optimal thing to do.

We thus get as the expectation:

$$\frac{14080}{19600}(10) + \frac{3168}{19600}(15) + \frac{2304}{19600}(30) + \frac{48}{19600}(60) = \$13.28$$