CS 6150 - Fall 2025 - HW 3
Randomized algorithms, Optimization formulations

Submission date: Wednesday, November 26, 2025 (11:59 PM)

This assignment has 5 questions for a total of 110 points. You will still be graded out of 100, and any points you earn above 100 will count as bonus and can compensate for a low score on other homeworks. Unless otherwise specified, complete and well-reasoned formal arguments will be expected in all answers.

Question	Points	Score
Superheroes	25	
Birthdays and applications	25	
Investment Analysis	20	
Checking feasibility vs optimization	20	
Minimum vertex cover revisited	20	
Total:	110	

Note: This homework has been released early. Some of the material required to solve these problems will be taught in class in the upcoming weeks. However, you should still be able to get started on a few questions using what has been taught so far.

Question 1: Superheroes ..... **[25]**

A cereal company has decided to give out superhero stickers with boxes of its cereal. There are n superheroes in total, and suppose that each cereal box you buy has a sticker of a uniformly random superhero. What is the expected number of boxes you need to buy so that you end up with at least one copy of *all* the n stickers?

There are many ways to do this analysis; let us see one of them. We would like to write down a recurrence for the expected value. Define $B(k)$ to be the expected number of boxes you need to buy to end up with all the stickers, *given that you have already seen k distinct stickers*. Thus, when $k = n$, we get $B(n) = 0$ as no more boxes need to be bought once you see all n stickers. When you begin buying boxes, you have seen no stickers so far (i.e. $k = 0$ in the beginning).

(a) **[10]** Use the law of conditional expectations to prove that

$$B(k) = \frac{n}{n-k} + B(k+1)$$

Define $X = T_k$ where T_k is the number of expected boxes of cereal we need to purchase after seeing k stickers

$$\mathbb{E}[X] = \mathbb{E}[T_k]$$

Define Y to be a random variable where

$Y = 1$ If the box purchased gives a unique sticker

$Y = 0$ If the box purchased gives a sticker we've already seen

If we take expectation over all possible values of Y we get the following

$$\mathbb{E}[X|Y] = \mathbb{E}[T_k|Y=1] * \Pr[Y=1] + \mathbb{E}[T_k|Y=0] * \Pr[Y=0] \quad (*)$$

Lets consider the case where we get a sticker we haven't seen and the case where we don't, and how that affects expectation

$$\mathbb{E}[T_k|Y=1] = 1 + T_{k+1} \rightarrow 1 + B(k+1)$$
$$\mathbb{E}[T_k|Y=0] = 1 + T_k \rightarrow 1 + B(k)$$

The 1 in both (1) and (2) come from the fact that we have purchased another box. After we purchase a box we update our random variable T_k to either have seen another one of the k stickers or to not have seen it.

Define Probabilities:

$$\Pr[Y=1] = \frac{n-k}{n}$$
$$\Pr[Y=0] = \frac{k}{n}$$

Plug all values those back into $(*)$

$$B(k) = \frac{n-k}{n} * (1 + B(k+1)) + \frac{k}{n} * (1 + B(k))$$
$$B(k) = 1 + \frac{n-k}{n}B(k+1) + \frac{k}{n}B(k)$$
$$B(k) - \frac{k}{n}B(k) = 1 + \frac{n-k}{n}B(k+1)$$
$$B(k)(1 - \frac{k}{n}) = 1 + \frac{n-k}{n}B(k+1)$$
$$B(k)(\frac{n}{n} - \frac{k}{n}) = 1 + \frac{n-k}{n}B(k+1)$$
$$B(k)(\frac{n-k}{n}) = 1 + \frac{n-k}{n}B(k+1)$$
$$B(k) = \frac{n}{n-k} + B(k+1)$$

(b) [**10**] Use the above result to answer the original question: What is the expected number of boxes you need to buy so that you end up with at least one copy of *all* the n stickers?
[*Hint:* $1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n} = \ln n + c$ for some $c \in (0, 1)$.]
Plug and Chug starting with k = 0. Base case $B(n) = 0$ (The case where all stickers are found)

$$\begin{aligned} B(0) &= \frac{n}{n} + B(1) \\ &= \frac{n}{n} + \frac{n}{n-1} + B(2) \\ &= \frac{n}{n} + \frac{n}{n-1} + \frac{n}{n-2} + B(3) \\ &= \frac{n}{n} + \frac{n}{n-1} + \frac{n}{n-2} + ... + \frac{n}{1} \end{aligned}$$

The final term comes from the fact that the last k before the base case is $n - (n-1)$. Lets now factor n out of everything

$$B(0) = n(\frac{1}{n} + \frac{1}{n-1} + \frac{1}{n-2} + ... + 1)$$
$$B(0) = n(1 + \frac{1}{2} + \frac{1}{3} + ... + \frac{1}{n})$$
$$B(0) = n(\ln n + c),\ c \in (0, 1)$$

(c) [**5**] Suppose $n \geq 3$. Prove that the probability of the event V is not more than 25%, where $V =$ seeing all n stickers on buying **exactly** $8n \ln n$ boxes.

[*Hint*: Markov's inequality can be generally stated as: $\Pr[X \geq p] \leq \frac{\mathbb{E}[X]}{p}$. Can you see why?]

$$\Pr[B(t) \geq 8nln(n)] \leq \frac{\mathbb{E}[X]}{8nln(n)}$$
$$\Pr[B(t) \geq 8nln(n)] \leq \frac{n(ln(n)+c)}{8nln(n)}$$
$$\Pr[B(t) \geq 8nln(n)] \leq \frac{ln(n)+c}{8ln(n)}$$
$$\Pr[B(t) \geq 8nln(n)] \leq \frac{ln(n)+c}{8ln(n)} \stackrel{?}{\leq} 0.25$$

Lets simplify this further

$$\frac{ln(n)+c}{8ln(n)} \stackrel{?}{\leq} 0.25$$
$$ln(n)+c \stackrel{?}{\leq} 0.25 * 8ln(n)$$
$$ln(n)+c \stackrel{?}{\leq} 2ln(n) \iff c \leq ln(n)$$

We know $c \in (0,1)$ and $n \geq 3$

$$ln(n) \geq ln(3) > 1 > c$$

Bringing it all together

$$\Pr[B(t) \geq 8nln(n)] \leq \frac{1}{4}$$

Question 2: Birthdays and applications ..... **[25]**

Suppose we have n people, each of whom has their birthday on some random day of the year. Suppose there are m days in the year, and let us pretend that this is some parameter.

(a) **[10]** What is the expected *number of pairs* (i, j) with $i < j$ such that person i and person j have the same birthday? For what value of n (as a function of m) does this number become 1?

Define $X_{i,j}$ s.t. $X_{i,j} = 1$ if person i and person j share a birthday and 0 otherwise.
Define X to be the total number of pairs with matching birthdays

$$X = \sum_{i<j} X_{i,j}$$

Next we will use linearity of expectation which works even if the random variables aren't indepdendent.

$$E[X] = \sum_{i<j} E[X_{i,j}]$$

Lets find $E[X_{i,j}]$

$$E[X_{i,j}] = 1 * Pr[\text{i,j same birthday}] + 0 * Pr[\text{i,j not same birthday}]$$
$$E[X_{i,j}] = 1 * Pr[\text{i,j same birthday}]$$

To find this probability start by picking i's birthday. This Date can be any date so the probability of that being successful is 1 and then give the probability that j's birthday is this specific value.

$$Pr[\text{i,j same brithday}] = 1 * \frac{1}{m}$$
$$E[X_{i,j}] = \frac{1}{m}$$

The summation happens over all pairs $i < j$ so lets count the number of pairs

$$E[X] = \sum_{i<j} E[X_{i,j}] = \sum_{i}^{n} \sum_{j=i+1}^{n} E[X_{i,j}]$$

Say $i = 1$ when this happens j can take on values $(2, n)$ or $n - 1$ values. When $i = 2$ then j can take on values $(3, n)$ or $n - 2$ values. If $i = n - 1$ then exactly one pair can be made. If $i = n$ no pairs can be made. If we are trying to count the number of possible pairs such that $i < j$ it can represented by the arithmetic sequence below

$$(n-1) + (n-2) + ... + 1$$

This is an arithmetic series

$$S_{n-1} = \frac{n}{2}[2a + ((n-1) - 1)d]$$
$$S_{n-1} = \frac{n}{2}[2 * 1 + ((n-1) - 1 - 1)1]$$
$$S_{n-1} = \frac{n}{2}[2 + (n-3)]$$
$$S_{n-1} = \frac{n}{2}[n-1]$$
$$S_{n-1} = \frac{n * (n-1)}{2}$$

So there are S_{n-1} terms. Using that information we can solve our $E[X_{i,j}]$

$$E[X] = \sum_{i<j} E[X_{i,j}] = \sum_{i}^{n} \sum_{j=i+1}^{n} E[X_{i,j}] = \frac{n(n-1)}{2}\frac{1}{m}$$
$$E[X] = \frac{n^2 - n}{2} * \frac{1}{m}$$
$$E[X] = \frac{n^2 - n}{2m}$$

Solve for the value of n that makes the equation 1

$$n^2 - n = 2m$$
$$n^2 - n - 2m = 0$$

Use the quadratic formula

$$n = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$
$$n = \frac{1 \pm \sqrt{1 - 4 * -2m}}{2}$$
$$n = \frac{1 \pm \sqrt{1 + 8m}}{2}$$

(b) **[15]** This idea has some nice applications in CS, one of which is in estimating the "support" of a distribution. Suppose we have a radio station that claims to have a library of one million songs, and suppose that the radio station plays these songs by picking at each step a uniformly random song from its library (with replacement), playing it, then picking the next song, and so on.

A listener who started listening when the station began, noticed that the first 400 songs were distinct, and then a song played earlier played again (i.e., the 401^{st} song was a repetition). Prove that with probability $> 85\%$, the station's claim of having a million song library is **false**.

This problem can be mapped to the problem above. Let n be the number of songs that have been played and let m be the total number of songs there are. Define $X_{i,j} = 1$ if song i and song j are the same song. Using this we can define the number of song pairs that are the same song.

$$X = \sum_{i<j} X_{i,j}$$
$$E[X] = \frac{n^2 - n}{2m}$$

From the problem statement we can assume that $n = 401$ and that $m = 10^6$. We can plug these values in.

$$E[X] = \frac{401^2 - 401}{2 * 10^6}$$
$$E[X] = \frac{160400}{2000000}$$
$$E[X] = 0.0802$$

We can use Markov. We want to find the probability that $X > a$. This means that there was at least one repeat song so far.

$$P(X \geq a) \leq \frac{E[X]}{a}$$
$$P(X \geq 1) \leq 0.0802$$

If the station had a million songs the probability that the listener heard a repeat song after 401 songs would be 0.0802. Therefore because we saw a repeat we have confidence of at least $1 - 0.0802 \rightarrow 0.9198 > 0.85$ that the song claim is false

Question 3: Investment Analysis ..... **[20]**

A computer science student opens a Roth IRA retirement account at age 20. The student takes a side job and invests \$7,000 at the start of each year in the account. Not knowing how finances work, the student picks a standard investment fund that follows the S&P500. Based on the historical data, the rate of return in the i^{th} year r_i is a uniformly random percentage that ranges between 9.5% and 14.1%, i.e., $r_i \sim \text{Uniform}(9.5, 14.1)$. Answer each of the following questions by showing in detail how you arrive to the result.

[*Hint*: You may use the fact that for independent random variables A and B, $\mathbb{E}[AB] = \mathbb{E}[A] \cdot \mathbb{E}[B]$.]

(a) **[5]** Assuming an early retirement at the age of 57, express the amount of money M the student has in the account. Note that this corresponds to 37 years in terms of the random variables $r_1, r_2, \ldots, r_{37}$.

Define M_i to be the amount of money (captial) in the account at the end of year i
Move the decimal over for each r_i two times to get a decimal percentage

$$M_i = (M_{i-1} + 7000)(1 + r_i)$$
$$M_{37} = (M_{36} + 7000) * (1 + r_{37})$$

Lets now use plug and chug on the reccurence to see if we can find a pattern

$$M_1 = ((M_0 + 7000)(1 + r_1)) \rightarrow (7000)(1 + r_1)$$
$$M_2 = ((M_1 + 7000)(1 + r_2)) \rightarrow (7000(1 + r_1) + 7000) * (1 + r_2)$$
$$M_3 = ((M_2 + 7000)(1 + r_3)) \rightarrow$$
$$= ((7000(1 + r_1) + 7000) * (1 + r_2) + 7000)(1 + r_3)$$

Notice that the deposit in year 1 is multiplied by $(1 + r_1)(1 + r_2)(...)(1 + r_{37})$
Notice that the deposit in year 2 is multiplied by $(1 + r_2)(...)(1 + r_{37})$

$$M_{37} = \sum_{i=1}^{37} 7000 \prod_{j=i}^{37} (1 + r_j)$$
$$M_{37} = M$$

(b) **[10]** Find the expected amount in the account at the end of 37 years, i.e. $\mathbb{E}[M]$.

$$E[M] = \sum_{i=1}^{37} E[7000 \prod_{j=i}^{37} (1 + r_j)]$$
$$E[M] = \sum_{i=1}^{37} 7000 E[\prod_{j=i}^{37} (1 + r_j)]$$

We can abuse the fact that all r_i values are i.i.d

$$E[M] = \sum_{i=1}^{37} 7000 \prod_{j=i}^{37} E[(1 + r_j)]$$
$$E[M] = \sum_{i=1}^{37} 7000 \prod_{j=i}^{37} (E[1] + E[r_j])$$

$$E[M] = \sum_{i=1}^{37} 7000 \prod_{j=i}^{37} (1 + E[r_j])$$
$$E[M] = \sum_{i=1}^{37} 7000(1 + E[r])^{37-i+1}$$

The expected value of a continuous normal distribution is defined by the min(a) and max (b) values of our uniform distribution

$$E[X] = \frac{a+b}{2}$$
$$E[r] = \frac{0.095 + 0.141}{2}$$
$$E[r] = 0.118$$

Pulling everything together

$$E[M] = \sum_{i=1}^{37} 7000(1 + 0.6247)^{37-i+1}$$
$$E[M] = 4,045,218.42$$

(c) **[5]** Let $\sigma^2 = \text{Var}(r_i)$ be the variance of the rate of return, and $\mu = \mathbb{E}[r_i]$ be its expected value. Give an upper bound on the probability that $r_i \geq \mu + 2\sigma$.
We will use Chebyshev's Innequality Let X be the random variable that represents the rate of return

$$Pr[|X - \mu| \geq k\sigma] \leq \frac{1}{k^2}$$
$$Pr[|r_i - \mu| \geq 2\sigma] \leq \frac{1}{4}$$

Chebyshev says that the probability that a random variable X is at least k standard deviations away from the expected value is $\leq \frac{1}{k^2}$

$$Pr[r_i \geq \mu + 2\sigma]$$
$$Pr[r_i - \mu \geq 2\sigma]$$

Here we notice that the problem is only asking for the probability that the random variable is 2 standard deviations GREATER than the random variable and regular Chebyshev tells us the probability that the random variable is 2 standard deviations greater and smaller.
We are told that the distribution that the random variable comes from is a uniform distribution. This means that values are symmetric around the expected value. This means that we can simply divide the value given by Chebyshev 2 to get the correct probability.

$$Pr[r_i - \mu \geq 2\sigma] \leq \frac{1}{4}/2$$
$$Pr[r_i \geq \mu + 2\sigma] \leq \frac{1}{8}$$

(d) [**0**] Ungraded general knowledge: What percentage of the money invested will be lost in taxes when money will be withdrawn from the Roth IRA account?

Question 4: Checking feasibility vs optimization ..... **[20]**

Some of the algorithms for linear programming (e.g. simplex) start off with one of the corner points of the feasible set. This turns out to be tricky in general. In this problem, we will see that **in general**, finding one feasible point is as difficult as actually performing the optimization!

Consider the following linear program (in n variables $x_1, \ldots, x_n$, represented by the vector x):

$$\begin{aligned} &\text{minimize } c^T x \text{ subject to} \\ &a_1^T x \geq b_1 \\ &a_2^T x \geq b_2 \\ &\cdots \\ &a_m^T x \geq b_m. \end{aligned}$$

Suppose you know that the optimum value (i.e. the minimum of $c^T x$ over the feasible set) lies in the interval $[-M, M]$ for some real number M (this is typically possible in practice). Suppose also that you have an **oracle** that can take any linear program and say whether it is feasible or not. Prove that using $O(\log(M/\epsilon))$ calls to the oracle, one can determine the optimum value of the LP above up to an error of $\pm\epsilon$, for any given accuracy $\epsilon > 0$. [*Hint:* can you write a new LP that is feasible only if the LP above has optimum value $\leq z$, for some z?]

Start by defining a function O. This function returns true if the linear program passed in is feasible and false otherwise.

The hint asks if we can write a new LP that is feasible only if the LP above has optimum value $\leq z$ for some z. To do this simply add a new constraint the problem

$$c^T x \leq z$$

Define the linear program that contains all the constraints above and the one we just defined to be LP'

Next we want to figure out a way to ensure the number of calls to the oracle is $O(log(\frac{M}{\epsilon}))$

Because we know that the optimum value is in the interval $[-M, M]$ we can utilize binary search to get the bound we want.

Algorithm 1 Linear Program Binary Search

```
l ← −M
h ← M
while h − l ≤ 2ϵ do
    z ← (l+h)/2
    res ← O(LP′)
    if res is true then
        h ← z
    else if res is false then
        l ← z
```

Once this algorithm completes any value in the interval $[l, h]$ is a valid solution to the problem

Correctness: We saw in class that binary search is correct. The reasoning comes from the fact that we are always shrinking our interval without ever exlcuding the correct answer to the query.

Runtime Analysis The algorithm starts with an interval length of $2M$ (This is simply $h-l$). We know from class that at every step binary search cuts the search space in half. This means that after k iterations the length of the interval can be defined as follows.

$$h - l = \frac{2M}{2^k}$$

The problem asks us to be accurat within $\pm\epsilon$ which means that when our algorithm finishes our interval should be of size

$$h - l \leq 2\epsilon$$

We know the size of our interval after k iterations so we can plug that value in and solve for k

$$\begin{aligned}
h - l &\leq 2\epsilon \\
\frac{2M}{2^k} &\leq 2\epsilon \\
\frac{M}{2^k} &\leq \epsilon \\
M &\leq \epsilon * 2^k \\
\frac{M}{\epsilon} &\leq 2^k \\
log_2(\frac{M}{\epsilon}) &\leq k \log_2 2 \\
log_2(\frac{M}{\epsilon}) &\leq k
\end{aligned}$$

Therefore after $log_2(\frac{M}{\epsilon})$ iterations the size of our interval is 2ϵ which is what the problem asked us to show.

Question 5: Minimum vertex cover revisited ..... **[20]**

Recall the street surveillance problem that we saw in HW 2. We are given an undirected graph $G = (V, E)$ and the goal is to select a subset S of the vertices of the smallest possible size that can "monitor" all the edges, i.e., for every edge $\{i, j\} \in E$, at least one of i, j is in S (so the objective is to minimize $|S|$ subject to the above).

We studied the linear programming (LP) relaxation for vertex cover. Recall that it is as follows:

$$\begin{aligned}
&\text{minimize } \sum_{u \in V} x_u \text{ subject to} \\
&0 \leq x_u \leq 1 \text{ for all } u \in V \\
&x_u + x_v \geq 1 \text{ for all } \{u, v\} \in E.
\end{aligned}$$

In class, we saw a rounding algorithm that takes a feasible solution $\{x_u\}$ to the LP above and produces a **feasible, binary** solution whose objective value is at most $2\sum_u x_u$. Now, suppose we were lucky and the LP solution had all the x_u satisfying $x_u \in [0, 0.2) \cup (0.8, 1]$. In this case, prove that rounding produces a feasible, binary solution whose cost is at most $(1.25)\sum_u x_u$.

Let $X = x_1 + x_2 + ... + x_n$ represent the optimal solution.

Let $Y = y_1 + y_2 + ... + y_n$ represent the rounded solution.

Because of the interval given $y_u = 1 \iff x_u \geq 0.8$

Consider the lowest value of X that can be cause $y_u = 1$. This value is 0.81. This gives the innequality $y_u < 1.25 * x_u$

The objective for the optimal solution can be represented as $\min \sum x_u$

The objective for the rounded solution can be represented as $\min \sum y_u$

Expanding out the rounded solution and plugging in the innequality between y_u and x_u derived above

$$y_1 + y_2 + ... + y_n < 1.25x_1 + 1.25x_2 + ... + 1.25x_n$$
$$y_1 + y_2 + ... + y_n < 1.25(x_1 + x_2 + ... + x_n)$$
$$\min \sum_u y_u < 1.25 \sum_u x_u$$