

Exam location: Front of 1 Pimentel, Second to last digit of SID - 1, 4, or 8

PRINT your student ID:

--	--	--	--	--	--	--	--

PRINT AND SIGN your name:

(last)

(first)

(signature)

PRINT your Unix account login: cs70-\_\_\_\_\_

PRINT your discussion section and GSI (the one you attend):

Name of the person to your left:

Name of the person to your right:

Name of someone in front of you:

Name of someone behind you:

## Section 0: Pre-exam questions (3 points)

- 1. What are your plans for fun this winter break? (1 pt)**
- 2. Describe a time when you overcame an obstacle and succeeded. (2pts)**

Do not turn this page until the proctor tells you to do so.

PRINT your name and student ID: \_\_\_\_\_

SOME APPROXIMATIONS AND OTHER USEFUL TRICKS THAT MAY OR MAY NOT COME IN HANDY:

$$n! \approx \sqrt{2\pi n} \left(\frac{n}{e}\right)^n$$


$$\binom{n}{k} \leq \left(\frac{ne}{k}\right)^k$$

$$\lim_{n \rightarrow \infty} \left(1 + \frac{x}{n}\right)^n = e^x$$

When  $x$  is small,  $\ln(1+x) \approx x$

When  $x$  is small,  $(1+x)^n \approx 1+nx$

**The Golden Rule of 70 (and Engineering generally) applies: if you can't solve the problem in front of you, state and solve a simpler one that captures at least some of its essence. You might get partial credit for doing so, and maybe you'll find yourself on a path to the solution.**



### Probability Content from $-\infty$ to $Z$

$z$	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

Source: <http://www.math.unb.ca/~knight/utility/NormTble.htm>

PRINT your name and student ID: \_\_\_\_\_

## Section 1: Straightforward questions (30 points)

*You get one drop: do 5 out of the following 6 questions. Bonus for getting all six perfectly. No partial credit will be given.*

### 3. Prove It

Prove by induction that  $5^n - 1$  is divisible by 4 for all integers  $n \geq 1$ .

**Solution:** **Base case:**  $P(1) : 5^1 - 1 = 4$ , which is clearly divisible by 4.

**Inductive Hypothesis:** Suppose  $P(k) : 5^k - 1$  is divisible by 4 for some integer  $k > 1$ .

**Inductive Step:** We want to show that  $P(k) \Rightarrow P(k+1)$ .

By the Inductive Hypothesis, since  $5^k - 1$  is divisible by 4,  $5^k - 1 = 4l$  for some integer  $l$ .

$5^{k+1} - 1 = 5 \cdot 5^k - 1 = 5(5^k - 1) + 4 = 5(4l) + 4 = 4(5l + 1)$ , which means  $5^{k+1} - 1$  is also divisible by 4.

By the Principle of Induction, the original statement must hold. Q.E.D.

PRINT your name and student ID: \_\_\_\_\_

#### 4. Equate It

Use a combinatorial argument to prove that the following combinatorial identity is true (*i.e.*, provide a story for why the identity should hold).

$$n2^{n-1} = \binom{n}{1} + 2\binom{n}{2} + \cdots + (n-1)\binom{n}{n-1} + n\binom{n}{n}$$

##### **Solution:**

We want to count the number of ways that we can select a team with a single leader out of  $n$  people. We don't allow an empty team because we insist on having a single leader.

Left-hand side: From  $n$  people, pick one team leader and then some (possibly empty) subset of other people on his team.

Right-hand side: First pick  $k$  people on the team, then pick the leader among them. For example, when  $k$  is 2, we first choose our team of 2 people in  $\binom{n}{2}$  ways, then pick a leader in 2 ways because there are only 2 members to choose from.

Since the left-hand side and the right-hand side count the same thing, they are equal. Q.E.D.

PRINT your name and student ID: \_\_\_\_\_

### 5. Bound It

A random variable  $X$  is always strictly larger than  $-100$ . You know that  $E[X] = -60$ . Give the best upper bound you can on  $P(X \geq -20)$ .

**Solution:** Notice that we do not have the variance of  $X$ , so Chebyshev's bound is not applicable here. There is no upper bound on  $X$ , so Hoeffding's inequality cannot be used. We know nothing else about its distribution so we cannot evaluate  $E[e^{sX}]$  and so Chernoff bounds are not available. Since  $X$  is also not a sum of other random variables, other bounds or approximations are not available. This leaves us with just Markov's Inequality. But Markov Bound only applies on a nonnegative random variable, whereas  $X$  can take on negative values.

This suggests that we want to "shift"  $X$  somehow, so that we can apply Markov's Inequality on it. Define a random variable  $Y = X + 100$ , which means  $Y$  is strictly larger than  $0$ , since  $X$  is always strictly larger than  $-100$ . Then,  $E[Y] = E[X + 100] = E[X] + 100 = -60 + 100 = 40$ . Finally, the upper bound on  $X$  that we want can be calculated via  $Y$ , and we can now apply Markov's Inequality on  $Y$  since  $Y$  is strictly positive.

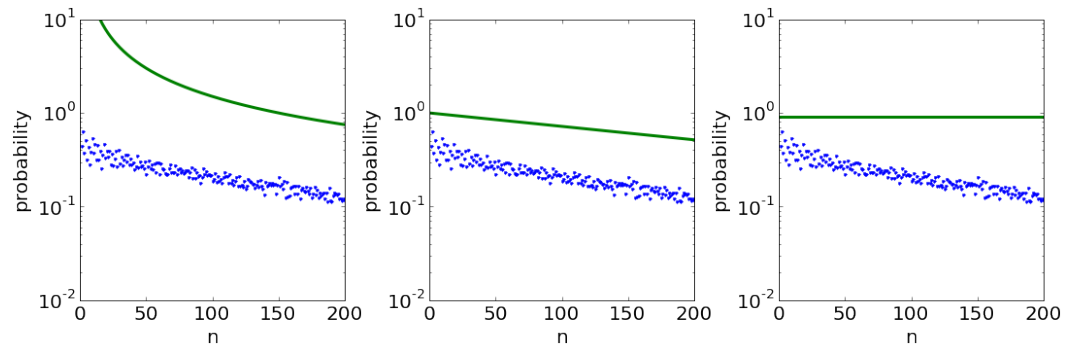
$$P(X \geq -20) = P(Y \geq 80) \leq \frac{E[Y]}{80} = \frac{40}{80} = \frac{1}{2}$$

Hence, the best upper bound on  $P(X \geq -20)$  is  $\frac{1}{2}$ .

## 6. Match It

We have  $n$  i.i.d. Bernoulli- $p$  random variables. Let  $A$  be the average of these random variables. The following plots are on log-scale and show various bounds and approximations for the probability of the average being more than 1.1 times the mean of  $A$  as a function of  $n$  (depicted in linear scale). The scatter points represent the actual probability of the deviation.

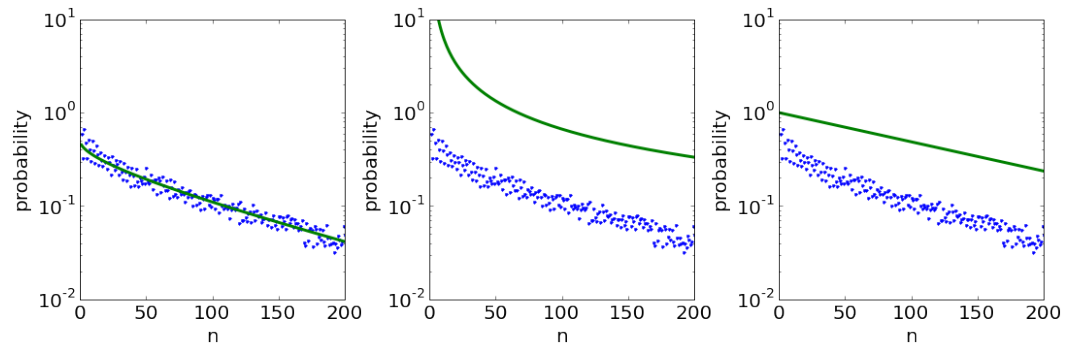
- (a) In this set of plots,  $p = 0.4$ . Label which one corresponds to Markov's Inequality, to Chebyshev's Inequality, and to a Chernoff Bound.



**Solution:** From left to right, Chebyshev's Inequality, Chernoff Bound, Markov's Inequality.

The dead give-away for Markov is that it doesn't get better with increasing  $n$ . The dead give-away for Chernoff is that it is a straight line of constant negative slope on such a plot with the horizontal axis in linear scale and the vertical axis in logarithmic scale. This leaves the first one as being Chebyshev's inequality. We can also see that Chebyshev's inequality can be larger than 1 while the other two will never be so when evaluating the probability of something being bigger than its mean.

- (b) In the next set of plots,  $p = 0.6$ . Label which one corresponds to Chebyshev's Inequality, to Hoeffding's Inequality, and to the Central Limit Theorem.



**Solution:** From left to right, Central Limit Theorem, Chebyshev's Inequality, Hoeffding's Inequality.

The dead give-away for Hoeffding is that it is a straight line of constant negative slope on such a plot with the horizontal axis in linear scale and the vertical axis in logarithmic scale. The dead give-away for the CLT is that it is an approximation and not a bound – so the points are on both sides of the CLT curve. This leaves the middle one as being Chebyshev's inequality. We can also see that Chebyshev's inequality can be larger than 1 while the other two will never be so when evaluating the probability of something being bigger than its mean.

PRINT your name and student ID: \_\_\_\_\_

### 7. Expect It

We call a couple a soulmate couple if both of them have each other as the first person on their preference lists. If there are  $n$  men and  $n$  women, and all of their preference lists are independently uniformly generated over all permutations, what is the expected number of soulmate couples?

**Solution:** This is similar to the *homeworks* example in Note 15. Let

$$X_i = \begin{cases} 1 & \text{if man } i \text{ is part of a soulmate couple} \\ 0 & \text{otherwise} \end{cases}$$

There is a probability of  $\frac{1}{n}$  that the woman at the top of man  $i$ 's list also has man  $i$  at the top of her list, therefore,

$$\begin{aligned} \mathbf{P}(X_i = 1) &= \frac{1}{n}, \\ \mathbf{E}(X_i) &= 1 \cdot \frac{1}{n} + 0 \cdot \frac{n-1}{n} = \frac{1}{n} \end{aligned}$$

The number of soulmate couples is equivalent to the number of men who belong to them, so the expected number of soulmate couples is

$$\begin{aligned} \mathbf{E}\left(\sum_{i=1}^n X_i\right) &= \sum_{i=1}^n \mathbf{E}(X_i) && \text{(By Linearity of Expectation)} \\ &= n \cdot \mathbf{E}(X_i) \\ &= n \cdot \frac{1}{n} \\ &= 1 \end{aligned}$$

□

PRINT your name and student ID: \_\_\_\_\_

### 8. Provoke It

In a demonstration/protest march, there are 2 kinds of people: civilians and undercover police. You know that 99% of the civilians are peaceful, and 1% advocate violence. You also know that 50% of the undercover police are agent provocateurs (who advocate violence) and 50% of them are peaceful. 2% of the protestors are undercover police. **Given that you see a person in the demonstration advocating violence, what is the probability that this person is undercover police?** (It is fine to leave the answer as a simplified fraction.)

**Solution:** We are given that people in the protest are either civilians or undercover police with the following probabilities.  $P(\text{Person is a civilian}) = P(C) = 0.98$  and  $P(\text{Person is an undercover police}) = P(UP) = 0.02$

We are also given that  $P(\text{Violent}|C) = 0.01$ ,  $P(\text{Peaceful}|C) = 0.99$ ,  $P(\text{Violent}|UP) = 0.5$ ,  $P(\text{Peaceful}|UP) = 0.5$ .

You spot a person advocating violence in the crowd. We want to find the probability that this person is an undercover police officer i.e.,  $P(UP|\text{Violent})$ .

Using Bayes' Rule we have

$$P(UP|\text{Violent}) = \frac{P(UP \cap \text{Violent})}{P(\text{Violent})} = \frac{P(\text{Violent}|UP) \times P(UP)}{P(\text{Violent})} \quad (1)$$

By the total probability rule we get,

$$P(\text{Violent}) = P(\text{Violent}|UP) \times P(UP) + P(\text{Violent}|C) \times P(C) = 0.5 \times 0.02 + 0.01 \times 0.98 = 0.0198 \quad (2)$$

Plugging Eq (2) and other given values to Eq (1), we get

$$P(UP|\text{Violent}) = \frac{0.5 \times 0.02}{0.0198} = \frac{50}{99}$$



PRINT your name and student ID: \_\_\_\_\_

## Section 2: Additional straightforward questions (27 points)

*You get one drop: do 3 out of the following 4 questions. Bonus for getting all of them perfectly. Very little partial credit will be given.*

### 9. Sections

An EECS class with  $3n$  students has three discussion sections, and each student attends exactly one of them by choosing one uniformly at random, independently of the others. **Use Stirling's approximation to estimate an  $n$  so that the probability of all three sections having *exactly* the same number of students is about 1.6%.**

It is fine to leave the answer as a simplified formula for  $n$ .

(Numerical hints for those who prefer actual numbers:  $2\pi \approx 6.28$ ,  $\sqrt{2\pi} \approx 2.51$ ,  $\sqrt{3} \approx 1.73$ .)

**Solution:** The number of ways that all three sections can have exactly the same number of students is the number of ways each section has  $3n/3 = n$  students,

$$\begin{aligned} \binom{3n}{n} \binom{2n}{n} \binom{n}{n} &= \frac{(3n)!}{n!n!n!} = \frac{(3n)!}{(n!)^3} \\ &\approx \frac{\sqrt{2\pi \cdot 3n} \left(\frac{3n}{e}\right)^{3n}}{\left(\sqrt{2\pi n} \left(\frac{n}{e}\right)^n\right)^3} && \text{(Stirling's Approximation)} \\ &= \frac{\sqrt{2\pi n} \cdot \sqrt{3} \cdot 3^{3n} \left(\frac{n}{e}\right)^{3n}}{\sqrt{2\pi n} \cdot (2\pi n) \cdot \left(\frac{n}{e}\right)^{3n}} \\ &= \frac{\sqrt{3} \cdot 3^{3n}}{2\pi n} \end{aligned}$$

The number of ways  $3n$  students can choose sections is  $3^{3n}$ , therefore,

$$\begin{aligned} \mathbf{P}(\text{All 3 sections have the same number of students}) &\approx \frac{\sqrt{3} \cdot 3^{3n}}{3^{3n} \cdot 2\pi n} \\ 0.016 &= \frac{\sqrt{3}}{2\pi n} \\ n &= \frac{\sqrt{3}}{2\pi \cdot 0.016} \\ n &\approx \frac{1.73}{6.28 \cdot 0.016} \approx 17.22 \end{aligned}$$

But since the number  $n$  has to be a positive integer, we take the floor and get 17. □

## 10. Two Face

Suppose you have two coins, one that has heads on both sides and another that has tails on both sides.

- (a) You pick one of the two coins uniformly at random and you flip that coin 400 times. Approximate the probability of getting more than 220 heads. Your answer should be a number that approximates this probability, accurate to 2 digits after the decimal point.

**Solution:** There are two equally likely outcomes of this experiment. Either you pick the double-headed coin, in which case you get 400 heads, or you pick the double-tailed coin, in which case you get 0 heads. Thus, the probability of getting more than 220 heads is 0.5.

- (b) You pick one of the two coins uniformly at random and flip it. You repeat this process 400 times, each time picking one of the two coins uniformly at random and then flipping it, for a total of 400 flips. Approximate the probability of getting more than 220 heads. Your answer should be a number that approximates this probability, accurate to 2 digits after the decimal point.

**Solution:** Let  $X$  be the number of heads. In each flip, you have probability  $\frac{1}{2}$  of picking the double-headed coin and getting heads and probability  $\frac{1}{2}$  of picking the double-tailed coin and not getting heads. The flips are independent, so  $X \sim \text{Bin}(400, \frac{1}{2})$ . Since  $\mathbb{E}(X) = 400 \cdot \frac{1}{2} = 200$  and  $\text{Var}(X) = 400 \cdot \frac{1}{2} \cdot \frac{1}{2} = 100$ . This gives a standard deviation of 10. So we are asking what is the probability of being two standard deviations or more away from the mean in the positive direction.

In other words, we can approximate  $X$  as a normal  $Y \sim N(400, 100)$ . Then

$$\Pr[X > 220] \approx \Pr[Y > 220] = \Pr\left[\frac{Y - 200}{10} > 2\right] = \Pr[Z > 2] = 1 - \Pr[Z \leq 2],$$

where  $Z$  is the standard normal. Thus,  $\Pr[X \geq 220] \approx 1 - 0.9772 = 0.0228$ .

PRINT your name and student ID: \_\_\_\_\_

- (c) Now you pick one of the two coins uniformly at random and flip it four times. You repeat this process 100 times, each time picking one of the two coins uniformly at random and then flipping it four times, for a total of 400 flips. Approximate the probability of getting more than 220 heads. Your answer should be a number that approximates this probability, accurate to 2 digits after the decimal point.

**Solution:** Let  $S$  be the number of times you pick the double-headed coin. Each time, you have probability  $\frac{1}{2}$  of picking the double-headed coin and probability  $\frac{1}{2}$  of picking the double-tailed coin, where the choice of coin is independent each time. Thus,  $S \sim \text{Bin}(100, \frac{1}{2})$ , and  $\mathbb{E}(S) = 100 \cdot \frac{1}{2} = 50$  and  $\text{Var}(S) = 100 \cdot \frac{1}{2} \cdot \frac{1}{2} = 25$ . So the standard deviation of  $S$  is 5. This means that the standard deviation for the number of heads  $4S$  is 20. The event we are interested in what is the chance that  $4S$  is more than 1 standard deviation away from the mean in the positive direction.

In other words, we can approximate  $S$  as a normal  $U \sim N(100, 25)$ . Let  $T$  be the total number of heads. Then  $T = 4S$ , so

$$\Pr[T > 220] = \Pr[S > 55] \approx \Pr[U > 55] = \Pr\left[\frac{U - 50}{5} > 1\right] = \Pr[Z > 1] = 1 - \Pr[Z \leq 1],$$

where  $Z$  is the standard normal. Thus,  $\Pr[T \geq 220] \approx 1 - 0.8413 = 0.1587$ .

On this question as well as the previous part, the indications to use the CLT were pretty clear. We want to approximate the probability, not bound it.

### 11. Estimating $\pi$

One can estimate  $\pi$  by playing darts with a special dartboard shown in figure 1. Assume every time you throw a dart, the dart will always be inside the square. The probability that your dart lands inside the circle is equal to the ratio of the area of the circle to the area of the square, i.e.,  $\frac{\pi}{4}$ . Let  $X_i$  be the random variable denoting whether your dart is within the circle after your  $i$ -th throw.

**How can you estimate  $\pi$  using this experiment? How many times should you throw to ensure your estimation error is within 0.01 with probability at least 95%? (You can just leave the numerical expression of the number of times but not compute the exact value.)**

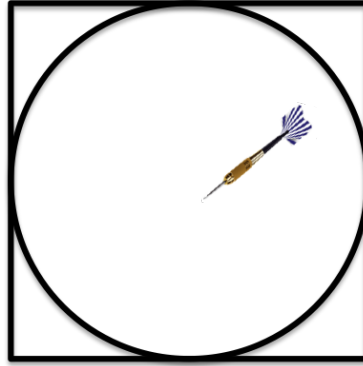


Figure 1: Dartboard.

**Solution:** Suppose we throw the dart  $n$  times in order to estimate  $\pi$ . We can estimate  $\pi$  by  $M_n = \frac{4}{n} \sum_{i=1}^n X_i$ .

Then we want  $P(|M_n - \pi| \geq 0.01) \leq 0.05$ .

$X_i$  is a bernolli random variable with  $P(X_i = 1) = \frac{\pi}{4}$ . The expectation of  $M_n$  is

$$\mathbf{E}[M_n] = \mathbf{E}\left[\frac{4}{n} \sum_{i=1}^n X_i\right] = \frac{4}{n} \sum_{i=1}^n \mathbf{E}[X_i] = \frac{4}{n} \cdot n \cdot \frac{\pi}{4} = \pi.$$

By Chebyshev's inequality:

The variance of  $M_n$  is

$$\text{Var}(M_n) = \text{Var}\left(\frac{4}{n} \sum_{i=1}^n X_i\right) = \frac{4^2}{n^2} \sum_{i=1}^n \text{Var}(X_i) = \frac{16}{n^2} \cdot n \cdot \frac{\pi}{4} \left(1 - \frac{\pi}{4}\right) = \frac{\pi(4-\pi)}{n}.$$

So

$$P(|M_n - \pi| \geq 0.01) \leq \frac{\text{Var}(M_n)}{0.01^2} = \frac{\pi(4-\pi)}{0.01^2 n}.$$

Let  $\frac{\pi(4-\pi)}{0.01^2 n} \leq 0.05$ , we get  $n \geq \frac{\pi(4-\pi)}{0.01^2 \cdot 0.05} = 539353.2$ . We have to throw at least 539354 times.

One may assume the exact value of  $\pi$  is unknown. Then we can use the inequality  $\pi(4-\pi) = -(\pi-2)^2 + 4 \leq 4$ . Let  $\frac{\pi(4-\pi)}{0.01^2 n} \leq \frac{4}{0.01^2 n} \leq 0.05$ , and we can get  $n \geq \frac{4}{0.01^2 \cdot 0.05} = 800000$ .

PRINT your name and student ID: \_\_\_\_\_

[Extra page. If you want the work on this page to be graded, make sure you tell us on the problem's main page.]

By **Hoeffding's inequality**:

$$P(|M_n - \pi| \geq 0.01) = P\left(\left|\frac{M_n}{4} - \frac{\pi}{4}\right| \geq 0.0025\right) \leq 2e^{-n \frac{2 \cdot 0.002^2}{(1-0)^2}}.$$

Let  $2e^{-n \frac{2 \cdot 0.0025^2}{(1-0)^2}} \leq 0.05$ , we get  $n \geq \ln \frac{0.05}{2} \cdot -\frac{(1-0)^2}{2 \cdot 0.0025^2} = 295110.36$ . We have to throw at least 295111 times.

## 12. Independent Shares

Consider a secret-sharing scheme over  $GF(5)$  that divides a secret among 4 people. Assume the secret is uniformly chosen from  $0,1,2,3,4$ . We use the standard polynomial-based secret-sharing scheme so that any 2 people can recover the secret. The linear coefficient in the polynomial is chosen uniformly from  $0,1,2,3,4$  and independently of the secret.

**Show that the shares (i.e. values obtained by evaluating the polynomial at their point) given to person 1 and person 2 are independent.**

**Solution:** Let the secret be  $X_0$  and the linear coefficient be  $X_1$ . These are chosen independently, so the sample space is the product-space of pairs  $(X_0, X_1)$ . Let the shares of person 1 and 2 be  $Y_1$  and  $Y_2$ , respectively. These are random variables, each a function of  $(X_0, X_1)$ . The secret-sharing scheme maps  $(X_0, X_1) \xrightarrow{f} (Y_1, Y_2)$  through some function  $f$  (ie, by evaluating the polynomial determined by  $X_0, X_1$ ). Notice that  $f$  is a **bijection**, since it maps between two unique representations of a degree-1 polynomial (the coefficient representation and the evaluation representation). Therefore, for given shares  $y_1, y_2$ , the event  $\{(Y_1, Y_2) = (y_1, y_2)\}$  is exactly the event  $\{f^{-1}(Y_1, Y_2) = f^{-1}(y_1, y_2)\} = \{(X_0, X_1) = f^{-1}(y_1, y_2)\}$ . Letting  $(x_0, x_1) = f^{-1}(y_1, y_2)$ , we find the distribution factors:

$$\Pr[(Y_1, Y_2) = (y_1, y_2)] = \Pr[(X_0, X_1) = (x_0, x_1)] = \Pr[X_0 = x_0] \Pr[X_1 = x_1] = 1/25$$

Where the second-to-last equality follows by the stipulated independence of the coefficients  $X_0$  and  $X_1$ . Therefore, the joint distribution of  $(Y_1, Y_2)$  is uniform on a square, so  $Y_1$  and  $Y_2$  are independent.

*Remarks:* The main idea here is that the two coefficients and two shares are equivalent ways of describing the same underlying “random degree-1 polynomial.” Any way of making this formal was acceptable (including more “brute-force” ways than the above.) It was insufficient to simply say “because one person’s share is hidden from another” or “because secret sharing works”, etc. The point of this problem was to prove exactly this.

PRINT your name and student ID: \_\_\_\_\_

### Section 3: True/False (30 points)

*You get one drop: do 3 out of the following 4 questions. Bonus for getting all four correct.*

*For each question in this section, determine whether the given statement is TRUE or FALSE. If TRUE, prove the statement. If FALSE, provide a counterexample or otherwise disprove it.*

#### 13. Conditional

If  $P(A) > P(B)$ , and  $P(C|A) > P(C|B)$ , then  $P(A|C) > P(B|C)$ .

Mark one: ☒ TRUE or FALSE.

**Solution:** The probability of any event lies in the range  $[0, 1]$ . Thus we have

$$P(A) > P(B) \geq 0 \tag{3}$$

$$P(C|A) > P(C|B) \geq 0 \tag{4}$$

Thus if we multiply the two inequalities Eq (3) and Eq (4), the sign of the inequality is preserved and we get,

$$P(A) \times P(C|A) > P(B) \times P(C|B)$$

$$P(A \cap C) > P(B \cap C)$$

$$P(C) \times P(A|C) > P(C) \times P(B|C)$$

Since  $P(A) > P(B) \geq 0$  and so  $P(C) \geq P(C \cap A) = P(A)P(C|A) > 0$ , we can divide through by  $P(C)$  from both sides of the inequality which preserving the sign to get

$$P(A|C) > P(B|C).$$

PRINT your name and student ID: \_\_\_\_\_

#### 14. Epsilon

Consider i.i.d. random variables  $\{X_i\}$  with mean  $\mu$  and variance  $\sigma^2$ . Let  $A_n = \frac{1}{n} \sum_{i=1}^n X_i$  be the average of  $n$  such random variables. Let  $g(\varepsilon, \delta)$  be the minimum  $n$  so that  $A_n$  is within  $\pm\varepsilon$  of  $\mu$  with probability at least  $(1 - \delta)$ . Then for  $\delta = 0.1$ , we can be certain that  $g(\frac{\varepsilon}{2}, \delta) \leq 2g(\varepsilon, \delta)$ .

Mark one: TRUE or FALSE.

##### Solution:

The conceptually easiest solution here makes a direct appeal to continuous random variables. (And this was accepted if done correctly.) Consider the case when the  $A_n$ s are exactly Gaussian/Normal random variables. For example, let  $X_i \sim N(0, 1)$ , so  $\mu = 0$ ,  $\sigma^2 = 1$ . Then  $A_n \sim N(0, 1/n)$ . Let  $\varepsilon$  be such that  $\Phi(\varepsilon) - \Phi(-\varepsilon) = 1 - \delta$ . That is, the probability of a standard normal being within  $\pm\varepsilon$  is  $1 - \delta$ . Therefore,  $g(\varepsilon, \delta) = 1$ . In order for  $A_n$  to be within  $\pm\varepsilon/2$  of its mean with probability  $1 - \delta$ , the standard deviation of  $A_n$  must be  $1/2$ . Notice the standard deviation of  $A_n$  is  $\frac{1}{\sqrt{n}}$ , so  $g(\frac{\varepsilon}{2}, \delta) = 4$ .

A much more elementary counterexample: Let each  $X_i$  be Bernoulli(1/2). So  $\mu = 1/2$ . Let  $\varepsilon = 1/2$ . Then  $g(\varepsilon, \delta) = 1$ , trivially. But notice that  $g(\varepsilon/2, \delta) > 2$ :

The distribution of  $A_2$  is: 0 w.p. 1/4; 1/2 w.p. 1/2; 1 w.p. 1/4.

So  $\Pr[|A_2 - \mu| \geq \varepsilon/2] = \Pr[|A_2 - 1/2| \geq 1/4] = 1/2 < 1 - \delta$ .

How could you have come up with the above counterexample? Just try fair coins as your default first example.

*Remarks:* Many students appealed to the  $\frac{1}{\sqrt{n}}$  scaling of the standard deviation, but argued using Chebyshev's inequality:  $\Pr[|A_n - \mu| \geq \varepsilon] \leq \frac{\sigma^2}{n\varepsilon^2}$ . Although this scaling turns out to be correct, Chebyshev can only give an  $n$  that is sufficiently high (not necessarily high), since it is a bound. We gave significant partial credit for such solutions, though they were not rigorous counterexamples.

In principle, it should be possible to give a generic counterexample to this by appealing to the Berry-Esseen inequality form of the CLT. But nobody seems to have done that.



PRINT your name and student ID: \_\_\_\_\_

### 15. Independence

If  $X, Y$  are random variables, and  $E[XY] = E[X]E[Y]$ , then  $X$  and  $Y$  are independent.

Mark one: TRUE or FALSE.

**Solution:** The converse is true, but the original statement is not true in general. Consider an example from Note 15, page 10. Let  $X$  be a fair coin toss that we consider as taking values  $+1$  and  $-1$  equally likely. Suppose  $Y$  is an independent fair coin toss that takes values  $+1$  and  $+2$ . The random variables  $X$  and  $Y$  are independent by construction.

Let's consider a new random variable  $Z = XY$ . Is  $Z$  independent of  $Y$ ? Obviously not.  $Z$  takes on four possible values  $-2, -1, +1, +2$  and the magnitude of  $Z$  reveals exactly what  $Y$  is. We also know that  $E[Y] = \frac{1}{2}(1+2) = \frac{3}{2}$ , and  $E[Z] = \frac{1}{4}(-2-1+1+2) = 0$ , so  $E[Y]E[Z] = 0$ . However,

$$\begin{aligned} E[YZ] &= \sum_y \sum_z yz \Pr[Y=y, Z=z] \\ &= \frac{1}{4}(1(-1) + 1(1) + 2(-2) + 2(2)) \\ &= \frac{1}{4}(-1 + 1 - 4 + 4) \\ &= 0, \end{aligned}$$

which means  $E[YZ] = E[Y]E[Z] = 0$ , but  $Y$  and  $Z$  are not independent. Hence, this is a counterexample to the original claim, and the statement must be false.

Many students made the mistake of using constant random variables. This cannot work because constants are independent of all variables. (Even themselves!) This is because  $0 \cdot x = 0, 1 \cdot x = x$ .

PRINT your name and student ID: \_\_\_\_\_

### 16. Gotta Get Them All

A person is trying to collect a set of  $2n$  cards:  $n$  distinct monster cards and  $n$  distinct spell cards. When buying a card, she gets a monster card with probability  $\frac{2}{3}$  and a spell card with probability  $\frac{1}{3}$ . Within each category, she will get a card uniformly at random. She keeps buying cards until she has a complete set (owns at least one of each card).

When  $n$  is large, the expected number of cards that she buys is less than or equal to  $3n(\ln(3n) + 1)$ .

Mark one: ☐ TRUE or FALSE.

**Solution:** Suppose for each distinct type of monster card, we can hypothetically imagine that it is either a female monster card or male monster card and they are with equal probability  $\frac{1}{2}$ . No matter whether you collect the female monster or the male monster of a given type of monster card, we count it as you have collected that type of monster card. Notice that in this model, the male monsters, female monsters, and spells are all equally likely.

Consider the following two scenarios:

1. The original problem becomes: we want to collect all distinct types of monster cards, **either** male or female monster in each type, and all spell cards. Let the number of cards we should buy is  $X$ .
2. Consider the case that we want to collect all distinct types of monster cards, including **both** male and female card in each type, and all spell cards. Let the number of cards we should buy is  $Y$ . So  $\mathbf{E}[Y] > \mathbf{E}[X]$ .

Instead of calculating  $\mathbf{E}[X]$ , let's calculate  $\mathbf{E}[Y]$ . The probability that you buy a female monster card of a given type is  $\frac{2}{3} \times \frac{1}{n} \times \frac{1}{2} = \frac{1}{3n}$ . Similarly, the probability that you buy a male monster card of a given type is also  $\frac{1}{3n}$ . The probability that you buy a given type of spell card is  $\frac{1}{3n}$  too. Then scenario 2 reduces to the classic coupon collector's problem: you have totally  $3n$  different type of cards ( $n$  female monster cards,  $n$  male monster cards and  $n$  spell cards) and the probability of buy each of them are both  $\frac{1}{3n}$ . So

$$\mathbf{E}[Y] \approx 3n(\ln(3n) + 0.5772) < 3n(\ln(3n) + 1).$$

Therefore  $\mathbf{E}[X] < \mathbf{E}[Y] < 3n(\ln(3n) + 1)$ .

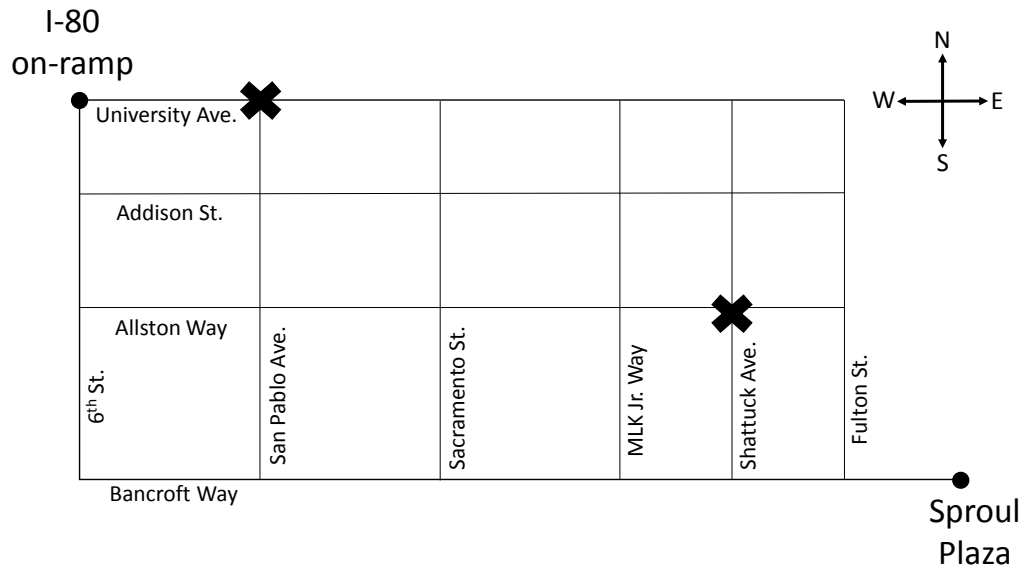
PRINT your name and student ID: \_\_\_\_\_

## Section 4: Free-form Problems (73 + 25 points)

### 17. Peaceful Paths (25 points)

A group of protesters are meeting at Sproul Plaza to march to the I-80 on-ramp. The group doesn't want to backtrack and lose momentum, so they will only move North or West.

For each part, simplify your answer.



- (a) (5 points) If the protesters stick to the main paths depicted in the map above, how many possible paths can they take to get from Sproul Plaza (marked by a dot) to the I-80 on-ramp (marked by another dot)?

**Solution:** Note that the protesters have no choice but to walk from Sproul to the intersection of Bancroft and Fulton. From there, they need to make 8 moves total, and need to choose 5 of them to be in the West direction, so there are  $\binom{8}{5} = \binom{8}{3} = 56$  possible paths.

PRINT your name and student ID: \_\_\_\_\_

- (b) (5 points) The police have set up blockades (as marked on the map by x's) at the intersection of Allston and Shattuck near the Berkeley BART station and at the intersection of University and San Pablo, just before the on-ramp. How many possible paths can the protesters take such that the police blockades are avoided?

**Solution:**  $\binom{2}{1} \binom{6}{4} = 30$  go through the first blockade,  $\binom{7}{4} = 35$  go through the second blockade, and  $\binom{2}{1} \binom{5}{2} = 20$  go through both. Therefore the total number of paths they can take is  $56 - (30 + 35 - 20) = 11$ .

- (c) (5 points) Assume the protesters do not know about the blockades, so they are equally likely to choose any path. What is the probability that they are intercepted by a police blockade?

**Solution:** Since the probability of taking each path is uniform, probability they don't hit a police blockade is  $\frac{11}{56}$ , and thus the probability that they do hit a police blockade is  $1 - \frac{11}{56} = \frac{45}{56}$ .

PRINT your name and student ID: \_\_\_\_\_

- (d) (5 points) In order to increase the chances of some protesters making it to the freeway, the protesters decide to split up into two groups. Each group independently chooses a path to take. As before, either group is equally likely to choose any of the paths. Let  $X$  be a random variable denoting the number of groups that make it to the I-80 on-ramp without being intercepted by police. **Write the probability distribution of  $X$ .**

(For this part, feel free to use the constants  $a, b, c$  to refer to the correct answers to parts a,b,c of this problem.)

**Solution:** Since there are two groups,  $X$  can take on the values 0, 1, or 2. We have

$$\begin{aligned}Pr(X = 0) &= \left(\frac{45}{56}\right)^2 = c^2 = \left(1 - \frac{b}{a}\right)^2 \\Pr(X = 1) &= 2 \left(\frac{11}{56}\right) \left(\frac{45}{56}\right) = 2c(1 - c) = 2\frac{b}{a} \left(1 - \frac{b}{a}\right) \\Pr(X = 2) &= \left(\frac{11}{56}\right)^2 = (1 - c)^2 = \frac{b^2}{a}\end{aligned}$$

- (e) (5 points) Now suppose the protesters want to split up into  $n > 2$  groups instead of just 2. As before, each of the  $n$  groups selects a path independently, and will choose each path with equal probability. **What is the minimum number of groups  $n$  needed such that the expected number of groups that reach the I-80 on-ramp without being blocked is at least 1?**

(For this part, feel free to use the constants  $a, b, c$  to refer to the correct answers to parts a,b,c of this problem.)

**Solution:** Let  $Y$  be the number of groups that reach the I-80 on-ramp without being blocked by police. We have that  $Y$  is Binomial( $n, 11/56$ ). We know the formula for the expected value of a Binomial R.V., so we have  $E(Y) = n(11/56)$ . If we want  $E(Y) \geq 1$ , we just solve for  $n$ :

$$\begin{aligned}E(Y) &= n \left(\frac{11}{56}\right) \geq 1 \\n &\geq \frac{56}{11} \\n &= \left\lceil \frac{56}{11} \right\rceil = 6\end{aligned}$$

PRINT your name and student ID: \_\_\_\_\_

[Extra page. If you want the work on this page to be graded, make sure you tell us on the problem's main page.]

We said you could write this in terms of answers to a,b,c. In this case, you'd get

$$\begin{aligned} E(Y) &= n(1-c) \geq 1 \\ n &\geq \frac{1}{1-c} \\ n &= \left\lceil \frac{1}{1-c} \right\rceil, \end{aligned}$$

or, since  $1-c = b/a$ ,

$$\begin{aligned} E(Y) &= n \left( \frac{b}{a} \right) \geq 1 \\ n &\geq \frac{a}{b} \\ n &= \left\lceil \frac{a}{b} \right\rceil. \end{aligned}$$

PRINT your name and student ID: \_\_\_\_\_

### 18. Couples (15 points)

You are trying to send a message of length 100, and every odd packet drops independently with probability 0.1, and every even packet drops independently with probability 0.5.

- (a) (10 points) If you want your message to be successfully received with probability approximately 95%, **about how many additional packets do you have to send using a Reed-Solomon error-correcting code?** Feel free to leave the answer in a form of a simple algebraic equation to be solved.

**Solution:** The name of the problem is an implicit hint. Look at the packets in couples: even and odd.

Suppose that we send  $m$  total packets (where  $m$  is even for convenience; if not, add one, since this is only an approximation). Half of the packets (the odd ones) drop with probability 0.1, and the other half drop with probability 0.5. So, the expected number of drops is  $0.1 \cdot \frac{m}{2} + 0.5 \cdot \frac{m}{2} = 0.3m$ .

Also, since each odd packet is a Bernoulli random variable (dropped or not), the variance for each odd packet is  $0.1 \cdot 0.9 = 0.09$ , and the variance for each packet is  $0.5 \cdot 0.5 = 0.25$ , for a total variance of  $0.09 \cdot \frac{m}{2} + 0.25 \cdot \frac{m}{2} = 0.17m$ , or a standard deviation of  $\sqrt{0.17m}$ .

Now, we can use the Central Limit Theorem to approximate the probability of dropping too many packets. Since we need at least 100 packets to make it through, we can drop at most  $m - 100$  packets. Examining the normal distribution table, we find that with 95% probability, a normal distribution takes on values at most 1.65 standard deviations above the mean. Therefore, we want  $m - 100$  packets to be 1.65 standard deviations above the mean: this will mean that with 95% probability, we drop at most  $m - 100$  packets (and successfully reconstruct). That is to say, we want

$$0.3m + 1.65\sqrt{0.17m} \approx m - 100$$

Also, the number of *additional* packets we send is just 100 less than the number of *total* packets; so if we send  $n$  additional packets, then  $n + 100 = m$ . Substituting in, we get:

$$0.3(n + 100) + 1.65\sqrt{0.17(n + 100)} \approx n$$

This is basically a quadratic equation to solve for  $n$ . The resulting  $n$  is about 155.

PRINT your name and student ID: \_\_\_\_\_

- (b) (5 points) **Is the answer to part (a) more or less than what it would be if the channel dropped packets with probability 0.3 at every time? Why?**

**Solution:** Suppose that we dropped packets with probability 0.3 every time. Then the expected number of drops is the same; however, the variance is  $0.3 \cdot 0.7m = 0.21m$ , which is higher. Of course, the standard deviation is therefore also higher. We need to send enough packets so that the expected number of received packets is 100, *plus 1.65 standard deviations*. So, a larger standard deviation corresponds to needing to send more packets. Therefore, we need to send more packets in part (b), i.e. less in part (a).

The actual  $n$  for part (b) is about 157. So, the difference is not that large. But it is there.



PRINT your name and student ID: \_\_\_\_\_

**19. Miley the Lumberjack (15 points)**

For a sequence of numbers  $a_1, a_2, \dots, a_n$ , the geometric mean is equal to  $(a_1 a_2 \cdots a_n)^{\frac{1}{n}}$ . Consider random variables  $X_i$  that take on the values

$$X_i = \begin{cases} 2 & \text{with probability } \frac{1}{2}(1-p) \\ 4 & \text{with probability } p \\ 8 & \text{with probability } \frac{1}{2}(1-p) \end{cases}$$

For an i.i.d. sequence of random variables  $X_1, X_2, \dots$  distributed as above with  $1 > p > 0$ ,

$$\lim_{n \rightarrow \infty} P \left( \frac{1}{1+\varepsilon} < \frac{(X_1 X_2 \cdots X_n)^{\frac{1}{n}}}{4} < 1+\varepsilon \right) = 0$$

for every  $\varepsilon > 0$ .

Mark one: TRUE or FALSE.

If TRUE, prove the statement. If FALSE, provide a counterexample or otherwise disprove it.

**Solution:** This problem is similar to Homework 14's *Wrecking Ball* in the sense that the expectation of  $\frac{(X_1 X_2 \cdots X_n)^{\frac{1}{n}}}{4}$  is not the value it is converging to. The reason is simple: expectation is a sum average and it is not meaningful when we are trying to predict the value of independent random variables being multiplying together. Students who drew conclusions about the converged value from the expectation of  $X_i$ 's alone will not get partial credit.

To use expectations, we can define a new variable that will lead to a summation that we can work on. Notice that the values of  $X_i$ 's are just powers of two, and when they are multiplied together, their exponents add up. Let  $Y_i$  be the relevant exponent corresponding to  $X_i$ , i.e.,  $Y_i = \log_2 X_i$ . We have,

$$\begin{aligned} \mathbf{E}(Y_i) &= 1 \cdot \frac{1}{2}(1-p) + 2p + 3 \cdot \frac{1}{2}(1-p) \\ &= 2 - 2p + 2p = 2 \end{aligned}$$

According to the definition of  $X_i$ ,  $\frac{(X_1 X_2 \cdots X_n)^{\frac{1}{n}}}{4}$  is always positive, so it is valid to define

$$Z = \log_2 \left( \frac{(X_1 X_2 \cdots X_n)^{\frac{1}{n}}}{4} \right) = \frac{1}{n} \left( \sum_{i=1}^n \log_2 X_i \right) - \log_2 4 = \frac{1}{n} \left( \sum_{i=1}^n Y_i \right) - 2,$$

then (aside: this is why Miley is a lumberjack in this problem. She likes logs.)

$$\begin{aligned} \mathbf{E}(Z) &= \mathbf{E} \left( \frac{1}{n} \sum_{i=1}^n Y_i - 2 \right) \\ &= \frac{1}{n} \sum_{i=1}^n \mathbf{E}(Y_i) - 2 = \frac{1}{n} \cdot n \cdot 2 - 2 = 0 \end{aligned}$$

PRINT your name and student ID: \_\_\_\_\_

[Extra page. If you want the work on this page to be graded, make sure you tell us on the problem's main page.]

$\mathbf{E}(Z) = 0$  gives us the intuition that  $\frac{(X_1 X_2 \cdots X_n)^{\frac{1}{n}}}{4}$  should converge to  $2^0 = 1$ , which is in the range  $(\frac{1}{1+\varepsilon}, 1+\varepsilon)$ . Formally, since logarithm is a monotonically increasing function and both  $1+\varepsilon$  and  $\frac{1}{1+\varepsilon}$  are positive,

$$\begin{aligned} \mathbf{P}\left(\frac{1}{1+\varepsilon} < \frac{(X_1 X_2 \cdots X_n)^{\frac{1}{n}}}{4} < 1+\varepsilon\right) &= \mathbf{P}\left(\log_2\left(\frac{1}{1+\varepsilon}\right) < \log_2\left(\frac{(X_1 X_2 \cdots X_n)^{\frac{1}{n}}}{4}\right) < \log_2(1+\varepsilon)\right) \\ &= \mathbf{P}\left(\log_2\left(\frac{1}{1+\varepsilon}\right) < Z < \log_2(1+\varepsilon)\right) \\ &= \mathbf{P}(-\log_2(1+\varepsilon) < Z < \log_2(1+\varepsilon)) \\ &= \mathbf{P}(|Z| < \log_2(1+\varepsilon)) \\ &= \mathbf{P}(|Z - \mathbf{E}(Z)| < \log_2(1+\varepsilon)) \quad (\mathbf{E}(Z) = 0) \end{aligned}$$

By the Law of Large Numbers,

$$\lim_{n \rightarrow \infty} \mathbf{P}(|Z - \mathbf{E}(Z)| \geq \log_2(1+\varepsilon)) = 0$$

Therefore,

$$\begin{aligned} \lim_{n \rightarrow \infty} \mathbf{P}\left(\frac{1}{1+\varepsilon} < \frac{(X_1 X_2 \cdots X_n)^{\frac{1}{n}}}{4} < 1+\varepsilon\right) &= \lim_{n \rightarrow \infty} \mathbf{P}(|Z - \mathbf{E}(Z)| < \log_2(1+\varepsilon)) \\ &= 1 - \lim_{n \rightarrow \infty} \mathbf{P}(|Z - \mathbf{E}(Z)| \geq \log_2(1+\varepsilon)) \\ &= 1 - 0 \\ &= 1 \end{aligned}$$

Hence, the claim is FALSE. □

**Note:** Instead of invoking the Law of Large Numbers, one can use Chebyshev inequality directly.

$$\begin{aligned} \mathbf{E}(Y_i^2) &= 1 \cdot \frac{1}{2}(1-p) + 4p + 9 \cdot \frac{1}{2}(1-p) \\ &= 5(1-p) + 4p = 5 - p, \\ \text{Var}(Y_i) &= \mathbf{E}(Y_i^2) - \mathbf{E}(Y_i)^2 \\ &= 5 - p - 2^2 = 1 - p \\ \text{Var}(Z) &= \text{Var}\left(\frac{1}{n} \sum_{i=1}^n Y_i - 2\right) = \frac{1}{n^2} \sum_{i=1}^n \text{Var}(Y_i) - \text{Var}(2) \quad (\text{By independence of } Y_i\text{'s}) \\ &= \frac{1}{n^2} \cdot n(1-p) - 0 = \frac{1-p}{n} \end{aligned}$$

By Chebyshev inequality,

$$\mathbf{P}(|Z - \mathbf{E}(Z)| \geq \log_2(1+\varepsilon)) \leq \frac{\text{Var}(Z)}{(\log_2(1+\varepsilon))^2} = \frac{1-p}{n(\log_2(1+\varepsilon))^2}$$

Since  $n$  is in the denominator,

$$\lim_{n \rightarrow \infty} \mathbf{P}(|Z - \mathbf{E}(Z)| \geq \log_2(1+\varepsilon)) = 0$$

□

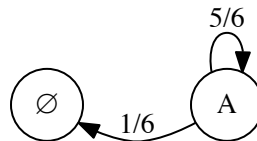
## 20. Markov Dice (18 points)

James Bond is playing a game that involves repeatedly rolling a fair standard 6-sided die.

- (a) (3 points) What is the expected number of rolls until he gets a 5?

**Solution:** Let  $X$  be the random variable denoting the number of rolls until 007 gets a 5, so  $E[X]$  is the expected number of rolls. Notice that he must roll at least once. On his first try, he gets a 5 with probability  $\frac{1}{6}$ , and fails with probability  $\frac{5}{6}$ , which means he has to start all over again (but with one additional roll under his belt).

Similar to the “Markov Conversations” problem from HW 12, this can be represented by the following Markov Chain (where state A represents “last roll not a 5”, and state  $\emptyset$  represents game termination): Starting from state A, we want to find the expected rolls until he hits  $\emptyset$ .



By the reasoning above, we can express  $E[X]$  in a recursive fashion and solve for it.

$$\begin{aligned}
 E[X] &= \frac{1}{6} \cdot 1 + \frac{5}{6}(E[X] + 1) \\
 &= \frac{1}{6} + \frac{5}{6} + \frac{5}{6}E[X] \\
 \frac{1}{6}E[X] &= 1 \\
 E[X] &= 6
 \end{aligned}$$

which means the expected number of rolls until 007 gets a 5 is  $\boxed{6}$ . Notice that it doesn't matter what the roll value he wants is (any other number between 1 and 6 that is not 5), it is expected that Bond tries 6 times before he gets the desired number.

You could solve this question by noting that the number of rolls is a geometric random variable with parameter  $p = \frac{1}{6}$ , and since we know that the expected value of a geometric random variable is  $\frac{1}{p}$ , you will also get the same correct answer of 6.

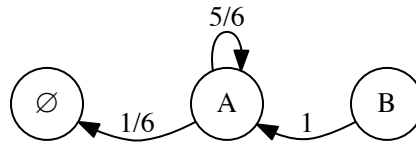
You could also solve this question by writing out the sum of a geometric series and calculating that sum, but this method would take the longest amount of time, given that this is only a 3-point question. The recursive approach will be the most useful way to think about this problem, as we will see in the next two parts.

- (b) (5 points) What is the expected number of rolls until the last two rolls sum to 7?

**Solution:** Intuitively, the answer to this question is the same as part (a), plus one. Why? First, notice that no matter what your “first” roll is, there is only a  $\frac{1}{6}$  chance of making the sum of that roll and the roll immediately after it equal 7. This is because every number between 1 and 6 has a unique “partner”

that sums to 7, e.g. 1 will be paired with 6, 2 with 5, etc. Second, notice that 007 cannot have a pair of rolls until he has rolled at least twice, so we need to always roll once first, and then we can calculate the desired expected number of rolls starting from the second roll.

Again, this can be modeled by the following Markov chain, where state A represents at least one previous roll has taken place, and state B represents the start of the game (no previous rolls). We want the expected rolls to hit  $\emptyset$ , starting from state B.



Mathematically, let  $\beta$  be the expected number of rolls, and  $\alpha$  be the expected number of rolls after the first roll. This implies that  $\beta = 1 + \alpha$ , as explained above. To calculate  $\alpha$ , note that there are two cases. The second roll of the pair either “matches” the first one (so their sum is 7) with probability  $\frac{1}{6}$  or it doesn’t (with probability  $\frac{5}{6}$ ). Either way it costs 007 an additional roll to find out. In the first case, the game ends, so the additional expected number of rolls is 0. In the second case, the additional expected number of rolls is just  $\alpha$ , since this so-called “failed” second roll is now the first roll and basically waits for the next roll to match it.

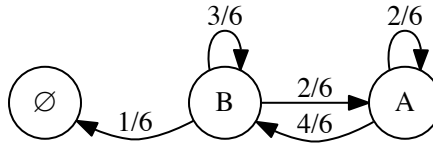
In other words,  $\alpha = 1 + \frac{5}{6}\alpha$ . Solving for  $\alpha$  gives 6 as expected. However, recall that we are interested in the number of rolls, and we must always roll once to actually start the game (since the event that we’re interested in is the sum of the last two rolls). Therefore, the expected number of rolls until the last two rolls sum to 7 is  $\beta = 1 + \alpha = 1 + 6 = \boxed{7}$ .

PRINT your name and student ID: \_\_\_\_\_

(c) (10 points) What is the expected number of rolls until the last two rolls sum to 9?

**Solution:** Notice that this question is trickier than the previous part. If 007 rolls a 1 or a 2 as his first roll, then no matter what the second roll's value is, the sum cannot be 9. Intuitively, the expected number of rolls in this question has to be larger than the answer we get for part (b). For the sum of the last two rolls to be 9, both rolls must be greater than or equal to 3.

Consider the following Markov Chain. State A represents the previous roll is 1,2, or nonexistent (at the beginning), and state B represents the previous roll is 3-6. We want the expected time to hit  $\emptyset$ , starting from state A.



Let's use the following two definitions:

- 1) Let  $\alpha$  be the average remaining number of rolls given you have just rolled a 1 or a 2.
- 2) Let  $\beta$  be the average remaining number of rolls given you have just rolled a number between 3 and 6.

When 007 first begins, which is the same as when he's at the  $\alpha$  "stage", he either continues to be stuck at  $\alpha$  with probability  $\frac{1}{3}$  (rolling a 1 or a 2) and one additional roll, or he moves on to the  $\beta$  "stage" with probability  $\frac{2}{3}$  (rolling a number between 3 and 6). Once he reaches  $\beta$ , there is a  $\frac{1}{6}$  probability of having the sum equal 9, a  $\frac{1}{3}$  chance that he goes back to  $\alpha$  if he rolls a 1 or a 2, and a  $\frac{1}{2}$  chance of continuing at  $\beta$  (rolling a number at least 3, but together with the previous roll does not sum up to 9). Hence, we can write the following recursive equations in terms of  $\alpha$  and  $\beta$ .

$$\alpha = (1/3)(\alpha + 1) + (2/3)(\beta + 1)$$

$$\beta = 1/6 \cdot 1 + (1/3)(\alpha + 1) + (1/2)(\beta + 1)$$

Solving this gives  $\alpha = 10.5$ ,  $\beta = 9$ .

Since 007 starts without any previous roll, it's the same as rolling a 1 or a 2 in the previous roll. In other words, it takes on average  $\alpha = \boxed{10.5}$  rolls for the sum to equal 9.

PRINT your name and student ID: \_\_\_\_\_

## 21. (Optional) Digits (25 points)

You choose a number from 0 to 999999 (inclusive) uniformly at random and you sum the digits up. For example if you were to choose 345, the sum of the digits would be 12.

(a) (5 points) What is the probability that the sum is 9?

**Solution:** Let  $x_1, x_2, \dots, x_6$  be variables corresponding to each digit in a number. Then we calculate the number of ways to assign these variables such that they sum to 9, i.e. the number of solutions to

$$x_1 + x_2 + x_3 + x_4 + x_5 + x_6 = 9, \quad 0 \leq x_1 \leq 9, \dots, 0 \leq x_6 \leq 9.$$

We have seen this problem in Discussion 10M, which suggests using stars and bars. We have 9 stars to be distributed, and 5 bars separating the 6 locations, so the number of 6-digit numbers whose sum is 9 is simply  $\binom{9+6-1}{9} = \binom{14}{9}$ . There are also 1,000,000 numbers between 0 and 999999, and we choose a number at random, so the probability that the sum of a randomly chosen 6-digit number is 9 is:

$$\frac{\binom{14}{9}}{1,000,000} = 2.002 \times 10^{-3}$$

You might be wondering what if the first digit is 0. For this problem (and the stars-and-bars approach), a number like 345 can be thought of as a 6-digit integer 000345. In balls-and-bins terminology, this is the same as throwing no balls into the first three bins, then throwing 3 balls into the fourth bin, and so on. In other words, it's completely fine to have any digit to be 0. If the chosen number is not 6-digit, think of it as adding extra imaginary zero(es) at the front to make the number 6-digit.

(b) (10 points) What is the probability that the sum is 19?

**Solution:** Let  $x_1, x_2, \dots, x_6$  be variables corresponding to each digit in a number. Then we calculate the number of ways to assign these variables such that they sum to 19, i.e. the number of solutions to

$$x_1 + x_2 + x_3 + x_4 + x_5 + x_6 = 19, \quad 0 \leq x_1 \leq 9, \dots, 0 \leq x_6 \leq 9.$$

Using stars and bars, the number of solutions is  $\binom{24}{5}$ . But this also counts the number of solutions for which some  $x_i$  is 10 or higher, and we can't have a digit value greater than 9. Thus we need to subtract the number of solutions for which some  $x_i$  is  $\geq 10$ . Notice that, because  $19 < 10 + 10$ , we cannot have two such  $x_i$ 's, so we just need to count the solutions  $(x_1, \dots, x_6)$  where exactly one of the  $x_i$ 's is 10 or higher.

Subtracting 10 from such an  $x_i$  and from the total sum, we obtain the equation

$$x'_1 + x'_2 + x'_3 + x'_4 + x'_5 + x'_6 = 9, \quad x'_1 \geq 0, \dots, x'_6 \geq 0.$$

There are  $\binom{14}{9}$  solutions to this equation, which we obtain from part (a). But since there are six choices for which variable is greater than 10, the total amount we over-counted by is  $6 \cdot \binom{14}{5}$ . Thus the total number of six digit numbers whose digits sum to 19 is  $\binom{24}{5} - 6 \cdot \binom{14}{5}$ . The probability that a random six digit number has digits that sum to 19 then is

$$\frac{\binom{24}{5} - 6\binom{14}{5}}{1,000,000} = \frac{42504 - 6 \times 2002}{1,000,000} = 0.030492$$

PRINT your name and student ID: \_\_\_\_\_

(c) (10 points) What is the probability that the sum is 29?

**Solution:** Let  $x_1, x_2, \dots, x_6$  be variables corresponding to each digit in a number. Then we calculate the number of ways to assign these variables such that they sum to 29, i.e. the number of solutions to

$$x_1 + x_2 + x_3 + x_4 + x_5 + x_6 = 29, \quad 0 \leq x_1 \leq 9, \dots, 0 \leq x_6 \leq 9.$$

Using stars and bars, the number of solutions is  $\binom{34}{5}$ . But this also counts the number of solutions for which some  $x_i$  is 10 or higher, and we can't have a digit value greater than 9. Thus we need to subtract the number of solutions for which some  $x_i$  is  $\geq 10$ .

Specifically, when  $x_k \geq 10$  we can express it as  $x_k = 10 + y_k$ . For all other  $j \neq k$  write  $y_j = x_j$ . The number of ways to arrange 29 amongst  $x_i$  when some  $x_k \geq 10$  is the same as the number of ways to arrange  $y_i$  so that  $\sum_{i=1}^6 y_i = 29 - 10 = 19$  is  $\binom{24}{5}$ . There are 6 possible ways for some  $x_k \geq 10$  so there are a total of  $6\binom{24}{5}$  ways for some digit to be greater than or equal to 10.

However, the above counts events multiple times. For instance,  $x_1 = x_2 = 10$  is counted both when  $x_1 \geq 10$  and when  $x_2 \geq 10$ . We need to account for these events that are counted multiple times. We can consider when two digits are greater than or equal to 10:  $x_j \geq 10$  and  $x_k \geq 10$  when  $j \neq k$ . Let  $x_j = 10 + y_j$  and  $x_k = 10 + y_k$  and  $x_i = y_i \forall i \neq j, k$ . Then the number of ways to distribute 29 amongst  $x_i$  when there are 2 greater than or equal to 10 is equivalent to the number of ways to distribute  $y_i$  when  $\sum_{i=1}^6 y_i = 29 - 10 - 10 = 9$ . There are  $\binom{14}{5}$  ways to distribute these  $y_i$  and there are  $\binom{6}{2}$  ways to choose the possible two digits that are greater than or equal to 10.

We are interested in when the sum of  $x_i$  is equal to 29. So we can have at most 2  $x_i$  greater than or equal to 10. So we are done. Thus there are  $\binom{34}{5} - 6\binom{24}{5} + \binom{6}{2}\binom{14}{5}$  numbers between 0 through 999999 whose digits sum up to 29. The probability that a random six digit number has digits that sum to 29 then is

$$\frac{\binom{34}{5} - 6\binom{24}{5} + \binom{6}{2}\binom{14}{5}}{1,000,000} = \frac{53262}{1,000,000} = 0.053262$$