EECS 70 Discrete Math and Probability
Spring 2020 UC Berkeley Midterm 2

1. Fun with Letters (10 pts)

(a) (5pts) **How many ways are there to arrange m A's, n B's and 1 C in a circle on a piece of paper?** Two letter arrangements are equivalent if and only if one can be rotated to obtain the other *without* flipping the paper over. **Clearly and concisely explain how you got your answer.**

(b) (5pts) Oski Bear is working on the above problem for $m = 3$ and $n = 2$ and says: "We can fix one of the B's at top of the cycle. We have 5 ways to choose a spot for the second B, and then 4 ways to choose a spot for C. Three A's go in the remaining 3 spots. So there are a total of $5 \times 4 = 20$ ways to arrange the letters." **Explain why Oski's approach does not work. In particular, is he undercounting or overcounting? If he is undercounting, give a specific instance that Oski's method didn't account for; if he is overcounting, give a specific example of an instance that is counted multiple times and briefly explain why it is counted more than once.**

2. Independence (14 pts)

(a) (8 pts) The villagers of Bararah want to select exactly *one* of their top-performing gymnasts to compete in the regional competitions next fall. The four gymnasts who have made it to the top are Audra, Bibi, Kirk, and Babak.

After intense deliberation, the villagers decide to select one of the four gymnasts uniformly at random, since the candidates are equally capable. That is, each candidate is as likely as any other candidate to be selected.

Let A denote the event, "the letter 'a' appears in the name of the selected gymnast." That is, $A = \{\text{Audra}, \text{Babak}\}$.

Let B denote the event, "the letter 'b' appears in the name of the selected gymnast." That is, $B = \{\text{Bibi}, \text{Babak}\}$.

And let K denote the event, "the letter 'k' appears in the name of the selected gymnast." That is, $K = \{\text{Kirk}, \text{Babak}\}$.

Determine whether the events A, B, and K are mutually independent. Explain your answer clearly using the definition of mutual independence.

(b) (6 pts) Consider two events A and B defined on a random experiment whose sample space is denoted by Ω. We know $\Pr(A \cap B) = \Pr(A)\Pr(B)$. Let A^c denote the complement of event A and B^c denote the complement of event B.

True or False? $\Pr(A^c \cap B^c) = (1 - \Pr(A))\Pr(B^c)$.

If you claim the assertion is true, provide a proof based on what is given. If you claim the assertion is false, provide a counterexample.

3. Babak's sneaky attempt to sneak into Cory Hall (30 pts)

On a typical day, the State of California hosts 40 million people ("Californians").

Of that population, 2,000,000 have a persistent cough.

Among those who have a persistent cough, 1,000 have muscle ache.

Another 4,000 have muscle ache that is *not* accompanied by a persistent cough.

Of those who have a persistent cough *and* muscle ache, 900 are infected with, and carry, the influenza (flu) virus.

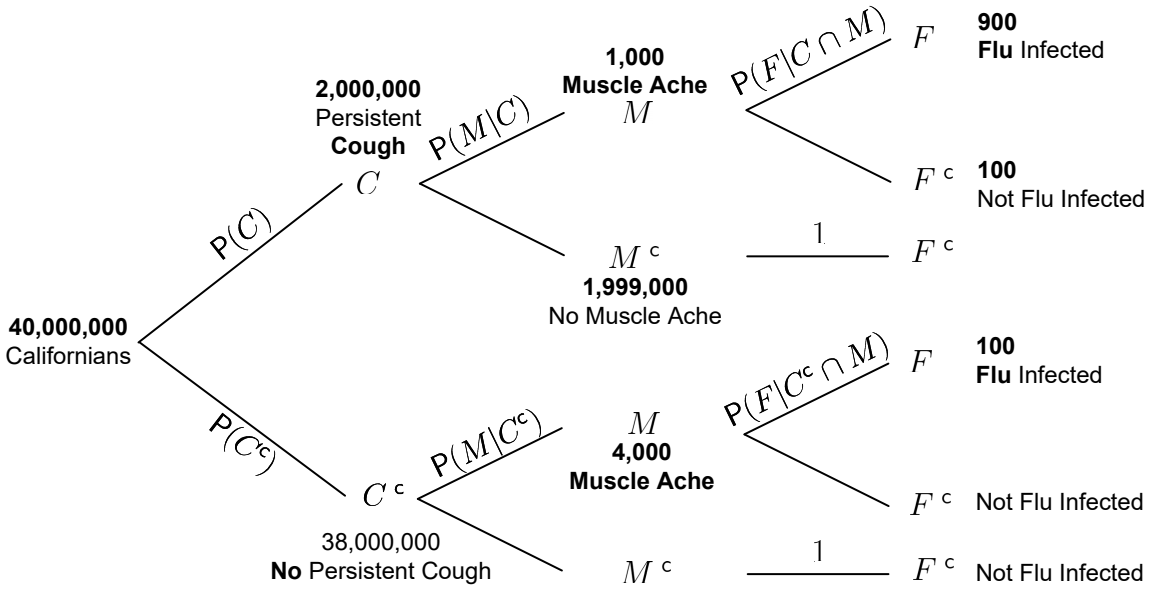
Anyone infected with the flu virus also suffers from muscle ache.

The number of Californians who are infected with, and carry, the flu virus, but who do *not* suffer from a persistent cough, is 100.

For convenience, we define the following events for a typical day in California:

- Event C denotes "A randomly-selected Californian suffers from a persistent cough."
- Event M denotes "A randomly-selected Californian suffers from muscle ache."
- Event F denotes "A randomly-selected Californian is infected with, and carries, the flu virus."

The diagram shown displays *part* of the statistics about a typical day in California. (Some relevant numbers might not be shown.)



The EECS Department at UC Berkeley has banned any person from entering EECS premises if they're likelier than not to carry the flu virus. In particular, a security guard has been assigned to each entrance of Cory Hall. The role of the guard is to assess the status of each prospective entrant and enforce department policy.

On the way to his office, Babak exhibits not only a persistent cough but also unmistakable signs of muscle ache. "You may not enter Cory Hall, Babak, because you're a likely carrier of the flu virus," the guard says.

Babak floods the guard with the California statistics and then says:

> *Fewer than the infinitesimal fraction of 1 in 2,000 Californians who suffer from a persistent cough are actually infected with, and carry, the flu virus. So, my persistent cough must not preclude me from entry into Cory Hall. Allow me in, please.*

Babak's trick succeeds. The guard — untrained in probability theory and intimidated by the numbers hurled at him — admits Babak into Cory Hall.

(a) (6 pts) **Explicitly evaluate $P(C)$, $P(M)$, and $P(C|M)$.**

(b) (6 pts) **Explicitly evaluate** $P(F)$, the probability that a randomly-selected Californian is infected with, and carries, the flu virus.

(c) (6 pts) **True or False? Events M and C are *conditionally independent*, given Event F.** Explain your answer.
Hint: Evaluate $P(M|F)$ and $P(M|F \cap C)$. This should not involve much work.

(d) (6 pts) **Explicitly evaluate** $P(F|C)$, the fraction of those afflicted with a persistent cough who are actually infected with, and carry, the flu virus. **Explain why this probability figure, an approximate upper bound for which Babak presented to the guard, should actually be irrelevant to the guard's decision.**

(e) (6 pts) **Explicitly evaluate $P(F|C \cap M)$ and explain why this is the probability figure that the guard should have taken into account while deciding.**

4. Probabilistic Cycles (20 pts)

Suppose the random undirected graph G is constructed as follows. G starts off as n distinct vertices, where n is even. Independently for each of the possible pair of vertices, we add an undirected edge between them with probability p. (i.e. We toss a biased coin with probability p of coming up heads for each distinct pair of vertices — if it comes up heads, an edge is introduced between them. If it comes up tails, there is no edge there. The different coin tosses are independent of each other.)

(a) (2 pts) Let $v_1, v_2, \ldots, v_n$ be the vertices of G.
What is the probability that the sequence $[v_1, v_2, v_3, \ldots, v_n, v_1]$ corresponds to a cycle in the graph? Clearly and concisely explain how you got your answer.
(HINT: How many distinct edges are involved in this cycle?)

(b) (10 pts) Using the union bound, **provide an upper bound on the probability that G has a Hamiltonian cycle.** Recall that a Hamiltonian cycle is a cycle which visits all the vertices of a graph exactly once.
Use Stirling's approximation ($n! \approx \sqrt{2\pi n}(\frac{n}{e})^n$) to argue about how small p has to be (as a function of n) for this bound to be nontrivial — i.e. for the bound to not be something bigger than 1. Feel free to make further approximations like $n-1 \approx n$, etc. if that helps simplify what is going on.
Clearly explain how you got your answers by showing all work.

(c) (8 pts) After the construction of G, we see that it has e edges in total. Given this information, what is the probability that the vertices v_1, v_2, and v_3 form a triangle? **In other words, what is the probability that v_1, v_2, and v_3 form a cycle within G given that there are e total edges in the graph.** Clearly and concisely explain how you got your answer.
(HINT: You know where three of the e edges have to go, how many ways are there for the other $e-3$ edges to be placed?)

5. Combinatorial Fun (16 pts)

(a) (6 pts) **Write the simplest expression potentially involving n, k, j that could go in the blank, and then prove the equality using a combinatorial proof.** You must give a combinatorial proof for full credit — algebra alone won't do it, even if it might help you figure out what is going on...

Assume $n > k > j$:

$$\text{\_\_\_\_\_} = \binom{n}{k-j}(k-j)!\binom{n-k+j}{j}j!$$

(b) (10 pts) Let $n \geq 2k$. **Write down the combinatorial identity that is proved by the following story.** *Some parts of the story are redacted and left for you to figure out.* Consequently, **you must explain your answer for full credit.**

There are n humans each with their dog. We want to select a party of $2k$ creatures from the n humans and n dogs to go on a hike, with the constraint that a dog cannot go without their owner, but an owner can go without their dog.

We can count the number of distinct valid parties in two ways: picking dogs first or picking people first.

If we select j of the dogs, then by the constraint, their owners have to be selected, so we only need to select $2k-2j$ more people from the remaining humans. The number of dogs selected ranges from 0 to k since at most half of the party can be dogs.

Alternatively, we could start with people. We select i people, then by the constraint, the dogs selected need to be owned by one of the i people. So there are *[redacted]* ways to pick the $2k-i$ dogs that are going to come along with the people we already selected. There clearly must be at least as many people as dogs selected, so we need to sum over i ranging from *[redacted]* to *[redacted]*.

6. Hotel Rooms and Beyond: Error-and-erasure correction leveraging the CRT (58 pts)

Note: for this entire problem, you can use properties of the Chinese Remainder Theorem that we discussed in lecture, notes, homework, and discussion without proof and without having to specify all the details.

Alice is staying at a hotel and she wants to share her room number with Bob by leaving a sequence of notes in a list of pre-arranged locations.

i) There are only 100 possible hotel rooms, labeled 0 to 99.

ii) Alice takes her room number p and computes the remainders $y_i = p \bmod p_i$. The specific p_i that she uses are $p_1 = 3, p_2 = 5, p_3 = 7, p_4 = 11$, and $p_5 = 13$.

iii) She writes y_i on the i-th note and places the note in location i. (Both Alice and Bob know which location corresponds to which numbers i and p_i.)

This sequence of notes can be viewed as a codeword $\vec{y}(p)$. For example, if Alice is in room 51, she sends the codeword $\vec{y}(51) = (0,1,2,7,12)$ since $51 \bmod 3 = 0$, $51 \bmod 5 = 1$, $51 \bmod 7 = 2$, $51 \bmod 11 = 7$, $51 \bmod 13 = 12$.

(a) (3 pts) Unfortunately, there is a chance that some of Alice's notes get blown away by the wind. Those missing notes will be treated as erasures and denoted by X.

It turns out that Alice is in room 51 as above, so she sends the codeword $(0,1,2,7,12)$. Bob gets the received symbols $(0,1,2,X,X)$. **Explain how Bob can leverage the CRT to identify Alice's room number.**

(b) (6 pts) Generalizing the previous part, **prove that in the case of up to two erasures anywhere, your scheme will always correctly identify the room number.** For this part, feel free to just use the properties of the CRT along with the following facts.

- $p_1 < p_2 < \cdots < p_5$.
- The number of rooms is less than $\prod_{i=1}^{3} p_i$.

(c) (5 pts) Would your scheme *always* work if there were 3 erasures and only 2 received notes? **If so, argue that it always works. If not, give an explicit counterexample where your scheme doesn't work.** (i.e. construct an example where there is an ambiguity about which room is Alice's room.)

(d) (8 pts) Now assume that instead of the wind blowing away a note, somebody malicious finds a note before Bob and replaces it with a corrupted note. Now the received notes r_i could differ from the true $y_i = p \bmod p_i$ in at most one location.

For such a small message-space, Bob decides to simply try all possible messages $p \in \{0,1,\ldots,99\}$, and see which one's $\vec{y}(p)$ differs from the received notes $\vec{r}$ in at most one location. If there is a unique such p, Bob just picks that one.

Could there be two different legitimate room numbers whose encodings both differ from the received notes in only one note? If so, give an example. If not, prove it.

(HINT: If someone told you which note was corrupt, could you uniquely figure out the room? What if they told you that two specific notes were possibly corrupt? Could you do it then? How does that help you think about the above situation?)

(e) (10 pts) Let's think about a more general case. Instead of 5 notes, assume that there are $n+2k$ of them. We have $n+2k$ large (think 100 binary digits or thereabouts) prime numbers $p_1 < p_2 < \cdots < p_n < p_{n+1} < \cdots < p_{n+2k}$. The message is the number p which is a natural number in the range $0 \leq p < \prod_{i=1}^{n} p_i$. Define the number $N = \prod_{i=1}^{n} p_i$ for this upper-bound quantity.

Suppose you receive a list $\vec{r}$ of received notes $r_1, r_2, \ldots, r_{n+2k}$.

At first, your friend is more relaxed than you about the possibility of corruptions. Your friend has a simple scheme in which she assumes that all $n+2k$ of the notes were not corrupted and just uses the Chinese Remainder Theorem to reconstruct a number r that satisfies all the $r_i = r \bmod p_i$ equations for all $i \in [1, n+2k]$. By the CRT construction, this r is always less than $M = \prod_{i=1}^{n+2k} p_i$. Your friend says that if there were no corruptions, then actually $r < N$ and she wouldn't have to think about anything and could just say that $p = r$.

Now your friend has received special information that the note in location ℓ might have been corrupted so that $r_\ell \neq y_\ell = p \bmod p_\ell$. Indeed your friend had already noticed that the r she naively reconstructed was in fact bigger than N.

She suggests the following scheme instead: she will take $e = p_\ell$ and multiply each of the r_i by e, taking the result mod p_i. In other words, she will compute $q_i = er_i \bmod p_i$ for all $i \in [1, n+2k]$. This way, $q_\ell = 0$ since she doesn't trust that corrupted note in location ℓ. Then, she will use the CRT to compute the unique non-negative integer q that satisfies all those equations and is less than M. With that q in hand, she will just divide q by e to get the original p.

Prove that your friend's scheme works if there is a corruption in the ℓ-th note. (You can assume that $k > 1$.)

(f) (6 pts) Continuing the setup of the previous part. Now suppose that your friend is told that there are corrupted notes in locations ℓ_1, ℓ_2, ℓ_3.

Show how you can adapt your friend's scheme to deal with this case. (You can assume that $k > 3$.) Explain briefly why this will work.

(HINT: You just have to change e to make everything work.)

(g) (12 pts) Now, nobody is going to tell you exactly which notes are corrupted. However, suppose that someone gives you a function that can magically find all integer solutions (e, q, z) to constrained equations of the form $ae + bq + gz = d$ where you get to pick interval constraints for the e, q, z of the form $e \in [0, e_u]$ and $q \in [0, q_u]$ while $z \in (-\infty, +\infty)$ is unconstrained. This function takes in integer arguments (a, b, g, d, e_u, q_u).

Describe how you could use this function exactly once to help you decode the underlying message p from a list of received notes $r_1, r_2, \ldots, r_{n+2k}$ where at most c of the notes have been corrupted away from their original values of $y_i = p \bmod p_i$. You know that $c < k$ in advance and if it helps, feel free to assume that both n and k are moderate sized numbers like 100 or 10, and think about c as being something like 3.

What arguments would you use for the integers a, b, g, d, e_u, q_u? How would you use the resulting integers e, q, z to find your p? (It is fine if you have to invoke calculations for these arguments, etc. Obviously, the received notes are going to be involved in some fashion.)

HINT: Think about what q and e mean in the previous part. What do you know about how big those q and e could be? Think about how you go back and forth between writing a system of CRT equations and writing something in mod math. Think about how you write an equality of the form $f \equiv g \pmod{M}$ in the form of an equality involving an unknown integer.

The numbers N and M defined in earlier parts might be useful as could the number $H = \prod_{i=n+2k-c+1}^{n+2k} p_i$.

Also, you might find it helpful to glance at the next part before starting this one.

(h) (8 pts) **Describe the relationship of what is happening in this problem with the Berlekamp-Welch approach to dealing with up to k malicious errors in the polynomial-based error-correcting codes that you learned about in the course.**

- **What are analogous to the $P(x), Q(x), E(x)$ polynomials?**
- **What are analogous to evaluations of polynomials?**
- **What is analogous to the maximum degrees of $P(x), Q(x), E(x)$?**

HINT: This entire part can be viewed as a hint for the previous part.

