445 Cheatsheet

Below is a quick refresher on some math tools and problem-solving techniques from 240 (or other prereqs) that we'll assume knowledge of for the PSets. Roadmap:

- Section 1 is a refresher on probability. I hope that most of it feels like a review. Section 1.9 may feel conceptually more advanced than the rest. We'll use a lot of probability in this course, so you may want to have this handy as a reference if you feel stuck on a problem because of probability.

- Section 2 is a refresher on continuous optimization. You don't need much for this course, but you may find it as a helpful quick reference.

- Section 3 contains quick tips on how to approach creative problem-solving. **I strongly recommend reading this section, especially Section 3.1**, which describes how I think of partial progress/credit.

- Section 4 contains my own (brief) thoughts on how to write solid proofs. This may be different than thoughts you've heard in previous courses. **I strongly recommend reading this section**, and in particular the examples demonstrating common pitfalls, and guidelines for breaking down complex proofs.

- Section 5 plays around with LLMs a little bit to give examples of how it might be (and not be) useful in this course. I so far found LLMs to be neither particularly useful, nor particularly bad for minor tasks (I asked it to explain a problem to me, polish my answer, and turn my text into LaTeX).[1] I found it to be very bad at solving a problem from scratch, and annotated why. This section will mostly just be fun if you find LLMs interesting (you may also find the annotation useful to get practice dissecting a bad proof).[2]

All of the notes here are intended for an audience who has seen these topics before, but may enjoy a refresher before using these tools to solve abstract problems. If you have not seen a topic before and want to learn it, I suggest revisiting the (significantly more thorough) course materials from 240.

1   Basic Probability

1.1   Discrete random variables

A *random variable* is a variable whose value is uncertain (i.e. the roll of a die). If X is a random variable that always takes non-negative, integer values, (we'll refer to this as a discrete random variable) then we can write the expected value of X as:

---

[1]**Note:** If you choose to use LLMs for any of these tasks, you are ultimately responsible for what you submit, and your own understanding of the problem. We can't promise that LLMs will do a good job.

[2]LLMs have gotten better since I wrote this section, and the best ones can probably solve the simple problems asked here now. This performance is more representative of what would happen if an LLM were asked to solve an actual PSet problem from scratch.

Definition of expected value, form 1: $\mathbb{E}[X] = \sum_{i=0}^{\infty} \Pr[X = i] \cdot i.$

Probably the above definition is familiar to most of you already. Another way to compute the expected value (which sometimes results in simpler calculations) is:

Definition of expected value, form 2: $\mathbb{E}[X] = \sum_{i=0}^{\infty} \Pr[X > i].$

Let's quickly see why the two definitions are equivalent:

$$\sum_{i=0}^{\infty} \Pr[X > i] = \sum_{i=0}^{\infty} \sum_{j>i} \Pr[X = j].$$

$$= \sum_{j=0}^{\infty} \sum_{i<j} \Pr[X = j]$$

$$= \sum_{j=0}^{\infty} j \cdot \Pr[X = j].$$

We obtain the second equality just by flipping the order of sums: the term $\Pr[X = j]$ is summed once for every $i < j$. The third equality is obtained by just observing that there are exactly j non-negative integers less than j.

Why bother knowing two definitions? Most students find it conceptually easier to remember form 1. I also find this to be the case. However, many problems you'll encounter in this course will be conceptually easier to solve using form 2. See the later sections for further discussion and a few examples.

1.2 Continuous random variables

Now let's consider a continuous, non-negative random variable with probability density function (PDF) $f(\cdot)$ and cumulative distribution function (CDF) $F(\cdot)$. What do all these words mean? You should imagine the following mapping:

- **Continuous** just means that the random variable might take any non-negative value. For instance, rather than the roll of a die, a random variable might be the number of seconds you spend reading this sentence.

- The **PDF** is just a formal way of discussing the probability that $X = x$. Because the random variable is continuous, the probability that $X = x$ is actually zero for all x (what is the probability that you spend exactly 3.4284203 seconds reading this sentence)? So we think of dx as being infinitesimally small (the same dx from your calculus classes), and think of $\Pr[X = x]$ as $f(x)dx$. Or, to be even more formal, this captures $f(x) = \lim_{\varepsilon \to 0} \Pr[X \in [x, x + \varepsilon]]/\varepsilon$.

- The **CDF** of a random variable is simpler to define, and just denotes $F(x) = \Pr[X \leq x]$.[3] Note that we therefore have $F(x) = \int_0^x f(y)dy$. Think of this as "summing" (integrating) over all $y \leq x$ the probability that $X = y$ ($f(y)dy$). Therefore $F'(x) = f(x)$ (by fundamental theorem of calculus).

So how do we take the expectation of a continuous random variable? We just need to map the definitions above into the new language.

Definition of expected value, continuous random variables, form 1: $\mathbb{E}[X] = \displaystyle\int_0^\infty xf(x)dx.$

You should parse exactly the same way as form 1 for discrete random variables, except we've replaced the sum with an integral, and $\Pr[X = i]$ is now "$f(x)dx \approx \Pr[X = x]$."[4] The equivalent definition for form 2 is also often easier to use in calculations:

Definition of expected value, continuous random variables, form 2: $\mathbb{E}[X] = \displaystyle\int_0^\infty \big(1-F(x)\big)dx.$

It is also quick to see that the two forms are equivalent, using the fact that $F'(x) = f(x)$ (it's not important that you're able to follow the calculations below — they're included in case you're curious and it helps to see).[5]

$$\int_0^\infty \big(1 - F(x)\big)dx = \int_0^\infty \left(\int_x^\infty f(y)dy\right)dx$$
$$= \int_0^\infty \left(\int_0^y f(y)dx\right)dy$$
$$= \int_0^\infty yf(y)dy$$

If $F(x) = \Pr[X \leq x]$, then $1 - F(x) = \Pr[X > x]$, so this is the same as form 2 for discrete random variables, except we've replaced the sum with an integral. For form 2, it is *crucial* that the integral start below at 0, even when the random variable only takes values (say) > 1. We'll see this in examples below.

1.2.1   Brief Discussion

Note also that the definition of expected value for continuous random variables, form 2, is actually well-defined for any non-negative random variable, whether it is discrete or continuous (or a mix of both). This is not always the mathematically cleanest way to compute the expectation of a complex random variable, but it is the conceptually most-straight-forward. **A safe approach for any random variable X you see in this class is to explicitly figure out, for any x, what is the probability that $X > x$? Then, write this integral and compute it.**

---

[3]CDFs, in contrast to PDFs, are well-defined for discrete random variables too. A nondecreasing $F : \mathbb{R} \to [0, 1]$ is a CDF if it is right-continuous and has left limits at every $x \in \mathbb{R}$, $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$. In particular, $\Pr[X \leq x]$ is a piecewise constant function (which satisfies these conditions) if X is discrete.

[4]Note that we used i for the values of a discrete random variable and x for those of a continuous one, but this isn't necessary. You can replace i with x in the discrete case if that makes it easier to map between the two.

[5]The first line follows from the fact that $\int_x^\infty f(y)dy = \lim_{y\to\infty} F(y) - F(x) = 1 - F(x)$. The second line follows from Fubini's theorem.

If you're more comfortable with probability, you may often be able to find a more clever approach which requires fewer calculations, but the above approach will always succeed.

Finally, also note that for continuous random variables, $\Pr[X = x] = 0$ for all x, so $\Pr[X \leq x] = \Pr[X < x]$. This is not true for discrete random variables, but observe that the evaluation of the integral is indifferent to whether we use $F(x) = \Pr[X \leq x]$ or $F(x) = \Pr[X < x]$.[6]

1.3   Three Examples

Example 1. *Let X be a random variable that is 4 with probability $1/2$, and 5 with probability $1/2$ (this is the uniform distribution over $\{4, 5\}$).*

Fact 1. *The expected value of X is 4.5 ($\mathbb{E}[X] = 4.5$).*

We provide two proofs of this, one using each form.

Proof: The expected value as computed by form 1 is:

$$\sum_{i=0}^{\infty} \Pr[X = i] \cdot i = 4 \cdot 1/2 + 5 \cdot 1/2 = 4.5. \qquad \blacksquare$$

Proof: The expected value as computed by form 2 is:

$$\sum_{i=0}^{\infty} \Pr[X > i] = \sum_{i=0}^{3} 1 + \sum_{i=4}^{4} 1/2 = 4.5. \qquad \blacksquare$$

Example 2. *Let X be a random variable drawn from the uniform distribution on the interval $[4, 5]$. That is, the PDF of X satisfies $f(x) = 1$ for all $x \in [4, 5]$, and $f(x) = 0$ otherwise. Observe that the CDF of X satisfies $F(x) = 0$ for all $x \in [0, 4]$, $F(x) = x - 4$ for all $x \in [4, 5]$, and $F(x) = 1$ for all $x \geq 5$.*

Fact 2. $\mathbb{E}[X] = 4.5$.

We again provide two proofs, one using each form.

Proof: We can compute the expected value by form 1 as:

$$\int_{0}^{\infty} x f(x) dx = \int_{4}^{5} x \, dx = x^2/2 |_{4}^{5} = 25/2 - 8 = 4.5. \qquad \blacksquare$$

Proof: We can also compute it using form 2 as:

$$\int_{0}^{\infty} \left(1 - F(x)\right) dx = \int_{0}^{4} 1 \, dx + \int_{4}^{5} (5 - x) dx + \int_{5}^{\infty} 0 \, dx = 4 + \left(5x - \frac{x^2}{2}\right)\Big|_{4}^{5} + 0 = 4.5. \quad \blacksquare$$

Note that it is *crucial* that we started the integral at 0 and not 4 for form 2, otherwise we would have incorrectly computed the expectation as $.5$ instead of 4.5. This isn't crucial for form 1, since all the terms in $[0, 4]$ drop out anyway as $f(x) = 0$.

Example 3. *Let X be defined as follows. A fair six-sided die is rolled. Let Y denote the roll, then $X = Y^2$ (that is, if the roll is 1, then $X = 1$. If the roll is 2, then $X = 4$, and so on).*

---

[6]This is because the two definitions differ on a set of measure zero, so the integrals must be identical. But don't worry about this if you're not familiar with this language.

Fact 3. $\mathbb{E}[X] = 91/6$.

Proof: This is an example of a random variable where it happens to be significantly cleaner to compute the expectation using form 1. Here we can write:

$$\mathbb{E}[X] = \sum_{i=1}^{\infty} i \cdot \Pr[X = i]$$

$$= \sum_{i=1}^{\infty} i \cdot \Pr[Y = \sqrt{i}] \quad \text{(because } X = Y^2\text{)}$$

$$= \sum_{j=1}^{6} j^2 \cdot \Pr[Y = j] \quad \text{(substituting } j = \sqrt{i}, \text{ and observing that } \Pr[Y = j] = 0 \text{ when } j \notin [6]\text{)}$$

$$= \sum_{j=1}^{6} \frac{1}{6} \cdot j^2 = \frac{91}{6}. \qquad \blacksquare$$

1.4   Linearity of Expectation

Linearity of expectation refers to the following simple, but surprisingly useful fact. Let X_1 and X_2 be two random variables. Then $\mathbb{E}[X_1 + X_2] = \mathbb{E}[X_1] + \mathbb{E}[X_2]$. The proof is immediate from the definitions above. We include the proof for the discrete case:

$$\mathbb{E}[X_1 + X_2] = \sum_{i=0}^{\infty} \Pr[X_1 + X_2 = i] \cdot i$$

$$= \sum_{i=0}^{\infty} \sum_{j=0}^{i} \Pr[X_1 = j] \cdot \Pr[X_2 = i - j] \cdot i$$

$$= \sum_{j=0}^{\infty} \sum_{i=j}^{\infty} \Pr[X_1 = j] \cdot \Pr[X_2 = i - j] \cdot i$$

$$= \sum_{j=0}^{\infty} \Pr[X_1 = j] \cdot \sum_{\ell=0}^{\infty} \Pr[X_2 = \ell] \cdot (\ell + j) \quad \text{(changing variables with } \ell = i - j\text{)}$$

$$= \sum_{j=0}^{\infty} \Pr[X_1 = j] \cdot \left(j + \sum_{\ell=0}^{\infty} \Pr[X_2 = \ell] \cdot \ell \right)$$

$$= \sum_{j=0}^{\infty} \Pr[X_1 = j] \cdot (j + \mathbb{E}[X_2])$$

$$= \mathbb{E}[X_1] + \mathbb{E}[X_2]. \qquad \text{(because } \sum_{j=0}^{\infty} \Pr[X_1 = j] = 1\text{)}$$

1.5   Bayes' Rule

Let's first recap the definition of conditional probability: the probability of an event A conditioned on another event B, denoted $\Pr[A|B]$, is equal to the probability of A and B divided by the probability of B (that is, $\Pr[A \wedge B] / \Pr[B]$). Think of this as the probability that A occurs, given that B

has occured. For a concrete example, consider the probability that a fair six-sided dice lands two (A), conditioned on it landing even (B). Then $\Pr[A \wedge B]$ is the probability that the die lands two *and* that it is even (which is just the probability that it is two), so $1/6$. $\Pr[B]$ is just the probability that the die is even, which is $1/2$, so the ratio is $1/3$. The probability that the roll is prime (A), conditioned on being even (B) can be computed similarly: the probability that the roll is prime *and* even is $1/6$ (the only even prime is two), and the probability that the roll is even is $1/2$. So again the ratio is $1/3$, and the probability of rolling prime conditioned on rolling even is $1/3$.

Sometimes, explicitly computing $\Pr[A \wedge B]$ might be challenging, but computing $\Pr[A|B]$ is not so bad. Bayes' rule simply manipulates the above equalities to write:

$$\Pr[A|B] = \frac{\Pr[A \wedge B]}{\Pr[B]} = \frac{\Pr[B|A] \cdot \Pr[A]}{\Pr[B]}.$$

Example 4. *Consider a coin whose probability of outputting heads is p, but p is a random variable. Specifically, p is either equal to $1/4$ with probability $1/2$, and equal to $3/4$ with probability $1/2$. You flip the coin once and it lands heads. What is the probability that the coin's bias is $3/4$?*

Fact 4. *The probability that the coin's bias is $3/4$ is $3/4$.*

Proof: We prove this using Bayes' rule. Let A denote the event that $p = 3/4$, and let B denote the event that the coin lands hands after one flip. Then the problem is asking us to compute $\Pr[A|B]$. This is conceptually quite tricky to reason about! But fortunately, Bayes' rule gives us a formula using three terms that are much simpler to think about.

- $\Pr[A|B]$ is the probability that the coin lands heads, conditioned on the bias being $3/4$. This is easy to compute, and is just $3/4$, immediately from the definition of bias.

- $\Pr[A]$ is the probability that the bias is $3/4$, which is just $1/2$, immediately by the problem setup.

- $\Pr[B]$ is the probability that the coin lands heads (without having seen any flips), which is also just $1/2$. This is because the coin's probability of landing on heads is always p, and $\mathbb{E}[p] = 1/2$.

Now, we can just apply Bayes' rule and write:

$$\Pr[A|B] = \frac{\Pr[A|B] \cdot \Pr[B]}{\Pr[A]} = \frac{(3/4) \cdot (1/2)}{(1/2)} = 3/4. \qquad \blacksquare$$

1.6   Conditional Expectation

The previous section covers conditional probability, what is the probability of A conditioned on B. We may also want to discuss conditional expectations: the expectation of a random variable X conditioned on an event A. Formally, think of this as first, "what is the probability that $X = x$, conditioned on A?" Then define a new random variable $X|A$, which is equal to x with this probability, and take its expectation.

This is related to the following notion: let $\mathbb{I}(A)$ denote a random variable which is one when event A occurs, and 0 otherwise. **Then $\mathbb{E}[X|A] = \mathbb{E}[X \cdot \mathbb{I}(A)]/\Pr[A]$. Sometimes, it may be mathematically simpler to reason about $\mathbb{E}[X \cdot \mathbb{I}(A)]$ then $\Pr[A]$ separately to compute $\mathbb{E}[X|A]$.**

1.6.1 An Example

Example 5. *Let X be a random variable that is drawn from a distribution with CDF $1 - e^{-x}$.[7] Observe that the PDF of this distribution is exactly e^{-x} (by taking the derivative of $1 - e^{-x}$). Compute $\mathbb{E}[X | X \in [5, 10]]$.*

Fact 5. $\mathbb{E}[X | X \in [5, 10]] = 6 - \frac{5}{e^5 - 1}$.

Proof: We propose doing this computation by first computing $\mathbb{E}[X \cdot \mathbb{I}(X \in [5, 10])]$, and then dividing by $\Pr[X \in [5, 10]]$.

Following the bold-faced suggestion in Section 1.2.1, we have a random variable whose expectation we wish to compute. So, let's do so by trying to directly compute $\Pr[X \cdot \mathbb{I}(X \in [5, 10]) > x]$, for all x.

Observation 1. *For all $x < 5$, $\Pr[X \cdot \mathbb{I}(X \in [5, 10]) > x] = e^{-5} - e^{-10}$.*

Proof: For $x < 5$, $X \cdot \mathbb{I}(X \in [5, 10]) > x$ whenever $X \in [5, 10]$ (because this guarantees that $X > x$, and also that the multiplier is one and not zero). Therefore,

$$\Pr[X \cdot \mathbb{I}(X \in [5, 10]) > x] = \Pr[X \in [5, 10]].$$

Observe also that $\Pr[X \in [5, 10]] = \Pr[X \geq 5] - \Pr[X > 10]$, so

$$\Pr[X \in [5, 10]] = 1 - F(5) - (1 - F(10)) = F(10) - F(5) = e^{-5} - e^{-10}. \qquad \blacksquare$$

Observation 2. *For all $x > 10$, $\Pr[X \cdot \mathbb{I}(X \in [5, 10]) > x] = 0$.*

Proof: Whenever $X > 10$, the multiplier $\mathbb{I}(X \in [5, 10]) = 0$, so the whole term can never be > 10 (because whenever $X > 10$, it is zeroed out). $\qquad \blacksquare$

Observation 3. *For $x \in [5, 10]$, $\Pr[X \cdot \mathbb{I}(X \in [5, 10]) > x] = e^{-x} - e^{-10}$.*

Proof: To see this, observe that when $x \in [5, 10]$, $X \cdot \mathbb{I}(X \in [5, 10]) > x$ if and only if $X \in [x, 10]$. This is because we first need $X > x$, but we also need $\mathbb{I}(X \in [5, 10]) = 1$. Therefore, we get that

$$\Pr\left[X \cdot \mathbb{I}(X \in [5, 10]) > x\right] = \Pr[X \in [x, 10]]$$
$$= \Pr[X \geq x] - \Pr[X \geq 10]$$
$$= e^{-x} - e^{-10}. \qquad \blacksquare$$

Now that we know $\Pr[X \cdot \mathbb{I}(X \in [5, 10]) > x]$ for all x, we can compute its expectation:

$$\mathbb{E}\left[X \cdot \mathbb{I}(X \in [5, 10])\right] = \int_0^\infty \Pr[X \cdot \mathbb{I}(X \in [5, 10]) > x] \, dx$$
$$= \int_0^5 \left(e^{-5} - e^{-10}\right) dx + \int_5^{10} \left(e^{-x} - e^{-10}\right) dx + \int_{10}^\infty 0 \, dx$$
$$= 5 \cdot \left(e^{-5} - e^{-10}\right) + \left(-e^{-x} - xe^{-10}\right)\big|_5^{10}$$
$$= 5 \cdot \left(e^{-5} - e^{-10}\right) + e^{-5} - e^{-10} - 5e^{-10}$$
$$= 6e^{-5} - 11e^{-10}.$$

---

[7]This is called an exponential distribution.

Finally,

$$
\begin{aligned}
\mathbb{E}\big[X | X \in [5, 10]\big] &= \frac{\mathbb{E}\big[X \cdot \mathbb{I}\big(X \in [5, 10]\big)\big]}{\Pr\big[X \in [5, 10]\big]} \\
&= \frac{6e^{-5} - 11e^{-10}}{e^{-5} - e^{-10}} \\
&= 6 - \frac{5}{e^5 - 1}.
\end{aligned}
$$ ∎

1.7 Coupling Arguments

Coupling arguments are typically useful to relate two probabilities. We'll first give a definition of a coupling argument, and then give several examples for why these arguments can be useful. In general, coupling arguments can be useful for saving significant energy compared to raw calculations (but you can typically replace any coupling argument with a brute-force calculation, if you find that approach preferable and are able to execute the calculations cleanly).

Definition 4 (Coupling). *Let D_1, D_2 be distributions. A coupling of D_1 and D_2 is a process to jointly draw two random variables X, Y such that:*

- *X is distributed according to D_1.*

- *Y is distributed according to D_2.*

The definition may feel abstract at first, but the following examples will help clarify it, and show where it can be useful.

Example 1: Biased Coins.[8] Consider two biased coins. Coin A lands heads with probability p, and coin B lands heads with probability $q > p$. Prove that for all $k, n \in \mathbb{N}$, the probability that n independent flips from coin B yield $\geq k$ heads is at least the probability that n independent flips from coin A yield $\geq k$ heads.

At first glance, the claim probably seems obvious: each flip of B is more likely to yield heads, so n flips of B should be more likely to yield more heads. But this is not a rigorous proof. One option for a rigorous proof is to explicitly compute the probability of seeing $\geq k$ heads from each coin, and prove that the term is larger for B (possibly by taking a derivative). There is nothing wrong with this approach, but it will wind up with a lot of calculations for what feels like a very obvious claim. A coupling argument, on the other hand, will more closely match the obvious intuition. Let's first try to set up the syntax for a coupling argument. Recall that a coupling argument starts with two distributions D_1 and D_2 that we want to draw jointly from a single process. So let's first be clear about what D_1 and D_2 might be for this claim.

- Let D_1 denote the following distribution: flip coin A n times, and output the list of results $\langle a_1, \ldots, a_n \rangle$, where a_i is the result of coin i. That is, D_1 draws a heads/tails vector $\langle a_1, \ldots, a_n \rangle$, where each a_i is independently heads with probability p, and tails with probability $1 - p$.

- Let D_2 denote the following distribution: flip coin B n times, and output the list of results $\langle b_1, \ldots, b_n \rangle$, where b_i is the result of coin i. That is, D_2 draws a heads/tails vector $\langle b_1, \ldots, b_n \rangle$, where each b_i is independently heads with probability q, and tails with probability $1 - q$.

---

[8]I found this example on the Wikipedia page for Coupling. The page is helpful, but perhaps notation-heavy.

- Observe that the claim being made is exactly that:

$$\Pr[\langle b_1, \ldots, b_n \rangle \text{ has} \geq k \text{ heads}] \geq \Pr[\langle a_1, \ldots, a_n \rangle \text{ has} \geq k \text{ heads}].$$

Now, we need to define a process to jointly draw random variables X and Y, so that X is distributed according to D_1, and Y is distributed according to D_2. But, we want to do this in a way that makes it easy to reason about the desired probabilities.

- Flip coin A n times, so that each is heads with probability p independently, and call the results $a_1, \ldots, a_n$.

- Use the labels $b_1, \ldots, b_n$ to denote the results of the n flips of coin B.

- For each i:

  - If a_i is heads, set b_i to heads.
  - If a_i is tails, set b_i to heads with probability $(q-p)/(1-p)$, and to tails with probability $(1-q)/(1-p)$.

Here is the logical flow we hope to pursue (after presenting the outline, we'll prove each step).

- The above procedure correctly has $\langle a_1, \ldots, a_n \rangle$ distributed according to D_1 (that is, each flip of the A coin is heads with probability exactly p, independently).

- The above procedure correctly has $\langle b_1, \ldots, b_n \rangle$ distributed according to D_2 (that is, each flip of the B coin is heads with probability exatcly q, independently).

- When we flip the coins in this manner, there are $\geq k$ heads among the flips of B *whenever there are $\geq k$ heads among the A flips* (maybe more often, too).

- Therefore, there are only three types of events: those where the A and B flips both have $\geq k$ heads (say that the probability that this happens is p_{AB}), those where neither set has $\geq k$ heads (say that the probability that this happens is p_0), and those where only the B flips have $\geq k$ heads (say that the probability that this happens is p_B).

- Then the probability that the A flips have $\geq k$ heads is p_{AB}, and the probability that the B flips have $\geq k$ heads is $p_{AB} + p_B$. Clearly, $p_B + p_{AB} \geq p_{AB}$, which proves the desired claim.

The important part of this process is the final bullet: we were able to use a trivial calculation to prove that $p_{AB} + p_B \geq p_{AB}$, *without ever computing p_{AB}, nor p_B*. This is the power of coupling arguments: it allows you to relate the probability of two events (perhaps showing that they occur with the same probability, or that one is more likely), *without ever computing the probability that of either*. Coupling arguments are conceptually harder to implement than direct calculations, but will *significantly* save on necessary calculations.

In the logical flow above, observe that it is crucial that we have a clear relation between our coupling experiment and the two probabilities that we want to analyze. It's therefore crucial in any proof that the first two steps are explicitly stated/proved. Now we proceed to prove each step in the outline.

First, it's indeed easy to see that this is a valid way of flipping coin A n times, because we explicitly flip coin A n times and each is heads with probability p (that is, $\langle a_1, \ldots, a_n \rangle$ is correctly distributed according to D_1. Next, observe that coin b_i is heads whenever a_i is heads (which occurs

with probability p), or $(q-p)/(1-p)$ of the time when a_i is tails (which occurs with probability $1-p$). Therefore, coin b_i is heads with probability $p + (1-p) \cdot (q-p)/(1-p) = p + q - p = q$, and this is done independently of all other coin Bs.Therefore, we have indeed flipped n coins independently that are heads with probability q, and this is a valid way to flip coin B n times (that is, $\langle b_1, \ldots, b_n \rangle$ is correctly distributed according to D_2).

Now, observe that there are $\geq k$ heads among $a_1, \ldots, a_n$ *only if* there are $\geq k$ heads among $b_1, \ldots, b_n$. This is simply because whenever coin a_i is heads, so is b_i. Therefore, the probability of seeing $\geq k$ heads among $b_1, \ldots, b_n$ is at least as large as the probability of seeing $\geq k$ heads among $a_1, \ldots, a_n$. By the work in the previous paragraph, the probability of seeing $\geq k$ heads among $b_1, \ldots, b_n$ is exactly the probability of seeing $\geq k$ heads among n flips of B, and the probability of seeing $\geq k$ heads among $a_1, \ldots, a_n$ is exactly the probability of seeing $\geq k$ heads among n flips of A. This completes the proof.

Again, note in the previous paragraph that all of our calculations concerned the coins $a_1, \ldots, a_n$, and $b_1, \ldots, b_n$. But we were asked to reason about n flips of the coin A, and n flips of the coin B. So we needed to explicitly state a connection between the two sets of coins in order to wrap up the final step.

Example 2: Random Walks. Consider an ant that starts at time $t = 0$ at location 0. Every timestep, they either move up $(+1)$ with probability $1/2$ or down (-1) with probability $1/2$, and they stop after n steps, where n is even. Let H_k denote the probability that the ant reaches $+k$ at *some point* during the walk, and let L_k denote the probability that the ant *finishes* their walk at $+k$ or higher. Prove that $H_k = 2L_k$ for all odd k.

For this problem, even the "raw calculations" approach is quite tricky (because it's cumbersome to write a closed form for H_k). However, there is an elegant coupling argument that makes the proof clean. Consider the following process for producing two random walks. To be explicit, we will let D_1 and D_2 be the same distribution over steps $R_0, \ldots, R_n$, where $R_0 = 0$, and each R_i is equal to $R_{i-1} + 1$ or $R_{i-1} - 1$ with probability $1/2$, independently. Now consider the following process to jointly draw two random walks:

- Let $R_0 = S_0 = 0$. For all i from 1 to n, let R_i be equal to $R_{i-1} + 1$ with probability $1/2$, and equal to $R_{i-1} - 1$ with probability $1/2$ (independently).

- If there exists an i such that $R_i = k$, let i^* be the largest such i. If there is no such i, let $S_i := R_i$ for all i. Observe that because n is even and k is odd, that we must have $k < n$.

- Otherwise, let $S_i := R_i$ for all $i \leq i^*$, and $S_i := k - (R_i - k)$ for all $i > i^*$. That is, "reflect" R_i over the horizontal line at k to get S_i when $i > i^*$.

Again, we need to show two things. First, we need to show that both R and S are correctly sampled random walks (that is, R is distributed according to D_1, and S is distributed according to D_2). Again, this will be trivial for R, and not-bad-but-not-trivial for S. Then, we need to use these two walks to reason about H_k and L_k.

First, it's clear that R is a correctly-drawn random walk, because we explicitly define it as such. It's also true that S is a correctly-drawn random walk. To see this, observe importantly that *reflecting from R to S after i^* does not change i^**. Indeed, after i^*, R is either completely above k, or completely below k (by definition). Reflecting R to S maintains $S_{i^*} = k$, and also maintains that S is completely below k or completely above k (the opposite of R). This means that the mapping from R to S is its own inverse (applying it twice returns the original R). In particular, this means that every possible random walk is equally likely to be drawn from R (if it is drawn directly) as it

is drawn from S (if its reflection is drawn from R). Therefore, S is also a correctly-drawn random walk.

Finally, observe that either: *neither R nor S ever reach k, or both R and S reach k.* Additionally, when neither R nor S reach k, neither R nor S finish above k. On the other hand, when both R and S reach k *exactly one of them finishes above k.* Because both R and S are correctly-drawn random walks, this means that H_k is exactly half the probability that R reaches k, plus half the probability that S reaches k. Similarly, L_k is exactly half the probability that R finishes above k, plus half the probability that S finishes above k. By the reasoning at the beginning of this paragraph, H_k is now exactly twice L_k.

Concluding Thoughts: Coupling arguments are a nice tool to make elegant arguments that better align with intuition than "raw calculations." You won't have to use them extensively in 445, but they will be relevant when we discuss information cascades, and it's a good tool to have in your toolkit to avoid messy calculations.

1.8   Stochastic Dominance

Stochastic Dominance is a useful relationship between two single-variable distributions, and also a good demonstration of coupling arguments. We won't use it much throughout the course, but it's generally a good concept to understand.

Definition 5 (Stochastic Dominance). *We say that a single-variable distribution D^+ stochastically dominates D if for all x, $\Pr_{v^+ \leftarrow D^+}[v^+ \geq x] \geq \Pr_{v \leftarrow D}[v \geq x]$. Equivalently, if F^+ is the CDF of D^+, and F is the CDF of D, then $F^+(x) \leq F(x)$ for all x.*[9]

We'll go through some examples shortly applying this definition, and a key fact (Fact 6). We'll first make a quick observation, state and prove the fact, and then go through the examples.

Observation 6. *Let D^+ stochastically dominate D. Let v^+ be a random variable drawn from D^+, and v be a random variable drawn from D. Then $\mathbb{E}[v^+] \geq \mathbb{E}[v]$.*

Proof: We have:

$$
\begin{aligned}
\mathbb{E}[v^+] &= \int_0^\infty 1 - F^+(x)dx \\
&\geq \int_0^\infty 1 - F(x)dx \\
&= E[v].
\end{aligned}
$$

The first and third equalities are just the definition of expected value. The middle line follows by stochastic dominance, as $1 - F^+(x) \geq 1 - F(x)$ for all x. ∎

Fact 6. *Let D^+ stochastically dominate D. Then there is a way to couple draws v^+ from D^+, and v from D so that $v^+ \geq v$ **with probability one**. That is, it is possible to jointly draw samples from D and D^+ so that the sample from D^+ is always bigger than the sample from D.*

---

[9]The equivalence between these two definitions isn't trivial, but you can use both versions interchangeably. If you're curious, the reverse implication is "easy" (using $\lim_{t \nearrow x} F(x) = \Pr[X < x]$ when the CDF of X is F); the direct one follows by taking right limits and the fact that CDFs have at most countably many points of discontinuity.

Proof: We'll prove the claim for continuous distributions with strictly monotone increasing CDF on its support (that is, the CDF is 0 on $[0, a]$, then strictly increasing on (a, b), then equal to 1 on $[b, \infty)$, for some $a < b$).[10] Recall that in order to properly establish a coupling argument, we must first define our random variables. To this end:

- Draw w^+ from D^+.

- Let $w := F^{-1}\big(F^+(w^+)\big)$. Note that our assumption on F implies F^{-1} exists (in the support $[a, b]$ of F), and we can set $w := 0$ if $F^+(w) = 0$ as well as $w := b$ if $F^+(w) = 1$.

- Set $v^+ := w^+$ and $v := w$.

We need to prove three claims:

- v^+ is distributed according to a draw from D^+. This is clearly true, as w^+ is drawn from D^+, and $v^+ := w^+$.

- v is distributed according to a draw from D. To see this, let's just explicitly compute the probability that $v \leq x$, for any x. First, note that $\Pr[v < a] = \Pr[v > b] = 0$. For x in the support $[a, b]$, we can write:

$$
\begin{aligned}
\Pr[v \leq x] &= \Pr\left[F^{-1}\big(F^+(w^+)\big) \leq x\right] \\
&= \Pr\left[w^+ \leq (F^+)^{-1}\big(F(x)\big)\right] \\
&= F^+\big((F^+)^{-1} \circ F(x)\big) \\
&= F(x).
\end{aligned}
$$

Above, the first line follows by definition of v. The second follows by just applying F and then $(F^+)^{-1}$ to both sides of the inequality (and using the fact that F^+ is also invertible in its support). The third line follows by definition of CDF ($F^+(x) := \Pr[w^+ \leq x]$, by definition of the CDF). The final line just cancels F^+ with $(F^+)^{-1}$.

- Finally, we have to ensure that $v^+ \geq v$ for all draws. To see this, observe that:

$$
\begin{aligned}
v &= F^{-1}(F^+(v^+)) \\
\Rightarrow F(v) &= F^+(v^+) \\
\Rightarrow F^+(v) &\leq F(v) = F^+(v^+) \\
\Rightarrow v &\leq v^+
\end{aligned}
$$

Here, the first line is just the definition of v. The second line applies F to both sides. The third line applies the definition of stochastic dominance. The fourth line follows because F^+ is an increasing function.

Now, we've defined a coupling, proved that both v^+ and v are drawn correctly, and also established that $v^+ \geq v$ with probability one. This completes the proof. ∎

Here are now a few examples of distributions that stochastically dominate each other, and the corresponding coupling in the proof of Fact 6.

---

[10]The proof for discrete distributions, or distributions with weakly monotone increasing CDF, follows the same outline, but requires some extra steps due to the fact that the CDF or its inverse is discontinuous. Due to the excessive calculations, the proof for the general case is omitted here. But, you may cite this fact for any distribution, discrete or continous, without proof.

Example One: Consider the uniform distribution on $[0, 1]$, which has CDF $F_1(x) := x$ on $x \in [0, 1]$, and $F_1(x) = 1$ for $x > 1$. Consider also the uniform distribution on $[0, 2]$, which has CDF $F_2(x) := x/2$ for $x \in [0, 2]$, and $F_2(x) = 1$ for $x > 2$. We can clearly see that $F_2(x) \leq F_1(x)$ for all x, and therefore the uniform distribution on $[0, 2]$ stochastically dominates the uniform distribution on $[0, 1]$.

Moreover, if we draw y uniformly at random from $[0, 1]$ and output the pair $(z_1, z_2) = (y, 2y)$, then z_1 is distributed according to uniform on $[0, 1]$. Also, z_2 is distributed according to uniform on $[0, 2]$. Also, $y \leq 2y$ with probability 1.

Example Two: Consider the exponential distribution with rate one, which has CDF $F_1(x) := 1 - e^{-x}$ on $[0, \infty)$, and the equal revenue curve, which has CDF $F_2(x) := 1 - 1/x$ on $[1, \infty)$, and $F_2(x) := 0$ on $[0, 1]$. Then $F_2(x) \leq F_1(x)$ for all x, as $0 \leq 1 - e^{-x}$ on $[0, 1]$, and $1 - 1/x \leq 1 - e^{-x}$ on $[1, \infty)$.[11] So the equal revenue curve stochastically dominates the exponential distribution with rate one.

Moreover, if we draw z_2 from the equal revenue curve and output $z_1 := \ln(z_2)$, then z_2 is distributed according to an equal revenue curve, and z_1 is distributed according to an exponential with rate one.[12] Also, $z_2 \geq z_1$ always, because $\ln(x) \leq x$ whenever $x \geq 1$.

And finally, here is an example of a simple proof that uses Fact 6. We'll reprove Observation 6 using Fact 6.

Alternate proof of Observation 6: Recall that D^+ stochastically dominates D. Therefore, by Fact 6, there is a way to couple draws (v^+, v) so that v^+ is a draw from D^+, v is a draw from D, and $v^+ \geq v$ with probability 1. It is now immediate that because $v^+ \geq v$ with probability 1, that $\mathbb{E}[v^+] \geq \mathbb{E}[v]$ as well. ∎

1.9 "Principle of Deferred Decisions"

The "principle of deferred decisions" isn't a formal theorem or definition, but a concept that will be useful throughout this class to simplify analyses. Consider the following example: say that there are n people who have purchased tickets for a plane. The plane has n seats, and each person purchases a ticket for a distinct seat. The passengers line up in order from 1 to n to board, but unfortunately the first passenger insists on taking a seat that is not their own, uniformly at random. Fortunately the remaining passengers are flexible: if their assigned seat is taken, they will happily take a uniformly random available seat (if their assigned seat is available, they will take it). We wish to understand: *what is the probability that the nth passenger sits in their own seat?*

At first glance, this seems like it should require some messy math. There's a lot of interaction going on between the different passengers (e.g. we first need to know which seat the first passenger takes in order to know which future passengers' decisions matter). But fortunately there's a (conceptually more challenging, but mathematically) simpler way to reason about this. To really appreciate the simplicity which will come later, try to analyze this probability without the tips below (we will not do this exercise in these notes).

Let's first make the following observation (which is not related to the principle of deferred decisions, but to this specific problem): the last passenger will sit either in their own seat, or in seat one. This is because no other seat can be empty: for all $i \in [2, n-1]$, if seat i is available when

---

[11] To quickly see this, observe that $1/x \geq e^{-x}$, as $x \leq e^x$ for all $x \in [1, \infty)$.

[12] To see this calculation, we have just applied the coupling used in the proof of Fact 6. We've set $z_1 := F_1^{-1}(F_2(z_2)) = -\ln(1 - F_2(z_2)) = -\ln(1 - (1 - 1/z_2)) = \ln(z_2)$.

passenger i boards, they will take it (and fill it). Otherwise, it is already filled. Clearly, exactly one of the two seats $\{1, n\}$ will be available when passenger n boards, and we just need to find out which one it is.

Now we can use the principle of deferred decisions. Every time a new passenger i boards the plane, their seat is either taken or not taken. If their seat is not taken, we ignore them. If seat n is already taken, we ignore them (because certainly passenger n will not sit in their own seat). If 1 is already taken, we ignore them (because now certainly passenger n will sit in their own seat). If their seat is taken and $1, n$ are not, then there are $n - i + 1$ remaining seats, and we need to pick one uniformly at random. One way to pick a uniformly random element from a set of size $n - i + 1$ is to just pick an element directly, each with probability $1/(n - i + 1)$. Another way (which may initially seem odd) is to first pick whether the element is in the set $\{1, n\}$ or not. The element is in the set $\{1, n\}$ with probability $2/(n - i + 1)$, and not otherwise. If the element is in the set $\{1, n\}$, then we are about to decide what happens with the last passenger! Conditioned on being in $\{1, n\}$, the probability of selecting 1 is $1/2$ (and that of selecting n is also $1/2$). So we see that **we are equally likely to select 1 or n, and therefore we are equally likely to have the last passenger sit in their own seat or** 1. If the element is not in the set $\{1, n\}$, we pick the seat uniformly at random from the remainder, and continue. Importantly, **at whatever step we finalize where passenger n will sit, it is equally likely that passenger n gets their own seat or** 1. Therefore, the probability that they get their own seat is exactly $1/2$.

Let's try to unpack the main idea used. One way to draw a uniformly random element of $\{1, \ldots, n\}$ is to just draw each element with probability $1/n$. Another option is to first draw whether the element is even or odd (say n is even). The element is even with probability $1/2$, and odd with probability $1/2$. Then, if we decide that the element is even, we could draw a uniformly random even number. This process will indeed select every even number with probability $1/n$, and also every odd number with probability $1/n$, as desired. The principle gets its name because we are *deferring* the decision of exactly what element to draw until after we have first decided whether it is even or odd. In the abstract, this probably seems not particularly useful, but it gave us a (conceptually challenging, but mathematically) simple proof above. We'll also revisit this in the "basic proof writing" section.

2 Basic continuous optimization

In COS 445, you'll occasionally need to use very basic continuous optimization. We say "basic" to emphasize that you do not need to know any fancy theorems, or remember any fancy approaches. The instances you'll need to solve in COS 445 will require you to be *clear and rigorous with the logic behind your approach*, but will not require that approach to be fancy.

As a quick recap, here is a reminder of some terms:

- The **global maximum** of a function $f: \mathbb{R} \to \mathbb{R}$ is a number y such that $f(y) \geq f(x)$ for all $x \in \mathbb{R}$. Similarly, the global maximum of a function $f: \mathbb{R}^n \to \mathbb{R}$ is a vector $\vec{y}$ such that $f(\vec{y}) \geq f(\vec{x})$ for all $\vec{x} \in \mathbb{R}^n$. The global minimum is the same, after flipping $\geq$ for $\leq$.

- The **constrained maximum** of a function $f: \mathbb{R} \to \mathbb{R}$ over an interval $[a, b]$ is a number $y \in [a, b]$ such that $f(y) \geq f(x)$ for all $x \in [a, b]$.[13] Similarly, the constrained maximum of a function $f: \mathbb{R}^n \to \mathbb{R}$ over a region $S \subseteq \mathbb{R}^n$ is a vector $\vec{y} \in S$ such that $f(\vec{y}) \geq f(\vec{x})$ for all $\vec{x} \in S$.[14] The constrained minimum is the same, after flipping $\geq$ for $\leq$.

---

[13]Note that f doesn't need to be defined over all of $\mathbb{R}$; it suffices for the domain D of f to contain $[a, b]$.
[14]Ditto, with domain $D \subseteq \mathbb{R}^n$ containing S.

So, whenever you are asked to find the (constrained or unconstrained) maximum, your goal is to find an $\vec{y}$ such that $f(\vec{y}) \geq f(\vec{x})$ for all relevant $\vec{x}$, and prove it. There is no requirement that you use any particular approach, as long as you can prove this claim. The rest of these notes give a reminder of some simple approaches.

2.1 Single-variable, unconstrained optimization

Say we want to find the global maximum of a continuous, differentiable function $f(\cdot)$. Here, you may remember that every global maximum must also be a local maximum. You may also remember that every local maximum must be a critical point, where $f'(x) = 0$. But, remember that not all critical points are local optima, and not all local optima are local maxima, and not all local maxima are global maxima. Moreover, recall that some functions may not even have a global maximum, depending on what happens as $\lim_{x \to \pm\infty} f(x)$.

Example 1: For example, say we want to find the global maximum of $f(x) = x^2$. There is a unique critical point at $x = 0$. Therefore, we can immediately conclude the following: *if $f(x)$ has a global maximum, it must be at $x = 0$*. However, $\lim_{x \to \infty} x^2 = \infty$, so the function doesn't attain its global maximum.

Example 2: Say we want to find the global maximum of $f(x) = 4x - x^2$. The derivative is $4 - 2x$, so there is a unique critical point at $x = 2$. So we can again conclude that if there is a global maximum, it must be $x = 2$. This time, we can observe that $f'(x) \geq 0$ when $x \leq 2$. This immediately implies that $f(2) \geq f(x)$ for all $x \leq 2$. Similarly, $f'(x) \leq 0$ when $x \geq 2$. This immediately implies that $f(2) \geq f(x)$ for all $x \geq 2$. Therefore, $x = 2$ is the global maximum.[15]

In the above example, observe that we didn't need a second-derivative test, or any other fancy math. We just needed any logical approach to prove that $f(2) \geq f(x)$ for all $x \in \mathbb{R}$.

2.2 Single-variable, constrained optimization

Say now we want to find the constrained maximum of a differentiable function $f(\cdot)$ over the interval $[a, b]$. Now, any value that is the constrained maximum must either be a critical point, or an endpoint of the interval. Here are a few approaches to find the constrained maximum:

- Find all critical points, compute $f(a)$, $f(b)$, $f(x)$ for all critical points x and output the largest.

- Confirm that $f'(a) > 0$ (that is, f is increasing at a) and $f'(b) < 0$. This proves that neither a nor b can be the global maximum. Then compute $f(x)$ for all critical points x and output the largest in (a, b).

- In either of the above, rather than directly comparing $f(x)$ to $f(y)$, one can instead prove that $f'(z) \geq 0$ on the entire interval $[x, y]$ to conclude that $f(y) \geq f(x)$.

- Prove that x is a global *unconstrained* maximum of $f(\cdot)$, and observe that $x \in [a, b]$.

---

[15]We can also verify that $x = 2$ is a local maximum by computing $f''(2) = -2$, but this isn't necessary — the proof above is complete.

There are many other approaches. The point is that at the end of the day, you must directly or indirectly compare all critical points and all endpoints. You don't have to directly compute $f(\cdot)$ at all of these values (the bullets above provide some shortcuts), but you must at least indirectly compare them. For this class, it is OK to just describe your approach without writing down the entire calculations (as in the following examples).

Example 3: Say we want to find the constrained maximum of $f(x) = x^2$ on the interval $[3, 8]$. f has no critical points on this range, so the maximum must be either 3 or 8. $f'(x) = 2x > 0$ on this entire interval, so therefore the maximum must be 8.

Example 4: Say we want to find the constrained maximum of $f(x) = 3x^2 - x^3$ on the interval $[-2, 3]$. $f'(x) = 6x - 3x^2$, and therefore f has critical points at 0 and 2. So we need to (at least indirectly) consider $-2, 0, 2, 3$. We see that $f'(x) \leq 0$ on $[-2, 0]$, so we can immediately conclude that $f(-2) \geq f(0)$. We also see that $f'(x) \leq 0$ on $[2, 3]$, so we can immediately conclude that $f(2) \geq f(3)$. Now, we only need to compare -2 and 2. We can also immediately see that $f(-x) > f(x)$ for all $x > 0$, and therefore $f(-2) > f(2)$, and $x = -2$ is the global constrained maximum.

Example 5: Say we want to find the constrained maximum of $f(x) = 4x - x^2$ on the interval $[-8, 5]$. We already proved above that $x = 2$ is the global *unconstrained* maximum (Example 2). Therefore, $f(2) \geq f(x)$ for all $x \in \mathbb{R}$, and certainly $f(2) \geq f(x)$ for all $x \in [-8, 5]$. Therefore $x = 2$ is also the global constrained maximum on $[-8, 5]$.

Warning! An incorrect approach. It might be tempting to try the following approach: First, find all local maxima of $f(\cdot)$. Call this set X. Then, check to see which elements of X lie in $[a, b]$. Call them Y. Then, output the argmax of $f(x)$ over all $x \in Y$. This approach *does not work*, and in fact we already saw a counterexample. Say we want to find the constrained maximum of $f(x) = 3x^2 - x^3$ on the interval $[-2, 3]$. Then $f'(x) = 6x - 3x^2$, and f has critical points at 0 and 2. We can verify that $x = 0$ is a local minimum and $x = 2$ is a local maximum. So $x = 2$ is the unique local maximum, and it also lies in $[-2, 3]$. But, we saw that it's incorrect to conclude that therefore $x = 2$ is the constrained global maximum.

In general, remember that your goal is to find a y such that $f(y) \geq f(x)$ for all $x \in [a, b]$. You can use shortcuts (you know that the maximum must be either a critical point, or an endpoint). You can also use derivatives to compare $f(x)$ vs. $f(y)$ without explicitly computing $f(x)$ or $f(y)$. You can also use all sorts of other tricks to save on calculations (and you will never need to use any fancy tricks in this course). But at the end of the day, you must provide a logically sound argument that $f(y) \geq f(x)$ for all $x \in [a, b]$. The above examples are arguments that are complete, and plenty rigorous for this course.

2.3 Multi-variable, unconstrained optimization

Say now we want to find the unconstrained global maximum of a differentiable multi-variate function $f(\cdot, \cdot, \ldots, \cdot)$. Again, any value that is the unconstrained maximum must be a critical point, where a critical point has $\frac{\partial f(\vec{x})}{\partial x_i} = 0$ for all i. Again, not all critical points are local optima/maxima, but all local maxima are definitely critical points. Recall also that some f don't achieve their global maximum, depending what happens when approaching ∞. Having a general approach that works

in all cases is *quite* tedious, but in this class we'll only see cases where a simple approach works.[16] Again, remember that your goal is just to provide a logically sound proof that $f(\vec{y}) \geq f(\vec{x})$ for all $\vec{x} \in \mathbb{R}^n$. Here are some examples that you might reasonably need to solve, and an approach to solve them:

Example 6: Say you want to maximize $f(x_1, x_2) = x_1 - x_1^2 - x_2^2$. We will solve this in two steps. First, we will think of x_1 as fixed, and try to first maximize $f(x_1, y)$ as a function of y. To this end, observe that $f(x_1, y) = x_1 - x_1^2 - y^2$. As a function of y, this is just a constant minus y^2, which is clearly maximized at $y = 0$. Therefore, we can immediately conclude that $f(x_1, 0) \geq f(x_1, x_2)$ for all x_2. This is the key first step, and we now know that the optimum must be of the form $(x_1, 0)$, for some x_1.

Now, we just need to optimize $f(x_1, 0) = x_1 - x_1^2$ over x_1. This is a single-variable optimization problem, which we can solve using the tools in Section 2.1. We notice that the derivative is $1 - 2x_1$, which is positive on $(-\infty, 1/2)$, and negative on $(1/2, \infty)$, meaning that $x_1 = 1/2$ is the global maximum. This means that the unconstrained global maximum is $(1/2, 0)$.

To recap the complete logic, what we've shown is that for any (x_1, x_2), we have: $f(1/2, 0) \geq f(x_1, 0) \geq f(x_1, x_2)$. The previous paragraph proves the first inequality, and the first paragraph proves the second.

Example 7: Say you want to maximize $f(x_1, x_2) = x_1 x_2 - x_1^2 - x_2^2$. We can again think of x_1 as fixed, and try to optimize $f(x_1, y)$ over y. Because we treat x_1 as fixed, this is again a single-variable optimization problem. The derivative with respect to y is $x_1 - 2y$. We conclude that the derivative is positive on $(-\infty, x_1/2)$, and negative on $(x_1/2, \infty)$, and therefore $y = x_1/2$ is the unconstrained optimum. This directly proves that for all x_1, $f(x_1, x_1/2) \geq f(x_1, x_2)$, for all x_2.

Now, we want to find the maximizer among points of the form $(x_1, x_1/2)$. Observe that now this is again a single-variable function, and its equal to $x_1 \cdot (x_1/2) - x_1^2 - (x_1/2)^2 = (-3/4)x_1^2$. This is clearly maximized at $x_1 = 0$, so the global maximizer is $(0, 0)$.

To again recap the complete logic, we've shown, for any (x_1, x_2) that $f(0, 0) \geq f(x_1, x_1/2) \geq f(x_1, x_2)$. The previous paragraph proves the first inequality, and the first paragraph proves the second.

Example 8: Say you want to maximize $f(\vec{x}) = \sum_i f_i(x_i)$. That is, the function you're trying to maximize is just the sum of single-variable functions (one for each coordinate of $\vec{x}$). Then we can simply maximize each $f_i(\cdot)$ separately, and let $x_i^* = \arg\max_{x_i}\{f_i(x_i)\}$. Observe that $\vec{x}^*$ must be the maximizer of $f(\vec{x})$.

To recap the complete logic, we've shown, for any $\vec{x}$, that $f(\vec{x}^*) = \sum_i f_i(x_i^*) \geq \sum_i f_i(x_i) = f(\vec{x})$. The middle inequality follows simply because every term in the sum on the left is at least the corresponding term in the sum on the right.

2.4 Multi-variable, constrained optimization

Finally, say we want to find the constrained global maximum of a differentiable multi-variate function $f(\cdot, \ldots, \cdot)$. Then the same rules as before apply: we must (at least indirectly) consider all critical points and all extreme points. Multi-variable constrained optimization in general is tricky, and would require an entire class to learn enough tricks to solve every instance. All of the instances that you'll need to solve in this class can be done by using a simple approach (building

---

[16]Sometimes you'll need to be clever, but ideally very few (if any) proofs will require very tedious calculations.

off approaches from the previous section). Just remember that at the end of the day, you just need a logically sound approach to compare your claimed optimum to all critical points and extreme points.

Example 9: For example, say you want to maximize $f(\vec{x}) = \sum_i x_i e^{-x_i}$, subject to the constraints $-5 \leq x_i \leq 5$ for all i. We can first try to find the unconstrained maximizer using the approach in Example 8, because this is a sum of single-variable functions. Indeed, the derivative of each single-variable function is of the form $e^{-x_i} - x_i e^{-x_i}$, which is positive when $x_i < 1$, and negative when $x_i > 1$. This means that the maximizer of this single-variable function is at $x_i = 1$. Therefore, $(1, \ldots, 1)$ is the global maximizer. We observe that $-5 \leq 1 \leq 5$, so $(1, \ldots, 1)$ also satisfies the constraints. So $(1, \ldots, 1)$ is also the constrained maximizer over $[-5, 5]^n$.

To recap the complete logic, we first showed that $f(1, \ldots, 1) \geq f(\vec{x})$ for all $\vec{x} \in \mathbb{R}^n$. So therefore, clearly $f(1, \ldots, 1) \geq f(\vec{x})$ for all $\vec{x} \in [-5, 5]^n$. Also, we confirmed that $(1, \ldots, 1) \in [-5, 5]^n$, so it is the global maximum.

Repeat Warning! Again, recall that it is **not** a valid approach to first find all critical points of f, and then see which critical points satisfy the constraints and only consider those (recall example at the end of Section 2.2).

3 Basic Problem Solving

The PSets for this class are "short", in that they're only three problems (four if you count the Strategy Designs). But some of the problems will be a full paragraph description, introduce new definitions, etc. **Part of the challenge is figuring out on your own how to break these problems down into tractable subparts.** Problem solving is more of an art, so I can't recommend a concrete step-by-step procedure. However, I can try to give general guidelines/tips. I will use the following problem as a running example for this section:

> Recall that a bipartite graph has two sets of nodes, L and R, with all edges having one endpoint in L and the other in R. Recall also that a perfect matching is a set of edges such that every node is in exactly one edge.
>
> Let G be a bipartite graph with n nodes on each side. Prove that if every node has degree $\geq n/2$, then G has a perfect matching.
>
> **Hint:** You may use Hall's Marriage Theorem. Recall that Hall's Marriage Theorem asserts that a bipartite graph has a perfect matching if and only if for every set $S \subseteq L$ of nodes on the left, we have $|N(S)| \geq |S|$, where $N(S)$ denotes the set of nodes with an edge to some node in S.

Step One: Understand what the question is asking. Just because you've taken 240 doesn't mean you're expected to remember off the top of your head exactly what a bipartite graph or perfect matching is, or Hall's Marriage Theorem. The first step is just understanding what all the words mean, and what the question is asking, **and this is part of the assignment**. Normally, all the terms will either be defined in the problem itself, or used in lectures. If that is not the case, you are free to use the Internet to look up terms, and can ask on Piazza if the Internet fails.

In this example, the second paragraph is the actual question. When I read the first sentence, I might get stuck because I don't know what a bipartite graph is. So I should revisit the definition

given above. If that's still not clear, I might Google "bipartite graph" to find the Wikipedia page or see some examples. I might even draw some small examples. But I definitely need to know what a bipartite graph is before I can move on to the second sentence, and I should take as much time as necessary to understand the definition. For the second sentence, I may also get stuck on what a "perfect matching" is. I should also refer to the definition given above. Again, if it's not immediately clear, I might draw some small graphs, and look at sets of edges and check whether they satisfy the definition or not. I might also try Google again to see if I can find some examples online.

There is also this hint, which references "Hall's Marriage Theorem." Probably I should know what that means before I start thinking. The statement itself might take some time to parse, because it introduces new notation $N(S)$. But after a few slow reads, I can figure out that every set $S \subseteq L$ must have at least as many neighbors in R as there are elements in S. I might also realize that this makes sense as a necessary condition: if $|N(S)| < |S|$ for any S, then clearly I can't match every node in S to a distinct neighbor, because there are only $< |S|$ possibilities. What's surprising is that this is a sufficient condition, but Hall proved that so we don't need to worry about it.

Some hints are useful to get your thinking started. Other hints are useful to help you formalize the proof once you get the right idea. It's your job to figure out how to use the hint (if you choose to use it at all). In this case, it feels like I probably want to use the hint to get started, since it gives an equivalent condition to G having a perfect matching.

In general, the main thing I want to emphasize here is that you should expect it to sometimes take several minutes just to understand what the question is asking, and that figuring out what the question is asking is part of the challenge. You may need to reread each sentence several times to really understand what it's saying before moving on, and you shouldn't expect to figure it out on your first read-through. It's of course possible to get stuck on this step, and it's fine to use office hours/Piazza for help getting unstuck.

Step Two: Try Stuff! Once you understand what the question is asking, there's no recipe for what to do next. So you have to just play around until inspiration hits. General tips are: try to construct a counterexample and see where you get stuck, or working through small examples (or both). **But it will generally not be obvious what you should try.**

Let's try a small example first, with $n = 2$. When $n = 2$, I can't help but notice that Hall's Theorem seems especially easy to use, since there are only two nodes on the Left (call them u and v, and let the Right nodes be $\{x, w\}$), and therefore only three non-empty subsets to consider. Moreover, we are guaranteed that u has at least one neighbor, so Hall's condition is satisfied. We are also guaranteed that v has at least one neighbor. But we aren't immediately guaranteed that $\{u, v\}$ has two neighbors: What if both have an edge to the same w on Right? But now we see that, if this were the case, then x would have zero neighbors, violating the hypothesis that every node has at least $n/2$ neighbors. Therefore, both x, w must be neighbors of $\{u, v\}$, and Hall's conditions are satisfied for all sets, so there must be a perfect matching.

This gives me some intuition for how to go about the general case. First, it leads me to guess that I should indeed try to use Hall's Theorem right away. Second, it suggests that I need to try and reason about different sets using different logic, because we reasoned about u and v differently than $\{u, v\}$. So let's jump to the general case and see what happens.

Consider an arbitrary n, and we want to show that when every node has degree $\geq n/2$, that every set S on the left has at least $|S|$ neighbors. I also can't help but notice, similarly to the $n = 2$ case, that some sets S will be easy to reason about: if $|S| \leq n/2$, then *just a single node in S* already has $\geq n/2$ neighbors, so clearly S has $\geq n/2$ neighbors and Hall's condition is satisfied. This seems exciting, so let's try thinking about $|S| > n/2$ now. This feels trickier: I do know that

there must be $\geq n/2$ neighbors of S, but what if every node in S has the same neighbors? Let's consider what must happen for this to occur. Let $N(S)$ denote the neighbors of S, and let T denote the remaining nodes in Right. We have at our disposal that $|S| > n/2$, and that $|N(S)| < |S|$, and therefore T is non-empty.

From here, I might again get quickly stuck, since the next step isn't obvious. Maybe I would try some random things, or maybe I would revisit the $n = 2$ case, or maybe I'd try to draw out an example with $n = 3$ or $n = 4$ with a particular S with $|S| > n/2$. But eventually I'd get to the realization that any node in T has *no neighbors in S*, which accounts for $> n/2$ of the nodes in Left. Therefore, any node in T has $< n/2$ neighbors, which would contradict our hypothesis. Therefore, T must be empty, which in turn means that we cannot have $|N(S)| < |S|$ for any S of size $> n/2$, which covers all remaining cases.

The two main ideas I want to convey with this example are: (a) you're not supposed to know exactly what the question is asking on your first read-through. It might take several minutes just to figure that out, and (b) you're not supposed to know exactly what step to take next to solve the problem, even if you understand the lectures perfectly. You're supposed to get stuck, and there's no recipe for getting unstuck. I tried to give a few tips above (work through small examples, try to handle special cases first), but it's mostly an art that you learn by doing (the PSets). While it may have taken you ten minutes to read the previous two pages, it took me *way* longer to come up with those ideas (in particular, I spent quite some time figuring out the idea from the paragraph "From here, I might again get quickly stuck, since the next step isn't obvious." An excellent student in 445 could easily be stuck at this stage for an hour or longer (and on an actual PSet question in the course, much longer) – that's part of the challenge!

Also, please note that the writing above is me trying to explain my thought process. This is not how I would write up a proof once I figure it out (in particular, I would skip the $n = 2$ case, write clearer statements, etc.). See Section 4 for how I would write up a solution to this problem!

3.1   Partial Progress and Partial Credit

Assignments are hard, and you're not expected to fully solve every problem. Here, I just want to elaborate on "productive" partial progress versus "unproductive" partial progress. Here is the key thought experiment to keep in mind: your solution will be evaluated through the lens of "how much would your writeup help another student solve this problem?" *not* "how hard does your writeup prove you worked?" or "how smart does your writeup prove you are?" Here are some implications of this:

- **Logical clarity matters!** If your solution is unclear or imprecise, it will be hard for the grader to figure out exactly what you proved. If your solution is clear and precise, even though it is not complete, the grader can easily figure out what you proved, and give you credit for it.

- Shoot for **Concrete Partial Progress**. Here are two paths (we'll give examples shortly).

  - **Fully Solve a Special Case.** For example, if the problem asks you to prove a claim for all n, prove it first for $n = 2$. Add $n = 3$, or larger n if you can.

  - Prove a **Clearly Useful and Clearly Stated Lemma.** For example, if the problem asks you to prove that X holds if and only if Y holds, prove that X implies Y.[17]

---

[17]There are also more creative ways to prove a clearly useful and clearly stated lemma. But to contribute to partial credit, it must be both clearly useful and clearly stated. You won't get credit for proving random claims that the grader finds unrelated to the problem.

- **Be up front about what your solution accomplishes**. The better the grader understands what your solution accomplishes, the more partial credit they can award. If you're up front about what your solution accomplishes, this will make it easier for the grader to give you partial credit. Additionally, the graders are instructed to give a couple points back for partial-credit solutions that were exceptionally easy to grade.

  – On a related note, **do not try to sneak something** false, vague, or otherwise wrong past the graders. The graders will always[18] catch it. Mistakes happen, and the graders aren't out for blood to punish mistakes. But if it's challenging for the graders to figure out what your solution accomplishes, you won't earn full presentation marks on the generic rubric (see below).

Here are two examples for the problem in Section 3.

> I was unable to solve the general case, but here is a proof for $n = 2$. Let u, v denote the two nodes on the left. Observe that Hall's Theorem says that a perfect matching exists if and only if: (a) u has at least one neighbor, (b) v has at least one neighbor, and (c) the set $\{u, v\}$ has two neighbors.
>
> Observe that, immediately because every node has degree at least $n/2$, that both u and v must have at least one neighbor, so this covers (a) and (b). To consider $\{u, v\}$, first observe that both right nodes have at least one neighbor (again, immediately from the problem statement). Because $\{u, v\}$ are the only possible neighbors this means that each right node has a neighbor in $\{u, v\}$. In particular, this means that the set $\{u, v\}$ has both right nodes as neighbors, which covers (c).
>
> In the general case, we can again seek to apply Hall's theorem. For any set S with $|S| \leq n/2$, observe that even a single node in S has $n/2$ neighbors, which implies that $|N(S)| \geq n/2 \geq |S|$, as desired. But I can't figure out how to handle sets with $|S| > n/2$.

This solution makes a lot of partial progress. First, the logical clarity is excellent — it is easy for a grader to figure out exactly what is proved. Additionally, it provides both examples of concrete partial progress: First, it provides a complete proof in the case of $n = 2$. Second, it provides the easy half of a proof for the general case. If I were to grade this solution, I might give it a 14/20.

> Below is a picture of a graph on 4 nodes without a perfect matching (imagine that a picture is given). We note that it has no perfect matching because it violates Hall's Theorem. We also note node u has < 2 neighbors, violating the problem hypothesis. It seems like it would be really challenging to add a neighbor to u without creating a perfect matching. For example, if we add any single edge to u, this yields a perfect matching. Of course, maybe we can add an edge to u and delete another edge, but this also seems unlikely to work.
>
> Also recall that Hall's Theorem can be proved using max-flow-min-cut. As such, we could also consider a proof approach by writing out a network with a source on the left with capacity 1 edge to all left nodes, infinite weight edges from left nodes to right nodes, and capacity 1 edges to a sink on the right. We could then try to prove that whenever each edge has degree at least $n/2$ that there is a flow of weight n, and therefore a perfect matching.

---

[18]Obviously not always, but it is *much* easier for graders to detect logical flaws and imprecise claims than it is to detect mistakes in less mathematical courses.

This solution conveys that the author clearly has a lot of ideas. I think it is a good idea to write down a concrete example and play around with it. I also think it's a good idea to see whether max-flow-min-cut helps at all with a proof. But unfortunately there's not much else here. The logical clarity of this solution is pretty good (there are no flaws, and I find it easy to read). And there's no false or frustrating claims. But I just can't find any concrete partial progress to give credit for. While I'm convinced that the author tried, and did come up with some ideas, none of the ideas make concrete progress towards a solution. Put another way, the solution does convince me that the author is further along towards a solution than when they first started, but the solution wouldn't help another problem-solver. If I were to grade this solution, I might give it a 5/20.

I hope that the examples above help display the difference between productive and unproductive partial progress.[19] Here are two last suggestions:

First, whenever you're stuck on a problem, I strongly recommend doing something small, but well (e.g. a proof for a special case, or a clear proof of a concrete stepping stone). It may be helpful to think of an analogy to programming: it's always best if you can provide well-documented code that correctly solves the task you were assigned. Short of that, it is *significantly* better to submit well-documented code that correctly solves a less ambitious task, than to submit buggy or unparseable code that may or may not solve the entire task.

Second, once you've solved as much as you can and are now writing it up, logical clarity plays a huge role in partial progress. If you submit a *well-executed* partial solution, that will surely help another student quite well! If your solution is poorly presented, it is harder for the grader to find concrete progress for partial credit. Note that I'm trying really hard to emphasize logical clarity, and not fancy formatting or fancy pictures. The grader needs to understand exactly what you're claiming (due to strong logical clarity), rather than feel that you worked really hard (due to good formatting and pictures).

3.1.1 A Generic Rubric

To help you understand in more detail how a typical problem is graded, here is a generic rubric template (for a 20-point problem). A majority (maybe super-majority) of problems you solve in this course will start with this as a rubric template, and fill in details with common examples of various levels of 'concrete partial progress'.

Quality of Progress. (Note the emphasis on *logical flow* below).

16 points The solution is fully correct. In particular, there is a concrete outline (perhaps implicit, not necessarily explicitly stated) that is correct, and each step of the outline is implemented correctly with an appropriate level of detail. In particular, the *logical flow* of the solution is fully correct and explained clearly.

13 points The solution clearly has all of the main ideas, but there are some minor (but non-trivial) issues with the *logical flow*. For example:

- Perhaps too many details are skipped to connect different parts of the solution's *logical flow* together.

- Perhaps the claimed *logical flow* of the solution is not fully concrete enough to form a connected outline.

---

[19]All examples in this document were made up by me for the purpose of explanation — these are not actual problems I've asked in 445, nor actual student solutions.

- Perhaps the solution makes a minor (but not trivial) mistake that requires a change in *logical flow* to resolve.

- Perhaps the solution is almost complete, except it cannot resolve some leftover cases.

10 points The solution makes significant, concrete, partial progress, but does not fully solve the problem. For example:

- Perhaps the solution proves a significant concrete lemma related to the problem, but does not connect it all the way to a complete solution.

- Perhaps the solution fully solves many cases, but there are also many cases unsolved.

- In general, whatever the solution contains is well-executed, and significant. But there is still significant work left to do to reach a full solution.

6 points The solution makes some concrete partial progress, but it is not very far towards a complete solution. For example:

- Perhaps the solution proves a concrete lemma related to the problem (that is novel, and not just a restatement of lemmas shown in class), but it is a not a significant fraction towards a complete solution.

- Perhaps the solution fully solves some very special cases, but most cases are unsolved (e.g. perhaps the problem asks for a solution 'for all n', and the solution is fully correct for $n \leq 5$).

- Whatever the solution contains is well-executed, non-trivial, and contains *something concrete, novel, and related to the problem.*

2 points The solution does not make any concrete partial progress, but demonstrates a good understanding of the problem being asked. For example, the solution may clearly state some solid ideas that the author tried, or demonstrate a strong understanding of the definitions posed in the problem, and some approaches to possibly solve it.

0 points The solution is absent or flawed.

Quality of Presentation. (Note the emphasis on *ease of evaluation* below).

4 points The solution is written clearly. It is easy to understand the solution's logical flow, and it is easy to see that each step of the outline is implemented with appropriate detail.

It is OK if the outline is not explicitly stated (provided that one can find the outline while reading the solution). If there are grammatical mistakes, typos, or calculation errors, this is OK *if they do not interfere with ability to process the logical flow*. If the mistakes are severe enough that it becomes hard to process the logical flow, a lower mark should be used.

Note: A blank solution is trivial to evaluate, and would receive 4 points. Similarly, a solution that clearly states there is no concrete partial progress and that the only goal of what's written is to demonstrate ideas is also trivial to evaluate, and would receive 4 points.

3 points The solution is written pretty clearly. It is generally easy to understand the solution's concrete outline, and it is not too bad to see that each step of the outline is implemented with appropriate detail. But, there may have been parts of the logical flow that were tricky to follow.

2 points The solution was written OK, but required some struggle to understand the logical flow. Perhaps a large part of the solution was challenging to understand, although there are also parts that are clear.

1 point Some parts of the solution were clear and easy to follow, but much of the solution requires some struggle to understand the logical flow.

0 points The solution was tough to evaluate. The solution may have made imprecise claims, and it was not easy to figure out what was intended. The solution may have made false claims, and it was not easy to see why they were false. It may have been a struggle to figure out the intended logical flow, or to figure out exactly what concrete claims the solution accomplished.

> **Note:** As a representative example to have in mind, any solution produced by first-gen LLMs is likely to get 0 presentation points — the logical flow is ultimately incorrect, but because the language is so smooth it is a *huge pain* to figure out exactly where the problems are.[20]

Using this rubric, I would have given the first sample solution a 10/16 + 4/4: the solution made significant partial progress, but does not have a clear outline for the full solution. However, it is very easy for me to read the solution and understand exactly what the solution accomplishes, so I'd give 4/4 on presentation points. I would have given the second solution a 2/16 + 3/4: I unfortunately can't find any concrete (even small!) claims to give partial credit for, although there is certainly good intuition demonstrated. The writing is fine, but it was not trivial to determine that there was no concrete content — it is hard for me to figure out where the solution is going and what it's trying to do. But, it is still pretty fine in terms of presentation.

See the following section for writing tips to help make it clear where your solution lies on the rubric, and also to help avoid presentation deductions.

4 Basic Proof Writing

This is a bit of an oversimplification, but I think there are two 'kinds' of proof-writing that this class will develop. First, you must be able to write rigorous, complete proofs of short claims. Second, you must be able to write a clear, rigorous outline for a complex proof, by breaking it down into concrete, rigorous claims. Section 4.1 deals with the first kind, and Section 4.2 deals with the second. **I strongly recommend reading both sections.**

4.1 Writing short proofs

I found the following source: `https://math.dartmouth.edu/archive/m31x12/public_html/Proof%20Writing.pdf` to be a good quick source for tips on writing a proof. In particular, this source notes that while a good proof should be written in complete sentences (and not a sequence of formal mathematical statements), it should still be possible for a reader to understand what is the sequence of formal mathematical statements which corresponds to your complete

---

[20]This is just an example to help give you an idea of what a "really poorly written" solution might look like — it is certainly possible to create such solutions without the help of first-gen LLMs, and the application of this rubric will be done agnostically to how you created your solution (and agnostic to however the grader may think you came up with your solution).

sentences. I'll also include below a few generals pitfalls I noticed in previous semesters. Note that most proofs you'll write in 445 are much longer than the examples below, but hopefully it is enough to give a sense of what these pitfalls might look like.

Pitfall One: False Implications.   The most common reason that a proof is incorrect is that there is a false implication along the way. For example, let's revisit the incorrect approach under single-variate optimization. If I were trying to find the maximum of $f(x) := 3x^2 - x^3$ on the interval $[-2, 3]$ and wrote the following:

> The derivative of $f(x)$ is $f'(x) = 6x - 3x^2$. There are two critical points: $x = 0$ and $x = 2$. As $f''(x) = 6 - 6x$, we see that $x = 0$ is a local minimum, and $x = 2$ is a local maximum. Because $x = 2$ is the only local maximum, it must also be the constrained global maximum.

This proof is "obviously" incorrect, because it claims that $x = 2$ is the constrained global maximum (when it is $x = -2$). Let me change the example slightly, so that we are trying to are trying to find the maximum of $f(x) := 3x^2 - x^3$ on the interval $[0, 3]$, and repeat the same proof, word for word:

> The derivative of $f(x)$ is $f'(x) = 6x - 3x^2$. There are two critical points: $x = 0$ and $x = 2$. As $f''(x) = 6 - 6x$, we see that $x = 0$ is a local minimum, and $x = 2$ is a local maximum. Because $x = 2$ is the only local maximum, it must also be the constrained global maximum.

Even though $x = 2$ is indeed the constrained global maximum (so the "solution" is correct), the proof is still incorrect, but it's now harder to see why. The very last line of the "proof" is combining two logical statements together. First, it is claiming that every constrained maximum must also be a local maximum. Second, it is observing that because there is only one local maximum, it must be the constrained global maximum (otherwise, the constrained maximum would not be a local maximum, contradicting the first claim). But the first of these logical claims is false. Indeed, we just saw an example above where $x = 2$ is the only local maximum, but is *not* the constrained global maximum (and therefore, the constrained global maximum is not a local maximum).

So to summarize the flaw here, when we mapped the last sentence of the proof into the corresponding logical claim, that claim was false, and therefore the proof is incorrect (even though the final conclusion happens to be true). As a general rule: if the same proof, verbatim, could be used to prove a statement that is false, then the proof is certainly incorrect. Similarly, if the same sentence, verbatim, could be inserted into an otherwise correct proof of a false statement, then that sentence is certainly incorrect.

Pitfall Two: Overly Vague Implications.   Another reason for proof to be incorrect is that it is not possible for the reader to map the complete sentences to logical claims. In particular, maybe the sentence is too imprecise, and it could reasonably be intending to make many possible logical claims, some of which are false. For the same example (again, proving that $x = 2$ is the constrained global maximum for $f(x) := 3x^2 - x^3$ on $[0, 3]$), consider the following sentence:

> $x = 2$ is a critical point, with negative derivative to the right and positive derivative to the left. Therefore, $x = 2$ is the constrained global maximum.

A favorable interpretation of this sentence is that the author is claiming that because the derivative is negative on $[2, 3]$ (to the right), and positive on $[0, 2]$ (to the left), then $x = 2$ must be the

constrained global maximum. This is a correct argument. However, an equally reasonable inter-pretation of this sentence is that the author is claiming that because the derivative is negative on some interval $[2, z)$ (to the right), and positive on some interval $(y, 2]$ (to the left), then $x = 2$ is a constrained global maximum. This argument is false (as it only proves that that x is a local maximum).

Depending on the surrounding context (e.g. if the author states prior to this sentence that "to the left" means "all the way left until the end of the interval"), maybe this particular sentence could be clear. But in isolation, the underlying logical claim is unclear. A similar general rule applies here: if your English sentence could reasonably be mapped (taking the surrounding context into account) into a logical claim that is false, then it is incorrect.[21]

Pitfall Three: Too Many Missing Steps. In 445, you are *certainly* not expected to "show your work" for mundane calculations. But it is still possible to pack too many logical claims into the same short sentence in a way that the reader has no hope of following. There's no objective measure for what counts as too many, but you should expect to learn throughout the semester (via the lecture notes, staff solutions to PSets, other students' solutions to PSets that you see on MTA) what is considered sufficient detail. In general, my goal is for the staff solutions to provide a little bit more detail than what is necessary for full credit.

For example, for 445, the following would be plenty sufficient to prove that the global maximum of $f(x) := 3x^2 - x^3$ on $[0, 3]$ is $x = 2$.

> $f'(x) = 6x - 3x^2$, and therefore the only critical points are 0 and 2. Note that $f'(x) \geq 0$ on $[0, 2]$, and $f'(x) \leq 0$ on $[2, 3]$. Therefore, $x = 2$ is the constrained global maximum.

Let me give a slightly different example, say that you are given the word problem: Alice sells apples, and knows that if she sets price $x \in [0, 3]$ per apple, then exactly $f(x) := 3x - x^2$ apples will be purchased. What price x should Alice set to maximize her total revenue from all sold apples? Then the following answer is *not* sufficient, because it skips too many steps:

> The constrained global max of $xf(x)$ on $[0, 3]$ is $x = 2$, so Alice should set price 2.

The above answer forces the reader to do too much in their head. There is nothing factually inaccurate about the above proof, but it essentially skips the first half of the problem (why is a global maximum of $xf(x)$ relevant?). A better solution is below:

> If Alice sets price x per apple, and $f(x)$ apples are purchased, then her revenue for setting price x is $xf(x) := 3x^2 - x^3$. As Alice wishes to maximize her revenue, she should set the price maximizing $xf(x)$ on $[0, 3]$, which is $x = 2$. To see this, observe that the derivative of $xf(x)$ is positive on $[0, 2]$, and negative on $[2, 3]$.

As a general rule, it should be *easy* for a 445 grader to read your proof, understand what you are saying, and whether or not it is correct.

---

[21]Of course, the 445 graders will try to read anything you write on the favorable side of reasonable. But it certainly does not mean that sentences with multiple interpretations will always be given the most favorable one.

4.1.1 A Final Example

Consider the following problem, which I first heard about here: `https://gilkalai.wordpress.com/2017/09/08/elchanan-mossels-amazing-dice-paradox-answers-to-tyi-30/`. You roll a fair six-sided die until it lands six. What is the expected number of rolls you make (included the one which lands six), conditioned on all rolls being even? Let's view two conflicting "proofs." In both, we'll use without proof the following fact:

Fact 7. *Let D be a distribution such that when random variable X is drawn from D, the probability that $X = x$ is p. Then if we repeatedly sample draws from D independently until we see one which is equal to x, the expected number of draws we make is $1/p$.*

Proof: This is also a good example where the math is simpler if we use "form 2" of the definition of expectation. The probability that we make strictly more than i draws is the probability that all of the first i draws are not equal to x. Because they are drawn independently, and equal to x with probability p, this is just $(1-p)^i$. So we get that the expected number of draws we make is $\sum_{i=0}^{\infty}(1-p)^i = 1/p$. ∎

Proof 1:

> We know that if we were to roll the die until we hit a six, it would take six rolls in expectation, by Fact 7 (because we have a $1/6$ chance of rolling a six each time). If instead we condition on all rolls being even, now there are only three possibilities instead of six, so the probability of rolling a six each time is $1/3$ instead of $1/6$. So by the same Fact 7, the expected number of rolls until we hit a six, conditioned on all rolls being even, is now three.

Proof 2:

> Consider instead repeatedly rolling a die in the following manner, using the principle of deferred decisions. First, decide if the die will land on two/four, or not on two/four (then decide exactly the roll, uniformly at random among the remaining possibilities). Stop as soon as the die lands not on two/four. Then the probability of terminating any given round is $2/3$, and so by Fact 7, the expected number of rolls is $3/2$. Moreover, observe that we can decide whether or not to stop rolling independently of whether the last roll is a six or odd. Therefore, the expected number of rolls until we hit a six, conditioned on all rolls being even, is $3/2$.

Both proofs seem tempting: the logic in proof 1 is pretty straight-forward to follow. Proof 2 may be extra tempting because it uses a fancy term that was introduced earlier. Proof 2, it turns out, is correct (but you should not typically associate correctness with fancy terms), and Proof 1 is not. The first sentence of Proof 1 is correct. The second sentence of Proof 1 is vague or incorrect. In particular, when the proof says "now there are only three possibilities instead of six," it seems to suggest that conditioning on all rolls being even is the same as independently rolling each die, and enforcing that each draw is even. These are not the same (likely the original motivation for this problem was to point out this misconception).

Indeed, let's consider instead a million-sided die. The point is that we are *extremely* unlikely to have a long run where all rolls are even, so conditioning on all rolls being even makes the length of the runs quite short. In particular, we shouldn't expect to have all even throws followed by a six for a long run at all, and most of the time when this happens, it's because we got a six very quickly. If

we repeat the argument in Proof 1, it would imply that the expected number of throws, conditioned on all throws being even, is 500000. Proof 2 instead suggests that the expected number of throws until we hit a roll which is either odd or six is $1000000/500001 \approx 2$. Hopefully that gives some intuition. We can also do the full calculation to confirm:

The probability that we roll exactly i times until hitting a six, and that the first $i-1$ rolls were all even (i.e. two or four), is $(1/3)^{i-1}/6$. So the probability that we roll all evens until hitting a six is:

$$\sum_{i \geq 1} (1/3)^{i-1}/6 = 1/4.$$

Also, the total number of rolls, only counting those from sequences from which we rolled evens until hitting a six is:

$$\sum_{i \geq 1} i(1/3)^{i-1}/6 = 9/24.$$

The conditional expectation then just divides these two to get $\frac{9/24}{1/4} = 9/6 = 3/2$.

4.2 Effectively breaking down long proofs

Let's revisit the problem from Section 3, and see how to write a clear proof of a complex claim. Recall first the problem:

> Recall that a bipartite graph has two sets of nodes, L and R, with all edges having one endpoint in L and the other in R. Recall also that a perfect matching is a set of edges such that every node is in exactly one edge.
>
> Let G be a bipartite graph with n nodes on each side. Prove that if every node has degree $\geq n/2$, then G has a perfect matching.
>
> **Hint:** You may use Hall's Marriage Theorem. Recall that Hall's Marriage Theorem asserts that a bipartite graph has a perfect matching if and only if for every set $S \subseteq L$ of nodes on the left, we have $|N(S)| \geq |S|$, where $N(S)$ denotes the set of nodes with an edge to some node in S.

Here is a solution I would write. Afterwards, I'll explain what I think of as the key points.

Solution. We will prove that G has a perfect matching using Hall's Marriage Theorem, and show that for all sets $S \subseteq L$, $|N(S)| \geq |S|$.

Let us first consider sets S where $|S| \leq n/2$.

Lemma 7. *Consider any $S \subseteq L$, with $|S| \leq n/2$. Then, $|N(S)| \geq |S|$.*

Proof: Let v be any node in S. Observe that, immediately by the definition of G, the degree of v is at least $n/2$. This immediately implies that v has at least $n/2$ neighbors, and therefore S has at least $n/2$ neighbors. Because $|S| \leq n/2$, we have that $|N(S)| \geq n/2 \geq |S|$, as desired. ∎

Next, we consider the case where $|S| > n/2$.

Lemma 8. *Consider any $S \subseteq L$, with $|S| > n/2$. Then, $|N(S)| \geq |S|$.*

Proof: In fact, we will prove an even stronger claim, that $|N(S)| = n$. To do this, assume for contradiction that $|N(S)| < n$. This means that there must exist some node $u \in R$ such that $u \notin N(S)$. In particular, this means that u has no neighbors in S. However, there are $> n/2$ nodes in S, and therefore $< n/2$ nodes in $L \setminus S$. This means that the degree of u is strictly less than $n/2$ (because all of u's neighbors must lie in $L \setminus S$). This contradicts the definition of G, as every node has degree at least $n/2$. ∎

Now, we can wrap up the proof. Lemmas 7 and 8 together prove that $|N(S)| \geq |S|$ for all $S \subseteq L$. Hall's marriage theorem now implies that G has a perfect matching, as desired.

Thoughts on this writeup. Here is a great test to see if you've successfully broken down your complex proof into clear subparts: Ignore the proofs of Lemma 7 and Lemma 8. Now, the entire proof is just a few sentences. Assuming that Lemma 7, Lemma 8, and Hall's Marriage Theorem are all correct, *is it easy to follow the logical flow?* This is always a bit subjective, but I'd argue that the logic is quite clear, and is entirely captured in the final two sentences. This is the key difference between writing a 'long' proof and a 'short' proof. For a long proof, it's impossible for a reader to follow multiple logical trails at once, so your job is to break it down into manageable short proofs, and also provide a single short proof to bring it all together. Try to think of it like a tree: the root is the outline which connects Lemma 7, Lemma 8 and Hall's Marriage Theorem to prove the claim. It has three children: Lemma 7, Lemma 8, and Hall's Marriage Theorem. Lemma 7 has a standalone proof, because the logic is short and coherent. Lemma 8 has a standalone proof, because the logic is short and coherent. Hall's Marriage Theorem is given to you, and does not require a proof. Each node in the tree should provide a clear, logically coherent proof (i.e. the proof uses just a few ideas, and can fit in the grader's head all at once) of the desired claim, assuming that the claims made in its children are correct. Here are some other bulleted thoughts:

- Is it crucial that the proof separates out the key claims using the Lemma environment in La-TeX? No. But, if you're new to proof-writing, this is a good structure to enforce on yourself. I personally try to use this structure whenever I write my own proofs.

- Is it crucial that Lemmas 7 and 8 are broken down into two lemmas, instead of just one? No. But, the two cases clearly use different logic. So absolutely, they should at least be broken up into separate paragraphs.

- It *is* crucial that each subpart has a clear, concrete, and formal statement. For example, it's crucial that it's easy to see that Lemmas 7 and 8 together cover all S, and connect to Hall's Theorem. Informal statements like "All sets have sufficient neighbors" (what is "sufficient"?) or "small sets satisfy $|N(S)| \geq |S|$" (what is "small"?) fail this, because the grader can't figure out exactly what's proved in these steps.

- Staff solutions and lecture notes will give you further examples of how to break down large proofs into smaller chunks. When you read them, try to go through the exercise of reading just the definitions/lemmas/conclusion, and confirming that the logic follows. Then, when reading each individual lemma, you just need to confirm that this individual lemma is correct.

4.3 And Two Quick Tips

Tip One: I still suggest this even to PhD students, and follow this advice myself: **after you write something up, read it yourself** (perhaps even out-loud-in-your-head). If you can't follow your own logic, don't be surprised when the grader can't either! You will likely be surprised at how much better your own writing can get just by iterating this process.

Tip Two: Remember that it should be *easy* for a 445 grader to follow your solution. Even if your solution is complicated, and it will take a long time to process everything you've written, you should try to write a concise outline, and/or try to state what you're about to do before diving into a pool of calculations.

5 Playing around with Large Language Models

In Spring 2023, ChatGPT was first becoming popular and many courses were creating LLM policies. We played around with LLMs for possible problems in this course to see if/how we should adjust the course policy. Below are some transcripts of my attempts to use LLMs for the bipartite matching problem in Section 3.

5.1 Asking ChatGPT to explain the problem

Overall, ChatGPT's responses here were accurate, and I can see how they might be helpful! In this case, I don't think the initial responses added anything beyond the problem statement, but the latter responses were actually pretty good (in case I was confused about bipartite matchings versus graphs). Nothing it said was incorrect. **Please remember:** if you use ChatGPT for this purpose, we can't promise it will be correct, and you are ultimately responsible for understanding whatever the problem is asking.

Matt: Here is a question that I want to ask. Please note that it's written in LaTeX.[22]

As a first step, could you please help me understand what this problem is asking?

Recall that a bipartite graph has two sets of nodes, L and R, with all edges having one endpoint in L and the other in R. Recall also that a perfect matching is a set of edges such that every node is in exactly one edge.

Let G be a bipartite graph with n nodes on each side. Prove that if every node has degree $\geq n/2$, then G has a perfect matching.

Hint: You may use Hall's Marriage Theorem. Recall that Hall's Marriage Theorem asserts that a bipartite graph has a perfect matching if and only if for every set $S \subseteq L$ of nodes on the left, we have $|N(S)| \geq |S|$, where $N(S)$ denotes the set of nodes with an edge to some node in S.

ChatGPT: The problem is asking to prove that if a bipartite graph G has n nodes on each side and every node has a degree of at least $n/2$, then G has a perfect matching. This can be proven using Hall's Marriage Theorem, which states that a bipartite graph has a perfect matching if and only if, for every subset of nodes on one side, there are at least as many nodes on the other side connected to the subset.

Matt: I don't remember what a bipartite matching is. Could you please remind me?

ChatGPT: A bipartite matching in a graph is a set of edges where each node is connected to at most one other node, and nodes in one set are only connected to nodes in the other set. In other words, the graph is divided into two sets of nodes, and edges only connect nodes between the sets, not within the same set.

A perfect matching in a bipartite graph is a matching where each node in the graph is connected to exactly one other node. In other words, every node in the graph is part of exactly one edge in the matching.

Matt: Thank you! Can you remind me what is a bipartite graph, and how that is different from a bipartite matching?

ChatGPT: A bipartite graph is a type of graph in which the vertices can be divided into two disjoint sets, such that all edges only connect nodes between the sets and not within the same set. In other words, no edges exist within the same set of nodes.

A bipartite matching is a specific type of matching in a bipartite graph, where each node is connected to at most one other node, and nodes in one set are only connected to nodes in the other set. In a perfect bipartite matching, every node in the graph is part of exactly one edge.

So to summarize, a bipartite graph is a graph structure with two sets of nodes, and a bipartite matching is a specific type of matching in a bipartite graph where edges only connect nodes between the two sets.

5.2 Asking ChatGPT to touch up my writing

Overall, ChatGPT didn't ruin any of the logical flow. I personally like my original solution better, but ChatGPT did make it more concise. It also kept the LaTeX properly formatted. **Please remember:** if you use ChatGPT for this purpose, we can't promise it will maintain your logical flow, and you are ultimately responsible for whatever you submit.

Matt: Now, I'm going to send you text for a solution I wrote to this problem, in LaTeX.[23] I was wondering if you could please: a) touch up the text for me, so that it flows better and b) touch up any LaTeX so that it will be prettier in LaTeX. In particular, please send me back LaTeX that I can copy/paste and compile.

Here is my solution: We will prove that G has a perfect matching using Hall's Marriage Theorem, and show that for all sets $S \subseteq L$, $|N(S)| \geq |S|$.

Let us first consider sets S where $|S| \leq n/2$.

Lemma 9. *Consider any $S \subseteq L$, with $|S| \leq n/2$. Then, $|N(S)| \geq |S|$.*

Proof: Let v be any node in S. Observe that, immediately by the definition of G, the degree of v is at least $n/2$. This immediately implies that v has at least $n/2$ neighbors, and therefore S has at least $n/2$ neighbors. Because $|S| \leq n/2$, we have that $|N(S)| \geq n/2 \geq |S|$, as desired. ∎

Next, we consider the case where $|S| > n/2$.

Lemma 10. *Consider any $S \subseteq L$, with $|S| > n/2$. Then, $|N(S)| \geq |S|$.*

Proof: In fact, we will prove an even stronger claim, that $|N(S)| = n$. To do this, assume for contradiction that $|N(S)| < n$. This means that there must exist some node $u \in R$ such that $u \notin N(S)$. In particular, this means that u has no neighbors in S. However, there are $> n/2$ nodes in S, and therefore $< n/2$ nodes in $L \setminus S$. This means that the degree of u is strictly less than $n/2$ (because all of u's neighbors must lie in $L \setminus S$). This contradicts the definition of G, as every node has degree at least $n/2$. ∎

---

[23]I copy/pasted the LaTeX, which LaTeX is properly formatting below — I sent the raw LaTeX to ChatGPT.

Now, we can wrap up the proof. Lemmas 9 and 10 together prove that $|N(S)| \geq |S|$ for all $S \subseteq L$. Hall's marriage theorem now implies that G has a perfect matching, as desired.

ChatGPT :

Proof: We will prove that G has a perfect matching using Hall's Marriage Theorem, and show that for all sets $S \subseteq L$, $|N(S)| \geq |S|$.

Lemma 11. *Consider any $S \subseteq L$, with $|S| \leq n/2$. Then, $|N(S)| \geq |S|$.*

Proof: Let v be any node in S. By the definition of G, the degree of v is at least $n/2$, implying that v has at least $n/2$ neighbors. Therefore, S has at least $n/2$ neighbors. Since $|S| \leq n/2$, we have $|N(S)| \geq n/2 \geq |S|$. ∎

Lemma 12. *Consider any $S \subseteq L$, with $|S| > n/2$. Then, $|N(S)| \geq |S|$.*

Proof: We will prove an even stronger claim, that $|N(S)| = n$. To do this, assume for contradiction that $|N(S)| < n$. This means that there must exist some node $u \in R$ such that $u \notin N(S)$. This implies that u has no neighbors in S. However, there are more than $n/2$ nodes in S, and therefore fewer than $n/2$ nodes in $L \setminus S$. This means that the degree of u is strictly less than $n/2$ (because all of u's neighbors must lie in $L \setminus S$), contradicting the definition of G as every node has degree at least $n/2$. ∎

Lemmas 11 and 12 together prove that $|N(S)| \geq |S|$ for all $S \subseteq L$. Hall's Marriage Theorem now implies that G has a perfect matching. ∎

5.3 Asking ChatGPT to turn text into LaTeX

Overall, ChatGPT made the raw text I gave it into reasonably formatted LaTeX (it's not quite as nicely-formatted as how I would do myself, but it's like 98% of the way there). It also did not alter the meaning of any text I gave it, and maintained a correct solution. **Please remember:** if you use ChatGPT for this purpose, we can't promise it will produce LaTeX that matches the text you input, and you are ultimately responsible for whatever you submit.

Matt: We will prove that G has a perfect matching using Hall's Marriage Theorem, and show that for all sets S subseteq L, —N(S)— geq —S—.

Let us first consider sets S where —S— leq n/2.

Lemma: Consider any S subseteq L, with —S— leq n/2. Then, —N(S)— geq —S—.

Proof: Let v be any node in S. Observe that, immediately by the definition of G, the degree of v is at least n/2. This immediately implies that v has at least n/2 neighbors, and therefore S has at least n/2 neighbors. Because —S— leq n/2, we have that —N(S)— geq n/2 geq —S—, as desired.

Next, we consider the case where —S— ¿ n/2.

Lemma: Consider any S subseteq L, with —S— ¿ n/2. Then, —N(S)— geq —S—.

Proof: In fact, we will prove an even stronger claim, that —N(S)— = n. To do this, assume for contradiction that —N(S)— ¡ n. This means that there must exist some node u in R such that u notin N(S). In particular, this means that u has no neighbors in S. However, there are ¿ n/2 nodes in S, and therefore ¡ n/2 nodes in L setminus S. This means that the degree of u is

strictly less than n/2 (because all of u's neighbors must lie in L setminus S). This contradicts the definition of G, as every node has degree at least n/2.

Now, we can wrap up the proof. The two lemmas together prove that —N(S)— geq —S— for all S subseteq L. Hall's marriage theorem now implies that G has a perfect matching, as desired.

ChatGPT: **Proof:** We will prove that G has a perfect matching using Hall's Marriage Theorem, and show that for all sets $S \subseteq L$, $|N(S)| \geq |S|$.

Let us first consider sets S where $|S| \leq n/2$.

Lemma: Consider any $S \subseteq L$, with $|S| \leq n/2$. Then, $|N(S)| \geq |S|$.

Proof: Let v be any node in S. Observe that, immediately by the definition of G, the degree of v is at least $n/2$. This immediately implies that v has at least $n/2$ neighbors, and therefore S has at least $n/2$ neighbors. Because $|S| \leq n/2$, we have that $|N(S)| \geq n/2 \geq |S|$, as desired.

Next, we consider the case where $|S| > n/2$.

Lemma: Consider any $S \subseteq L$, with $|S| > n/2$. Then, $|N(S)| \geq |S|$.

Proof: In fact, we will prove an even stronger claim, that $|N(S)| = n$. To do this, assume for contradiction that $|N(S)| < n$. This means that there must exist some node u in R such that $u \notin N(S)$. In particular, this means that u has no neighbors in S. However, there are $> n/2$ nodes in S, and therefore $< n/2$ nodes in $L \setminus S$. This means that the degree of u is strictly less than $n/2$ (because all of u's neighbors must lie in $L \setminus S$). This contradicts the definition of G, as every node has degree at least $n/2$.

Now, we can wrap up the proof. The two lemmas together prove that $|N(S)| \geq |S|$ for all $S \subseteq L$. Hall's marriage theorem now implies that G has a perfect matching, as desired. ∎

5.4 Asking ChatGPT to solve the problem

The chat below is representative of how ChatGPT seems to interact with math. You'll notice that the language structure pattern matches a correct proof shockingly well, but the content is just unfortunately mostly nonsense. I copied the chat first below, and then annotated ChatGPT's response. This is from state-of-the-art LLMs in 2023.

Matt: Hi! Can you write a solution to the following problem? Note that it's written in LaTeX. It would be great if you can also write your solution in LaTeX.

Recall that a bipartite graph has two sets of nodes, L and R, with all edges having one endpoint in L and the other in R. Recall also that a perfect matching is a set of edges such that every node is in exactly one edge.

Let G be a bipartite graph with n nodes on each side. Prove that if every node has degree $\geq n/2$, then G has a perfect matching.

Hint: You may use Hall's Marriage Theorem. Recall that Hall's Marriage Theorem asserts that a bipartite graph has a perfect matching if and only if for every set $S \subseteq L$ of nodes on the left, we have $|N(S)| \geq |S|$, where $N(S)$ denotes the set of nodes with an edge to some node in S.

33

ChatGPT: Here's the solution to the problem:

Let $G = (L, R, E)$ be a bipartite graph with n nodes on each side, where every node has degree $\geq n/2$. To prove that G has a perfect matching, we will use Hall's Marriage Theorem.

Suppose for contradiction that there exists a set $S \subseteq L$ of nodes on the left such that $|N(S)| < |S|$. Then, $|S| > n/2$, since each node in S has degree $\geq n/2$. But this contradicts the assumption that each node has degree $\geq n/2$, since $|S| > n/2$ and $|N(S)| < |S|$.

Therefore, for every set $S \subseteq L$ of nodes on the left, we have $|N(S)| \geq |S|$. By Hall's Marriage Theorem, G has a perfect matching.

Solution in LaTeX:

Proof: Let $G = (L, R, E)$ be a bipartite graph with n nodes on each side, where every node has degree $\geq n/2$. To prove that G has a perfect matching, we will use Hall's Marriage Theorem.

Suppose for contradiction that there exists a set $S \subseteq L$ of nodes on the left such that $|N(S)| < |S|$. Then, $|S| > n/2$, since each node in S has degree $\geq n/2$. But this contradicts the assumption that each node has degree $\geq n/2$, since $|S| > n/2$ and $|N(S)| < |S|$.

Therefore, for every set $S \subseteq L$ of nodes on the left, we have $|N(S)| \geq |S|$. By Hall's Marriage Theorem, G has a perfect matching. ∎

Annotating ChatGPT's response. Here's the solution to the problem:

Let $G = (L, R, E)$ be a bipartite graph with n nodes on each side, where every node has degree $\geq n/2$. To prove that G has a perfect matching, we will use Hall's Marriage Theorem.

> Off to a great start! This first bit not only pattern matches the language structure, but also has reasonable content. But, this is basically just stating "I will now solve the problem using the hint."

Suppose for contradiction that there exists a set $S \subseteq L$ of nodes on the left such that $|N(S)| < |S|$. Then, $|S| > n/2$, since each node in S has degree $\geq n/2$. But this contradicts the assumption that each node has degree $\geq n/2$, since $|S| > n/2$ and $|N(S)| < |S|$.

> The first sentence is a reasonable way to start a proof, go for contradiction! The second sentence is a true statement, and the logic is actually sound (I think this is also a reasonable level of detail to give). But then the third sentence makes absolutely no sense. How does $|S| > n/2$ contradict the assumption that each node has degree $\geq n/2$? In particular, how does $|S| > n/2$ and $|N(S)| < |S|$ contradict that each node has degree $n/2$? Note, however, that the language patterns absolutely match those of a correct proof.

Therefore, for every set $S \subseteq L$ of nodes on the left, we have $|N(S)| \geq |S|$. By Hall's Marriage Theorem, G has a perfect matching.

> The wrapup at the end is OK, but it is basically just restating Hall's Marriage Theorem. If I were to evaluate this on the generic rubric, I would give this 2/16 for quality of progress. The solution demonstrate a good understanding of Hall's Marriage Theorem, and the solution also demonstrates an understanding that if one node has $> n/2$ neighbors, then any set S also has $> n/2$ neighbors. But, there is no actual concrete

progress here. If I were to evaluate the quality of presentation, I would give 1/4. The solution is so short that I can't really claim it was all that tough to evaluate. The first and last paragraphs were accurate and well-written, but ultimately had no content. The middle paragraph is the only one with content. Here, the third sentence is very flawed, but it actually took me a while to convince myself that it was indeed flawed, and that it wasn't just a more concise version of ultimately correct logic. Because I'm not sure whether this is written by some math genius who just thinks everything is obvious (and I'm just the fool who's not sharp enough to see why), or a chatbot with no mathematical ability whatsoever, I have to really sit down and dissect the logic before ultimately determining that the sentence is actually very flawed.

If the middle paragraph were omitted, and the solution clearly stated that it did not fully solve the problem, I would have given the solution 2/16 progress (for demonstrating a good understanding of Hall's marriage theorem), and 4/4 presentation (it was otherwise very easy to understand what was given, but the problematic middle paragraph really made me sit down and think before declaring the solution incorrect).

Solution in LaTeX:

I don't understand what was not "in LaTeX" about the solution written above, but it seems as though ChatGPT has just rewritten the above solution inside a proof environment. There's nothing particulary good/bad about that, but just an interesting observation of how it parsed my request.

Proof: Let $G = (L, R, E)$ be a bipartite graph with n nodes on each side, where every node has degree $\geq n/2$. To prove that G has a perfect matching, we will use Hall's Marriage Theorem.

Suppose for contradiction that there exists a set $S \subseteq L$ of nodes on the left such that $|N(S)| < |S|$. Then, $|S| > n/2$, since each node in S has degree $\geq n/2$. But this contradicts the assumption that each node has degree $\geq n/2$, since $|S| > n/2$ and $|N(S)| < |S|$.

Therefore, for every set $S \subseteq L$ of nodes on the left, we have $|N(S)| \geq |S|$. By Hall's Marriage Theorem, G has a perfect matching. ∎

5.5 Asking ChatGPT to solve a different problem

I also tried asking ChatGPT to solve a slightly longer problem (Problem 4 on PS0). This one went horribly (in fact, it went so horribly, that I'm comfortable sharing ChatGPT's solution with you while still asking you Problem 4 on PS0, because it contributes absolutely no helpful content).

If I were evaluating this solution, I'd absolutely have given a 0/4 on presentation because of how challenging it was to identify the logical flaws in some places, and because of the extreme level of nonsense in others. Unfortunately, I also would've given it a 0/16 on progress. It definitely has no concrete progress I can point to. And unfortunately, because there are so many mistakes, it's hard for me to agree that the solution 'demonstrates a good understanding' of anything.

This is from state of the art LLMs in 2023. 2024 LLMs did much better at solving this problem, which initially alarmed me. However, I was told that this is not surprising because it's a well-known problem with many solutions on the Internet, and so it's success is not likely to replicate across PSets (indeed, this was the case).

Matt: Could you solve the following problem for me? It's written in LaTeX. It would be great if your solution was written in LaTeX.

You're trying to collect all n distinct cards from your favorite trading card game. The only way to obtain new cards is to purchase a sealed pack of one uniformly random card. After you buy a pack, you open it, see the card inside, and add it to your collection. If you have at least one copy of each of the n cards, you stop. Otherwise, you purchase a new pack. Prove that the expected number of packs you purchase is $\Theta(n \log n)$.[24][25][26]

ChatGPT: Let E_n be the expected number of packs you need to purchase to complete your collection of n distinct cards. We can start with an upper bound on E_n.

> This is off to a sensible start! It's defining a variable for what it eventually hopes to bound, and does one side of the bound first.

Let X_i be the number of packs you need to purchase in order to obtain the ith distinct card. Note that $E[X_i] = 1/(n - i + 1)$ because, at this point, you still need to obtain $n - i + 1$ distinct cards, and each pack has a probability of $1/(n - i + 1)$ of being the ith distinct card. Since the X_i are all independent, we have

$E_n = \sum_{i=1}^{n} E[X_i] = \sum_{i=1}^{n} \frac{1}{n-i+1} \leq \sum_{i=1}^{n} \frac{1}{i} \leq \ln(n + 1) + 1$

> This is where things start unraveling, and it gets worse with each line. Let's go one at a time:
>
> – The first sentence is sloppy: it defines X_i to be the number of packs needed to obtain the ith distinct card. But what does this mean? Does it mean what it literally says: the number of packs needed *from the start* to obtain the ith distinct card? Or, does it mean the number of packs needed *after finding the $(i-1)^{th}$ distinct card* to obtain the ith distinct card? These two definitions are very different! It's important for the proof to clearly state what the definition is. That being said, it's *possible* this gets cleared up by context in subsequent sentences, if the subsequent sentences are particularly clear. Personally, if I were evaluating a solution with a sentence like this, I wouldn't take off any points *if the confusion got cleared up by subsequent (or preceding) context*. But if it was hard for me to figure out what definition was actually being used here (and especially if it felt like one definition was used in one place, and another in a second), I would hit this pretty hard on quality of presentation — a hugely important aspect of your proof is that the grader at minimum knows what you're trying to say!
>
> – The second sentence is unfortunately complete nonsense! Neither potential definition of X_i has $E[X_i] = 1/(n - i + 1)$. First, $1/(n - i + 1)$ is certainly < 1, and X_i is always an integer ≥ 1, so this can't possibly both be true. Second, it's true that "you still need to obtain $n - i + 1$ distinct cards", but this does not imply the preceding claim (claiming to compute $E[X_i]$) nor the subsequent claim. Finally, each pack does have *some* probability of being the ith distinct card, but this is not $1/(n - i + 1)$. So, this whole sentence is just a mishmosh of wrong: the individual claims are incorrect, and also the

---

[24] Recall that $f(n) = \Theta(g(n))$ if $f(n) = O(g(n))$ and $f(n) = \Omega(g(n))$. Recall further that $f(n) = O(g(n))$ if there exist absolute constants C, n_0 such that $f(n) \leq C \cdot g(n)$ for all $n \geq n_0$. $f(n) = \Omega(g(n))$ if there exist absolute constants C, n_0 such that $f(n) \geq c \cdot g(n)$ for all $n \geq n_0$.

[25] You may use without proof the fact that $\ln(n + 1) + 1 > \sum_{i=1}^{n} 1/i > \ln(n + 1)$.

[26] You may also use the following fact without proof: if a coin is heads with probability p (independently on each flip), then the expected number of flips until seeing a heads is $1/p$.

three claims don't tie together correctly. This is unfortunately really not good. If I were grading this assignment, I would hit this really hard for quality of presentation, because I had to read through this sentence ten times to convince myself it was actually wrong (if I'm being completely honest, the first time I read it I was like "oh wow, it actually did pretty good!" because the sentence structure is so smooth. But it gets worse after each successive read).

- The third sentence (including the math) is interesting. Setting aside the mistakes of the previous sentence, it's not necessarily wrong? But it's still odd in a few places. The good news is that only one definition of X_i from the confusion earlier satisfies that X_is are independent, and also this is the definition that satisfies $E_n = \sum_{i=1}^{n} E[X_i]$. So that's good, and reinforces that probably the solution intended to define X_i in a consistent way. Given that I had to keep track of two possible definitions for three sentences until it got cleared up, I would have dinged this one mark on the clarity of presentation scale. The other good news is that the chain of math equations are all correct. But aside from this, things are a bit odd. First, none of the math in the chain of equations uses the fact that the X_is are independent, so this is superfluous. Second, the third inequality could actually just be an equality. Finally, the last inequality seems to state that $E_n \leq \ln(n + 1) + 1$. This would unfortunately contradict that $E_n = \Theta(n \log n)$, which is what the problem asks to prove. Something about this should be a major red flag? But the solution seems to just ignore it and press on. This sentence has no false math, but its conclusion is a major red flag (and also it has two superfluous distracting parts).

- To recap: this paragraph contributes no concrete progress towards the solution. Also, it takes effort to even tell that this is nonsense! To recap the places where I'd take off points:
  * The first sentence gives an ambiguous definition that's cleared up five logical steps later. This is annoying to keep track of both definitions for five logical steps, but not a dealbreaker. So this would make a minor contribution to quality of presentation.
  * There are two superfluous claims (using independence when it's not necessary, $\leq$ instead of $=$). This is a little annoying when reading (because it's one more thing I have to think about — when the reason I think a claim is true is not the reason the author claims, I need to spend more effort confirming the claim is still true), but two small superfluous claims would not cause me to take off presentation points.
  * The mess of nonsense in the second sentence is really bad! I had to reread this ten times to convince myself that the logic is truly flawed, and I'm not the one making a mistake. This would probably tank the presentation all the way to 0 (this doesn't mean that there aren't worse presentations out there, but this is really consuming an abnormal amount of energy to evaluate, compared to a typical solution submitted by students).

Next, we find a lower bound on E_n.

Great! This would complete both bounds.

We consider the event that, after purchasing the ith distinct card, we have not yet completed

our collection. This occurs if and only if we have not yet obtained each of the n distinct cards. In this case, we have E_{n-i} expected purchases remaining. Hence, we have

$E_n \geq \sum_{i=1}^{n} \frac{1}{n-i+1} + E_{n-i} = \sum_{i=1}^{n} \frac{1}{n-i+1} + \sum_{j=1}^{n-i} \frac{1}{j} \geq \ln n + 1$

Let's begin with the first two sentences. These aren't actually incorrect. But, they're just kind of useless? If we've only just now purchased the ith distinct card, haven't we by definition not yet completed our collection? So, I'm not sure what's going on here. I wouldn't take off points for writing these sentences, because it's easy to see that they don't contribute everything. But it's not a good sign of where the logic might later go.

The third sentence has a lot going on. First, my best interpretation of E_{n-i} is "the expected number of packs purchased when there are $n - i$ total packs in order to get all $n - i$ cards." This would be the most natural interpretation, as this just substitutes $n - i$ for n in the definition of E_n. For this definition, the claim is just false! At no point during the process are we magically going to have only $n - i$ total packs to choose from. The solution seems to now be using E_{n-i} to refer to "the expected number of packs purchased when there are n total packs in order to get all cards in a specific set of $n - i$ cards." For this definition, the claim that there are E_{n-i} purchases remaining is correct. But now the problem is that the solution has done absolutely no math related to this definition of E_{n-i}, so whatever follows is completely unjustified (and also, it's a huge stretch to expect the reader to think this is the intended definition). At this point, I'd dock 1-2 presentation marks just for the confusing definition alone, and possibly more if the later logic is also incorrect.

Thus, we have $E_n = \Theta(n \log n)$.

The confidence with which the last line is stated is audacious!! I hope that, even if it's been a while since you took 226/240 and you may have forgotten some big-oh notation, a huge red flag jumps out at you that a number $\geq \ln(n) + 1$ and $\leq \ln(n+1) + 1$ is *absolutely not* $\Theta(n \log n)$ (it is $\Theta(\log n)$). I would hit this with maybe two lower marks in quality of presentation. The good news is that it's easy to see that something is clearly wrong. The bad news is that because it's stated with such confidence and no justification, it's very hard for me to see what the author's intent might have been instead, so I have to go back and reread the proof to see if I'm missing something.