445 Math Cheatsheet

Below is a quick refresher on some math tools from 340 that we'll assume knowledge of for the PSets.

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:

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

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:

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

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

$$\begin{aligned}\sum_{i=0}^{\infty} \Pr[X > i] &= \sum_{i=0}^{\infty} \sum_{j>i} \Pr[X = j]. \\ &= \sum_{j=0}^{\infty} \sum_{i<j} \Pr[X = j] \\ &= \sum_{j=0}^{\infty} j \cdot \Pr[X = j].\end{aligned}$$

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.

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 probably 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 $f(x)dx$.
- The **CDF** of a random variable is simpler to define, and just denotes $F(x) = \Pr[X \leq x]$. 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: $\mathbb{E}[X] = \int_0^\infty xf(x)dx.$

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 "$f(x)dx \approx \Pr[X = x]$." The equivalent definition for form 2 is also often easier to use in calculations:

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

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.

1.3 Examples

Consider the uniform distribution on the set $\{4, 5\}$ (4 w.p. 1/2, 5 w.p. 1/2). Then the expected value as computed by form 1 is:

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

The expected value as computed by form 2 is:

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

Now consider the uniform distribution on the interval $[4, 5]$ (equally likely to be any real number in $[4, 5]$). Then the PDF associated with this distribution is $f(x) = 1, x \in [4, 5]$, $f(x) = 0, x \notin [4, 5]$. And we can compute the expected value by form 1 as:

$$\int_0^\infty xf(x)dx = \int_4^5 xdx = x^2/2|_4^5 = 25/2 - 8 = 4.5.$$

We can also compute it using form 2 as:

$$\int_0^\infty (1 - F(x))dx = \int_0^4 1dx + \int_4^5 (x-4)dx + \int_5^\infty 0dx = 4 + (x^2/2 - 4x|_4^5) + 0 = 4.5.$$

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 $f(x) = 0$.

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:

$$\begin{aligned}
\mathbb{E}[X_1 + X_2] &= \sum_{i=0}^{\infty} \Pr[X_1 + X_2 = i] \cdot i \\
&= \sum_{i=0}^{\infty} \sum_{j=0}^{i} \Pr[X_1 = j] \cdot \Pr[X_2 = i - j] \cdot i \\
&= \sum_{j=0}^{\infty} \sum_{i=j}^{\infty} \Pr[X_1 = j] \cdot \Pr[X_2 = i - j] \cdot i \\
&= \sum_{j=0}^{\infty} \Pr[X_1 = j] \cdot \sum_{\ell=0}^{\infty} \Pr[X_2 = \ell] \cdot (\ell + j) && \text{(changing variables with } \ell = i - j) \\
&= \sum_{j=0}^{\infty} \Pr[X_1 = j] \cdot \left(j + \sum_{\ell=0}^{\infty} \Pr[X_2 = \ell] \cdot \ell \right) \\
&= \sum_{j=0}^{\infty} \Pr[X_1 = j] \cdot (j + \mathbb{E}[X_2]) \\
&= \mathbb{E}[X_1] + \mathbb{E}[X_2]. && \text{(because } \sum_{j=0}^{\infty} \Pr[X_1 = j] = 1)
\end{aligned}$$

1.5 Bayes' Rule

Let's first recap the definition of conditional probability: the probability of an event X conditioned on another event Y, denoted by $\Pr[X|Y]$ is equal to the probability of X and Y divided by the probability of Y (that is, $\Pr[X \wedge Y]/\Pr[Y]$. Think of this as the probability that X occurs, given that Y has occured. For a concrete example, consider the probability that a fair six-sided dice lands two (X), conditioned on it landing even (Y). Then $\Pr[X \wedge Y]$ is the probability that the die lands two *and* that it is even (which is just the probability that it is two), so $1/6$. $\Pr[Y]$ is just the probability that hte die is even, which is $1/2$, so the ratio is $1/3$. The probability that the roll is prime (X), conditioned on being even (Y) 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 $\Pr[X \wedge Y]$ might be challenging, but computing $\Pr[Y|X]$ is not so bad. Bayes' rule simply manipulates the above equalities to write:

$$\Pr[X|Y] = \Pr[X \wedge Y]/\Pr[Y] = \Pr[Y|X] \cdot \Pr[X]/\Pr[Y].$$

For example, say that you have a coin whose probability of outputting heads is p, and p is either $1/4$ or $3/4$ with equal probability. Say now that you flip the coin once and it lands heads. What is the probability that the coin's bias was $3/4$ (call this event X), conditioned on one flip landing heads (call this event Y)? This is conceptually tricky to reason about, but Bayes' rule suggests that we compute $\Pr[Y|X], \Pr[X]$, and $\Pr[Y]$. $\Pr[Y|X]$ is the probability that the coin lands heads, conditioned on the bias being $3/4$. This is just $3/4$. $\Pr[X]$ is the probability that the bias is $3/4$, which is just $1/2$. $\Pr[Y]$ is the probability that the coin lands heads (without having seen any flips), which is also just $1/2$. So $\Pr[X|Y] = (3/4) \cdot (1/2)/(1/2) = 3/4$.

2 Basic continuous optimization

2.1 Single-variable, unconstrained optimization

Say we want to find the global maximum of a continuous, differentiable function $f(\cdot)$. Any value that is a global maximum must also be a critical point, point where $f'(x) = 0$. Not all critical points are local optima, and not all local optima are local maxima, but all local maxima are critical points. One also needs to confirm that $f(\cdot)$ indeed achieves its global maximum by examining $\lim_{x\to\pm\infty} f(x)$.

For example, say we want to find the global maximum of $f(x) = x^2$. There is a unique critical point at $x = 0$. So if the function attains its 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.

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 if there is a global maximum, it must be $x = 2$. We can verify that $\lim_{x\to\pm\infty} = -\infty$, so $x = 2$ must be the global maximum.[1]

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 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 $x \in [a, b]$.

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

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).

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.

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 rule out 0. We also see that $f'(x) \leq 0$ on $[2, 3]$, so we can immediately rule out 3, and we only need to compare -2 and 2. We can also immediately see that $f(-x) > f(x)$ for all $x > 0$, and therefore $f(-2)$ is the global constrained maximum.

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. 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.

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. One also needs to confirm that $f(\cdot)$ indeed achieves its global maximum by examining limits towards ∞. Doing this formally can sometimes be tedious, but in this class we'll only see cases where this is straight-forward.[2] Sometimes, it might also be helpful to think of some variables as being fixed, and solve successive single-variable optimization problems. Here are some examples that you might reasonably need to solve:

Say you want to maximize $f(x_1, x_2) = x_1 - x_1^2 - x_2^2$. We can immediately see that for any x_1, $f(x_1, x_2)$ is maximized at $x_2 = 0$ (this is what we mean by thinking of x_1 as fixed and solving a single-variable optimization problem for x_2). Once we've set $x_2 = 0$, we now just want to maximize $x_1 - x_1^2$, which is achieved at $x_1 = 1/2$. So the unconstrained maximizer is $(1/2, 0)$.

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 see that $\frac{\partial f}{\partial x_2} = x_1 - 2x_2$, and so for fixed x_1, the unique maximizer is at $x_2 = x_1/2$. We can then just optimize $x_1(x_1/2) - x_1^2 - (x_1/2)^2 = (-3/4) \cdot x_1^2$, which is clearly maximized at $x_1 = 0$. So the unique global maximizer is $(0, 0)$.

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

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

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 $f(\vec{x})$. Most (possibly all) of the instances you will need to solve in the PSets will be of this format.

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. Most (possibly all) of the instances you will need to solve in the PSet will be solvable by finding an unconstrained maximizer of f, $\vec{x}^*$, and observing that $\vec{x}^*$ satisfies the constraints.

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 find the unconstrained maximizer by observing that $\frac{\partial f}{\partial x_i} = e^{-x_i} - x_i e^{-x_i}$, which is positive when $x_i < 1$, and negative when $x_i > 1$. So the unique maximizer is at $x_i = 1$. So $(1, \ldots, 1)$ is the unique 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.

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

3 Basic Proof Writing

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 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 interpretation 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.[3]

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.

[3] 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.

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 $f'(x) \leq 0$ on $[0, 2]$, and $f'(x) \geq 0$ 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 maximum 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.