TAKEAWAYS FROM UNDERGRADUATE MATH CLASSES

STEVEN J. MILLER

ABSTRACT. Below we summarize some items to take away from various undergraduate classes. In particular, what are one time tricks and methods, and what are general techniques to solve a variety of problems, as well as what have we used from various classes. The goal is to provide a brief summary of what parts of subjects are used where. Comments and additions welcome!

CONTENTS

1. CALCULUS I AND II (MATH 103 AND 104)

We use a variety of results and techniques from 103 and 104 in higher level classes:

Date: December 3, 2012.

(1) **Standard integration theory:** One of the most important technique is integration by parts; one of many places it is used is in computing the moments of the Gaussian in probability theory. Integration by parts is a very powerful technique, and is frequently used. While most of the time it is clear how to choose the functions u and dv, sometimes we need to be a bit clever. For example, consider the second moment of the standard normal (if you don't know what this is, no worries; just treat this as an integral you want to evaluate): $(2\pi)^{-1/2}\int_{-\infty}^{\infty} x^2\exp(-x^2/2)dx$. The natural choices are to take $u = x^2$ or $u = \exp(-x^2/2)$, but neither of these work as they lead to choices for dv that do not have a closed form integral. What we need to do is split the two 'natural' functions up, and let $u = x$ and $dv = \exp(-x^2/2)xdx$. The reason is that while there is no closed form expression for the anti-derivative of the standard normal, once we have xdx instead of dx then we can obtain nice integrals. One final remark on integrating by parts: it is a key ingredient in the 'Bring it over' method (which will be discussed below).

(2) **Definition of the derivative:** Recall

$$f'(x) \ = \ \lim_{h\to 0}\frac{f(x+h)-f(x)}{h}.$$

In upper level classes, the definition of the derivative is particularly useful when there is a split in the definition of a function. For example, consider

$$f(x) \ = \ \begin{cases}\exp(-1/x^2) & \text{if } x\neq 0\\ 0 & \text{if } x = 0.\end{cases}$$

This function has all derivatives zero at $x = 0$, but is non-zero for $x \neq 0$. Thus the Taylor series (see below) does not converge in a neighborhood of positive length containing the origin. This function shows how different real analysis is from complex analysis. Explicitly, here we have an infinitely differentiable function which is not equal to its Taylor series in a neighborhood of $x = 0$; if a complex function is differentiable once it is infinitely differentiable and it equals its derivative in a neighborhood of that point.

The proofs of all the standard differentiation lemmas (for a sum, for a difference, for a product, for a quotient, for chained variables, ...) all start with the definition of the derivative, applied to an appropriate function. For a product it is $A(x) = f(x)g(x)$, for a sum it is $A(x) = f(x) + g(x)$. In practice we don't want to go back to the definition every time we need a derivative; the point is to isolate out common occurrences / expressions but with *general* inputs; then we just plug in the values specific to our problem. Thus, no one evern creates a pre-computed list of derivatives involving $x^{127043252525213523} - 345353534x^{43535}$, but we can quickly get this from our rules. This idea is used in many higher courses: go back to the definition and choose appropriate values, and then isolate out results that will be of great use again and again. One of my favorite examples of this are identities for Moment Generating Functions of combinations of random variables in probability.

(3) **Taylor series:** Taylor expansions are very useful, allowing us to replace complicated functions (locally) by simpler ones. The moment generating function of a random variable is a Taylor series whose coefficients are the moments of the distribution. Another instance is in proving the Central Limit Theorem from probability. **Taylor's Theorem:** *If f is differentiable at least $n+1$ times on $[a,b]$, then for all $x \in [a,b]$, $f(x) = \sum_{k=0}^{n}\frac{f^{(k)}(a)}{k!}(x-a)^k$ plus*

an error that is at most $\max_{a \le c \le x} |f^{(n+1)}(c)| \cdot |x-a|^{n+1}$.

(4) **L'Hopital's Rule:** This is one of the most useful ways to compare growth rates of different functions. It works for ratios of differentiable functions such that either both tend to zero or both tend to $\pm\infty$. We used this in class to see that, as $x \to \infty$, $(\log x)^A \ll x^B \le e^x$ for any $A, B > 0$. (Recall $f(x) \ll g(x)$ means there is some C such that for all x sufficiently large, $|f(x)| \le Cg(x)$.) We also used L'Hopital to take the derivatives of the troublesome function $h(x) = \exp(-1/x^2)$ for $x \neq 0$ and 0 otherwise (this function is the key to why real analysis is so much harder than complex analysis). We can also use L'Hopital's Rule to determine whether or not certain sequences converge.

2. MULTIVARIABLE CALCULUS (MATH 105/106)

(1) **Dot product, Cross product:** If $\overrightarrow{v} = (v_1, \dots, v_n)$ and $\overrightarrow{w} = (w_1, \dots, w_n)$ then the dot product is $\overrightarrow{v} \cdot \overrightarrow{w} = v_1w_1 + \cdots + v_nw_n$, and the angle θ between the two vectors is given by $\frac{\overrightarrow{v}\cdot\overrightarrow{w}}{||\overrightarrow{v}||\,||\overrightarrow{w}||}$, where $||\overrightarrow{v}||$ is the length of $\overrightarrow{v}$:

$$||\overrightarrow{v}|| \;=\; (v_1^2 + v_2^2 + \cdots + v_n^2)^{1/2}.$$

If $n = 3$, then the cross product is defined by

$$\begin{vmatrix} \overrightarrow{i} & \overrightarrow{j} & \overrightarrow{k} \\ v_1 & v_2 & v_3 \\ w_1 & w_2 & w_3 \end{vmatrix} \;=\; (v_2w_3 - v_3w_2, v_3w_1 - v_1w_3, v_1w_2 - v_2w_1).$$

The cross product gives the area of the parallelogram generated by $\overrightarrow{v}$ and $\overrightarrow{w}$.

(2) **Definition of the Derivative: One Variable:** Let $f : \mathbb{R} \to \mathbb{R}$ be a function. We say f is differentiable at x_0, and denote this by $f'(x_0)$ or df/dx, if the following limit exists:

$$f'(x_0) \;=\; \lim_{h\to 0} \frac{f(x_0+h) - f(x_0)}{h}.$$

We may also write this limit by

$$f'(x_0) \;=\; \lim_{x\to x_0} \frac{f(x_0+h) - f(x_0)}{h},$$

or as

$$\lim_{x\to x_0} \frac{f(x_0+h) - f(x_0) - f'(x_0)h}{h} \;=\; 0.$$

(3) **Definition of the Derivative: Several Variables, One Output:** Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function of n variables $x_1, \dots, x_n$. We say the partial derivative with respect to x_i exists at the point $a = (a_1, \dots, a_n)$ if

$$\lim_{h\to 0} \frac{f(\overrightarrow{a} + h\overrightarrow{e}_i) - f(\overrightarrow{a})}{h}$$

exists, where

$$\overrightarrow{a} + h\overrightarrow{e}_i \;=\; (a_1, \dots, a_{i-1}, a_i + h, a_{i+1}, \dots, a_n).$$

Let $f : \mathbb{R}^2 \to \mathbb{R}$. The tangent plane approximation to f at (x_0, y_0) is given by

$$z \;=\; f(x_0, y_0) + \frac{\partial f}{\partial x}(x_0, y_0)(x - x_0) + \frac{\partial f}{\partial y}(x_0, y_0)(y - y_0),$$

provided of course the two partial derivatives exist (and this naturally generalizes to more variables).

Finally, let $f : \mathbb{R}^2 \to \mathbb{R}$. We say f is differentiable at (x_0, y_0) if the tangent plane approximation tends to zero significantly more rapidly than $||(x, y) - (x_0, y_0)||$ tends to 0 as $(x, y) \to (x_0, y_0)$. Specifically, f is differentiable if

$$\lim_{(x,y)\to(x_0,y_0)} \frac{f(x,y) - f(x_0,y_0) - \frac{\partial f}{\partial x}(x_0,y_0)(x-x_0) - \frac{\partial f}{\partial y}(x_0,y_0)(y-y_0)}{||(x,y) - (x_0,y_0)||} \;=\; 0.$$

Note the above is truly the generalization of the derivative in one variable. The distance $x - x_0$ is replaced with $||(x, y) - (x_0, y_0)||$; while this is always positive, the fact that the limit must equal zero for the function to be differentiable means we could have used $|x-x_0|$ in the denominator in the definition of the derivative of one variable. Also note that the last two parts of the tangent plane approximation can be written as a dot product of two vectors:

$$\frac{\partial f}{\partial x}(x_0, y_0)(x - x_0) + \frac{\partial f}{\partial y}(x_0, y_0)(y - y_0) \ = \ \left(\frac{\partial f}{\partial x}(x_0, y_0), \frac{\partial f}{\partial y}(x_0, y_0)\right) \cdot (x - x_0, y - y_0).$$

(4) **Gradient:** The gradient of a function $f : \mathbb{R}^n \to \mathbb{R}$ is the vector of the partial derivatives with respect to each variable. We write

$$\text{grad}(f) \ = \ \nabla f \ = \ \left(\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n}\right).$$

The gradient points in the direction of maximum change for the function f.

(5) **Definition of the Derivative: Several Variables, Several Outputs:** Let $f : \mathbb{R}^n \to \mathbb{R}^m$; we may write

$$f(\overrightarrow{x}) \ = \ (f_1(\overrightarrow{x}, \ldots, f_m(\overrightarrow{x}))\,.$$

By $(Df)(\overrightarrow{x_0})$ we mean the matrix whose first row is $(\nabla f_1)(\overrightarrow{x})$, whose second row is $(\nabla f)(\overrightarrow{x})$, and so on until the last row, which is $(\nabla f_m)(\overrightarrow{x})$. In full glory, we have

$$(Df)(x_0) \ = \ \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(\overrightarrow{x}) & \cdots & \frac{\partial f_1}{\partial x_n}(\overrightarrow{x}) \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1}(\overrightarrow{x}) & \cdots & \frac{\partial f_m}{\partial x_n}(\overrightarrow{x}) \end{pmatrix}.$$

Note $(Df)(\overrightarrow{x})$ is a matrix with m rows and n columns. We say f is differentiable at $\overrightarrow{a}$ if the tangent hyperplane approximation for each component tends to zero significantly more rapidly than $||\overrightarrow{x} - \overrightarrow{a}||$ tends to 0 as $\overrightarrow{x} \to \overrightarrow{a}$. Specifically, f is differentiable if

$$\lim_{\overrightarrow{x} \to \overrightarrow{a}} \frac{f(\overrightarrow{x}) - f(\overrightarrow{a}) - (Df)(\overrightarrow{a}) \cdot (\overrightarrow{x} - \overrightarrow{a})}{||\overrightarrow{x} - \overrightarrow{a}||} \ = \ \overrightarrow{0},$$

where we regard $\overrightarrow{x} - \overrightarrow{a}$ as a column vector being acted on by the matrix $(Df)(\overrightarrow{a})$.

(6) **Main Theorem on Differentiation** The following implications hold (note the reverse implications may fail): (1) implies (2) implies (3), where

- ◇ (1) The partial derivatives of f are continuous.
- ◇ (2) The function f is differentiable.
- ◇ (3) The partial derivatives of f exist.

For counterexamples when reversing the implication, consider $f(x) = x^2 \sin(1/x)$ if $x \neq 0$ and 0 if $x = 0$, and $g(x, y) = (xy)^{1/3}$.

(7) **Chain Rule** Let $g : \mathbb{R}^n \to \mathbb{R}^m$ and $f : \mathbb{R}^m \to \mathbb{R}^p$ be differentiable functions, and set $h = f \circ g$ (the composition). Then

$$(Dh)(\overrightarrow{x}) \ = \ (Df)(g(\overrightarrow{x}))(Dg)(\overrightarrow{x}).$$

Important special cases are:

$\diamond$ Let $c : \mathbb{R} \to \mathbb{R}^3$ and $f : \mathbb{R}^3 \to \mathbb{R}$, and set $h(t) = f(c(t))$. Then

$$\frac{dh}{dt} \;=\; (\nabla f)(c(t)) \cdot c'(t) \;=\; \frac{\partial f}{\partial x}\frac{dx}{dt} + \frac{\partial f}{\partial y}\frac{dy}{dt} + \frac{\partial f}{\partial z}\frac{dz}{dt}.$$

Note that we could have written $\partial f/\partial x$ for df/dx.

$\diamond$ Let $g(x_1, \dots, x_n) = (u_1(x_1, \dots, x_n), \dots, u_m(x_1, \dots, x_n)$ and set $h(x_1, \dots, x_n) = f(g(x_1, \dots, x_n))$, where $f : \mathbb{R}^m \to \mathbb{R}$. Then

$$\frac{\partial h}{\partial x_i} \;=\; \frac{\partial f}{\partial u_1}\frac{\partial u_1}{\partial x_i} + \frac{\partial f}{\partial u_2}\frac{\partial u_2}{\partial x_i} + \cdots + \frac{\partial f}{\partial u_m}\frac{\partial u_m}{\partial x_i}.$$

(8) **Equality of Mixed Partial Derivatives:** Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function of class $\mathcal{C}^2$ (which means that all the partial derivatives of order at most 2 exist and are continuous). Then for any two variables x_i and x_j we have

$$\frac{\partial^2 f}{\partial x_i \partial x_j} \;=\; \frac{\partial^2 f}{\partial x_j \partial x_i}.$$

(9) **Tricks for Taylor Series Expansions:** We give a few examples of some powerful tricks to find Taylor series expansions. The idea is to use Taylor series expansions in one-variable. These work when we have functions such as $\sin(x + y)$ but not $\sin(\sqrt{x + y})$.

$\diamond$ $\cos(x + y) = 1 - \frac{(x+y)^2}{2!} + \frac{(x+y)^4}{4!} - \cdots$.

$\diamond$ $\cos x \sin y = (1 - \frac{x^2}{2!} + \cdots)(y - \frac{y^3}{3!} + \cdots)$.

$\diamond$ $e^{x-y}\cos(x + y) = (1 + (x - y) + \frac{(x-y)^2}{2!} + \cdots)(1 - \frac{(x+y)^2}{2!} + \cdots)$.

(10) **Method of Lagrange Multipliers:** Let $f, g : U \to \mathbb{R}$, where U is an open subset of $\mathbb{R}^n$. Let S be the level set of value c for the function g, and let $f|_S$ be the function f restricted to S (in other words, we only evaluate f at $\overrightarrow{x} \in U$). Assume $(\nabla g)(\overrightarrow{x}_0) \neq \overrightarrow{0}$. Then $f|_S$ has an extremum at $\overrightarrow{x}_0$ if and only if there is a λ such that $(\nabla f)(\overrightarrow{x}_0) = \lambda(\nabla g)(\overrightarrow{x}_0)$. Briefly, the reason this is true is similar to how we find max/min in one variable. If the first derivative is non-negative, we are increasing if we move one way and decreasing if we move another, so the only candidate for an extremum has the first derivative vanish (i.e., is a critical point). This is the natural generalization. We are saying all directional derivatives vanish when confined to the hyper-surface. Equivalently, if we try to 'flow' in any direction on the surface, the derivative in that direction is zero. We can look at $c(t)$ as some path in the surface with $c(0)$ at the point we care about. We look at $\mathcal{F}(t) = f(c(t))$. This is a function of one variable, and its derivative must be zero at an extremum. By the multivariable chain rule, $\mathcal{F}'(0) = (\nabla f)(c(0)) \cdot c'(0)$ (remember $c'(0)$ is a vector tangent to the surface at $c(0)$). We see that $(\nabla f)(c(0))$ is perpendicular to all vectors tangent to $c(0)$ in the hyper-surface, which forces $(\nabla f)(c(0))$ to be in the one remaining direction, which is the normal to the surface.

(11) **Method of Least Squares:** Given a set of observations

$$(x_1, y_1),\ (x_2, y_2),\ \ldots,\ (x_N, y_N)$$

and a proposed linear relationship between x and y, namely

$$y \ = \ ax + b,$$

then the best fit values of a and b (according to the Method of Least Squares) are given by minimizing the error function given by

$$E(a, b) \ = \ \sum_{n=1}^{N} \left(y_n - (ax_n + b)\right)^2 .$$

The best fit values are

$$a \ = \ \frac{\sum_{n=1}^{N} 1 \sum_{n=1}^{N} x_n y_n - \sum_{n=1}^{N} x_n \sum_{n=1}^{N} y_n}{\sum_{n=1}^{N} 1 \sum_{n=1}^{N} x_n^2 - \sum_{n=1}^{N} x_n \sum_{n=1}^{N} x_n}$$

$$b \ = \ \frac{\sum_{n=1}^{N} x_n \sum_{n=1}^{N} x_n y_n - \sum_{n=1}^{N} x_n^2 \sum_{n=1}^{N} y_n}{\sum_{n=1}^{N} x_n \sum_{n=1}^{N} x_n - \sum_{n=1}^{N} x_n^2 \sum_{n=1}^{N} 1}. \tag{2.1}$$

Frequently by taking logarithms we can use this method for non-linear relations. For example, if $T = BL^a$, then if $\mathcal{T} = \log T$, $\mathcal{L} = \log L$ and $b = \log B$ then $\mathcal{T} = a\mathcal{L} + b$, a linear relation.

In a probability or a stats class you'll learn the associated matrix is always invertible (i.e., we are not dividing by zero when finding a and b) so long as the x_i's are not all the same value. This is because we may interpret the denominators as a non-zero multiple of the variance of the x_i's, and variances are non-negative.

(12) **Metric dependence of answers:** A very important fact, made clear in the previous subject, is that depending on the metric used to evaluate / answer a problem one can reach different conclusions. What do we 'mean' by best-fit line? Depending on how we measure the data (ranging from just summing the signed errors to absolute values to squares), we can get a different answer. It is very important to be aware of these situations.

(13) **Monte Carlo Integration:** Let D be a nice region in $\mathbb{R}^n$, and assume for simplicity that it is contained in the n-dimensional unit hypercube $[0, 1] \times [0, 1] \times \cdots \times [0, 1]$. Assume further that it is easy to verify if a given point $(x_1, \ldots, x_n)$ is in D or not in D. Draw N points from the n-dimensional uniform distribution; in other words, each of the n coordinates of the N points is uniformly distributed on $[0, 1]$. Then as $N \to \infty$ the n-dimensional volume of D is well approximated by the number of points inside D divided by the total number of points.

(14) **Fubini Theorem (or Fubini-Tonelli):** Frequently we want to / need to justify interchanging two integrals (or an integral and a sum). Doing such interchanges is one of the most frequent tricks in mathematics; whenever you see a double sum, a double integral, or a sum and an integral you should consider this. While we cannot always interchange orders, we can if the double sum (or double integral) of the absolute value of the summand (or the integrand) is

finite. For example,

$$\begin{aligned}\int_{y=0}^{1}\left[\int_{x=0}^{1}e^{-xy}xdx\right]dy &= \int_{x=0}^{1}\left[\int_{y=0}^{1}e^{-xy}xdy\right]dx\\ &= \int_{x=0}^{1}e^{-xy}\Big|_{1}^{0}dx\\ &= \int_{x=0}^{1}\left(1-e^{-x}\right)dx \;=\; 2-e^{-x}. \end{aligned} \tag{2.2}$$

Note how much easier it is when we integrate with respect to y first – we bypass having to use Integration by Parts. For completeness, we state:

Fubini's Theorem: *Assume f is continuous and*

$$\int_a^b\int_c^d |f(x,y)|dxdy \;<\; \infty. \tag{2.3}$$

Then

$$\int_a^b\left[\int_c^d f(x,y)dy\right]dx \;=\; \int_c^d\left[\int_a^b f(x,y)dx\right]dy. \tag{2.4}$$

Similar statements hold if we instead have

$$\sum_{n=N_0}^{N_1}\int_c^d f(x_n,y)dy, \qquad \sum_{n=N_0}^{N_1}\sum_{m=M_0}^{M_1} f(x_n,y_m). \tag{2.5}$$

(15) ***Whenever you have a theorem, you should always explore what happens if you remove a condition. Frequently (though not always) the claim no longer holds; sometimes the claim is still true but the proof is harder. Rarely, but it can happen, removing a condition causes you to look at a problem in a new light, and find a simpler proof.*** We apply this principle to Fubini's theorem; specifically, we remove the finiteness condition and construct a counter-example.

For simplicity, we give a sequence a_{mn} such that $\sum_m(\sum_n a_{m,n}) \neq \sum_n(\sum_m a_{m,n})$. For $m,n \geq 0$ let

$$a_{m,n} \;=\; \begin{cases} 1 & \text{if } n=m\\ -1 & \text{if } n=m+1\\ 0 & \text{otherwise.}\end{cases} \tag{2.6}$$

We can show that the two different orders of summation yield different answers; if we sum over the columns first we get 0 for each column, and then doing the sum of the column sums gives 0; however, if we do the row sums first, than all the row sums vanish but the first (which is 1), and hence the sum of the row sums is 1, *not* 0. The reason for this difference is that the sum of the absolute value of the terms diverges.

(16) **Interchanging derivatives and sums:** It is frequently useful to interchange a derivative and an infinite sum. The first place this is met is in proving the derivative of e^x is e^x; using the series expansion for e^x, it is trivial to find the derivative *if* we can differentiate term by

term and then add.

Interchanging differentiation and integration: *Let $f(x,t)$ and $\partial f(x,t)/\partial x$ be continuous on a rectangle $[x_0,x_1]\times[t_0,t_1]$ with $[a,b]\subset[t_0,t_1]$. Then*

$$\frac{d}{dx}\int_{t=a}^{b} f(x,t)dt \;=\; \int_{t=a}^{b}\frac{\partial f}{\partial x}(x,t)dt. \tag{2.7}$$

Frequently one wants to interchange differentiation and summation; this leads to the method of differentiating identities, which is extremely useful in computing moments of probability distributions. For example, consider the identity

$$(p+q)^n \;=\; \sum_{k=0}^{n}\binom{n}{k}p^k q^{n-k}. \tag{2.8}$$

Applying the operator $p\frac{d}{dp}$ to both sides we find

$$p\cdot n(p+q)^{n-1} \;=\; \sum_{k=0}^{n} k\binom{n}{k}p^k q^{n-k}. \tag{2.9}$$

Setting $q=1-p$ yields the mean of a binomial random variable:

$$np \;=\; \sum_{k=0}^{n} k\binom{n}{k}p^k (1-p)^{n-k}. \tag{2.10}$$

It is very important that initially p and q are distinct, free variables, and only at the end do we set $q=1-p$.

(17) **Dangers when interchanging:** One has to be very careful in interchanging operations. Consider, for example, the family of probability densities[1] $f_n(x)$, where f_n is a triangular density on $[1/n,3/n]$ with midpoint (i.e., maximum value) n. While each f_n is continuous (as is the limit $f(x)$, which is identically 0), each f_n is a probability density (as each integrates to 1); however, the limit density is identically 0, and thus not a density! We can easily modify our example so that the limit is not continuous:

$$g_n(x) \;=\; \begin{cases} n|x| & \text{if } 0\le|x|\le 1/n \\ 1 & \text{if } 1/n\le|x|\le 1/2 \\ n\left(\frac{1}{2}+\frac{1}{n}-|x|\right) & \text{if } 1/2\le x\le 1/2+1/n \\ 0 & \text{otherwise.}\end{cases} \tag{2.11}$$

Note that $g_n(0)=0$ for all n, but as we approach 0 from above or below, in the limit we get 1.

(18) **Change of Variables Theorem:** *Let V and W be bounded open sets in $\mathbb{R}^n$. Let $h:V\to W$ be a 1-1 and onto map, given by*

$$h(u_1,\ldots,u_n) \;=\; \left(h_1(u_1,\ldots,u_n),\ldots,h_n(u_1,\ldots,u_n)\right). \tag{2.12}$$

[1] A function p is a probability density if $p(x)\ge 0$ and p integrates to 1.

Let $f : W \to \mathbb{R}$ be a continuous, bounded function. Then

$$\begin{aligned}&\int \cdots \int_W f(x_1, \ldots, x_n) dx_1 \cdots dx_n \\ &= \int \cdots \int_V f\left(h(u_1, \ldots, u_n)\right) |J(u_1, \ldots, u_v)| \, du_1 \cdots du_n, \end{aligned} \tag{2.13}$$

*where J is the **Jacobian***

$$J \ = \ \begin{vmatrix} \frac{\partial h_1}{\partial u_1} & \cdots & \frac{\partial h_1}{\partial u_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial h_n}{\partial u_1} & \cdots & \frac{\partial h_n}{\partial u_n} \end{vmatrix}. \tag{2.14}$$

We used this result to simplify the algebra in many problems by passing to an easier set of variables.

(19) **Counting two different ways / telling a story:** Calculating something two different ways is one of the most important ideas in math. A good part of combinatorics is to note that there are two ways to compute something, one of which is easy and one of which is not. We then use our knowledge of the easy calculation to deduce the hard. For example, $\sum_{k=0}^{n} \binom{n}{k}^2 = \binom{2n}{n}$; the right side is easy to compute, the left side not so clear. Why are the two equal? It involves finding a story. Imagine we have $2n$ people, say n Red Sox fans and n Yankee fans. How many ways are there to form a group of n people from the $2n$ people, if order doesn't matter? One way is to note it is just $\binom{2n}{n}$. Another way is to look at how many Red Sox fans we have in our group of n. Let k be the number of Red Sox fans we choose; we may take any integer k from 0 to n. There are $\binom{n}{k}$ ways to choose k Red Sox fans from n, and thus $\binom{n}{n-k} = \binom{n}{k}$ ways to choose $n-k$ Yankee fans (if we want n people total, if we have k Red Sox fans we need $n-k$ Yankee fans, and of course no one likes both teams!). Thus the number of ways to form our group is just $\sum_{k=0}^{n} \binom{n}{k}\binom{n}{n-k} = \sum_{k=0}^{n} \binom{n}{k}^2$, which must equal $\binom{2n}{n}$. See how easy it is to evaluate the sum *if* we can tell the right story!

(20) **Memoryless process:** When proving the geometric series formula by playing a basketball game, we used the fact that after two misses it was as if we just started playing the game then. This idea is used in many problems.

(21) **Ratio, root, integral and comparison tests:** These are used to determine if a series or integral converges. We frequently used the geometric series formula $\sum_{n=0}^{\infty} x^n = 1/(1-x)$ if $|x| < 1$.

◇ **Comparison Test** Let $\{b_n\}_{n=1}^{\infty}$ be a sequence of non-negative terms (so $b_n \geq 0$). Assume the series converges, and $\{a_n\}_{n=1}^{\infty}$ is another sequence such that $|a_n| \leq b_n$ for all n. Then the series attached to $\{a_n\}_{n=1}^{\infty}$ also converges.

◇ **Ratio Test** Consider a sequence $\{a_n\}_{n=1}^{\infty}$ of positive terms. Let

$$r \ = \ \lim_{n\to\infty} \frac{a_{n+1}}{a_n}.$$

If r exists and $r < 1$ then the series converges, while if $r > 1$ then the series diverges; if $r = 1$ then this test provides no information on the convergence or divergence of the series.

◇ **Root Test** Consider a sequence $\{a_n\}_{n=1}^{\infty}$ of positive terms. Let

$$\rho \;=\; \lim_{n\to\infty} a_n^{1/n},$$

the n^{th} root of a_n. If $\rho < 1$ then the series converges, while if $\rho > 1$ then the series diverges; if $\rho = 1$ then the test does not provide any information.

◇ **Integral Test** Consider a sequence $\{a_n\}_{n=1}^{\infty}$ of non-negative terms. Assume there is some function f such that $f(n) = a_n$ and f is non-increasing. Then the series

$$\sum_{n=1}^{\infty} a_n$$

converges if and only if the integral

$$\int_1^{\infty} f(x)dx$$

converges.

3. DIFFERENTIAL EQUATIONS (MATH 209)

(1) **The method of Divine Inspiration and Difference Equations:** Difference equations, such as the Fibonacci equation $a_{n+1} = a_{n+1} + a_n$, arise throughout nature. There is a rich theory when we have linear recurrence relations. To find a solution, we 'guess' that $a_n = r^n$ and take linear combinations.

Specifically, let k be a fixed integer and $c_1, \ldots, c_k$ given real numbers. Then the general solution of the difference equation

$$a_{n+1} \ = \ c_1 a_n + c_2 a_{n-1} + c_3 a_{n-2} + \cdots + c_k a_{n-k+1}$$

is

$$a_n \ = \ \gamma_1 r_1^n + \cdots + \gamma_k r_k^n$$

if the characteristic polynomial

$$r^k - c_1 r^{k-1} - c_2 r^{k-2} - \cdots - c_k \ = \ 0$$

has k distinct roots. Here the $\gamma_1, \ldots, \gamma_k$ are any k real numbers; if initial conditions are given, these conditions determine these γ_i's. If there are repeated roots, we add terms such as nr^n, $\ldots$, $n^{m-1}r^n$, where m is the multiplicity of the root r.

For example, consider the equation $a_{n+1} = 5a_n - 6a_{n-1}$. In this case $k = 2$ and we find the characteristic polynomial is $r^2 - 5r + 6 = (r-2)(r-3)$, which clearly has roots $r_1 = 2$ and $r_2 = 3$. Thus the general solution is $a_n = \gamma_1 2^n + \gamma_2 3^n$. If we are given $a_0 = 1$ and $a_1 = 2$, this leads to the system of equations $1 = \gamma_1 + \gamma_2$ and $2 = \gamma_1 \cdot 2 + \gamma_2 \cdot 3$, which has the solution $\gamma_1 = 1$ and $\gamma_2 = 0$.

Applications include population growth (such as the Fibonacci equation) and why double-plus-one is a bad strategy in roulette.

4. REAL ANALYSIS (MATH 301)

(1) **Continuity:** General continuity properties, in particular some of the $\epsilon - \delta$ arguments to bound quantities, are frequently used to prove results. Often we use these to study moments or other properties of densities. Most important, however, was probably when we can interchange operations, typically interchanging integrals, sums, or an infinite sum and a derivative. For the derivative of the geometric series, this can be done by noting the tail is another geometric series; in general this is proved by estimating the contribution from the tail of the sum). See the multivariable calculus section for more comments on these subjects.

(2) **Proofs by Induction:** Induction is a terrific way to prove formulas for general n *if* we have a conjecture as to what the answer should be. Assume for each positive integer n we have a statement $P(n)$ which we desire to show is true for all n. $P(n)$ is true for all positive integers n if the following two statements hold: (i) **Basis Step:** $P(1)$ is true; (ii) **Inductive Step**: whenever $P(n)$ is true, $P(n+1)$ is true. Such proofs are called proofs by induction or induction (or inductive) proofs.

The standard examples are to show results such as $\sum_{k=0}^{n} k = \frac{n(n+1)}{2}$. It turns out that $\sum_{k=0}^{n} k^m$ is a polynomial in n of degree $m+1$ with leading coefficient $1/(m+1)$ (one can see that this is reasonable by using the integral test to replace the sum with an integral); however, the remaining coefficients of the polynomial are harder to find, and without them it is quite hard to run the induction argument for say $m = 2009$.

(3) **Dirichlet's Pigeonhole principle:** Let $A_1, A_2, \ldots, A_n$ be a collection of sets with the property that $A_1 \cup \cdots \cup A_n$ has at least $n+1$ elements. Then at least one of the sets A_i has at least two elements. We frequently use the Pigeonhole principle to ensure that some event happens.

5. COMPLEX ANALYSIS (MATH 302)

5.1. **Complex Differentiability.** Similar to one-dimensional real variable calculus, everything in complex analysis follows from the definition of the derivative. What drastically changes the subject from a real variable is the geometry of the space. In the real line, you can approach a point essentially in only two ways: from above or from below. In the complex plane, there are an infinitude of paths, ranging from along the axes to spirals to what Cam and Kayla draw. This leads to the definition

Complex differentiability: A function of a complex variable is said to be complex differentiable at z if

$$\lim_{h\to 0} \frac{f(z+h)-f(z)}{h}$$

exists as $h \neq 0$ tends to 0 along any path.

Functions such as the polynomials $\sum_{k=0}^{n} c_k z^k$ are differentiable, while functions such as $\overline{z}^k$ are not (remember $\overline{z} = x - iy$ if $z = x + iy$).

If $f(x+iy) = u(x,y) + iv(x,y)$, then f is holomorphic (i.e., complex differentiable) if and only if it satisfies the **Cauchy-Riemann equations**:

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y} \text{ and } \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x}.$$

5.2. **Cauchy's Theorem.** We say a function f has a **primitive** F if $F'(z) = f(z)$. From the theory of line integrals, we see that if γ is a simple closed curve and f has a primitive, then $\int f(z)dz = 0$. The main result is Cauchy's Theorem:

Cauchy's Theorem: Let f be a holomorphic function and γ a simple closed curve. Then $\int_\gamma f(z)dz = 0$.

There are many ways to prove this. A popular one is to first prove **Goursat's Theorem**: if f is holomorphic on an open set containing a triangle T, then $\int_{\partial T} f(z)dz = 0$ (where ∂T is the boundary of the triangle). The key step in proving this is to keep reducing the line integral into four smaller line integrals; geometrically this is doable as we can divide a triangle easily into four similar triangles. We then use some compactness arguments to finish the proof. From Goursat's Theorem, we can then prove that any holomorphic function on an open disk has a primitive on the disk. We do this by taking polygonal paths with components parallel to the x and y-axes.

There are many consequences of Cauchy's Theorem. The fist is to allow us to evaluate many integrals. Another is the

Integral Representation Theorem: If f is holomorphic on an open set that contains a closed curve γ, then for any z in the set we have

$$f(z) = \frac{1}{2\pi i}\int_\gamma \frac{f(\zeta)}{\zeta - z}dz.$$

Notice that this means that if we know f on the boundary and if we know f is holomorphic then there is a unique extension to the interior. This formula has a multitude of consequences as well. In particular, it gives us a series expansion for a holomorphic function, and shows that holomorphicity implies analyticity (the converse is straightforward, as we can differentiate power series term by term). Recall that a function is **analytic** if it has a convergent series expansion.

Holomorphic equals Analytic: If a complex function is differentiable once, it is infinitely differentiable and it equals its Taylor series. Specifically,

$$f(z) \;=\; \sum_{n=0}^{\infty} a_n (z-z_0)^n, \quad a_n \;=\; f^{(n)}(z_0)/n!$$

where

$$f^{(n)}(z) \;=\; \frac{n!}{2\pi i}\int_\gamma \frac{f(\zeta)}{(\zeta - z)^{n+1}} dz,$$

where as always γ is a simple closed curve. The proof follows by using the geometric series formula to expand the denominator in the integral representation; specifically

$$\zeta - z \;=\; (\zeta - z_0) - (z - z_0) \;=\; (\zeta - z_0)\cdot\left(1 - \frac{z-z_0}{\zeta - z_0}\right);$$

for z close to z_0, the fraction above is less than one and we may expand the reciprocal of the above with the geometric series formula; note the extra factor of $\zeta - z_0$ in front is what is responsible for the exponent being $n+1$ and not n.

This is remarkably different than real analysis (remember the function $f(x) = \exp(-1/x^2)$ for $x \neq 0$ and 0 for $x = 0$; this function is infinitely differentiable, but only equals its Taylor series at $x = 0$ (which is not impressive, as by definition all functions equal their Taylor series at the expansion point!).

Another consequence are the **Cauchy Inequalities**, which state that if f is holomorphic on a set containing a circle with boundary C centered at z_0 with radius R then

$$|f^{(n)}(z_0)| \;\leq\; \frac{n!||f||_C}{R^n},$$

where $||f||_C$ denotes the largest value of f on C.

From the Cauchy Inequalities we immediately obtain **Liouville's Theorem** (also known as the first big theorem without Cauchy's name in it): if f is entire (i.e., holomorphic on all of $\mathbb{C}$) and bounded then f is entire. The proof follows by using the Cauchy Inequalities on larger and larger circles; as f is bounded the numerators are uniformly bounded while the denominators tend to zero with R. From Liouville's Theorem we obtain the **Fundamental Theorem of Algebra**, which states that any degree n polynomial with complex coefficients has exactly n roots.

We end with two other important concepts.

Analytic continuation: Given a function f defined in some subset of the complex plane, its analytic continuation is a new function which agrees with the old in the original region, but makes

sense elsewhere. The standard example is the geometric series formula: $\sum_{n=0}^{\infty} x^n = 1/(1-x)$; the right hand side makes sense for all values of $x \neq 1$, while the left hand side is only defined if $|x| < 1$. This leads to the interpretation that $1 + 2 + 4 + 8 + 16 + \cdots = -1$!

Accumulation points: Let f be a complex differentiable function defined on an open set U; assume $f(z_n) = 0$ for some sequence of points $\{z_n\}_{n=1}^{\infty}$ that has an accumulation point in U (i.e., there is some z^* such that a subsequence of the z_n's converge to z^*). Then f is identically zero! Again, this is very different than real analysis: the function $f(x) = x^3 \sin(1/x)$ for $x \neq 0$ and 0 for $x = 0$ is zero whenever $x = 1/n\pi$, and is zero at $x = 0$; however, clearly this function is not identically zero even near the origin (just consider $x = 2/n\pi$ for n odd). In probability, this result is used to study the moment problem, namely, how many moments are needed to uniquely determine a probability density. The proof involves the equivalence between holomorphicity and analyticity. We Taylor expand our function about the accumulation point, and note that all the derivatives vanish there (some work is required to show that).

5.3. **The Residue Formula.** Perhaps the most important result in complex analysis (from an applications standpoint, though this arises in numerous theoretical investigations as well) is

The Cauchy Residue Formula: Suppose f is holomorphic in an open set containing a simple closed curve γ except for finitely many poles (at $z_0, \ldots, z_n$) with residues $\mathrm{Res}_{z_j}(f)$. Then

$$\frac{1}{2\pi i} \int_{\gamma} f(z)dz \; = \; \sum_{j=1}^{n} \mathrm{Res}_{z_j}(f).$$

One proves this in a similar manner to extensions of Green's theorem in the plane, taking contours where we approach one of the poles, circle around it, and then retrace our steps back to the main curve. The point is to convert the integral over γ to n integrals over circles centered at the poles.

The residue of f at z_0 is the negative first coefficient in its Laurent expansion at z_0 (the Laurent expansion is similar to the Taylor expansion, except now we allow z to be raised to negative integer powers as well). A useful way to compute residues is the following:

Computing Residues: Assume $f(z) = g(z)/h(z)$ where g and h are holomorphic and h has a simple zero at z_0 (i.e., the zero has multiplicity one). Then the residue of f at z_0 is $g(z_0)/h'(z_0)$.

We can use the residue theorem to evaluate many integrals, especially real integrals. We complete the contours, carefully choosing our completion to exploit the decay in the function. For some examples, see §5.6.

We list a few applications of the Residue Theorem:

Argument Principle: If f is meromorphic (holomorphic except at finitely many places where it has poles) on an open set containing some simple closed curve γ then

$$\frac{1}{2\pi i} \int_{\gamma} \frac{f'(z)}{f(z)} dz$$

equals the number of zeros of f inside γ minus the number of poles inside γ. The formula can be generalized to include a factor $g(z)$ in the integrand, which results in the zeros and poles receiving different weights which are a function of g at these points. This weighting is quite useful in number theory.

Rouche's Theorem: If f and g are holomorphic on an open set containing a simple closed curve γ and $|f(z)| > |g(z)|$ for all z on γ then f and $f+g$ have the same number of zeros inside γ. The idea is to use the argument principle, and note the integral of $h'(z)/h(z)$ is a continuous and integer valued. If we take the family $h_t(z) = f(z) + tg(z)$, we see that the value at $t = 0$ equals the value at $t = 1$.

Rouche's theorem provides a nice way to reach several useful results. We say a map is **open** if it maps open sets to open sets. Note many simple real valued functions are not open; for example, $f(x) = x^2$ is not open as it maps $(-1, 1)$ to $[0, 1)$. We have

Open Mapping Theorem: If f is holomorphic and non-constant in a region Ω, then f is open.

We may also prove the Open Mapping Theorem by noting that a holomorphic function is analytic, and then analyzing the series expansion. This is more work if you only want the Open Mapping Theorem, but this perspective is useful for other problems, as it gives a better sense of what the map is going. This point of view can surface in studying the Riemann Mapping Theorem.

Maximum Modulus Principle: If f is non-constant and holomorphic on an open set, then f cannot attain its maximum in the open set (if we assume f is continuous on the boundary of our open set, then f attains its maximum on the boundary). The Open Mapping Theorem is the key ingredient in the proof; as f is open, if it attained a maximum at an interior point z_0 then the image of a neighborhood of that point includes a ball about $f(z_0)$, and hence includes a point with larger absolute value.

Finally, we note that the complex logarithm exists, though it does not have all the properties of the real logarithm. It can be defined on any simply connected set that is not all of $\mathbb{C}$.

5.4. **Weierstrass Products.** It is convenient to represent a function as a product. This is especially true if we are going to consider its logarithmic derivative, $f'(z)/f(z) = \frac{d}{dz}\log f(z)$ (which the argument principle tells us is a natural item to study). An infnite product $\prod(1 + a_n)$ converges if $\sum |a_n| < \infty$ (it may of course converge even if this sum diverges).

Let $E_0(z) = 1 - z$ and $E_k(z) = (1 - z)\exp(z + z^2/2 + \cdots + z^k/k!)$ for $k \geq 1$. Writing $1 - z$ as $\exp(\log(1 - z))$, we see the exponential factor is deliberately chosen to cancel the first k terms of $\log(1 - z)$. These are called the **canonical factors**.

Weierstrass Products: Given any sequence $\{a_n\}$ of complex numbers with $|a_n| \to \infty$ as $n \to \infty$, there is an entire function f that vanishes at each a_n and nowhere else. If g also only vanishes at these points, then there is an entire $h(z)$ such that $f(z) = g(z)\exp(h(z))$.

The key ingredient in the proof is to use these canonical factors to make sure our product converges. Specifically, if we want to have m zeros at the origin then we look at $f(z) = z^m \prod_{n=1}^{\infty} E_n(z/a_n)$. This is wasteful; note the degree of the canonical factor is growing. In many instances we can always take $n = 1$ or $n = 2$.

5.5. **The Riemann Mapping Theorem.** Undoubtable one of the gems of the subject is the Riemann Mapping Theorem. The proof is quite involved, and requires numerous concepts from analysis. The main application of this beautiful result is that we can reduce the analysis of many problems to the study of an equivalent problem in the unit disk. This is a lot like the change of base formulas for logarithms: $\log_b x = \log_c x / \log_c b$, which implies that if we know logarithms in one base we know them in any base.

We say a map $f : U \to V$ is **conformal** if f is bijective and holomorphic. If that is the case, we say U and V are **conformally equivalent**; one can show this is in fact an equivalence relation (the hard part is showing that the inverse of f is holomorphic, but we can do this through the chain rule).

Below are some useful facts. Remember $\mathbb{D}$ is the unit disk.

(1) The Schwarz lemma: if $f : \mathbb{D} \to \mathbb{D}$ is holomorphic and $f(0) = 0$ then (i) $|f(z)| \leq |z|$ for all $z \in \mathbb{D}$; (ii) If $|f(z_0)| = |z_0|$ for some $z_0 \in \mathbb{D}$ then f is a rotation (i.e., $f(z) = e^{i\theta} z$ for some $\theta \in \mathbb{R}$; (iii) $|f'(0)| \leq 1$ and if it equals 1 then f is a rotation. The proof involves looking at f as a power series (holomorphic implies analytic) and using the maximum modulus principle.

(2) Automorphisms of the unit disk: Letting $\psi_\alpha(z) = (\alpha - z)/(1 - \overline{\alpha} z)$ for $\alpha \in \mathbb{C}$ with $|\alpha| < 1$, we find that if f is an automorphism of $\mathbb{D}$ then there is a $\theta \in \mathbb{R}$ and an α with absolute value less than 1 such that $f(z) = e^{i\theta} \psi_\alpha(z)$. The proof uses the Schwarz lemma repeatedly. From this we can enumerate all automorphisms of any set conformally equivalent to $\mathbb{D}$.

(3) If Ω is an open subset of $\mathbb{C}$ and $\mathcal{F}$ is a family of holomorphic functions, then the family is said to be **normal** on Ω if every sequence in $\mathcal{F}$ has a subsequence that converges uniformly on every compact subset of Ω (note the limit function need not be in $\mathcal{F}$). The family is said to be **uniformly bounded on compact subsets of** Ω if for any compact $K \subset \Omega$ there is a constant $B_K > 0$ such that for any f in the family we have $|f(z)| \leq B_K$ for all $z \in K$. The family is **equicontinuous** on a compact set K if given any $\epsilon > 0$ there is a $\delta > 0$ such that for any f in the family and any $z, w \in K$ with $|z - w| < \delta$ then $|f(z) - f(w)| < \epsilon$.

The main result we need from all of this is

Montel's Theorem: Let $\mathcal{F}$ be a family of holomorphic functions on Ω. Assume the family is uniformly bounded on compact subsets of Ω. Then (1) The family is equicontinuous on compact subsets of Ω; (2) the family is normal.

The last part of this theorem is often called the Arzela-Ascoli Theorem. The proof of (1) involves the Cauchy integral formula, which shows why a similar statement does not hold in the real case (as we do not have an analogue of this integral representation). The second part requires us to find a countable, dense subset of Ω, which is possible by looking at those $z = x + iy$ in Ω with $x, y \in \mathbb{Q}$. This part does not involve complex analysis, and the corresponding result holds in the real case.

Riemann Mapping Theorem: Any open, proper (i.e., not all of $\mathbb{C}$) subset Ω of $\mathbb{C}$ that is simply connected is conformally equivalent to the unit disk $\mathbb{D}$.

There are three main steps to the proof. (1) We first show that Ω is conformally equivalent to a subset of $\mathbb{D}$. If Ω misses an entire ball, we simply invert about that ball and rescale. If not, we use the complex logarithm to first map Ω to a set that misses an entire ball and then invert. (2) We consider the family of holomorphic injective maps from $\Omega \to \mathbb{D}$ with $f(0) = 0$ (recall now we are able to assume $\Omega \subset \mathbb{D}$). We use Montel's theorem, and find a function in the family whose absolute value of its derivative at 0 is equal to the supremum of the absolute values of the derivatives at the origin of all holomorphic functions in our family. (3) We then show that the function found in (2) is surjective, completing the proof; if it weren't surjective, we explicitly show how to construct a new function that has larger absolute value of its derivative at the origin.

It is a natural question to ask when one can extend the map to the boundary; this can be done in some cases. In fact, if our region is a nice polygon, we can write down explicitly the conformal equivalence.

Note how amazing this result is – we have the existence of an infinitely differentiable map from the unit square to the unit disk (as well as more complicated regions!).

5.6. **Examples of Contour Integrals.** The following are the solutions to some contour integrals from the midterm, and highlight many of the techniques.

- (a: 20 points)
$$\int_{-\infty}^{\infty} \frac{x^2}{x^4+1}\, dx.$$
- (b: 20 points)
$$\oint_{\gamma} \frac{\exp(-z^2)}{z^2}\, dz$$
where γ is the diamond given by $|x| + |y| = 2010$.
- (c: 20 points)
$$\int_{-\infty}^{\infty} \frac{\cos x}{e^x + e^{-x}}\, dx.$$
- (d: 20 points)
$$\int_0^{2\pi} \frac{1}{a + \sin\theta}\, d\theta$$
where $a > 1$. *Hint: on the unit circle,* $e^{-i\theta} = 1/e^{i\theta} = 1/z$. *Try writing* $\sin\theta$ *as* $g(z, 1/z)$ *for some function* g.

(a) Let γ_1 be the part of the real line from $-R$ to R and γ_2 be a counter-clockwise oriented semicircle of radius R from R to $-R$. Two poles lie in the region enclosed by these paths: $z = e^{\pi i/4}$ and $e^{3\pi i/4}$. The residues are
$$\lim_{z \to e^{\pi i/4}} \frac{(z - e^{\pi i/4} z^2)}{z^4+1} = \frac{e^{\pi i/2}}{(e^{\pi i/4} - e^{3\pi i/4})(e^{\pi i/4} - e^{5\pi i/4})(e^{\pi i/4} - e^{7\pi i/4})}$$
$$= \frac{\sqrt{2}}{8} - \frac{\sqrt{2}}{8} i$$

and

$$\lim_{z\to e^{3\pi i/4}} \frac{(z-e^{3\pi i/4}z^2)}{z^4+1} = \frac{e^{3\pi i/2}}{(e^{3\pi i/4}-e^{\pi i/4})(e^{3\pi i/4}-e^{5\pi i/4})(e^{3\pi i/4}-e^{7\pi i/4})}$$
$$= -\frac{\sqrt{2}}{8} - \frac{\sqrt{2}}{8}i$$

So

$$\int_{\gamma_1} \frac{z^2}{z^4+1}\,dz + \int_{\gamma_2} \frac{z^2}{z^4+1}\,dz = 2\pi i(-\frac{\sqrt{2}}{4}i) = \frac{\pi\sqrt{2}}{2}$$

Now we can calculate

$$\left|\int_{\gamma_2} \frac{z^2}{z^4+1}\,dz\right| \le \pi R \cdot \frac{R^2}{R^4-1} = \frac{\pi R^3}{R^4-1}$$

which clearly goes to zero as R goes to infinity. So we then have

$$\int_{-\infty}^{\infty} \frac{z^2}{z^4+1}\,dz = \frac{\pi\sqrt{2}}{2}$$

by taking the limit of γ_1 as $R \to \infty$.

(b) Since the curve is already closed, we need only find the resdiues. There is clearly only one pole, a pole of order 2 at $z = 0$. The residue at $z = 0$ is 0. One can see this by noting that the Taylor series expansion of $\exp(-z^2)/z^2$ contains only even powers of z; hence the coefficient of $1/z$ is 0. Since the integral is $2\pi i$ times the sum of the residues, the integral itself must just be 0.

(c) This problem is a bit tricky. The problem is the numerator is $\cos z = (e^{iz}+e^{-iz})/2$ while the denominator is $e^z + e^{-z}$. If we take $z = x + ix$ with $x \to \infty$, then our function is essentially equal to 1/2, and thus it is *not* decaying to infinity! This tells us that a semi-circular contour is probably not goign to be a good idea. For problems like this, when you aren't given the contour, it's best to try and get a feel for the size of the function in different places. Here we see

$$\frac{e^{iz}+e^{-iz}}{e^z+e^{-z}} = \frac{e^{ix}e^{-y}+e^{-ix}e^{y}}{e^xe^{iy}+e^{-x}e^{-iy}},$$

and then taking $x = y$ shows that this ratio is essentially 1/2 for x large.

What contour should we choose? Seeing exponential functions like these, a rectangular one is a natural guess. Why? We have periodicity in the exponential functions, and there is thus a chance of the 'top' and 'bottom' being simply multiples of each other. For this one, we take a rectangle with vertices $-R, R, R + i\pi$ and $-R + i\pi$. It's not too bad to show that the integral over the two vertical sides tends to zero as $R \to \infty$, and then a little algebra relates the top contribution to the bottom. All that remains is to find the poles (which probably should have been done earlier). We need $e^z + e^{-z} = 0$, or $e^{2z} = -1$. As $e^{i\theta} = -1$ for $\theta = \pi \pm 2\pi n$, we see the poles are located at $z = \pi/2 \pm \pi n$, which means that there is only one pole inside the region. After doing all the algebra, one finds the answer is $\pi/(e^{\pi/2} + e^{-\pi/2})$.

(d) If z is a point on the unit circle, say, $z = e^{i\theta}$, then $e^{-i\theta} = 1/z$. Then $\sin\theta = (z - 1/z)/2i$ (the i in the denominator is very important – you will get an answer that doesn't make sense if you forget it – more on that observation later). Further, since $z = e^{i\theta}$ we have

$$dz \;=\; ie^{i\theta}d\theta \;=\; iz d\theta$$

or equivalently

$$d\theta \;=\; -\frac{id\theta}{z}.$$

Therefore

$$\int_0^{2\pi} \frac{1}{a+\sin\theta}\,d\theta = \int_\gamma \frac{1}{a+(z-1/z)/2i}\cdot -\frac{i}{z}\,dz = \int_\gamma \frac{2}{z^2+2iaz-1}\,dz$$

where γ is the unit circle centered at 0. The integrand has poles at $-ia \pm i\sqrt{a^2-1}$. The only one inside the unit circle, however, is $-ia + i\sqrt{a^2-1}$ (because $|a| > 1$), which gives the residue as

$$\frac{2}{(-ia+i\sqrt{a^2-1}) - (-ia-i\sqrt{a^2-1})} = \frac{1}{i\sqrt{a^2-1}}$$

So the integral is just $2\pi i$ times the above; that is,

$$\int_0^{2\pi} \frac{1}{a+\sin\theta}\,d\theta = \frac{2\pi}{\sqrt{a^2-1}}.$$

Note that the answer has a very good property – it is undefined when $a^2 \le 1$. This makes sense, as our original integral is only well-defined for $a^2 > 1$, because for smaller a^2 the denominator can vanish. It is very important to be able to glance at a solution and test for reasonableness.

Another way to do this problem is to use the geometric series expansion, assuming you know the integral of $\sin x$ to any integer power. We have

$$\frac{1}{a+\sin\theta} \;=\; \frac{1}{a}\frac{1}{1-\left(-\frac{\sin\theta}{a}\right)} \;=\; \frac{1}{a}\left(1 - \frac{\sin\theta}{a} + \frac{\sin^2\theta}{a^2} - \cdots\right).$$

Of course, even after doing the integral of $\sin^{2k}\theta$ we're not done – we then have to recognize the Taylor series expansion of $2\pi(a^2-1)^{-1/2}$, a highly non-trivial observation to make. That said, this does show that, in principle, this integral *could* be done without resorting to complex analysis (and gives an appreciation of the power of complex analysis!).

Let's spend a little more time thinking about this problem. *We can 'see' many features of the solution. As $a \to \infty$, the denominator is essentially a, and thus the integral tends to $2\pi/a$ as $a \to \infty$. We know the integral makes sense for $|a| > 1$ and not for $|a| \le 1$; the denominator is zero when $a = \pm 1$. Thus it is reasonable to guess that the denominator of the integral looks like $\sqrt{a^2-1}$. Why? This is bad for $|a| \le 1$, and tends to a (if we take the appropriate square-root) for a large. This is not a* proof, *but it suggests that the answer should be something like $2\pi/\sqrt{a^2-1}$.*

6. FOURIER ANALYSIS (MATH 3XX)

(1) **Integral transforms:** If $K(s,t)$ and $g(t)$ are nice functions, we define the integral transform of g with kernel K to be $\int_{-\infty}^{\infty} g(t)K(s,t)dt$. What this does is, given a function as input, generates a new function. Two particularly useful transforms are the Fourier transform ($\widehat{f}(y) = \int_{-\infty}^{\infty} f(x)e^{-2\pi ixy}dx$) and the Laplace transform ($(\mathcal{L}f)(s) = \int_0^{\infty} f(t)e^{-st}dt$). Depending on the problem, it may be worthwhile to take a transform of both sides, as often the transformed quantity is easier to analyze. For example, if X and Y are independent random variables with densities f_X and f_Y, then the density of their sum is the convolution

$$f_{X+Y}(t) \ = \ (f_X * f_Y)(t) \ = \ \int_{-\infty}^{\infty} f_X(u)f_Y(t-u)du.$$

As the Fourier transform of a convolution is the pointwise product of the Fourier transforms, we have

$$\widehat{f_{X+Y}}(t) \ = \ \widehat{f_X}(t) \cdot \widehat{f_Y}(t);$$

thus the convolution integral has been replaced with standard multiplication (the integration has not vanished – we must take the Fourier transforms of f_X and f_Y, and then we must take the inverse Fourier transform to recover f_{X+Y}; however, this is still often progress). There are many other nice properties of the Fourier transform. For example, let p be a probability density. Then

$$\widehat{p}(y) \ = \ \int_{-\infty}^{\infty} p(x)e^{-2\pi ixy}dx.$$

Taking the derivative yields

$$\widehat{p}'(y) \ = \ \int_{-\infty}^{\infty} p(x) \cdot (-2\pi ix)e^{-2\pi ixy}dx,$$

and then setting $y = 0$ yields

$$\widehat{p}'(0) \ = \ -2\pi i \int_{-\infty}^{\infty} xp(x)dx \ = \ -2\pi i\mathbb{E}[X].$$

We note two important items: the Fourier transform of $-2\pi ix$ times the function p is the derivative of the Fourier transform of p, and the derivative of the Fourier transform at 0 is a simple multiple of the mean (and a generalization holds for higher moments).

(2) **Complex differentiability:** A function of a complex variable is said to be complex differentiable at z if

$$\lim_{h \to 0} \frac{f(z+h) - f(z)}{h}$$

exists as $h \neq 0$ tends to 0 along any path. Functions such as the polynomials $\sum_{k=0}^{n} c_k z^k$ are differentiable, while functions such as $\overline{z}^k$ are not (remember $\overline{z} = x - iy$ if $z = x + iy$). If a complex function is differentiable once, it is infinitely differentiable and it equals its Taylor series; this is remarkably different than real analysis (remember the function $f(x) = \exp(-1/x^2)$ for $x \neq 0$ and 0 for $x = 0$; this function is infinitely differentiable, but only equals its Taylor series at $x = 0$ (which is not impressive, as by definition all functions equal their Taylor series at the expansion point!).

(3) **Analytic continuation:** Given a function f defined in some subset of the complex plane, its analytic continuation is a new function which agrees with the old in the original region, but makes sense elsewhere. The standard example is the geometric series formula: $\sum_{n=0}^{\infty} x^n = 1/(1-x)$; the right hand side makes sense for all values of $x \neq 1$, while the left hand side is only defined if $|x| < 1$. This leads to the interpretation that $1 + 2 + 4 + 8 + 16 + \cdots = -1$!

(4) **Accumulation points:** Let f be a complex differentiable function defined on an open set U; assume $f(z_n) = 0$ for some sequence of points $\{z_n\}_{n=1}^{\infty}$ that has an accumulation point in U (i.e., there is some z^* such that a subsequence of the z_n's converge to z^*). Then a beautiful result from complex analysis says that f is identically zero! Again, this is very different than real analysis: the function $f(x) = x^3 \sin(1/x)$ for $x \neq 0$ and 0 for $x = 0$ is zero whenever $x = 1/n\pi$, and is zero at $x = 0$; however, clearly this function is not identically zero even near the origin (just consider $x = 2/n\pi$ for n odd). In probability, this result is used to study the moment problem, namely, how many moments are needed to uniquely determine a probability density.

(5) **Poisson summation:** for nice functions, $\sum f(n) = \sum \widehat{f}(n)$. Often this allows us to replace a long sum of slowly decaying terms with a short sum of rapidly decaying terms. We used this in obtaining very good estimates on the probability of being far from the mean for normal random variables, as well as proving the functional equation of the Riemann zeta function.

(6) A nice function can be uniformly approximated by a trigonometric polynomial (Fejer's theorem). One great use of this is in $n^k\alpha \bmod 1$, as trig functions are particularly nice to work with.

7. PROBABLITY THEORY (MATH 341)

For probability, many of the results from Multivariable Calculus (see §2) and Real Analysis (see §4) are useful.

7.1. **Pavlovian Responses.**

(1) **Logarithms:** If you see a `product`, you should have an immediate, Pavlovian response: `take a logarithm`! We don't have classes on products, but we do have classes on sums. The logarithm of a product is the sum of the logarithms; frequently this makes the problem a lot more amenable. Great examples include attacking the Moment Generating Functions in the proof of the Central Limit Theorem, and deriving estimates like Stirling's formula for $n!$.

(2) **Double Counting:** It is very easy to double count; avoid that temptation! Try to do a problem multiple ways. If you get the same answer that's reassuring (but not a proof that you're right). Think hard about whether or not order matters, and whether or not you've introduced order.

(3) **Generating Functions:** If you have a sequence that is important to you that you want to understand better, build a generating function. Unfortunately there are many different ways to build generating functions, but you may be able to find one that gives valuable information. The idea is to combine local information cleverly and get a global object that has a nice closed form, and allows you to extract information about the local data. A great example is Binet's formula (using the generating function of the Fibonacci numbers to get a closed form solution for the n^{th} Fibonacci number). We start by saying $g(x) = \sum_{n=0}^{\infty} F_n x^n$, and from this get a closed form expression for F_n (here $F_{n+1} = F_n + F_{n-1}$ and $F_0 = 0, F_1 = 1$). Other good generating functions are the Riemann zeta function, where for $\mathrm{Re}(s) > 1$ we have

$$\zeta(s) \ := \ \sum_{n=1}^{\infty} \frac{1}{n^s};$$

it turns out this equals the product over all primes p of $(1 - p^{-s})^{-1}$, and thus we can extract information about the primes by understanding this sum over the integers. There are many other examples of the power of generating functions.

7.2. **Combinatorics.**

(1) **Combinatorics:** There are several items to remember for combinatorial problems. The first is to be careful and avoid double counting. The second is that frequently a difficult sum can be interpreted two different ways; one of the interpretations is what we want, while the other is something we can do. We have seen many examples of this. One is that

$$\sum_{k=0}^{n} \binom{n}{k}^2 \ = \ \sum_{k=0}^{n} \binom{n}{k}\binom{n}{n-k}$$

is the middle coefficient of $(x+y)^{2n}$, and thus equals $\binom{2n}{n}$. To see this, note

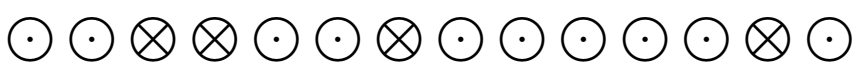
$$\begin{aligned}\sum_{\ell=0}^{2n}\binom{2n}{\ell}x^{\ell}y^{2n-\ell} &= (x+y)^{2n}\\ &= (x+y)^{n}(x+y)^{n}\\ &= \sum_{i=0}^{n}\binom{n}{i}x^{i}y^{n-i}\cdot\sum_{j=0}^{n}\binom{n}{n-j}x^{n-j}y^{j}.\end{aligned}$$

As the two expressions are equal, the coefficients of $x^a y^{2n-a}$ on the left must equal that on the right. Let's look at the middle coefficient, $x^n y^n$. To get that on the left, we just take $\ell = n$ and get $\binom{2n}{n}x^n y^n$. On the right, we need to choose i and j so that $i + n - j = n$, or $i = j$, giving us $\sum_{k=0}^{n}\binom{n}{k}x^k y^{n-k}\cdot\binom{n}{n-k}x^{n-k}y^k = \sum_{k=0}^{n}\binom{n}{k}\binom{n}{n-k}x^n y^n$, proving the claim.

We can also see this by telling a story. Imagine there are $2n$ people, n want to be mathematicians and n want to be physicists. How many ways can we choose n people when order does not matter? The answer is just the definition of the binomial coefficient, or $\binom{2n}{n}$. We can look at this another way: when we form our group, it must have *some* number of mathematicians. Let's say there are k mathematicians; note k can be any number from 0 to n. Then the number of physicists must be $n - k$. How many ways are there to form a group (order does not matter) with exactly k of the n mathematicians and exactly $n - k$ of the n physicists? That's just $\binom{n}{k}\binom{n}{n-k}$. Summing over k, this count must equal $\binom{2n}{n}$, and we regain the formula.

(2) **'Auxiliary lines':** In geometry, one frequently encounters proofs where the authors add an auxiliary line not originally in the picture; once the line is added things are clear, but it is often a bit of a mystery as to how someone would think of adding a line in that place. In combinatorics we have an analogue of this. Consider the classic cookie problem: we wish to divide 10 identical cookies among 5 distinct people. One simple way to do this is to imagine we have 14 ($14 = 10 + 5 - 1$) cookies, and eat 4 of them. This partitions the remaining cookies into 5 sets, with the first set going to the first person and so on.

For example, if we have 10 cookies and 5 people, say we choose cookies 3, 4, 7 and 13 of the $10 + 5 - 1$ cookies:



This corresponds to person 1 receiving two cookies, person 2 receiving zero, person 3 receiving two, person 4 receiving five and person 5 receiving one cookie.

This implies that the answer to our problem is $\binom{10+5-1}{5-1}$, or in general $\binom{C+P-1}{P-1}$.

(3) **Find an interpretation:** Consider the following sum: $\sum_{c=0}^{C}\binom{c+P-1}{P-1}$. By the arguments above, we are summing the number of ways of dividing c cookies among P people for $c \in \{0,\ldots,C\}$ (or we divide C cookies among P people, but we do not assume each cookie is given). A nice way to solve this is to imagine that there is a $P+1^{\text{st}}$ person who receives $C - c$ cookies, in which case this sum is now the same as counting the number of ways of dividing C cookies among $P + 1$ people where each cookie must be assigned to a person, or $\binom{C+P}{P}$. (See also the 'tell a story' entry in §7.3 and the 'convolution' entry in

§7.4.)

(4) **Inclusion - Exclusion Principle:** Suppose $A_1, A_2, \ldots, A_n$ is a collection of sets. Then the *Inclusion-Exclusion Principle* asserts that

$$\left|\bigcup_{i=1}^{n} A_i\right| \;=\; \sum_i |A_i| - \sum_{i,j} |A_i \cap A_j| + \sum_{i,j,k} |A_i \cap A_j \cap A_k| - \cdots .$$

This has many uses for counting probabilities. We used it to determine the probability of a generic integer is square-free, as well as the probability a random permutation of $\{1, \ldots, n\}$ returns at least one element to its initial location.

(5) **Binomial Theorem:** We have

$$(x+y)^n \;=\; \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k} \;=\; \sum_{k=0}^{n} \binom{n}{k} x^{n-k} y^k;$$

in probability we usually take $x = p$ and $y = 1 - p$. The coefficients $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ have the interpretation as counting the number of ways of choosing k objects from n when order does not matter. A better definition of this coefficient is

$$\binom{n}{k} \;=\; \frac{n(n-1)\cdots(n-(k-1))}{k(k-1)\cdots 1}.$$

The reason this definition is superior is that $\binom{3}{5}$ makes sense with this definition, and is just zero. One can easily show $\binom{n}{k} = 0$ whenever $k > n$, which makes sense with our combinatorial interpretation: there is no way to choose k objects from n when $n < k$, regardless of whether or not order matters.

7.3. General Techniques of Probability.

(1) **Differentiating Identities:** Equalities are the bread and butter of mathematics; differentiating identities allows us to generate infinitely many more from one, which is a very good deal! For example, consider the identity

$$(p+q)^n \;=\; \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k}. \tag{7.1}$$

Applying the operator $p\frac{d}{dp}$ to both sides we find

$$p \cdot n(p+q)^{n-1} \;=\; \sum_{k=0}^{n} k\binom{n}{k} p^k q^{n-k}. \tag{7.2}$$

Setting $q = 1 - p$ yields the mean of a binomial random variable:

$$np \;=\; \sum_{k=0}^{n} k\binom{n}{k} p^k (1-p)^{n-k}. \tag{7.3}$$

It is very important that initially p and q are distinct, free variables, and only at the end do we set $q = 1 - p$. Another example is differentiating $\sum_{n=0}^{\infty} x^n = 1/(1-x)$ by applying the operator $x\frac{d}{dx}$ gives $\sum_{n=0}^{\infty} nx^n \;=\; x/(1-x)^2$. While we can prove the $2m^{\text{th}}$ moment of the

standard normal is $(2m-1)!!$ by induction, we can also do this with differentiating identities.

(2) **Law of Total Probability:** This is perhaps one of the most useful observations: $\text{Prob}(A^c) = 1 - \text{Prob}(A)$, where A^c is the complementary event. It is frequently easier to compute the probability that something does not happen than the probability it does. Standard examples include hands of bridge or other card games. The Birthday Problem is a great example: assuming each day of the year is equally likely to be someone's birthday, all people in the room have independent birthdays (i.e., no identical twins invited), and no one is ever born on February 29th (so all years have 365 days), how many people do we need before there is at least a 50% chance that at least two share a birthday? If we look at all the ways people could share birthdays, it's a computational nightmare! We could have only two share a birthday, or maybe three, or maybe two pairs of people sharing birthdays.... If instead we look at the probability that n people have n distinct birthdays, that's much simpler. It's just

$$\frac{365-0}{365}\,\frac{365-1}{365}\,\cdots\,\frac{365-(n-1)}{365};$$

now we just take 1 minus this to get the probability that at least two people in n share a birthday.

(3) **Fundamental Theorem of Calculus (cumulative distribution functions and densities):** One of the most important uses of the Fundament Theorem of Calculus is the relationship between the cumulative distribution function F_X of a random variable X and its density f_X. We have

$$F_X(x) \;=\; \text{Prob}(X \le x) \;=\; \int_{-\infty}^{x} f_X(t)dt.$$

In particular, the Fundamental Theorem of Calculus implies that $F_X'(x) = f_X(x)$. This means that if we know the cumulative distribution function, we can essentially deduce the density. For example, let X have the standard exponential density (so $f_X(x) = e^{-x}$ for $x \ge 0$ and 0 otherwise) and set $Y = X^2$. Then for $y \ge 0$ we have

$$F_Y(y) \;=\; \text{Prob}(Y \le y) \;=\; \text{Prob}(X^2 \le y) \;=\; \text{Prob}(X \le \sqrt{y}) \;=\; F_X(\sqrt{y}).$$

We now differentiate, using the Fundamental Theorem of Calculus and the Chain Rule, and find that for $y \ge 0$

$$f_Y(y) \;=\; F_X'(\sqrt{y}) \cdot \frac{d}{dy}\left(\sqrt{y}\right) \;=\; f_x(\sqrt{y}) \cdot \frac{1}{2\sqrt{y}} \;=\; \frac{e^{-\sqrt{y}}}{2\sqrt{y}}.$$

(4) **Binary (or indicator) random variables:** For many problems, it is convenient to define a random variable to be 1 if the event of interest happens and 0 otherwise. This frequently allows us to reduce a complicated problem to many simpler problems. For example, consider a binomial process with parameters n and p. We may view this as flipping a coin with probability p of heads a total of n times, and recording the number of heads. We may let $X_i = 1$ if the i^{th} toss is heads and 0 otherwise; then the total number of heads is $X = X_1 + \cdots + X_n$. In other words, we have represented a binomial random variable with parameters n and p as a sum of n independent Bernoulli random variables. This facilitates calculating quantities such as the mean or variance, as we now have $\mathbb{E}[X] = n\mathbb{E}[X_i] = np$

and $\mathrm{Var}(X) = n\mathrm{Var}(X_i) = np(1-p)$. Explicitly, to compute the mean we need to evaluate $\mathbb{E}[X_i] = 1 \cdot p + 0 \cdot (1-p)$ and then multiply by n; this is significantly easier than directly evaluating the mean of the binomial random variable, which requires us to determine $\sum_{k=0}^{n} k \cdot \binom{n}{k} p^k (1-p)^{n-k}$.

(5) **Linearity of Expectation:** One of the worst complications in probability is that random variables might not be independent. This greatly complicates the analysis in a variety of cases; however, if all we care about is the expected value, these difficulties can vanish! The reason is that the expected values of a sum is the sum of the expected values; explicitly, if $X = X_1 + \cdots + X_n$ then $\mathbb{E}[X] = \mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n]$. One great example of this was in the coupon or prize problem. Imagine we have c different prizes, and each day we are randomly given one and only of the c prizes. We assume the choice of prize is independent of what we have, with each prize being chosen with probability $1/c$. How long will it take to have one of each prize? If we let X_i denote the random variable which is how long we must wait, given $i-1$ prizes, until we obtain the next new prize, then X_i is a geometric random variable with parameter $p_i = 1 - \frac{i-1}{c}$ and expected value $\frac{1}{p_i} = \frac{c}{c-(i-1)}$. Thus the expected number of days we must wait until we have one of each prize is simply

$$\mathbb{E}[X] \;=\; \sum_{i=1}^{c-1} \mathbb{E}[X_i] \;=\; \sum_{i=1}^{c-1} \frac{c}{c-(i-1)} \;=\; c\sum_{i=1}^{c} \frac{1}{i} \;=\; cH_c,$$

where $H_c = 1/1 + 1/2 + \cdots + 1/c$ is the c^{th} harmonic number (and $H_c \approx \log c$ for c large). Note we do not need to consider elaborate combinations or how the prizes are awarded. Of course, if we want to compute the variance or the median, it's a different story and we can't just use linearity of expectation.

(6) **Bring it Over:** We have seen two different applications of this method. One is in evaluating integrals. Let I be a complicated integral. What often happens is that, after some number of integration by parts, we obtain an expression of the form $I = a + bI$; so long as $b \neq 1$ we can rewrite this as $(1-b)I = a$ and then solve for I ($I = \frac{a}{1-b}$). This frequently occurs for integrals involving sines and cosines, as two derivatives (or integrals) basically returns us to our starting point. We also saw applications of this in memoryless games, to be described below.

(7) **Memoryless games / processes:** There are many situations where to analyze future behavior, we do not need to know how we got to a given state or configuration, but rather just what the current game state is. A terrific example is playing basketball, with the first person to make a basket winning. Say A shoots first and always gets a basket with probability p, and B shoots second and always makes a basket with probability q. A and B keep shooting, A then B then A then B and so on, until someone makes a basket. What is the probability A wins? The long was is to note that the probability A wins on her n^{th} shot is $((1-p)(1-q))^{n-1}\, p$, and thus

$$\mathrm{Prob}(A \text{ wins}) \;=\; \sum_{n=0}^{\infty} ((1-p)(1-q))^{n-1}\, p;$$

while we can evaluate this with the geometric series, there is an easier way. How can A win? She can win by making her first basket, which happens with probability p. If she

misses, then to win she needs B to miss as well. At this point, it is A's turn to shoot again, and it is as if we've just started the game. It does not matter that both have missed! Thus

$$\text{Prob}(A \text{ wins}) \;=\; p + (1-p)(1-q)\text{Prob}(A \text{ wins}).$$

Note this is exactly the set-up for using 'Bring it over', and we find

$$\text{Prob}(A \text{ wins}) \;=\; \frac{p}{1-(1-p)(1-q)};$$

in fact, we can use this to provide a proof of the geometric series formula! The key idea here is that once both miss, it is as if we've just started the game. This is a very fruitful way of looking at many problems.

(8) **Standardization:** Given a random variable X with finite mean and variance, it is almost always a good idea to consider the standardized random variable $Y = (X - \mathbb{E}[X])/\text{StDev}(X)$, especially if X is a sum of independent random variables. The reason is that Y now has mean 0 and variance 1, and this sets us up to compare quantities on the same scale. Equivalently, when we discuss the Central Limit Theorem everything will converge to the same distribution, a standard normal. We thus will only need to tabulate the probabilities for one normal, and not a plethora or even an infinitude. The situation is similar to logarithm tables. We only need to know logarithms in one base to know them in all, as the Change of Base formula gives $\log_c x = \log_b x / \log_b c$ (and thus if we know logarithms in base b, we know then in base c).

(9) **Tell a story:** One of our exam questions was whether or not $f(n) = \binom{n+k-1}{n}(1-p)^n p^k$ for $n \in \{0, 1, 2, \dots\}$, $p \in (0,1)$ is a probability mass function. One way to approach a problem like this is to try and tell a story. How should we interpret the factors? Well, let's make p the probability of getting a head when we toss a coin, or we could let it denote the probability of a success. Then $(1-p)^n p^k$ is the probability of a string with exactly n failures and k successes. There are $\binom{n}{k}$ ways to choose which n of $n+k$ places to be the failures; however, we have $\binom{n+k-1}{n}$. What's going on? The difference is that we are not considering all possible strings, but only strings where the *last* event is a success. Thus we must have exactly n failures (or exactly $k-1$ successes) in the first $n+k-1$ tosses followed by a success on trial $n+k$. By finding a story like this, we know it is a probability mass function; it is possible to directly sum this, but that is significantly harder. (See also the 'find an interpretation' entry in §7.2 and the 'convolution' entry in §7.4.)

(10) **Probabilistic Models:** We can often gain intuition about complex but deterministic phenomena by employing a random model. For example, the Prime Number Theorem tells us that there are about $x/\log x$ primes at most x, leading to the estimation that any n is prime with probability about $1/\log n$ (this is known as the Cramer model). Using this, we can estimate various number theoretic quantities. For example, let X_n be a random binary indicator variable which is 1 with probability $\frac{1}{\log n}$ and 0 with probability $1 - \frac{1}{\log n}$. If we want to estimate how many numbers up to x start a twin prime pair (i.e., n and $n+2$ are both prime) then the answer would be given by the random variable $X = X_2X_4 + X_3X_5 + \cdots + X_{n-2}X_n$.

As everything is independent and $\mathbb{E}[X_k] = \frac{1}{\log k}$, we have

$$\mathbb{E}[X] \;=\; \sum_{k=2}^{n-2} \mathbb{E}[X_k]\mathbb{E}[X_{k+2}] \;=\; \sum_{k=2}^{n-2} \frac{1}{\log(k)\log(k+2)} \;\approx\; \int_2^{n-2} \frac{dt}{\log^2 t} \;\approx\; \frac{x}{\log^2 x}.$$

The actual (conjectured!) answer is about $C_2 x/\log^2 x$, where

$$C_2 \;=\; \prod_{\substack{p\geq 3\\ p \text{ prime}}} \frac{p(p-2)}{(p-1)^2} \;\approx\; .66016.$$

What's important is to note that the simple heuristic *did* capture the correct x dependence, namely a constant times $x/\log^2 x$. Of course, one must be very careful about how far one pushes and trusts these models. For example, it would predict there are about $C_3 x/\log^3 x$ prime triples $(n, n+2, n+4)$ up to x for some non-zero C_3, whereas in actuality there is only the triple $(3,5,7)$! The problem is this model misses arithmetic, and in any three consecutive odd numbers exactly one of them is divisible by 3.

(11) **Simplifying sums:** Often we encounter a sum which is related to a standard sum; this is particularly true in trying to evaluate moment generation functions. Some of the more common (and important) identities are

$$\begin{aligned}
e^x &= 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots \;=\; \sum_{n=0}^{\infty} \frac{x^n}{n!}\\
\frac{1}{1-x} &= 1 + x + x^2 + x^3 + \cdots \;=\; \sum_{n=0}^{\infty} x^n\\
\frac{1}{(1-x)^2} &= 1 + 2x + 3x^3 + 4x^3 \;=\; \sum_{n=0}^{\infty} \binom{n}{1} x^{n-1}\\
\frac{1}{(1-x)^k} &= \sum_{n=0}^{\infty} \binom{n}{k} x^{n-k}\\
(x+y)^n &= x^n + nx^{n-1}y + \frac{n(n-1)}{2}x^{n-2}y^2\\
&\qquad = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k} \;=\; \sum_{k=0}^{n} \binom{n}{k} x^{n-k} y^k.
\end{aligned}$$

The goal is to 'see' a complicated expression is one of the above (for a special choice of x). For example, let X be a Poisson with parameter λ; thus $f_X(n) = x\lambda^n e^{-n}/n!$ if $n \in \{0, 1, 2, \dots\}$ and 0 otherwise. Then

$$M_X(t) \;=\; \mathbb{E}[e^{tX}] \;=\; \sum_{n=0}^{\infty} e^{tn} \cdot \frac{\lambda^n e^{-\lambda}}{n!}.$$

Fortunately, this looks like one of the expressions above, namely the one for e^x. Rearranging a bit gives

$$M_X(t) \;=\; e^{-\lambda}\sum_{n=0}^{\infty}\frac{(\lambda e^t)^n}{n!} \;=\; e^{-\lambda}\cdot\exp\left(\lambda e^t\right) \;=\; \exp\left(\lambda e^t-\lambda\right).$$

7.4. **Moments.**

(1) **Convolution:** Let X and Y be independent random variables with densities f_X and f_Y. Then the density of $X+Y$ is

$$f_{X+Y}(u) \;=\; (f_X * f_Y)(u) \;:=\; \int_{-\infty}^{\infty} f_X(u)f_Y(t-u)du;$$

we call $f_X * f_Y$ the convolution of X and Y. While we can prove by brute force that $f_X * f_Y = f_Y * f_X$, a faster interpretation is obtained by noting that since addition is commutative, $X+Y = Y+X$ and hence $f_{X+Y} = f_{Y+X}$, which implies convolution is commutative. Convolutions give us a handle on the density for sums of independent random variables, and is a key ingredient in the proof of the Central Limit Theorem.

(2) **Generating Functions:** Given a sequence $\{a_n\}_{n=0}^{\infty}$, we define its generating function by

$$G_a(s) \;=\; \sum_{n=0}^{\infty} a_n s^n$$

for all s where the sum converges. For discrete random variables that take on values at the non-negative integers, an excellent choice is to take $a_n = \mathrm{Prob}(X=n)$, and the result is called the generating function of the random variable X. Using convolutions, we find that if X_1 and X_2 be *independent* discrete random variables taking on non-negative integer values, with corresponding probability generating functions $G_{X_1}(s)$ and $G_{X_2}(s)$, then $G_{X_1+X_2}(s) = G_{X_1}(s)G_{X_2}(s)$.

(3) **Moment Generating Functions:** For many probability problems, the moment generating function $M_X(t)$ is more convenient to study than the generating function. It is defined by $M_X(t) = \mathbb{E}[e^{tX}]$, which implies (if everything converges!) that

$$M_X(t) \;=\; 1+\mu_1' t+\frac{\mu_2' t^2}{2!}+\frac{\mu_3' t^3}{3!}+\cdots,$$

where $\mu_k' \;=\; d^k M_X(t)/dt^k\Big|_{t=0}$ is the k^{th} moment of X. Key properties of the moment generating function are: (i) Let α and β be constants. Then

$$M_{\alpha X+\beta}(t) \;=\; e^{\beta t}M_X(\alpha t).$$

(ii) if $X_1,\ldots,X_N$ are independent random variables with moment generating functions $M_{X_i}(t)$ which converge for $|t|<\delta$, then

$$M_{X_1+\cdots+X_N}(t) = M_{X_1}(t)M_{X_2}(t)\cdots M_{X_N}(t).$$

If the random variables all have the same moment generating function $M_X(t)$, then the right hand side becomes $M_X(t)^N$. Unfortunately the moment generating function does not always exist in a neighborhood of the origin (this can be seen by considering the Cauchy

distribution); this is rectified by studying the characteristic function, $\mathbb{E}[e^{itX}]$, which is essentially the Fourier transform of the density (that is $\mathbb{E}[e^{-2\pi itX}]$).

(4) **Moment Problem:** When does a sequence of moments uniquely determine a probability density? If our distribution is discrete and takes on only finitely many (for definiteness, say N) values, then only finitely many moments are needed. If the density is continuous, however, infinitely many might not be enough. Consider

$$\begin{aligned} f_1(x) &= \frac{1}{\sqrt{2\pi x^2}}\, e^{-(\log^2 x)/2} \\ f_2(x) &= f_1(x)\left[1+\sin(2\pi\log x)\right]. \end{aligned}$$

These two densities have the same integral moments (their k^{th} moments are $e^{k^2/2}$ for k a non-negative integer); while they also have the same half-integral moments, all other moments differ (thus there is no sequence of moments where they agree which has an accumulation point; see §6). Thus it is possible for two densities to have the same integral moments but differ.

7.5. **Approximations and Estimations.**

(1) **Cauchy-Schwarz inequality:** For complex-valued functions f and g,

$$\int_0^1 |f(x)g(x)|dx \;\le\; \left(\int_0^1 |f(x)|^2dx\right)^{\frac{1}{2}} \cdot \left(\int_0^1 |g(x)|^2dx\right)^{\frac{1}{2}}.$$

One of my favorite applications of this was proving the absolute value of the covariance of X and Y is at most the product of the square-roots of the variances. The key step in the proof was writing the joint density $f_{X,Y}(x,y)$ as $\sqrt{f_{X,Y}(x,y)}\cdot\sqrt{f_{X,Y}(x,y)}$ and putting one factor with $|x-\mu_X|$ and one with $|y-\mu_Y|$. The reason we do this is we cannot directly integrate x^2 or $|x-\mu_X|^2$; we need to hit it with a probability density in order to have a chance of getting a finite value. This explains why we write the density as a product of its square root with its square root; it allows us to use Cauchy-Schwarz.

(2) **Stirling's Formula:** Almost any combinatorial problem involves factorials, either directly or through binomial coefficients. It is essential to be able to estimate $n!$ for large n. Stirling's formula says

$$n! \;=\; n^n e^{-n}\sqrt{2\pi n}\left(1+\frac{1}{12n}+\frac{1}{288n^2}-\frac{139}{51840n^3}+\cdots\right);$$

thus for n large, $n! \approx (n/e)^2\sqrt{2\pi n}$. There are many ways to prove this, the most common being complex analysis or stationary phase. We can get a ballpark estimate by 'summifying'. We have $n! = \exp(\log n!)$, and

$$\log n! = \sum_{k=1}^{n}\log k \;\approx\; \int_1^n \log t dt.$$

As the anti-derivative of $\log t$ is $t\log t$, we find $\log n! \approx n\log n - n$, so $n! \approx e^{n\log n - n} = n^n e^{-n}$, which is off by a factor of $\sqrt{2\pi n}$ (while this is a large number, it is small relative to $n^n e^{-e}$. If we wanted, using the integral test and a better job of estimate upper and lower

sums (the Euler-Maclaurin formula), we could get a better approximation for $n!$.

(3) **Chebyshev's Theorem:** Chebyshev's theorem (or inequality) is a mixed blessing; it is terrific in the sense that it works for any density that has finite mean and variance; however, in many applications its estimates are far from the truth. The reason is that it works for *all* such densities, and thus cannot exploit any specific properties of the density to get decay. (This is similar to the difference between using Divide and Conquer or Newton's Method to find a zero of a function; Newton's method is magnitudes faster because it assumes more about the function, namely differentiability, and thus it exploits that to get better estimates.) Chebyshev's theorem states

$$\text{Prob}(|X - \mu| \geq k\sigma) \ \leq \ \frac{1}{k^2}.$$

Note the event $|X - \mu| \geq k\sigma$ is a very natural event to consider: we are seeing how far X is from its expected value, and measuring this difference in terms of the natural units, the standard deviation. The assumptions for Chebyshev's theorem are a little weaker than those for the Central Limit Theorem, and there are situations where crude bounds suffice (for example, some of the problems we studied in additive number theory).

(4) **The Central Limit Theorem:** The Central Limit Theorem (CLT) states that if $X_1, \dots, X_n$ are independent, identically distributed random variables with mean μ and variance σ^2, then in many instances we have

$$Z_n \ := \ \frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \ = \ \frac{\frac{X_1+\cdots+X_n}{n} - \mu}{\sigma/\sqrt{n}}$$

converges to having the standard normal distribution as $n \to \infty$. If the moment generating function exists in a neighborhood containing the origin, that suffices for the CLT to hold (though with additional work we the conclusion holds under weaker assumptions about the X_i's). In practice one often uses the normal approximation once $n \geq 30$. One application is to use the CLT to estimate sums of random variables. Another is for hypothesis testing; there key thresholds are that if Z has the standard normal distribution, the $\text{Prob}(|Z| \leq 1) \approx 68.3\%$, $\text{Prob}(|Z| \leq 1.96) \approx 95.0\%$ and $\text{Prob}(|Z| \leq 2.575) \approx 99.0\%$.

(5) **Taylor Series:** See the section from Calculus I and II. For us, particularly important Taylor series are

$$\begin{aligned}
\log(1+x) &= x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots \\
\log(1-x) &= -\left(x + \frac{x^2}{2} + \frac{x^3}{3} + \frac{x^4}{4} + \cdots\right) \\
e^x &= 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots = \lim_{n\to\infty}\left(1 + \frac{x}{n}\right)^n \\
e^{-x} &= 1 - x + \frac{x^2}{2!} - \frac{x^3}{3!} + \cdots = \lim_{n\to\infty}\left(1 - \frac{x}{n}\right)^n \\
\frac{1}{1-x} &= 1 + x + x^2 + x^3 + \cdots.
\end{aligned}$$

7.6. **Applications.**

(1) **Benford's Law:**
(2) **Additive Number Theory:**
(3) **Economics:**
(4) **Gambling:**
(5) **Sabermetrics:**
(6) **Monte Carlo Integration:**

8. NUMBER THEORY (MATH 308 AND 406; MATH 238 AT SMITH)

Many of the techniques of number theory are covered in great detail in other sections (especially probability and complex analysis), and thus this section will be brief. See also the general techniques in §10.

(1) **Elementary functions:** $e^{i\theta} = \cos(\theta) + i\sin(\theta)$, $\phi(q)$ is the number of positive integers at most q that are relatively prime to q, ...

(2) **The Prime Number Theorem or the Siegel-Walfisz Theorem:** we used these frequently in analyzing prime sums as these yield unconditional estimates. Sometimes we can get by with Chebyshev's theorem, which says there are A, B with $0 < A < 1 < B < \infty$ such that $Ax/\log x \le \pi(x) \le Bx/\log x$ for all x sufficiently large (with $\pi(x)$ the number of primes). For example, we have good enough values on A, B to prove Bertrand's postulate (always a prime in $[n, 2n)$ as well as the sum of the reciprocals of the primes diverge. We also have Dirichlet's theorem for Primes in Arithmetic Progression: if a and b are relatively prime, then there are infinitely many primes congruent to a modulo b (in fact, each residue class, to first order, has the same number of primes: $\pi(x)/\phi(b)$, where ϕ is Euler's totient function, counting the number of numbers relatively prime and less than the argument).

(3) **Overcounting:** A nice application of Dirichlet's theorem (see above) is the proof that there is *at least one prime congruent to a modulo b*; it's interesting that the only way to prove this in general is to prove that there are infinitely many, and thus there must be at least one. This overcounting approach is often very useful.

(4) **Counting two different ways / telling a story:** Calculating something two different ways is one of the most important ideas in math. A good part of combinatorics is to note that there are two ways to compute something, one of which is easy and one of which is not. We then use our knowledge of the easy calculation to deduce the hard. For example, $\sum_{k=0}^{n} \binom{n}{k}^2 = \binom{2n}{n}$; the right side is easy to compute, the left side not so clear. Why are the two equal? It involves finding a story. Imagine we have $2n$ people, say n Red Sox fans and n Yankee fans. How many ways are there to form a group of n people from the $2n$ people, if order doesn't matter? One way is to note it is just $\binom{2n}{n}$. Another way is to look at how many Red Sox fans we have in our group of n. Let k be the number of Red Sox fans we choose; we may take any integer k from 0 to n. There are $\binom{n}{k}$ ways to choose k Red Sox fans from n, and thus $\binom{n}{n-k} = \binom{n}{k}$ ways to choose $n-k$ Yankee fans (if we want n people total, if we have k Red Sox fans we need $n-k$ Yankee fans, and of course no one likes both teams!). Thus the number of ways to form our group is just $\sum_{k=0}^{n} \binom{n}{k}\binom{n}{n-k} = \sum_{k=0}^{n} \binom{n}{k}^2$, which must equal $\binom{2n}{n}$. See how easy it is to evaluate the sum *if* we can tell the right story!

(5) **Partial summation:** allows us to pass from one known sum to another. For example, knowing $\sum_{p\le x} \log p \sim x$ we can then evaluate $\sum_{p\le x} 1$.

(6) **Dirichlet's Pidgeonhole principle:** this was very useful in studying $n^k\alpha \bmod 1$, and gave us very good rational approximations to irrationals. This says that if we must put $n+1$ objects into n boxes, at least one box must get at least 2 elements. More generally, at least

one box must get at least the average number of elements per box.

(7) **Unique factorization of the integers:** this was crucial in proving $\zeta(s) = \sum 1/n^s$ also equals $\prod_p (1-p^{-1})^{-1}$; as we know where the integers are, the hope is that we can use this knowledge to deduce information about the primes.

(8) **Logarithms:** Whenever we have a product, it's a good idea to take a logarithm and convert to a sum, frequently using Taylor series to further the analysis.

(9) **Multiplicative functions:** Often functions are multiplicative, so knowing the values at the primes and prime powers allow us to deduce the values everywhere.

(10) **Complex analysis:** Even though statements such as the Prime Number Theorem don't involve complex numbers, it is frequently useful to expand our view, as then powerful results are at our disposal.

(11) **Generating functions:** Frequently we can encode what we want to study into a function which we can then attack with a variety of methods. See the corresponding entries from the probability section. Here we've seen applications from the Riemann zeta function (relating the primes, which are hard to study, to the integers, whose distribution is well understood), to twin primes and Goldbach's problem. Another great example is Binet's formula (using the generating function of the Fibonacci numbers to get a closed form solution for the n^{th} Fibonacci number).

(12) **Dyadic decompositions:** It often helps to break the analysis over a large interval into analysis over many smaller intervals. We can write $[1, 2^k)$ as $[1, 2) \cup [2, 2^2) \cup \cdots \cup [2^{k-1}, 2^k)$. The advantage is that there is less variation in the function over these smaller intervals.

(13) **Efficiency of algorithms:** It's often not enough to be able to do things; we need to be able to do them quickly! We saw lots of instances of this, ranging from fast exponentiation for RSA to the Euclidean algorithm.

9. MATH 416: ADVANCED APPLIED LINEAR ALGEBRA

The main theme of this class is that just because a calculation can be done, this does not mean it can be done feasibly in a reasonable amount of time. Often a lot of time, effort and work is required to get a calculation down to something that can be done in the allotted time. Some things we found over the semester:

(1) **Dimensional analysis:** The best example was our dimensional analysis proof of the Pythagorean Theorem, but this technique permeates modern mathematics. Looking at units / dimensions can often shed light on critical behavior.

(2) **Canonical Form:** There are lots of advantages of putting something into canonical form. Perhaps the earliest example from your careers is the quadratic formula; we have removed thought, we can now just crank out answers from the quadratic formula. We put all linear programming problems into canonical form for a variety of reasons; a major one is that we know everything will be in a certain form, and it helps in future arguments to know that. This is related to standardization (big in probability). We put our constraints in a certain way, we make assumptions about our variables, et cetera.

(3) **Duality:** Often it is a lot easier to solve a related problem than the original. Our best example was the chess problem: putting down 5 queens so that 3 pawns are safe on a 5×5 board is the same as putting down 3 queens so that 5 pawns are safe, and then exchanging queens and pawns. Often the 'dual' problem is easier.

(4) **Optimization:** A lot of times the optimal values are on the 'boundary', and we can reduce the size of the space we must search. Also note that sometimes the best way to reach optimal behavior is not to do locally what's optimal. We saw examples of this in the 'do dogs know calculus' in that sometimes the best path for a dog in water to a stick in water is to run to land; even though locally this is not the best thing to do at that instant, the short term cost of going to land is amply paid back by a greater speed there. There were issues along these lines in the traffic presentation. We had the following: Hypothesis: if there is an additional stream that doesn't introduce conflict, is it efficient to add the stream?

(5) **Linearization:** Linear problems are typically easier than non-linear ones. We can often convert non-linear problems to linear problems, though at a cost of more variables. This is frequently a good idea, as it is often easier to study / analyze a lot of simpler problems and then combine these than to attack head on a complex problem.

(6) **Definitions and Special Cases:** We can define anything we want. Sadly most of the time our definitions won't be useful, but occasionally they help us make great progress. A terrific example is the notion of basic feasible (or basic optimal) solution. These are nice special cases of feasible (or optimal) solutions, but have additional nice properties that facilitate the analysis.

Another example of special cases comes from our random matrix theory unit. It's often easier to prove results for diagonalizable or triangularizable matrices, and then conjugate a general matrix to one of these special forms and then show the property we care about

survives this conjugation.

(7) **The Simplex Method's Key Idea:** I love the 'proof' of the simplex method: we have Phase I and Phase II, we show how to do Phase I given Phase II and then use Phase I to prove Phase II! It's not a circular loop as we only need Phase I in a simple case, where the existence of a basic feasible solution is clear. I find this similar to drawing auxiliary lines in geometry (the difficulty is knowing where to put these lines).

(8) **Efficiencies:** A lot of times there are significantly faster ways to do a problem then we might expect. Our experience can mislead us. Examples include fast exponentiation, Horner's algorithm (evaluating a polynomial), the Euclidean algorithm (finding greatest common divisors), factorization (no – just seeing if you're paying attention), the Strassen algorithm (for matrix multiplication), ... The list is long and distinguished. Just because you've always done something one way doesn't mean you should continue to do so in that way. Try to see if there's another approach. A great example is the Walmart versus Kmart story, where the entire class thought the Kmart expansion strategy seemed more sound.

(9) **Appropriate Definitions:** We saw two great examples of the importance of good definitions. The first was the Method of Least Squares. We chose to measure errors by squares rather than absolute values so that calculus would be available. Another example was from the Kepler Conjecture presentation, where a good scoring function was needed.

(10) **Method of Divine Inspiration:** It's great to know the answer! Of course, one cannot always count on this.

(11) **Generating functions:** see the entry in §7.1 of the probability section. We had some nice applications of this to the Catalan numbers in the random matrix theory unit.

(12) **Be wary of generalizations:** For real numbers x and y we have $e^x e^y = e^{x+y}$, but unless A and B are square matrices that commute, $e^A e^B \neq e^{A+B}$. From Mark Twain (one of my favorites): *The cat, having sat upon a hot stove lid, will not sit upon a hot stove lid again. But he won't sit upon a cold stove lid, either.* From each lesson in life extract just what you should, and no more.

(13) **Comparing the incomparable:** We have to compare objects, so we need to have them on similar scales. This was very important in multi-objective linear programming. We have to find a way to compare how much we value cost saving measures with how much we value taste. If you are unsure why, look at the picture of the optimal diet (Figure 1)! Another example comes from the environmental programming lecture, where we had to weigh the benefits of production against the costs of pollution.

(14) **Counting and accounting:** A good amount of the random matrix theory unit was simple counting arguments to show the number of configurations of a certain form times their contribution was negligible in the limit. Don't spend too much time analyzing something that is washed away in the limit.

FIGURE 1. Optimal Diet – it keeps you alive for less than a dollar a day!

(15) **Generalize in Stages:** Start with a problem you can do or a model you can solve, and slowly add in complications. A great example is the oil problem in the stochastic programming presentation, where we took a problem we could solve (the standard problem) and examined how to allow the parameters to vary. The goal was to return to something we could do.

(16) **Test special cases:** It's very important to test special cases to get a sense of the problem. This can range from building intuition to seeing whether or not your proposed solution is reasonable. For example, in the quadratic programming lecture we saw that if we allowed quadratic constraints then we could get all integer programming problems (to get binary, have $x(x-1) \leq 0$ and $x(x-1) \geq 0$, which forces $x(x-1) = 0$ or $x \in \{0,1\}$). Thus, if we could handle quadratic constraints we'd be able to do binary integer programming, which is conjectured to be very hard. Thus, when generalizing linear programming, it is unlikely to have an easy generalization to quadratic constraints, but possibly we can adjust the objective function.

10. General Techniques (for many classes)

These are techniques that appear in several different classes I've taught, over and over. The notes below are written from the point of view of a student who has taken these classes, and thus some of the passages below may be hard to follow / may refer to advanced material.

(1) **If all you have is a hammer, pretty soon every problem looks like a nail.** There are a lot of ways to interpret this. One way is to know your strengths. If you have a technique you are good at and other people either are not or are not familiar with it, perhaps you can successfully apply it in their fields to solve some of their problems (i.e., find the land of the screwdriver people!). The other interpretation is that whenever you encounter a problem try and recast it so that your hammer will work. Thus in one the problem comes to you, in the other you go searching for problems.

(2) **It is often easier to solve a lot of simple problems and then combine than it is to solve one complex problem.**

(3) **Being algebraically lazy:** Another common theme is that we try to do as little work as possible to get as good of an estimate as needed. For example, we computed the moment generating function of the standard normal by completing the square, and found $M_X(t) = \mathbb{E}[e^{tX}] = e^{t^2/2}$. Later we needed to Fourier transform of the standard normal; while we could attack the integral which arises, it is far easier to note the Fourier transform at y is the same as the moment generating function at $-2\pi iy$. While we need to use some results from complex analysis to justify this argument, we now get the Fourier transform.

(4) **Be wary of generalizations:** For real numbers x and y we have $e^x e^y = e^{x+y}$, but unless A and B are square matrices that commute, $e^A e^B \neq e^{A+B}$. From Mark Twain (one of my favorites): *The cat, having sat upon a hot stove lid, will not sit upon a hot stove lid again. But he won't sit upon a cold stove lid, either.* From each lesson in life extract just what you should, and no more.

(5) **Problem formulation and blinders:** We've also seen on a few problems how the way the problem is formulated can influence how one attempts to solve it. For example, recall the function $x^3 \sin(1/x)$. The oscillation is bounded by two cubics; however, if we just look at the part above the x-axis, the plot looks like a parabola. It is thus a good idea, if you're stuck, to try and think of alternative ways of looking at a problem. Other examples include the graph coloring problem from the HW (vertices are 2 through N and are connected if they share a divisor; the HW problem was to show the coloring number is at least 13, which can be done by looking at powers of 2, but it's actually at least 5000, from looking at even numbers) and the following (for each $n > 1$ finding an $m > 1$ such that nm only has 0s and 1s base 10; one proof is similar to the pidgeonhole problem of a subset of $\{a_1, \dots, a_n\}$ has a sum divisible by n). It is amazing how often one can get trapped at looking at a problem in a certain way; this is something to be aware of.

(6) **Choosing approaches.** Certain functions become natural choices in studying certain problems. For example, for $n^k\alpha \bmod 1$ we use the exponential function. The reason this is so useful is that $\exp(2\pi i n^k \alpha) = \exp(2\pi i (n^k\alpha \bmod 1))$. Thus we may drop the difficult

modulo 1 condition and sum more easily. Depending on the problem, different functions and expansions will be more useful than others. The ease at which the exponential function handles the modulo 1 condition suggests the usefulness of applying Fourier analysis.

(7) **Counting two different ways / telling a story:** Calculating something two different ways is one of the most important ideas in math. A good part of combinatorics is to note that there are two ways to compute something, one of which is easy and one of which is not. We then use our knowledge of the easy calculation to deduce the hard. For example, $\sum_{k=0}^{n} \binom{n}{k}^2 = \binom{2n}{n}$; the right side is easy to compute, the left side not so clear. Why are the two equal? It involves finding a story. Imagine we have $2n$ people, say n Red Sox fans and n Yankee fans. How many ways are there to form a group of n people from the $2n$ people, if order doesn't matter? One way is to note it is just $\binom{2n}{n}$. Another way is to look at how many Red Sox fans we have in our group of n. Let k be the number of Red Sox fans we choose; we may take any integer k from 0 to n. There are $\binom{n}{k}$ ways to choose k Red Sox fans from n, and thus $\binom{n}{n-k} = \binom{n}{k}$ ways to choose $n-k$ Yankee fans (if we want n people total, if we have k Red Sox fans we need $n-k$ Yankee fans, and of course no one likes both teams!). Thus the number of ways to form our group is just $\sum_{k=0}^{n} \binom{n}{k}\binom{n}{n-k} = \sum_{k=0}^{n} \binom{n}{k}^2$, which must equal $\binom{2n}{n}$. See how easy it is to evaluate the sum *if* we can tell the right story!

(8) **Adding zero / multiplying by one:** This is perhaps **the** most important technique to learn, though it is one of the hardest to master. The difficult part of these methods is figuring out how to 'do nothing' in an intelligent way. The first example you might remember is proving the product rule from calculus. Let $A(x) = f(x)g(x)$. Then

$$\begin{aligned}
A'(x) &= \lim_{h\to 0} \frac{A(x+h) - A(x)}{h} \\
&= \lim_{h\to 0} \frac{f(x+h)g(x+h) - f(x)g(x)}{h} \\
&= \lim_{h\to 0} \frac{f(x+h)g(x+h)\textbf{-f(x)g(x+h) + f(x)g(x+h)} - f(x)g(x)}{h} \\
&= \lim_{h\to 0} \left[\frac{f(x+h)g(x+h) - f(x)g(x+h)}{h} + \frac{f(x)g(x+h) - f(x)g(x)}{h}\right] \\
&= \lim_{h\to 0} \frac{f(x+h) - f(x)}{h} g(x+h) + \lim_{h\to 0} f(x) \frac{g(x+h) - g(x)}{h} \\
&= \lim_{h\to 0} \frac{f(x+h) - f(x)}{h} \lim_{h\to 0} g(x+h) + f(x) \lim_{h\to 0} \frac{g(x+h) - g(x)}{h} \\
&= f'(x)g(x) + f(x)g'(x).
\end{aligned}$$

My favorite example was probably in proving the multinomial distribution is a density.

(9) **Summifying or summification:** We frequently replace $\prod a_n$ with $\exp\left(\log \prod a_n\right)$, as this converts the product to a sum, and we have a much better understanding of sums. Probably the most important use was in proving the Central Limit Theorem, where we replaced

studying $\prod_i M_{X_i}(t)$ with studying $\sum_i \log M_{X_i}(t)$. We also used it to obtain an approximation for Stirling's formula, replacing $n!$ with $\sum_{\ell \le n} \log \ell$ (which we evaluated by using the integral test). We used this to provide a good lower bound for the singular series $\mathfrak{S}(N) = \prod_{p|N} \left(1 - \frac{1}{(p-1)^2}\right)$ in the Circle Method (writing odd numbers as the sum of three primes). We also used it to get a good lower bound for $\phi(q)$, which allowed us to see that $q/\log\log q \ll \phi(q) \ll q-1$. Basically, if you see a sum, you should have an immediate, Pavlovian response: take a logarithm! We don't have classes on products, but we do have classes on sums. The logarithm of a product is the sum of the logarithms; frequently this makes the problem a lot more amenable. Great examples include attacking the Moment Generating Functions in the proof of the Central Limit Theorem, and deriving estimates like Stirling's formula for $n!$.

(10) **L^2-norms:** in the Circle Method we had the generating function $F_N(x) = \sum_{p \le N} \log p \cdot \exp(2\pi i p x)$. We are able to get a very good bound for $\int_0^1 |F_N(x)|^2 dx$ as $|F_N(x)|^2 = F_N(x)F_N(-x)$, and the only terms that survive the integration are when we have reinforcement. More generally, it is often easy (or at least easier) to get reasonable estimates for quantities such as $\int |F(x)|^{2^k} dx$.

(11) **Removing conditions:** Whenever you have a theorem, you should always explore what happens if you remove a condition. Frequently (though not always) the claim no longer holds; sometimes the claim is still true but the proof is harder. Rarely, but it can happen, removing a condition causes you to look at a problem in a new light, and find a simpler proof.

(12) **Efficient algebra:** It is frequently worthwhile to think about whether or not we can approach a tedious algebra problem another way. Some examples from previous courses: to compute A^n for n large, diagonalize A if possible, say $A = S\Lambda S^{-1}$ with Λ the diagonal matrix of eigenvalues. Then $A^n = S\Lambda^n S^{-1}$, and Λ^n is readily computed. Another example is telescoping series, $(a_1 - a_0) + (a_2 - a_1) + \cdots + (a_n - a_{n-1}) = a_n - a_0$; this is a key ingredient in many proofs of the Fundamental Theorem of Calculus. Frequently in probability we combine these approaches with recognizing and exploiting an identity; for example, if we had to evaluate $\binom{n}{2}2^2 + \binom{n}{3}2^3 + \cdots + \binom{n}{n}2^n$, we might notice that this is almost the binomial expansion of $(1+2)^n$; it would be, but we're missing the first two terms. The solution is to add zero by adding and subtracting those terms, which gives

$$\begin{aligned} \binom{n}{2}2^2 + \binom{n}{3}2^3 + \cdots + \binom{n}{n}2^n &= \left(\sum_{k=0}^{n} \binom{n}{k} 1^{n-k}2^k\right) - \left(\binom{n}{0} + \binom{n}{1}2\right) \\ &= (1+2)^n - (n + n(n-1)) \;=\; 3^n - n^2; \end{aligned}$$

note we included the factor 1^{n-k} to make this match the standard binomial theorem expansion.

(13) **Illuminating algebra:** It is very easy to obtain complicated expressions involving the parameters of interest; while the answer is correct, the final product is not illuminating. It is worthwhile to see if the answer can be simplified. For example, consider the sabermetrics (baseball math) problem where we had Team X scores runs from a geometric distribution

with parameter p (in this case, $\text{Prob}(X = m) = (1-p)p^m$ for $m \in \{0, 1, 2, \dots\}$ and allows runs to Team Y with a geometric distribution with parameter q; we assume the two random variables are independent. The mean number of runs Team X scores is denoted RS, and equals $\text{RS} = \frac{p}{1-p}$ which implies $p = \frac{\text{RS}}{\text{RS}+1}$; we let RA denote the runs allowed, and $\text{RA} = \frac{q}{1-q}$ which implies $q = \frac{\text{RA}}{\text{AS}+1}$. After some algebra we found the probability Team X wins is

$$\frac{p(1-q)}{p(1-q)+q(1-p)}.$$

No one, however, thinks in terms of the decay probability from scoring m to scoring $m+1$ runs; we want a formula in terms of runs scored RS and runs allowed RA. Substituting for p and q yields

$$\frac{\left(1-\frac{\text{RA}}{1+\text{RA}}\right)\text{RS}}{(1+\text{RS})\left(\frac{\left(1-\frac{\text{RA}}{1+\text{RA}}\right)\text{RS}}{1+\text{RS}}+\frac{\text{RA}\left(1-\frac{\text{RS}}{1+\text{RS}}\right)}{1+\text{RA}}\right)},$$

a most unilluminating formula! With some work, we can simplify this to the nice answer we'll describe below; however, what is important about this problem (for us – major league baseball would beg to differ!) is not the result, but how to reach it efficiently. We know that $\frac{p}{1-p}$ is a nice expression, namely RS, and similarly for $\frac{q}{1-q}$. Thus we should take our expression and multiply by 1 in the form $(1/(1-p)(1-q)) \Big/ (1/(1-p)(1-q))$. Doing so yields

$$\frac{p(1-q)}{p(1-q)+q(1-p)} \cdot \frac{\frac{1}{(1-p)(1-q)}}{\frac{1}{(1-p)(1-q)}} = \frac{\frac{p}{1-p}}{\frac{p}{1-p}+\frac{q}{1-q}} = \frac{\text{RS}}{\text{RS}+\text{RA}}.$$

Note we obtain a very nice formula very quickly.

(14) **Numerical exploration:** When given a problem, one can frequently build intuition by running numerical experiments. For example, one of our problems concerned a person who made 40% of all their shots. We wanted to know the probability that the number of shots required to make 341 baskets was within 35 of the mean number of shots required. We came up with an answer by seeing that this was equivalent to the sum of 341 independent geometric random variables with parameter $p = .4$, and thus the Central Limit Theorem is applicable to estimate the probability.

To test our predictions, consider the person shooting until they get 341 baskets a staggering 10,000 times (see Figure 2). Note the numerical data is quite close to theory. If you can program in some environment, you can quickly gather numerical data to help elucidate the answer. The Mathematica code for this problem is:

```
num = 10000;
count = {};
prob = 0;
mean = 852.5;
For[n = 1, n <=  num, n++,
  {
   numfound = 0;
   counter = 0;
```

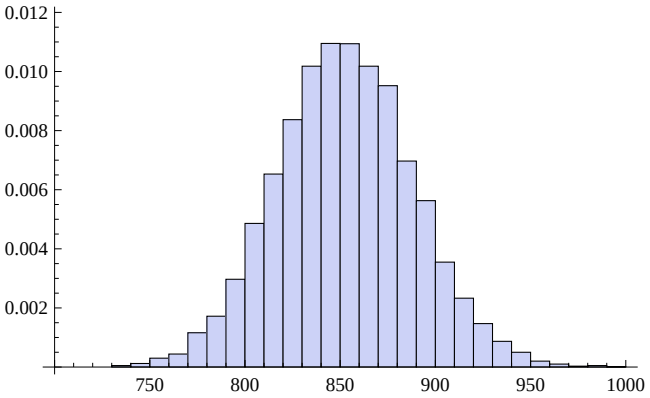


FIGURE 2. Histogram plot of number of shots to make 341 baskets given a 40% chance of making a shot. The data was obtained by playing the game 10,000 times and recording how long it took. The sample mean is 852.058 (which is quite close to the predicted 852.5), the sample standard deviation is 35.8092 (quite close to the predicted 35.7596), and 67.4% of the time the number of shots was within 35 of 852.5 (quite close to our prediction).

```
   While[numfound < 341,
    {
     counter = counter + 1;
     If[Random[] <= .4, numfound = numfound + 1];
     }];
   count = AppendTo[count, counter];
   If[Abs[counter - mean] <=  35, prob = prob + 1];
   }];
Histogram[count, Automatic, "Probability"]
Print[prob 100.0/num];
```

Of course, sometimes we are fortunate enough that, instead of settling for numerical answers, programs like Mathematica can find the exact answer. For example, consider the following difference equation, which arises in a problem related to a random walk with boundaries:

$$T_{i+1} \ = \ \frac{1}{p}T_i - \frac{1-p}{p}T_{i-1} - \frac{1}{p}.$$

Typing

```
Simplify[RSolve[{T[i] == p (T[i + 1] + 1) + (1 - p) (T[i - 1] + 1),
T[0] == 0,  T[M] == 0}, T[i], i]]
```

into Mathematica yields

$$T_i \ = \ \frac{i + M\left(\left(\frac{1-p}{p}\right)^i - 1\right) - i\left(\frac{1-p}{p}\right)^M}{\left(\left(\frac{1-p}{p}\right)^M - 1\right)(2p-1)}.$$

(15) **Test functions:** You should always consider testing the limits of a theorem, conjecture or intuition. Does it hold for the standard normal? For the Cauchy? How important is the finiteness of moments? Usually a result is false if you remove a condition; however, when you are trying to figure out what the conditions should be in a theorem, you're in a different mindset. In this case, it is worthwhile to play with various functions and see what happens.

(16) **Check for reasonableness:** Whenever we have a formula, it is a very good idea to check special cases to see if it is reasonable. For example, consider the sabermetrics formula from the previous point: if a team scores on average RS runs per game and allows on average RA per game (with RS and RA independent geometric random variables with respective means RS and RA), then its probability of winning is $\mathrm{RS}/(\mathrm{RS}+\mathrm{RA})$. Is this formula reasonable? There are many checks we can do. The first is that we always get a number between 0 and 1 (which is a must for a probability!). Further, if RS is zero or if RA tends to infinity than we have no chance of winning, exactly as we would expect. If we score on average more runs than we allow, our winning percentage is greater than 50%, while if we score and allow the same number on average than the winning percentage is 50%, again quite reasonable.

For another example, imagine we flip a fair coin with probability p of heads and $1-p$ of tails n times, and we ask how many runs (alterations between heads and tails) there are; for example, if the outcome were HHTTHTHTTTTTHTHHHH then there were 18 tosses, 9 heads and 9 tails and 9 runs, the shortest being a run of length 1 and the longest being a run of length 5. The expected number of runs is $1+(n-1)2p(1-p)$. Is this formula reasonable? Note that if $p=0$ or $p=1$ then because of the factor $p(1-p)$ the expected number of runs is 1; we should be shocked if this is not the case, as if the coin always lands on heads, how could there ever be an alteration? A little calculus shows that the maximum expected value is when $p=1/2$, which also seems reasonable. Finally, in the special case $p=1/2$ the expected number is essentially $n/2$; there are n tosses and each toss has a 50% chance of being different than the previous (and thus starting a run), so again our answer makes sense.

(17) **Check all conditions:** Whenever you want to use a theorem, make sure all the conditions are satisfied. For example, if you are summing the geometric series $1+x+x^2+x^3+\cdots$ then you better have $|x|<1$. If you are asked whether or not something is a probability distribution, it must satisfy both requirements (non-negative and sums to 1; it is not enough to just sum to one). If you want something to be a group, it must satisfy all four properties (closure, identity, associativity, inverse). Frequently some but not all of the conditions are met.

E-mail address: sjm1@williams.edu, Steven.Miller.MC.96@aya.yale.edu

DEPARTMENT OF MATHEMATICS AND STATISTICS, WILLIAMS COLLEGE, WILLIAMSTOWN, MA 01267