Booth Math Camp Lecture Notes

Karthik Srinivasan
September 6, 2022

Contents

Credit and Resources

Preliminary Notes for Booth Math Camp 2021. Please do not distribute. Send corrections to ks@chicagobooth.edu.

You may find the first year Dropbox as well as the collaborative latex editor Overleaf helpful. When trying to learn latex commands, I recommend detexify. To convert from a math expression in a pdf to latex code, you may find Mathpix to be a useful OCR service (100 uses per month are free for students). Additionally, though unrelated to Math Camp, you may want these links to Department seminars and Booth seminars.

This set of notes builds on a previous notes written by former Booth Math Camp TAs Jianfei Cao and Ali Goli. Ali Goli, Jianfei Cao, Malaina Brown, Jeffery Russell, and the 4^{th} year Booth PhD cohort all provided helpful advice for designing this course. Nothing but respect for my co-teacher, Walter.

The Real Analysis section of these notes is based on An Introduction to Real Analysis by John K. Hunter. The Probability section of these notes borrows heavily from

MITOpenCourseware's Fundamentals of Probability course. Some material is taken from Professor Joshua Wilde's math camp notes. I also learned from Mark Walker's 2021 math camp which is available in full on YouTube.

These notes also include links to Wikipedia, Khan Academy, and a variety of videos on YouTube all of which helped me better understand the material.

Warm Up

Logs/Exponents

Exponent Rules

$e = 2.71828...$

$$exp(a+b) = exp(a)exp(b)$$

$$exp(a)^b = exp(ab)$$

Log Rules

$$log(ab) = log(a) + log(b)$$

$$log(a^n) = nlog(a)$$

Context: Macro One standard production function is Cobb-Douglas, where output Y is modeled as a function of total factor productivity A, labor L, capital K, and labor and capital elasticities α and β with the equation $Y = AL^{\alpha}K^{\beta}$. When estimating this equation using a regression, it's common to take logs:

$$\begin{aligned} Y &= AL^{\alpha}K^{\beta} \\ log(Y) &= log(AL^{\alpha}K^{\beta}) \\ log(Y) &= log(A) + log(L^{\alpha}) + log(K^{\beta}) \\ log(Y) &= log(A) + \alpha log(L) + \beta log(K) \end{aligned}$$

Linear Algebra

Why should you care? Many complex problems can be approximated by linear systems of equations, and matrix algebra gives us a way to think about these systems. Additionally, most empirical work relies on linear regression techniques. Your data is represented in a matrix, and all proofs of the properties of how linear regressions work rely on matrix algebra.

Matrix Multiplication

Definition: The **dot product (inner product)** of two vectors a and b of length n is given by $\langle x, y \rangle = \sum_{i \in 1}^{n} a_i b_i$.

Matrix Multiplication: Let A be an $m \times n$ dimensional matrix and B be an $s \times r$ dimensional matrix. AB is defined (a valid operation) if the number of columns of A (n) is the same as the number of rows of B (s). The resulting matrix will have dimension $m \times r$. If we index the entries of the matrix AB by their column number i and row number j, then the element x_{ij} is given by the dot product of the i^{th} row of A and the j^{th} column of B.

If you need a refresher on how matrix multiplication works, please take a look at this Khan academy video.

Matrix Operation Properties: Recall that the commutative law does not hold for matrix multiplication: $AB \neq BA$ in general. However, the associative and distributive laws do hold (assuming that the multiplication is defined).

Associative: $A(BC) = (AB)C$

Scalar Associative: Let α be a scalar. $(\alpha A)B = \alpha(AB) = A(\alpha B)$

Distributive: $A(B + C) = AB + AC$, $(A + B)C = AC + BC$

Definition: Let A be a $m \times n$ matrix. The **transpose** of A denoted A^T (also denoted A' which is said "A prime") is determined by 'flipping' the matrix across it's upper-left to lower-right diagonal. Formally, the element x_{ij} of the matrix A is the element x_{ji} of the matrix A^T.

Example: Row Reduction

Vector Spaces

Definition: A non-empty set $X \in \mathbb{R}^n$ is a **vector space** if it's closed under addition and scalar multiplication (i.e. for $x, y \in X$ and $c \in \mathbb{R}$, $x + y \in X$ and $cx \in X$).

Example: $\mathbb{R}^2$ is a vector space.

Definition: For some set of vectors $v_1, ...v_n \in \mathbb{R}^m$, a **linear combination** of these vectors is any expression of the form $c_1 v_1 + ... + c_n v_n$ where $c_1, ...c_n \in \mathbb{R}$ are scalars.

Definition: For a collection of vectors $v_1, ...v_n$, the **span** of these vectors is the set of all linear combinations of the vectors.

Example: The span of $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$ is the y axis, and the span of $\begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ is $\mathbb{R}^2$.

Note: The span of a set of vectors defines a vector space.

Definition: A set of vectors is $x_1, ..., x_n$ is **independent** if $c_1x_1 + ... + c_nx_n = 0$ implies $c_1 = ... = c_k = 0$.

Example: The vectors $\begin{bmatrix}1 & 0\end{bmatrix}$ and $\begin{bmatrix}0 & 1\end{bmatrix}$ are independent. So are the vectors $\begin{bmatrix}5 & 2\end{bmatrix}$ and $\begin{bmatrix}3 & 4\end{bmatrix}$. However, $x_1 = \begin{bmatrix}7 & -3\end{bmatrix}$ and $x_2 = \begin{bmatrix}-14 & 6\end{bmatrix}$ are not independent since $2x_1 + 1x_2 = 0$.

Theorem: Consider a set of n vectors $V = [v_1, ..., v_n] \in \mathbb{R}^n$. Then, $Span(V) = \mathbb{R}^n \iff v_1, ..., v_n$ are independent.

Definition: A set of vectors S forms a **basis** for the vector space X if S is independent and spans X.

Definition: The **column (row) space** of a matrix A is the span of the columns (rows) of A.

Definition: A set of vectors $x_1, ..., x_n$ is **orthogonal** if $\forall i, j$ we have $\langle x_i, x_j \rangle = 0$.

Note: This definition implies that orthogonal vectors form a right (90 degree) angle.

Example: The vectors $\begin{bmatrix}1 & 0\end{bmatrix}$ and $\begin{bmatrix}0 & 1\end{bmatrix}$ are independent and orthogonal. The vectors $\begin{bmatrix}1 & 0\end{bmatrix}$ and $\begin{bmatrix}1 & 1\end{bmatrix}$ are independent, but they are not orthogonal.

Note: The Standard Basis Vectors (e.g. $\begin{bmatrix}1 & 0 & 0\end{bmatrix}$, $\begin{bmatrix}0 & 1 & 0\end{bmatrix}$ and $\begin{bmatrix}0 & 0 & 1\end{bmatrix}$ for $\mathbb{R}^3$) are orthogonal, independent and form a basis for $\mathbb{R}^n$.

Invertability

Definition: Let A be an $n \times n$ square matrix. We call A **invertible (non-singular)**if there exists some matrix B such that $AB = I_n$, where I_n is the $n \times n$ identity matrix.

Finding Inverse of 2x2 Matrix: The formula for the 2x2 inverse of the matrix $A = \begin{bmatrix}a & b\\ c & d\end{bmatrix}$ is $A^{-1} = \frac{1}{ad-bc}\begin{bmatrix}d & -b\\ -c & a\end{bmatrix}$. Recall that the **determinant** of the matrix A (written $det(A)$ or $|A|$) is given by $ad - bc$, so we could rewrite the inverse expression as $A^{-1} = \frac{1}{|A|}\begin{bmatrix}d & -b\\ -c & a\end{bmatrix}$.

Theorem (Invertible Matrix Theorem (Partial)): Let A be an $n \times n$ square matrix over R^n. The following statements are equivalent:

- A is invertible.
- A has a left or right inverse (i.e. $AB = I$ or $CA = I$).
- The columns (rows) of A span R^n (columns are linearly independent, column space is $\mathbb{R}^n$).
- The transpose A^T is invertible.
- $det(A) \neq 0$

- The number 0 is not an eigenvalue of A.

Projections

Definition: For the vector spaces X and Y, a function $f : X \to Y$ is a **linear transformation (linear mapping)** if for $x_1, x_2 \in X$ and scalar $c \in \mathbb{R}$, $f(x_1 + x_2) = f(x_1) + f(x_2)$ and $f(cx_1) = cf(x_1)$.

Example: Consider $f(x) = \beta_1 x$. Let's try multiplying by a scalar. $f(cx) = \beta_1 cx$ and $c * f(x) = c * \beta_1 x$ so this function allows you to "pull out" scalars. Similarly, $f(a + b) = \beta_1(a + b) = \beta_1 a + \beta_1 b$ and $f(a) + f(b) = \beta_1 a + \beta_1(b)$ so adding the arguments and then applying the function is equivalent to applying the function to each argument and then adding.

Theorem: For any linear transformation $f : \mathbb{R}^k \to \mathbb{R}^l$, there is a unique matrix $M \in \mathbb{R}^{l \times k}$ such that $f(x) = Mx \; \forall \; x \in \mathbb{R}^k$.

Definition: A linear transformation $f(x) = Ax$ is called a **projection (idempotent)** if $A^2 = A$.

For example, the identity matrix is a projection, as for any matrix B, $IB = I^2 B$. Another example (which will become relevant soon) is the $X(X'X)^{-1}X'$. Briefly,

$$\begin{aligned}(X(X'X)^{-1}X') &= X(X'X)^{-1}X'X(X'X)^{-1}X' \\ &= X(X'X)^{-1}X'\end{aligned}$$

since we can cancel out the first $(X'X)^{-1}$ with the following $(X'X)$.

Context: Econometrics (Geometric Interpretation of OLS)

To start to build intuition for the relationship between linear algebra and regression analysis, think of a dataset containing an outcome vector Y (sales) as well as a matrix X containing two columns (marketing spend, quality). If we write

$$Y = \beta_1 X_1 + \beta_2 X_2 + \epsilon,$$

we can see that the set of all possible values for $\beta_1 X_1 + \beta_2 X_2$ is the column space of X.

Writing an OLS regression in matrix notation, we have $Y = X\beta + U$. Here, $Y \in R^{n \times 1}, X \in \mathbb{R}^{n \times k}, \beta \in \mathbb{R}^{k \times 1}$ and $U \in \mathbb{R}^{n \times 1}$ where n is the number of observations and k is the number of variables plus 1 for the constant. (Note: The trick to including a constant in this form is that the first column of X is all 1's.)

The OLS estimator is defined as $\beta = (X'X)^{-1}X'Y$.

The fitted values are given by $\hat{Y} = X\hat{\beta} = X(X'X)^{-1}X'Y$, and the residuals are given by $\hat{Y} - X\hat{\beta} = Y - X(X'X)^{-1}X'Y = (I - X(X'X)^{-1}X')Y$.

Note that $X(X'X)^{-1}X$ is a projection onto the span of the column vectors of X. So we can interpret $\hat{Y}$ as the closets vector to Y within the span of X. Additionally, notice that $\hat{U}$ is orthogonal to the span of X.

To build intuition, let's graph a simple example. Imagine we want to video game sales (Y) of three games (Mario Kart, Pokemon, Zelda) and we have information on marketing spend (X1) and video game review scores (X2). Let the sales be given by $(3, 2, 1)$, the marketing spend be given by $(2, 2, 1)$, and the scores be given by $(1, 1, 1)$. We can draw each of these vectors in three dimensional space. Taking any linear combination of the two X variables create a plane in $\mathbb{R}^3$ (the column space of X). In this case, the Y vector does not fall within that plane. We can interpret the OLS projection $X\beta = X(X'X)^{-1}X'Y$, which geometrically is a linear transformation that maps the Y vector to the closest Y vector in the column space of X. We call that output the fitted values, $\hat{Y}$. Similarly, if we think about the error, which in OLS will be given by $Y - X\beta = Y - X(X'X)^{-1}X'Y = Y(I - X(X'X)^{-1}X')$, this will capture a vector that is orthogonal to the column space of X, and maps from $\hat{Y}$ to Y.

Eigenvalues and Eigenvectors

Consider a 2x2 matrix $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$. Any 2x2 matrix can be used to define a transformation from $T : \mathbb{R}^2 \to \mathbb{R}^2$. In particular, if we start with two co-ordinates in $\mathbb{R}^2$ and define a vector containing them $v = \begin{bmatrix} x \\ y \end{bmatrix}$, we can define a function $T(v) = Av$. In our example, $T(v) = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} ax + by \\ cx + dy \end{bmatrix}$.

Definition: A vector v is called an **eigenvector** and a scaler λ is called an **eigenvalue** of the matrix A if $Av = \lambda v$.

In words, we're looking for a vector v such that the transformation Av does not change the direction of v, but only scales the magnitude by λ.

To find the eigenvalues, rewrite the definition in the following way:

$$\begin{aligned} Av &= \lambda v \\ Av &= \lambda I v \\ Av - \lambda I v &= 0 \\ (A - \lambda I)v &= 0 \\ det(A - \lambda I) &= 0 \\ det(\begin{bmatrix} a - \lambda & b \\ c & d - \lambda \end{bmatrix}) &= 0 \\ (a - \lambda)(d - \lambda) - bc &= 0 \end{aligned}$$

How can we make the step from line 4 to line 5 switching to the determinant? Notice that if $A - \lambda I$ is invertible, then the only solution is the trivial $v = 0$ which we can

see by multiplying both sides by $(A - \lambda I)^{-1}$. If $A - \lambda I$ is not is not invertible, then the determinant of $A - \lambda I$ must be equal to 0 by the Invertible Matrix Theorem.

Solving this equation for λ will give you the eigenvalues associated with A. If we call one of these eigenvalues λ^* and call the two elements of the eigenvector v that we're trying to solve for x and y, we can plug back into the defining equation to get

$$\begin{aligned} Av &= \lambda^* v \\ \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} &= \lambda^* \begin{bmatrix} x \\ y \end{bmatrix} \\ \begin{bmatrix} ax + by \\ cx + dy \end{bmatrix} &= \begin{bmatrix} \lambda^* x \\ \lambda^* y \end{bmatrix} \end{aligned}$$

These equations should pin down a ratio of x to y, and any non-zero multiple of that ratio is an eigenvector (geometrically, all of these eigenvectors are points on a line, and the transformation A scales takes vectors on that line and scales them along the line). If you would like to see a geometric representation of eigenvectors and eigenvalues, take a look at this interactive explainer.

Diagonalization

Definition: A square matrix M is called a **diagonal** matrix if all non-zero elements lie within the diagonal running from the upper left corner to the lower right.

For example, $\begin{bmatrix} 5 & 0 \\ 0 & -3 \end{bmatrix}$ and $\begin{bmatrix} -1 & 0 \\ 0 & 0 \end{bmatrix}$ are diagonal matrices.

Definition: A matrix M is **diagonalizable** if there exists an invertible matrix P such that PMP^{-1} is a diagonal matrix.

Connection between Diagonalization and Eigenvalues and Eigenvectors

The defining equation for eigenvalues and eigenvectors is $Av = \lambda v$. Consider the 2x2 case, and suppose we have two eigenvalues λ_1, λ_2 and two corresponding eigenvectors v_1, v_2 for the matrix A. By definition, we have the following two equations:

$$\begin{aligned} Av_1 &= \lambda_1 v_1 \\ Av_2 &= \lambda_2 v_2 \end{aligned}$$

If we define a eigenvector matrix where each column of the matrix is one of the eigenvectors of A (i.e. $V = \begin{bmatrix} v_1 & v_2 \end{bmatrix}$) and a diagonal eigenvalue matrix (i.e. $\Lambda = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$), then we can represent those two equations in matrix notation as $AV = V\Lambda$.

Working with this matrix equation, we can write $V^{-1}AV = \Lambda$, and since we know that Λ is a diagonal matrix, we know A is diagonalizable by definition.

Additionally, we could rewrite the equation as $A = V\Lambda V^{-1}$. A nice property that can be leveraged here is that $A^n = V\Lambda^n V^{-1}$, so while in general computing A^n is burdensome, if you can rewrite A in this way, the task is substantially easier. To see why this works, notice that $A^2 = (V\Lambda V^{-1})(V\Lambda V^{-1})$ and the inner $V^{-1}V$ cancels. Also, notice that Λ^n is easy to compute as $\Lambda^n = \begin{bmatrix} \lambda_1^n & 0 \\ 0 & \lambda_2^n \end{bmatrix}$.

If you would like a video explanation of these concepts, see this MIT OpenCourseware lecture.

Theorem: An $n \times n$ matrix A is diagonalizable in R^n if and only if the eigenvectors of A span R^n.

The takeaway of this theorem is that the diagonalization procedure outlined above does not always work! It only works if the eigenvectors associated with the matrix span the relevant space. A sufficient condition for this to work is that the eigenvalues are distinct, but note that this condition is not necessary.

Example: Simple Dynamic System

Context: Macro *(Simple Dynamic System)*

It is common to simplify complex macroeconomic models to linear approximations using a technique called "log-linearization." For example, consider a dynamic model where the level of consumption c_t and capital k_t in time period t affect the levels of consumption c_{t+1} and capital k_{t+1} in the next period (time $t+1$) according to a potentially complex equation $F(c_t, k_t, c_{t+1}, k_{t+1}) = 0$. Using log-linearization, we can obtain a simpler linear dynamic system

$$\begin{bmatrix} \dot{c_{t+1}} \\ \dot{k_{t+1}} \end{bmatrix} = M \begin{bmatrix} \dot{c_{t+1}} \\ \dot{k_{t+1}} \end{bmatrix}$$

where $\dot{c_t} = log(c_t) - log(c^*) = log(\frac{c_t}{c^*})$ and c^* represents the "steady state" of consumption in the long run, so we can interpret $\dot{c_t}$ as the percentage difference between the amount of consumption in period t and the amount of consumption in the steady state. M is a matrix that transforms capital and labor in this period to capital and labor in the next period. For example, if $M = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, then $\dot{c_{t+1}} = \dot{c_t}$ and $\dot{k_{t+1}} = \dot{k_t}$.

In this model, we assume that in equilibrium the economy converges to a steady state, so we have $\dot{c_t}, \dot{k_t} \to 0$. Our goal is to find an expression for the initial level of consumption c_0 as a function of the initial level of capital k_0.

Using the linear algebra techniques we just covered and assuming that M is diagonalizable (it's eigenvectors span the space $\mathbb{R}^2$), we can apply an eigenvalue decomposition

to M so we have $M = P\Lambda P^{-1}$, and plugging this into our model, we have

$$\begin{bmatrix} \dot{c_{t+1}} \\ \dot{k_{t+1}} \end{bmatrix} = P\Lambda P^{-1} \begin{bmatrix} \dot{c_{t+1}} \\ \dot{k_{t+1}} \end{bmatrix}$$
$$P^{-1} \begin{bmatrix} \dot{c_{t+1}} \\ \dot{k_{t+1}} \end{bmatrix} = \Lambda P^{-1} \begin{bmatrix} \dot{c_{t+1}} \\ \dot{k_{t+1}} \end{bmatrix}$$
$$Z_{t+1} = \Lambda Z_t$$

where in the last line we define $Z_t = P^{-1} \begin{bmatrix} \dot{c_{t+1}} \\ \dot{k_{t+1}} \end{bmatrix}$. Now, using this expression to iterate backwards in time, we have

$$Z_t = \Lambda Z_{t-1} = \Lambda^2 Z_{t-1} = ... = \Lambda^t Z_0$$

Since Λ is a diagonal matrix, we have $\Lambda^t = \begin{bmatrix} \lambda_1^t & 0 \\ 0 & \lambda_2^t \end{bmatrix}$. If we define $P^{-1} = \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix}$, then we rewrite our expression as

$$Z_t = \Lambda^t Z_0$$
$$Z_t = \begin{bmatrix} \lambda_1^t & 0 \\ 0 & \lambda_2^t \end{bmatrix} \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix} \begin{bmatrix} \dot{c_0} \\ \dot{k_0} \end{bmatrix}$$
$$Z_t = \begin{bmatrix} \lambda_1^t (p_{11}\dot{c_0} + p_{12}\dot{k_0}) \\ \lambda_2^t (p_{21}\dot{c_0} + p_{22}\dot{k_0}) \end{bmatrix}$$

We have assumed that the left hand side of this equation should go to 0 in equilibrium, as it's a function of $\dot{c}, \dot{k}$ and as time goes on and the model approaches steady state. The exact solution depends on the eigenvalues; suppose that $|\lambda_1| > 1$ and $|\lambda_2| < 1$. We know that the lower equation will approach 0 as t grows large, but the upper equation will grow exponentially unless $p_{11}\dot{c_0} + p_{12}\dot{k_0} = 0$, and since we have assumed $Z_t \to 0$, we have this conclusion. Substituting in for the initial definition of $\dot{c}$ and

isolating c_0, we have

$$\begin{aligned}
0 &= p_{11}\dot{c_0} + p_{12}\dot{k_0} \\
0 &= p_{11}(log(c_0) - log(c^*)) + p_{12}(log(k_0) - log(k^*)) \\
p_{11}(log(c_0) - log(c^*)) &= -p_{12}(log(k_0) - log(k^*)) \\
p_{11}log(c_0) &= p_{11}log(c^*) - p_{12}(log(k_0) - log(k^*)) \\
log(c_0) &= log(c^*) - \frac{p_{12}}{p_{11}}(log(k_0) - log(k^*)) \\
c_0 &= exp(log(c^*) - \frac{p_{12}}{p_{11}}(log(\frac{k_0}{k^*})) \\
c_0 &= exp(log(c^*))exp(-\frac{p_{12}}{p_{11}}(log(\frac{k_0}{k^*}))) \\
c_0 &= c^*exp(-\frac{p_{12}}{p_{11}}log(\frac{k_0}{k^*}))
\end{aligned}$$

which is what we were looking for (namely, an expression of initial labor as a function of initial capital).

Calculus and Real Analysis

Motivation

In this section, we will cover just enough ideas from real analysis to formalize calculus concepts like differentiability and continuity. Then, we'll briefly review frequently used calculus techniques. These techniques are the bread and butter of first year coursework: integration and differentiation are useful for solving economic models and proving econometrics results.

In contrast, I think real analysis can sometimes seem far away from day to day coursework. To give some examples of where the concepts we are about to study will come up in econometrics:

- Sets are an important concept for understanding confidence intervals: intuitively, a confidence interval is a set of points that contains the true value of a parameter 95% of the time, for example.
- Limits are important when proving asymptotic properties of estimators. If you want to prove that your estimator is consistent (the error approaches zero as the number of observations grows large), you need to take the limit of that expression.

Additionally, real analysis provides a context to practice the mechanics of formal proofs.

Mathematical Logic

You will often be asked to prove a statement. Here, I will briefly discuss mathematical logic and how to go about that task.

A **statement** is any claim about the world that is either true or false. For example, "10 is an odd number" or "the sky is blue." In logic, it's common to use the letters p and q to refer to statements.

A standard kind of claim that you will have to prove is an **if/then** statement, or an **implication**: if p, then q. Sometimes this is written $p \implies q$.

An **if and only if** statement $p \iff q$ means that $p \implies q$ and $q \implies p$. If we know that $p \iff q$, then we say that p and q are **equivalent**.

Notice that $p \implies q$ does not imply that $q \implies p$. For example, Karthik likes all cheese (p) implies that Karthik likes Gouda (q), but the fact that Karthik likes Gouda does not mean that Karthik likes all cheese.

Once you have proven an $\iff$ statement, it functions 'like a definition'.

The logical operator **not** p, sometimes called the negation of p, is denoted $\neg p$ and takes on the opposite value of p. So if p is true, not p is false and vice versa.

The **contraposative** of the implication $p \implies q$ is $\neg q \implies \neg p$. The implication is true if and only if the contrapositive is true. If Karthik does not like Gouda, then Karthik does not like all cheese.

Loosely speaking, there are four ways to prove the claim $p \implies q$:

First, a direct proof (proof by construction). Assume p is true. Use p to show q. (Simple, right!)

My advice: when trying to create a proof, start by trying to fully understand each of the definitions of objects .

Second, proof by contradiction. Assume p and $\neg$ q. Show that the pair of assumptions means that some statement is inconsistent (both true and false), which implies that one of our assumptions is wrong. Typically, we either assume or know that p is true, so this gives us that q must be true.

Third, proof by induction. Suppose you want to show $p(x) \implies q(x)$ for some set of statements indexed by the natural numbers. Show that the statement is true for the basic case $p(1) \implies q(1)$. Then, show that if $p(n) \implies q(n)$, then $p(n+1) \implies q(n+1)$. Arguably, this is a specific case of direct proof.

Fourth, proof by contrapositive. Prove $\neg q \implies \neg p$. Then, since the contrapositive and the original implication are equivalent statements, if $\neg q \implies \neg p$ then $p \implies q$.

Sets

Definition: A **set** is a (potentially infinite) collection of objects (e.g. numbers, vectors, sides of dice, words). These objects are called the **members** or **elements** of the set.

The set with no elements is called the **empty set** and is denoted $\emptyset$.

The **natural numbers** are denoted $\mathbb{N} = \{1, 2, 3, ...\}$. If we include 0 in this set, it's instead referred to as the **whole numbers** and can be denoted $\mathbb{N}_0$.

The **integers** are denoted $\mathbb{Z} = \{..., -2, -1, 0, 1, 2, ...\}$.

The **rational numbers** are denoted $\mathbb{Q}$ and are the set of all numbers that can be represented by the ratio of two integers. Formally, $\mathbb{Q} = \{\frac{p}{q} : p, q \in \mathbb{Z} \text{ and } q \neq 0\}$.

The **real numbers** are denoted $\mathbb{R}$. Intuitively, the real numbers are all numbers that we can find on the number line. Their definition is complicated, but we can think of them as a completion of the rational numbers defined by decimal expansion (e.g. 3, 3.1, 3.14, 3.141, 3.1415... converges to a particular real number π. All real numbers can be thought of as potentially infinite sequences of decimals.)

Definition: The **set difference** of two sets B and A is the set of elements that are in B but not in A. Formally, $B \setminus A = \{x \in B : x \notin A\}$.

Definition: For a set $A \subset X$, we define the **complement** of A as $A^C = X \setminus A$.

The **irrational numbers** are denoted $\mathbb{P}$ and can be thought of as $\mathbb{R}$ set minus $\mathbb{Q}$.

For any set X, we can form a finite ordered list of elements called an **n-tuple** where n gives the number of elements in the list.

Example: In a simple economic model of job search, we might imagine that there is a set of wage offers $\{H, M, L\}$. We could think about a worker who gets an ordered list of job (L, M, L, M, H).

Because the notation is similar, it's easy to confuse sets and n-tuples. There are two key differences. The elements of sets are *unique (repetition is not okay)* and *order does not matter*. In contrast, the elements of n-tuples are *not unique (repetition is okay)* and *order maters*.

Definition: A set A is a **subset** of X, denoted $A \subset X$ if every element of A belongs to X. Formally, $x \in A \implies x \in X$.

Definition: The **power set** of X, denoted $\mathcal{P}(X)$ is the set of all subsets of X.

Example: If $X = \{1, 2, 3\}$, then $\mathcal{P}(X) = \{\{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}$.

Definition: The **intersection** of sets A and B, denoted $A \cap B$ is the set of elements that belong to both A and B. That is, $x \in A \cap B$ if and only if $x \in A$ and $x \in B$.

De Morgan's Laws: For sets $A, B \subset X$ we have

$$(A \cup B)^c = A^c \cap B^c$$

and

$$(A \cap B)^c = A^c \cup B^c$$

Question: Prove De Morgan's laws graphically, using Venn diagrams.

Definition: A set $G \subset \mathbb{R}$ is **open** in $\mathbb{R}$ if for every $x \in G$ there exists $\delta > 0$ such that $(x - \delta, x + \delta) \subset G$.

Example: Consider the open interval $(0, 1)$. For any point $x \in (0, 1)$, we can construct delta by setting $\delta = \frac{min(|x|,|x-1|)}{2}$. This construction is essentially half the distance between the point and the closest edge, and so we know that $(x - \delta, x + \delta) \in (0, 1)$ for any x we choose if we construct δ in this way.

Definition: A set $G \subset \mathbb{R}$ is **closed** if it's complement G^c is open.

Question: Proof or counter-example: A set cannot be both closed and open.

Question: Proof or counter-example: A set cannot be neither closed nor open.

Proposition: An arbitrary union of open sets is open. A finite intersection of open sets is open.

Example: Consider sets $A_n = \{(-\frac{1}{n}, \frac{1}{n}) : n \in \mathbb{N}\}$. The infinite intersection is the point 0, and that single point is not an open set.

Proposition: An arbitrary intersection of closed sets is closed, and a finite union of closed sets is closed.

Definition: The **Cartesian product** of sets X and Y is the set of tuples (ordered pairs) (x, y) such that $x \in X$ and $y \in Y$. For sets $X_1, ...X_n$, the Cartesian product $\Pi_{i=1}^n X_i = \{(x_1, ..., x_n) | x_i \in X_i \; \forall \; i \in 1, ..., n\}$.

Example: A Cartesian product you will often work with is $\mathbb{R}^2$ (the Cartesian plane). More generally, we frequently think about $\mathbb{R}^n$.

Definition: $A \subset \mathbb{R}$ is **bounded from above** if there exists a real number $M \in \mathbb{R}$ called an upper bound of A, such that $x \leq M \; \forall \; x \in A$. A is **bounded from below** if there exists an $m \in \mathbb{R}$, called a lower bound of A, such that $x \geq m \; \forall \; x \in A$. A is **bounded** if it is bounded from above and below.

Definition: $A \subset \mathbb{R}$ is a set of real numbers. If $M \in \mathbb{R}$ is an upper bound of an such that $M \leq M'$ for every upper bound M' of A, then M is called the least upper bound or **supremum** of A, denoted $sup(A)$.

Similarly, if $m \in \mathbb{R}$ is a lower bound of A such that $m \geq m'$ for every lower bound m' of A, then m is called the greatest lower bound or **infimum** of A, denoted $inf(A)$.

Intuitively, the reason we want to define an infimum and supremum is because open sets in $\mathbb{R}$ typically do not have well defined maximums and minimums within the set. For example, the set $A = (0, 1)$ has no minimum or maximum, but the infimum is 0 and the supremum is 1.

Another example to consider is $A = \{\frac{1}{n} : n \in \mathbb{N}\}$. For this set, $max(A) = sup(A) = 1$. $inf(A) = 0$ and the minimum does not exist.

Question: Let $A \subset \mathbb{R}$, $A = \{x \in \mathbb{R} : x^2 < 2\}$. Does A have a maximum? Supremum? What if $A \subset \mathbb{Q}$?

Functions

Definition: A **function** $f : X \to Y$ for sets X and Y is a subset $f \subset X \times Y$ that satisfies $\forall\, x \in X \exists$ a unique $y \in Y : (x, y) \in f$. We write $f(x) = y$ if $(x, y) \in f$.

Intuitively, a function assigns elements $x \in X$ to elements $f(x) \in Y$.

The set X is called the **domain**, and the set $Y = \{y : y = f(x) \text{ for some } x \in X\}$ is called the **co-domain**.

Definition: The **image** or **range** of a set $S \subset X$ under f is $f(X) := \{f(x) | x \in S\}$.

Definition: The **preimage** of a set $V \subset Y$ under f is $f^{-1}(V) := \{x \in X | f(x) \in V\}$.

Definition: A function $f : X \to Y$ is **onto (surjective)** if it's range is all of Y. That is, if $\forall\, y \in Y \exists x \in X$ such that $f(x) = y$.

Definition: A function $f : X \to Y$ is **one-to-one (injective)** if it maps distinct elements of X to distinct elements of Y. That is, if $x_1 \neq x_2$, than $f(x_1 \neq x_2)$.

Example: $f : X \to Y$ with $f(x) = x^2$. Consider $X = [0, 2]$ and $X = [-2, 2]$ with $Y = [0, 4]$

Question: $f : \mathbb{R}^2 \to \mathbb{R}^2$ with $f(x) = Ax$, $A = \begin{bmatrix} 1 & 1 \\ \frac{1}{2} & 2 \end{bmatrix}$. What is the determinant of A? What is the image of the unit square $C = [0, 1] \times [0, 1]$ under f? Is f one-to-one? Onto?

Question: $f : \mathbb{R}^2 \to \mathbb{R}^2$ with $f(x) = Ax$, $A = \begin{bmatrix} 2 & -1 \\ -2 & 1 \end{bmatrix}$. What is the determinant of A? What is the image of the unit square $C = [0, 1] \times [0, 1]$ under f? Is f one-to-one? Onto?

Definition: A function $f : X \to Y$ is called **bijective** if it is both one-to-one and onto.

Theorem: A bijective function $f : X \to Y$ has an inverse $f^{-1} : Y \to X$ such that $f^{-1}(y) = x$ if and only if $f(x) = y$.

Definition: A function $f : A \to \mathbb{R}$ is **bounded** on $B \subset A$ if there exists $M \geq 0$ such that

$$|f(x)| \leq M \; \forall\, x \in B$$

Definition: A set X is **indexed** by I if there is an onto function $f : I \to X$. We can write $X = \{x_i : i \in I\}$.

Question: Can you index $\mathbb{Q}$ with the natural numbers $\mathbb{N}$? What about indexing $\mathbb{R}$ with $\mathbb{N}$?

Proposition: (De Morgan) If $\{X_i \subset X : i \in I\}$ is a collection of subsets of a set X, then

$$\left(\bigcup_{i\in I} X_i\right)^c = \bigcap_{i\in I} X_i^c, \quad \left(\bigcap_{i\in I} X_i\right)^c = \bigcup_{i\in I} X_i^c$$

Definition: A **binary relation** R on sets X and Y defines a relationship between elements of X and elements of Y. We write xRy if $x \in X$ and $y \in Y$ are related.

The key point here is a relation is a generalization of a function (functions are a subset of relations). A relation allows each element of x to be mapped to potentially many y's. In particular, a function is a relation such that an element in x and an element in y are related if and only if $f(x) = y$.

Limits

Definition: Let $f : A \to \mathbb{R}$, $A \subset \mathbb{R}$, $c \in \mathbb{R}$. Then, the **limit** of f(x) as x approaches c, denoted

$$\lim_{x\to c} f(x) = L$$

if for every $\epsilon > 0$ there exists $\delta > 0$ such that

$$0 < |x - c| < \delta \implies |f(x) - L| < \epsilon$$

Definition: Let $f : A \to \mathbb{R}$, $A \subset \mathbb{R}$, and $c \in \mathbb{R}$. The the **right limit** equals L, denoted

$$\lim_{x\to c^+} f(x) = L$$

if for every $\epsilon > 0$ there exists a $\delta > 0$ such that $c < x < c + \delta$ and $x \in A$ implies that $|f(x) - L| < \epsilon$ and the **left limit** equals L, denoted

$$\lim_{x\to c^-} f(x) = L$$

if for every $\epsilon > 0$ there exists a $\delta > 0$ such that $c - \delta < x < c$ and $x \in A$ implies that $|f(x) - L| < \epsilon$.

Definition: Let $f : A \to \mathbb{R}$, $A \subset \mathbb{R}$. If A is not bounded from above, the the **limit as x approaches infinity** equals L, denoted

$$\lim_{x\to\infty} f(x) = L$$

if for every $\epsilon > 0$ there exists an $M \in \mathbb{R}$ such that

$$x > M \text{ and } x \in A \implies |f(x) - L| < \epsilon$$

If A is not bounded from below, then the limit as x approaches negative infinity equals L, denoted

$$\lim_{x \to -\infty} f(x) = L$$

if for every $\epsilon > 0$ there exists an $M \in \mathbb{R}$ such that

$$x < M \text{ and } x \in A \implies |f(x) - L| < \epsilon$$

Definition: Let $f : A \to \mathbb{R}$, $A \subset \mathbb{R}$, $c \in \mathbb{R}$. Then we say that f(x) **diverges** to infinity, denoted

$$\lim_{x \to c} f(x) = \infty$$

if for every $M \in \mathbb{R}$ there exists $\delta > 0$ such that

$$0 < |x - c| < \delta \text{ and } x \in A \implies f(x) > M$$

and

$$\lim_{x \to c} f(x) = -\infty$$

if for every $M \in \mathbb{R}$ there exists $\delta > 0$ such that

$$0 < |x - c| < \delta \text{ and } x \in A \implies f(x) < M$$

Theorem: If the limit of a function exists, then the limit is unique.

Question: Prove this theorem.

Theorem (Limits Preserve Order): Suppose $f, g : A \to \mathbb{R}$. If $f(x) < g(x) \; \forall \; x \in A$,

$$f(x) < g(x) \; \forall \; x \in A$$

and the limits exist, then

$$\lim_{x \to c} f(x) < \lim_{x \to c} g(x)$$

Theorem (Algebraic Properties of Limits): Let $f, g : A \to \mathbb{R}$, $k \in \mathbb{R}$, and let the limits

$$\lim_{x \to c} f(x) = L \qquad \lim_{x \to c} g(x) = M$$

exist. Then,

$$\lim_{x \to c} kf(x) = kL$$
$$\lim_{x \to c} f(x) + g(x) = L + M$$
$$\lim_{x \to c} f(x)g(x) = LM$$
$$\lim_{x \to c} \frac{f(x)}{g(x)} = \frac{L}{M} \text{ if } M \neq 0$$

Question: Prove the portion of this theorem related to multiplying limits.

Theorem (Sandwich/Squeeze): Let $f, g, h : A \to \mathbb{R}$, $c \in \mathbb{R}$. If

$$f(x) \leq g(x) \leq h(x) \; \forall \; x \in A$$

and

$$\lim_{x \to c} f(x) = \lim_{x \to c} h(x) = L$$

then the limit of $g(x)$ as $x \to c$ exists and

$$\lim_{x \to c} g(x) = L$$

Note: This theorem is often used to show that certain functions converge to zero.

Example: The limit

$$\lim_{x \to 0} x^2 \sin(\frac{1}{x})$$

cannot be found using simple limit rules because $\lim_{x \to 0} \sin(\frac{1}{x})$ does not exist.

Instead, we can notice that $x^2 \sin(\frac{1}{x})$ is bounded by $-x^2$ and x^2 since $|\sin(x)| \leq 1$. Then, since $\lim_{x \to 0} x^2 = \lim_{x \to 0} -x^2 = 0$, we know that

$$\lim_{x \to 0} x^2 \sin(\frac{1}{x}) = 0$$

by the squeeze theorem.

Continuity

Intuitively, continuous functions are functions that take on nearby values at nearby points.

Definition: Let $f : A \to \mathbb{R}$, $c \in A \subset \mathbb{R}$. The function f is **continuous at c** if for every $\epsilon > 0$ there exists $\delta > 0$ such that for $x \in A$

$$|x - c| < \delta \implies |f(x) - f(c)| < \epsilon$$

Proposition: Let $f : A \to \mathbb{R}$, $c \in A \subset \mathbb{R}$. The function f is continuous at c if and only if

$$\lim_{x \to c} f(x) = f(c)$$

Definition: Let $f : A \to \mathbb{R}$. We say that the function if is **continuous** on A if the function is continuous for all $c \in A$.

Question: Prove that the function $f(x) = \sqrt{x}$ is continuous on $[0, \infty)$.

Theorem: If $f, g : A \to \mathbb{R}$ are continuous at $c \in A$ and $k \in R$, then kf, $f + g$, and fg are continuous at c. Moreover, if $g(c) \neq 0$, then $\frac{f}{g}$ is continuous at c.

Question: A polynomial function is a function of the form $P(x) = a_0 + a_1x + a_2x^2 + ...$ for coefficients $a_n \in \mathbb{R}$. Prove that every polynomial function is continuous.

Theorem: Let $f : A \to \mathbb{R}$ and $g : B \to \mathbb{R}$ where $f(A) \subset B$. If f is continuous at $c \in A$ and g is continuous at $f(c) \in B$, then the composition $g(f(x)) : A \to \mathbb{R}$ is continuous at c.

Definition: Let $f : A \to \mathbb{R}$ where $A \subset \mathbb{R}$. f is **uniformly continuous** on A if for every $\epsilon > 0$ $\exists$ $\delta > 0$ such that

$$|x - y| < \delta \text{ and } x, y \in A \implies |f(x) - f(y)| < \epsilon$$

The key point here is that we need to be able to pick a δ based only on ϵ and not as a function of the points x and y. For example, $f(x) = x^2$ is uniformly continuous on any bounded interval, but is only continuous on $\mathbb{R}$.

Connection between continuous functions and open sets: Consider $f : \mathbb{R} \to \mathbb{R}$, $f(x) = x^2$. Consider the open interval $I = (1, 4)$. The image of this set $f(I) = (1, 16)$ is open. The inverse image $f^{-1}(-2, -1) \cup (1, 2)$ is open.

However, if we instead consider the set $J = (-1, 1)$ then the image $f(J) = [0, 1)$ and the inverse image $f^{-1}(J) = (-1, 1)$, so the inverse image is open but the image is not.

Theorem: A function $f : A \to \mathbb{R}$ is continuous on A if and only if $f^{-1}(V)$ is open in A for every set V that is open in $\mathbb{R}$.

Intermediate Value Theorem (Special Case): Suppose that $f : [a, b] \to \mathbb{R}$ is a continuous function on a closed, bounded interval. If $f(a) < 0$ and $f(b) > 0$, then there is a point $a < c < b$ such that $f(c) = 0$.

Question: Sketch a proof of this theorem.

Intermediate Value Theorem: Suppose that $f : [a, b] \to \mathbb{R}$ is a continuous function on a closed, bounded interval. For every d between $f(a)$ and $f(b)$ there is a point $a < c < b$ such that $f(c) = d$.

Proof: Let $g(x) = f(x) - d$. Then, $g(a) < 0$ and $g(b) > 0$, so $g(c) = 0$ for some c by the special case of the theorem, which means $f(c) = d$.

Context (Price Theory): This theorem can be used to show the existence of solutions when a closed form solution is difficult. To give a trivial example, if you know that for some continuous marginal profit function, the marginal profit is negative and at another point it's positive, then you know there must be a point where it equals 0.

Differentiability

Definition: Suppose that $f : (a, b) \to \mathbb{R}$ and $a < c < b$. Then f is **differentiable** at c with derivative $f'(c)$ if

$$\lim_{h \to 0} \left[\frac{f(c+h) - f(c)}{h} \right] = f'(c)$$

The domain of f' is the set of points $c \in (a, b)$ for which this limit exists. If the limit exists for every $c \in (a, b)$ then we say that f is differentiable on (a, b).

This definition can also be written

$$\lim_{x \to c} \left[\frac{f(x) - f(c)}{x - c} \right] = f'(c)$$

Question: Give an example of a function that is continuous but not differentiable.

Theorem: If $f : (a, b) \to \mathbb{R}$ is differentiable at $c \in (a, b)$ then f is continuous at c.

Question: Prove this theorem.

Definition: A function $f : (a, b) \to \mathbb{R}$ is **continuously differentiable** on (a, b) denoted $f \in C^1(a, b)$, if it is differentiable on (a,b) and $f' : (a, b) \to \mathbb{R}$ is continuous.

An example of a function that is differentiable but not continuously differentiable is

$$f(x) = \begin{cases} x^2 sin(\frac{1}{x}) & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{cases}$$

as

$$f'(x) = 2xsin(\frac{1}{x}) - cos(\frac{1}{x})$$

and this function is not continuous at x = 0.

Theorem: (Algebraic Properties of Derivatives) If $f, g : (a, b) \to \mathbb{R}$ are differentiable at $c \in (a, b)$ and $k \in \mathbb{R}$, then kf and $f + g$ are differentiable at c with

$$kf'(c) = kf'(c)$$
$$(f + g)'(c) = f'(c) + g'(c)$$

Context (Optimization): We use derivatives to locate maximum and minimum points.

Theorem: If $f : A \subset \mathbb{R} \to \mathbb{R}$ has a local extreme value at an interior point $c \in A$ and f is differentiable at c, then $f'(c) = 0$.

Convexity/Concavity

Definition: A twice-differentiable, single variable function is **convex** (**concave**) if and only if its second derivative is non-negative (non-positive) on its entire domain. More generally, a real-valued function defined on an n-dimensional interval is convex (concave) if the line segment between any two points on the graph of the function lies on or above (below) the graph.

Convex Function Examples: $f(x) = x^2$, $f(x) = e^x$

Tip: (Remembering which is which) For convex, V looks like x^2, a convex function. For concave, think of a cave entrance, which looks like a concave function.

Context: Price Theory One common use case for invoking convex and concave functions in economics is to ensure that two functions cross: for example, if a person's utility from consuming coconuts is concave (diminishing marginal utility), and the cost of labor from climbing up trees to get coconuts is convex, then we know that at some point the cost of climbing up a tree to get an additional coconut will exceed the benefits of consuming an additional coconut, so the total number of coconuts consumed is finite.

Differentiation Rules

Derivatives to Know

$$\frac{\partial}{\partial x} ln(x) = \frac{1}{x}$$

$$\frac{\partial}{\partial x} sin(x) = cos(x)$$

$$\frac{\partial}{\partial x} cos(x) = -sin(x)$$

Power Rule

$$\frac{\partial}{\partial x} x^n = nx^{n-1}$$

Product Rule

$$\frac{\partial}{\partial x} f(x)g(x) = f'(x)g(x) + g'(x)f(x)$$

Example: How can we find the derivative $h'(x)$ for $h(x) = xsin(x)$?

Write $h(x) = f(x)g(x)$ where $f(x) = x$ and $g(x) = sin(x)$. Then, applying the product rule, we have $h'(x) = sin(x) + xcos(x)$.

Quotient Rule

$$\frac{\partial}{\partial x} \frac{f(x)}{g(x)} = \frac{f'(x)g(x) - g'(x)f(x)}{g^2}$$

Chain Rule

$$\frac{\partial}{\partial x} f(g(x)) = f'(g(x))g'(x)$$

Example: How can we find the derivative $\frac{\partial}{\partial x} h(x) = (7x+3)^4$?

Write $h(x) = f(g(x))$ where $f(x) = x^4$ and $g(x) = 7x + 4$. Then, applying the chain rule formula we have $h'(x) = 3(7x+3)^3 \times 7 = 21(7x+3)^3$.

L'Hospital's Rule

Definition: (L'Hospital's Rule) If $\lim_{x\to c} f(x) = 0$ and $\lim_{x\to c} g(x) = 0$ and $\lim_{x\to c} \frac{f'(x)}{g'(x)} = K$, then $\lim_{x\to c} \frac{f(x)}{g(x)} = K$. Similarly, if $\lim_{x\to c} f(x) = \pm\infty$ and $\lim_{x\to c} g(x) = \pm\infty$ and $\lim_{x\to c} \frac{f'(x)}{g'(x)} = K$, then $\lim_{x\to c} \frac{f(x)}{g(x)} = K$.

Example: Consider $\lim_{x\to 0} \frac{sin(x)}{x}$. Notice that if we just plug in 0 for x, we get $\frac{0}{0}$ which is undefined. Instead, we can take derivatives so $f(x) = sin(x) \Rightarrow f'(x) = cos(x)$ and $g(x) = x \Rightarrow g'(x) = 1$. Evaluating the limit of the derivatives of the numerator and denominator functions, we have $\lim_{x\Rightarrow 0} \frac{cos(x)}{1} = \frac{1}{1} = 1$. Since this is defined, we can apply L'Hospital's Rule to get that $\lim_{x\to 0} \frac{sin(x)}{x} = 1$.

Riemann Integration

Definition: The **supremum** of a function $f : \mathbb{R} \to \mathbb{R}$, denoted $\sup f(x)$, is the supremum of the range of the function (i.e. $\sup Y$ where $y \in Y$ if $y = f(x)$ for some x in the domain of f.)

Definition: Two intervals are **almost disjoint** if they are disjoint or intersect only at a common endpoint.

Definition: A **partition** P of an interval I is a finite collection $I_1, ...I_N$ of almost disjoint, non-empty, compact subintervals whose union is I.

Consider a function $f : [a, b] \to \mathbb{R}$ and a compact interval $I = [a, b]$. Define $M = \sup_I f$ and $m = \inf_I$. For a partition $P = \{I_1, ..., I_N\}$, define $M_k = \sup_{I_k} f$ and $m = \inf_{I_k} f$. Denote the length of a sub-interval I_k as $|I_k|$.

Definition: The **upper Riemann sum** of f with respect to the partition P is

$$U(f, P) = \sum_{k=1}^{N} M_k |I_k|$$

and the **lower Riemann sum** of f with respect to P is

$$L(f, P) = \sum_{k=1}^{N} m_k |I_k|$$

Let $\Pi(a,b)$ denote the set of all partitions of $[a,b]$.

Definition: The **upper Riemann integral** of f on $[a,b]$ is given by

$$U(f) = \inf_{P\in\Pi} U(f,P)$$

and the **lower Riemann integral** of f on $[a,b]$ is given by

$$L(f) = \sup_{P\in\Pi} L(f,P)$$

Definition: A function $f : [a,b] \to \mathbb{R}$ is **Riemann integrable** on $[a,b]$ if it is bounded and it's upper integral $U(f)$ and lower integral $L(f)$ are equal. In that case, the integral, denoted

$$\int_a^b f(x)dx = U(f) = L(f).$$

If the function is unbounded, or the upper and lower integrals disagree, we say the function is **not integrable** (or that the integral is not defined).

Question: Proof or Counterexample: If a function is discontinuous, it is not integrable.

Example: The Dirichlet function $f : [0,1] \to \mathbb{R}$ defined by

$$f(x) = \begin{cases} 0 \text{ if } x \in \mathbb{Q} \\ 1 \text{ otherwise} \end{cases}$$

is bounded but not integrable as for any interval in the partition, $\inf_{I_k} f = 0$ and $\sup_{I_k} f = 1$.

Theorem (Algebraic Properties of Integrals):

- Linearity:

$$\int_a^b cf = c\int_a^b f$$

$$\int_a^b f+g = \int_a^b f + \int_a^b g$$

- Monotnoicity: If $f \leq g$, then

$$\int_a^b f \leq \int_a^b g$$

- Additivity: For $a \leq c \leq b$

$$\int_a^c f + \int_c^b f = \int_a^b f$$

Integration by Parts

Applying the product rule to the functions U and V, we have

$$\frac{\partial}{\partial x}UV = Udv + Vdu$$
$$\int (\frac{\partial}{\partial x}UV) = \int (Udv + Vdu) \text{ (integrating both sides)}$$
$$UV = \int Udv + \int Vdu$$
$$\int Udv = UV - \int Vdu$$

Rule of Thumb: In general, it's best to choose the part of the integrand that gets simpler/smaller when you take the derivative as your 'U', leaving the rest of the expression as 'V'.

LIATE Rule of Thumb: LIATE is an acronym that stands for Logs, Inverse trig (e.g. arcsin, inv cos), Algebraic (e.g. polynomials), Trig (e.g. sin, cos, tan), Exponentials. LIATE helps you choose your 'U': the functions that occur earlier in the acronym should be chosen as the 'U' compared to the functions that occur later in the acronym. So, for example, if you're trying to integrate an inverse sin $\times$ a polynomial, choose the inverse sin as the 'U' since it's an inverse trig function, and I comes before A in LIATE.

Example: Consider $f(x) = \int xe^x$. Let $U = x$, $dv = e^x dx$. Then, $du = 1dx$ and $V = e^x$. So, we can write $f(x) = xe^x - \int e^x dx = xe^x - e^x + C$.

Change of Variables

In the single-dimensional case, the change of variables method is often refereed to as **u-substitution**. Briefly, this technique works by defining a part of the integrand u (a new variable), and rewriting the rest of the integrand in terms of u.

Rule of Thumb: U substitution is typically useful when one part of the integrand can be thought of as a derivative of another part of the integrand. Often, the u is the part of the integrand that is "inside" another part of the function (i.e. typically, u will be the part of the integrand that's raised to a power, in the denominator, or exponentiated).

Example: Consider $f(x) = \int 3x^2(x^3+5)^7 dx$. Define $u = x^3 + 5$. Differentiating, we have that $du = 3x^2 dx$. Then, we can rewrite the expression $f(x) = \int u^7 du = \frac{u^8}{8} + C$, and substituting back in order rewrite the expression in terms of x, we have $\frac{(x^3+5)^8}{8} + C$.

Fundamental Theorem of Calculus

Intuition: The derivative of the integral of a function is the function, so derivatives are like inverse integrals (and vice-versa).

Theorem: (First Fundamental Theorem of Calculus) Let f be a continuous real-valued function defined on a closed interval $[a, b]$. Let F be the function defined, $\forall\ x \in [a, b]$, by

$$F(x) = \int_a^x f(t)dt$$

Then, F is uniformly continuous on $[a, b]$, differentiable on the open interval (a, b) and $F'(x) = f(x)\ \forall\ x \in (a, b)$.

Note: It's often convenient to think of the conclusion to this theorem as

$$\frac{\partial}{\partial x}\int_a^x f(t)dt = f(x).$$

Example: Let $g(x) = \int_{19}^x t^{\frac{1}{3}}$. What is $g'(27)$?

Since $t^{\frac{1}{3}}$ is everywhere continuous, it will be be continuous on the interval from 19 to x. Then, applying the second fundamental theorem of calculus, $g'(x) = x^{\frac{1}{3}}$, so $g'(27) = 3$.

Example: Find the derivative of $f(x) = \int_1^{\sqrt{x}} \frac{t^2}{t^2+1}dt$.

This calls for applying the chain rule and the fundamental theorem of calculus. Specifically, think of f(x) as composed of $g(x) = \int_1^x \frac{t^2}{t^2+1}dt$ and $h(x) = \sqrt{x}$, so $f(x) = g(h(x))$. Then, we have $f'(x) = g'(h(x))h'(x)$

$$= \frac{\sqrt{x}^2}{(\sqrt{x}^2+1)}\frac{1}{2}x^{-\frac{1}{2}}$$
$$= \frac{\sqrt{x}}{2x+2}$$

Theorem: (Second Fundamental Theorem of Calculus) Let f be a real valued function on a closed interval $[a, b]$, and F an antiderivitive of f in (a, b):

$$F'(x) = f(x).$$

If f is Riemann Integrable on $[a, b]$, then

$$\int_a^b f(x)dx = F(b) - F(a)$$

Example: What is $\int_0^3 x^2$?

If $f(x) = x^2$, then $F(x) = \frac{x^3}{3} + C$. Then, $F(3) - F(0) = \frac{3^3}{3} - 0 = 9$.

Note: What if x appears in the lower integral? Recall that

$$\int_x^a f(t)dt = -\int_a^x f(t)dt$$

See this link for a nice explainer.

Separable Differential Equations

Differential equations are equations that relate functions to their derivatives. In contrast to standard algebraic equations, the solution to a differential equation is a function or set of functions.

In order to solve simple separable differential equations, we will use a slight abuse of notation and treat the differential terms dx and dy as variables. Loosely speaking, we call a differential equation separable if we can write the derivative as $\frac{dy}{dx} = f(y)g(x)$, so you have "separated" the algebraic expression into a function of y and a function of x. The only technique that I'll cover for solving differential equations is integrating both sides. The idea is to rewrite the differential equation in the form $f(y)dy = f(x)dx$, and then integrate away the differential terms to get an expression that is just in terms of x and y.

For example, if $\frac{dy}{dx} = 6y^2x$, then

$$\begin{aligned} \frac{1}{y^2}dy &= 6xdx \\ \int y^{-2}dy &= \int 6xdx \\ -y^{-1} &= 3x^2 + C \end{aligned}$$

Does this work all the time? No. The functions you end up with may not be integrable. Additionally, you may have a differential equation that is not separable. However, this kind of problem hopefully shouldn't show up in your first year coursework!

Sequences

Definition: A **sequence** (x_n) of real numbers is a potentially infinite ordered list of numbers $x_n \in \mathbb{R}$ indexed by the natural numbers $n \in \mathbb{N}$.

We can think of every sequence in $\mathbb{R}$ as defined by a function that maps from $\mathbb{N}$ to $\mathbb{R}$.

Definition: A sequence x_n of real numbers converges to a limit $x \in \mathbb{R}$, denoted

$$x = \lim_{n\to\infty} x_n \text{ or } x_n \to x \text{ as } n \to \infty$$

if for every $\epsilon > 0$ there exists $N \in \mathbb{N}$ such that

$$|x_n - x| < \epsilon \; \forall \; n > N$$

Proposition: A sequence must either converge to some unique limit $x \in \mathbb{R}$, otherwise it diverges.

Definition: A sequence x_n of real numbers is **bounded from above (below)** if there exists $M \in \mathbb{R}$ such that $x_n \leq M \; \forall \; n \in \mathbb{N}$ ($x_n \geq M \; \forall \; n \in \mathbb{N}$). A sequence is **bounded** if it is bounded from above and below.

Proposition: A convergent sequence is bounded.

Question: Prove this proposition.

Definition: A sequence of real numbers x_n is **increasing** if $x_{n+1} > x_n \; \forall \; n \in \mathbb{N}$, **decreasing** if $x_{n+1} < x_n \; \forall \; n \in \mathbb{N}$, and **monotone** if it is either increasing or decreasing.

Definition: (Lim Sup/Lim Inf) Let x_n be a sequence of real numbers. Define $y_n = \sup\{x_k : k \geq n\}$ and $z_n = \inf\{x_k : k \geq n\}$. Then the $\limsup_{n\to\infty} x_n = \lim_{n\to\infty} y_n$ and the $\liminf_{n\to\infty} x_n = \lim_{n\to\infty} z_n$

Theorem: A sequence x_n of real numbers converges if and only if $\liminf_{n\to\infty} x_n = \limsup_{n\to\infty} x_n = x$.

Definition: A sequence x_n of real numbers is a **Cauchy sequence** if for every $\epsilon > 0$ there exists $N \in \mathbb{N}$ such that

$$|x_m - x_n| < \epsilon \; \forall \; m, n > N$$

What is special about this definition? It defines a necessary and sufficient condition for the convergence of a sequence using only terms in the sequence and not the eventual limit.

Theorem: A sequence of real numbers converges if and only if it is a Cauchy sequence.

Theorem: Let $f : A \to \mathbb{R}$, $A \subset \mathbb{R}$, $c \in \mathbb{R}$. Then

$$\lim_{x\to c} f(x) = L$$

if and only if

$$0 < |x_n - c| < \delta \implies |f(x) - L| < \epsilon$$

for every sequence $(x_n) \in A \setminus \{c\}$ such that

$$\lim_{n\to\infty} x_n = c$$

Question: Why do we need to exclude c as an element of the sequence x_n in the above definition?

This theorem provides a tool for demonstrating that a limit does not exist. For example, if $f(x_n)$ does not converge, or if $f(x_n) \neq f(y_n)$ for two sequences x_n and y_n that both approach c, then we know that the limit as x approaches c does not exist.

Example: What is the limit of $f(x) = sin(\frac{1}{x})$ as $x \to 0$? For this example, it's useful to look at a graph. The limit is undefined.

To show this, consider the two sequences $x_n = \frac{1}{2\pi n}$ and $y_n = \frac{1}{2\pi n + \frac{\pi}{2}}$. Both of these sequences converge to 0 as n grows large, but the limits are different:

$$\lim_{n\to\infty} f(x_n) = 0 \text{ but } \lim_{n\to\infty} f(y_n) = 1$$

Theorem: If $f : A \to \mathbb{R}$ and $c \in A$, then f is continuous at c if and only if

$$\lim_{n\to\infty} f(x_n) = f(c)$$

for every sequence (x_n) in A such that $x_n \to c$ as $n \to \infty$.

Example: A sequence characterization is useful for proving that a function is not continuous. Consider the sign function,

$$sign(x) = \begin{cases} -1 & \text{if } x < 0 \\ 0 & \text{if } x = 0 \\ 1 & \text{if } x > 0 \end{cases}$$

To show that this function is discontinuous at $x = 0$, we can consider the sequences $x_n = \frac{1}{n}$ and $y_n = -\frac{1}{n}$ which both converge to zero, though $\lim_{n\to\infty} f(x_n) = -1 \neq 1 = \lim_{n\to\infty} f(y_n)$.

Question:

$$f(x) = \begin{cases} \sin(\frac{1}{x}) & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{cases}$$

Is f(x) continuous?

$$g(x) = \begin{cases} x\sin(\frac{1}{x}) & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{cases}$$

Is g(x) continuous?

Proposition: A function $f : A \to \mathbb{R}$ is **not** uniformly continuous on A if and only if there exists $\epsilon_0 > 0$ and sequences x_n, y_n such that

$$\lim_{n\to\infty} |x_n - y_n| = 0 \text{ and } |f(x_n) - f(y_n)| \geq \epsilon_0 \; \forall \; n \in \mathbb{N}$$

Example: Define $f : [0, 1] \to \mathbb{R}$ by

$$f(x) = \begin{cases} sin(\frac{1}{x}) & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{cases}$$

Now consider the two sequences $x_n = \frac{1}{2\pi n}$ and $y_n = \frac{1}{2\pi n + \frac{\pi}{2}}$., we can show that the difference between these two sequences that converge to 0 is constant (1).

$$|f(x_n) - f(y_n)| = |sin(2n\pi + \frac{\pi}{2}) - sin(2n\pi)| = 1 \; \forall \; n \in \mathbb{N}$$

Definition: For a sequence x_n, a subsequence x_{n_k} is a sequence constructed by selecting a subset (infinite) of the terms of x_n and maintaining the order of those terms.

Example: If we consider the natural numbers as a sequence $1, 2, 3, 4, 5, ...$ the even numbers and odd numbers are subsequences.

Theorem (Bolzano-Weierstrass): Every bounded sequence of real numbers has a convergent subsequence.

Question: Sketch a proof for this theorem.

Bolzano-Weierstrauss is a theorem that is about the "compactness" of $\mathbb{R}$. It guarantees us a convergent subsequence of any bounded sequence without knowing anything about the limit of that sequence.

Proposition: A set $F \subset \mathbb{R}$ is closed if and only if the limit of every convergent sequence in F belongs to F.

Question: Prove this proposition.

Definition: A set $K \subset \mathbb{R}$ is **sequentially compact** if every sequence in K has a convergent subsequence whose limit belongs to K. Typically, we just say the set is compact.

Theorem (Bolzano-Weierstrass): A subset of $\mathbb{R}$ is sequentially compact if and only if it is closed and bounded.

Example: Every closed interval in $\mathbb{R}$ is compact.

Example: $(0, 1)$ is not compact. Consider the sequence $\frac{1}{n}$.

Example: The set N in $\mathbb{R}$ is closed, but not compact. Consider the sequence $x_n = n$ which diverges.

Definition: Let $A \subset \mathbb{R}$. A cover of A is a collection of sets $\{A_i \subset \mathbb{R} : i \in I\}$ such that

$$A \subset \bigcup_{i \in I} A_i$$

Example: Consider $A_i = (\frac{1}{n}, 2)$. This covers the interval $(0, 1]$.

Definition: Suppose that C is a cover of $A \subset \mathbb{R}$. A subcover S of C is a sub-collection $S \subset C$ that covers A, meaning that

$$S = \{A_{i_k} \in C : k \in J\} \text{ with } A \subset \bigcup_{k \in J} A_{i_k}$$

Definition (Topological Compactness): A set $K \subset \mathbb{R}$ is compact if every open cover of K has a finite subcover.

Theorem (Heine-Borel): A subset of $\mathbb{R}$ is compact if and only if it is closed and bounded.

If Heine-Borel and Bolzano-Weierstrass sound the same to you, that's good, because they are in $\mathbb{R}$! (In other spaces, these two definitions of compactness can be different, but this should not come up.)

Example: Consider trying to cover the natural numbers $\mathbb{N}$. We could pick intervals around N. For example, $A_i = \{i - 1, i + 1 : i \in \mathbb{N}\}$, but there is no finite sub-cover that will cover all of the natural numbers.

Hopefully, this example gives you the intuition for the relatedness of 'finite sub-covers' and boundedness.

Continuous functions on compact sets:

Continuous functions on compacts sets are bounded, they attain their max and min, and they are uniformly continuous. The class of functions we are typically thinking about here are $f : [a, b] \to \mathbb{R}$.

Theorem: If $K \subset \mathbb{R}$ is compact and $f : K \to \mathbb{R}$ is continuous, then $f(K)$ is compact.

Question: Prove this theorem.

Theorem: If $f : K \to \mathbb{R}$ is continuous and $K \subset \mathbb{R}$ is compact, then f is uniformly continuous on K.

Weierstrauss Extreme Value Theorem: If $f : K \to \mathbb{R}$ is continuous and $K \subset \mathbb{R}$ is compact, then f attains its maximum and minimum.

Example: Consider $f(x) = \begin{cases} \frac{1}{x} & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{cases}$. This shows a case where the theorem doesn't apply because the function is not continuous.

Rolle's Theorem: Suppose that $f : [a, b] \to \mathbb{R}$ is continuous on the closed, bounded interval $[a, b]$ and differentiable on the open interval (a, b), and $f(a) = f(b)$. Then there exists some $a < c < b$ such that $f'(c) = 0$.

This is an application of the Weierstrauss extreme value theorem: the function must attain it's max and min. If the endpoints are both the max and the min, then the function is flat between a and b so the derivative is 0 everywhere. Otherwise, at least one of the max and the min is attained somewhere in between a and b, so we can use the above result about local maxima and minima to ensure that the derivative is zero somewhere in the interval.

Mean Value Theorem: Suppose that $f : [a, b] \to \mathbb{R}$ is continuous on the closed, bounded interval $[a, b]$ and differentiable on the open interval (a, b). Then there exists some $a < c < b$ such that

$$f'(c) = \frac{f(b) - f(a)}{b - a}$$

Aside: The geometric series trick: $\sum_0^\infty a^n = \frac{1}{1-a}$.

Proof:

$$\begin{aligned} S &= \sum_0^\infty a^n \\ aS &= \sum_0^\infty a * a^n = \sum_1^\infty a^n \\ S - aS &= 1 \\ S &= \frac{1}{1-a} \end{aligned}$$

Sequences of Functions

Definition: Suppose that f_n is a sequence of functions $f_n : A \to \mathbb{R}$ and $f : A \to \mathbb{R}$. Then we say that f_n **converges pointwise** to f if $f_n(x) \to f(x)$ as $n \to \infty$ for every $x \in A$.

We can write this $f(x) = \lim_{n\to\infty} f_n(x)$.

Question: Consider $f_n(x) : (0, 1) \to \mathbb{R}$, $f_n(x) = \frac{n}{nx+1}$. Does the limit exist, and if so, what is it?

Definition: Suppose that a sequence of functions $f_n : A \to \mathbb{R}$ and $f : A \to \mathbb{R}$. Then, $f_n \to f$ uniformly on A if, for every $\epsilon > 0$ there exists $N \in \mathbb{N}$ such that

$$n > N \implies |f_n(x) - f(x)| < \epsilon \; \forall \; x \in A$$

Definition: A sequence f_n of functions is uniformly Cauchy on A if for every $\epsilon > 0$ there exists $N \in \mathbb{N}$ such that

$$m, n > N \implies |f_m(x) - f_n(x)| < \epsilon \; \forall \; x \in A$$

Theorem: A sequence of functions converges uniformly on A if and only if it is Cauchy on A.

Taylor Expansion

In general, a series is the sum of terms from a sequence. The Taylor Expansion is a sum of a sequence of functions.

You can think of a Taylor series as a way to approximate a complicated function using a polynomial. Consider approximating a function $g(x)$ around the point $x = 0$ using a 2nd degree polynomial $f(x) = c_0 + c_1x + c_2x^2$.

When we're expanding around the point 0, the nice thing about this kind of polynomial approximation is that the constant c_0 controls the value, the constant c_1 controls the first derivative, and the constant c_2 controls the second derivative. To see this, write:

$$\begin{aligned} f(x) &= c_0 + c + 1x + c_2 * x^2 \\ f'(x) &= c_1 + 2c_2x \\ f''(x) &= 2c_2 \end{aligned}$$

and plugging in $x = 0$, we have

$$\begin{aligned} f(0) &= c_0 \\ f'(0) &= c_1 \\ f''(0) &= 2c_2 \end{aligned}$$

By repeatedly applying the power to take derivatives, we can see that only the n^{th} constant affects the n^{th} derivative. To be clear, this is because terms c_mx^m with $m < n$ drop out when you've taken n derivatives, and the choice of $x = 0$ means that terms c_kx^k with $k > n$ drop out, which ensures that only the constant on the n^{th} term affects the value of the n^{th} derivative.

The idea of a Taylor series is to create a good approximation of a function near a point by 'matching derivatives', so for example, if we want to approximate the function $cos(x)$ around the point 0, we know that $cos(0) = 1$. The first derivative of $cos(x)$ is $-sin(x)$, and $-sin(0) = 0$. The second derivative of $cos(x)$ is $-cos(x)$, and $-cos(0) = -1$. So, if we want to construct a 2nd degree polynomial whose derivatives match $cos(x)$ around the point zero, we can set $c_0 = 1$, $c_1 = 0$, and $2c_2 = -1 \implies c_d = \frac{-1}{2}$, yielding $f(x) = 1 - \frac{1}{2}x^2$.

What is the general rule for selecting the constant on the n^{th} term, c_n? To see this, think about applying the power rule to the polynomial $f(x) = c_nx^n$ n times. If we were thinking of the 5th term, then we'd have the first derivative of $f'(x) = 5x^4$, the second derivative $f''(x) = 5 * 4x^3$, and so on, so in general, the power rule will give the solution $f^n(x) = n!c_n$. Since we're matching derivatives of the function $g(x)$, what

we want is to choose the constant c_n so that the n^{th} derivative of our polynomial approximation $f^n(x)$ equals the n^{th} derivative of $g(x)$, so we pick $c_n = \frac{g^n(x)}{n!}$.

What if you want to approximate a function around the point a, where $a \neq 0$? To maintain this nice property where one constant controls only the n^{th} constant affects the n^{th} derivative, you want it to be the case that $x = a$ ensures that all terms of power greater than n drop out, so choosing a polynomial "centered around a" does the trick. Specifically, consider $f(x) = c_0 + c_1(x-a) + c_2(x-a)^2 + c_3(x-a)^3 + ...$ When we take derivatives with respect to x, lower order terms will drop out like before, and evaluation the function at a will ensure that higher order terms also drop out.

Putting these two ideas together, we have the general equation for a Taylor Series f(x) approximating a function g(x) given by:

$$f(x) = \sum_{n=0}^{\infty} c_n(x-a)^n$$

where

$$c_n = \frac{g^{(n)}(a)}{n!}$$

If you would like a video explanation of Taylor series, check out this pretty comprehensive 3Blue1Brown YouTube video.

Metric Spaces and the Triangle Inequality

Definition: A **metric space** is an ordered pair (M, d) where M is a set and d is a metric on M. That is, $d : M \times M \to \mathbb{R}$ wit the following properties:

- Identity of Indiscernibles: $d(x, y) = 0 \iff x = y$
- Symmetry: $d(x, y) = d(y, x)$
- Triangle Inequality: $d(x, z) \leq d(x, y) + d(y, z)$

Intuitively, a metric is just a distance function defined over a set that takes on the standard properties we associate with distances. These three properties imply a fourth property, which is also intuitive: distance must be positive. To see this,

$$\begin{aligned} d(x,y) + d(y,x) &\geq d(x,x) \text{ by triangle inequality} \\ d(x,y) + d(x,y) &\geq d(x,x) \text{ by symmetry} \\ 2d(x,y) &\geq 0 \text{ by identity of indiscernibles} \\ d(x,y) &\geq 0 \end{aligned}$$

so we have non-negativity.

The "simplest" metric is the **discrete metric**, which takes on the value $d(x, y) = 0$ if $y = x$ and $d(x, y) = 1$ otherwise. This metric demonstrates that for any set, there is at least one valid metric. In probability space, there is the **Wasserstein** metric,

which is sometimes referred to as the earth mover's distance. If we visualize two probability distributions as piles of dirt, this metric is given by the minimum "cost" to convert one pile to the other, which is assumed to be the amount of dirt that needs to be moved times the distance it has to be moved.

While this is a set of definitions that makes sense for any space, we are typically thinking about Euclidean space (sometimes, we care about function space). Recall our definition from Linear Algebra:

Definition: A non-empty set $X \in \mathbb{R}$ is a vector space if it's closed under addition and scalar multiplication (i.e. for $x, y \in X$ and $c \in \mathbb{R}$ we have $x+y \in X$ and $cx \in X$).

Definition: A **norm** is a function $p : X \to \mathbb{R}$ satisfying the following three properties:

- Triangle Inequality (subadditivity): $p(u + v) \leq p(u) + p(v)$
- Homogeneous of Degree One: $p(av) = |a|p(v)$
- If $p(v) = 0$ then $v = 0$, the zero vector

We call this pair of a space V and a norm p a **normed vector space**. A normed vector space is a particular type of metric space that you're used to using. You are likely familiar with a few kinds of norms.

In $\mathbb{R}$, the absolute value distance metric is given by $d(x, y) = |x - y|$. The Euclidean distance metric in n dimensional space for two points $p, q \in \mathbb{R}^n$ is given by $d(p, q) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + ... + (q_n - p_n)^2}$. This distance metric is the the reason we call the triangle inequality the triangle inequality.

When do we use the triangle inequality? When we want to bound things. We used it a few times over the course of the last few lectures, such as in the multiplying limits problem. The version of it that is common to use in simple bounding exercises is given by: $|x + y| \leq |x| + |y|$

A vector space only requires addition and scalar multiplication to be defined. If we also define the dot product (inner product), then we have a **inner product space**.

Recall our definition of a dot product:

Definition: The **dot product (inner product)** of two vectors a and b of length n is given by $\langle x, y\rangle = \sum_{i\in 1}^{n} a_i b_i$.

For a graphical intuition, note that this is equivalent to $||a||||b||cos(\theta)$ where θ is the angle between the points. The $cos(\theta)$ comes from projecting a onto b or vice versa.

Cauchy-Schwartz Inequality: For all vectors u and v, $\langle x, y\rangle^2 \leq \langle x, x\rangle\langle y, y\rangle$.

If we take square roots, we could write this inequality in terms of the Euclidean norm: $|\langle x, y\rangle| \leq ||x||||y||$

Intuition: In $\mathbb{R}^2$, we can think about the dot product in terms of the angle θ between vectors: $\langle x, y\rangle^2 = (||x||||y||cos(\theta))^2 \leq (||x||||y||)^2$ which makes sense because $cos(x) \leq$

1.

Question: Prove that the triangle inequality is a consequence of the Cauchy-Schwartz inequality.

Example: Fixed Point Theorems

Contraction Mapping Theorem (Banach-Caccioppoli Fixed-point Theorem): Let (X, d) be a complete metric space. We say the function $T : X \to X$ is a **contraction mapping** on X if there exists $q \in [0, 1)$ such that

$$d(T(x), T(y)) \leq qd(x, y)$$

for all $x, y \in X$

Then, T admits a unique fixed-point x^* in X (i.e. $T(x^*) = x^*$). Furthermore, we can find x^* as follows: start with an arbitrary element of X, x_0 and consider the sequence $x_n = T(x_{n-1})$ for $n \geq 1$. Then, $x_n \to x^*$.

Proof: The idea of this proof is to show that x_n is a Cauchy sequence. Take any two points x_m and x_n. Let $\epsilon > 0$.

$$\begin{aligned} d(x_m, x_n) &\leq d(x_m, x_{m-1}) + ... + d(x_{n+1}, x_n) \\ &\leq q^{m-1}d(x_1, x_0) + q^{m-2}(x_1, x_0) + ... + q^n(x_1, x_0) \\ &= q^n d(x_1, x_0) \sum_{k=0}^{m-n-1} q^k \\ &\leq q^n d(x_1, x_0) \sum_{k=0}^{\infty} q^k \\ &= q^n d(x_1, x_0)(\frac{1}{1-q}) \end{aligned}$$

Choose N large enough that

$$q^N < \frac{\epsilon(1-q)}{d(x_1, x_0)} d(x_1, x_0)(\frac{1}{1-q}) = \epsilon$$

so we have that x_n is Cauchy. Since (X, d) is complete, the sequence limit is in X. Since the contraction mapping is continuous, $x^* = lim_{n\to\infty} x_n = lim_{n\to\infty} T(x_{n-1}) = T(lim_{n\to\infty} x_{n-1}) = T(x^*)$. Since this is a convergent, Cauchy sequence, we know that the limit is unique.

Context: This theorem can be used to prove the existence and uniqueness of Cournot competition (duopoly competition over output), and other dynamic economic models. The idea is that the strategy set that the firm takes when best responding is a contraction mapping.

Definition: A **convex** set is a set such that every for every pair of points in the set, the line connecting them is contained within the set.

The relationship between this definition and our original definition of a convex function is that the epigraph (the set of points above the graph of the function) on a fixed interval forms a convex set.

Definition: A **continuous mapping** is a a function that maps convergent sequences to convergent sequences. That is, if $x_n \to x$ then $g(x_n) \to g(x)$.

A simple example of a continuous mapping is flipping or rotating.

Brouwer's Fixed-point Theorem: For any continuous function f, mapping a compact convex set to itself there is a point x_0 such that $f(x_0) = f(x_0)$.

Intuition: mixing around a cup of coffee.

This proof is difficult, so we won't dwell on it here.

Kakutani Fixed-point Theorem: Let S be a non-empty, compact and convex subset of some Euclidean space $\mathbb{R}^n$. Let $\phi : S \to 2^S$ be a set-valued function on S with the following properties:

- ϕ has a closed graph
- $\phi(x)$ is non-empty and convex for all $x \in S$

Then, ϕ has a fixed point.

Historical Note: Proved by Shizuo Kakutani in 1941, and used by Nash in his description of Nash equillibria. Eventuallly used in the prof of existence of general equilibrium in market economies (Arrow-Debreu markets, covered by Reny in Price Theory II).

Probability

If you find yourself in need of a more formal treatment of probability theory, the following set of lectures from MITOpenCourseware is good but dense. Not recommended unless you actually need very formal probability tools.

Probability Spaces

We want to think about uncertain outcomes. For example, the roll of a dice, the flip of a coin. In more natural context, we might want to know how many people will buy a product, or how much rain will fall tomorrow.

We will define a **probability space** using a the triple $\Omega, \mathcal{F}, P$ which provides a formal model of a random process, sometimes called an experiment.

Ω is the sample space, which is the set of all possible outcomes. For example, for a single roll of die, this would be then numbers 1 through 6. An out come is the result of single execution of the model.

$\mathcal{F}$ is called a σ-algebra, and is the space of all events. Typically, we're thinking about the set of all subsets of the sample space. For example, we might want to know the probability that a die lands on a 5 which is the set $\{5\}$, but we also might want to know the probability that a die lands on an even number, which is the set $\{2, 4, 6\}$. Formally

- $\mathcal{F}$ contains the sample space: $\Omega \in \mathcal{F}$
- $\mathcal{F}$ is closed under complements: if$A \in \mathcal{F}$, then $\Omega \setminus A \in \mathcal{F}$
- $\mathcal{F}$ is closed under countable unions: If $A_i \in \mathcal{F}$, then $U_{i=1}^{\infty} \in \mathcal{F}$

You could apply De Morgan's laws to show that $\mathcal{F}$ is closed under countable intersections as well.

$P : \mathcal{F} \to [0, 1]$ is a function that assigns a probability to each event in the event space. By definition, probabilities are numbers between 0 and 1 that describe how likely an event is to occur. In order for P to be a valid probability function, it must satisfy two requirements:

- Countable Additivity: The countable union of mutually exclusive events must be equal to the sum of the probabilities of those events.
- The probability measure of the entire sample space Ω must be equal to 1.

Random Variables

Loosely speaking, a **random variable** is a *function* that maps from the set of possible (random or chance-based) events to the real numbers. For example, consider the coin flip. X is a random variable that maps from the outcome of a random event (heads or tails) to a numerical value:

$$X = \begin{cases} 1 & \text{if Heads} \\ 0 & \text{if Tails} \end{cases}$$

For example, consider tossing a coin five times, and denote tales as 0 and heads as 1. We can represent the set of possible outcomes as vectors of zeros and ones, e.g. $\Omega = \{0, 1\}^5$.

Notice, this is not a variable in the algebraic sense! It is also not random!!

A closely related concept is a **probability distribution function (pdf)** (called probability density function in the continuous case).

Definition: The **probability density function (pdf)**, also called the density of a random variable, is a function that represents the relative likelihood of a random

variable taking on a particular value. The probability density function is non-negative everywhere, and the integral over it's support must be equal to 1.

Notice: For a continuous function, the probability that a random variable takes on any particular value is 0.

The integral from any a to b of the pdf gives the probability that the random variable will take on a value from a to b.

Definition: The **cumulative distribution function (cdf)** of a real-valued random variable X, evaluated at x, is the probability that X will take a value less than or equal to x. Formally, $F_X(x) = Pr(X \leq x)$. Every CDF is non-decreasing and right-continuous. Furthermore, $\lim_{x \to -\infty} F_X(x) = 0$ and $\lim_X(x) = 1$.

If X is a discrete random variable, then it's CDF will be discontinuous at the values it attains. If X is a continuous random variable, then $F_X(a) - F_X(b) = Pr(a < X \leq b)$.

Example: Shifting CDFs

Example: Uniform distribution

Definition: A **random vector** is a vector where each entry is a random variable.

Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of random vectors and X be a random vector on $\mathbb{R}^k$.

Definition (Convergence in Distribution): $X_n \xrightarrow{d} X$ if $Pr(X_n < x) \to Pr(X < x)$ for all continuous points of $x \to P(X \leq x)$.

Definition (Convergence in Probability): $X_n \xrightarrow{p} X$ if $Pr(|X_n - X| \geq \epsilon) \to 0 \forall \epsilon > 0$.

Definition (Almost Sure Convergence): $X_n \xrightarrow{as} X$ if $Pr(\lim_{n \to \infty} X_n = X) = 1$

Theorem: If $X_n \xrightarrow{as} X$, then $X_n \xrightarrow{p} X$. If $X_n \xrightarrow{p} X$ then $X_n \xrightarrow{d} X$. The reverse directions are not true in all cases.

Example: To see that convergence in distribution does not necessarily imply convergence in probability, think about defining two variables that take on opposite values for a coin flip. To be specific, consider

$$X = \begin{cases} 1 & \text{if heads} \\ 0 & \text{if tails} \end{cases}$$

$$Y = \begin{cases} 0 & \text{if heads} \\ 1 & \text{if tails} \end{cases}$$

If we write out these two random variables' CDFs, they will be identical, so X and Y will converge in distribution. However, $|X - Y| = 1$ because the variables are perfectly uncorrelated, so they won't converge in probability.

Example: Normal Distributions. Let $X \sim N(0, 1)$. Define $Y = -X$. Y converges in distribution to X, but Y does not converge in probability to X. Notice that this

flipping of the normal distribution preserves convergence in distribution because the normal distribution is symmetric around 0. To see this, let's write out how these two points are distributed:

$$\begin{bmatrix} X_n \\ Y_n \end{bmatrix} = \begin{bmatrix} X \\ -X \end{bmatrix} \xrightarrow{d} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}\right)$$

but

$$\begin{bmatrix} X \\ X \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}\right)$$

Expectations

Definition (Expectation): The expected value of a random variable is the average of it's realizations weighted by their probability. For a discrete random variable, this is given by $\mathbb{E}[X] = \sum_{i=1}^{k} x_i p_i = x_1 p_1 + x_2 p_x + ...x_k p_k$. Generalizing to a continuous random variable, we have, $\mathbb{E}[X] = \int x f(x) dx$ where f(x) is the pdf of X.

Thinking about this definition in terms of probability space, we have $\mathbb{E}[X] = \int_\Omega X(\omega) dP(\omega)$.

Definition (Moment): The n^{th} moment of a real valued continuous function f(x) is given by $\mathbb{E}[x^n] = \int x^n f(x) dx$.

The first moment of a distribution is it's mean. The second moment is it's variance. The third moment is it's skewness. This should give you a sense that moments tell you something about the "shape" of a distribution.

Definition (Centered Moment): The n^{th} centered moment is given by $\mathbb{E}[(X - \mathbb{E}[X])^n]$.

Definition: We say that **a moment exists** if $\mathbb{E}[|X^n|] < \infty$.

Proposition: The existence of higher moments implies the existence of lower moments. Formally, let X be a random vector. Then,

$$\mathbb{E}[|X|^k] < \infty \implies \mathbb{E}[|X|^j] < \infty$$

Working with expectations: $\mathbb{E}[cX] = c\mathbb{E}[X]$. $Var[cX] = c^2 Var[X]$. You can prove this to yourself by writing out the moment definitions and substituting cX for X.

Definition: Intuitively, a conditional expectation is the expectation of a variable given that certain conditions are met. Notationally, $\mathbb{E}[X|Y = y]$ for example. Formally, we can write this expectation

$$\mathbb{E}[X|H] = \int x dP(x|H)$$

Law of Iterated Expectations: Suppose X and Y are random variables and $\mathbb{E}[X]$ exists. Then $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]]$

Intuition: Let's say X represents the quantity of rain and Y represents whether it's a cloudy day.

The average quantity of rain is equal to the quantity of rain on cloudy days times the probability it's a cloudy day + the average quantity of rain given that it's not a cloudy day times the quantity of rain when it's not a cloudy day.

Useful Inequalities

Cauchy-Schwartz Inequality (Again!): Let X, Y be two random variables. Then,

$$|\mathbb{E}[XY]|^2 \leq \mathbb{E}(X^2)\mathbb{E}(Y^2)$$

This implies that $Cov(X,Y) \leq Var(X)Var(Y)$, which is sometimes called the covariance inequality.

Jensen's Inequality: Let $I \subset \mathbb{R}$ be a convex set and $f : I \to \mathbb{R}$ be a convex function. Then, for any random variable X such that $Pr(x \in I) = 1$, $\mathbb{E}[|X|] < \infty$ and $ex[|f(X)|] < \infty$, then

$$f(\mathbb{E}[X]) < \mathbb{E}[f(X)]$$

If f is a concave function, then the above inequality is reversed.

Example: Think about the one dimensional case (for example, $f(x) = x^2$). This theorem says that for a convex function, a line between any two points a and b on the curve lies above the curve. The line represents the expectation (expectations are linear averages), and the curve represents what you would get by averaging first and then evaluating the function at the average point. If we think about a concave function (for example $f(x) = -x^2$), we can see why the inequality reverses: now, the line between any two points on the graph lies below the graph.

Question: Prove that higher moments imply the existence of lower order moments.

Markov's Inequality: Let X be a random variable. Then,

$$P(|X| \geq \epsilon) \leq \frac{\mathbb{E}[|X|^q]}{\epsilon^q} \; \forall \; q, \epsilon > 0$$

where $|\cdot|$ is the Euclidean norm.

First, we can convert probability to the expectation of the indicator function: $Pr(|X| > \epsilon) = \mathbb{E}[\mathbb{I}(|X| > \epsilon)]$. So we need to show:

$$\mathbb{I}(|X| > \epsilon) \leq \frac{|X|^q}{\epsilon^q}$$

We can prove this going case by case, since the indicator function takes on the values 0 and 1.

Case 1: The indicator function takes on the value 0. We want to show that $\mathbb{I}(|X| > \epsilon) \leq \frac{|X|^q}{\epsilon^q}$. Since the indicator equals 0, we want to show that $0 \leq \frac{|X|^q}{\epsilon^q}$. We can show that the right hand side of the inequality is weakly positive by noting that $\frac{|X|}{\epsilon} > 0$ since $|X| > 0$ and $\epsilon > 0$. Then, note that raising a weakly positive number to a positive power $q > 0$ gives a weakly positive number.

Case 2: The indicator function takes on the value 1. Again, we want to show that $\mathbb{I}(|X| > \epsilon) \leq \frac{|X|^q}{\epsilon^q}$. Here, we know that the left hand side of the inequality equals one, so we want to show that $1 \leq \frac{|X|^q}{\epsilon^q}$. Since thei ndicator function is 1, we know that $|X| > \epsilon$. Then, $\frac{|X|}{\epsilon} > 1$. So $\frac{|X|^q}{\epsilon^q} > 1$ since raising numbers above 1 to any positive power yields numbers above 1.

Since the inequality holds in both cases, it holds in general. To complete the proof, take expectations of both sides:

$$P(|X| \geq \epsilon) = \mathbb{E}[\mathbb{I}(|X| > \epsilon)] \leq \mathbb{E}[\frac{|X|^q}{\epsilon^q}] = \frac{\mathbb{E}[|X|^q]}{\epsilon^q}$$

When do we use Markov's inequality? The key point is that it moves us from an inequality to an expectation. Often, when you're trying to prove something about convergence in probability, you'll find Markov's inequality useful because you can move from $Pr(|X_n - X| > \epsilon) \leq \frac{\mathbb{E}[X - X_n]}{\epsilon}$.

Key Distributions

Uniform Distribution

A uniform distribution is defined along some interval (a,b), and is typically denoted $U(a, b)$. All points $x \in (a, b)$ have an equal probability. The pdf of a uniform distribution is given by

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } x \in (a, b) \\ 0 & \text{otherwise} \end{cases}$$

$$var(x) = (x - \mu_x)^2$$

Binomial Distribution

Definition: A **Bernoulli process** is a sequence of random variablies X_n such that the value of X is either 0 or 1, and the probability that X = 1 is constant, and given by $p \in [0, 1]$.

Coin flipping is a Bernoulli process. This process is memoryless, so knowing what happens at X_n gives you no information about X_{n+1}. (This relates to the 'hot hand' fallacy.) The expected value of a Bernouli random variable is p.

A Binomial distribution is defined by the probability of getting exactly k successes in n independent Bernoulli trials. The pdf of this object is given by

$$f(k, n, p) = \binom{n}{k} p^k (1-p)^{n-k}$$

where

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

The expected value of a Binomial random variable is np.

The variance is given by $np(1-p)$.

Normal Distribution

The pdf of a normal distribution is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

A **standard normal** distribution is a distribution with a mean of 0 and a standard deviation of 1.

Let's say $X \sim N(\mu, \sigma^2)$. How is c +X distributed? $c + X \sim N(\mu + c, \sigma^2)$.

How is cX distributed? $cX \sim N(c\mu, c^2\sigma^2)$

Poisson Distribution

A Poisson distribution is a discrete distribution which is typically used for modeling the number of times a particular event occurs in a given interval of time. (e.g. how many floods per year, how many patients will arrive in the emergency room in the next 20 minutes).

The pmf is given by

$$f(k, \lambda) = Pr(X = k) = \frac{\lambda^k e^- \lambda}{k!}$$

This is a valid model of events occurring if the events are discrete, the events are independent, and two events cannot occur at exactly the same time. Another way of saying this is that the Poisson distribution is memoryless.

Exponential Distribution

The pdf of the exponential distribution is given by

$$f(x, \lambda) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

One way of thinking about the exponential distribution is that it measures the time between events in a Poisson process. Similarly, it is a memoryless distribution.

The expectation of this distribution is $\frac{1}{\lambda}$ and the variance is $\frac{1}{\lambda^2}$.

Note: This is distinct from the idea of the class of distributions called the exponential family, which comes up in the second quarter of econometrics and includes all the previously mentioned distributions.

T1EV (Gumbel) Distribution

This distribution models the maximum of a number of samples of various distributions. It is frequently used in structural models, which show up in macro, marketing, and industrial organization, where assuming that the error term takes this form results in nice properties.

The CDF of this distribution is $F(x, \mu, \beta) = e^{-e^{\frac{-x-\mu}{\beta}}}$.

Probability Theorems

Definition: Two random variables are **independent** if $F_{X,Y}(x, y) = F_X(x)F_Y(y)$.

Another way of saying this definition is $\mathbb{E}[X|Y] = \mathbb{E}[X]$. In words, knowing the value of Y doesn't tell us anything about the value of X.

Definition: Two random variables are **identically distributed** if $F_x(X) = F_y(Y)$ for all x.

Definition: We say variables are **i.i.d.** if they're independent and identically distributed.

Weak Law of Large Numbers (WLLN): If X_n is an i.i.d. sequence of random vectors such that $\mathbb{E}[|X|] < \infty$, then $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{i=n} X_i \xrightarrow{p} \mathbb{E}[X_i]$ as $n \to \infty$.

Definition: We say that a sequence of random vectors X_n is consistent for μ if $X_n \xrightarrow{p} \mu$.

Strong Law of Large Numbers: If X_n is a i.i.d. sequence of random vectors such that $\mathbb{E}[|X|] < \infty$, then $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{i=n} X_i \xrightarrow{as} \mathbb{E}[X_i]$ as $n \to \infty$.

Central Limit Theorem: If X_n is an i.i.d. sequence of random vectors such that $Var[X] < \infty$, then

$$\sqrt{n}(\bar{X}_n - \mathbb{E}[X_i]) \xrightarrow{d} N(0, Var[X_i])$$

as $n \to \infty$.

O Notation

O notation describes the limiting behavior of a function, sometimes referred to as the order of the function.

Definition: A function $f(x) = O(g(x))$ if for some $x_0 \in \mathbb{R}$ and some $M \in \mathbb{R}$, $f(x) < Mg(x)$ for all $x \geq x_0$.

Big O Notation Rules: If a function f(x) is composed of several terms, the term with the largest growth rate is kept and all others are omitted. If a function is the product of several terms, any constants can be omitted.

Example: If we want to write the function $f(x) = 12x^2 - 23x + 14$ in big O notation, we notice that the x^2 has the largest growth rate, and so dropping the constant we have $f(x) = O(x^2)$.

Definition: A function $f(x) = o(g(x))$ if for all $\epsilon > 0$ there exists N such that

$$|f(x)| \leq \epsilon g(x)$$

for all $x \geq N$.

Another way of writing this is that

$$\lim_{x \to \infty} \frac{f(x)}{g(x)} = 0$$

Intuitively, little o notation says that $f(x)$ grows much more slowly than $g(x)$. For the same order, this means little o is a stronger statement than big o: $f(x) = o(g(x)) \implies f(x) = O(g(x))$.

These two concepts can also be defined in the context of probability and random variables. In particular:

Definition (o_p): X_n is $o_p(1)$ if $X_n \xrightarrow{p} 0$. Moreover, we say X_n is $o_p(R_n)$ if $X_n = Y_n R_n$ for some Y_n that is $o_p(1)$.

Definition: A sequence of random vectors on $\mathbb{R}^k$ is **tight** if, for any $\epsilon > 0$, there exists $B > 0$ such that

$$\inf_n Pr(|X_n| < B) \geq 1 - \epsilon$$

Example: if $X_n \sim U[0,1]$ for even numbers and $X_n \sim U[2,3]$ for odd numbers, than choosing $B = 3$ proves the definition of tightness. This should show you that tightness is something like "bounded with probability approaching 1."

Definition (O_p): We say that $X_n = O_p(1)$ if X_n is tight. We say that $X_n = O_p(R_n)$ if $X_n = Y_n R_n$ for some Y_n that is $O_p(1)$.

Markov Chains

Markov Chains are a way of modeling probabilistic events that depend on the state of the world. For example, we can use Markov Chains to think about probability of rain today as depending on whether or not it rained yesterday.

A Markov model is defined by a set of possible **states** of the world as well as a set of **transition probabilities**.

Example: Grad School

Note that Markov models can be rewritten as matricies, with the rows denoting the current states and the columns denoting the future states, so each entry is the transition probability between the current and future state of the world.

A key feature of a Markov process is 'memorylessness': we can model events that happen tomorrow as a function of today's state and no other information - as long as we know what happened yesterday, we don't learn anything additional from the entire history of events before yesterday.

This construction emphasizes how Markov models can help us understand the state of the future, as we can just raise the transition matrix to the n^{th} power to understand the expected future state.

Additionally, this provides a new use case for the diagonalization technique we learned in linear algebra which allowed us to raise matricies to the n^{th} power.

Question: Consider two TA's Walter and Karthik, each of whom is given a fair coin. We ask Walter to continue tossing the coin until he sees $\{Heads, Tails\}$ and call the number of tosses it took for Walter to observe this patter A. We ask Karthik to toss the coin until he sees $\{Heads, Heads\}$ and call the number of tosses B. In expectation, is A or B greater (or are they the same)?

Note: This question is taken from this fun Twitter thread!

Statistical Inference II: Endogeneity

Warm Up

Recall:

Definition: For a random variable X, the expectation is given by $\mathbb{E}[X] = \int Xf(x)dx$.

Definition: For a random variable X, variance is given by $Var[X] = \mathbb{E}[(X - \mathbb{E}[X])^2]$.

Question: Consider a sequence of i.i.d. draws X_i from some distribution with mean μ and variance σ^2.

- Provide a consistent, unbiased estimator for μ. Provide a justification for why the estimator is consistent and unbiased.
- Provide a consistent, unbiased estimator for σ^2. Provide a justification for why the this estimator is consistent and unbiased.

For today, we'll be using some Variance and Covariance math. Recall:

$Var(aX) = a^2 Var(X)$

Definition: For random variables X and Y, covariance is given by $Cov(X, Y) = (X - \mathbb{E}[X])(Y - \mathbb{E}[Y])$

$Cov(A + B, C + D) = Cov(A, C) + Cov(A, D) + Cov(B, C) + Cov(B, D)$

OLS

Recall the setup of Ordinary Least Squares (OLS) linear regression. We have a dataset $\{Y_i, X_i\}_{i=1}^n$, and we consider a linear model. In matrix notation,

$$Y = X'\beta + U$$

or in algebraic notation,

$$Y = x_0 + \beta_1 x_1 + ... + \beta_n x_n + U$$

We have a dependent (outcome) variable $Y \in \mathbb{R}^n$, a set of independent variables denoted $X \in \mathbb{R}^{n \times k}$. We want to estimate the value of $\beta \in \mathbb{R}^k$, a set of parameters that weight the relative importance of each variable in predicting or causing Y, depending on our assumptions ('holding all other things constant, a 1 unit increase in X_i predicts or causes a β_i unit increase Y').

If we assume the conditional expectation of Y is linear in X ($\mathbb{E}[Y|X] = X'\beta$), then $\mathbb{E}[U|X] = 0 \implies \mathbb{E}[XU] = 0$. From here, we can derive the OLS estimator

$$\begin{aligned} 0 &= \mathbb{E}[X'U] \\ 0 &= \mathbb{E}[X'(Y - X\beta)] \\ 0 &= \mathbb{E}[X'Y] - \mathbb{E}[X'X\beta] \\ \beta\mathbb{E}[X'X] &= \mathbb{E}[X'Y] \\ \beta &= \mathbb{E}[X'X]^{-1}\mathbb{E}[X'Y] \end{aligned}$$

If we relax the requirement that the conditional expectation is linear, we can instead think about OLS as the solution to

$$\beta = \underset{\beta \in \mathbb{R}^k}{\arg\min}\, \mathbb{E}[(Y - X\beta)^2]$$

We could show (though we will not) that the solution to this problem is the OLS estimator. This interpretation accords with our graphical intuition for 'minimizing the sum of squared errors' and is the namesake for OLS.

Theorem (Guass-Markov): The Ordinary Least Squares (OLS) estimator has the lowest variance within the class of linear unbiased estimators.

Note: Sometimes this is referred to as the BLUE property of OLS for Best Linear Unbiased Estimator.

Question: Give an example of a more efficient (lower variance) estimator than OLS when we relax the requirement of unbiasedness.

Finally, and most relevant to economics, we could take a casual view of OLS. In this case, we want it to be the case that $\mathbb{E}[XU] = 0$, but that isn't necessarily true.

When justifying OLS as an approach, we're making 4 assumptions:

1. The model is correctly specified. In particular, the *parameters* β enter the specification linearly. Notice that it's not a problem if a *variable* X does not enter linearly.

Example: If X enters quadratically, we could define $Z = X^2$, and run $Y = \beta Z$.

The key issue is if the value of X affects the value of β: we cannot deal with $Y = \beta(X)X$.

Recall our discussion of Cobb-Douglas estimation by taking logs.

Sometimes, when people talk about model specification, they also mean not omitting relevant variables. As we'll see later, this is technically covered in our next assumption.

2. Strict exogeneity: The error is conditionally independent of the data: $\mathbb{E}[U|X] = 0$. Note that mean independence implies mean-zero errors $\mathbb{E}[U] = 0$ and uncorrelatedness $\mathbb{E}[X'U] = 0$.

The primary purpose of this lecture is to explore this assumption, so we'll come back to discuss it in great detail in a bit.

3. The columns of X are linearly independent. This condition is also called "no perfect multicolinearity."

Why do we need this condition? We will show that without linear independence of the columns of X, the OLS estimator will be undefined.

$X \in \mathbb{R}^{n \times k}$ $X'X \in \mathbb{R}^{k \times k}$

Suppose the columns of X are not linearly independent. By definition, there exists some non-zero $v \in \mathbb{R}^k$ such that

$$Xv = 0$$

Then, by the scalar associative property of matrix multiplication,

$$X'Xv = X'(Xv) = 0$$

This means the columns of $X'X$ are not linearly independent, so by the Invertable Matrix Theorem, $X'X$ won't be invertible. But, notice that the OLS estimator $\beta = \mathbb{E}[X'X]^{-1}X'Y$ is not defined if $X'X$ is not invertible.

When does this go wrong in practice? A typical example is forgetting to drop one of your dummy variables. For example, years of schooling, age and date of birth can be colinear if date of birth determines the age you start school.

4. Spherical Errors: $Var[\epsilon|X] = \sigma^2 I$. This condition implies homoskedasticity and rules out autocorrelation. Note that the presence of non-spherical errors does not bias OLS, but it does make it inefficient such that it's no longer the lowest variance estimator (and therefore, Guass-Markov no longer holds).

Example: Census data quality improvement.

Types of Endogeneity

Definition: For a particular economic model, we call a variable is **exogenous** if it's value determined by factors outside of the model, and taken as a "given" from the perspective of the model.

Definition: For a particular economic model, we call a variable **endogenous** if it's value is determined by the model.

Definition: In the context of econometrics, **endogeneity** refers to the idea that there is a correlation between the error term of a model U and the independent (right hand side) variables. A variable X_j is exogenous if $\mathbb{E}[X_j U] = 0$ and is endogenous if $\mathbb{E}[X_j U] \neq 0$.

We will consider three potential causes of endogeneity:

- Measurement Error
- Omitted Variables
- Reverse Causality (Simultaneity)

Measurement Error

We're thinking about measurement error in the independent (right hand side) variable. For example, consider trying to understand the effect of wages on worker productivity. If we survey workers for their annual income, they may choose to round their responses creating a measurement error the true wage X and the reported wage $\tilde{X}$.

Suppose we want to estimate $Y = \beta X + U$. However, we don't have data on X. Instead, we have data on $\tilde{X} = X + \epsilon$ where $\epsilon \sim N(0, 1)$. We assume that the error is independent: $\mathbb{E}[\epsilon|X] = 0$.

Why is this endogeneity? Note that in the regression that we actually run with mismeasured $\tilde{X}$, the regression is given by $Y = \beta\tilde{X} + u$, so $u = Y - \beta\tilde{X}$. The strict exogeneity requirement is $\mathbb{E}[u|X] = 0$, which implies that $Cov[u, X] = 0$. Computing

the covariance, we have

$$\begin{aligned} Cov[u, \tilde{X}] &= Cov[Y - \beta(X + \epsilon), X + \epsilon] \\ &= Cov[Y, X] - \beta Var[X] - \beta Var[\epsilon] \\ &= \beta\sigma_\epsilon^2 \\ &\neq 0 \end{aligned}$$

where we use the fact that $Cov[X, Y] = Cov[X, \beta X + U] = \beta Var[X]$ because $\mathbb{E}[U|X] = 0$.

How does this bias the estimator? Plugging in our mis-measured variable, we have

$$\begin{aligned} \hat{\beta} &\xrightarrow{p} \mathbb{E}[\tilde{X}'\tilde{X}]^{-}1\mathbb{E}[\tilde{X}'Y] \\ &= \frac{Cov[\tilde{X}, Y]}{Var[\tilde{X}]} \end{aligned}$$

Additionally, we can compute $Var[\tilde{X}] = Cov[\tilde{X}, \tilde{X}] = Cov[X + \epsilon, X + \epsilon] = Var[X] + Var[\epsilon]$.

Then, we have

$$\begin{aligned} &= \frac{Cov[\tilde{X}, Y]}{Var[\tilde{X}]} \\ &= \frac{\beta Var[X]}{Var[X] + Var[\epsilon]} \\ &= \frac{\beta}{1 + \frac{\sigma_\epsilon^2}{\sigma_X^2}} \end{aligned}$$

so we know that this estimator must be less than the true β.

Intuition: In the extreme case, the error is very large and essentially "all noise." In this case, we would expect no relationship between the outcome variable Y and the mis-measured independent variable $\tilde{X}$. This kind of bias is called **attenuation bias**.

Reverse Causality and Simultaneity

Revisiting our example concerning worker productivity and wages, consider the following two intuitive claims:

- Higher productivity workers will tend to seek and secure jobs that pay higher wages.
- Paying higher wages may induce higher worker satisfaction leading to higher productivity

In a simple cross-section of data where we observe the true wages and productivity of workers at different firms, we may still have difficulty estimating the causal effect of productivity on wages or wages on productivity because the observed data is the product of both relationships simultaneously.

To formalize this, consider:

$$y_i = \beta_1 x_i + \gamma_1 z_i + u_i$$
$$z_i = \beta_2 x_i + \gamma_2 y_i + v_i$$

Now, suppose we want to estimate the regression for y_i. First, let's show that this expression will display endogeneity. To do this, we need to solve for z_i. Plugging our expression for y_i into the equation for z_i and grouping terms, we get

$$z_i = \frac{\beta_2 + \gamma_2\beta_1}{1 - \gamma_1\gamma_2} + \frac{1}{1 - \gamma_1\gamma_2}v_i + \frac{\gamma_2}{1 - \gamma_1\gamma_2}u_i$$

Now, when we write out the expression for exogeneity, we'll have $\mathbb{E}[u|Z] = 0 \implies Cov(U, Z) = 0$, but

$$\begin{aligned} Cov[u, Z] &= Cov[u, \frac{\beta_2 + \gamma_2\beta_1}{1 - \gamma_1\gamma_2} + \frac{1}{1 - \gamma_1\gamma_2}v_i + \frac{\gamma_2}{1 - \gamma_1\gamma_2}u_i] \\ &= \frac{\gamma_2}{1 - \gamma_1\gamma_2}\sigma_u^2 \end{aligned}$$

Example: Demand and Supply

Suppose $Q_d = \alpha_0 + \alpha_1 P_t + U_t$ and $Q_s = \beta_0 + \beta_1 P_t + V_t$, and we have the market clearing restriction that $Q_d = Q_s$. The, rewriting our equations in terms of $Q_d = Q_s = Q_t$, we have

$$Q_t = \alpha_0 + \alpha_1 P_t + U_t$$
$$Q_t = \beta_0 + \beta_1 P_t + V_t$$

Solving for P_t, we have

$$P_t = \frac{\beta_0 - \alpha_0}{\alpha_1 - \beta_1} + \frac{V_t - U_t}{\alpha_1 - \beta_1}$$

so we can see that P_t is a function of the error terms, so it will be correlated with both error terms. When we take covariances, we have

$$Cov(P_t, U_t) = -\frac{Var[U_t]}{\alpha_1 - \beta_1}$$
$$Cov(P_t, V_t) = \frac{Var[V_t]}{\alpha_1 - \beta_1}$$

What happens if we just regress price on quantity? The OLS estimator $\beta = \frac{Cov[Q,P]}{Var[P]}$. If we plug in the complex expressions for P and Q in terms of U and V as well as the coefficients $\alpha_0, \alpha_1, \beta_0, \beta_1$, we have

$$\tilde{\beta} = \frac{\alpha_1 Var[V] + \beta_1 Var[U]}{Var[V_t] + Var[U_t]}$$

We don't get an estimate of either the demand or the supply curve! Instead, we get a linear combination of the two parameters.

Note: This example was taken from a set of notes written by Professor Douglas Steigerwald, though the notes are no longer online.

Omitted Variable Bias

Suppose the true model is $y = a + \beta x + \delta z + u$, but we do not have data on z. For example, maybe the monetary returns to schooling are a function of your test taking ability, as measured by test scores, and how competitive you are, which we cannot measure. In this example, competitiveness influences your wages and also influences your test scores.

Example: DAG

If we omit the variable z, we will have an endogeneity. Recall, the strict exogeneity requirement is $\mathbb{E}[u|X] = 0$. The regression we run is $y = a + \tilde{\beta} x + \epsilon$, where we have $\epsilon = cz + u$. If we write out the covariance, we have:

$$\begin{aligned} Cov[\epsilon, x] &= Cov[\delta z + u, x] \\ &= \delta Cov[z, x] + Cov[u, x] \\ &= \delta Cov[z, x] \end{aligned}$$

Let's find an expression for the bias. We have the estimator $\beta = \mathbb{E}[X'X]^{-1}\mathbb{E}[X'Y]$. Plugging in for Y, we have

$$\begin{aligned} \tilde{\beta} &= \mathbb{E}[X'X]^{-1}\mathbb{E}[X'(X\beta + Z\delta + U)] \\ &= \mathbb{E}[X'X]^{-1}\mathbb{E}[X'X]\beta + \mathbb{E}[X'X]^{-1}\mathbb{E}[X'Z]\delta + \mathbb{E}[X'X]^{-1}\mathbb{E}[X'U] \\ &= \beta + \mathbb{E}[X'X]^{-1}\mathbb{E}[X'Z]\delta \end{aligned}$$

Where the last term depends on $\delta = \frac{Cov[Y,Z]}{Var[Z]}$, the true relationship between Y and Z and $Cov[X, Z]$, which describes the relationship between X and Z.

Remark: Even though we are often interpreting regressions causally, the bias here comes from correlations between Z and X and Y (a weaker claim than saying that Z is causally related to X or Y), so intuitively this omitted variable requirement is easier to violate than it might seem at first.

Example: Collider Bias in CS: Suppose attractiveness and talent are both normally distributed in the population and uncorrelated. That is $A \sim N(0,1)$, $T \sim N(0,1)$, and $Cov[A,T] = 0$.

Now, suppose I have a sample of actors, and it's the case that to be an actor $AT \geq c$ so if you're very attractive you don't need to be talented, or if you're very talented you don't need to be attractive.

Then, in the population of actors $Cov[A,T] \neq 0$. So, if we were to run $A = \beta T$ in the population, we would get $\beta = 0$, but if we were to run $A = \beta_A T$ in the population of actors, we would get a negative correlation.

Why is this interesting? Being an actor is not an omitted variable in the traditional sense, as being an actor does not cause you to be talented or attractive. Rather, being talented and attractive both contribute to being an actor. If we draw this out in a DAG, we can get a sense of a different way that not controlling for a variable can create a bias that operates similarly to omitted variable bias.

Instrumental Variables

Suppose we have some equation we want to estimate, $Y = X\beta + U$, where $\mathbb{E}[XU] \not\emptyset$. In all three cases we've discussed, the OLS estimator is biased and inconsistent. What's the solution? We use "instruments."

Example: Do Political Protests Matter? (QJE 2013) by Madestam, Shoag, Veuger, and Yanagizawa-Drott

Y = Vote Share

X = Size of Tea Party Rally

Z = Rainfall in City

Recall,

$$\begin{aligned} 0 &= \mathbb{E}[X'U] \\ 0 &= \mathbb{E}[X'(Y - X\beta)] \\ 0 &= \mathbb{E}[X'Y] - \mathbb{E}[X'X\beta] \\ \mathbb{E}[X'X]\beta &= \mathbb{E}[X'Y] \\ \beta &= \mathbb{E}[X'X]^{-1}\mathbb{E}[X'Y] \end{aligned}$$

so we can see why $\mathbb{E}[X'U] \neq 0$ will create a problem for our OLS estimator.

Suppose we have an instrument Z such that:

- $\mathbb{E}[Zu] = 0$
- $\mathbb{E}[ZZ']$ and $\mathbb{E}[ZX']$ exist
- $\mathbb{E}[ZX']$ is invertible

- Z includes all exogenous X_j
- No perfect collinearity in Z

What conditions do we need an instrument to meet:

- First Stage: $Cov(x, z) \neq 0$ - There is some relationship between x and z.
- Exclusion Restriction: $Cov(z, u) = 0$ - Z is not endogenous. Z is only related related to Y through X.

Then

$$\begin{aligned} 0 &= \mathbb{E}[ZU] \\ 0 &= \mathbb{E}[Z(Y - X'\beta)] \\ \mathbb{E}[ZX]\beta &= \mathbb{E}[ZY] \\ \beta_{IV} &= \mathbb{E}[ZX']^{-1}\mathbb{E}[ZY] \end{aligned}$$

This β is the **IV estimator**.

Sometimes we have multiple instruments to estimate the same endogenous X. Then, we might want to use both of those instruments in order to have a more efficient estimator of β. In this case we use **two-stage least squares**.

Two stage least squares works by regressing X on Z, so we use the instruments to get a prediction for X. Then, we call this prediction $\hat{X}$, and we substitute in $\hat{X}$ for X in the original regression of Y on X.

Formally, $\beta_{TSLS} = (\Pi'\mathbb{E}[ZZ']\Pi)^{-1}\Pi'\mathbb{E}[ZY]$ where $\Pi = \mathbb{E}[ZZ']^{-1}\mathbb{E}[ZX']$.

In our previous example, we could imagine using both rain and an indicator for the weather being "very hot."

Note: While IV and TSLS provide consistent estimators of β, they are biased. The reason is because $\mathbb{E}[\frac{1}{X}] \neq \frac{1}{\mathbb{E}[X]}$

- Measurement Error: A second, independent measurement is an IV
- Simultaneity: For demand and supply, use supply shocks (e.g. weather) as an instrument for demand curve, and demand curve shifts (e.g. from changing price of substitutes) as an instrument for supply curve
- Omitted Variable Bias: Political protests example from above

Question Answers

Question: $f : \mathbb{R}^2 \to \mathbb{R}^2$ with $f(x) = Ax$, $A = \begin{bmatrix} 1 & 1 \\ \frac{1}{2} & 2 \end{bmatrix}$. What is the determinant of A? What is the image of the unit square $C = [0, 1] \times [0, 1]$ under f? Is f one-to-one? Onto?

The determinant of A is given by $1 * 2 - \frac{1}{2} * 1 = 1.5$. We can sketch the image of the unit square by mapping each of it's edge points: $(0, 0)$ is mapped to $(0, 0)$, $(1, 0)$ is mapped to $(1, \frac{1}{2})$, $(0, 1)$ is mapped to $(1, 2)$, and $(1, 1)$ is mapped to $(2, 2.5)$. f is one-to-one and onto.

Question: $f : \mathbb{R}^2 \to \mathbb{R}^2$ with $f(x) = Ax$, $A = \begin{bmatrix} 2 & -1 \\ -2 & 1 \end{bmatrix}$. What is the determinant of A? What is the image of the unit square $C = [0, 1] \times [0, 1]$ under f? Is f one-to-one? Onto?

The determinant of A is given by $2 * 1 - (-2) * (-1) = 2 - 2 = 0$. We can sketch the image of the unit square by mapping each of it's edge points: $(0, 0)$ is mapped to $(0, 0)$, $(1, 0)$ is mapped to $(-2, 2)$, $(0, 1)$ is mapped to $(-1, 1)$, and $(1, 1)$ is mapped to $(1, -1)$. f is not one-to-one or onto.

Question: Can you index $\mathbb{Q}$ with the natural numbers $\mathbb{N}$? What about the real numbers $\mathbb{R}$?

Yes, you can index the rational numbers. See this diagram.

No, you cannot index the real numbers. To see why, consider Cantor's diagonal argument.

Key takaways: hopefully this exercise gives you a clear idea of what an index is, and you remember that not all sets can be indexed with the natural numbers.

Not particularly useful for our purposes, but more as trivia, this notion of being indexed by the natural numbers defines the idea of whether a set is countably infinite (the rational numbers) or uncountably infinite (the real numbers). It's relates to the broader notion of the cardinality of sets.

Question: Let $A \subset \mathbb{Q}$, $A = \{x \in \mathbb{Q} : x^2 < 2\}$. Does a have a maximum? Supremum?

No and no. While every member of the set of A is less than $\sqrt{(2)}$, the issue is that $\sqrt{(2)}$ is not a member of $\mathbb{Q}$.

Proposition: A convergent sequence is bounded.

Question: Prove this proposition.

Let x_n be a convergent sequence. Then, there exists some $N \in \mathbb{N}$ such that $|x_n - x| < 1$ for all $n > N$ by the definition of convergence. Applying the triangle inequality, we have $|x_n| \leq |x_n - x| + |x| < 1 + |x|$.

Now, consider the max of all elements before x_n as well as the number $x_n + 1$, $M = max\{|x_1|, |x_2|, ..., |x_n + 1|\}$. We have that $|x_n| \leq M$ for all $n \in \mathbb{N}$ so x_n is bounded.

Theorem (Bolzano-Weierstrass): Every bounded sequence of real numbers has a convergent subsequence.

Question: Sketch a proof for this theorem.

The idea here is that we can continue to divide the bounded interval into halves. Since we're working with a sequence, there are infinite terms, and since the sequence is bounded, those terms must lie somewhere in the bound. So, in particular, there must be infinite terms in at least one of the two halves (potentially both). Choose one element from the half that has infinite terms. This whole argument can now be reapplied starting from the element we've chosen, so we will be able to find elements in each half interval, with the halfs getting smaller and smaller (eventually converging).

Proposition: A set $F \subset \mathbb{R}$ is closed if and only if the limit of every convergent sequence in F belongs to F.

Question: Prove this proposition.

Direction 1: F Closed $\implies$ the limit of ever convergent sequence in F belongs to F.

F is closed so F^C is open. Since F^C is open, for every point $x \in F^C \exists \delta$ such that $(x-\delta, x+\delta) \in F^C$. We want to show that the limit of every convergent sequence in F belongs to F. For a contradiction, assume that x_n is a sequence in F and it converges to a limit $x \in F^C$. Since $x_n \to x$, for every $\delta > 0 \exists x_n \in (x-\delta, x+\delta)$. That is, every open ball around x contains points in F. So, x cannot be in F^C since F^C is open, but this is a contradiction since we assumed $x \in F^C$.

Direction 2: The limit of every convergent sequence in F belongs to F $\implies$ F Closed

We know the limit of every convergent sequence in F belongs to F. For a contradiction, suppose F is not closed. Then, F^C is not open. So, there exists at least one element $x \in F^C$ such that every open ball around x contains a point in F. Due to this property, we can construct a sequence by choosing balls of radius $\frac{1}{n}$ around x and selecting points in F. This creates a sequence $x_n \in F$ that converges to $x \in F^C$ which contradicts the assumption that every convergent sequence in F belongs to F.

Then, for any point in F, there is an open ball around that point contained within F. Consider a sequence in F $x_n \to x \in F$.

Question: Why do we need the requirement that $x_n \neq c$?

Because f(c) may not exist. The definition of a limit does not depend on the idea that the point itself is exists or is nearby. For example, consider the sign function which takes on -1 for negative values, 0 for 0 and 1 for positive values. If we take the sequence $\frac{1}{n}$, we know that every value in that sequence is 1, so the sequence converges to 1, but $f(0)$ is defined to be zero. For the limit convergence definition to capture the idea we want it to, we need to avoid sequences that are evaluated at $f(0)$.

Theorem (Algebraic Properties of Limits): Let $f, g : A \to \mathbb{R}$, $k \in \mathbb{R}$, and let the limits

$$\lim_{x \to c} f(x) = L \qquad \lim_{x \to c} g(x) = M$$

exist. Then,

$$\lim_{x\to c} kf(x) = kL$$
$$\lim_{x\to c} f(x) + g(x) = L + M$$
$$\lim_{x\to c} f(x)g(x) = LM$$
$$\lim_{x\to c} \frac{f(x)}{g(x)} = \frac{L}{M} \text{ if } M \neq 0$$

Question: Prove the portion of this theorem related to multiplying limits.

Since the area around a limit is bounded, we can choose some $\delta_0 > 0$ and $K > 0$ such that $|g(x)| \leq K \forall x \in A$ with $0 < |x - c| < \delta_0$.

Choose δ_1 and δ_2 such that $0 < |x - c| < \delta_1$ and $x \in A$ implies that $|f(x) - L| < \epsilon/(2K)$ $0 < |x - c| < \delta_2$ and $x \in A$ implies that $|g(x) - M| < \epsilon/(2|L| + 1)$ Let $\delta = \min(\delta_0, \delta_1, \delta_2) > 0$. Then for $0 < |x - c| < \delta$ and $x \in A$

$$\begin{aligned}
|f(x)g(x) - LM| &= |f(x)g(x) + Lg(x) - L(g(x) - LM| \\
&= |(f(x) - L)g(x) + L(g(x) - M)| \\
&\leq |f(x) - L||g(x)| + |L||g(x) - M| \\
&< \frac{\epsilon}{2K} \cdot K + |L| \cdot \frac{\epsilon}{2|L| + 1} \\
&< \epsilon
\end{aligned}$$

which proves that $\lim(fg) = \lim f \lim g$.

Proposition: If $K \subset \mathbb{R}$ is compact, then K has a maximum and minimum.

Question: Prove this proposition.

We'll use the squeeze theorem. Consider the $sup(K) = M$. Consider the two sequences M and $M - \frac{1}{n}$. Let x_n be a sequence such that $M - \frac{1}{n} < x_n < M$. By the squeeze theorem, this will converge to M. Every convergent subsequence is in K, so M is in K. Then M is the max.

Question: Prove that the function $f : [0, \infty) \to \mathbb{R}$ given by $f(x) = \sqrt(x)$ is continuous. First, note that we're trying to find an expression for δ so that given an $\epsilon > 0$, we can ensure that the function value is with epsilon if the x is within delta around some point c.

$$|f(x) - f(c)| = |\sqrt(x) - \sqrt(c)| = |\frac{x - c}{\sqrt{x} + \sqrt{c}}| \leq \frac{1}{\sqrt{c}}|x - c|$$

so given $\epsilon > 0$ we can choose $\delta - \sqrt{c}\epsilon > 0$ is the definition of continuity.

To prove this for 0, choose $\delta < \epsilon^2$. Then $|f(x) - f(0) = \sqrt(x) < \epsilon$.

Question: A polynomial function is a function of the form $P(x) = a_0 + a_1x + a_2x^2 + ...$ for coefficients $a_n \in \mathbb{R}$. Prove that every polynomial function is continuous.

This follows immediately from the fact that $f(x) = 1$ is continuous and $g(x) = x$ is continuous on $\mathbb{R}$. Now, we just combine these functions arbitrarily by multiplying by scalars and each other to generate any polynomial.

Question: Is

$$f(x) = \begin{cases} sin(\frac{1}{x}) & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{cases}$$

continuous?

Is

$$f(x) = \begin{cases} xsin(\frac{1}{x}) & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{cases}$$

continuous?

For the first one, it is continuous at points not equal to 0 and discontinuous at 0. We can see this because it's the composition of two functions that are continuous when not at 0, but $\frac{1}{x}$ is discontinuous at 0.

For the second one, again, the functions are continuous at points not equal to 0. To show that this function is continuous at 0, we need to use the squeeze theorem:

$$|f(x) - f(0)| = |xsin(\frac{1}{x}) \leq |x|$$

where the second line comes from the fact that $x \leq 1$. Applying the squeeze theorem, we know that the function will converge to 0.

Theorem: If $K \subset \mathbb{R}$ is compact and $f : K \to \mathbb{R}$ is continuous, then $f(K)$ is compact.

Question: Prove this theorem.

Let y_n be a sequence in $f(K)$. So, $y_n = f(x_n)$ for some $x_n \in K$. Since K is compact, x_n has a convergent subsequence (x_{n_k}) such that $\lim_{k\to\infty} x_{n_k} = x$ and crucially x is in K. Since f is continuous on K, we know that $f(x_{n_k}) = f(x)$. So, since $y = f(x)$, we have $y \in f(K)$ and $\lim_{k\to\infty} y_{n_k} = y$.

Intermediate Value Theorem - Special Case: Suppose that $f : [a, b] \to \mathbb{R}$ is a continuous function on a closed, bounded interval. If $f(a) < 0$and $f(b) > 0$, then there is a point $a < c < b$ such that $f(c) = 0$.

Question: Sketch a proof of this theorem.

Proof Sketch: Consider the set $E = \{x \in (a, b) : f(x) < 0\}$. Define $c = sup(E)$. Claim $f(c) = 0$. Since f is continuous, for points near c, $|f(x) - f(c)| < \frac{1}{2}|f(c)|$ if we choose δ small enough.

If $f(c) < 0$, then $c \neq b$ and $f(x) = f(c) + f(x) - f(c) < f(c) - \frac{1}{2}f(c)$ so $f(x) > \frac{1}{2}f(c) > 0$. The contradiction here is that there are points $x \in E$ with $x > c$ such that $x < 0$ but $x > c$ so c is not an upper bound.

Question: Consider $f_n(x) : (0,1) \to \mathbb{R}$, $f_n(x) = \frac{n}{nx+1}$. Does the limit exist, and if so, what is it.

Yes. $\frac{1}{x}$. We can see this by dividing both the denominator and the numerator by n.

Question: Prove that the triangle inequality is a consequence of the Cauchy-Schwartz Inequality.

$$\begin{aligned} ||x+y||^2 &= \langle x+y, x+y\rangle \\ &= ||x||^2 + \langle x,y\rangle + \langle y,x\rangle + ||y||^2 \\ &= ||x^2|| + \langle x,y\rangle + \langle y,x\rangle + ||y^2|| \\ &\leq ||x||^2 + 2||x||||y|| + ||y||^2 \end{aligned}$$

Taking square roots, we have

$$||x+y|| \leq ||x|| + ||y||$$

Question: Prove that higher moments imply the existence of lower order moments.

We want to show that $\mathbb{E}[|X^j|] < \infty$. We know that $0 < j < k$ and $\mathbb{E}[|X^k|] < \infty$. Define $f(x) = x^{k/j}$.

$$\begin{aligned} \infty > \mathbb{E}[|X^k|] &= \mathbb{E}[|f(X^j)|] \\ &\geq f(\mathbb{E}[|X^j|]) = (\mathbb{E}[|X^j|])^{\frac{k}{j}} \end{aligned}$$

by Jensen's inequality since f(x) is a convex function so we have

$$\mathbb{E}[|X^k|] \geq (\mathbb{E}[|X^j|])^{\frac{k}{j}}$$

and raising both sides to the $\frac{j}{k}$ gives

$$\begin{aligned} \mathbb{E}[|X^k|]^{\frac{j}{k}} &\geq \mathbb{E}[|X^j|])^{(\frac{k}{j})(\frac{j}{k})} \\ \mathbb{E}[|X^k|]^{\frac{j}{k}} &\geq \mathbb{E}[|X^j|]) \end{aligned}$$

where the inequality is preserved because the expectations are weakly positive and we know $\frac{k}{j}$ is positive.

Finally, notice that since $\frac{j}{k} \leq 1$, if $\mathbb{E}[|X^k|] \geq 1$, then

$$\infty > \mathbb{E}[|X^k|] \geq \mathbb{E}[|X^k|]^{\frac{j}{k}} \geq \mathbb{E}[|X^j|])$$

If $\mathbb{E}[|X^k|] < 1$, then $\infty > \mathbb{E}[|X^k|]^{\frac{j}{k}}$ since raising anything less than 1 to a power is less than 1 so it's finite.

Question: Consider two TA's Walter and Karthik, each of whom is given a fair coin. We ask Walter to continue tossing the coin until he sees $\{Heads, Tails\}$ and call the number of tosses it took for Walter to observe this patter A. We ask Karthik to toss the coin until he sees $\{Heads, Heads\}$ and call the number of tosses B. In expectation, is A or B greater (or are they the same)?

To answer this question, let's write out the Markov chain for Walter and Karthik.

Consider three states for Walter: (start), (tails, waiting for heads), (win) Consider three states for Karthi: (start), (heads, waiting for heads), (win)

The key point is that once Walter gets Tails, he cannot go back to the start. However, if Karthik gets tails, he must go back to the start. This means $A < B$, and if we were to solve this question fully, we could learn that in expectation $A = 4$ and $B = 6$.

Question: Consider a sequence of i.i.d. draws X_i from some distribution with mean μ and variance σ^2

- Provide a consistent, unbiased estimator for μ. Provide a justification for why the estimator is consistent and unbiased.
- Provide a consistent, unbiased estimator for σ^2. Provide a justification for why the this estimator is consistent and unbiased.

Recall our definitions for biasness and consistency of estimators:

Definition: An estimator is $\hat{X}_n$ **unbiased** for a statistical object μ if $\mathbb{E}[\hat{X}_n] = \mu$.

Definition: An estimator $\hat{X}_n$ is **consistent** for a statistical object μ if $\hat{X}_n \xrightarrow{p} \mu$ as $n \to \infty$.

Recall the definition of mean and variance.

$$\mu = \mathbb{E}[X_i]$$

$$\sigma^2 = \mathbb{E}[X_i^2] - \mathbb{E}[X_i]^2$$

For an estimator of the mean, consider the sample analog $\frac{1}{n}\sum_{i=1}^{n} X_i$. To show con-

sistency, we can invoke the weak law of large numbers. For unbiasedness, note

$$\begin{aligned}\mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} X_i] &= \frac{1}{n}\mathbb{E}[\sum_{i=1}^{n} X_i] \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] \\ &= \frac{1}{n}n\mathbb{E}[X_i] \\ &= \mathbb{E}[X_i] \\ &= \mu\end{aligned}$$

For an estimator of the variance, consider the sample analog

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}{n-1}$$

Let's prove this estimator is unbiased.

We'll need expressions for $\mathbb{E}[X_i^2]$ and $\mathbb{E}[\bar{X_n}^2]$ in our proof.

To find an expression for $\mathbb{E}[X_i^2]$, recall from the definition of variance that $\text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$, so

$$\begin{aligned}\mathbb{E}[X_i^2] &= \text{Var}(X) + \mathbb{E}[X]^2 \\ &= \sigma^2 + \mu^2\end{aligned}$$

To find an expression for $\mathbb{E}[\bar{X_n}^2]$, let's first compute an expression for $Var(\bar{X_n}^2)$.

$$\begin{aligned}\text{Var}(\bar{X_n}^2) &= \text{Var}(\frac{1}{n}\sum_{i=1}^{n} X_i) && (1)\\ &= \frac{1}{n^2}\text{Var}(\sum_{i=1}^{n} X_i) && (2)\\ &= \frac{1}{n^2}n\text{Var}(X_i) \text{ (by independence)} && (3)\\ &= \frac{\sigma^2}{n} && (4)\end{aligned}$$

Now, applying the same idea to the formula for variance of $\bar{X}_n$, we have

$$\begin{aligned}\mathbb{E}[\bar{X_n}^2] &= \text{Var}(\bar{X}_n) + \mathbb{E}[\bar{X}_n]^2 \\ \mathbb{E}[\bar{X_n}^2] &= \frac{\sigma^2}{n} + \mu^2\end{aligned}$$

With these two expressions, we can now work out the expectation of the numerator:

$$\begin{aligned}
\mathbb{E}[\sum_{i=1}^{n}(X_i - \bar{X_n})^2] &= \mathbb{E}[\sum_{i=1}^{n} X_i^2 + \bar{X_n}^2 - 2X_i\bar{X_n}] \\
&= \mathbb{E}[\sum_{i=1}^{n} X_i^2 + \sum_{i=1}^{n} \bar{X_n}^2 - \sum_{i=1}^{n} 2X_i\bar{X_n}] \\
&= \mathbb{E}[\sum_{i=1}^{n} X_i^2 + n\bar{X_n}^2 - 2\bar{X_n}\sum_{i=1}^{n} X_i] \\
&= \mathbb{E}[\sum_{i=1}^{n} X_i^2 + n\bar{X_n}^2 - 2\bar{X_n}n\bar{X_n}] \text{ since } \bar{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i \\
&= \mathbb{E}[\sum_{i=1}^{n} X_i^2 + n\bar{X_n}^2 - 2n\bar{X_n}^2] \\
&= \mathbb{E}[\sum_{i=1}^{n} X_i^2 - n\bar{X_n}^2] \\
&= \sum_{i=1}^{n} \mathbb{E}[X_i^2] - n\mathbb{E}[\bar{X_n}^2] \\
&= \sum_{i=1}^{n} (\sigma^2 + \mu^2) - n(\frac{\sigma^2}{n} + \mu^2) \\
&= n\sigma^2 + n\mu^2 - \sigma^2 - n\mu^2 \\
&= (n-1)\sigma^2
\end{aligned}$$

Finally, we can consider the expectation of the whole estimator

$$\begin{aligned}
\mathbb{E}[\frac{\sum_{i=1}^{n}(X_i - \bar{X_n})^2}{n-1}] &= \frac{1}{n-1}\mathbb{E}[\sum_{i=1}^{n}(X_i - \bar{X_n})^2] \\
&= \frac{1}{n-1}(n-1)\sigma^2 \\
&= \sigma^2
\end{aligned}$$