MS Math Boot Camp Lecture Notes

Instructor: Naz Koont (nkoont23@gsb.columbia.edu)

Contents

0 Goals

I hope that this course will be useful to you in one of two ways:

1. If you have taken math courses in these topics previously, these lectures can help remind you of the key takeaways from these courses as you prepare for your masters degree curriculum.

2. If you have not taken these math courses, you can treat these lectures as a survey overview. I will make you aware of different tools that will be useful for your upcoming courses, and present a "map" of mathematics while emphasizing the connections to finance and economics and the motivations for learning a given mathematical concept for our applied purposes.

At the end of each section I will also provide resources for deeper learning and practice for topics in which you feel you need additional review.

1 Fundamentals

The tools of mathematics allow economists to make simple assumptions and represent complex phenomena as stylized mathematical relationships in order to make useful deductions about the world. For example, the **Black-Scholes Model** derives the prices of options contracts under certain assumptions, including that prices follow a lognormal distribution. The **Solow-Swan Model** explains long-run economic growth using a nonlinear system of a single ordinary differential equation to model the evolution of the per-person stock of capital. The **Capital Asset Pricing Model (CAPM)** determines the appropriate required rate of return on a given asset under assumptions about how the market functions and how risk can be measured.

In order to understand, use, and create new and better economic models, we must be fluent in the mathematics used to formulate them. To begin our survey of mathematics that are used in economic and financial analysis, let us review:

1. Two essential objects in the language of mathematics: **sets** and **functions**.
2. How the language of mathematics allows logical deductions from a set of assumptions via proofs.

1.1 Sets

A set is a Many that allows itself to be thought of as a One – Georg Cantor

Definition 1.1 (Set)**.** *A set B is any collection of items (elements) x thought of as a whole.*

In economics we will usually think of sets with elements in $\mathbb{R}$. The principal concept of set theory is that of *belonging*. Sets can be defined by enumerating elements $x \in B$, or by stating a property.

$$B = \{2, 4, 6, 8, 10\} = \{x | x \text{ is an even number between 1 and 11}\}$$

Then e.g. $2 \in B$ but $3 \notin B$. Note that sets can have infinitely many elements. For example, the set of real numbers $\mathbb{R}$ is an uncountably infinite set.

Definition 1.2 (Subset)**.** *If all the elements of a set A are also elements of a set B then, A is a subset of B, $A \subseteq B$. Note that if $C \subseteq B$ and $B \subseteq C$, $\implies C = B$.*

$$A = \{2, 4\} \implies A \subset B$$

1.1.1 Examples of Sets

Let us further define a few important sets:

Definition 1.3 (Null Set). *The empty set or the null set is the set with no elements* $\emptyset$. $\emptyset \subset S, \forall S$

Definition 1.4 (Natural numbers). *The natural numbers* $\mathbb{N} = \{1, 2, 3, ...\}$

Definition 1.5 (Integers). *The integers* $\mathbb{Z} = \{..., -2, -1, 0, 1, 2, ...\}$

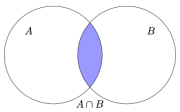
Definition 1.6 (Rationals). *The rational numbers* $\mathbb{Q} = \{\frac{p}{q} \quad | \quad \forall p, q \in \mathbb{Z}, q \neq 0\}$

Definition 1.7 (Real numbers). *The real numbers* $\mathbb{R}$ *are the set of numbers that can be represented as a (possibly infinite) decimal expansion, e.g.* $\pi = 3.14159...$

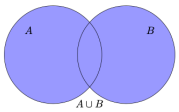
Often the set of real numbers $\mathbb{R}$ is visualized as a straight line, *the Real Line*, on which numbers are *points*. Note that $\mathbb{N}, \mathbb{Z}, \mathbb{Q}$ are all strict subsets of $\mathbb{R}$.

1.1.2 Set Operations

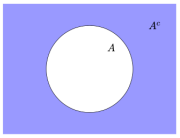
Definition 1.8 (Intersection). *The intersection* C *of two sets* A *and* B *is the set of elements that are in* ***both*** A *and* B: $C = A \cap B = \{x | x \in A \wedge x \in B\}$



Definition 1.9 (Union). *The union* D *of two sets* A *and* B *is the set of elements that are in* ***either*** A *or* B $D = A \cap B = \{x | x \in A \vee x \in B\}$



Definition 1.10 (Complement). *The complement* A' *of a set* A *is the set of elements (in our "universal" set* $U = \mathbb{R}$ *or* $\mathbb{N}$*) that are not in* A: $A' = \{x | \notin A\}$



Definition 1.11 (Set Difference*). *The set difference of sets* A *and* B*, denoted* $A \setminus B$ *is the set of elements* x *in* A *that are not also in* B: $A \setminus B = \{x | x \in A \wedge x \notin B\}$

Definition 1.12 (Partition*). *A partition* $\{S_i\}_{i \in n}$ *of a set* $\mathbb{R}$*, is a collection of disjoint subsets of* R *(intersection of any* $S_i \cap S_j = \emptyset$*) such that their union* $\bigcup_{i \in n} S_i = \mathbb{R}$

Definition 1.13 (Power Set*). *A power set* $P(S)$ *of a set* S *is the set of* ***all possible subsets*** *of* S: $P(S) = \{A | A \subseteq S\}$

1.1.3 Properties of Point Sets in $\mathbb{R}^n$

Definition 1.14 (Closed Interval). $[a, b] = \{x \in \mathbb{R} | a \leq x \leq b\}$ *contains boundary points*

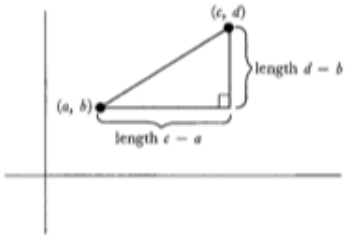
Definition 1.15 (Half-Open Interval). $(a, b] = \{x \in \mathbb{R} | a < x \leq b\}$

Definition 1.16 (Open Interval). $(a, b) = \{x \in \mathbb{R} | a < x < b\}$ *excludes boundary points*

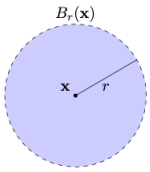
Definition 1.17 (Cartesian Plane $\mathbb{R}^2$). $\mathbb{R}^2$ *is the set of* ***ordered*** *pairs* (x, y) *formed by the cartesian product of* $\mathbb{R} \otimes \mathbb{R}$

Similarly, we can think about $\mathbb{R}^3$ or $\mathbb{R}^n$ for $n \in \mathbb{N}$. We will use geometric pictures for the purpose of aiding intuition, and define distance as follows so that the Pythagorean theorem is built into our geometry:

Definition 1.18 (Euclidean Distance). *The Euclidean distance between two elements of* $\mathbb{R}^n$ *is given by the length of the line segment between them:* $d(a, b) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2}$

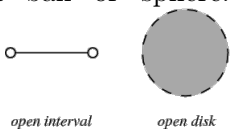


Definition 1.19 (ϵ-Neighborhood). *An* ϵ*-neighborhood of a point* $x_0 \in \mathbb{R}^n$ *is given by the set* $N_\epsilon(x_0) = \{x \in \mathbb{R}^n | d(x_0, x) < \epsilon\}$

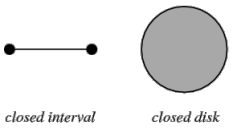


Of course, when in dimension $n = 1$, this is just an open interval. Similarly, for $n = 2$, this is disc, or circle. In higher dimensions, it can be referred to as a ball or sphere.

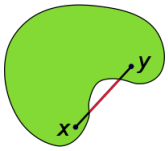
Definition 1.20 (Open Set). *A set* $X \subset \mathbb{R}^n$ *is open if for every* $x \in X$, *there exists an* $\epsilon > 0$ *such that* $N_\epsilon(x) \subset X$



Definition 1.21 (Closed Set). *A set* $X \subset \mathbb{R}^n$ *is* ***closed*** *if its complement* X' *is an open set.*



Definition 1.22 (Convex*). *A set* $X \subset \mathbb{R}^n$ *is convex if for every pair of points* $x, x' \in X$ *and any* $\lambda \in [0, 1]$, *the point* $\bar{x} = \lambda x + (1 - \lambda)x'$ *is in* X *(convex combination)*



Definition 1.23 (Bounded). *A set* $X \subset \mathbb{R}^n$ *is bounded if for every* $x_0 \in X$, *there exists* $\epsilon < \infty$ *such that* $X \subset N_\epsilon(x_0)$

In particular, a set A of real numbers is **bounded above** if there is a number x such that $x \geq a$ for every $a \in A$. Such a number x is called an **upper bound** for A. The set of real numbers $\mathbb{R}$ and the natural numbers $\mathbb{N}$ are two examples of sets that are **not** bounded above.

An example of a set that is bounded above is $A = \{x | 0 \leq x < 1\}$. Further, 1 is the **least upper bound** or **supremum** of A, i.e. the smallest such bound. Analogously, we can think of sets **bounded below** and **greatest lower bounds** or **infimums**.

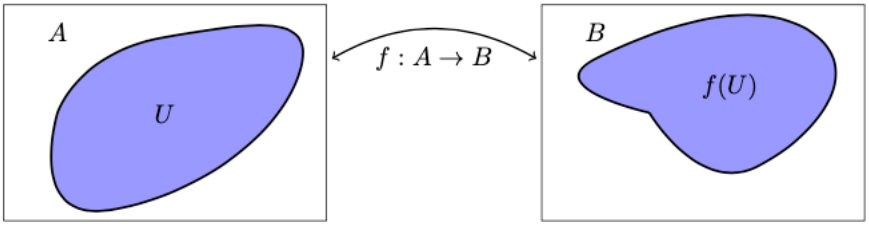
Theorem 1.1 (Least Upper Bound Property or The Completeness Axiom)**.** *If A is a set of real numbers, $A \neq \emptyset$, and A is bounded above, then A has a least upper bound.*

1.2 Functions

> Undoubtedly the most important concept in all of mathematics is that of a function – in almost every branch of modern mathematics functions turn out to be the central objects of investigation.
> – Michael Spivak, Calculus

Functions are an object of great generality[1] but for applications to finance and economics, we can restrict our attention to a small subset of functions.

Definition 1.24 (Function)**.** *Given two sets A, B a **function** $f : A \mapsto B$ is a map which assigns to every element in A a **unique** element in B.*



If $a \in A$, we usually denote the corresponding element of B by $f(a)$. A is the **domain** and B is the **codomain** of function f.

Definition 1.25 (Domain)**.** *The **domain** of function f is the set $\{a \in A | \exists! b \in B \text{ s.t. } b = f(a)\}$.*

Definition 1.26 (Range)**.** *The **range** of function f is the set $\{b \in B | \exists a \in A \text{ s.t. } b = f(a)\}$.*

Definition 1.27 (Injective Function)**.** *A function f is **injective** or **one-to-one** if $f(x) = f(y) \implies x = y$*

Definition 1.28 (Surjective Function)**.** *A function f is **surjective** if the range of f is B.*

We can define a **composite** mapping of two functions $f : A \mapsto B$ and $g : B \mapsto C$ as $g \circ f : A \mapsto C$

1.2.1 Examples of Functions

Let us further define a few important types of functions:

Definition 1.29 (Linear Function)**.** *$f : \mathbb{R} \mapsto \mathbb{R}$: $f(x) = ax + b = y$ for $a, b \in \mathbb{R}$ where a is the slope and b is the intercept*

[1] In fact, a function can be defined as a special kind of set: a collection of pairs of elements with the following property. If (a, b) and (a, c) are both members of the set, then it must be that $b = c$.

Definition 1.30 (Quadratic Function)**.** $f : \mathbb{R} \mapsto \mathbb{R}$*:* $f(x) = ax^2 + bx + c = y$ *for* $a, b, c \in \mathbb{R}$

The simplest quadratic function is the **parabola**, $f(x) = x^2$.

Definition 1.31 (Polynomial Function)**.** $f : \mathbb{R} \mapsto \mathbb{R}$*:* $f(x) = a_n x^n + a_{n-1}x^{n-1} + ... + a_0$ *for* $a_i \in \mathbb{R}$

The simplest polynomial function is the **power function**, $f(x) = x^n$ for $n \in \mathbb{N}$.

Definition 1.32 (Graph)**.** *We can draw the graph of a function* $(x, f(x))$ *on the Cartesian Plane* $\mathbb{R}^2$

In contrast, some of the simplest and most important subsets of the Cartesian Plane are not the graphs of functions. These include the circle, or ϵ-neighborhood defined above. Similarly, ellipses and hyperbolas are not the graphs of a function. *Why?*

1.3 Proofs

Pure mathematics is, in its way, the poetry of logical ideas. –Albert Einstein

A proof is a mathematical argument intended to convince us that a result is correct. A proof of a theorem is thus a series of logical deductions using the assumptions of the theorem, the definitions of the terms involved, and previous results that have been proven. In mathematics, a statement (or proposed theorem, etc.) is an assertion that is either **true** or **false**.

Many mathematical theorems can be expressed symbolically in the form $P \implies Q$, i.e. P *implies* Q. The statement P is the *assumption* of the theorem, and statement Q is the *conclusion*. Equivalence theorems can be expressed as $P \iff Q$, which means the same thing as $P \implies Q$ and $Q \implies P$.

Similarly, economic analysis is mainly concerned with deductive statements that require a logical proof, such as *if the money supply increases, then the price level will rise*. Or, even stronger statements such as *the price level rises* ***if and only if*** *the money supply increases*.

Often to prove a statement, we will use one of the standard proof techniques:

1. Direct Proof: the conclusion is established by logically combining the axioms, definitions, and earlier theorems.

   Theorem 1.2 (Example of Direct Proof)**.** *The sum of two even integers is always even, where even is defined to be an integer that has a factor of 2.*

   Proof. Consider two even integers x, y. Then we know that they can be written as $x = 2a$ and $y = 2b$ respectively, for $a, b \in \mathbb{Z}$. Then $x + y = 2(a + b)$ has 2 as a factor, and thus by definition is even. □

2. Contrapositive Proof Method: $P \implies Q$ is proved by showing the logically equivalent contrapositive statement, $\neg Q \implies \neg P$.

   Theorem 1.3 (Example of Contrapositive Proof)**.** *Given an integer* x*, if* x^2 *is even, then* x *is even:*

Proof. Suppose x is not even. Then x is odd. The product of two odd numbers is odd, thus $x^2 = x * x$ is odd. Thus x^2 is not even. We have shown the contrapositive, which implies that if x^2 is even it must be that x is even. □

3. Proof by Contradiction, or *reductio ad absurdum*: it is shown that if some statement is assumed true, a logical contradiction occurs, and hence the statement must be false.

   Theorem 1.4 (Example of Proof by Contradiction)**.** $\sqrt{2}$ *is irrational*

   Proof. Suppose $\sqrt{2} \in \mathbb{Q}$. Then it could be written in lowest terms as $\sqrt{2} = \frac{a}{b}$ where $a, b \in \mathbb{Z} \setminus 0$ with no common factor. Thus $b\sqrt{2} = a \implies 2b^2 = a^2$.
   Since this equality shows that a^2 is even, it must be that a is even by the result in the example above of the contrapositive proof, so that 2 is a factor of a. Thus we can write $a = 2c$ for $c \in \mathbb{Z}$. Substituting this in, we get $2b^2 = (2c)^2 \implies b^2 = 2c^2$ and so b must also be even, so that 2 is a factor of b.
   This **contradicts** that a, b had no common factor. Thus it must be that $\sqrt{2} \notin \mathbb{Q}$, $\sqrt{2}$ is irrational. □

4. Proof by Induction: In proof by mathematical induction, a base case is provided, and an induction rule is proved that establishes that any arbitrary case implies the next case. Since the induction rule can be applied repeatedly, it follows that all cases are provable.

   Theorem 1.5 (Example of Proof by Induction)**.** *All positive integers in the form of* $2n - 1$ *are odd. Let* $P(n)$ *denote the statement "*$2n - 1$ *is odd".*

   Proof.

   (a) Base Case: For $n = 1$, $2n - 1 = 2(1) - 1 = 1$ and is odd, since it does not have a factor of 2. Thus $P(1)$ is true.

   (b) Induction: For any n, if $2n - 1$ is odd such that $P(n)$ is true, then $(2n - 1) + 2$ must also be odd, because adding 2 to an odd number results in an odd number. But $(2n-1)+2 = 2n+1 = 2(n+1) - 1$ which is $P(n+1)$. Thus, $P(n) \implies P(n+1)$.

   Thus, $2n - 1$ is odd, for all positive integers n. □

1.4 Resources

At the end of each section I will list some resources for additional reading and for practice problems. In this first section I will also list some general math tools that are available online.

1.4.1 General Math Resources

1. For solving equations: https://www.wolframalpha.com/

2. A database of questions and answers related to mathematics: https://math.stackexchange.com/

3. For basic mathematical concepts: https://www.khanacademy.org/

4. For typing up problem sets: LaTeX, https://www.overleaf.com/learn/latex/Tutorials and see the .tex version of these lecture notes for an example.

5. Wikipedia entries of a given mathematical concept are generally relatively high quality.

1.4.2 Textbooks

1. *Mathematics for Economics by Hoy et al.* is a great resource for mathematics geared towards economic applications, with many applied practice problems. Fundamentals are covered in chapters 2.1—2.5.

1.4.3 Web Resources

1. Notes on Proofs and Logic: MAT 102 http://home.tykenho.com/index.html?notes

2. Set theory: https://www.math.uh.edu/~dlabate/settheory_Ashlock.pdf

1.4.4 Practice Problems with Solutions

1. Proofs: https://www.math.cmu.edu/~mradclif/teaching/127S19/Notes/Logic%20and%20Proof.pdf

---

2 Single Variable Calculus

Much of financial and economic analysis is concerned with marginal analysis, i.e. how a change in the level of one variable x determines a change in the level of another variable y. Calculus provides tools to analyze these changes.

2.1 Limits

> The concept of a limit is surely the most important, and probably the most difficult one in all of Calculus.
> – Michael Spivak, Calculus

Intuitively, the function f approaches the limit l near a, if we can make $f(x)$ as close as we like to l by requiring that x be sufficiently close, but unequal to, a.

Definition 2.1 (Limit)**.** *The function f approaches the limit l near a if for every $\epsilon > 0$, there is some $\delta > 0$ such that, for all x, $0 < |x - a| < \delta$, then $|f(x) - l| < \epsilon$*

This definition gives rise to the *epsilon-delta method* of determining whether a limit exists. It is useful to think about what is true when the function f does **not** approach a limit l near a. This means that there is **some** $\epsilon > 0$, such that **for every** $\delta > 0$, there is **some** x which satisfies $0 < |x - a| < \delta$ but not $|f(x) - l| < \epsilon$. An example of this flavor of reasoning is found in the proof below that limits are unique.

Theorem 2.1 (Limits are unique)**.** *A function cannot approach two different limits near a. If f approaches l near a, and f approaches m near a, then $l = m$.*

Proof. Let us assume that $l \neq m$. We are given that f approaches l near a, so by the definition of a limit, we have that for any $\epsilon > 0$, there is some number $\delta_1 > 0$ such that, for all x,

$$\text{if } 0 < |x - a| < \delta_1, \text{ then } |f(x) - l| < \epsilon$$

Similarly, we know that f approaches m near a, so again have that there is some $\delta_2 > 0$ such that for all x,

$$\text{if } 0 < |x - a| < \delta_2, \text{ then } |f(x) - m| < \epsilon$$

Taking the minimum of δ_1 and δ_2, we then have that there exists δ such that

$$\text{if } 0 < |x - a| < \delta, \text{ then } |f(x) - l| < \epsilon \text{ and } |f(x) - m| < \epsilon$$

This holds for any ϵ. Since we have assumed that $l \neq m$, we know that $|l - m| > 0$ and so can choose $\epsilon = \dfrac{|l - m|}{2}$. This gives that

$$\text{if } 0 < |x - a| < \delta, \text{ then } |f(x) - l| < \frac{|l - m|}{2} \text{ and } |f(x) - m| < \frac{|l - m|}{2}$$

So for $0 < |x - a| < \delta$, we have

$$|l - m| = |l - f(x) + f(x) - m| \leq |l - f(x)| + |f(x) - m| < |l - m|$$

Contradiction. So, $l = m$. □

Theorem 2.2 (Operations with Limits)**.** *If $\lim_{x \to a} f(x) = l$ and $\lim_{x \to a} g(x) = m$ then:*

1. $\lim_{x \to a}(f + g)(x) = l + m$
2. $\lim_{x \to a}(f \cdot g)(x) = l \cdot m$
3. $\lim_{x \to a} \dfrac{1}{g}(x) = \dfrac{1}{m}$ *if* $m \neq 0$.

2.2 Continuous Functions

If f is an arbitrary function, it is not necessarily true that $\lim_{x\to a} f(x) = f(a)$. There are many ways in which this can fail to be true. Functions for which this condition holds are said to be continuous. Intuitively, a function f is continuous if the graph contains no breaks, jumps, or wild oscillations.

Definition 2.2 (Continuous Function)**.** *The function f is continuous at a if* $\lim_{x\to a} f(x) = f(a)$

Theorem 2.3 (Operations with Continuous Functions)**.** *If f and g are continuous at a then:*

1. *$f + g$ is continuous at a*
2. *$f \cdot g$ is continuous at a*
3. *If $g(a) \neq 0$, then $1/g$ is continuous at a.*

Theorem 2.4 (Compositions Continuous Functions)**.** *If g is continuous at a, and f is continuous at $g(a)$, then $f \circ g$ is continuous at a.*

We have defined continuity of functions at a single point a. However, the concept of continuity is particularly useful to us if we focus our attention on functions that are continuous at all points within some interval. These functions are usually regarded to be especially well behaved.

Theorem 2.5 (Intermediate Value Theorem)**.** *If f is continuous on $[a, b]$ and $f(a) < c < f(b)$, then there is some number x in $[a, b]$ such that $f(x) = c$.*

This theorem is very useful in economics e.g. to find an equilibrium market-clearing price.

Theorem 2.6 (Continuous Functions on a Closed Interval are Bounded)**.** *If f is continuous on $[a, b]$, then f is bounded above on $[a, b]$.*

Theorem 2.7 (Extreme Value Theorem)**.** *If f is continuous on $[a, b]$, then there is a number y in $[a, b]$ such that $f(y) \geq f(x)$ for all x in $[a, b]$. I.e., f attains a **maximum** (and minimum) on the closed interval.*

Theorem 2.8 (Archimedean Property of $\mathbb{N}$)**.** *$\mathbb{N}$ is not bounded above.*

Proof. Suppose that $\mathbb{N}$ were bounded above. Since $\mathbb{N} \neq \emptyset$, we know that there exists a least upper bound α for $\mathbb{N}$. Then, $\alpha \geq n$ for all $n \in \mathbb{N}$. And also, $\alpha \geq n + 1$ for all $n \in \mathbb{N}$.

This is because if $n \in \mathbb{N}$ then $n + 1 \in \mathbb{N}$. However, rearranging, we find that $\alpha - 1 \geq n$ for all $n \in \mathbb{N}$. This means that $\alpha - 1$ must also be an upper bound for $\mathbb{N}$, **contradicting** that α was the last upper bound. Thus $\mathbb{N}$ is not bounded above. □

Theorem 2.9 (Formulating a Small ϵ)**.** *For any $\epsilon > 0$, there is a natural number $n \in \mathbb{N}$ with* $\dfrac{1}{n} < \epsilon$.

Proof. Suppose not. Then $\dfrac{1}{n} \geq \epsilon$ for all $n \in \mathbb{N}$. Thus $n \leq 1/\epsilon$ for all $n \in \mathbb{N}$. But this would mean that ϵ is a least upper bound for $\mathbb{N}$, which **contradicts** the Archimedean property of $\mathbb{N}$ that it is not bounded above. □

2.3 Derivatives

The most useful results for us about functions will be obtained once we restrict our attention even further than simply looking at continuous functions. Even continuous functions can have peculiarities such as sharp edges at a point, e.g. $f(x) = |x|$ at 0. At such points, it may be that a unique line that is tangent to the graph of the function cannot be drawn.

Definition 2.3 (Differentiable Function)**.** *The function f is differentiable at a if $\lim_{h\to 0} \frac{f(a+h)-f(a)}{h}$ exists. In this case the limit is denoted by $f'(a)$ and is called the derivative of f at a. We further say that f is differentiable if f is differentiable at a for every a in the domain of f.*

Accordingly, for any function f, we denote by f' the function whose domain is the set of all numbers a such that f is differentiable at a, and whose value at such a number a is precisely $\lim_{h\to 0} \frac{f(a+h)-f(a)}{h}$. The function f' is called the **derivative** of f.

Theorem 2.10 (Differentiable if Continuous)**.** *If f is differentiable at a, then f is continuous at a. (The converse is not true, e.g. continuous and nowhere differentiable function).*

Thus, differentiability places more restrictions on functions that continuity. Yet, we can actually use the concept of a derivative to formulate even more restrictive conditions on functions. For any function f, we obtain, by taking the derivative, a new function f' whose domain may be smaller than that of f. The notion of differentiability may be applied to the function f', yielding another function f'', called the **second derivative** of f. Similarly, we can define the **nth derivative** for any $n \in \mathbb{N}$, collectively referred to as higher order derivatives for f.

Theorem 2.11 (Derivative of a Constant)**.** *If $f(x) = c$ for $c \in \mathbb{R}$ then $f'(x) = 0$*

Proof. $f'(a) = \lim_{h\to 0} \frac{f(a+h)-f(a)}{h} = \lim_{h\to 0} \frac{c-c}{h} = 0$ □

Theorem 2.12 (Derivative of a Linear Function)**.** *If $f(x) = mx + b$ for $m, b \in \mathbb{R}$, then $f'(x) = m$*

Proof. $f'(a) = \lim_{h\to 0} \frac{f(a+h)-f(a)}{h} = \lim_{h\to 0} \frac{m(a+h)+b-(ma+b)}{h} = \lim_{h\to 0} \frac{mh}{h} = m$ □

Theorem 2.13 (Derivative of a Power Function)**.** *If $f(x) = x^n$ then $f'(x) = nx^{n-1}$*

Proof. Proof for $n = 2$:
$f'(a) = \lim_{h\to 0} \frac{f(a+h)-f(a)}{h} = \lim_{h\to 0} \frac{(a+h)^2-a^2}{h} = \lim_{h\to 0} \frac{a^2+2ah+h^2-a^2}{h} = \lim_{h\to 0} 2a + h = 2a$ □

Theorem 2.14 (Derivative of an Exponential Function*)**.** *If $f(x) = e^x$ then $f'(x) = e^x$*

Theorem 2.15 (Derivative of a Log Function*)**.** *If $f(x) = log(x)$ then $f'(x) = 1/x$*

Theorem 2.16 (Derivative of constant times a function*)**.** *If $f(x) = a * g(x)$ then $f'(x) = a * g'(x)$*

Theorem 2.17 (Sum Rule*)**.** *If $f(x) = g(x) + h(x)$ then $f'(x) = g'(x) + h'(x)$*

Theorem 2.18 (Product Rule). *If* $f(x) = g(x)h(x)$ *then* $f'(x) = g'(x)h(x) + g(x)h'(x)$

Theorem 2.19 (Quotient Rule). *If* $f(x) = \dfrac{g(x)}{h(x)}$ *then* $f'(x) = \dfrac{g'(x)h(x) - g(x)h'(x)}{h(x)^2}$ *for* $h(x) \neq 0$

Theorem 2.20 (Chain Rule). $(f \circ g)'(a) = f'(g(a)) * g'(a)$

Theorem 2.21 (L'Hopital's Rule). $\lim_{x\to a} \dfrac{f(x)}{g(x)} = \lim_{x\to a} \dfrac{f'(x)}{g'(x)}$

L'hopital's rule is particularly useful in obtaining the limit of expressions in which both the numerator and denominator of a function diverge to $+\infty$ or 0. For example,

$$\lim_{x\to\infty} \frac{\ln x}{\sqrt{x}} = \frac{\infty}{\infty} = ?$$ is difficult to evaluate directly. Applying L'hopital's,

$$\lim_{x\to\infty} \frac{(\ln x)'}{(\sqrt{x})'} = \lim_{x\to\infty} \frac{1/x}{1/2\sqrt{x}} = \lim_{x\to\infty} \frac{2}{\sqrt{x}} = 0$$

Theorem 2.22 (Mean Value Theorem). *If* f *is continuous on* $[a, b]$ *and differentiable on* (a, b)*, then there is a number* x *in* (a, b) *such that* $f'(x) = \dfrac{f(b) - f(a)}{b - a}$

Corollary 2.22.1. *If* f *is defined on an interval and* $f'(x) = 0$ *for all* x *in the interval, then* f *is constant on the interval.*

Corollary 2.22.2. *If* $f'(x) > 0$ *for all* x *in an interval, then* f *is increasing on the interval. If* $f'(x) < 0$ *for all* x *in an interval, then* f *is decreasing on the interval.*

2.4 Inverse Functions*

Recall that a function $f : A \to B$ is **injective** or **one-to-one** if $f(a) \neq f(b)$ whenever $a \neq b$. A simple example of an injective function is the Identity function $f(x) = x$. The key defining feature of a function is that it maps every element in its domain A to a unique element in its codomain B.

If we have an injective function, we can imagine reversing the mapping without violating this key principle. In other words, we can think of a function $g : B \to A$ defined by $g(b) = a$ where $f(a) = b$.

Definition 2.4 (Inverse Function). *For any function* f*, the inverse of* f*, denoted by* f^{-1}*, is the set of all pairs* (a, b) *for which the pair* (b, a) *is in* f*.*

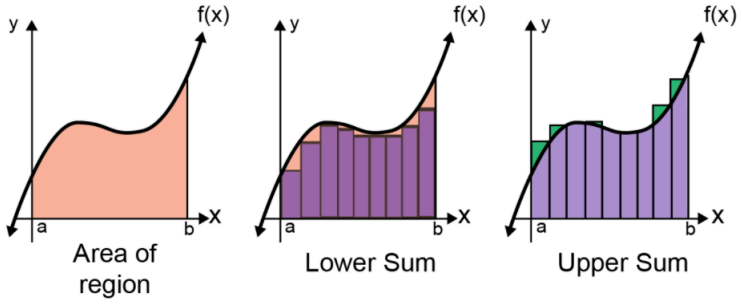
Theorem 2.23 (Inverse is a Function). f^{-1} *is a function if and only if* f *is injective.*

Theorem 2.24 (Inverse Function Theorem). *Let* f *be a continuous injective function defined on an interval, and suppose that* f *is differentiable at* $f^{-1}(b)$*, with derivative* $f'(f^{-1}(b)) \neq 0$*. Then* f^{-1} *is differentiable at* b*, and the derivative of the inverse function at* b *is the reciprocal of the derivative of* f *at* a*:*

$$(f^{-1})'(b) = \frac{1}{f'(f^{-1}(b))}$$

2.5 Integrals

Intuitively, an integral of a function gives the area beneath its graph. This concept is very useful in economics, as it allows us to model the relationship between stocks and flows (e.g. investment in a given year and the accumulation of capital stock) and marginal and total concepts (e.g. the marginal cost of producing an additional unit vs the total cost until that point).



To think about approximating the area under an arbitrary curve, we may find it easier to think about adding up the area of rectangles — a more manageable calculation — that fit under (lower sum) or just over (upper sum) the graph of the function. As we make these rectangles smaller and smaller (or "skinnier"), our approximate calculation of the area under the curve approaches the true value. This is the intuition behind the **Reimann Integral**. Loosely speaking, this integral is the limit of this sum as the rectangles get smaller and smaller (i.e. as the *partitions* get finer).

Definition 2.5 (Integrable Function)**.** *A function f which is bounded on $[a,b]$ is (Reimann) integrable on $[a,b]$ if the limit of its lower and upper sum are equal. In this case, this common number is called the* ***integral*** *of f on $[a,b]$ and is denoted by*

$$\int_a^b f(x)dx = \int_a^b f$$

Where a and b are called the lower and upper limits of integration, $f(x)$ is called the integrand, and dx is referred to as "integrating with respect to x".

Theorem 2.25 (Integrability over subsets)**.** *Let $a < c < b$. If f is integrable on $[a,b]$, then f is integrable on $[a,c]$ and on $[c,b]$. Conversely, if f is integrable on $[a,c]$ and on $[c,b]$ then f is integrable on $[a,b]$. Further,*

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

Theorem 2.26 (Integrability of Sums)**.** *If f and g are integrable on $[a,b]$, then $f+g$ is integrable on $[a,b]$ and:*

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

Theorem 2.27 (Integrability and Scalar Multiplication)**.** *If f is integrable on $[a, b]$, then for any number $c \in \mathbb{R}$, the function cf is integrable on $[a, b]$ and*

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

Theorem 2.28 (Bounding the Integral)**.** *Suppose f is integrable on $[a, b]$ and that $m \leq f(x) \leq M$ for all x in $[a, b]$. Then,*

$$m(b - a) \leq \int_a^b f \leq M(b - a)$$

Note that a function need not be continuous in order to be integrable. However, it turns out that the function defined as the integral of a given integrable function is indeed continuous:

Theorem 2.29 (Integral is Continuous)**.** *If f is integrable on $[a, b]$ and F is defined on $[a, b]$ by $F(x) = \int_a^x f$, then F is continuous on $[a, b]$.*

What if the original function f is continuous? This gives rise to a famous theorem in Calculus:

Theorem 2.30 (Fundamental Theorem of Calculus)**.** *Let f be integrable on $[a, b]$ and define F on $[a, b]$ by*

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

If f is continuous at c on $[a, b]$, then F is differentiable at c, and $F'(c) = f(c)$.

Corollary 2.30.1. *If f is continuous on $[a, b]$ and $f = g'$ for some function g, then $\int_a^b f = g(b) - g(a)$. (In fact, this is true even if f is only integrable on $[a, b]$).*

2.5.1 Techniques of Integration

Nature laughs at the difficulties of integration. – Pierre-Simon Laplace

In general, a function F satisfying $F' = f$ is called a **primitive** of f. Of course, a continuous function f always has a primitive,

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

However, we may be interested in writing out a primitive as an elementary function (i.e. one that can be obtained by addition, multiplication, division, and composition from the rational and trigonometric functions, and their inverses). Elementary primitives for arbitrary functions cannot usually be found.

Note that in the following discussion, we are interested in the primitive as a **function**, and thus will write integrals without limits to represent this, otherwise known as an indefinite integral. In contrast, so far we have been looking at definite integrals, which represents a number when the upper and lower limits are constants.

Nevertheless, it is useful to discuss methods for finding elementary primitives of given elementary functions. In general, the techniques to evaluate a derivative consists of two parts:

1. Have a good reservoir of known integrals.
2. Be fluent with integration techniques (linearity of integration, substitution, integration by parts).

$$\int cf(x)\,dx = c\int f(x)\,dx \qquad \int [f(x)+g(x)]\,dx = \int f(x)\,dx + \int g(x)\,dx$$

$$\int k\,dx = kx + C$$

$$\int x^n\,dx = \frac{x^{n+1}}{n+1} + C \quad (n \neq -1) \qquad \int \frac{1}{x}\,dx = \ln|x| + C$$

$$\int e^x\,dx = e^x + C \qquad \int b^x\,dx = \frac{b^x}{\ln b} + C$$

$$\int \sin x\,dx = -\cos x + C \qquad \int \cos x\,dx = \sin x + C$$

$$\int \sec^2 x\,dx = \tan x + C \qquad \int \csc^2 x\,dx = -\cot x + C$$

$$\int \sec x \tan x\,dx = \sec x + C \qquad \int \csc x \cot x\,dx = -\csc x + C$$

$$\int \frac{1}{x^2+1}\,dx = \tan^{-1}x + C \qquad \int \frac{1}{\sqrt{1-x^2}}\,dx = \sin^{-1}x + C$$

$$\int \sinh x\,dx = \cosh x + C \qquad \int \cosh x\,dx = \sinh x + C$$

Figure 2: Table of Indefinite Integrals [Stewart]

Theorem 2.31 (Sum of Primitives)**.**

$$\int [f(x)+g(x)]dx = \int f(x)dx + \int g(x)dx$$
$$\int cf(x)dx = c\int f(x)dx$$

As an example, let us evaluate the following integral:

$$\begin{aligned}
&\int (1+x^2)^2 + 9e^x + \frac{\pi}{x}dx \\
&= \int (1+x^2)^2 dx + 9\int e^x dx + \pi\int \frac{1}{x}dx \\
&= \int dx + 2\int x^2 dx + \int x^4 dx + 9\int e^x dx + \pi\int \frac{1}{x}dx \\
&= x + \frac{2}{3}x^3 + \frac{1}{5}x^5 + 9e^x + \pi\ln x
\end{aligned}$$

A useful theorem involves viewing a function as a product of a function f with a simple derivative, and a function that is in the form of g':

Theorem 2.32 (Integration by Parts)**.** *If f' and g' are continuous, then*

$$\int fg' = fg - \int f'g$$
$$\int f(x)g'(x)dx = f(x)g(x) - \int f'(x)g(x)dx$$
$$\int_a^b f(x)g'(x)dx = f(x)g(x)|_a^b - \int_a^b f'(x)g(x)dx$$

There are two special tricks which often work with integration by parts. The first is to consider the function g' to be 1, which can always be written in:

$$\int log(x)dx = \int \underbrace{1}_{g'} * \underbrace{log(x)}_{f} dx = xlog(x) - \int x * \frac{1}{x}dx = x(log(x)) - x$$

The second trick is to use integration by parts to find $\int h$ in terms of $\int h$ again, and then solve for $\int h$:

$$\int \underbrace{\frac{1}{x}}_{g'} \underbrace{log(x)}_{f} dx = log(x)log(x) - \int \frac{1}{x}log(x)dx$$
$$\implies 2\int \frac{1}{x}log(x)dx = (log(x))^2$$
$$\implies \int \frac{1}{x}log(x) = \frac{(log(x))^2}{2}$$

Another important method of integration is a consequence of the chain rule.

Theorem 2.33 (The Substitution Formula)**.** *If f' and g' are continuous, then*

$$\int_{g(a)}^{g(b)} f = \int_a^b (f \circ g)g'$$
$$\int_{g(a)}^{g(b)} f(u)du = \int_a^b f(g(x))g'(x)dx$$

To use this method, it is useful to use the following procedure:

1. Let $u = g(x)$ and $du = g'(x)dx$ so that only the letter u appears, not x
2. Find a primitive as an expression involving u
3. Substitute $g(x)$ back for u

2.6 Infinite Sequences*

Definition 2.6 (Infinite Sequence)**.** *An infinite sequence of real numbers $\{a_1, a_2, a_3, ...\}$ is a function $f : \mathbb{N} \to \mathbb{R}$, i.e. whose domain is $\mathbb{N}$.*

Definition 2.7 (Limit of a Sequence)**.** *A sequence $\{a_n\}$ converges to l, $\lim_{n\to\infty} a_n = l$, if for every $\epsilon > 0$, there is a natural number $\mathbb{N}$ such that, for all natural numbers n,*

$$n > N \implies |a_n - l| < \epsilon$$

A sequence is said to **converge** if it converges to l for some l, and to **diverge** if it does not converge.

Theorem 2.34 (Sum and Product of Limits of Sequences)**.** *If $\lim_{n\to\infty} a_n$ and $\lim_{n\to\infty} b_n$ both exist, then*

$$\lim_{n\to\infty} (a_n + b_n) = \lim_{n\to\infty} a_n + \lim_{n\to\infty} b_n$$
$$\lim_{n\to\infty} (a_n * b_n) = \lim_{n\to\infty} a_n * \lim_{n\to\infty} b_n$$

Moreover, if $\lim_{n\to\infty} b_n \neq 0$, then $b_n \neq 0$ for all n greater than some N, and

$$\lim_{n\to\infty} \frac{a_n}{b_n} = \frac{\lim_{n\to\infty} a_n}{\lim_{n\to\infty} b_n}$$

Theorem 2.35 (Limits of Functions and Sequences)**.** *Let f be a function defined in an open interval containing c, except perhaps at c itself, with $lim_{x\to c} f(x) = l$. Then, suppose that $\{a_n\}$ is a sequence such that*

1. *Each a_n is in the domain of f*
2. *Each $a_n \neq c$*
3. *$\lim_{n\to\infty} a_n = c$*

Then, the sequence $\{a_n\}$ satisfies $\lim_{n\to\infty} f(a_n) = l$. Conversely, if this is true for every sequence $\{a_n\}$ satisfying the above conditions, then $\lim_{x\to c} f(x) = l$.

Theorem 2.36 (Bounded increasing sequences converge)**.** *If $\{a_n\}$ is nondecreasing and bounded above, then $\{a_n\}$ converges.*

Definition 2.8 (Subsequence)**.** *A subsequence of a sequence $\{a_n\}$ is a sequence of the form $\{a_{n_1}, a_{n_2}, a_{n_3}, ...\}$ where $n_j \in \mathbb{N}$ such that $n_1 < n_2 < n_3 < ....$*

Theorem 2.37 (Monotone Subsequence)**.** *Any sequence $\{a_n\}$ contains a subsequence which is either nondecreasing or nonincreasing.*

Corollary 2.37.1 (Bolzano-Weierstrass Theorem)**.** *Every bounded sequence has a convergent subsequence.*

Definition 2.9 (Cauchy Sequence)**.** *A sequence $\{a_n\}$ is a Cauchy sequence if for every $\epsilon > 0$, there is a natural number N such that, for all m and n,*

$$m, n > N \implies |a_n - a_m| < \epsilon \iff \lim_{m,n\to\infty} |a_m - a_n| = 0$$

Theorem 2.38 (Cauchy Convergence)**.** *If a subsequence of a Cauchy sequence converges, then the Cauchy sequence itself converges.*

2.7 Infinite Series

We introduced infinite sequences in the previous section so that we could consider their "sums" now, i.e. $a_1 + a_2 + a_3 + ...$ This is not fully straightforward, since the sum of infinitely many numbers needs to be defined. What we can easily define are the **partial sums** $s_n = a_1 + ... + a_n$.

If we hope to compute the infinite sum $a_1 + a_2 + a_3 + ...$, it must be the case that the partial sums s_n represent closer and closer approximations as n grows. Thus, we may define the "infinite sum" to be exactly this limit. This approach will necessarily leave the sum of many sequences undefined, since the sequence $\{s_n\}$ may fail to have a limit (e.g. $\{a_n\} = 1, -1, 1, -1, ...$ has sums $s_1 = 1, s_2 = 0, s_3 = 1, s_4 = 0, ...$ for which $\lim_{n\to\infty} s_n$ does not exist).

Definition 2.10 (Sum of Infinite Sequence)**.** *A sequence $\{a_n\}$ is summable if the sequence $\{s_n\}$ converges, where $s_n = a_1 + ... + a_n$. In this case, $\lim_{n\to\infty} s_n$ is denoted by* ***the infinite series***

$$\sum_{n=1}^{\infty} a_n \text{ or } a_1 + a_2 + a_3 + ...$$

and is called the sum of the sequence $\{a_n\}$.

Theorem 2.39 (Cauchy Criterion)**.** *The sequence $\{a_n\}$ is summable iff $\lim_{m,n\to\infty} a_{n+1} + ... + a_m = 0$*

Although the Cauchy criterion provides an equivalence for the summability of any given sequence, it is not particularly useful for figuring out if a particular sequence is summable in practice. Instead, there are various conditions that are used in order to determine summability:

Theorem 2.40 (The Vanishing Condition)**.** *If $\{a_n\}$ is summable, then $\lim_{n\to\infty} a_n = 0$*

This condition is necessary but not sufficient. For example, the *Harmonic Series* $\lim_{n\to\infty} \frac{1}{n} = 0$, yet its corresponding sequence is not summable, in fact it is not bounded.

Theorem 2.41 (The Boundedness Criterion)**.** *A nonnegative sequence $\{a_n\}$ is summable if and only if the set of partial sums s_n is bounded.*

This theorem gives rise to multiple tests:

Theorem 2.42 (Tests for Summability)**.**

1. *(Comparison Test) Suppose that $0 \le a_n \le b_n$ for all n. Then if $\sum_{n=1}^{\infty} b$ converges, so does $\sum_{n=1}^{\infty} a_n$.*
2. *(Limit Comparison Test) If $a_n, b_n > 0$ and $\lim_{n\to\infty} \frac{a_n}{b_n} = c \neq 0$, then $\sum_{n=1}^{\infty} a_n$ converges iff and only if $\sum_{n=1}^{\infty} b$ converges.*
3. *(Ratio Test) Let $a_n > 0$ for all n and suppose that $\lim_{n\to\infty} \frac{a_{n+1}}{a_n} = r$. Then $\sum_{n=1}^{\infty} a_n$ converges if $r < 1$ and diverges if $r > 1$.*

2.7.1 Examples of Series

Definition 2.11 (Geometric Series)**.** *The most important of all infinite series are the geometric series.*

$$\sum_{n=0}^{\infty} r^n = 1 + r + r^2 + r^3 + r^4 + ...$$

Only the cases $|r| < 1$ are interesting, since the individual terms do not approach 0 otherwise. These series can be managed because the partial sums can be evaluated in simple terms:

$$\begin{aligned} s_n &= 1 + r + ... + r^n \\ &\implies rs_n = r + r^2 + ...r^{n+1} \\ &\implies s_n(1-r) = 1 - r^{n+1} \\ &\implies s_n = \frac{1 - r^{n-1}}{1 - r} \end{aligned}$$

Then it follows that

$$\sum_{n=0}^{\infty} r^n = \lim_{n\to\infty} \frac{1 - r^{n-1}}{1 - r} = \frac{1}{1 + r} \text{ for } |r| < 1$$

For example,

$$\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + ... = \sum_{n=1}^{\infty} \left(\frac{1}{2}\right)^n = \sum_{n=0}^{\infty} \left(\frac{1}{2}\right)^n - 1 = \frac{1}{1 - \frac{1}{2}} - 1 = 1$$

2.7.2 Present Value of a Stream of Payments*

In many economic settings, we need to compute the present value of a series of future cash flows. Consider an individual who makes annual payments at the end of each year in amount V for T years, with an interest rate of r. Then, we can calculate the present value of this stream of payments as:

$$PV = \frac{V}{(1+r)^1} + \frac{V}{(1+r)^2} + ... + \frac{V}{(1+r)^T} = \sum_{t=1}^{T} \frac{V}{(1+r)^t}$$

2.7.3 Approximation by Polynomial Functions

It is useful to be able to reduce the calculation of a function f to the evaluation of a polynomial function. The method depends on finding polynomial functions which are close approximations to f.

Consider a polynomial function $p(x)$:

$$p(x) = a_0 + a_1 x + ... + a_n x^n$$

It is interesting to note that the coefficients a_i can be expressed in terms of the value of p and its various

derivatives evaluated at 0.

$$p'(x) = a_1 + 2a_2x + ... + na_nx^{n-1}$$
$$p'(0) = a_1$$

We can further differentiate again,

$$p''(x) = 2a_2 + 3 * 2a_3x + ... + n(n-1)a_nx^{n-2}$$
$$p''(0) = 2a_2$$

In general, we arrive at the following expression for a given coefficient on a polynomial function:

$$a_k = \frac{p^{(k)}(0)}{k!}$$

We can also consider a polynomial around a point a rather than 0, to write:

$$p(x) = a_0 + a_1(x-a) + ... + a_n(x-a)^n$$
$$a_k = \frac{p^{(k)}(a)}{k!}$$

Definition 2.12 (Taylor Polynomial of degree n for f at a)**.** *Suppose that f is a function such that its first n derivatives $f^{(1)}(a), ..., f^{(n)}(a)$ at a all exist. Let*

$$a_k = \frac{f^{(k)}(a)}{k!} \quad , 0 \leq k \leq n$$

And define

$$P_{n,a}(x) = a_0 + a_1(x-a) + ... + a_n(x-a)^n$$

Then $P_{n,a}(x)$ is called the **Taylor polynomial of degree n for f at a**, and has been defined in such a way so that

$$P_{n,a}^{(k)}(a) = f^{(k)}(a) \text{ for } 0 \leq k \leq n$$

Thus we have found a polynomial that has the same first n derivatives as f at a. We argue that this is a good approximation for f at a. In order to justify this, we need to look at the error term $r_{n,a}(x) = f(x) - p_{n,a}(x)$. In particular, in order for $p_{n,a}(x)$ to represent a good k-th order approximation to f, we should require that the remainder tends to 0 faster than k-th order, i.e.

$$\lim_{x \to a} \frac{r_{n,a}(x)}{(x-a)^k} = 0$$

Theorem 2.43 (Taylor's Theorem)**.** *Suppose that f is $n+1$ times differentiable on the interval I with*

$a \in I$. For each $x \in I$, there is a point c between a and x such that

$$r_{n,a}(x) = \frac{f^{(n+1)}(c)}{(n+1)!}(x-a)^{n+1}$$

Corollary 2.43.1. *Suppose that f is $n+1$ times differentiable on the interval I with $a \in I$. Then,*

$$\lim_{x \to a} \frac{r_{n,a}(x)}{|x-a|^n} = 0$$

This corollary implies that the Taylor polynomial is a good approximation, since the error vanishes faster than order n. The Taylor formula opens up the way for most of the calculations of applied analysis, and is extremely important from a practical point of view.

Notice further that we can make a Taylor polynomial of as high a degree as we'd like. The **Taylor Series** is just the Taylor polynomial with infinite degree. Let us consider an example, and find the Taylor series for $f(x) = e^x$ at $a = 1$. Since all derivatives of $f(x)$ are e^x, we have that $f^{(n)}(1) = e$ for all $n \geq 0$. Thus, its Taylor series at 1 is:

$$\sum_{n=0}^{\infty} \frac{e}{n!}(x-1)^n$$

2.8 Resources

2.8.1 Textbooks

1. *Calculus by Michael Spivak* for a beautiful and rigorous exposition of single variable calculus. These lecture notes draw from this text.

2. *Calculus by James Stewart* is the standard undergraduate single variable calculus text.

3. *Mathematics for Economics by Hoy et al.* chapters 3.1—3.5; 4.1—4.3; 5.1—5.6; 16.1—16.5.

2.8.2 Web Resources

1. Integration Techniques: http://math.mit.edu/~nehcili/data/mat136_integration.pdf

2. Lebesgue Integration: http://williamchen-mathematics.info/lnilifolder/ili04.pdf

3. Videos on the essence of calculus by 3blue1brown: https://www.youtube.com/playlist?list=PLZHQObOWTQDMsr9K-rj53DwVRMYO3t5Yr

4. Videos on differential equations: https://www.youtube.com/playlist?list=PLZHQObOWTQDNPOjrT6KVlfJuKtYTftqH6

5. General calculus notes: https://tutorial.math.lamar.edu/

6. General calculus notes: MAT 137 http://home.tykenho.com/index.html?notes

7. Implicit differentiation: https://tutorial.math.lamar.edu/classes/calci/implicitdiff.aspx

2.8.3 Practice Problems with Solutions

1. MIT: https://ocw.mit.edu/courses/mathematics/18-01sc-single-variable-calculus-fall-2010/syllabus/

2. Khan Academy: https://www.khanacademy.org/math/calculus-1

3. Practice Problems: https://tutorial.math.lamar.edu/Problems/CalcI/CalcI.aspx

---

3 Linear Algebra

In many financial or economic problems, a single linear equation identifies or characterizes a relationship between two variables, x and y. Often, however, there are two or more equations that must be satisfied *simultaneously*. **Linear Algebra** provides powerful methods for finding solutions to two or more linear equations. A linear equation is any equation of the form

$$c_1x_1 + c_2x_2 + ... + c_nx_n = b \text{ for } c_1, ...c_n, b \in \mathbb{R}$$

where c_i are coefficients of the linear equation and b is the constant term.

3.1 Vectors and the Geometry of Linear Algebra

To understand the intuition and theory behind the tools that Linear Algebra provides for solving systems of linear equations, it is useful to think about the geometry underlying the results.

Previously, we have considered points in Euclidean space as ordered n-tuples of numbers. We can define these n-tuples as *vectors*. In fact, $\mathbb{R}^n$ admits a vector space structure, i.e. our standard rules of addition, subtraction, and scalar multiplication will continue to behave nicely when considering vectors rather than numbers, which will lead to useful calculation methods.

Definition 3.1 (Vector)**.** *A **vector** x is an array of real numbers, where the corresponding row vector is x' ("x prime").*

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \qquad x' = \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix}$$

3.1.1 Vector Addition and Scalar Multiplication

Definition 3.2 (Vector Addition)**.** *To add or subtract two vectors, add or subtract the corresponding components.*

$$x + y = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} x_1 + y_1 \\ x_2 + y_2 \\ x_3 + y_3 \end{bmatrix} = y + x$$

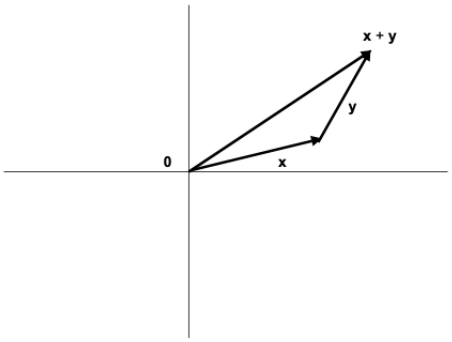


Figure 6.1: Vector Addition

Theorem 3.1 (Properties of Vector Addition)**.**

1. *Vector addition is commutative, i.e.* $x + y = y + x$
2. *For* $n \neq m$*, cannot add* $x^n + y^m$

Definition 3.3 (Scalar Multiplication)**.** *For* $k \in \mathbb{N}$*,*

$$k * x = k * \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} k * x_1 \\ k * x_2 \\ k * x_3 \end{bmatrix} = x * k$$

Theorem 3.2 (Properties of Scalar Multiplication)**.**

1. *Scalar multiplication is commutative, i.e.* $k * x = x * k$
2. *Scalar multiplication is distributive over vector addition, i.e.* $k(x + y) = k * x + k * y$ *for all* $x, y \in \mathbb{R}^n; k \in \mathbb{N}$
3. *Scalar multiplication is distributive over scalar addition, i.e.* $(h + k)x = h * x + k * x$ *for all* $x \in \mathbb{R}^n; h, k \in \mathbb{N}$

3.1.2 Vector Length and Distance

Given our understanding of vectors, we want to be able to think about concepts of length and distance. We will define these concepts largely as a generalization of the familiar concept of Euclidean distance as defined in the Fundamentals section.

As a starting point, the concept of an *inner product* will be useful for this purpose.

Definition 3.4 (Inner Product)**.** *An Inner Product is a linear function that assigns a number given two vectors, i.e.* $f : \mathbb{R}^2 \times \mathbb{R}^2 \mapsto \mathbb{R}$ *(and satisfies certain nice properties). The standard inner product, or "Dot*

Product" of two vectors $x, y \in \mathbb{R}^n$ is:

$$x \cdot y = x'y = \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} * \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = x_1 * y_1 + x_2 * y_2 + x_3 * y_3 = \sum_{i=1}^{n} x_i * y_i$$

Theorem 3.3 (Properties of Inner Product)**.**

1. *The Inner Product is commutative, i.e.* $x \cdot y = y \cdot x$
2. *For* $x, y \in \mathbb{R}^n$ *and* $k \in \mathbb{R}$*:* $k(x'y) = (k * x)'y = x'(k * y)$

Given this, we can define the concept of the length, or *norm* of a vector.

Definition 3.5 (Norm)**.** *Given an Inner Product* $x \cdot y$*, we can define the norm or length of any vector as* $||x|| = \sqrt{x \cdot x}$*.*

The norm of a vector satisfies several important properties:

Theorem 3.4 (Properties of a Norm)**.**

1. $||c * x|| = |c| \cdot ||x||$ *for* $c \in \mathbb{R}$
2. $||x|| = 0 \iff x = 0$ *and for all* x, $||x|| \geq 0$
3. *Cauchy-Shwarz Inequality:* $|x \cdot y| \leq ||x|| \cdot ||y||$
4. *Triangle Inequality:* $||x + y|| \leq ||x|| + ||y||$

Finally, we can define a notion of distance, or a *metric*, between two vectors.

Definition 3.6 (Distance Between Vectors)**.** *Given an Inner Product* $x \cdot y$*, we can define a metric, or distance function, between any two vectors as* $d(x, y) = ||x - y|| = \sqrt{(x - y) \cdot (x - y)}$

Notice that the definitions of length and distance coincide with their Euclidean formulations with the standard inner product.

3.1.3 Angle Between Two Vectors*

If we think of two vectors x and y emanating from the same point, we can speak of the *angle* θ between them.

$$\theta = arccos\left(\frac{x \cdot y}{||x||||y||}\right)$$

In particular, the directions of the two vectors x and y are said to be perpendicular to each other if and only if $x \cdot y = 0$. In this case the vectors are said to be orthogonal to each other.

3.2 Matrices

Definition 3.7 (Matrix)**.** *A matrix A is a rectangular array of numbers. A number appearing in a matrix is called an entry of A. If the array has n rows and m columns, we say that A has size n by m, or that A is an n by m matrix. We denote the entry of A appearing in the ith row and jth column as a_{ij}, where i is the row index and j is the column index of this entry.*

$$A \in \mathbb{R}^3 \times \mathbb{R}^3 = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

3.2.1 Matrix Algebra

Definition 3.8 (Matrix Addition and Subtraction)**.** *Addition and subtraction of matrices is well defined only if the matrices involved are of the same size. The sum of two matrices is a matrix, the elements of which are the sums of the corresponding elements of matrices (and analogously for matrix subtraction).*

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} + \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \begin{bmatrix} a_{11} + b_{11} & a_{12} + b_{12} \\ a_{21} + b_{21} & a_{22} + b_{22} \end{bmatrix}$$

Definition 3.9 (Scalar Multiplication)**.** *In matrix algebra, real numbers are called scalars. Multiplying a matrix by a scalar is called scalar multiplication, and it is carried out by multiplying each element of the matrix by the scalar.*

$$3A = \begin{bmatrix} 3a_{11} & 3a_{12} \\ 3a_{21} & 3a_{22} \end{bmatrix}$$

Definition 3.10 (Matrix Multiplication)**.** *Multiplication of matrices is well defined only if the matrices involved are conformable. This is the case if the number of columns of the first matrix are the same as the number of rows of the second matrix. The product of two matrices $A \in \mathbb{R}^{m \times n} * B \in \mathbb{R}^{n \times q}$ is a matrix $C \in \mathbb{R}^{m \times q}$, and its ijth element c_{ij} is obtained by multiplying the elements of the ith row of A by the elements of the jth column of B and adding the resulting products.*

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} * \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{bmatrix} = \begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} + a_{13}b_{31} & a_{11}b_{12} + a_{12}b_{22} + a_{13}b_{33} \\ a_{21}b_{11} + a_{22}b_{21} + a_{23}b_{31} & a_{21}b_{12} + a_{22}b_{22} + a_{23}b_{32} \end{bmatrix} \in \mathbb{R}^{2 \times 2}$$

In general, the product matrix can be expressed as $c_{ij} = \sum_{k=1}^{n} a_{jk} b_{kj}$.

Notice the similarity between the expression for a given element in the resulting matrix c_{ij}, and the notion of taking a dot product between two vectors. In fact, now that we have defined the concept of matrix multiplication we can further define the **outer product** of two vectors:

Definition 3.11 (Outer Product*)**.** *Given two vectors u, v of size $m \times 1$ and $n \times 1$ respectively, their outer*

product $u \otimes v$ is the $m \times n$ matrix A obtained via matrix multiplication:

$$\begin{bmatrix} u_1 \\ ... \\ u_m \end{bmatrix} * \begin{bmatrix} v_1 & ... & v_n \end{bmatrix} = \begin{bmatrix} u_1v_1 & u_1v_2 & ... & u_1v_n \\ u_2v_1 & u_2v_2 & ... & u_2v_n \\ ... & ... & ... & ... \\ u_mv_1 & u_mv_2 & ... & u_mv_n \end{bmatrix} \in \mathbb{R}^{m \times n}$$

Theorem 3.5 (Properties of Matrix Multiplication)**.**

1. *Matrix multiplication is not commutative. In fact, even if $A * B$ exists, $B * A$ may not exist depending on the dimensions of the matrices.*
2. *Matrix multiplication is distributive with respect to addition: $A(B+C) = AB + AC$ and $(B+C)A = BA + CA$*
3. *Matrix multiplication is associative: $A(BC) = (AB)C$*
4. *Matrix multiplication is commutative with respect to scalars: $A(\lambda B) = \lambda(AB)$*
5. *The unit matrix is the neutral element of matrix multiplication: $IA = AI = A$*
6. *The zero matrix is absorbent: $0A = A0 = 0$*

3.2.2 Examples of Matrices

Definition 3.12 (Square Matrix)**.** *A matrix that has the same number of rows and columns is called a square matrix*

$$A \in \mathbb{R}^2 \times \mathbb{R}^2 = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

Definition 3.13 (Triangular Matrix)**.** *A matrix A that has 0s in every element above or below its diagonal is a triangular matrix. If all elements below the diagonal are 0, then A is an upper-triangular matrix. If all elements above the diagonal are 0, then A is a lower-triangular matrix.*

$$U = \begin{bmatrix} a & b & c \\ 0 & d & e \\ 0 & 0 & f \end{bmatrix} \quad L = \begin{bmatrix} a & 0 & 0 \\ b & c & 0 \\ d & e & f \end{bmatrix}$$

Definition 3.14 (Diagonal Matrix)**.** *A square matrix that has only nonzero entries on the main diagonal and zeroes everywhere else is known as a diagonal matrix. A diagonal matrix is also triangular.*

$$D \in \mathbb{R}^3 \times \mathbb{R}^3 = \begin{bmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{bmatrix}$$

Definition 3.15 (Identity Matrix I)**.** *A special case of a diagonal matrix is the identity matrix, where the diagonal entries are all 1s. We will see that the identity matrix plays the same role in matrix algebra as the number 1 does in the algebra of real numbers.*

$$I \in \mathbb{R}^3 \times \mathbb{R}^3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Definition 3.16 (Null Matrix)**.** *A square matrix with all its entries being 0 is known as the null matrix. The null matrix plays a similar role in matrix algebra as does 0 in the algebra of real numbers.*

$$0 \in \mathbb{R}^2 \times \mathbb{R}^2 = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

Definition 3.17 (Transpose Matrix)**.** *The transpose matrix A^T is the original matrix A with its rows and columns interchanged*

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \qquad A^T = \begin{bmatrix} a & c \\ b & d \end{bmatrix}$$

1. *A matrix that is equal to its transpose is called a symmetric matrix.*
2. *The transpose of a transpose matrix is the original matrix: $(A^T)^T = A$.*
3. *The transpose of a sum of matrices is the sum of transposes $(A + B)^T = A^T + B^T$*
4. *The transpose of a product of matrices: $(AB)^T = B^T * A^T$*

Definition 3.18 (Orthogonal Matrix)**.** *Two vectors $x, y \in \mathbb{R}^n$ are **orthogonal** if $x^T y = 0$ (i.e., if their dot product is 0). A vector $x \in \mathbb{R}^n$ is **normalized** if its norm is 1, $|x| = 1$. A square matrix $U \in \mathbb{R}^{n \times n}$ is orthogonal if all of its columns are orthogonal to each other and are normalized. The columns are then referred to as being orthonormal. Then,*

$$U^T U = U U^T = I_n$$

3.2.3 Solving Systems of Linear Equations*

Solving a system of linear equations is probably the most important application of linear algebra for our purposes. For systems of two linear equation, finding solutions by simple substitution of unknowns works well, but for systems with a larger number of unknowns we need a new tractable solution method. One method of doing so involves reformulating a given system of linear equations into a form that has the same solution but is easier to solve. There are three types of **row operations** that are used to simplify the original system:

1. Interchange the order of any two equations in the system.

2. Multiply any equation in the system by a nonzero constant.

3. Add a constant multiple of any equation to another equation in the system.

And we would like to arrive at a system that has the following properties:

1. The first nonzero coefficient of each equation is one.

2. If an unknown is in the first unknown with a nonzero coefficient in some equation, then that unknown occurs with a 0 coefficient in each of the other equations.

3. The first unknown with a nonzero coefficient in any equation has a larger subscript than the first unknown with a nonzero coefficient in any preceeding equation.

Lets see an example of this method. We are given the following system of equations:

$$\begin{aligned} a_1 - 2a_2 + 2a_4 - 3a_5 &= 2 && (1)\\ 2a_1 - 4a_2 + 2a_3 + 8a_5 &= 6 && (2)\\ a_1 - 2a_2 + 3a_3 - 3a_4 + 16a_5 &= 8 && (3) \end{aligned}$$

As a first step, let us eliminate a_1 from every equation except the first by using rule 3. We add $-2 * (1)$ to equation (2), and $-1 * (1)$ to equation (3). While doing this, it will so happen that we also eliminate a_2 from every equation except the first.

$$\begin{aligned} a_1 - 2a_2 + 2a_4 - 3a_5 &= 2 && (1)\\ 2a_3 - 4a_4 + 14a_5 &= 2 && (2)\\ 3a_3 - 5a_4 + 19a_5 &= 6 && (3) \end{aligned}$$

We now want to make the coefficient of a_3 in equation (2) equal to 1, and then eliminate a_3 from equation (3). We use rule 2 to multiply equation (2) by $\frac{1}{2}$, and then add $-3 * (2)$ to equation (3).

$$\begin{aligned} a_1 - 2a_2 + 2a_4 - 3a_5 &= 2 && (1)\\ a_3 - 2a_4 + 7a_5 &= 1 && (2)\\ a_4 - 2a_5 &= 3 && (3) \end{aligned}$$

We continue by eliminating a_4 by every equation except the (3). We do so by adding $-2 * (3)$ to equation (1) and adding $2 * (3)$ to equation (2). This yields:

$$\begin{aligned} a_1 - 2a_2 + a_5 &= -4 && (1)\\ a_3 + 3a_5 &= 7 && (2)\\ a_4 - 2a_5 &= 3 && (3) \end{aligned}$$

We have arrived in a system of linear equations that is easy to solve. We can solve for the first unknown present in each equation in terms of the other unknowns (a_2, a_5). This results in:

$$a_1 = 2a_2 - a_5 - 4 \tag{1}$$
$$a_3 = -3a_5 + 7 \tag{2}$$
$$a_4 = 2a_5 + 3 \tag{3}$$

Thus for any choice of $a_2, a_5 \in \mathbb{R}$, a vector of the following form is a solution to this system.

$$(a_1, a_2, a_3, a_4, a_5) = (2a_2 - a_5 - 4, a_2, -3a_5 + 7, 2a_5 + 3, a_5)$$

Notice that this system has infinitely many solutions. In contrast, a system can also have a unique solution, or no solution at all (if these operations ever result in an equation of the form $0 = c$ for $c \neq 0$). A system of equations which yields no solution is said to be **inconsistent**. A system with infinitely many solutions is said to be **underdetermined**, i.e. there are *free variables* such as a_2, a_5 in the example above.

Theorem 3.6 (Solutions of a System of Linear Equations)**.** *A system of linear equations has either no solution, exactly one solution, or infinitely many solutions.*

3.2.4 Elementary Matrix Operations and Gaussian Elimination*

We can express a system of linear equations as a single matrix equation by writing the constants of the system into a matrix. In this representation of the system, the three operations detailed above for solving systems of linear equations are the "elementary row operations" for matrices. These operations provide a convenient computational method for determining all solutions to a system of linear equations.

Definition 3.19 (Elementary Row (Column) Operations)**.** *Let A be an $n \times n$ matrix. Any one of the following three operations on the rows or columns of A is called an elementary row (column) operation:*

1. *Interchanging any two rows (columns) of A*
2. *Multiplying any row (column) of A by a nonzero constant*
3. *Adding any constant multiple of a row (column) of A to another row (column).*

Just as before, it is the case that these elementary row operations preserve the solution set of the system of linear equations. Using these operations, we can convert the matrix into a form that is useful for solving the system.

Definition 3.20 (Reduced Row Echelon Form)**.** *A matrix is said to be in reduced row echelon form if the following three conditions are satisfied:*

1. *Any row containing a nonzero entry precedes any row in which all the entries are 0 (if any)*
2. *The first nonzero entry in each row is the only nonzero entry in its column.*

3. *The first nonzero entry in each row is 1 and it occurs in a column to the right of the first nonzero entry in the preceding row.*

Theorem 3.7 (Unique Reduced Row Echelon Form)**.** *Every matrix admits a unique reduced row echelon form.*

Let us see an example of this. Consider the following system:

$$3x_1 + 2x_2 + 3x_3 - 2x_4 = 1 \tag{1}$$
$$x_1 + x_2 + x_3 = 3 \tag{2}$$
$$x_1 + 2x_2 + x_3 - x_4 = 2 \tag{3}$$

Let us rewrite this system into a matrix and use the elementary matrix operations.

$$\left(\begin{array}{cccc|c} 3 & 2 & 3 & -2 & 1 \\ 1 & 1 & 1 & 0 & 3 \\ 1 & 2 & 1 & -1 & 2 \end{array}\right)$$

Let us create a 1 in a_{11} by interchanging the first and third rows.

$$\left(\begin{array}{cccc|c} 1 & 2 & 1 & -1 & 2 \\ 1 & 1 & 1 & 0 & 3 \\ 3 & 2 & 3 & -2 & 1 \end{array}\right)$$

Let us obtain 0s in a_{21} and a_{31} by adding -1 times the first row to the second row, and -3 times the first row to the third row.

$$\left(\begin{array}{cccc|c} 1 & 2 & 1 & -1 & 2 \\ 0 & -1 & 0 & 1 & 1 \\ 0 & -4 & 0 & 1 & -5 \end{array}\right)$$

Let us obtain a 1 in a_{22} by multiplying the second row by -1.

$$\left(\begin{array}{cccc|c} 1 & 2 & 1 & -1 & 2 \\ 0 & 1 & 0 & -1 & -1 \\ 0 & -4 & 0 & 1 & -5 \end{array}\right)$$

Let us obtain a 0 in a_{32} by adding 4 times the second row to the third row.

$$\left(\begin{array}{cccc|c} 1 & 2 & 1 & -1 & 2 \\ 0 & 1 & 0 & -1 & -1 \\ 0 & 0 & 0 & -3 & -9 \end{array}\right)$$

Let us obtain a 1 in a_{33} by multiplying the third row by $-\frac{1}{3}$.

$$\left(\begin{array}{cccc|c} 1 & 2 & 1 & -1 & 2 \\ 0 & 1 & 0 & -1 & -1 \\ 0 & 0 & 0 & 1 & 3 \end{array}\right)$$

We have an *upper triangular matrix* i.e., the lower left triangle of the matrix below the diagonal is populated with 0s. Now we want to make the first nonzero entry in each row the only nonzero entry in its column. To do this, lets work upwards (i.e. via *backwards substitution*) and first add 1 times the third row to the second and first row.

$$\left(\begin{array}{cccc|c} 1 & 2 & 1 & 0 & 5 \\ 0 & 1 & 0 & 0 & 2 \\ 0 & 0 & 0 & 1 & 3 \end{array}\right)$$

Now we add -2 times the second row to the first row, and have created a matrix in row echelon form from applying elementary matrix operations to our original system.

$$\left(\begin{array}{cccc|c} 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 2 \\ 0 & 0 & 0 & 1 & 3 \end{array}\right)$$

This matrix corresponds to a system of linear equations that can be easily solved.

$$x_1 + x_3 = 1 \tag{1}$$
$$x_2 = 2 \tag{2}$$
$$x_4 = 3 \tag{3}$$

A vector of the following form is a solution to this system:

$$(1 - t, 2, t, 3) \quad \text{for } t \in \mathbb{R}$$

3.2.5 Linear Dependence and Properties of Linear Systems

Definition 3.21 (Linear Dependence)**.** *A set of vectors $\{x_1, ...x_m\} \subset \mathbb{R}^n$ is said to be linearly independent if no vector can be represented as a linear combination of the remaining vectors. Conversely, if one vector belonging to the set* ***can*** *be represented as a linear combination of the remaining vectors, then the vectors are said to be linearly dependent. That is, if for example*

$$x_m = \sum_{i=1}^{m-1} \alpha_i x_i \text{ for some scalar values } \alpha_1, ..., \alpha_{n-1} \in \mathbb{R}$$

Then we say that the vectors $\{x_1, ...x_m\}$ are linearly dependent.

Definition 3.22 (Rank of a Matrix)**.** *Given a matrix $A \in \mathbb{R}^{m \times n}$, the size of the largest subset of columns of A that constitute a linearly independent set is the* ***column rank*** *of A. With some abuse of terminology, this is often referred to as the number of linearly independent columns of A. Similarly, the* ***row rank*** *is the largest number of rows of A that constitute a linearly independent set. For any matrix $A \in \mathbb{R}^{m \times n}$, the column rank* ***equals*** *the row rank, and so both quantities are referred to collectively as the* ***rank*** *of A, denoted as $rank(A)$.*

Theorem 3.8 (Properties of Matrix Rank)**.**

1. *For $A \in \mathbb{R}^{m \times n}$, $rank(A) \leq min(m, n)$. If $rank(A) = min(m, n)$, then A is said to be full rank.*
2. *For $A \in \mathbb{R}^{m \times n}$, $rank(A) = rank(A^T)$.*
3. *For $A \in \mathbb{R}^{m \times n}$, $B \in \mathbb{R}^{n \times p}$, $rank(AB) \leq min(rank(A), rank(B))$*
4. *For $A, B \in \mathbb{R}^{m \times n}$, $rank(A + B) \leq rank(A) + rank(B)$*

Definition 3.23 (Span)**.** *The span of a set of vectors $\{x_1, ...x_n\}$ is the set of all vectors that can be expressed as a linear combination of $\{x_1, ...x_n\}$. That is,*

$$span(\{x_1, ...x_n\}) = \left\{ v | v = \sum_{i=1}^{n} \alpha_i x_i, \alpha_i \in \mathbb{R} \right\}$$

Definition 3.24 (Range)**.** *The range (or columnspace) of a matrix $A \in \mathbb{R}^{m \times n}$, denoted $R(A)$, is the span of the columns of A.*

Definition 3.25 (Nullspace)**.** *The nullspace of a matrix $A \in \mathbb{R}^{m \times n}$, denoted $N(A)$, is the set of all vectors that equal 0 when multiplied by A, i.e.*

$$N(A) = \{x \in \mathbb{R}^n | Ax = 0\}$$

Note that vectors in $R(A)$ are of size m, while vectors in $N(A)$ are of size n, so vectors in $R(A^T)$ and $N(A)$ are both in $\mathbb{R}^n$. In fact, $R(A^T)$ and $N(A)$ are **orthogonal complements**, i.e. disjoint subsets that together span the entire space of $\mathbb{R}^n$. We denote this as $R(A^T) = N(A)^\perp$.

Definition 3.26 (Basis)**.** *A set of vectors $\{x_1, ...x_n\}$ is a basis for some subset $V \subseteq \mathbb{R}^n$ if and only if each vector $v \in V$ can be uniquely expressed as a linear combination of the vectors $\{x_1, ...x_n\}$.*

Theorem 3.9 (Solution to a System of Equations)**.** *The system of equations $Ax = b$ for $A \in \mathbb{R}^{n \times n}$ has a solution iff $Rank([A|b] = Rank(A)$. When the system has a solution,*

1. *The solution is unique iff the columns of A are linearly independent, i.e. $Rank(A) = n$.*
2. *The system has infinitely many solutions iff $Rank(A) < n$.*

3.2.6 Determinants and the Inverse Matrix

We have already defined the operations of addition, subtraction, and multiplication on matrices. What about division? Can we define rules for diving one matrix by another? The answer is yes, but only under certain restrictions. Division is restricted to square matrices that satisfy a condition known as **nonsingularity**, which is equivalent to a square matrix having **full rank**. The reason for all of this can again be traced to the relation between matrix algebra and the problem of solving a system of simultaneous linear equations.

To define the matrix inverse, it is useful to think about the division of real numbers. We can write division by a number b as multiplication by the inverse of b, $\frac{1}{b} = b^{-1}$. This inverse further satisfies that $b * b^{-1} = b^{-1} * b = 1$.

Definition 3.27 (Inverse Matrix)**.** *The inverse matrix A^{-1} of a square matrix A of order n is the matrix that satisfies the condition that $A * A^{-1} = A^{-1}A = I_n$. Note that the inverse is defined only for square matrices, but that not every square matrix has an inverse.*

Theorem 3.10 (Properties of Matrix Inverse)**.**

1. $(A^{-1})^{-1} = A$
2. $(AB)^{-1} = B^{-1}A^{-1}$
3. $(A^{-1})^T = (A^T)^{-1}$

When we deal with real numbers, we know that for any nonzero b, the inverse exists. However for A^{-1} to exist, it is not sufficient simply to assume that A is a square matrix that is different from the null matrix.

Definition 3.28 (Singular Matrix)**.** *Any matrix A for which A^{-1} does not exist is known as a singular matrix. The matrix A for which A^{-1} exists is known as a nonsingular matrix.*

Consider the following example. The matrix equation $Ax = b$, where $A \in \mathbb{R}^{n\times n}; x, b \in \mathbb{R}^{n\times 1}$ defines a system of n simultaneous linear equations in n unknowns. Let us solve for x.

$$\begin{aligned} Ax &= b \\ A^{-1}Ax &= A^{-1}b \\ I_n x &= A^{-1}b \\ x &= A^{-1}b \end{aligned}$$

If A and b are known, this solves for the unknown vector x, provided that A^{-1} exists. We can see from this example that the existence of the inverse matrix is equivalent to being able to solve a linear system.

Definition 3.29 (Inverse of a 2×2 Matrix)**.**

$$A \in \mathbb{R}^2 \times \mathbb{R}^2 = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \implies \quad A^{-1} = \frac{1}{ab - cd}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix} \equiv= \frac{1}{|A|}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

We can verify that $AA^{-1} = A^{-1}A = I_2$. *Clearly in order for the matrix inverse to exist, it must be the case that* $|A| \neq 0$.

Notice in the example of a 2×2 matrix, the inverse exists if and only iff $|A| = ab - cd \neq 0$. In fact, this value is referred to as the **determinant** of the matrix A and can be defined generally for a square matrix of size n. The existence of a nonzero determinant gives an equivalence for the invertibility of a given matrix.

Definition 3.30 (Determinant)**.** *For a square matrix* A*, its determinant, denoted as* $\det(A)$*, is a value defined inductively in the following way:*

1. *For a* 1×1 *matrix* $A = a_{11}$*, define its determinant as* $\det(A) = a_{11}$
2. *For an* $n \times n$ *matrix where* $n \geq 2$*, define its determinant as:*

$$\det(A) \equiv \sum_{j=1}^{n} (-1)^{1+j} a_{1j} det(A_{-i,j})$$

where $A_{-i,j}$ *is the matrix* A *with the* i*th row and* j*th column eliminated (referred to as the* (i,j)*th cofactor of* A*).*

Theorem 3.11 (Invertibility and Determinant)**.**

1. *A square matrix* A *is invertible iff* $\det(A) \neq 0$.
2. *For two* $n \times n$ *matrices, we have that* $\det(AB) = \det(A)\det(B)$.
3. *A matrix and its transpose have the same determinant:* $\det(A) = \det(A^T)$.

Theorem 3.12 (Properties of the Determinant)**.**

1. $\det(I_n) = 1$
2. *The determinant of a triangular matrix* A *is the product of its diagonal elements:* $\det(A) = \Pi_{i=1}^{n} a_{ii}$.
3. *If any two columns of* A *are equal, then* $\det(A) = 0$.
4. *Antisymmetry: If two columns of* A *are interchanged, then the determinant changes sign.*
5. *If one adds a scalar multiple of one column to another, the determinant does not change.*

3.2.7 Eigenvalues, Eigenvectors

Definition 3.31 (Eigenvalues, Eigenvectors)**.** *Let* A *be an* $n \times n$ *matrix. A scalar* λ *is said to be an* ***eigenvalue*** *of* A *if and only if there exists a nonzero vector* x *in* $\mathbb{R}^n$ *such that* $Ax = \lambda x$*. A nonzero vector* x *in* $\mathbb{R}^n$ *is said to be an* ***eigenvector*** *of* A *if and only if there exists a scalar* λ *such that* $Ax = \lambda x$*.*

Theorem 3.13 (Equivalent Condition for Eigenvalues)**.** $\lambda \in \mathbb{R}$ *is an eigenvalue of* A *if and only if* $\det(\lambda I_n - A) = 0$.

Theorem 3.14 (Eigenvalues and Linear Independence)**.** *Let $A \in \mathbb{R}^{n\times n}$ with n distinct eigenvalues $\lambda_1, \lambda_2, ...\lambda_n \in \mathbb{R}$. Let $x_1, ..., x_n \in \mathbb{R}^n$ the corresponding (nonzero) eigenvectors. Then $x_1, ..., x_n$ are linearly independent.*

Theorem 3.15 (Eigenvalues and Invertibility)**.** *A matrix A is invertible if and only if 0 is not one of its eigenvalues.*

Eigenvalues and eigenvectors are one of the most important applications of linear algebra, since there is a sense in which a matrix is effectively determined by these values. We will see this when we look at the application of principal component analysis.

3.2.8 Quadratic Forms

Definition 3.32 (Quadratic Form)**.** *A quadratic form on $\mathbb{R}^n$ is a function $Q : \mathbb{R}^n \to \mathbb{R}$ that can be represented by a unique symmetric matrix A as follows:*

$$Q(x) = x^T Ax = \sum_{i=1}^{n} x_i (Ax)_i = \sum_{i=1}^{n} x_i \left(\sum_{j=1}^{n} A_{ij} x_j \right) = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} x_i x_j$$

The following theorem provides a necessary and sufficient characterization of positive/negative semi-definite matrices using eigenvalues.

Theorem 3.16 (Positive/Negative Semi-Definite Matrix)**.** *Let $A \in \mathbb{R}^{n\times n}$ a symmetric matrix. The matrix A is:*

1. *Positive definite iff all of its eigenvalues are positive.*
2. *Negative definite iff all of its eigenvalues are negative.*
3. *Positive semi-definite iff all of its eigenvalues are non-negative.*
4. *Negative semi-definite iff all of its eigenvalues are non-positive.*
5. *Indefinite iff it has both positive and negative eigenvalues.*

3.3 Resources

3.3.1 Textbooks

1. *Linear Algebra and Its Applications by Lay et al* is a standard reference.
2. *Linear Algebra by Friedberg et al.* is a rigorous exposition.
3. *Mathematics for Economics by Hoy et al.* chapters 7.1—7.2; 8.1—8.4; 9.1—9.4; 10.1—10.3.

3.3.2 Web Resources

1. Videos on the essence of linear algebra by 3blue1brown: https://www.youtube.com/playlist?list=PLZHQObOWTQDPD3MizzM2xVFitgF8hE_ab
2. Notes on Linear Algebra: MAT 223,224 http://home.tykenho.com/index.html?notes

3.3.3 Practice Problems with Solutions

1. Exercises and Problems: https://web.pdx.edu/~erdman/LINALG/Linalg_pdf.pdf
2. Practice problems: https://tutorial.math.lamar.edu/Problems/Alg/Alg.aspx

---

4 Multivariable Calculus*

In our study of analysis, so far we have focused on functions of one variable. In practice, it is often necessary to deal also with functions depending on two, three, or more variables.

4.1 Derivatives

Recall our definition for differentiability for a function of a single variable in $\mathbb{R}$:

Definition 4.1 (Differentiable Function)**.** *The function f is differentiable at a if $\lim_{h\to 0} \frac{f(a+h)-f(a)}{h}$ exists. In this case the limit is denoted by $f'(a)$ and is called the derivative of f at a. We further say that f is differentiable if f is differentiable at a for every a in the domain of f.*

It turns out that this definition is not the most useful when it comes to dealing with functions from $\mathbb{R}^n \to \mathbb{R}^m$. Instead, we can think of a function being differentiable at a point if it is *approximately linear* near that point. Geometrically, this corresponds to the graph of the function f having a tangent line at $x = a$. In particular, if h is sufficiently small, one would hope that there exists an m such that $f(a+h) = f(a) + mh + error(h)$ where $error(h)$ is the corresponding error in the linear approximation. For the approximation to be good, the error should go to zero faster than linearly in h, i.e.

$$\lim_{h\to 0} \frac{error(h)}{h} = 0$$

Using this intuition, we can arrive at an equivalent definition of differentiability:

Definition 4.2 (Differentiable Function)**.** *The function f is differentiable at a if there exists an $m \in \mathbb{R}$ such that*

$$\lim_{h\to 0} \frac{f(a+h)-f(a)-mh}{h} = 0$$

And in this case, $m = f'(a)$.

4.1.1 Vector Valued $\gamma : \mathbb{R} \to \mathbb{R}^n$

The first and simplest generalization of the derivative comes from looking at vector valued functions $\gamma : \mathbb{R} \to \mathbb{R}^n$. These functions are differentiable at a point if each of their component functions are differentiable, and the derivative may be computed by differentiating each component separately.

4.1.2 Multivariable $f : \mathbb{R}^m \to \mathbb{R}^n$

The simplest notion of derivative for a function of several variables $f : \mathbb{R}^n \to \mathbb{R}$ is that of *partial derivatives*, which are just the derivatives of the function with respect to each of its variables when the others are held fixed. Once we fix the values of other variables, then our function of multiple variables becomes again a single variable function.

Definition 4.3 (Partial Derivative)**.** *The partial derivative of a function $f(x_1, ..., x_n)$ with respect to the variable x_j is as below, provided that the limit exists.*

$$\lim_{h \to 0} \frac{f(x_1, ..., x_j - h, ..., x_n) - f(x_1, ..., x_j, ..., x_n)}{h}$$

The partial derivative is denoted as $\dfrac{\partial f}{\partial x_j}$, f_{x_j}, f_j, $\partial_{x_j} f$, or $\partial_j f$.

Further, just as in differentiation of a single variable, we can consider *higher order* partial derivatives. Interestingly though, there are now many different ways of computing a second derivative. We can consider differentiating with respect to the same variable again *(pure partial derivatives)*, or differentiating with respect to a different variable *(mixed partial derivatives)*. In general, given a function $f : \mathbb{R}^n \to \mathbb{R}$, there are n^2 different second-order partial derivatives.

Theorem 4.1 (Young; Schwarz)**.** *If a function f is twice-differentiable at x, then for any $i, j \in \{1, ..., n\}$, both $\dfrac{\partial^2 f}{\partial x_j \partial x_i}(x)$ and $\dfrac{\partial^2 f}{\partial x_i \partial x_j}$ exist and further are equal.*

To think about the total derivative of a multivariable function at a point $a \in \mathbb{R}^n$, the generalization of f being *approximately linear* to multiple dimensions is that f should behave like an *n-plane* near a. The following theorem reveals the relation between the total derivative and the partial derivatives. Namely, the total derivative is a matrix that collects all partial derivatives as its entries.

Definition 4.4 (Total Derivative)**.** *We say that a function $f : \mathbb{R}^n \to \mathbb{R}^m$ is differentiable at a point $a \in \mathbb{R}^n$ if there exists a an $m \times n$ matrix D such that*

$$\lim_{h \to 0} \frac{f(a + h) - f(a) - D \cdot h}{||h||} = 0$$

In this case, define the total derivative of f at x as the matrix D, denoted as $f'(x)$ or $Df(x)$. Furthermore,

we have:

$$Df(x) \in \mathbb{R}^{m\times n} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(x) & \frac{\partial f_1}{\partial x_2}(x) & ... & \frac{\partial f_1}{\partial x_n}(x) \\ \frac{\partial f_2}{\partial x_1}(x) & \frac{\partial f_2}{\partial x_2}(x) & ... & \frac{\partial f_2}{\partial x_n}(x) \\ ... & ... & ... & ... \\ \frac{\partial f_m}{\partial x_1}(x) & ... & ... & \frac{\partial f_m}{\partial x_n}(x) \end{bmatrix}$$

4.1.3 Chain Rule for Partial Derivatives

The general definition of the Chain Rule for multivariable functions can be written as follows:

Theorem 4.2 (Multivariable Chain Rule)**.** *Let $g : \mathbb{R}^k \to \mathbb{R}^n$ and $f : \mathbb{R}^n \to \mathbb{R}^m$. If g is differentiable at $a \in \mathbb{R}^k$ and f is differentiable at $g(a) \in \mathbb{R}^n$, then $f \circ g$ is differentiable at a, and moreover its derivative can be written as*

$$D(f \circ g)(a) = Df(g(a))Dg(a)$$

Notably, if $g : \mathbb{R} \to \mathbb{R}^n$, and $f : \mathbb{R}^n \to \mathbb{R}$, so that $f \circ g : \mathbb{R} \to \mathbb{R}$, by the chain rule we have

$$\frac{d}{dt}(f \circ g) = \frac{\partial y}{\partial x_1}\frac{\partial x_1}{\partial t} + ...\frac{\partial y}{\partial x_n}\frac{\partial x_n}{\partial t}$$

4.2 Resources

4.2.1 Textbooks

1. *Multivariable Calculus by James Stewart* for a simple exposition.
2. *Analysis on Manifolds by James Munkres* for a more advanced treatment.
3. *Mathematics for Economics by Hoy et al.* chapters 11.1—11.6; 17.1; 18.1—18.2; 21.2; 14.1—14.2.

4.2.2 Web Resources

1. List of infinite series and their sums: https://en.wikipedia.org/wiki/List_of_mathematical_series
2. Notes on Multivariate Calculus: MAT 237 http://home.tykenho.com/index.html?notes
3. Implicit function theorem: https://www.math.ucla.edu/~archristian/teaching/32a-f16/implicit-function-theorem.pdf

4.2.3 Practice Problems with Solutions

1. Practice problems: https://tutorial.math.lamar.edu/problems/calciii/calciii.aspx

---

5 Optimization

Many economic models are based on the idea that an individual decision maker makes an *optimal choice* from some set of alternatives. To formalize this idea, we interpret optimal choice as maximizing or minimizing the value of some function. For example, a firm is assumed to minimize costs of producing each level of output and maximize profit, a consumer maximizes their utility, a policy maker maximizes welfare or GDP, etc. Thus, the mathematics of optimization are of central importance in economics. Luckily, we have already developed the machinery that will be used to optimize functions.

Given some function $f(x)$, we optimize it by finding a value of x at which it takes on a maximum or minimum value, i.e. an *extreme value*. If we are optimizing over the entire real line $\mathbb{R}$, we say that we are *unconstrained*, whereas if we focus on some strict subset $[a, b] \subset \mathbb{R}$, we say that we are conducting *constrained* optimization.

It is possible that a function does not have a minimum or maximum value. An example is a linear function $y = a + bx$ for $a, b > 0$. Similarly, a parabola $y = x^2$ has a minimum at $x = 0$ but no maximum.

Definition 5.1 (Global and Local Maximum)**.** *At a global maximum x^*,*

$$f(x^*) \geq f(x), \forall x \in \mathbb{R}$$

Whereas at a local maximum x^*,

$$f(x^*) \geq f(x), \forall x \in B_\epsilon(x) \text{ for } \epsilon > 0$$

While we are generally interested in finding a global maximum, the methods we have for optimizing functions will only guarantee that we are able to find local maximum (if they exist). In practice, many economic problems have built in assumptions to ensure that there is only one local maximum, in which case it is also the global maximum.

5.1 Finding Extreme Values of f

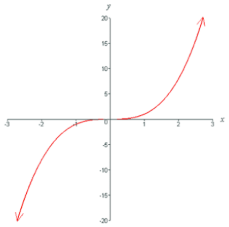
5.1.1 First Order Conditions: Necessary

Using the first derivative of a function, we can determine some necessary conditions for the existence of a local extreme point.

Theorem 5.1 (First Order Conditions)**.** *Let f a function defined on (a, b). If x is a maximum or minimum point for f on (a, b), and f is differentiable at x, then $f'(x) = 0$. (Converse is not true)*

The intuition here is that if the derivative is not 0, then on some small interval containing x, $B_\epsilon(x)$, the function will be strictly increasing or decreasing, contradicting that x can be an extreme point of f.

To see however that first order conditions are not sufficient to guarantee the existence of an extreme value, consider the function $f(x) = x^3$, the cubic function. We have that $f'(x)|_{x=0} = 3x^2 = 0$, however 0 is an *inflection point* for f, not a local minima or maxima.



5.1.2 Second Order Conditions: Sufficient

So far we have only used information about the first derivative of f. By bringing in information about higher order derivatives, namely the second derivative, we can distinguish whether a given point is or is not an extreme value of the function.

Theorem 5.2 (Second Order Conditions)**.** *Suppose $f'(a) = 0$. If $f''(a) > 0$, then f has a local minimum at a, and if $f''(a) < 0$ then f has a local maximum at a.*

5.1.3 Convexity and Concavity

Definition 5.2 (Convex Function)**.** *A function f is convex on an interval, if for all a and b in the interval, the line segment joining $(a, f(a))$ and $(b, f(b))$ lies above the graph of f. Equivalently, if for all x such that $a < x < b$ we have $\frac{f(x) - f(a)}{x - a} < \frac{f(b) - f(a)}{b - a}$*

Theorem 5.3 (Convex Functions)**.** *If f is differentiable and f' is increasing, then f is convex.*

In economic applications, convex functions are used to ensure that local minima are global. Any local minimum of a convex function is also a global minimum. A strictly convex function will have at most one global minimum.

5.1.4 Multivariable Optimization*

For multivariable functions, the first and second order conditions have analogs in terms of partial derivatives.

Definition 5.3 (First Order Conditions for $\mathbb{R}^n$)**.** *A stationary value of the function f over the domain $\mathbb{R}^n$ occurs at a point $x \in \mathbb{R}^n$ at which the n equalities below hold simultaneously:*

$$f_1(x_1, ..., x_n) = 0$$

$$...$$

$$f_n(x_1, ..., x_n) = 0$$

As before, this is a necessary but not sufficient condition to find a local maxima or minima.

Similarly, we can think of second order conditions for multivariable functions by requiring that small changes in the x vector along any dimension from this point must reduce the value of f if x is to be a maximum.

Definition 5.4 (Second Order Conditions for $\mathbb{R}^n$)**.** *It is sufficient for x to yield a local maximum of the continuous function f that $f_i(x) = 0, \forall i 1, ..., n$, and the quadratic form*

$$d^2y(x) = \sum_i \sum_j f_{ij}(x) dx_i d_x j < 0, \quad i, j = 1, ..., n$$

That is, d^2y is negative definite.

5.2 Constrained Optimization

So far we have been dealing with unconstrained optimization, in which the solution to the optimization can lie anywhere on the real line $\mathbb{R}$. Often in economics however, this is unacceptably general. A key idea of economics is the idea of *scarcity of resources*– the gap between limited resources and theoretically limitless wants. The optimization of interest for real world applications is that of allocating finite resources efficiently, i.e. in a way the maximizes utility or welfare for the agents involved. In order to formalize this mathematically, we consider optimization over a constrained set, such as an interval in $\mathbb{R}$ for functions of one variable.

5.2.1 Optimization Over an Interval

When we optimize a function f subject to the constraint that the value lies within some closed interval $[a, b]$, it is no longer a necessary condition that the derivative $f'(x) = 0$ at the extreme value. This is because the extreme value can additionally occur at the endpoints of the interval, a or b, rather than an *interior solution*. In fact, there are now three possibilities for a local maximum (or minimum) x:

1. $x = a$. In this case, we must have $f'(x) \leq 0$. *Why?*
2. $a < x < b$. In this case we must have $f'(x) = 0$
3. $x = b$. In this case, we must have $f'(x) \geq 0$.

5.2.2 Lagrangian Method

Suppose that we wish to maximize a function $f(x_1, x_2)$ for strictly concave and continuous f. And further, that we impose the constraint $g(x_1, x_2) = 0$ for continuous g. This means that we are only allowed to consider as possible solutions to the problem, pairs of x_1, x_2 that satisfy the equation $g = 0$.

We proceed by introducing a new variable λ, the **Lagrange Multiplier**, and by forming the **Lagrangian**:

$$L(x_1, x_2, \lambda) = f(x_1, x_2) + \lambda g(x_1, x_2)$$

We then carry out the unconstrained maximization of L with respect to x_1, x_2, λ to get the following first

order conditions:

$$\begin{aligned}\frac{\partial L}{\partial x_1} &= f_1(x_1, x_2) + \lambda g_1(x_1, x_2) = 0\\ \frac{\partial L}{\partial x_2} &= f_2(x_1, x_2) + \lambda g_2(x_1, x_2) = 0\\ \frac{\partial L}{\partial \lambda} &= g(x_1, x_2) = 0\end{aligned}$$

Thus we have three equations to solve for the three unknowns, x_1, x_2, λ. The Lagrange multiplier procedure is a way of defining an unconstrained optimization problem, which delivers in a routine way the tangency conditions for an optimal solution. Let us see an example:

Example Lagrangian Problem: Solve the constrained maximization problem:

$$\max y = x_1^{0.25} x_2^{0.75} \quad \text{such that } 100 - 2x_1 - 4x_2 = 0$$

We write out the Lagrangian function for this problem and take the first-order conditions:

$$\begin{aligned}L &= x_1^{0.25} x_2^{0.75} + \lambda(100 - 2x_1 - 4x_2)\\ \frac{\partial L}{\partial x_1} &: 0.25 x_1^{-0.75} x_2^{0.75} - 2\lambda = 0\\ \frac{\partial L}{\partial x_2} &: 0.75 x_1^{0.25} x_2^{-0.25} - 4\lambda = 0\\ \frac{\partial L}{\partial \lambda} &: 100 - 2x_1 - 4x_2 = 0\end{aligned}$$

Let us solve this system of equations via substitution:

$$\begin{aligned}\frac{1}{8}\left(\frac{x_2}{x_1}\right)^{0.75} = \lambda = \frac{3}{16}\left(\frac{x_1}{x_2}\right)^{0.25}&\\ \frac{x_2}{x_1} = \frac{3}{2}&\\ x_2 = \frac{3}{2}x_1&\\ 100 - 2x_1 - 6x_1 = 0 \implies \frac{100}{8} = x_1 \implies x_1 = \frac{25}{2} \text{ and } x_2 = \frac{75}{4}&\end{aligned}$$

5.2.3 Interpretation of λ

We apparently introduced λ just as a placeholder to help generate the conditions which give us the solution to the original constrained optimization problem. It turns out however, that λ has a very important and interesting economic interpretation in constrained optimization problems. The value of the Lagrange multiplier λ at the optimal solution tells us the effect of the optimized value of the function f of a small relaxation of the constraint, i.e. the *shadow price* of the constraint.

5.3 Applications to Finance and Economics

5.3.1 Derivation of Mean-Variance Portfolio

A key topic in financial economics is that of an investor's portfolio choice across assets. An investor may wish to maximize the average return of their portfolio while minimizing the corresponding risk, or volatility of their returns. We can combine the tools that we have learned so far to solve the classic Markowitz (1952) model of choosing N risky assets to minimize variance for a given level of mean return.

We define $R \in \mathbb{R}^n$ as the vector of mean returns on the n risky assets, Σ as the variance-covariance matrix of returns, ω as the vector of portfolio weights, and 1 as a vector of ones. The problem we wish to solve is the following optimization problem:

$$\min_{\omega} \frac{1}{2}\omega'\Sigma\omega \quad \text{such that } R'\omega = A \text{ and } 1'\omega = 1$$

We set up the Lagrangian:

$$\begin{aligned} L(\omega, \lambda_1, \lambda_2) &= \frac{1}{2}\omega'\Sigma\omega + \lambda_1(A - R'\omega) + \lambda_2(1 - 1'\omega) \\ \frac{\partial L}{\partial \omega} &= \Sigma\omega - \lambda_1 R - \lambda_2 1 = 0 \\ \frac{\partial L}{\partial \lambda_1} &= A - R'\omega \\ \frac{\partial L}{\partial \lambda_2} &= 1 - 1'\omega \\ \implies \omega &= \lambda_1\Sigma^{-1}R + \lambda_2\Sigma^{-1}1 \end{aligned}$$

We can then solve for the Lagrange multipliers using the two constraints to solve for two unknowns:

$$\begin{aligned} A = R'\omega &= \lambda_1 R'\Sigma^{-1}R + \lambda_2 R'\Sigma^{-1}1 \\ 1 = 1'\omega &= \lambda_1 1'\Sigma^{-1}R + \lambda_2 1'\Sigma^{-1}1 \end{aligned}$$

This gives:

$$\begin{aligned} \lambda_1 &= \frac{(1'\Sigma^{-1}1)A - R'\Sigma^{-1}1}{(R'\Sigma^{-1}R)(1'\Sigma^{-1}1) - (R'\Sigma^{-1}1)^2} \\ \lambda_2 &= \frac{(R'\Sigma^{-1}R - (R'\Sigma^{-1}1)A}{(R'\Sigma^{-1}R)(1'\Sigma^{-1}1) - (R'\Sigma^{-1}1)^2} \end{aligned}$$

5.3.2 Principal Component Analysis

Principal Component Analysis (PCA) is one of a family of techniques for taking high-dimensional data, and using the dependencies between the variables to represent it in a more tractable, lower dimensional form, without loosing too much information, i.e. to reduce the dimension of the data.

We begin with p-dimensional vectors with mean 0, and want to summarize them by projecting down into a q-dimensional shape, for $p > q$. Our summary will be the projection of the original vectors on to q

directions, the **principal components**, which span the subspace. There are multiple equivalent ways to find these principal components:

1. We can find the projections that maximize the variance.
2. We can find the projection with the smallest mean-squared distance between the original vectors and their projections onto the principal components.

Via maximizing variance: If we stack our n data vectors into an $n \times p$ matrix x, then the projections are given by $xw \in \mathbb{R}^{n \times 1}$. The variance is:

$$\sigma_\omega^2 = \frac{1}{n}(xw)^T(xw) = \frac{1}{n}w^T x^T x w = w^T \frac{x^T x}{n} w = w^T v w$$

Thus we want to choose a unit vector w to maximize σ_ω^2. Thus our optimization problem is the following:

$$\max_w w^T v w \quad \text{such that} w^T w = 1$$

So we can set up our Lagrangian:

$$\begin{aligned} L(w, \lambda) &= w^T v w + \lambda(1 - w^T w) \\ \frac{\partial L}{\partial w} &: 2vw - 2\lambda w = 0 \implies vw = \lambda w \\ \frac{\partial L}{\partial \lambda} &: w^T w = 1 \end{aligned}$$

Thus, the desired vector w is an **eigenvector** of the covariance matrix v, and the maximizing vector will be the one associated with the largest **eigenvalue** λ. Since $v \in \mathbb{R}^{p \times p}$, we know that it will have p different eigenvectors. Further, since $v \geq 0$ as a covariance matrix, the eigenvalues of v must all be weakly positive. The eigenvectors of v are the **principal components** of the data.

5.4 Resources

5.4.1 Textbooks

1. *Mathematics for Economics by Hoy et al.* chapters 5.5; 6.1—6.3; 11.4; 12.1—12.2; 13.1—13.2; 14.3; 15.1—15.2.

5.4.2 Web Resources

1. PCA: http://www.stat.cmu.edu/~cshalizi/uADA/12/lectures/ch18.pdf

5.4.3 Practice Problems with Solutions

1. Economics: https://www.math24.net/optimization-problems-economics

6 Probability and Statistics

6.1 Probabilities

> The most important questions of life are, for the most part, really only problems of probability.
> – Pierre-Simon Laplace

We have already developed the tools that we need to formalize the notion of probability. Let us consider the following triplet of mathematical objects:

Definition 6.1 (Probability Space)**.** *A probability space $(\Omega, \mathcal{A}, \mathbb{P})$ consists of three elements:*

1. *A **sample space** Ω, which is the set of all possible outcomes.*
2. *An **event space** $\mathcal{A}$, which a set of events i.e. sets of outcomes in the sample space.*
3. *A **probability measure** (function) $\mathbb{P}$, which assigns each event in the event space $\mathcal{A}$ a probability, $\mathbb{P} : \mathcal{A} \to [0, 1]$.*

Theorem 6.1 (Properties of $\mathbb{P}$)**.** *Probability measure $\mathbb{P}$ must satisfy the following properties:*

1. *Positivity: $\mathbb{P}(A) \geq 0, \forall A \in \mathcal{A}$*
2. *Normalization: $\mathbb{P}(\Omega) = 1$*
3. *Additivity: If $A \cap B = \emptyset$, then $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$ and in general, $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$*
4. *Monotonicity: If $A \subseteq B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$*

Definition 6.2 (σ-Algebra)**.** *If the nonempty collection $\mathcal{A}$ of subsets of $\Omega \neq \emptyset$ contains all complements and unions of its sets (i.e. is "closed" under complements and countable unions), then it is a σ-algebra of Ω.*

Definition 6.3 (Measurable Space)**.** *The pair $(\Omega, \mathcal{A})$, where $\mathcal{A}$ is a σ-algebra of subsets of Ω is called a measurable space. The subsets of $\mathcal{A}$ are called the events of Ω.*

6.1.1 Random Variables

Along with a probability measure, a central concept in probability and statistics is that of a *random variable*. Informally, this is a function whose values depend on the outcomes of random phenomena.

Definition 6.4 (Random Variable)**.** *Let $(\Omega, \mathcal{A})$ and $(\mathbb{R}, \mathcal{B})$ be two measurable spaces. The map $X : \Omega \to \mathbb{R}$ is called real-valued random variable, if for all $B \in \mathcal{B}$, $X^{-1}(A)$ is **measurable** for all $B \in \mathcal{B}$, i.e.:*

$$X^{-1}(B) = \{\omega \in \Omega | X(\omega) \in B\} \in \mathcal{A}$$

Thus by definition a random variable takes events from one sample space into well-defined events in another space.

If a random variable $X : \Omega \to \mathbb{R}$ defined on the probability space $(\Omega, \mathcal{A}, \mathbb{P})$ is given, we can ask questions like, "How likely is it that the value of X is equal to 2?". This is the same as the probability of the event $\{\omega | X(\omega) = 2\} \equiv \mathbb{P}(X = 2) = \mathbb{P}_X(2)$. In other words, since $\mathbb{P}$ specifies the probability of every subset of Ω, it also induces probabilities on events expressed in terms of X.

Recording all of these probabilities of output ranges of a real-valued random variable X yields the **probability distribution** of X. The probability distribution "forgets" about the particular probability space used to define X and only records the probabilities of various values of X. Such a probability distribution can always be captured by its **cumulative distribution function**.

Definition 6.5 (Cumulative Distribution Function)**.** *The cumulative distribution function (c.d.f) of a real-valued random variable $X : \Omega \to \mathbb{R}$ is defined as the real-valued function $F : \mathbb{R} \to [0, 1]$ given by:*

$$F_X(x) = \mathbb{P}\{\omega \in \Omega | X(\omega) \leq x\} = \mathbb{P}\{X^{-1}(-\infty, x]\}$$

Thus it is simply its induced probability measure $\mathbb{P}_X$ evaluated at sets of the form $(-\infty, x]$.

Theorem 6.2 (Properties of the CDF)**.** *If F_x is the CDF of a random variable $X : \Omega \to \mathbb{R}$, then*

1. *F_X is non-decreasing.*
2. *$\lim_{x \uparrow \infty} F_X(x) = 1$*
3. *$\lim_{x \downarrow -\infty} F_X(x) = 0$*
4. *$\lim_{h \to 0^+} F_X(x + h) = F_X(x)$ i.e. the CDF is right-continuous.*

Definition 6.6 (Probability Density Function)**.** *For continuous random variables, assuming that F_X is differentiable, by the fundamental theorem of calculus we have that:*

$$F_X(x) = \int_{-\infty}^{x} p_x(x) dx$$

The function $p_x(x)$ is called the probability density function (PDF) of X. Note that $P_X(x) = \mathbb{P}(X = x)$.

Definition 6.7 (Probability Mass Function)**.** *Scalar random variables that take values in a discrete set such as $\{1, 2, ...\}$ do not have a density function. Instead they have a probability mass function (PMF), with the same interpretation as the density.*

$$p_X(x) = \sum_i \mathbb{P}(X = x_i) 1\{x = x_i\}$$

The function $p_X(x)$ is called the probability mass function (PMF) of X. Note that $P_X(x) = \mathbb{P}(X = x)$.

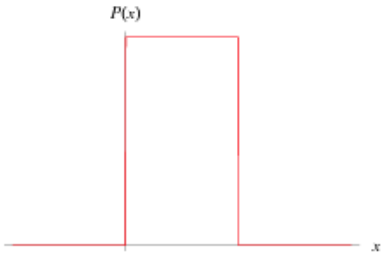
6.1.2 Examples of CDFs for $X : \Omega \to \mathbb{R}$

Definition 6.8 (Bernoulli Distribution)**.** *The random variable X is said to have a Bernoulli distribution with parameter $p \in [0,1]$ if the CDF is given by*

$$F_X(x) = \begin{cases} 0 & x \leq 0 \\ 1-p & x \in [0,1) \\ 1 & x \geq 1 \end{cases}$$

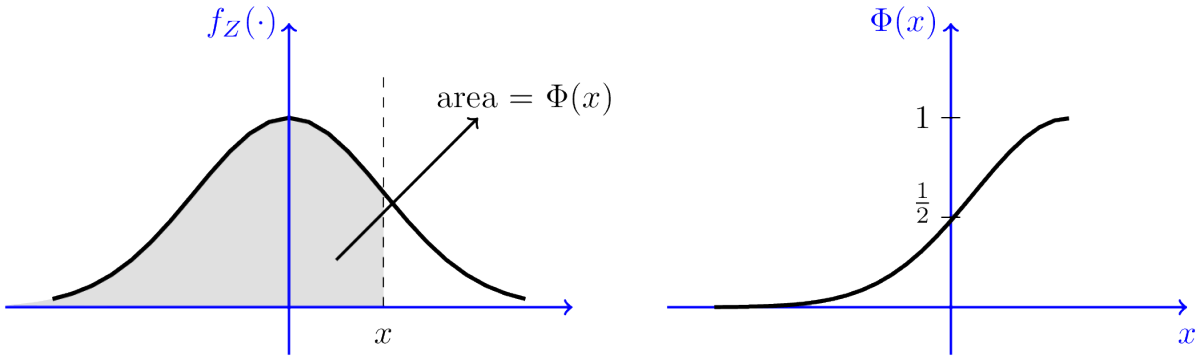
Definition 6.9 (Uniform Distribution)**.** *The random variable X is said to have a uniform distribution in $[a,b]$ if the CDF is given by*

$$F_X(x) = \begin{cases} 0 & x \leq a \\ \frac{x-a}{b-a} & x \in [a,b) \\ 1 & x \geq b \end{cases}$$



Definition 6.10 (Gaussian (Normal) Distribution)**.** *The random variable X is said to have a Normal distribution with parameters (μ, σ^2) if the CDF is given by*

$$\Phi(x) = F_X(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(u-\mu)^2\right) du$$



Note that a Bernoulli random variable is discrete, whereas Uniform and Normal random variables have continuous support.

6.1.3 Moments of $X : \Omega \to \mathbb{R}$

Let us introduce the definition of expectation for discrete and (absolutely) continuous random variables.

Definition 6.11 (Expectation of continuous random variable)**.** *Let $X : \Omega \to \mathbb{R}$ an absolutely continuous random variable with CDF F and PDF $f(x)$. Let $g : X \to \mathbb{R}$ a real-valued function. The expectation of*

$g(X)$ under F is defined as:

$$\mathbb{E}_F[g(X)] = \int_{\mathbb{R}} g(x)f(x)dx$$

Definition 6.12 (Expectation of discrete random variable)**.** *Let $\{x_1, x_2, ...\}$ the support of discrete random variable $X : \Omega \to \mathbb{R}$. Let $g : X \to \mathbb{R}$ a real-valued function. The expectation of $g(X)$ under PMF p is defined as:*

$$\mathbb{E}_F[g(X)] = \sum_{x_i \in Supp} g(x_i)p(x_i)$$

Some important moments are the following:

Definition 6.13 (Moments)**.**

1. *Mean of a RV: $\mathbb{E}_F[X] = \mu$ is called the **mean** or **first moment** of X.*
2. *Variance of RV: $\mathbb{E}_F[(X - \mu)^2] = \sigma^2$ is called the **variance** or **second centered moment** of X.*
3. *k-th moment of a RV: $\mathbb{E}_F[X^k] = \mu$ is called the **k-th uncentered moment** of X.*
4. *Probability: $\mathbb{P}(X \in \mathcal{A}) = \mathbb{E}[1\{X \in \mathcal{A}\}]$*

Important properties of moments:

Theorem 6.3 (Properties of Moments)**.**

1. *Jensen's Inequality: For any convex function g: $\mathbb{E}_F[g(X)] \geq g(\mathbb{E}[X])$*
2. *Markov's Inequality: For any random variable X and $c \geq 0$, $c\mathbb{P}_X[|X| > c] \leq \mathbb{E}_F[|X|]$*

Definition 6.14 (Moment Generating Function*)**.** *The random variable X is said to have a moment generating function (MGF) $m_x(t)$ if*

$$m_x(t) \equiv \mathbb{E}_F[exp(tX)] \leq \infty \quad \text{for all } t \in (-\epsilon, \epsilon) \text{ for some } \epsilon > 0$$

Two random variables with the same moment generating function have the same distribution. As the name implies, the MGF can be used to compute a distribution's moments: the nth moment about 0 is the nth derivative of the MGF, evaluated at 0.

6.1.4 Conditional Probability and Expectations

If we could only analyze one random variable at a time, the questions that we would be able to answer would not be very interesting. Most of the empirical questions in economics boil down to *what is the causal effect of X on Y?*. So, we need a way to describe the relationship between at least two random variables.

Until now, we have been considering probability density and mass functions assuming that we have *no* information about the random variable of interest. We might be able to do better if we observe something that is informative about that variable, and incorporate that information into our expectation.

Definition 6.15 (Conditional Probability)**.** *Consider two events $A, B \in \mathcal{A}$. The conditional probability that A occurs given that B occurs is:*

$$P(A|B) \equiv \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

Theorem 6.4 (Bayes Rule)**.** *Recall the conditional probability $\mathbb{P}(A|B)$. Bayes Rule is a formula for the "inverse" conditional probability, $\mathbb{P}(B|A)$:*

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A)}$$

Definition 6.16 (Independence)**.** *Two events A and B are said to be independent if $\mathbb{P}(A|B) = \mathbb{P}(A)$. In other words, B provides no information about whether A has occurred.*

6.1.5 Multivariate Distributions

Definition 6.17 (Multivariate or Joint CDF)**.** *To fully understand the relationship between two random variables X, Y, we need to look at something that fully characterizes their joint distribution, such as their joint CDF.*

$$F_{X,Y}(x, y) = \boldsymbol{P}\left[(X \leq x) \cap (Y \leq y)\right]$$

From this joint distribution, we can recover the marginal univariate PDF of X for example by integrating out Y:

$$f_x(x) = \int_{S_Y} f_{X,Y}(x, y)dy$$

Theorem 6.5 (Bayes Rule For Joint PDF)**.** *The Bayes Rule applied to a joint pdf becomes:*

$$f_{X|Y}(x; y) = \frac{f_{Y|X}(y; x)f_y(y)}{f_X(x)}$$

6.2 Statistics

> By a small sample, we may judge the whole piece.
> – Miguel De Cervantes, Don Quixote

6.2.1 Basics of Estimators

The frequentist approach to estimation starts with the premise that there is a *population parameter* θ that determines how we observe data. We observe a *sample* of this data, $\{x_i\}_{i=1}^N = X$ and use this sample to produce an *estimate* of θ, the property of the population that we would like to know about. Our prime objective is to estimate properties of the population, using a sample and an estimator.

Definition 6.18 (Estimator)**.** *For a random variable X, an estimator $\hat{\theta} : X \to \mathbb{R}^n$ is a function that maps the sample space to a set of sample estimates.*

- *The estimator itself is a random variable.*
- *The estimate for a given realization of the data is a fixed number.*

There are various types of estimators. A **point estimator** outputs a single number that can be regarded as the most plausible value of θ. An **interval estimator** outputs a range of numbers, called a confidence interval. This range can be regarded as likely containing the true value of θ.

An important example of a point estimator is the sample mean, which is useful to arrive at point estimates of a given parameter.

Definition 6.19 (Sample Mean)**.** *The sample mean is the average of the values of a random variable in a sample, used as an estimator for the population mean of a random variable, μ.*

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

It turns out that this estimator is a good one for the population mean, as it gets very close to the population mean as the sample size grows.

Theorem 6.6 (Law of Large Numbers)**.** *The sample mean $\bar{x}$ converges in probability towards μ as $n \to \infty$.*

$$\lim_{n\to\infty} \mathbb{P}(|\bar{x}_n - \mu| > \epsilon) = 0 \text{ for all } \epsilon > 0$$

Given a point estimator $\hat{\theta}$, we can define the standard error of the estimator by recalling that the estimator is itself a random variable:

Definition 6.20 (Standard Error)**.** *The standard error of an estimator $\hat{\theta}$ is its standard deviation:*

$$SE(\hat{\theta}) = \sqrt{Var(\hat{\theta})}$$

In addition to a point estimate, we may want to have a confidence interval around a given estimate of a parameter. For this, the following theorem is very useful:

Theorem 6.7 (Central Limit Theorem)**.** *If $X_1, ...X_n$ are n random samples drawn from a population with overall mean μ and finite variance σ^2, and if $\bar{X}$ is the sample mean, then the limiting form of the following distribution is $N(0,1)$:*

$$Z = \lim_{n\to\infty} \sqrt{n} \left(\frac{\bar{X}_n - \mu}{\sigma} \right) \sim N(0,1)$$

This theorem is very useful in economic regression analysis as the assumption of normally distributed errors simplifies much analysis and can be justified by the central limit theorem– regarding the error as

the sum of many independent errors. A similar logic can be used to explain why so many variables in real-world data sets seem to be normally distributed (although not always in financial data– see the Time Series Analysis section below).

6.2.2 Inference and Hypothesis Tests

Now we can give an important example of an interval estimator, a **confidence interval**.

Definition 6.21 (Confidence Interval)**.** *A confidence interval $C(X)$ for a parameter θ is a set of possible values which contains θ with some specified probability.*

- *A higher confidence interval generates a wider (less precise) confidence interval.*

6.2.3 CEF And Ordinary Least Squares

We are interested in explaining economic relationships, i.e. in explaining "systematic randomness". Thus, the object that we would like to estimate is the population relationship between two random variables X and Y. This relationship is conveniently summarized by the *Conditional Expectation Function.*

Definition 6.22 (Conditional Expectation)**.** *The conditional expectation for a "dependent" variable Y_i, given a set of covariates X_i, is the expectation, or population average, of Y_i, with X_i held fixed.*

$$E[Y_i|X_i = x]$$

The conditional expectation function is useful because it is the "best" predictor of Y_i given X_i, in the sense that it solves a minimum mean squared error (MMSE) prediction problem.

Theorem 6.8 (The CEF Prediction Property)**.**

$$E[Y_i|X_i] = \arg\min_{m(X_i)} \mathbb{E}[(Y_i - m(X_i))^2]$$

There is a tight link between the regression function and the CEF. The regression function is the "best" linear approximation to the CEF, the "best" linear predictor of Y_i given X_i in the sense that it solves a MMSE prediction problem, and is equivalent to the CEF if the CEF is linear. *Therefore, one should be interested in regression parameters if one is interested in the CEF.*

To arrive at the population regression function, we are seeking a vector of population regression coefficients, defined as a solution to a population least squares problem. Let the $K \times 1$ regression coefficient vector be defined by solving:

$$\beta = \arg\min_{b} \mathbb{E}[(Y_i - X_i'b))^2]$$

To optimize this function, we find the first order condition:

$$\mathbb{E}[X_i(Y_i - X_i'b)] = 0$$

And thus the solution can be written:

$$\beta = \mathbb{E}[X_i X_i']^{-1}\mathbb{E}[X_i Y_i]$$

And so the slope coefficient for the kth regressor x_k is:

$$\beta_k = \frac{Cov(Y_i, \tilde{x}_{ik})}{Var(\tilde{x}_{ik})}$$

Where $\tilde{x}_{ik}$ is the residual of a regression of $x_{i,k}$ on all the other covariates, $x_{i,-k}$, i.e. $x_{i,k}$ after *partialling out* the effects of other regressors.

6.2.4 Standard Errors Under Different Assumptions About ϵ

We have some parameter which we estimate:

$$\beta_k = \frac{Cov(Y_i, \tilde{x}_{ik})}{Var(\tilde{x}_{ik})}$$

As we have discussed, this is an estimate of the true parameter β, and we are interested in understanding the distribution of estimate if we repeatedly drew random samples from the same population and estimated β. Thus, we want the variance of the sample average effect.

$$\begin{aligned} var(\hat{X}) &= var\left(\frac{1}{n}\sum X\right) = \frac{1}{n^2}\sum var(X) = \frac{1}{n}var(X) \\ se(\hat{X}) &= \sqrt{var(\hat{X})} = \frac{\sigma_X}{\sqrt{n}} \end{aligned}$$

So in order to calculate the standard error of $\hat{\beta}$, first we would like to get the variance of $\hat{\beta}$:

$$\begin{aligned} var(\hat{\beta}) &= var\left[\frac{cov(X,Y)}{var(Y)}\right] \\ &= var\left[\frac{\sum(X_i-\mu_X)(Y_i-\mu_Y)}{\sum(X_i-\mu_X)^2}\right] \\ &= var\left[\frac{(\sum(X_i-\mu_X)\epsilon_i}{\sum(X_i-\mu_X)^2}\right] \\ &= \sigma_\epsilon^2\left[\frac{\sum(X_i-\mu_X)}{\sum(X_i-\mu_X)^2}\right] \qquad (*) \\ &= \frac{\sigma_\epsilon^2}{\sigma_X^2} \\ &\implies SE(\hat{\beta}) = \frac{\sigma_\epsilon}{\sqrt{n}*\sigma_X} \end{aligned}$$

Here in the starred line, we have assumed *homoskedasticity*, i.e. that there exists some constant σ_ϵ^2, i.e. that the variance of ϵ does not vary with X. In contrast, we can instead derive *White* standard errors, or robust standard errors that do not require the homoskedasticity assumption.

6.3 Time Series Analysis*

Time series data, such as stock market prices and returns, is very important in financial analysis. This data violates many of the common assumptions used in statistical analysis and require different techniques. For details, take the PhD Timeseries Analysis course offered at CBS[2]. For one salient example, see below on the non-stationarity of prices.

6.3.1 How Variance of Returns Scales with Time

A basic property we would like to have of our data is that the data properties should not be changing too much over time, or in ways that imply there are fundamental differences between the data collected yesterday and today. *Stationarity* intuitively refers to something that is not changing much over time. We usually apply this concept to moments in the data, such as means and variances.

Let us consider an example to see whether different types of financial data satisfy this property. Given a process for returns with mean μ and standard deviation σ, consider the log value of a portfolio p_t that earns a return each period and reinvests all wealth.

$$p_t = p_{t-1} + r_t = p_{t-1} + \mu + \sigma\epsilon_t$$

This value process is said to follow a *Random Walk with Drift*, i.e. is a process with unforecastable increments, except for a constant drift term μ. Simulated data from this model approximates the properties of stock return data. This is the original Efficient Markets model of Gene Fama (1970): if markets are efficient, returns are not forecastable apart from the constant risk premium component.

In this model, prices are *non-stationary* whereas returns are *stationary*, in the sense of the stability of their covariances over time. Consider the variance of p_t:

$$\lim_{t\to\infty} Var[p_t|p_0] = \lim_{t\to\infty} t\sigma^2 = \infty$$

Instead, consider the return process $r_t = \mu + \sigma\epsilon_t$:

$$\lim_{t\to\infty} Var[r_t] = \lim_{t\to\infty} Var[\mu + \sigma\epsilon_t] = \sigma^2, \forall t$$

6.4 Resources

6.4.1 Textbooks

1. *Mostly Harmless Metrics by Angrist & Pischke* is a popular gentle exposition of applied microeconometrics.

6.4.2 Web Resources

1. Probability lecture notes: https://www.stat.berkeley.edu/~aldous/134/gravner.pdf

[2]The lecture notes in this section are taken from the notes in Michael Johannes' course material.

2. Statistics lecture notes: https://people.richland.edu/james/lecture/m170/

3. Econometrics lecture notes: https://people.stern.nyu.edu/wgreene/Econometrics/Notes.htm

6.4.3 Practice Problems with Solutions

1. Khan Academy: https://www.khanacademy.org/math/statistics-probability/probability-library