ROB 501
Mathematics for Robotics

M | ROBOTICS

Jessy Grizzle
Director, Michigan Robotics

Cover design by Dan Newman, Head of Communications, Michigan Robotics

JERRY W. AND CAROL L. LEVIN PROFESSOR OF ENGINEERING
ELMER G. GILBERT DISTINGUISHED UNIVERSITY PROFESSOR
PROFESSOR OF EECS AND ME

Composed as lecture notes and piloted between August 2014 and December 2018 for use in ROB 501, Mathematics for Robotics.

First release, August 31, 2022.

Contents

Preface

This collection of course notes is dedicated to all the students who have taken ROB 501 Mathematics for Robotics at the University of Michigan. The first class of students in Fall 2014 took on one of two tasks: typesetting in Latex a lecture or creating a numerically oriented HW problem in MATLAB. For their dedication, enthusiasm, and patience with the pilot version of the course, I tip my hat to Pedro Fernando, Jimmy, Vittorio, Kevin, Yuxiao, Katelyn, Omar, Ross, Jakob, Xianan, Mohammad, Jeffrey, Bo, Kurt, Josh, Xiangyu, Sulbin, Connie, Meghan, Su-Yang, Katie, Il Sun, Mia, Yong, Yunxiang, Hiroshi, Yevgeniy, Ming-Yuan, Pengcheng, and Zhiyuan. I have taken your work and expanded it into these notes. Thank you!

Jessy Grizzle
Winter Term, 2022

Philosophy of the Course

The Robotics Graduate Program was designed in 2013 and launched in 2014. The vision was to take in students from all engineering backgrounds, plus many STEM backgrounds, such as Mathematics and Physics, and prepare M.S. and Ph.D. students for fruitful careers in Industry and Academia. We knew the incoming students would have very heterogeneous preparation in Mathematics, Programming, and Hardware. To form them into a more homogeneous cohort, two first-semester courses were created, ROB 501 Mathematics for Robotics and ROB 550 Robotics Laboratory[1]. After taking these two courses, the students would have adequate preparation to take courses in Sensing, Acting and Reasoning, as laid out here https://robotics.umich.edu/academic-program/courses/. The Robotics Graduate Program has worked remarkably well due to the enthusiasm of the students, staff, and faculty.

I accepted the charge of creating the mathematical fundamentals course, ROB 501. I interviewed the Robotics faculty in Spring 2014 to find out key topics they wanted covered. Over Summer 2014, I pared the list down and organized it into a coherent set of topics that would mostly meet the needs as identified by the faculty. To the set of topics enumerated by my colleagues, I added one additional goal, to break the students' reliance on example-based (e.g., physics-based) reasoning, and instill in them the ability to conduct **abstract reasoning**. The ability to abstract problems from a physical instantiation to a mathematical formulation provides clarity of thinking as well as being the first step in making problems amenable to algorithmic solutions. Once a student has written a problem down mathematically, they are able to search the literature for potential solutions, or understand that they have stumbled upon a new question that requires new investigations.

To be clear, breaking a student's reliance on physics-based reasoning really means augmenting their reasoning ability, it means giving them confidence to deal with abstract formulations of problems. The key word here is **confidence**. ROB 501 does this by doing proof after proof in every lecture. By practice and repetition, the student learns to read (relatively) complex mathematical statements, negate them, and otherwise appreciate common mathematical symbols as a precise and effective form of shorthand. To make the repetition of definitions and proofs tenable, they first practice them in the friendly environment of linear algebra. We take advantage of the fact that the vast majority of the students taking the course will be comfortable with matrix-vector arithmetic and MATLAB. While Chapter 1 introduces mathematical notation and proof techniques, these topics are not really learned until they are exercised in Chapter 2 with an abstract treatment of linear algebra, with everything proved, either in class or HW. Chapter 3 adds more abstraction, inner product spaces, but starts to provide a pay off, namely deterministic least squares problems. The students start to see how concrete optimization problems can be solved easily and precisely with the normal equations, and through HW sets, that "practical" problems involving functions can be solved with no essential changes to the methods. In addition, as a setup for the Kalman Filter, we derive a recursive solution to the standard (batch) least squares problem, giving us RLS.

Chapter 4 is meant as a mental break. While students are completing HW sets and/or exams on the previous material, we use Gram Schmidt to develop the QR Factorization, and we use orthogonal matrices and eigenvectors to derive the SVD and pause to illustrate its practical importance. While typing up these notes, I added in the LU Factorization, in part because I used it as a building block in the undergrad course, ROB 101, Computational Linear Algebra.

Chapter 5 sets the stage for minimum variance estimation, or non-deterministic least squares problems. We build slowly because if there is one topic with which a typical first-semester engineering graduate student struggles, it's probability. The students themselves are not to blame for this, nor are their undergraduate instructors. The topic is fundamentally very challenging. As soon we leave the confines of discrete random variables (with a finite number of possible outcomes) and enter the realm of continuous random variables, we have a choice: go down the measure theory rabbit hole or fake it. To be sure, we fake it. To clear our conscience, we are upfront about it and explain the restrictions we make to allow probabilities of events to be computed via an integral, that is, via a density and Riemann integration. This allows random vectors to be defined and the important quantities, the mean of a random vector, its covariance, and its variance. Armed with these concepts, we revisit a standard least squares problem and add a measurement noise

[1]While ROB 550 was initially charged with preparing students in the areas of embedded programming, inverse kinematics, camera calibration, path planning, encoders, motor drives, and more, later, in 2019, ROB 502 Programming for Robotics was created to relive some of the burden from ROB 550.

model, leading to the Best Linear Unbiased Estimator (BLUE), and then a measurement noise model and a stochastic model for the unknown, x, leading to the Minimum Variance Estimator (MVE). From here, the goal is the Kalman filter. To get there, conditional probability is needed, leading to a treatment of conditional normal random vectors, which, once distilled to a set of key facts, yields the Kalman filter.

Chapter 6 is motivated by considering the tools required to understand the convergence of sequences and the existence of minima and maxima of functions. It's also a last pass through our proof techniques, because Chapters 4 and 5 were proof-light, being more focused on important applications of the math we had already done. We introduce the basic notions of topology, open and closed sets, in the context of normed spaces. Open and closed sets can be characterized with the notion of distance and through the convergence of sequences, showing the interplay of things that may not seem so related at a student's first glance. The contraction mapping principle is covered as are standard results on compact sets and continuous functions on compact sets. This provides general conditions that are easy to check in many engineering situations for the existence of extrema of continuous real-valued functions. These notes conclude with brief remarks on convex functions, QPs, and LPS for minimizing the one-norm and the max-norm.

Jessy Grizzle Ann Arbor, Michigan USA

Chapter 1

Introduction to Mathematical Arguments

Learning Objectives

- Establish notation
- Cover basic proof techniques, some of which may be a review
- Set the stage for cool things to come.

Outcomes

- Learn how to read mathematical statements such as "$\forall\ \epsilon > 0, \exists\ \delta > 0$ such that $\forall\ x \in B_\delta(x_0), ||f(x) - f(x_0)|| < \epsilon$. " This is the definition of a function being continuous at a point x_0, by the way.
- Learn how to negate mathematical statements such as the above.
- Review or learn methods of proofs that we will use on a daily basis in the course.
- In particular, overcome reluctance to use "proof by contradiction".

1.1 Mathematical Notation

$\mathbb{N} = \{1, 2, 3, \cdots\}$ Natural numbers or counting numbers

$\mathbb{Z} = \mathcal{Z} = \{\cdots, -3, -2, -1, 0, 1, 2, 3, \cdots\}$ Integers or whole numbers

$\mathbb{Q} = \left\{ \frac{m}{q} \mid m, q \in \mathbb{Z}, q \neq 0, \text{no common factors (reduce all fractions)} \right\}$ Rational numbers

$\mathbb{R}$ = Real numbers

$\mathbb{C}$ = $\{\alpha + j\beta \mid \alpha, \beta \in \mathbb{R}, j^2 = -1\}$ Complex numbers

$\forall$ means "for every", "for all", "for each"

$\exists$ means "for some", "there exist(s)", "there is/are", "for at least one"

$\in$ means "element of" as in "$x \in A$", i.e., x is an element of the set A

$\sim$ denotes "logical not". You will also often see $\neg$. We'll use both in these notes.

$p \implies q$ means "if the logical or mathematical statement p is true, then the statement q is true"

$p \iff q$ means "p is true if, and only if , q is true". While p iff q is another way to write $p \iff q$, we will mostly avoid using it in these notes.

$p \iff q$ is logically equivalent to

(a) $p \implies q$ and
(b) $q \implies p$

The *contrapositive* of $p \implies q$ is $\sim q \implies \sim p$

The *converse* of $p \implies q$ is $q \implies p$. It is very important to note that in general, $(p \implies q)$ *DOES NOT IMPLY* $(q \implies p)$, and vice-versa. If they did, we would not need $p \iff q$.

Relation: $(p \implies q) \Leftrightarrow (\sim q \implies \sim p)$. The two statements are logically equivalent, and hence in principle, we do not need both of them. We will see, however, that sometimes one of them is easier to use in a proof than the other.

Logical and: $p_1 \wedge p_2$ is read p_1 and p_2. It is true when both p_1 and p_2 are true.

Logical or: $p_1 \vee p_2$ is read p_1 or p_2. It is true when at least one of p_1 and p_2 is true. We do not use "exclusive or" in this course. Hence, if **T** stands for true and **F** for false, then $\mathbf{T} \vee \mathbf{T} = \mathbf{T} \vee \mathbf{F} = \mathbf{F} \vee \mathbf{T} = \mathbf{T}$.

Q.E.D. or QED is an abbreviation of the Latin "quod erat demonstrandum" which means "thus it was demonstrated"; it is used to alert the reader that a proof has been completed. Nowadays, you more frequently see □ or ■ instead of QED.

Warning: In the beginning, it is quite frequent that students confuse the meanings of *contrapositive* and *converse*. Just be careful. With practice, it becomes second nature. The fact that they both start with "con-" is not helpful!

1.2 Vocabulary

The following definitions borrow liberally from Math 299, Michigan State University, https://users.math.msu.edu/users/duncan42/AxiomNotes.pdf

Meanings:

- Definition : An explanation of the mathematical meaning of a word.
- Theorem : A statement that has been proven to be true.
- Proposition : A less important but nonetheless interesting true statement.

- Lemma: A true statement used in proving other true statements (that is, a less important theorem that is helpful in the proof of other results).
- Claim: A true statement that is sometimes made as a step toward proving a theorem, in which case it is similar to a lemma. It is also sometimes used to highlight a property of a mathematical object that a reader might miss. For example, one might define a matrix A to be *invertible* if there exists a second matrix B such that $A \cdot B = B \cdot A = I$, where I is the identity matrix. Nowhere in this definition have we explicitly stated that A must be square. Hence, you might then *Claim:* An invertible matrix is square. And then prove the claim by using the rules (i.e., definition) of matrix multiplication.
- Corollary: A true statement that is a "simple" deduction from a theorem or proposition.
- Proof : The explanation of why a statement is true.
- Conjecture: A statement believed to be true, but for which we have no proof. (A statement that is being proposed to be a true statement).
- Axiom: A basic assumption about a mathematical situation, or said another way, a statement Mathematicians assume to be true. One example comes from Euclid " two distinct parallel lines in the plane never intersect". This is a fundamental tenet of Euclidean geometry and is false in geometries where lines are curved. Another example is the axiom that for any two integers a and b, the symbol "+" called sum has the property $a + b = b + a$, that is, the sum of a and b is equal to the sum of b and a. In this sense, axioms and definitions are very similar. Axioms are reserved for the "bedrock" definitions in a logical or mathematical system, the definitions on which everything else is based.

1.3 Review of Proof Techniques

When constructing a proof of a statement, axioms, definitions, theorems, lemmas, propositions, claims, and corollaries all have the same "power" because they consist of true statements. It is common practice to place the most fundamental description of an idea as its definition and then to provide theorems that show how to check, compute, or apply the property in the definition.

1.3.1 Direct Proofs

A proof is *direct* if the result is obtained by directly applying "simple rules of logic", such as $p \implies q$, to the given assumptions, definitions, axioms, and already known theorems. Employing the method of "proof by contradiction" would not be a direct proof. A better elaboration is given here https://en.wikipedia.org/wiki/Direct_proof

Definition 1.1 *An integer n is* even *if $n = 2k$ for some integer k; n is* odd *if $n = 2k + 1$ for some integer k.*

Remark 1.2 *In a definition, the convention is that "if" means "if, and only if".* It would be considered bad style (form) to write *the above definition as follows: An integer n is* even *if, and only if $n = 2k$ for some integer k; n is* odd *if, and only if $n = 2k + 1$ for some integer k. What you have to understand is that the two meanings are identical when used in a definition. In a lemma, theorem, claim, proposition, corollary, etc., the two meanings are very different. Yes, this takes time before it becomes second nature.*

Example 1.3 *Provide a direct proof that the sum of two odd integers is even.*

Proof: Let n_1 and n_2 be odd integers. Then by the definition of odd, there exist integers k_1 and k_2 such that

$$n_1 = 2k_1 + 1$$
$$n_2 = 2k_2 + 1.$$

Then using the rules of arithmetic,

$$n_1 + n_2 = (2k_1 + 1) + (2k_2 + 1) = 2(k_1 + k_2 + 1).$$

Because $k_1 + k_2 + 1$ is the sum of three integers, it is also an integer, and therefore $2(k_1 + k_2 + 1)$ is by definition, an even integer. Because $n_1 + n_2 = 2(k_1 + k_2 + 1)$, it is even. ■

Did we really need to say the last line in the proof, namely, "Because $n_1 + n_2 = 2(k_1 + k_2 + 1)$, it is even"? It's a matter of taste. A proof is supposed to convince a skeptical reader that something is true. Sometimes, when writing a paper, you do have to restate the obvious to ensure that a reviewer does not miss it.

1.3.2 Proof by Contrapositive:

A proof by *contrapositive* means that to establish $p \implies q$, we prove its logical equivalent, $\sim q \implies \sim p$. Once you have written down what is $\sim q$ and what is $\sim p$, the proof often proceeds like a direct proof.

Example 1.4 *Let n be an integer. Prove that if n^2 is even, then n is even.*

Proof: In the beginning of your proof writing career, it is highly recommended to explicitly write down what is p and what is q. This helps you to understand what you are trying to show and what is/are the given hypothesis(eses).

- $p = (n^2$ is even). Hence, $\sim p = (n^2$ is odd).
- $q = (n$ is even). Hence, $\sim q = (n$ is odd).

Our proof of $p \implies q$ is to show $\sim q \implies \sim p$, that is, if n is odd, then n^2 is odd. Hence, let n be an odd integer. By the definition of n being odd, there exists an integer k such that $n = 2k + 1$. Therefore,

$$n^2 = (2k+1)^2 = 4k^2 + 4k + 1 = 2(2k^2 + 2k) + 1.$$

Because $(2k^2 + 2k)$ is an integer, we are done. ■

Here, we simply took it as "obvious to the most casual observer" that once we arrived at $n^2 = 2m + 1$ for $m = 2k^2 + 2k$, that it was game over, we've said enough to convince "anyone" that n^2 is odd. Was it enough for you?

1.3.3 Proof by Exhaustion

Proof by *exhaustion* means to reduce the proof to a finite number of cases, and then prove each case separately. As its name suggests, these proofs can be tedious at times. The *Famous Four Color Problem* https://en.wikipedia.org/wiki/Four_color_theorem was proved by using computer algebra to reduce the problem to checking a finite number of map cases, and then checking them one by one. In particular, 1,834 map configurations (later reduced to 1,482) had to be checked one by one and took a computer of its day over a thousand hours.

On occasion, we will do a proof by exhaustion. For us, four cases will already be a lot!

1.3.4 Proofs by Induction

There are two forms of proof by *induction*, that are in fact equivalent. Most engineers only know one, the first one:

First Principle of Induction (Standard Induction): Let $P(n)$ denote a statement about the natural numbers with the following properties:

(a) **Base case:** $P(1)$ is true

(b) **Induction hypothesis:** If $P(k)$ is true, then $P(k+1)$ is true.

Then $P(n)$ is true for all $n \geq 1$.

Remark 1.5 *Suppose the base case involves an integer $k_0 \neq 1$. Then you can re-index things and reduce it to the base case having $k_0 = 1$. Alternatively, you assume that $P(k)$ true for $k \geq k_0$ implies $P(k+1)$ is true, and then you get $P(n)$ is true for all $n \geq k_0$. A common mistake is to not use the correct base case. For an example, you should read about how "to prove by induction" that all horses are the same color* https://en.wikipedia.org/wiki/All_horses_are_the_same_color.

Example 1.6 *Let's prove the **Claim:** For all $n \geq 1$, $1 + 3 + 5 + \cdots + (2n-1) = n^2$.*

Proof:

- Step 0: Write down $P(k)$: $1 + 3 + 5 + \cdots + (2k-1) = k^2$.

- Step 1: Check the base case, $P(1)$: For $k = 1$, we have that $1 = 1^2$, and hence the base case is true.
- Step 2: Show the induction hypothesis is true. That is, using the fact that $P(k)$ is true, show that $P(k+1)$ is true. Often, this involves re-writing $P(k+1)$ as a sum of terms that show up in $P(k)$ and another term.

For us,

$$P(k+1) : 1 + 3 + 5 + \cdots + (2k-1) + (2(k+1)-1) = (k+1)^2.$$

For the induction step, we assume that

$$P(k) := 1 + 3 + 5 + \cdots + (2k-1) = k^2$$

is true and thus $P(k+1)$ is true if, and only if

$$k^2 + (2(k+1)-1) = (k+1)^2.$$

Using the known (and accepted) rules of algebra, we check that

$$k^2 + (2(k+1)-1) = k^2 + 2k + 2 - 1 = k^2 + 2k + 1 = (k+1)^2,$$

and hence $P(k+1)$ is true. Because we have shown that $P(1)$ is true and for all $k \geq 1$, $P(k) \implies P(k+1)$, by the Principle of Induction, we conclude that for all $k \geq 1$,

$$1 + 3 + 5 + \cdots + (2k-1) = k^2.$$

■

Remark 1.7 *When the base case $P(1)$ seems so totally trivial that you are unsure whether there is anything to show, it's OK to check $P(2)$, just to convince yourself that it is true too. In our case you would check $P(2) : 1 + 3 = 2^2$; since this is easily established to be true, you may now be more confident of your proof. If you do one more, $P(3) : 1 + 3 + 5 = 3^3$, you are now on a roll and ready to attack the general case by induction. Bottom line: In the beginning, it's natural to be tentative when you do a proof. It takes practice to learn the art of proving things. When you write a proof, we will not take off points for doing a bit more work than is strictly required. We understand that you are slowly building up your confidence.*

Warning: When you seek to establish the *induction hypothesis*, that $P(k+1) \implies P(k)$, you need to make sure that your reasoning works for all k, including the base case, $k = 1$. In the infamous "all horse are the same color proof (spoof)", the statement $P(1) \implies P(2)$ fails. People overlook it by starting at $P(2) \implies P(3)$ or they make a mistake on $P(1) \implies P(2)$. Oooops!

Second Principle of Induction (Strong Induction): Let $P(n)$ be a statement about the natural numbers with the following properties:

(a) **Base Case:** $P(1)$ is true.

(b) **Induction hypothesis:** If $P(j)$ is true for all $1 \leq j \leq k$, then $P(k+1)$ is true.

Then $P(n)$ is true for all $n \geq 1$ (or, n greater than or equal to the n_0 used in the Base Case).

Remark 1.8 *You can see why it is sometimes called* Strong Induction: *we have access to all of the logical statements $P(j)$ up to, and including, $P(k)$, when we are trying to prove the induction step $P(k+1)$ is true. We will show a bit later that the two principles of induction are logically equivalent. Nevertheless, sometimes one method is easier to apply than the other.*

Definition 1.9 *A natural number $n \geq 2$ is* composite *if it can be factored as $n = a \cdot b$, where a and b are natural numbers satisfying $1 < a, b < n$. Otherwise, n is* prime.

Remark 1.10 *It follows from the above definition that a natural number greater than or equal to two is either prime or composite, and it cannot be both. The first prime number is 2. What is the number 1?*

Example 1.11 *Let's prove the **Theorem**: (Fundamental Theorem of Arithmetic) Every natural number $n \geq 2$ can be factored as a product of one or more primes.*

Proof:

- Step 0: We write down the statements. For $k \geq 2$, $P(k)$: there exist $i_k \geq 1$ and prime numbers $p_1, p_2, \ldots, p_{i_k}$ such that the product $p_1 \cdots p_{i_k} = k$.
- Step 1: Check the base case, $P(2)$: For $k = 2$, we have that $2 = 2$, and hence the base case is true.
- Step 2: Show the induction hypothesis is true. That is, using the fact that $P(j)$ is true for $1 \leq j \leq k$, show that $P(k+1)$ is true, that is, $k+1$ can be expressed as a product of primes. There are two cases:
  (a) Case 1: $k+1$ is prime. In this case, we are done because $k+1$ is already the product of one prime, namely itself.
  (b) Case 2: $k+1$ is composite. Then, there exist two natural numbers a and b, $2 \leq a, b \leq k$, such that $k+1 = a \cdot b$.

  Because a and b are natural numbers that are greater than or equal to 2 and less than or equal to k, by the induction step:

$$\begin{aligned} P(a) &\implies a = p_1 \cdot p_2 \cdot \cdots \cdot p_{i_a}, \text{ for some primes } p_i \\ P(b) &\implies b = q_1 \cdot q_2 \cdot \cdots \cdot q_{j_b}, \text{ for some primes } q_j \end{aligned}$$

  Hence, $a \cdot b = (p_1 \cdot p_2 \cdot \cdots \cdot p_{i_a}) \cdot (q_1 \cdot q_2 \cdot \cdots \cdot q_{j_b})$, which is a product of primes.

■

Strong Induction was useful here because we needed to "reach back" and use our statements $P(j)$ for values of j not equal to k. If we relied only on Ordinary Induction, we would have been stuck. This raises the question again, is Strong Induction really more powerful than Ordinary Induction? The answer is NO, it is just sometimes more convenient.

Equivalence of Strong and Ordinary Induction: Let $P(k)$ be the set of logical statements that are used with Strong Induction. Then the induction step is equivalent to

$$(P(1) \wedge P(2) \wedge \cdots \wedge P(k)) \implies P(k+1), \tag{1.1}$$

because we assume that $P(j)$ is true for $1 \leq j \leq k$. Next, you can note that (1.1) is equivalent to

$$P(1) \wedge P(2) \wedge \cdots \wedge P(k) \implies P(1) \wedge P(2) \wedge \cdots \wedge P(k) \wedge P(k+1), \tag{1.2}$$

because if $P(1) \wedge P(2) \wedge \cdots \wedge P(k) = \mathbf{T}$, then

$$(P(1) \wedge P(2) \wedge \cdots \wedge P(k) \wedge P(k+1) = \mathbf{T}) \iff (P(k+1) = \mathbf{T}).$$

It follows that Ordinary Induction on

$$Q(k) := P(1) \wedge P(2) \wedge \cdots \wedge P(k)$$

is equivalent to Strong Induction on $P(k)$.

If we return to our proof of the Fundamental Theorem of Arithmetic, then what is $Q(k)$? Well, $Q(k)$ true means that $P(1), P(2), \ldots, P(k)$ are all true, and hence

$$Q(k): \text{ for all integers } 2 \leq j \leq k, \text{ there exist primes } p_1, p_2, \ldots, p_{i_j}, \text{ such that } j = p_1 \cdot p_2 \cdots p_{i_j}.$$

The proof of the induction step $Q(k) \implies Q(k+1)$ is then identical to the proof we gave for $P(1) \wedge \cdots \wedge P(k) \implies P(k+1)$.

1.3.5 Proof by Contradiction

Definition 1.12 *A* logical contradiction*, or* contradiction *for short, is a logical statement R such that $R \wedge (\sim R) = \boldsymbol{T}$.*

Initially, students tend to dislike proofs by contradiction. It seems to be a contradiction to logical thinking that such a method of proof can even be correct! A statement that is both true and false is called a *contradiction*. The basis for a *proof by contradiction* is that if you start with only true statements and correctly apply the rules of logic, you cannot generate a contradiction. Hence, if you start with a statement, say p, and through valid application of the rules of logic, arrive at a second statement, say R that is both **T** and **F**, then the statement p must be false. We can write this as

$$\big(p \implies (\exists\, R \text{ such that } R \wedge (\sim R) = \mathbf{T})\big) \iff p = \mathbf{F}.$$

Only false statements can generate contradictions.

The game plan for a proof by contradiction: We want to show that a statement p is true. We assume instead that the statement is false. On the basis of p being false, we derive a "contradiction", meaning some statement R that we show to be both true and false. We conclude that $\sim p$ is false, because it led to a contradiction. Hence, p must be true.

Once we do a few examples, it becomes much easier to digest.

Example 1.13 *Use proof by contradiction to show that $\sqrt{2}$ is an irrational number.*

Proof: Our statement is $p : \sqrt{2}$ is irrational. We assume $\sim p : \sqrt{2}$ is rational. We seek to show that this leads to the existence of a statement R that is both true and false, a contradiction.

If $\sqrt{2}$ is rational, then there exist natural numbers m and n such that

- m and n have no common factors,
- $n \neq 0$, and

$$\sqrt{2} = \frac{m}{n}. \tag{1.3}$$

All we have done is apply the definition of a rational number. Next, we square both sides of (1.3) to arrive at

$$\left(2 = \frac{m^2}{n^2}\right) \implies \left(2n^2 = m^2\right) \implies \left(m^2 \text{ is even}\right).$$

From our result in Example 1.4, we deduce that m must be even, and hence there must exist an integer k such that $m = 2k$.

From $2n^2 = m^2$, we deduce that

$$\left(2n^2 = (2k)^2\right) \implies \left(2n^2 = 4k^2\right) \implies \left(n^2 = 2k^2\right) \implies n^2 \text{ is even.}$$

Once again appealing to our result in Example 1.4, we deduce that n must be even, and hence there must exist an integer j such that $n = 2j$.

Because both m and n are even, they have 2 as a common factor, which is a contradiction to m and n have no common factors.

Because we arrived at this contradiction from the statement "$\sqrt{2}$ is rational", we deduce that "$\sqrt{2}$ is rational" must be false. Hence, $\sqrt{2}$ is irrational. ■.

Here is a blow by blow recap of the proof.

- We define $p : \sqrt{2}$ is an irrational number.
- We start with the assumption $(\sim p = \mathbf{T})$, that is, $\sqrt{2}$ is a rational number.
- Based on that assumption, we can deduce there exist integers m and n, $n \neq 0$, such that $\sqrt{2} = \frac{m}{n}$ and m and n do not have a common factor.
- We now define ($R : m$ and n do not have a common factor) and know that $R = \mathbf{T}$.
- However, from $\sqrt{2} = \frac{m}{n}$, we show that m and n have 2 as a common factor.
- We now have $\sim R = \mathbf{T}$.
- Hence, $(R \wedge (\sim R)) = \mathbf{T}$, **which is a contradiction**.
- Conclusion: $\sim p = \mathbf{T}$ is impossible, and therefore $\sim p = \mathbf{F}$.
- Hence, $p = \mathbf{T}$ and we have proved that $\sqrt{2}$ is irrational. Pretty cool!

Remark 1.14 *You can also use* proof by contradiction *to prove $p \implies q$. What you do is start with $p \wedge (\sim q)$, that is, you assume p is true and $\sim q$ is true. This gives you an easy way of generating a statement that you believe to be false, and hence should lead to a contradiction.*

1.3.6 Summary:

In conclusion, we have the following proof techniques.

- Direct Proof: $p \implies q$
- Proof by Contrapositive: $\sim q \implies \sim p$. (Start with the conclusion being false, that is $\sim q$ and do logical steps to arrive at $\sim p$)
- Proof by Contradiction Version I: To show that p is a true statement, you assume instead that $\sim p$ is true and seek a statement R such that both R and $\sim R$ are true, which is a contradiction. Deduce that $\sim p = \mathbf{F}$ and hence $p = \mathbf{T}$.
- Proof by Contradiction Version II: Start with $p \wedge (\sim q)$ (assume p is true and q is false. Find a statement R such that both R and $\sim R$ are true, which is a contradiction. Deduce that $\sim (p \wedge (\sim q)) = \mathbf{T}$, that is, $(p \wedge (\sim q)) = \mathbf{F}$, and hence, if $p = \mathbf{T}$ then $q = \mathbf{T}$. That is, $p \implies q$.
- Hence, we have that $(p \implies q) \iff \sim (p \wedge (\sim q))$.
- Proof by Induction, which can be done in two forms.
- Proof by Exhaustion, where we enumerate a finite set of cases and check them one by one.
- To show $p \iff q$ is true, we need to show **both** $p \implies q$ **and its converse**, $q \implies p$.

  A rookie mistake is to prove "both" $p \implies q$ and its contrapositive , $\sim q \implies \sim p$. The problem here is that $\sim q \implies \sim p$ is logically equivalent to $p \implies q$ and hence all you have done is prove the same thing two different ways, which is not what you wanted to do!

1.4 Truth Tables

Logic Tables or *Truth Tables* are simply a list of the possible logical values of a statement's inputs followed by the corresponding value of the statement's output. Here is a *truth table* for the negation operation, which has one input, p and one output $\neg p$.

p	$\neg p$
T	F
F	T

Here is a *truth table* for $p \wedge q$, which has two inputs, p and q, and one output $p \wedge q$.

p	q	$p \wedge q$
T	T	T
T	F	F
F	T	F
F	F	F

Here is a *truth table* for $p \implies q$ using *proof by contradiction*. The table has two inputs p and q, and one output $\sim (p \wedge (\sim q))$, the last column. We include several intermediate columns required to compute the output. The input and output are highlighted in blue.

p	q	$\neg q$	$p \wedge \neg q$	$\neg(p \wedge \neg q)$
T	T	F	F	T
T	F	T	T	F
F	T	F	F	T
F	F	T	F	T

It is remarkably easy to check that the above table is correct because it only involves logical negation and logical and. The hard part is accepting that $(p \implies q) \iff \neg(p \wedge \neg q)$.

Now we give you a partially filled *truth table* for $p \implies q$, which has two inputs, p and q, and one output $p \implies q$ and ask you to complete it.

p	q	$p \implies q$
T	T	T
T	F	F
F	T	?
F	F	?

Remark 1.15 *Here is one way to think about it that I once found on the web and have lost the link: We define p : (you score 100% on all exams and assignments) and q : (I assign you an A^+ grade for the course). We all agree that $p \implies q$ must be a correct statement. In fact, it should be so deep into the bedrock of grading that we can call it an axiom! It's just that until now, you maybe never thought to complete a truth table for it! So, consider the following four scenarios, which correspond to the four lines in the truth table:*

*p = you score 100% and q = your grade is an A^+, then $(p \implies q) =$ **T** (easy)*

*p = you score 100% and q = your grade is an A^-, then $(p \implies q) =$ **F** (easy)*

*p = you score 85% and q = your grade is an A^+, then $(p \implies q) =$ **?** (ask yourself, does this invalidate the statement?)*

*p = you score 85% and q = your grade is an A^-, then $(p \implies q) =$ **?** (ask yourself, does this invalidate the statement?)*

*To get the answers, look at the truth table of $\neg(p \wedge \neg q)$. If this makes your head spin, don't worry about it. After we do a bunch of proofs, you can come back to this riddle. The reasoning is, once you fail to live up to the 100% standard, then $p =$ **F**. Then, no matter what grade I give you, I am not invalidating the promise, $p \implies q$. Hence, the questions marks cannot be **F**. Because we are using binary logic, $\neg$**F** = **T**. Therefore, the question marks must be replaced with **T**. Pretty cool, right?*

1.5 Negating Logical Statements

By now, you've noticed that several of the proof techniques require you to negate a logical statement. Some are super easy, such as, when $p : x > 0$, you easily see that $\sim p : x \leq 0$. In the beginning, for more complex statements, we recommend you first "translate them to English (or your preferred language), negate that statement, and then "translate" back into math. The math symbols are simply shortcuts for word phrases, so there is nothing illogical or wrongheaded about doing negations this way. It does take more time, and with practice, you will learn to skip the "translation" step, altogether.

Example 1.16 *Let $p : \forall\, x \in \mathbb{R},\ f(x) > 0$. Compute its negation.*

Solution:

- Math form: $p : \forall\, x \in \mathbb{R},\ f(x) > 0$
- Word form: p : for all $x \in \mathbb{R}$, $f(x) > 0$
- Negate: $\sim p$: not (for all $x \in \mathbb{R}$, $f(x) > 0$)
- Equivalent: $\sim p$: for some $x \in \mathbb{R}$, not[$f(x) > 0$]
- Equivalent: $\sim p$: for some $x \in \mathbb{R}$, $f(x) \leq 0$
- Math form: $\sim p$: $\exists\, x \in \mathbb{R}$, such that $f(x) \leq 0$

■

Example 1.17 *Let $y \in \mathbb{R}$ and define $p : \forall\ \delta > 0, \exists x \in \mathcal{Q}$ such that $|x - y| < \delta$. Determine its negation. Recall, $\mathcal{Q}$ is the set of rational numbers.*

Solution:

- Math form: $p : \forall\ \delta > 0, \exists x \in \mathcal{Q}$ such that $|x - y| < \delta$
- Word form: p : for all $\delta > 0$, there exists $x \in \mathcal{Q}$ such that $|x - y| < \delta$
- Negate: $\sim p$: not (for all $\delta > 0$, there exists $x \in \mathcal{Q}$ such that $|x - y| < \delta$)
- Equivalent: $\sim p$: for some $\delta > 0$, not[there exists $x \in \mathcal{Q}$ such that $|x - y| < \delta$]
- Equivalent: $\sim p$: for some $\delta > 0$, there does not exist $x \in \mathcal{Q}$ such that $|x - y| < \delta$
- Equivalent: $\sim p$: there exists $\delta > 0$, such that for all $x \in \mathcal{Q}$, not[$|x - y| < \delta$]
- Equivalent: $\sim p$: there exists $\delta > 0$, such that for all $x \in \mathcal{Q}$, $|x - y| \geq \delta$
- Math form: $\sim p$: $\exists\ \delta > 0, \forall x \in \mathcal{Q}, |x - y| \geq \delta$

■

When negating statements with $\exists$ and $\forall$, here are the logical equivalents of their negations:

$$\sim \forall \iff \exists$$
$$\sim \exists \iff \forall$$

Remark 1.18 *It is better to avoid $\not\forall$ and $\not\exists$, but they are legal symbols.*

Example 1.19 *Negate the statement $p : \forall y \in \mathbb{R}, \forall\ \delta > 0, \exists x \in \mathcal{Q}$ such that $|x - y| < \delta$.*

Solution: $\sim p : \exists y \in \mathbb{R}$ and $\exists \delta > 0$ such that, $\forall\ x \in \mathcal{Q}, |x - y| \geq \delta$. ■

1.6 Key Properties of Real Numbers

Let A be a subset of the reals, $\mathbb{R}$.

Definition 1.20 *An element $b \in A$ is a* maximum *of A if $x \leq b$ for all $x \in A$. We note that in the definition, b must be an element of A. We denote it by*

$$\max\ A \text{ or } \max\{A\}.$$

It is very important to note that a *max of a set may not exist*! Indeed, the set $A = \{x \in \mathbb{R} \mid 0 < x < 1\}$ does not have a maximum element. We will see later that it does not have a minimum either. This is what motivates the notions of supremum and infimum.

Definition 1.21 *An element $b \in \mathbb{R}$ is an* upper bound *of A if $x \leq b$ for all $x \in A$. We say that A is* bounded from above.

Remark 1.22 *We note that in the definition of upper bound, b does NOT have to be an element of A.*

Definition 1.23 *An element $b^* \in \mathbb{R}$ is the* least upper bound *of A if*

(a) b^* is an upper bound, that is $\forall\ x \in A$, $x \leq b^*$, and

(b) if $b \in \mathbb{R}$ satisfies $x \leq b$ for all $x \in A$, then $b^* \leq b$. (This means that there is no other upper bound that is strictly smaller than b^*.)

Notation and Vocabulary 1.24 *The least* least upper bound *of A is also called the* supremum *of A and is denoted*

$$\sup A \quad \text{or} \quad \sup\{A\}$$

Key Theorem If $A \subset \mathbb{R}$ is bounded from above, then $\sup\{A\}$ exists in $\mathbb{R}$.

Remark 1.25 *The rational numbers $\mathcal{Q}$ do not satisfy the above property, namely, sets that are bounded from above may not have a supremum within the rational numbers. Let $A \subset \mathcal{Q}$ be defined as*

$$A := \{x \in \mathcal{Q} \mid \exists\, n \geq 1, x + \frac{1}{n} \leq \sqrt{2}\}.$$

The number $1.5 \in \mathcal{Q}$ is an upper bound of A, but there is no rational number that is a least upper bounded. *Of course, viewed as a subset of the real numbers, A has a least upper bound, namely $\sqrt{2}$. This is a subtle but important difference. In fact, one can construct the real numbers from the rational numbers as the "smallest set that (i) contains the rationals and (ii) all subsets that are bounded from above have a least upper bound, that is, a supremum.*

Example 1.26

- $A = \{x \in \mathbb{R} \mid 0 < x < 1\}$. *Then* $\sup A = 1$.
- $A = \{x \in \mathbb{R} \mid x^2 \leq 2\}$. *Then* $\sup A = \sqrt{2}$.

Remark 1.27 ***(Déjà Vu All Over Again)*** *The existence of a* supremum *is a special property of the real numbers. If you are working within the rational numbers, an upper bounded set may not have a <u>rational</u> supremum. If you view the set as a subset of the reals, it will then have a supremum, but the supremum may be an irrational number.*

Example 1.28

- $A = \{x \in \mathbb{Q} \mid 0 < x < 1\}$. *Then* $\sup A = 1$. *Indeed, 1.0 is a rational number, it is an upper bound, and it less than or equal to any other upper bound; hence it is the supremum.*
- $A = \{x \in \mathbb{Q} \mid x^2 \leq 2\}$. *Then* $(1.42)^2 = 2.0164$*, and thus $b = 1.42$ is a rational upper bound. Also* $(1.415)^2 = 2.002225$*, and thus $b = 1.1415$ is a smaller rational upper bound. However, there is no least upper bound within the set of rational numbers. When we view the set A as being a subset of the real numbers, then there is a real number that is a least upper bound and we have* $\sup A = \sqrt{2}$*. This is what we mean when we say that the existence of a supremum is a special or distinguishing property of the real numbers.*

Important Fact: Suppose $A \subset \mathbb{R}$. If the *supremum* of A is in the set A, then it is equal to the *maximum*. In fact,

$$b^* = \max\{A\} \iff (b^* = \sup\{A\}) \wedge (b^* \in A).$$

Everything we have done above can be repeated with *minimum* replacing *maximum*, *greatest lower bound* replacing *least upper bound*, and *infimum* replacing *supremum*. Consider once again a set $A \subset \mathbb{R}$.

Definition 1.29 *An element $b \in A$ is a* minimum *of A if $b \leq x$ for all $x \in A$. We note that in the definition, b <u>must</u> be an element of A. We denote it by* $\min\ A$ *or* $\min\{A\}$.

It is important to note that a *min* of a set may not exist! Indeed, the set $A = \{x \in \mathbb{R} \mid 0 < x < 1\}$ does not have a minimum element.

Definition 1.30 *An element $b \in \mathbb{R}$ is a* lower bound *of A if $b \leq x$ for all $x \in A$. We say that A is* bounded from below.

Remark 1.31 *We note that in the definition of lower bound, b does NOT have to be an element of A.*

Definition 1.32 *An element $b^* \in \mathbb{R}$ is the* greatest lower bound *of A if*

1. *b^* is a lower bound, that is $\forall\, x \in A$, $b^* \leq x$, and*
2. *if $b \in \mathbb{R}$ satisfies $b \leq x$ for all $x \in A$, then $b^* \geq b$.*

Notation and Vocabulary 1.33 *The greatest lower bound of A is also called the* ***infimum*** *and is denoted*

$$\inf A \quad \textit{or} \quad \inf\{A\}$$

Key Theorem If the set A is bounded from below, then $\inf\ A$ exists.

Example 1.34

- $A = \{x \in \mathbb{R} \mid 0 < x < 1\}$. *Then* $\inf A = 0$.
- $A = \{x \in \mathbb{R} \mid x^2 \leq 2\}$. *Then* $\inf A = -\sqrt{2}$.

Remark 1.35 *If the* infimum *is in the set A, then it is equal to the* minimum.

Definition 1.36 *If a set $A \subset \mathbb{R}$ is not bounded from above, we define* $\sup\ A = +\infty$. *If $A \subset \mathbb{R}$ is not bounded from below, we define* $\inf\ A = -\infty$. The symbols $+\infty$ and $-\infty$ are not real numbers. *The* ***extended real numbers*** *are sometimes defined as*

$$\mathbb{R}_e := \{-\infty\} \cup \mathbb{R} \cup \{+\infty\}.$$

Remark 1.37 ***(Final)*** *Michigan's MATH 451* constructs *the real numbers from the rational numbers. This is a very cool process to learn. Unfortunately, we do not have the time to go through this material in ROB 501. The existence of supremums and infimums for bounded subsets of the real numbers is a consequence of how the real numbers are defined (i.e., constructed).*

Chapter 2

Some Highlights of Abstract Linear Algebra (or Practicing Proofs in a Safe Environment)

Learning Objectives

- Establish key definitions for fields, vectors, vector spaces, linear combinations, linear independence, subspaces, basis vectors, linear operators, matrix representations, eigenvalues and eigenvectors.
- Initial practice with the proof techniques of the previous chapter.

Outcomes

- Understand a very general notion of vector spaces, where the vectors range from the standard column vectors in $\mathbb{R}^n$ to matrices and functions.
- Understand how the notion of representing a vector with respect to a set of basis vectors establishes one-to-one relations with columns of numbers that can be used in numerical computations.
- Understand a related notion for associating linear operators to matrices.
- Set you up for Chapter 3 which solves best approximation problems for generalizations of the Euclidean norm.
- Have your mind blown by the real numbers forming an infinite dimensional vector space over the field of rational numbers.

2.1 Fields and Vector Spaces

Mostly in this course, we work with the real numbers or the complex numbers. However, in this Chapter, as part of our push to tear you away from the familiar and confront you with more abstraction than you may have seen before, we define a general *field*, $\mathcal{F}$. When you are first learning the abstract definition, it is good to keep a " canonical example" in mind, and that would be $\mathcal{F} = \mathbb{R}$.

Definition 2.1 *(Chen, 2nd edition, page 8) : A **field** consists of a set, denoted by $\mathcal{F}$, of elements called **scalars** and two operations called addition "$+$" and multiplication "$\cdot$"; the two operations are defined over $\mathcal{F}$ such that they satisfy the following conditions:*

1. *To every pair of elements α and β in $\mathcal{F}$, there correspond an element $\alpha + \beta$ in $\mathcal{F}$ called the* sum *of α and β, and an element $\alpha \cdot \beta$ (or simply $\alpha\beta$) in $\mathcal{F}$ called the* product *of α and β.*
2. *Addition and multiplication are respectively commutative: For any α and β in $\mathcal{F}$,*
$$\alpha + \beta = \beta + \alpha \qquad\qquad \alpha \cdot \beta = \beta \cdot \alpha$$
3. *Addition and multiplication are respectively associative: For any α, β, γ in $\mathcal{F}$,*
$$(\alpha + \beta) + \gamma = \alpha + (\beta + \gamma) \qquad\qquad (\alpha \cdot \beta) \cdot \gamma = \alpha \cdot (\beta \cdot \gamma)$$
4. *Multiplication is distributive with respect to addition: For any α, β, γ in $\mathcal{F}$,*
$$\alpha \cdot (\beta + \gamma) = (\alpha \cdot \beta) + (\alpha \cdot \gamma)$$
5. *$\mathcal{F}$ contains an element, denoted by 0, and an element, denoted by 1, such that $\alpha + 0 = \alpha$ and $1 \cdot \alpha = \alpha$ for every α in $\mathcal{F}$.*
6. *To every α in $\mathcal{F}$, there is an element β in $\mathcal{F}$ such that $\alpha + \beta = 0$. The element β is called the* additive inverse.
7. *To every α in $\mathcal{F}$ which is not the element 0, there is an element γ in $\mathcal{F}$ such that $\alpha \cdot \gamma = 1$. The element γ is called the* multiplicative inverse.

To show a given set is a field, you must check all seven axioms. To show something is not a field, you only need to show that one of the axioms fails.

Examples of Fields	Non-examples
$\mathbb{R}$	Irrational (Fails axiom 1)
$\mathbb{C}$	2×2 matrices, real coeff. (Fails axiom 2)
$\mathbb{Q}$	2×2 diagonal matrices real coeff. (Fails axiom 7)

Definition 2.2 *(Chen 2nd Edition, page 9) A **vector space (or, linear space)** over a field $\mathcal{F}$, denoted by $(\mathcal{X}, \mathcal{F})$, consists of a set, denoted by $\mathcal{X}$, of elements called **vectors**, a field $\mathcal{F}$, and two operations called **vector addition** and **scalar multiplication**. The two operations are defined over $\mathcal{X}$ and $\mathcal{F}$ such that they satisfy all the following conditions:*

1. *To every pair of vectors v^1 and v^2 in $\mathcal{X}$, there corresponds a vector $v^1 + v^2$ in $\mathcal{X}$, called the sum of v^1 and v^2* [1].
2. *Addition is commutative: For any v^1, v^2 in $\mathcal{X}$, $v^1 + v^2 = v^2 + v^1$.*
3. *Addition is associative: For any v^1, v^2, and v^3 in $\mathcal{X}$, $\left(v^1 + v^2\right) + v^3 = v^1 + \left(v^2 + v^3\right)$.*
4. *$\mathcal{X}$ contains a vector, denoted by $\mathbf{0}$, such that $\mathbf{0}+v = v$ for every v in $\mathcal{X}$. The vector $\mathbf{0}$ is called the zero vector or the origin.*
5. *To every v in $\mathcal{X}$, there is a vector $\bar{v}$ in $\mathcal{X}$, such that $v + \bar{v} = 0$.*
6. *To every α in $\mathcal{F}$, and every v in $\mathcal{X}$, there corresponds a vector $\alpha \cdot v$ in $\mathcal{X}$ called the* scalar product *of α and v.*
7. *Scalar multiplication is associative: For any α, β in $\mathcal{F}$ and any v in $\mathcal{X}$, $\alpha \cdot (\beta \cdot x) = (\alpha \cdot \beta) \cdot v$.*

[1] We use superscripts v^1, v^2, v^3 to denote different *vectors*. The superscripts *do not* denote powers.

8. *Scalar multiplication is distributive with respect to vector addition: For any α in $\mathcal{F}$ and any v^1, v^2 in $\mathcal{X}$, $\alpha \cdot \left(v^1 + v^2\right) = \alpha \cdot v^1 + \alpha \cdot v^2$.*

9. *Scalar multiplication is distributive with respect to scalar addition: For any α, β in $\mathcal{F}$ and any v in $\mathcal{X}$, $(\alpha + \beta) \cdot v = \alpha \cdot v + \beta \cdot v$.*

10. *For any v in $\mathcal{X}$, $1 \cdot v = v$, where 1 is the element 1 in $\mathcal{F}$.*

Example 2.3 *$\mathcal{F}$ = field, $\mathcal{X}$ = set of vectors*

1. *Every field forms a vector space over itself. $(\mathcal{F}, \mathcal{F})$. Examples: $(\mathbb{R}, \mathbb{R})$, $(\mathbb{C}, \mathbb{C})$, $(\mathbb{Q}, \mathbb{Q})$.*

2. *$(\mathbb{C}, \mathbb{R})$, meaning $\mathcal{X} = \mathbb{C}$, $\mathcal{F} = \mathbb{R}$, is a vector space. This works because a real scalar $\alpha \in \mathcal{F} = \mathbb{R}$ times a complex number $v \in \mathcal{X} = \mathbb{C}$ yields a complex number $\alpha \cdot v \in \mathcal{X} = \mathbb{C}$.*

3. *$\mathcal{F} = \mathbb{R}$, and $\mathcal{X} = \{f : D \to \mathbb{R}\} = \{$functions from D to $\mathbb{R}\}$, where $D \subset \mathbb{R}$ (examples: $D = [a, b]$; $D = (0, \infty)$; $D = \mathbb{R}$) with vector addition and scalar times vector multiplication defined as follows:*

   (a) *$\forall\ f, g \in \mathcal{X}$, define $f + g \in \mathcal{X}$ by $\forall t \in D,\ (f + g)(t) := f(t) + g(t)$;*

   (b) *$\forall\ f \in \mathcal{X}$ and $\alpha \in \mathbb{R}$, define $\alpha \cdot f \in \mathcal{X}$ by $\forall t \in D, (\alpha \cdot f)(t) := \alpha \cdot f(t)$.*

   Note that we are using the known properties of addition and multiplication of real numbers to define what we mean by the sum of two functions and the product of a real number and a function. To prove that $(\mathcal{X}, \mathcal{F})$ is a vector space, you must check all ten of the axioms. We'll check #8 to illustrate what as to be done: we need to show that scalar multiplication is distributive with respect to vector addition:

$$\alpha \cdot (f + g) = \alpha \cdot f + \alpha \cdot g.$$

   Our method of proof is to show that the left hand side (LHS) equals the right hand side (RHS) when

   - *we use the definition of a function evaluated at a point t, and*
   - *we subsequently apply the known definitions for sums and products of real numbers.*

   Let $t \in D$, then

   (a) *LHS: $[\alpha \cdot (f + g)](t) := \alpha \cdot [f + g](t) = \alpha \cdot [f(t) + g(t)] = \alpha \cdot f(t) + \alpha \cdot g(t)$.*

   (b) *RHS: $[\alpha \cdot f + \alpha \cdot g](t) := [\alpha \cdot f](t) + [\alpha \cdot g](t) = \alpha \cdot f(t) + \alpha \cdot g(t)$*

   (c) *Hence, LHS = RHS and we are done.*

4. *Let $\mathcal{F}$ be a field and define $\mathcal{X} := \mathcal{F}^n$ the set of n-tuples written as columns*

$$\mathcal{F}^n := \left\{ \left. \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix} \right| \alpha_i \in \mathcal{F}, 1 \leq i \leq n \right\}$$

   Then $(\mathcal{F}^n, \mathcal{F})$ is a vector space with the operations

   (a) *Vector Addition:* $\begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix} + \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_n \end{bmatrix} = \begin{bmatrix} \alpha_1 + \beta_1 \\ \vdots \\ \alpha_n + \beta_n \end{bmatrix}$

   (b) *Scalar Multiplication:* $\alpha \cdot x = \begin{bmatrix} \alpha x_1 \\ \vdots \\ \alpha x_n \end{bmatrix}$

5. *In a similar manner, one can take $\mathcal{X} := \mathcal{F}^{n \times m} = \{n \times m$ matrices with coefficients in $\mathcal{F}\}$ with vector addition and scalar times vector multiplication defined in the obvious ways. Hence, matrices can be viewed as vectors.*

6. *$(\mathbb{R}, \mathbb{Q})$ is a vector space. It is not particular useful in Robotics. It's kind of cool to think about. The set of vectors is $\mathbb{R}$. Clearly, the sum any two real numbers is again a real number, and hence a vector. The usual scalar times vector multiplication works because the product of a rational number and a real number is a real number, and hence is a vector.*

Non-Example 2.4

1. *Take $\mathcal{X} = \mathbb{R}, \mathcal{F} = \mathbb{C}$. Then $(\mathcal{X}, \mathcal{F}) := (\mathbb{R}, \mathbb{C})$ fails the definition of scalar multiplication (and others).*
2. *$\mathcal{X} = \{x \geq 0, x \in \mathbb{R}\}, \mathcal{F} = \mathbb{R}$. Then $(\mathcal{X}, \mathcal{F})$ fails the definition of scalar multiplication (and others).*
3. *$(\mathbb{Q}, \mathbb{R})$ is a not vector space. The set of vectors is $\mathbb{Q}$. Clearly, the sum any two rational numbers is again a rational number, and hence a vector. The usual scalar times vector multiplication fails, however, because the product of a real number and a rational number is a real number, and hence is not a vector.*

2.2 Subspaces

Notation 2.5 *Let A and B be sets. Then*

- $(A \subset B) \iff (a \in A \implies a \in B)$
- $(A = B) \iff (A \subset B \text{ and } B \subset A)$

In ROB 501, we do not use the notion of A being a *strict subset* of B, which in some books is denoted as $A \subsetneq B$ or $A \subsetneqq B$. Please do not use such notation in your HWs or exams.

Definition 2.6 *Let $(\mathcal{X}, \mathcal{F})$ be a vector space, and let $\mathcal{Y}$ be a subset of $\mathcal{X}$. Then $\mathcal{Y}$ is a **subspace** if using the rules of vector addition and scalar multiplication defined in $(\mathcal{X}, \mathcal{F})$, we have that $(\mathcal{Y}, \mathcal{F})$ is a vector space.*

Remark 2.7 *To show that a set is a subspace using the definition, you have to check axioms 1 to 10. To show that a set is NOT a subspace, you just need to show that one of the axioms is violated. The first thing you should always check is that $0 \in \mathcal{Y}$.*

Proposition 2.8 *(Tools to check that something is a subspace) Let $(\mathcal{X}, \mathcal{F})$ be a vector space and $\mathcal{Y} \subset \mathcal{X}$. Then, the following are equivalent (TFAE):*

(a) $(\mathcal{Y}, \mathcal{F})$ is a subspace of $(\mathcal{X}, \mathcal{F})$.

(b) $\forall v^1, v^2 \in \mathcal{Y}, v^1 + v^2 \in \mathcal{Y}$ (closed under vector addition), and $\forall y \in \mathcal{Y}$ and $\alpha \in \mathcal{F}, \alpha y \in \mathcal{Y}$ (closed under scalar multiplication).

(c) $\forall v^1, v^2 \in \mathcal{Y}$, $\forall \alpha \in \mathcal{F}, \alpha \cdot v^1 + v^2 \in \mathcal{Y}$.

(d) $\forall v^1, v^2 \in \mathcal{Y}$, $\forall \alpha_1, \alpha_2 \in \mathcal{F}, \alpha_1 \cdot v^1 + \alpha_2 \cdot v^2 \in \mathcal{Y}$.

Because (a) through (d) are equivalent, you can use any of (b) through (d) to show that (a) is true.

Example 2.9

- *$(\mathcal{X}, \mathcal{F}) := (\mathbb{R}^2, \mathbb{R})$ and $\mathcal{Y} := \left\{ \begin{bmatrix} \beta \\ 2\beta \end{bmatrix} \mid \beta \in \mathbb{R} \right\} \subset \mathcal{X}$. For $v^1, v^2 \in \mathcal{Y}$, we have*

$$\underbrace{\begin{bmatrix} \beta_1 \\ 2\beta_1 \end{bmatrix}}_{v^1} + \underbrace{\begin{bmatrix} \beta_2 \\ 2\beta_2 \end{bmatrix}}_{v^2} = \begin{bmatrix} \beta_1 + \beta_2 \\ 2(\beta_1 + \beta_2) \end{bmatrix} \in \mathcal{Y}$$

and for $v \in \mathcal{Y}$ and $\alpha \in \mathbb{R}$,

$$\alpha \underbrace{\begin{bmatrix} \beta \\ 2\beta \end{bmatrix}}_{v} = \begin{bmatrix} \alpha\beta \\ 2(\alpha\beta) \end{bmatrix} \in \mathcal{Y}.$$

Hence, by the Proposition, $\mathcal{Y}$ is a subspace.

- $(\mathcal{X}, \mathcal{F})$, $\mathcal{F} = \mathbb{R}$, *where* $\mathcal{X} = \{f : \mathbb{R} \to \mathbb{R}\}$, *and we define*

$$\mathcal{Y} := \mathcal{P}(t) := \{\textit{polynomials in t with real coefficients}\}$$

is a subspace because we know that the sum of two polynomials with real coefficients is another polynomial with real coefficients, and the product of a real number and a polynomial with real coefficients is once again a polynomial with real coefficients.

- *In a similar manner, you can check that* $\widetilde{\mathcal{Y}} := \left\{f : \mathbb{R} \to \mathbb{R} \mid f \textit{ is differentiable }, \frac{d}{dt}f \equiv 0\right\}$ *is a subspace of* $\mathcal{X}$.

Non-Example 2.10 *We take* $(\mathcal{X}, \mathcal{F}) := (\mathbb{R}^2, \mathbb{R})$ *and define* $\mathcal{Y} := \left\{ \begin{bmatrix} \beta \\ 2\beta \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} \mid \beta \in \mathbb{R} \right\} \subset \mathcal{X}$. *Then* $0 \notin \mathcal{Y}$ *and hence* $\mathcal{Y}$ *is not a subspace. In a similar manner,* $\widehat{\mathcal{Y}} := \{f : \mathbb{R} \to \mathbb{R} \mid f(2) = 1.0\}$ *is not a subspace because it does not contain the zero vector, that is, a function that is zero for all* $t \in \mathbb{R}$.

2.3 Linear Combinations and Linear Independence

Definition 2.11 *Let* $(\mathcal{X}, \mathcal{F})$ *be a vector space. A* ***linear combination*** *is a* ***finite sum*** *of the form* $\alpha_1 v^1 + \alpha_2 v^2 + \ldots + \alpha_n v^n$ *where* $n \geq 1, \alpha_i \in \mathcal{F}, v^i \in \mathcal{X}$. **To be extra clear, a sum of the form** $\sum_{i=1}^{\infty} \alpha_i v^i$ **is not a linear combination because it is not finite.**

Definition 2.12 *A finite set of vectors* $\{v^1, \ldots, v^k\}$ *is* ***linearly dependent*** *if* $\exists \alpha_1, \ldots \alpha_k \in \mathcal{F}$ ***not all zero*** *such that* $\alpha_1 v^1 + \alpha_2 v^2 + \ldots + \alpha_k v^k = 0$. *Otherwise, the set is* ***linearly independent****.*

Remark 2.13 *Suppose* $\{v^1, \ldots, v^k\}$ *is a linearly dependent set. Then,* $\exists \alpha_1, \ldots, \alpha_k$ *are not all zero such that*

$$\alpha_1 v^1 + \alpha_2 v^2 + \ldots + \alpha_k v^k = 0.$$

Suppose $\alpha_k \neq 0$. *Then,*

$$\begin{aligned} \alpha_k v^k &= -\alpha_1 v^1 - \alpha_2 v^2 - \cdots - \alpha_{k-1} v^{k-1} \\ v^k &= -\frac{\alpha_1}{\alpha_k} v^1 - \frac{\alpha_2}{\alpha_k} v^2 - \cdots - \frac{\alpha_{k-1}}{\alpha_1} v^{k-1} \end{aligned}$$

and therefore v^k *is a linear combination of the set* $\{v^1, \ldots, v^{k-1}\}$. *Using this observation, you can prove the following.*

Claim 2.14 *For a finite set of vectors,* $\mathcal{S} := \{v^1, \ldots, v^k\}$, *TFAE*

(a) The set is $\mathcal{S}$ *linearly independent.*

(b) $\{v^1, \ldots, v^{k-1}\}$ *is linearly independent and* v^k *cannot be written as a linear combination of* $\{v^1, \ldots, v^{k-1}\}$.

(c) Every (finite) subset of $\mathcal{S}$ *is linearly independent.*

The equivalence of (a) and (c) motivates the following definition for linear independence of sets that may have an infinite number of elements.

Definition 2.15 *An arbitrary set of vectors* $\mathcal{S} \subset \mathcal{X}$ *is* ***linearly independent*** *if every finite subset is linearly independent.*

Example 2.16 *Let* $\mathcal{F} = \mathbb{R}$ *and* $\mathcal{X} = \mathbb{P}(t) = \{$ *set of polynomials with real coefficients* $\}$. *Prove the following* ***Claim:*** *The monomials are linearly independent. In particular, for each* $n \geq 0$, *the set* $\{1, t, \ldots, t^n\}$ *is linearly independent.*

Direct Proof: Suppose that $p(t) := \alpha_0 + \alpha_1 t + \cdots + \alpha_n t^n = 0$ is the zero polynomial. We need to show that $\alpha_0 = \alpha_1 = \cdots = \alpha_n = 0$. From Calculus, we know that a polynomial of degree n is identically zero if, and only, if, $p(0) = 0$ and $\frac{d^k p(t)}{dt^k}|_{t=0} = 0$ for

$k = 1, 2, \ldots, n$. Armed with this notion, we check that

$$\begin{aligned} 0 = p(0) &\implies \alpha_0 = 0 \\ 0 = \frac{dp(t)}{dt}|_{t=0} = (\alpha_1 + 2\alpha_2 t + \cdots + n\alpha_n t^{n-1})|_{t=0} &\implies \alpha_1 = 0 \\ 0 = \frac{d^2p(t)}{dt^2}|_{t=0} = (2\alpha_2 + 6\alpha_3 t + \cdots + n(n-1)\alpha_n t^{n-2})|_{t=0} &\implies \alpha_2 = 0 \\ &\vdots \\ 0 = \frac{d^n p(t)}{dt^n}|_{t=0} = n!\alpha_n &\implies \alpha_n = 0. \end{aligned}$$

Proof by Induction: Let $k \geq 0$, and define the property $P(k)$ by

$$P(k) : \{1, t, \ldots, t^k\} \text{ is linearly independent.}$$

Base Case: $P(0)$ is true; that is, the set $\{1\}$ is linearly independent. You can check this on your own.

Induction Step: For $k \geq 0$, we assume that $P(k)$ is true and we must show that $P(k+1)$ is true.

By a previous remark, it is enough to show that t^{k+1} cannot be written as a linear combination of $\{1, t, \ldots, t^k\}$. Suppose to the contrary that there exist $\alpha_0, \alpha_1, \ldots, \alpha_k$ such that

$$t^{k+1} = \alpha_0 + \alpha_1 t + \cdots + \alpha_k t^k.$$

Appealing to basic Calculus, we differentiate both sides $k+1$ times and arrive at

$$(k+1)! = 0,$$

which is clearly false. Hence t^{k+1} cannot be written as a linear combination of $\{1, t, \ldots, t^k\}$ and therefore $P(k+1)$ is true.

Note that we did a proof by contradiction to show the induction step. ■

Example 2.17 *Let $\mathcal{F} = \mathbb{R}$ and $\mathcal{X} = \mathbb{R}^{2\times 3}$, the set of 2×3 matrices with real coefficients. Let $v^1 = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 0 & 0 \end{bmatrix}$, $v^2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$, $v^3 = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$, $v^4 = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}$. Show that $\{v^1, v^2\}$ is a linearly independent set and that $\{v^1, v^2, v^4\}$ is a linearly dependent set.*

Solution: To show independence, we must show that the only linear combination resulting in the zero vector in $\mathcal{X}$ is a trivial linear combination. Hence, we check

$$\alpha_1 v^1 + \alpha_2 v^2 = 0_{2\times 3} \iff \begin{bmatrix} \alpha_1 & 0 & 0 \\ 2\alpha_1 & 0 & 0 \end{bmatrix} + \begin{bmatrix} \alpha_2 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \iff \alpha_1 = \alpha_2 = 0.$$

Hence, $\{v^1, v^2\}$ is a linearly independent set.

For $\{v^1, v^2, v^4\}$, we form a linear combination and seek coefficients resulting in the zero vector,

$$\alpha_1 v^1 + \alpha_2 v^2 + \alpha_4 v^4 = 0_{2\times 3} \iff \begin{bmatrix} \alpha_1 & 0 & 0 \\ 2\alpha_1 & 0 & 0 \end{bmatrix} + \begin{bmatrix} \alpha_2 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ \alpha_4 & 0 & 0 \end{bmatrix} = \begin{bmatrix} \alpha_1 + \alpha_2 & 0 & 0 \\ 2\alpha_1 + \alpha_4 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

One then checks that $\alpha_1 = 1, \alpha_2 = -1, \alpha_4 = -2$ is a nontrivial solution. There are infinite number of other nontrivial solutions. We only needed to find one to show that $\{v^1, v^2, v^4\}$ is a linearly dependent set of vectors.

■

Example 2.18 *Let $\mathcal{F} = \mathbb{R}$ and $\mathcal{X} = \mathbb{R}^{2\times 3}$, the set of 2×3 matrices with real coefficients. Are the vectors*

$$A_1 = \begin{bmatrix} 1 & 0 & 4 \\ 3 & -1 & 2 \end{bmatrix}, \; A_2 = \begin{bmatrix} 4 & 1 & 0 \\ 6 & 0 & 6 \end{bmatrix}$$

linearly independent?

Solution: We form a linear combination of A_1 and A_2 and check for a nontrivial solution.

$$\alpha_1 A_1 + \alpha_2 A_2 = \begin{bmatrix} \alpha_1 + 4\alpha_2 & \alpha_2 & 4\alpha_1 \\ 3\alpha_1 + 6\alpha_2 & -\alpha_1 & 2\alpha_1 + 6\alpha_2 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \iff \begin{cases} \alpha_1 + 4\alpha_2 &= 0 \\ \alpha_2 &= 0 \\ 4\alpha_1 &= 0 \\ 3\alpha_1 + 6\alpha_2 &= 0 \\ -\alpha_1 &= 0 \\ 2\alpha_1 + 6\alpha_2 &= 0 \end{cases}$$

We wrote out all of the equations to emphasize that when you set a 2×3 matrix equal to the zero matrix, each of its entries must be zero. We only have two unknowns, so we could have gotten by with just noting that the second and fifth equations together imply that $\alpha_1 = \alpha_2 = 0$. We conclude that the set $\{A_1, A_2\}$ is linearly independent. ■

Remark 2.19 *The field $\mathcal{F}$ is important when determining whether a set is linearly independent or not. For example, let $\mathcal{X} = \mathbb{C}$ and $v^1 = 1$, $v^2 = j := \sqrt{-1}$. v^1 and v^2 are linearly independent when $\mathcal{F} = \mathbb{R}$. However, they are linearly dependent when $\mathcal{F} = \mathbb{C}$.*

Definition 2.20 *Let $\mathcal{S}$ be a subset of a vector space $(\mathcal{X}, \mathcal{F})$. The **span** of $\mathcal{S}$, denoted span$\{\mathcal{S}\}$, is the set of all linear combinations of elements of $\mathcal{S}$. That is,*

$$\text{span}\{\mathcal{S}\} := \{x \in \mathcal{X} \mid \exists n \geq 1, \alpha_1, \ldots, \alpha_n \in \mathcal{F}, v^1, \ldots, v^n \in \mathcal{S}, \text{ s.t. } x = \alpha_1 v^1 + \cdots + \alpha_n v^n\}.$$

Remark 2.21 *By construction, $\text{span}\{\mathcal{S}\}$ is closed under linear combinations. Hence, it is a subspace of $\mathcal{X}$.*

Example 2.22 *Let $\mathcal{F} = \mathbb{R}$ and $\mathcal{X} = \{f : \mathbb{R} \to \mathbb{R}\}$.*

(a) What is the span of $\mathcal{S} = \{1, t, t^2, \ldots\}$?

(b) Is $e^t \in \text{span}\{\mathcal{S}\}$?

Solution:

(a) $\text{span}\{\mathcal{S}\} = \text{span}\{t^k, k \geq 0\} = \mathbb{P}(t)$, the set of polynomials in t with real coefficients.

(b) $e^t \notin \text{span}\{\mathcal{S}\}$ because e^t is not a polynomial. Even though e^t has a Taylor series expansion, the number of components in that sum is infinite, while, by definition, a linear combination has to be finite. Is it possible to write e^t in a different way as a polynomial? The answer is no. From Calculus, we know that $\frac{d}{dt}e^t = e^t$. Moreover, because differentiation of a polynomial reduces its degree by one, there is no non-zero polynomial $p(t)$ that satisfies $\frac{d}{dt}p(t) = p(t)$.

■

2.4 Basis Vectors and Dimension

Definition 2.23 *A set of vectors $\mathcal{B}$ in $(\mathcal{X}, \mathcal{F})$ is a **basis** for $\mathcal{X}$ if*

(a) $\mathcal{B}$ is linearly independent.

(b) span$\{\mathcal{B}\} = \mathcal{X}$.

Example 2.24 *We provide several examples of bases.*

*(a) Consider $(\mathcal{F}^n, \mathcal{F})$ where $\mathcal{F}$ is $\mathbb{C}$, $\mathbb{R}$ or $\mathbb{Q}$. The set $\{e^1, e^2, \ldots, e^n\}$ is called the **natural basis**, where*

$$e^1 := \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, e^2 := \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}, \ldots, e^n := \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}.$$

The expression

$$\alpha_1 e^1 + \alpha_2 e^2 + \cdots + \alpha_n e^n = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{bmatrix}$$

immediately shows that $\{e^1, e^2, \ldots, e^n\}$ is linearly independent and its span is $\mathcal{F}^n$. Therefore, it is a basis.

(b) $\{je^1, je^2, \ldots, je^n\}$ *is also a basis for* $(\mathbb{C}^n, \mathbb{C})$.

(c) $\{v^1, v^2, \ldots, v^n\}$ *is also a basis for* $(\mathcal{F}^n, \mathcal{F})$*, where*

$$v^1 = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, v^2 = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 0 \end{bmatrix}, \ldots, v^n = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}.$$

You can use the relationships $e^1 = v^1, e^2 = v^2 - v^1, \ldots, e^n = v^n - (v^1 + \cdots + v^{n-1})$ *to show this.*

(d) $\{e^1, e^2, \ldots, e^n, je^1, je^2, \ldots, je^n\}$ *is a basis for* $(\mathbb{C}^n, \mathbb{R})$.

(e) The infinite set $\{1, t, \ldots, t^n, \ldots\}$ *is a basis for* $(\mathbb{P}(t), \mathbb{R})$*. We have already shown the set is linearly independent. It spans* $\mathbb{P}(t)$ *by the very definition of* $\mathbb{P}(t)$.

Example 2.25 *We provide two non-examples of bases.*

(a) $\{e^1, e^2, \ldots, e^n\}$ *is NOT a basis for* $(\mathbb{C}^n, \mathbb{R})$ *because* $\text{span}\{e^1, e^2, \ldots, e^n\} \neq \mathbb{C}^n$*. Indeed,* $je^1 \notin \text{span}\{e^1, e^2, \ldots, e^n\}$ *when the field is the real numbers!*

(b) The set $\{e^1, e^2, \ldots, e^n, je^1, je^2, \ldots, je^n\}$ *is not linearly independent in* $(\mathbb{C}^n, \mathbb{C})$ *and is therefore not a basis of* $(\mathbb{C}^n, \mathbb{C})$.

Definition 2.26 *The vector space* $(\mathcal{X}, \mathcal{F})$ *has* ***finite dimension*** $n > 0$ *if*

- *there exists a set with* n *linearly independent vectors, and*
- *any set with* $n + 1$ *or more vectors is linearly dependent.*

Definition 2.27 $(\mathcal{X}, \mathcal{F})$ *is* ***infinite dimensional*** *if for every* $n > 0$*, there is a linearly independent set with* n *or more elements in it.*

Remark 2.28 *Because subspaces are vectors spaces in their own right, the above definitions apply to assigning their dimensions. By* ***convention****, the subspace consisting of the zero vector,* $\{0\}$*, has dimension zero.*

Example 2.29 *We provide a few examples.*

(a) $\dim(\mathcal{F}^n, \mathcal{F}) = n$.

(b) $\dim(\mathbb{C}^n, \mathbb{R}) = 2n$.

(c) $\dim(\mathbb{P}(t), \mathbb{R}) = \infty$.

(d) $\dim(\mathbb{R}, \mathbb{Q}) = \infty$.

Remark 2.30 *It is common to define the dimension of a vector space (or subspace) as the* ***cardinality of a basis****. We provided bases for examples (a), (b), and (c) in Example 2.24. But what about (d)? For (d), an explicit basis cannot be written down. You can find proofs online, based on other types of arguments, that show* $\dim(\mathbb{R}, \mathbb{Q}) = \infty$*, such as* http://www2.math.ou.edu/~aroche/courses/LinAlg-Fall2011/solutions1.pdf*. The vector space* $(\mathbb{R}, \mathbb{Q})$ *does not have any particular importance in Robotics. It's just cool every once in a while to think about non-intuitive ideas, such as creating an infinite dimensional vector space where the vectors belong to a set that your intuition is screaming "these must be one dimensional objects"!*

Theorem 2.31 *Let* $(\mathcal{X}, \mathcal{F})$ *be an* n*-dimensional vector space (*always means n is finite*). Then, any set of* n *linearly independent vectors is a basis.*

Proof: Let $\{v^1, \ldots, v^n\}$ be a linearly independent set in $(\mathcal{X}, \mathcal{F})$. To show it is a basis, we need to show the "span" property, namely,

$$\forall x \in \mathcal{X}, \exists\, \alpha_1, \ldots, \alpha_n \in \mathcal{F} \text{ such that } x = \alpha_1 v^1 + \cdots + \alpha_n v^n.$$

Let $x \in \mathcal{X}$ be given. Because $(\mathcal{X}, \mathcal{F})$ is n-dimensional, the set $\{x, v^1, \ldots, v^n\}$ is a linearly dependent set, because, otherwise, $\dim(\mathcal{X}) > n$. Hence, $\exists\, \beta_0, \beta_1, \ldots, \beta_n \in \mathcal{F}$, NOT ALL ZERO, such that $\beta_0 x + \beta_1 v^1 + \cdots + \beta_n v^n = 0$.

Claim 2.32 $\beta_0 \neq 0$.

Proof: We do a proof by contradiction. Suppose that $\beta_0 = 0$. Then,

(a) At least one of $\beta_1, \ldots, \beta_n$ is non-zero, and

(b) $\beta_1 v^1 + \cdots + \beta_n v^n = 0$.

(a) and (b) imply that $\{v^1, \ldots, v^n\}$ is a linearly dependent set. However, by assumption, $\{v^1, \ldots, v^n\}$ is a basis and hence linearly independent. This is a contradiction. Hence, $\beta_0 = 0$ cannot hold. □

Because $\beta_0 \neq 0$, we complete the proof by

$$\begin{aligned} \beta_0 x &= -\beta_1 x^1 - \ldots - \beta_n v^n \\ &\Downarrow \\ x &= \left(\frac{-\beta_1}{\beta_0}\right) v^1 + \ldots + \left(\frac{-\beta_n}{\beta_0}\right) v^n \end{aligned}$$

and therefore, $\alpha_1 := \frac{-\beta_1}{\beta_0}, \ldots, \alpha_n := \frac{-\beta_n}{\beta_0}$ are the required coefficients in $\mathcal{F}$. ■

Proposition 2.33 *Let $(\mathcal{X}, \mathcal{F})$ be a vector space with basis $\{v^1, \ldots, v^n\}$ and let $x \in \mathcal{X}$. Then, there exist* unique *coefficients $\alpha_1, \ldots, \alpha_n$ such that*

$$x = \alpha_1 v^1 + \alpha_2 v^2 + \ldots + \alpha_n v^n.$$

Proof: Suppose x can also be written as $x = \beta_1 v^1 + \beta_2 v^2 + \ldots + \beta_n v^n$. We need to show: $\alpha_1 = \beta_1, \alpha_2 = \beta_2, \ldots, \alpha_n = \beta_n$. To do so, write

$$0 = x - x = (\alpha_1 - \beta_1) v^1 + \ldots + (\alpha_n - \beta_n) v^n.$$

By the linear independence of $\{v^1, \ldots, v^n\}$, we obtain that

$$\alpha_1 - \beta_1 = 0, \ldots, \alpha_n - \beta_n = 0.$$

Hence, $\alpha_1 = \beta_1, \ldots, \alpha_n = \beta_n$, that is, the coefficients are unique. ■

Proposition 2.34 *Let$(\mathcal{X}, \mathcal{F})$ be an n-dimensional vector space and let $\{v^1, \cdots, v^k\}$ be a linearly independent set with k strictly less than n. Then, $\exists\, v^{k+1} \in \mathcal{X}$ such that $\{v^1, \cdots, v^k, v^{k+1}\}$ is linearly independent.*

Proof: We use proof by contradiction. Suppose that $\{v^1, \cdots, v^k\}$ is a linearly independent and $k < n$, but no such v^{k+1} exists. Then, $\forall x \in \mathcal{X}$, $x \in \text{span}\{v^1, \cdots, v^k\}$, and therefore, $\mathcal{X} \subset \text{span}\{v^1, \ldots, v^k\}$. This in turn implies that $n = \dim(\mathcal{X}) \leq \dim(\text{span}\{v^1, \ldots, v^k\}) = k$, which contradicts $k < n$. Hence, there must exist $v^{k+1} \in \mathcal{X}$ such that $\{v^1, \ldots, v^k, v^{k+1}\}$ is linearly independent. ■

Corollary 2.35 *In a finite dimensional vector space, any linearly independent set can be completed to a basis. More precisely, let $\{v^1, \cdots, v^k\}$ be linearly independent, $n = \dim(\mathcal{X})$ and $k < n$. Then, $\exists v^{k+1}, \ldots, v^n$ such that $\{v^1, \cdots, v^k, v^{k+1}, \ldots, v^n\}$ is a basis for $\mathcal{X}$.*

Proof: Previous Proposition plus induction. ■

2.5 Representations of Vectors and the Change of Basis Matrix

Definition 2.36 *Let $(\mathcal{X}, \mathcal{F})$ be a vector space with basis $v := \{v^1, \ldots, v^n\}$ and write $x \in \mathcal{X}$ as a unique linear combination of the basis vectors, $x = \alpha_1 v^1 + \ldots + \alpha_n v^n$. Then*

$$[x]_v := \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{bmatrix} \in \mathcal{F}^n$$

*is **the representation** of x with respect to the basis v.*

Remark 2.37 *Just to be absolutely 100% clear, we note*

$$[x]_v := \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{bmatrix} \iff x = \alpha_1 v^1 + \alpha_2 v^2 + \cdots + \alpha_n v^n.$$

Once a basis is specified, you can work with vectors as if they were n-tuples. This can be handy when doing numerical computations.

Example 2.38 *We take $\mathcal{F} = \mathbb{R}$ and $\mathcal{X} = \mathbb{R}^{2\times 2}$. Compute the representation of*

$$x = \begin{bmatrix} 5 & 3 \\ 1 & 4 \end{bmatrix}$$

in each of the bases given below.

Basis 1: $v^1 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, v^2 = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, v^3 = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, v^4 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$

Basis 2: $w^1 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, w^2 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, w^3 = \begin{bmatrix} 0 & 1 \\ -1 & 1 \end{bmatrix}, w^4 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$

Solution:

Basis 1: This one can be done by inspection because the basis is so simple:

$$x = \begin{bmatrix} 5 & 3 \\ 1 & 4 \end{bmatrix} = 5v^1 + 3v^2 + 1v^3 + 4v^4 \iff [x]_w = \begin{bmatrix} 5 \\ 3 \\ 1 \\ 4 \end{bmatrix} \in \mathbb{R}^4.$$

Basis 2: We'll work this one out

$$\alpha_1 w^1 + \alpha_2 w^2 + \alpha_3 w^3 + \alpha_4 w^4 = \begin{bmatrix} \alpha_1 & \alpha_2 + \alpha_3 \\ \alpha_2 - \alpha_3 & \alpha_3 + \alpha_4 \end{bmatrix} = \begin{bmatrix} 5 & 3 \\ 1 & 4 \end{bmatrix}.$$

This gives us four equations in four unknowns, which we express in matrix form as

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{bmatrix} = \begin{bmatrix} 5 \\ 3 \\ 1 \\ 4 \end{bmatrix}.$$

The solution is, $\alpha_1 = 5, \alpha_2 = 2, \alpha_3 = 1, \alpha_4 = 3$. Therefore,

$$\begin{bmatrix} 5 & 3 \\ 1 & 4 \end{bmatrix} = 5w^1 + 2w^2 + 1w^3 + 3w^4 \iff [x]_w = \begin{bmatrix} 5 \\ 2 \\ 1 \\ 3 \end{bmatrix} \in \mathbb{R}^4.$$

■

Fact 2.39 *Representations of vectors in a finite dimensional vector space.*

1. Addition of vectors in $(\mathcal{X}, \mathcal{F}) \longleftrightarrow$ Addition of the representations in $(\mathcal{F}^n, \mathcal{F})$.

$$[x + y]_v = [x]_v + [y]_v$$

2. Scalar multiplication in $(\mathcal{X}, \mathcal{F}) \longleftrightarrow$ Scalar multiplication with the representations in $(\mathcal{F}^n, \mathcal{F})$.

$$[\alpha x]_v = \alpha [x]_v$$

3. Once a basis $v := \{v^1, \ldots, v^n\}$ is chosen , an n-dimensional vector space $(\mathcal{X}, \mathcal{F}) \overset{v}{\longleftrightarrow} (\mathcal{F}^n, \mathcal{F})$.

Question Let $\{u^1, \cdots, u^n\}$ and $\{\overline{u}^1, \cdots, \overline{u}^n\}$ be two bases for $(\mathcal{X}, \mathcal{F})$. Is there a relation between $[x]_u$ and $[x]_{\overline{u}}$? There is, via the **change of basis matrix**.

Theorem 2.40 *There exists an invertible matrix P, with coefficients in $\mathcal{F}$, such that $\forall x \in (\mathcal{X}, \mathcal{F})$, $[x]_{\overline{u}} = P[x]_u$, where, $P = \begin{bmatrix} P_1 & P_2 & \cdots & P_n \end{bmatrix}$ and its i^{th} column is given by $P_i := [u^i]_{\overline{u}} \in \mathcal{F}^n$, and $[u^i]_{\overline{u}}$ is the representation of u^i with respect to $\overline{u}$. Similarly, there exists an invertible matrix $\overline{P} = \begin{bmatrix} \overline{P}_1 & \overline{P}_2 & \cdots & \overline{P}_n \end{bmatrix}$ with $\overline{P}_i$=$[\overline{u}^i]_u$, the representation of $\overline{u}^i$ with respect to u, and $P \cdot \overline{P} = \overline{P} \cdot P = I$.*

Proof: We can express $x \in \mathcal{X}$ in terms of both bases, $x = \alpha_1 u^1 + \cdots + \alpha_n u^n = \overline{\alpha}_1 \overline{u}^1 + \cdots + \overline{\alpha}_n \overline{u}^n$, so that

$$\alpha := \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{bmatrix} := [x]_u \text{ and } \overline{\alpha} := \begin{bmatrix} \overline{\alpha}_1 \\ \overline{\alpha}_2 \\ \vdots \\ \overline{\alpha}_n \end{bmatrix} = [x]_{\overline{u}}$$

From the linearity of the representation operation,

$$\overline{\alpha} := [x]_{\overline{u}} = \Big[\sum_{i=1}^{n} \alpha_i u^i \Big]_{\overline{u}} = \sum_{i=1}^{n} \alpha_i [u^i]_{\overline{u}} = \sum_{i=1}^{n} \alpha_i P_i = P\alpha. \tag{2.1}$$

Therefore, $\overline{\alpha} := P\alpha = P[x]_u$. Similarly,

$$\alpha = [x]_u = \Big[\sum_{i=1}^{n} \overline{\alpha}_i \overline{u}^i \Big]_u = \sum_{i=1}^{n} \overline{\alpha}_i [\overline{u}^i]_u = \sum_{i=1}^{n} \overline{\alpha}_i \overline{P}_i = \overline{P}\overline{\alpha}, \tag{2.2}$$

yielding $\alpha = \overline{P}\overline{\alpha}$. Combining (2.1) and (2.2) gives $\alpha = \overline{P}P\alpha$ and $\overline{\alpha} = P\overline{P}\overline{\alpha}$. Because this holds for all x, and hence for all $\alpha =$ and $\overline{\alpha}$, we deduce $P\overline{P} = \overline{P}P = I$.

In conclusion, $\overline{P}$ is the inverse of P ($\overline{P} = P^{-1}$). ■

Example 2.41 *For $\mathcal{F} = \mathbb{R}$ and $\mathcal{X} = \mathbb{R}^{2\times 2}$, find the change of basis matrices.*

$$u = \left\{ \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \right\}$$
$$\overline{u} = \left\{ \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \right\}$$

Solution: We have following relations:

$$\alpha = P\overline{\alpha}, P_i = [u^i]_{\overline{u}}, \quad \overline{\alpha} = \overline{P}\alpha, \overline{P}_i = [\overline{u}^i]_u$$
$$\overline{P}^{-1} = P, P^{-1} = \overline{P}$$

Typically, one computes the easier of P and $\overline{P}$, and then computes the other by matrix inversion. For this example, we choose to compute $\overline{P}$ because the required representations can be computed by inspection.

$$\overline{P}_1 = [\overline{u}^1]_u = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \qquad \overline{P}_2 = [\overline{u}^2]_u = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \end{bmatrix}$$

$$\overline{P}_3 = [\overline{u}^3]_u = \begin{bmatrix} 0 \\ 1 \\ -1 \\ 0 \end{bmatrix} \qquad \overline{P}_4 = [\overline{u}^4]_u = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

Therefore, $\overline{P}=\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$ and $P = (\overline{P})^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & .5 & .5 & 0 \\ 0 & .5 & -.5 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$

What if we did it the other direction?

$$P_1 = [u^1]_{\overline{u}} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \leftrightarrow \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} = 1 \cdot \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} + 0 \cdot \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} + 0 \cdot \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} + 0 \cdot \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$$

$$P_2 = [u^2]_{\overline{u}} = \begin{bmatrix} 0 \\ .5 \\ .5 \\ 0 \end{bmatrix} \leftrightarrow \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} = 0 \cdot \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} + 0.5 \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} + .5 \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} + 0 \cdot \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$$

$$P_3 = [u^3]_{\overline{u}} = \begin{bmatrix} 0 \\ .5 \\ -.5 \\ 0 \end{bmatrix} \leftrightarrow \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} = 0 \cdot \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} + 0.5 \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} - .5 \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} + 0 \cdot \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$$

$$P_4 = [u^4]_{\overline{u}} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \leftrightarrow \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} = 0 \cdot \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} + 0 \cdot \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} + 0 \cdot \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} + 1 \cdot \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$$

Therefore, $P = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & .5 & .5 & 0 \\ 0 & .5 & -.5 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$ and $\overline{P} = P^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$.

■

2.6 Linear Operators and Matrix Representations

Definition 2.42 *Let $(\mathcal{X}, \mathcal{F})$ and $(\mathcal{Y}, \mathcal{F})$ be vector spaces. $\mathcal{L} : \mathcal{X} \to \mathcal{Y}$ is a **linear operator** if for all $x, z \in \mathcal{X}$, $\alpha, \beta \in \mathcal{F}$,*

$$\mathcal{L}(\alpha x + \beta z) = \alpha \mathcal{L}(x) + \beta \mathcal{L}(z).$$

Equivalently, you can check $\mathcal{L}(x + z) = \mathcal{L}(x) + \mathcal{L}(z)$ and $\mathcal{L}(\alpha x) = \alpha \mathcal{L}(x)$.

Example 2.43 *We provide one example based on matrices and one from Calculus.*

(a) Let A be an $n \times m$ matrix with coefficients in $\mathcal{F}$. Show that $\mathcal{L} : \mathcal{F}^m \to \mathcal{F}^n$ by $\mathcal{L}(x) = Ax$, is a linear operator.

(b) Let $\mathcal{X} = \{$ polynomials of degree ≤ 3 $\}, \mathcal{F} = \mathbb{R}, \mathcal{Y} = \mathcal{X}$. Show that $\mathcal{L} : \mathcal{X} \to \mathcal{Y}$ by $p \in \mathcal{X}$, $\mathcal{L}(p) := \frac{d}{dt}p(t)$, is a linear operator.

Solution:

(a) $\mathcal{L}(\alpha x + \beta z) := A(\alpha x + \beta z) = \alpha Ax + \beta Az =: \alpha \mathcal{L}(x) + \beta \mathcal{L}(z)$.

(b) From Calculus, we note that $\frac{d}{dt}(\alpha p(t) + \beta q(t)) = \alpha \frac{d}{dt}p(t) + \beta \frac{d}{dt}q(t)$, and hence the result.

■

Definition 2.44 *Let $(\mathcal{X}, \mathcal{F})$ and $(\mathcal{Y}, \mathcal{F})$ be finite dimensional vector spaces, and $\mathcal{L} : \mathcal{X} \to \mathcal{Y}$ be a linear operator. A **matrix representation** of $\mathcal{L}$ with respect to a basis $u := \{u^1, \ldots, u^m\}$ for $\mathcal{X}$ and $v := \{v^1, \ldots, v^n\}$ for $\mathcal{Y}$ is an $n \times m$ matrix A, with coefficients in $\mathcal{F}$, such that $\forall x \in \mathcal{X}$, $\left[\mathcal{L}(x)\right]_v = A \left[x\right]_u$.*

Theorem 2.45 *Let $(\mathcal{X}, \mathcal{F})$ and $(\mathcal{Y}, \mathcal{F})$ be finite dimensional vector spaces, $\mathcal{L} : \mathcal{X} \to \mathcal{Y}$ a linear operator, $u := \{u^1, \ldots, u^m\}$ a basis for $\mathcal{X}$ and $v := \{v^1, \ldots, v^n\}$ a basis for $\mathcal{Y}$, then $\mathcal{L}$ has a matrix representation $A = \begin{bmatrix} A_1 & \cdots & A_m \end{bmatrix}$, where the i^{th} column of A is given by*

$$A_i := \left[\mathcal{L}(u^i)\right]_v, \;\; 1 \leq i \leq m.$$

Proof: $x \in \mathcal{X}$, we write $x = \alpha_1 u^1 + \cdots + \alpha_m u^m$ so that its representation is

$$\left[x\right]_u = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_m \end{bmatrix} \in \mathcal{F}^m.$$

As in the theorem, we define

$$A_i = \left[\mathcal{L}(u^i)\right]_v, \;\; 1 \leq i \leq m.$$

Using linearity

$$\begin{aligned} \mathcal{L}(x) &= \mathcal{L}(\alpha_1 u^1 + \cdots + \alpha_m u^m) \\ &= \alpha_1 \mathcal{L}(u^1) + \cdots + \alpha_m \mathcal{L}(u^m). \end{aligned}$$

Hence, computing representations, we have

$$\begin{aligned} [\mathcal{L}(x)]_v &= [\alpha_1 \mathcal{L}(u^1) + \cdots + \alpha_m \mathcal{L}(u^m)]_v \\ &= \alpha_1 [\mathcal{L}(u^1)]_v + \cdots + \alpha_m [\mathcal{L}(u^m)]_v \\ &= \alpha_1 A_1 + \cdots + \alpha_m A_m \\ &= \begin{bmatrix} A_1 & A_2 & \cdots & A_m \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_m \end{bmatrix} \\ &= A \left[x\right]_u. \end{aligned}$$

Hence, $\left[\mathcal{L}(x)\right]_v = A \left[x\right]_u$. ■

Example 2.46 *$\mathcal{F} = \mathbb{R}, \mathcal{X} = P_3(t) = \{$polynomials of degree $\leq 3\}$, and $\mathcal{Y} = P_3(t)$. Use the same basis on $\mathcal{X}$ and $\mathcal{Y}$, namely $u := v := \{1, t, t^2, t^3\}$ and define $\mathcal{L} : \mathcal{X} \to \mathcal{Y}$ by $L(p) := \frac{d}{dt}p$. Find A, the matrix representation of $\mathcal{L}$, which will be a 4×4 real matrix.*

Solution: We compute A column by column, where $A = \begin{bmatrix} A_1 & A_2 & A_3 & A_4 \end{bmatrix}$. Then,

$$A_1 = \left[\mathcal{L}(1)\right]_{\{1,t,t^2,t^3\}} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \qquad A_2 = \left[\mathcal{L}(t)\right]_{\{1,t,t^2,t^3\}} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$$A_3 = \left[\mathcal{L}(t^2)\right]_{\{1,t,t^2,t^3\}} = \begin{bmatrix} 0 \\ 2 \\ 0 \\ 0 \end{bmatrix} \qquad A_4 = \left[\mathcal{L}(t^3)\right]_{\{1,t,t^2,t^3\}} = \begin{bmatrix} 0 \\ 0 \\ 3 \\ 0 \end{bmatrix}$$

and thus

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 3 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

To check that it makes sense, we let

$$p(t) = a_0 + a_1 t + a_2 t^2 + a_3 t^3$$

and note that

$$\big[p(t)\big]_{\{1,t,t^2,t^3\}} = \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{bmatrix}$$

$$A[p(t)]_{\{1,t,t^2,t^3\}} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 3 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} a_1 \\ 2a_2 \\ 3a_3 \\ 0 \end{bmatrix}$$

Does this correspond to differentiating the polynomial $p(t)$? From Calculus, we have that

$$\frac{d}{dt}p(t) = a_1 + 2a_2 t + 3a_3 t^2$$

$$[\frac{d}{dt}p(t)]_{\{1,t,t^2,t^3\}} = \begin{bmatrix} a_1 \\ 2a_2 \\ 3a_3 \\ 0 \end{bmatrix}$$

and thus, yes indeed,

$$A[p(t)]_{\{1,t,t^2,t^3\}} = [\frac{d}{dt}p(t)]_{\{1,t,t^2,t^3\}}.$$

■

Example 2.47 *Let $(\mathcal{X}, \mathcal{F})$ be a finite dimensional vector space, with bases $u := \{u^1, \ldots, u^n\}$ and $v := \{v^1, \ldots, v^n\}$. Let $\mathcal{L} : \mathcal{X} \to \mathcal{X}$ be the **identity operator**, that is, $\forall\, x \in \mathcal{X}$, $\mathcal{L}(x) := Id(x) := x$. Then the matrix representation of $\mathcal{L}$ is the change of basis matrix.*

Solution: For $x \in \mathcal{X}$, we write $x = \alpha_1 u^1 + \cdots + \alpha_n u^n$ so that its representation is

$$\big[x\big]_u = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{bmatrix} \in \mathcal{F}^m.$$

As in Theorem 2.45, we define

$$A_i = \big[\mathcal{L}(u^i)\big]_v, \;\; 1 \le i \le n.$$

Using linearity

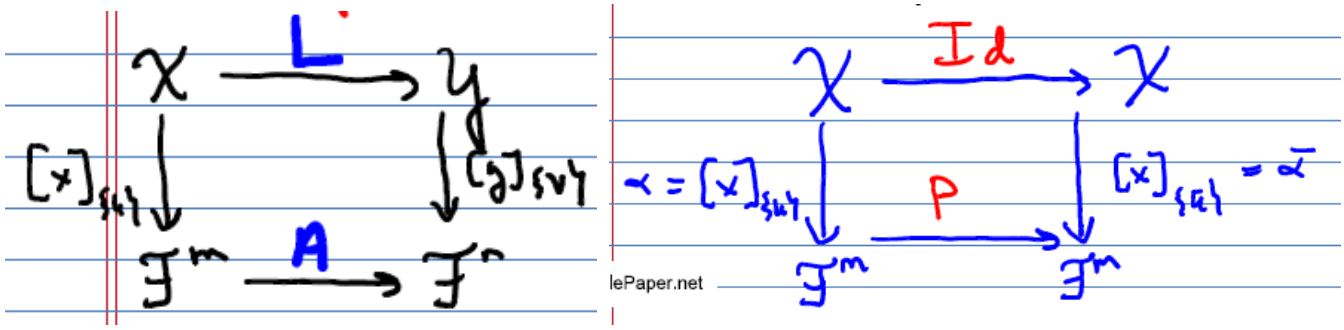
$$\begin{aligned} [\mathcal{L}(x)]_v &= \alpha_1[\mathcal{L}(u^1)]_v + \cdots + \alpha_n[\mathcal{L}(u^m)]_v \\ &= \alpha_1 A_1 + \cdots + \alpha_n A_n \\ &= \begin{bmatrix} A_1 & A_2 & \cdots & A_n \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{bmatrix} \\ &= A\big[x\big]_u. \end{aligned}$$

Hence, $\forall\, x \in \mathcal{X}$,

$$\begin{array}{c} \big[\mathcal{L}(x)\big]_v = A\big[x\big]_u \\ \Updownarrow \\ \big[Id(x)\big]_v = A\big[x\big]_u \\ \Updownarrow \\ \big[x\big]_v = A\big[x\big]_u \\ \Updownarrow \\ A = P \text{ the change of basis matrix.} \end{array}$$

■

Remark 2.48 *A shorter proof is just to note that the i-th column of the change of basis matrix satisfies $P_i := [u_i]_v$, and that the i-th column of the matrix representation of $\mathcal{L}$ satisfies $A_i := [\mathcal{L}(u_i)]_v = [u_i]_v$ when $\mathcal{L}$ is the identity operator. The relationship is summarized in Fig. 2.1.*



(a) (b)

Figure 2.1: Diagram chasing. Commuting diagrams used to illustrate the matrix representation of a linear operator and the change of basis matrix. When $\mathcal{X} = \mathcal{Y}$ and $\mathcal{L} = Id$, they become one and the same. Hence, there is really only one idea to remember.

Example 2.49 *Let $(\mathcal{X}, \mathcal{F}) = \left(\mathbb{R}^2, \mathbb{R}\right)$, and define $L : \mathbb{R}^2 \to \mathbb{R}^2$ by $L(e_1) = 3e_1 + 4e_2$, $L(e_2) = -e_1 + 6e_2$, where $e_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and $e_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ are the canonical basis elements.*

(a) What is the matrix representation of L with respect to $\{e_1, e_2\}$?

(b) What is the matrix representation of L with respect to $\left\{v^1, v^2\right\}$ where $v^1 = e_1 + e_2$, $v^2 = 3e_1 - 4e_2$?

Solution:

(a) Let A =matrix representation of L. Then the i^{th} column of $A = [L(e_i)]_{\{e_1, e_2\}}$. Hence,

$$[L(e_1)]_{\{e_1,e_2\}} = \begin{bmatrix} 3 \\ 4 \end{bmatrix}$$
$$[L(e_2)]_{\{e_1,e_2\}} = \begin{bmatrix} -1 \\ 6 \end{bmatrix}$$
$$\implies A = \begin{bmatrix} 3 & -1 \\ 4 & 6 \end{bmatrix}.$$

(b) Let P be the change of coordinates from $\{e_1, e_2\}$ to $\left\{v^1, v^2\right\}$, and $\overline{P}$ be the change of coordinates from $\left\{v^1, v^2\right\}$ to $\{e_1, e_2\}$. Note that the i^{th} column of $\overline{P}$ is just the representation of v^i in $\{e_1, e_2\}$. That is,

$$\overline{P} = \begin{bmatrix} 1 & 3 \\ 1 & -4 \end{bmatrix}.$$

Recall that $\overline{P} = P^{-1}$, so

$$P = \left(\overline{P}\right)^{-1} = \frac{-1}{7}\begin{bmatrix} -4 & -3 \\ -1 & 1 \end{bmatrix} = \frac{1}{7}\begin{bmatrix} 4 & 3 \\ 1 & -1 \end{bmatrix}.$$

Therefore, if $\overline{A}$ is the representation of L in $\left\{v^1, v^2\right\}$, then

$$\overline{A} = PAP^{-1} = \frac{1}{7}\begin{bmatrix} 4 & 3 \\ 1 & -1 \end{bmatrix}\begin{bmatrix} 3 & -1 \\ 4 & 6 \end{bmatrix}\begin{bmatrix} 1 & 3 \\ 1 & -4 \end{bmatrix} = \frac{1}{7}\begin{bmatrix} -38 & 16 \\ -8 & 25 \end{bmatrix}.$$

Note: $\overline{P}$ was readily available, not P, as you may have guessed!! Just to check, let's do the same thing the "long way":

$$\begin{aligned} L\left(v^1\right) &= L\left(e_1+e_2\right) \\ &= L\left(e_1\right)+L\left(e_2\right) \\ &= \left(3e_1+4e_2\right)+\left(-e_1+6e_2\right) \\ &= 2e_1+10e_2 \\ L\left(v^2\right) &= L\left(3e_1-4e_2\right) \\ &= 3L\left(e_1\right)-4L\left(e_2\right) \\ &= 3\left(3e_1+4e_2\right)-4\left(-e_1+6e_2\right) \\ & 13e_1-12e_2 \end{aligned}$$

$\left[L\left(v^1\right)\right]_{\{v^1,v^2\}} =?$ To find it, write

$$\begin{bmatrix} 2 \\ 10 \end{bmatrix} = \overline{a}_{11} \underbrace{\begin{bmatrix} 1 \\ 1 \end{bmatrix}}_{v^1} + \overline{a}_{21} \underbrace{\begin{bmatrix} 3 \\ -4 \end{bmatrix}}_{v^2} = \begin{bmatrix} 1 & 3 \\ 1 & -4 \end{bmatrix} \begin{bmatrix} \overline{a}_{11} \\ \overline{a}_{12} \end{bmatrix}$$

$$\implies \begin{bmatrix} \overline{a}_{11} \\ \overline{a}_{12} \end{bmatrix} = \frac{1}{7} \begin{bmatrix} 38 \\ -18 \end{bmatrix}$$

Similarly

$$\begin{bmatrix} 13 \\ -12 \end{bmatrix} = \overline{a}_{12} \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \overline{a}_{22} \begin{bmatrix} 3 \\ -4 \end{bmatrix}$$

$$\implies \begin{bmatrix} \overline{a}_{12} \\ \overline{a}_{22} \end{bmatrix} = \frac{1}{7} \begin{bmatrix} 16 \\ 25 \end{bmatrix}$$

and hence

$$\overline{A} = \frac{1}{7} \begin{bmatrix} 38 & 16 \\ -8 & 25 \end{bmatrix}.$$

■

2.7 Eigenvalues, Eigenvectors, and Diagonalization

Definition 2.50 *Let A be an $n \times n$ matrix with complex coefficients. A scalar $\lambda \in \mathbb{C}$ is an **eigenvalue** (e-value) of A, if there exists a non-zero vector $v \in \mathbb{C}^n$ such that $Av = \lambda v$. Any such vector v is called an **eigenvector** (e-vector) associated with λ.*

Eigenvectors are not unique because if $Av = \lambda v$, then for all $\alpha \neq 0$, $A(\alpha v) = \lambda(\alpha v)$, and thus αv is also an e-vector. To find eigenvalues, we need to know conditions under which $\exists v \neq 0$ such that $Av = \lambda v$.

$$\exists v \neq 0 \text{ s.t. } Av = \lambda v \iff \exists v \neq 0 \text{ s.t. } (\lambda I - A)v = 0 \iff \det(\lambda I - A) = 0$$

Example 2.51 *Find the e-values and e-vectors for $A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$.*

Solution: $A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \implies \det(\lambda I - A) = \lambda^2 + 1 = 0$. Therefore, the eigenvalues are $\lambda_1 = j, \lambda_2 = -j$. To find eigenvectors, we need to solve $(A - \lambda_i I)v^i = 0$. The eigenvectors are

$$v^1 = \begin{bmatrix} 1 \\ j \end{bmatrix}, v^2 = \begin{bmatrix} 1 \\ -j \end{bmatrix}.$$

Note that both eigenvalues and eigenvectors are complex conjugate pairs. ■

Definition 2.52 *$\Delta(\lambda) := \det(\lambda I - A)$ is called the **characteristic polynomial**. $\Delta(\lambda) = 0$ is called the **characteristic equation**. By the Fundamental Theorem of Algebra, $\Delta(\lambda)$ can be factored as*

$$\Delta(\lambda) = (\lambda - \lambda_1)^{m_1}(\lambda - \lambda_2)^{m_2} \cdots (\lambda - \lambda_p)^{m_p}$$

*where $\lambda_1, \cdots, \lambda_p$ are the distinct eigenvalues (roots), m_i is the **algebraic multiplicity** of λ_i, and $m_1 + m_2 + \cdots + m_p = n$. The **geometric multiplicity** of λi is defined as $\eta_i := \dim \text{null}(A - \lambda_i I)$.*

Theorem 2.53 *Let A be an $n \times n$ matrix with coefficients in $\mathbb{R}$ or $\mathbb{C}$. If the e-values $\{\lambda_1, \ldots, \lambda_n\}$ are distinct, that is, $\lambda_i \neq \lambda_j$ for all $1 \leq i \neq j \leq n$, then the e-vectors $\{v^1, \ldots, v^n\}$ are linearly independent in ($\mathbb{C}^n$,$\mathbb{C}$).*

Remark 2.54 *Restatement of the theorem: If the e-values $\{\lambda_1, \ldots, \lambda_n\}$ are distinct then $\{v^1, \ldots, v^n\}$ is a basis for ($\mathbb{C}^n$,$\mathbb{C}$).*

Proof: We prove the contrapositive: if $\{v^1, \ldots, v^n\}$ is linearly dependent then there is a repeated e-value ($\lambda_i = \lambda_j$ for some $i \neq j$).

$\{v^1, \ldots, v^n\}$ linearly dependent $\implies \exists\, \alpha_1, \ldots, \alpha_n \in \mathbb{C}$, not all zero, such that

$$\alpha_1 v^1 + \cdots + \alpha_n v^n = 0. \tag{2.3}$$

Without loss of generality, we can suppose $\alpha_1 \neq 0$. (that is, we can always reorder of e-values and e-vectors so that the first coefficient α_1 is nonzero).

For all $1 \leq i \leq n$ and $1 \leq j \leq n$, because v^i is an e-vector,

$$(A - \lambda_j I)v^i = Av^i - \lambda_j v^i = \lambda_i v^i - \lambda_j v^i = (\lambda_i - \lambda_j)v^i.$$

Using this fact, it is an easy exercise to show

$$(A - \lambda_2 I)(A - \lambda_3 I) \cdots (A - \lambda_n I)v^i = (\lambda_i - \lambda_2)(\lambda_i - \lambda_3) \cdots (\lambda_i - \lambda_n)v^i, \text{ for } 1 \leq i \leq n. \tag{2.4}$$

Plugging in now for i yields,

for $i = 1$

$$(A - \lambda_2 I)(A - \lambda_3 I) \cdots (A - \lambda_n I)v^1 = (\lambda_1 - \lambda_2)(\lambda_1 - \lambda_3) \cdots (\lambda_1 - \lambda_n)v^1;$$

for $i = 2$

$$(A - \lambda_2 I)(A - \lambda_3 I) \cdots (A - \lambda_n I)v^2 = (\lambda_2 - \lambda_2)(\lambda_2 - \lambda_3) \cdots (\lambda_2 - \lambda_n)v^2 = 0 \text{ because } (\lambda_2 - \lambda_2) = 0;$$

$$\vdots$$

for $i = n$

$$(A - \lambda_2 I)(A - \lambda_3 I) \cdots (A - \lambda_n I)v^n = (\lambda_n - \lambda_2)(\lambda_n - \lambda_3) \cdots (\lambda_n - \lambda_n)v^2 = 0 \text{ because } (\lambda_n - \lambda_n) = 0.$$

Combining the above with (2.3), we obtain

$$\begin{aligned} 0 &= (A - \lambda_2 I)(A - \lambda_3 I) \cdots (A - \lambda_n I)(\alpha_1 v^1 + \cdots + \alpha_n v^n) \\ &= \alpha_1(\lambda_1 - \lambda_2)(\lambda_1 - \lambda_3) \cdots (\lambda_1 - \lambda_n)v^1 \end{aligned}$$

We know $\alpha_1 \neq 0$, as stated above, and $v^1 \neq 0$, by definition of e-vectors. Therefore,

$$0 = (\lambda_1 - \lambda_2)(\lambda_1 - \lambda_3) \cdots (\lambda_1 - \lambda_n),$$

and hence, at least one of the terms $(\lambda_1 - \lambda_k)$, $2 \leq k \leq n$ must be zero. Therefore, there is a repeated e-value $\lambda_1 = \lambda_k$ for some $2 \leq k \leq n$. ■

2.8 A Few Additional Properties of Matrices

Definition 2.55 *Two $n \times n$ matrices A and B are **similar** if there exists an invertible $n \times n$ matrix P such that $B = P \cdot A \cdot P^{-1}$. P is called a **similarity** matrix.*

Definition 2.56 *An $n \times n$ matrix A is said to have a **full set of e-vectors** if there exists a basis $\{v^1, v^2, \ldots, v^n\}$ of $(\mathbb{C}^n, \mathbb{C})$ such $Av^i = \lambda_i v^i$, $1 \leq i \leq n$.*

Theorem 2.57 *An $n \times n$ matrix A has a full set of e-vectors if, and only if, it is similar to a diagonal matrix. And when this happens, the entries on the diagonal matrix are e-values of A.*

Proof: We assume that $\{v^1, \ldots, v^n\}$ is a basis for $(\mathbb{C}^n, \mathbb{C})$ and that for $1 \leq i \leq n$, $Av^i = \lambda_i v^i$. Define two $n \times n$ matrices

$$\begin{aligned} M &:= \begin{bmatrix} v^1 & v^2 & \cdots & v^n \end{bmatrix} \\ \Lambda &:= \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n). \end{aligned}$$

Then

$$\begin{aligned} A \cdot M &:= \begin{bmatrix} Av^1 & Av^2 & \cdots & Av^n \end{bmatrix} \\ &= \begin{bmatrix} \lambda_1 v^1 & \lambda_2 v^2 & \cdots & \lambda_n v^n \end{bmatrix} \\ &= M \cdot \Lambda. \end{aligned}$$

We'll leave as an Exercise,

$$M\alpha = \begin{bmatrix} v^1 & v^2 & \cdots & v^n \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{bmatrix} = \alpha_1 v^2 + \alpha_2 v^2 + \cdots + \alpha_n v^n,$$

and hence M is invertible if, and only if, $\{v^1, \ldots, v^n\}$ is linearly independent. Therefore we have

$$A = M\Lambda M^{-1} \text{ and } \Lambda = M^{-1} A M,$$

proving that $\{v^1, \ldots, v^n\}$ is a basis implies A is similar to a diagonal matrix.

The other direction is straightforward and left to the reader. You need to recognize the columns of the "similarity matrix" as being e-vectors of A. ■

Fact 2.58 *If A and B are similar, they have the same e-values. Moreover, the e-values have the same algebraic and geometric multiplicities.*

Definition 2.59 *Let A be an $n \times m$ matrix with coefficients in $\mathbb{R}$ or $\mathbb{C}$. The **rank** of A is equal to the number of linearly independent columns.*

Fact 2.60 *If M is square, then $\operatorname{rank}(M)$ equals the number of non-zero e-values.*

Fact 2.61 *For an $n \times m$ real matrix A, $\operatorname{rank}(A) = \operatorname{rank}(A^\top A) = \operatorname{rank}(AA^\top) = \operatorname{rank}(A^\top)$. Hence, $A^\top A$ and $AA^\top$ have the same number of non-zero e-values. For a proof, see Chapter 10 of the ROB 101 textbook and see Lemma 2.63 below.*

Corollary 2.62

- *# of linearly independent rows = # of linearly independent columns.*
- $\operatorname{rank}(A) \leq \min(n, m)$

Lemma 2.63 *Suppose that A is a real $n \times m$ matrix. Then,*

- *If λ is a non-zero e-value of $\left(A^\top A\right)$ with e-vector v, then λ is also an e-value of $\left(AA^\top\right)$ with e-vector Av.*
- *If λ is a non-zero e-value of $\left(AA^\top\right)$ with e-vector v, then λ is also an e-value of $\left(A^\top A\right)$ with e-vector $A^\top v$.*

Proof: We only prove the first statement. Suppose that $\left(A^\top A\right) v = \lambda v$ and both λ and v are non-zero. Then $Av \neq 0$. Next, we form

$$\left(AA^\top\right)(Av) = A\left(A^\top A\right) v = A(\lambda v) = \lambda(Av),$$

and thus λ is an e-value of $\left(AA^\top\right)$ with e-vector Av.

■

Corollary 2.64 *$AA^\top$ and $A^\top A$ have the same non-zero e-values. Because they have different sizes, they may have different number of zero e-values.*

Definition 2.65 ***(Trace of a Matrix)*** *Let C be an $n \times n$ matrix. Then $\mathrm{tr}(C) := \sum_{i-1}^{n} C_{ii}$.*

Exercise 2.66 *Suppose that A is $n \times m$ and B is $m \times n$. Then*

$$\mathrm{tr}(A \cdot B) = \mathrm{tr}(B \cdot A).$$

Fact 2.67 ***(A lesser known way doing matrix multiplication, the outer product formula.)*** *Suppose that A is $n \times k$ and B is $k \times m$ so that the two matrices are compatible for matrix multiplication. Then*

$$A \cdot B = \sum_{i=1}^{k} a_i^{\mathrm{col}} \cdot b_i^{\mathrm{row}},$$

the "sum of the columns of A multiplied by the rows of B". A more precise way to say it would be "the sum over i of the i-th column of A times the i-th row of B."

Why: To see why this is true, let's first consider two 2×2 matrices A and B, where

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \text{ and } B = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix}.$$

Then, using the standard rows of A times the columns of B formulation of matrix multiplication yields

$$\begin{aligned}
A \cdot B :=& \begin{bmatrix} a_1^{\mathrm{row}} \cdot b_1^{\mathrm{col}} & a_1^{\mathrm{row}} \cdot b_2^{\mathrm{col}} \\ a_2^{\mathrm{row}} \cdot b_1^{\mathrm{col}} & a_2^{\mathrm{row}} \cdot b_2^{\mathrm{col}} \end{bmatrix} \\
=& \begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} \end{bmatrix} \text{ (next, we take the sum outside the matrix)} \\
=& \begin{bmatrix} a_{11}b_{11} & a_{11}b_{12} \\ a_{21}b_{11} & a_{21}b_{12} \end{bmatrix} + \begin{bmatrix} a_{12}b_{21} & a_{12}b_{22} \\ a_{22}b_{21} & a_{22}b_{22} \end{bmatrix} \text{ (next, we recognize each term)} \\
=& \begin{bmatrix} a_{11} \\ a_{21} \end{bmatrix} \cdot \begin{bmatrix} b_{11} & b_{12} \end{bmatrix} + \begin{bmatrix} a_{12} \\ a_{22} \end{bmatrix} \cdot \begin{bmatrix} b_{21} & b_{22} \end{bmatrix} \\
=& a_1^{\mathrm{col}} \cdot b_1^{\mathrm{row}} + a_2^{\mathrm{col}} \cdot b_2^{\mathrm{row}}
\end{aligned}$$

In a similar manner, we can treat the general case,

$$
\begin{aligned}
A \cdot B := & \begin{bmatrix} \boxed{a_{11} \;\; a_{12} \;\; \cdots \;\; a_{1k}} \\ \boxed{a_{21} \;\; a_{22} \;\; \cdots \;\; a_{2k}} \\ \vdots \\ \boxed{a_{n1} \;\; a_{n2} \;\; \cdots \;\; a_{nk}} \end{bmatrix} \cdot \begin{bmatrix} \boxed{\begin{matrix} b_{11} \\ b_{21} \\ \vdots \\ b_{k1} \end{matrix}} & \boxed{\begin{matrix} b_{12} \\ b_{22} \\ \vdots \\ b_{k2} \end{matrix}} & \cdots & \boxed{\begin{matrix} b_{1m} \\ b_{2m} \\ \vdots \\ b_{km} \end{matrix}} \end{bmatrix} \\
= & \begin{bmatrix} \sum_{i=1}^{k} a_{1i}b_{i1} & \sum_{i=1}^{k} a_{1i}b_{i2} & \cdots & \sum_{i=1}^{k} a_{1i}b_{im} \\ \sum_{i=1}^{k} a_{2i}b_{i1} & \sum_{i=1}^{k} a_{2i}b_{i2} & \cdots & \sum_{i=1}^{k} a_{2i}b_{im} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{k} a_{ni}b_{i1} & \sum_{i=1}^{k} a_{ni}b_{i2} & \cdots & \sum_{i=1}^{k} a_{ni}b_{im} \end{bmatrix} \text{ (next, we pull the sum outside the matrix)} \\
= & \sum_{i=1}^{k} \begin{bmatrix} a_{1i}b_{i1} & a_{1i}b_{i2} & \cdots & a_{1i}b_{im} \\ a_{2i}b_{i1} & a_{2i}b_{i2} & \cdots & a_{2i}b_{im} \\ \vdots & \vdots & \ddots & \vdots \\ a_{ni}b_{i1} & a_{ni}b_{i2} & \cdots & a_{ni}b_{im} \end{bmatrix} \text{ (next, we recognize what this is)} \\
= & \sum_{i=1}^{k} a_i^{\text{col}} \cdot b_i^{\text{row}} \\
= & \begin{bmatrix} \boxed{\begin{matrix} a_{11} \\ a_{21} \\ \vdots \\ a_{n1} \end{matrix}} & \boxed{\begin{matrix} a_{12} \\ a_{22} \\ \vdots \\ a_{n2} \end{matrix}} & \cdots & \boxed{\begin{matrix} a_{1k} \\ a_{2k} \\ \vdots \\ a_{nk} \end{matrix}} \end{bmatrix} \cdot \begin{bmatrix} \boxed{b_{11} \;\; b_{12} \;\; \cdots \;\; b_{1m}} \\ \boxed{b_{21} \;\; b_{22} \;\; \cdots \;\; b_{2m}} \\ \vdots \\ \boxed{b_{k1} \;\; b_{k2} \;\; \cdots \;\; b_{km}} \end{bmatrix}.
\end{aligned}
\tag{2.5}
$$

■

Fact 2.68 ***(Matrix Inversion Lemma)*** *Suppose that A, B, C and D are compatible[2] matrices. If A, C, and $(C^{-1} + DA^{-1}B)$ are each square and invertible, then $A + BCD$ is invertible and*

$$(A + BCD)^{-1} = A^{-1} - A^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1}$$

Remark 2.69 *In many important applications, the inverse of A may be already known or easy to compute. Here is a made up example, but it gets the point across: By hand, evaluate $(A + BCD)^{-1}$ when*

$$A = \text{diag}([1,\ 0.5,\ 0.5,\ 1,\ 0.5]),\ B = \begin{bmatrix} 1 \\ 0 \\ 2 \\ 0 \\ 3 \end{bmatrix}, C = 0.2, D = B^{\top}$$

[2]The sizes are such the matrix products and sum in $A + BCD$ make sense.

Chapter 3

Abstract Inner Product Spaces for a Clear Vision of Deterministic Least Squares Problems

Learning Objectives

- Learn that a norm is a general means of measuring the length of a vector.
- Learn how the notions of an inner product and an inner product space generalize the dot product on $\mathbb{R}^n$ to a wide range of useful settings.
- Apply the tools of inner product spaces to best approximation problems.

Outcomes

- Norms and normed spaces as settings where best approximation problems can be posed.
- Learn that orthogonality, the Pythagorean Theorem, and Gram-Schmidt work just as well in abstract inner product spaces as they do in $\mathbb{R}^n$.
- The "normal equations" provide a systematic means to compute solutions to best approximation problems in any finite-dimensional inner product space.
- While we are on the subject of "orthogonality", we'll dive into real symmetric matrices and see that their eigenvectors have the amazing property that they can be selected to form an orthonormal basis for $\mathbb{R}^n$.
- We solve two very important least squares problems for overdetermined systems, underdetermined systems, and provide a recursive solution for the case of overdetermined systems. The latter is meant as a preview of the Kalman Filter (KF).

3.1 Preliminaries on Norms and Normed Spaces

Definition 3.1 *Let $(\mathcal{X}, \mathcal{F})$ be a vector space where the field $\mathcal{F}$ is either $\mathbb{R}$ or $\mathbb{C}$. A function $\|\cdot\|: \mathcal{X} \to \mathbb{R}$ is a **norm** if it satisfies*

(a) $\|x\| \geq 0$, $\forall x \in \mathcal{X}$ *and* $\|x\| = 0 \iff x = 0$.

(b) Triangle inequality: $\|x + y\| \leq \|x\| + \|y\|, \forall x, y \in \mathcal{X}$

(c) $\|\alpha x\| = |\alpha| \cdot \|x\|$, $\forall x \in \mathcal{X}, \alpha \in \mathcal{F}$, $\begin{cases} \text{If } \alpha \in \mathbb{R}, |\alpha| \text{ means the absolute value} \\ \text{If } \alpha \in \mathbb{C}, |\alpha| \text{ means the magnitude} \end{cases}$.

Example 3.2

(a) $\mathcal{F} := \mathbb{R}$ *or* $\mathbb{C}$, $\mathcal{X} := \mathcal{F}^n$.

- $\|x\|_2 := \left(\sum_{i=1}^{n} |x_i|^2 \right)^{\frac{1}{2}}$, *Euclidean norm or 2-norm*
- $\|x\|_p := \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}}, 1 \leq p < \infty$, *p-norm*
- $\|x\|_\infty := \max_{1 \leq i \leq n} |x_i|$, *max-norm, sup-norm,* ∞*-norm*

(b) $\mathcal{F} := \mathbb{R}$, $\mathcal{D} \subset \mathbb{R}$, $\mathcal{D} := [a, b]$, $a < b < \infty$, *and* $\mathcal{X} := \{f : \mathcal{D} \to \mathbb{R} \mid f \text{ is continuous}\}$.

- $\|f\|_2 := (\int_a^b |f(t)|^2 \, \mathrm{d}t)^{\frac{1}{2}}$
- $\|f\|_p := (\int_a^b |f(t)|^p \, \mathrm{d}t)^{\frac{1}{p}}$, $1 \leq p < \infty$
- $\|f\|_\infty := \max_{a \leq t \leq b} |f(t)|$, *which is also written* $\|f\|_\infty := \sup_{a \leq t \leq b} |f(t)|$

Definition 3.3 $(\mathcal{X}, \mathcal{F}, \|\cdot\|)$ *is called a **normed space**.*

Notation and Vocabulary 3.4 ***(Notions of Distance and Best Approximation)*** *For $x, y \in \mathcal{X}$,*

- $d(x, y) := \|x - y\|$ *is called the **distance from** x to y. We note that $d(x, y) = d(y, x)$.*
- ***Distance to a set:*** *Let $S \subset \mathcal{X}$ be a subset.*
$$d(x, S) := \inf_{y \in S} \|x - y\|$$
- *If $\exists x^* \in S$ such that $d(x, S) = \|x - x^*\|$, then x^* is called **a best approximation of x by elements of S.** We sometimes write $\widehat{x}$ for x^* because we are really thinking of the solution as an approximation.*

Question 3.5 ***(Important questions for this chapter)***

(a) When does a best approximate x^* exist?

(b) How to characterize (compute) x^* such that $\|x - x^*\| = d(x, S)$, $x^* \in S$?

(c) If a solution exists, is it unique?

Notation 3.6 ***(arg min)*** *When x^* exists and is unique, we write $x^* := \arg\min_{y \in S} \|x - y\|$ or $\widehat{x} := \arg\min_{y \in S} \|x - y\|$. It means that x^* is the **argument of the minimum function that achieves the minimum value.***

Remark 3.7

*(a) $e := x - y$ is the **error** when x is approximated by $y \in S$.*

(b) $\inf_{y \in S} \|x - y\|$ is the smallest value that of the norm of the error that can be achieved over all elements $y \in S$.

(c) When $\exists \widetilde{y} \in S$ such that $\|x - \widetilde{y}\| = \inf_{y \in S} \|x - y\|$, we can write $\|x - \widetilde{y}\| = \min_{y \in S} \|x - y\|$. While it may seem natural to denote the best y by y^*, we typically denote it by x^* because we are trying to best approximate x by elements of S.

(d) $x^* := \arg\min_{y \in S} \|x - y\|$ *is the value of* $y \in S$ *that best achieves the approximation of* x *in the sense of minimizing the norm.*

(e) $x^* := \arg\min_{y \in S} \|x - y\|$ *only makes sense when* $\inf_{y \in S} \|x - y\| = \min_{y \in S} \|x - y\|$*, that is, the infimum is actually achieved.*

(f) *Furthermore, if you really want to be careful, you should also check that there is a unique minimum. Otherwise, the correct notation is* $x^* \in \arg\min_{y \in S} \|x - y\|$*. In engineering publications, you rarely see this much care being taken. C'ést la vie, baby !*

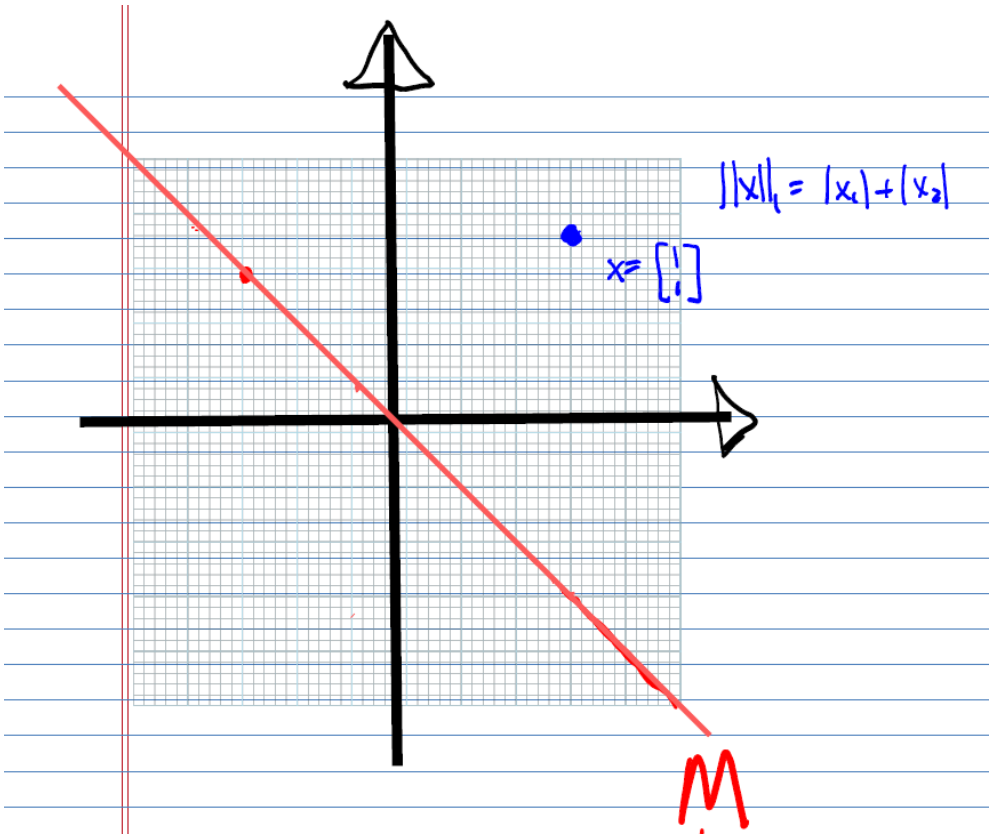


Figure 3.1: In HW, we'll introduce the concept of "strict" norms. The $||\bullet||_1$ is not strict and consequently, the minimum distance problem (or best approximation problem) does not have a unique answer.

Remark 3.8 *Figure 3.1 shows an example having* ***nonunique solutions****. Indeed, the set*

$$S := \arg\min_{y \in M} \|x - y\|_1$$

contains an uncountable number of elements. Specifically, every element of

$$S = \left\{ \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \mid x_2 = -x_1, |x_1| \leq 1 \right\}$$

is a minimizing vector for $x = [1 \quad 1]^\top$*. That is,*

$$\forall\, \widehat{x} \in S,\ ||x - \widehat{x}||_1 = 2 = \inf_{m \in M} ||x - m||_1.$$

To see this, you need to show that, $\forall\, \widehat{x} \in S$*,*

$$\begin{aligned} ||x - \widehat{x}||_1 &= |1 - \hat{x}_1| + |1 + \hat{x}_1|,\ -1 \leq \hat{x}_1 \leq 1 \\ &= (1 - \hat{x}_1) + (1 + \hat{x}_1) \\ &= 2 \end{aligned}$$

Exercise 3.9 *Consider $(\mathbb{R}^n, \mathbb{R})$ with the p-norm. Show that for all $x \in \mathbb{R}^n$,*

$$\lim_{p\to\infty} ||x||_p = ||x||_{\max}.$$

Hints: (i) Prove the result when $||x||_{\max} = 1$. (ii) When $x \neq 0$, consider $\bar{x} = x/||x||_{\max}$. (iii) For any non-negative real number a, $\lim_{p\to\infty} \sqrt[p]{a} = 1$.

3.2 Inner Product Spaces

Recall 3.10 *For $z = \alpha + j\beta \in \mathbb{C}$, $\alpha, \beta \in \mathbb{R}$, we define $\mathrm{Re}\{z\} := \alpha$ and $\mathrm{Im}\{z\} := \beta$. Note that $\mathrm{Im}\{z\} \in \mathbb{R}$. The complex conjugate of z is $\overline{z} := \alpha - j\beta$ and $|z| := \sqrt{z \cdot \overline{z}} = \sqrt{\alpha^2 + \beta^2}$. Note that $z \in \mathbb{R} \iff (z = \overline{z}) \iff (\mathrm{Im}\{z\} = 0)$. For any complex number z, $\mathrm{Re}\{z\} \leq |z|$. Finally, $z = 0 \iff \overline{z} = 0 \iff |z| = 0$.*

Definition 3.11 *Let $(\mathcal{X}, \mathbb{C})$ be a vector space. A function $\langle \cdot, \cdot \rangle : \mathcal{X} \times \mathcal{X} \to \mathbb{C}$ is an inner product if*

(a) $\langle a, b\rangle = \overline{\langle b, a\rangle}$.

(b) $\langle \alpha_1 x_1 + \alpha_2 x_2, y\rangle = \alpha_1\langle x_1, y\rangle + \alpha_2\langle x_2, y\rangle$, *linear in the left argument. Some books place the linearity on the right side.*

(c) $\langle x, x\rangle \geq 0$ *for any $x \in \mathcal{X}$, and $\langle x, x\rangle = 0 \iff x = 0$. (See below: $\langle x, x\rangle$ is a real number and therefore it can be compared to 0.)*

Remark 3.12

1. $\langle x, x\rangle = \overline{\langle x, x\rangle}$, *by (a) and hence, $\langle x, x\rangle$ is always a real number.*
2. *If the vector space is defined as $(\mathcal{X}, \mathbb{R})$, replace (a) with* $(a^{'})$ $\langle a, b\rangle = \langle b, a\rangle$
3. *What about linear combinations on the right side? From (a) and (b)*

$$\begin{aligned}\langle x, \beta_1 y_1 + \beta_2 y_2\rangle &= \overline{\langle \beta_1 y_1 + \beta_2 y_2, x\rangle} \\ &= \overline{\beta_1\langle y_1, x\rangle + \beta_2\langle y_2, x\rangle} \\ &= \overline{\beta_1}\ \overline{\langle y_1, x\rangle} + \overline{\beta_2}\ \overline{\langle y_2, x\rangle} \\ &= \overline{\beta_1}\ \langle x, y_1\rangle + \overline{\beta_2}\ \langle x, y_2\rangle\end{aligned}$$

4. *If the field is the real numbers, then the above reduces to $\langle x, \beta_1 y_1 + \beta_2 y_2\rangle = \beta_1\ \langle x, y_1\rangle + \beta_2\ \langle x, y_2\rangle$*

Example 3.13 *Common inner products:*

(a) $(\mathbb{C}^n, \mathbb{C})$, $\langle x, y\rangle := x^\top \overline{y} = \sum_{i=1}^{n} x_i \overline{y_i}$.

(b) $(\mathbb{R}^n, \mathbb{R})$, $\langle x, y\rangle := x^\top y = \sum_{i=1}^{n} x_i y_i$.

(c) $\mathcal{F} = \mathbb{R}$, $\mathcal{X} = \{A \,|\, n \times m$ *real matrices*$\}$, $\langle A, B\rangle := \mathrm{tr}(AB^\top) = \mathrm{tr}(A^\top B)$.

(d) $\mathcal{X} = \{f : [a, b] \to \mathbb{R}, f$ *continuous*$\}$, $\mathcal{F} = \mathbb{R}$, $\langle f, g\rangle := \int_a^b f(t)g(t)\,\mathrm{d}t$.

Theorem 3.14 ***(Cauchy-Schwarz Inequality)*** *Let $(\mathcal{X}, \mathcal{F}, \langle \cdot, \cdot \rangle)$ be an inner product space, with $\mathcal{F}$ either $\mathbb{R}$ or $\mathbb{C}$. Then, for all $x, y \in \mathcal{X}$*

$$|\langle x, y\rangle| \leq \langle x, x\rangle^{1/2}\langle y, y\rangle^{1/2}.$$

Proof: We will first do the proof for $\mathcal{F} = \mathbb{R}$. We note that if $y = 0$, the result is clearly true. Hence, we assume $y \neq 0$ and let $\lambda \in \mathbb{R}$ be a scalar that is to be chosen. Then,

$$\begin{aligned}0 &\leq ||x - \lambda y||^2 \\ &= \langle x - \lambda y,\ x - \lambda y\rangle \\ &= \langle x,\ x - \lambda y\rangle - \lambda\langle y,\ x - \lambda y\rangle \\ &= \langle x, x\rangle - \lambda\langle x, y\rangle - \lambda\langle y, x\rangle + \lambda^2\langle y, y\rangle \\ &= \langle x, x\rangle - 2\lambda\langle x, y\rangle + \lambda^2\langle y, y\rangle.\end{aligned}$$

We'll now make a choice for λ that minimizes $\langle x,x\rangle - 2\lambda\langle x,y\rangle + \lambda^2\langle y,y\rangle$. Taking the derivative with respect to λ and setting it equal to zero yields $\lambda = \langle x,y\rangle/\langle y,y\rangle$. In case you are curious whether this is a max or min, note that $\langle y,y\rangle > 0$ because $y \neq 0$.

Substituting in this value for λ gives

$$\begin{aligned}
0 &\leq \min_{\lambda\in\mathbb{R}} \ \langle x-\lambda y, x-\lambda y\rangle \\
&= \langle x-\lambda y, x-\lambda y\rangle|_{\lambda=\langle x,y\rangle/\langle y,y\rangle} \\
&= \langle x,x\rangle - 2|\langle x,y\rangle|^2/\langle y,y\rangle + |\langle x,y\rangle|^2/\langle y,y\rangle \\
&= \langle x,x\rangle - |\langle x,y\rangle|^2/\langle y,y\rangle \\
&\Downarrow \\
0 &\leq \langle x,x\rangle\cdot\langle y,y\rangle - |\langle x,y\rangle|^2
\end{aligned}$$

Therefore, we can conclude that $|\langle x,y\rangle|^2 \leq \langle x,x\rangle\langle y,y\rangle \implies |\langle x,y\rangle| \leq \langle x,x\rangle^{1/2}\langle y,y\rangle^{1/2}$ and the proof is done.

We quickly outline the steps for $\mathcal{F} = \mathbb{C}$. Because the inner product of a vector with itself is always a non-negative real number, for all scalars $\lambda \in \mathbb{C}$,

$$0 \leq \ \langle x-\lambda y, x-\lambda y\rangle = \langle x,x\rangle - \lambda\langle y,x\rangle - \overline{\lambda}\langle x,y\rangle + |\lambda|^2\langle y,y\rangle.$$

For the particular choice $\lambda = \frac{\langle x,y\rangle}{\langle y,y\rangle}$, direct calculation shows

$$0 \leq \ \langle x,x\rangle - \frac{|\langle x,y\rangle|^2}{\langle y,y\rangle},$$

which gives

$$|\langle x,y\rangle| \leq \sqrt{\langle x,x\rangle\langle y,y\rangle} = \langle x,x\rangle^{1/2}\cdot\langle y,y\rangle^{1/2},$$

and the proof is done.

Note: With the above choice of λ, we have

$$\lambda\cdot\langle y,x\rangle = \frac{|\langle x,y\rangle|^2}{\langle y,y\rangle}, \ \ \overline{\lambda}\cdot\langle x,y\rangle = \frac{|\langle x,y\rangle|^2}{\langle y,y\rangle}, \text{ and } |\lambda|^2\cdot\langle y,y\rangle = \frac{|\langle x,y\rangle|^2}{\langle y,y\rangle}.$$

■

Corollary 3.15 *Let $(\mathcal{X},\mathcal{F},\langle\cdot\,,\cdot\rangle)$ be an inner product space, with $\mathcal{F}$ either $\mathbb{R}$ or $\mathbb{C}$. Then,*

$$\|x\| := \langle x,x\rangle^{1/2} = \sqrt{\langle x,x\rangle}$$

*is a **norm**.*

Proof: As before, for clarity of exposition, we first assume $\mathcal{F} = \mathbb{R}$. We will only check the triangle inequality $\|x+y\| \leq \|x\| + \|y\|$, which is equivalent to showing $\|x+y\|^2 \leq \|x\|^2 + \|y\|^2 + 2\|x\|\cdot\|y\|$. The other parts are left as an exercise.

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

where the last step uses the Cauchy-Schwarz inequality.

We'll now quickly do the changes required to handle $\mathcal{F} = \mathbb{C}$. The triangle inequality is $\|x+y\| \leqslant \|x\| + \|y\|$, which is equivalent to showing $\|x+y\|^2 \leqslant \|x\|^2 + 2\|x\|\,\|y\| + \|y\|^2$. Brute force computation yields,

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

where $\mathrm{Re}\{\langle x, y\rangle\}$ denotes the real part of the complex number $\langle x, y\rangle$. However, for any complex number α, $\mathrm{Re}\{\alpha\} \leq |\alpha|$, and thus we have

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

where the last inequality is from the Cauchy-Schwarz Inequality.

■

Definition 3.16 *Orthogonal and orthonormal vectors.*

*(a) Two vectors x and y are **orthogonal** if $\langle x, y\rangle = 0$. **Notation:** $x \perp y$*

*(b) **A set of vectors S is orthogonal** if*

$$\forall x,y \in S, x \neq y \implies \langle x, y\rangle = 0 \text{ (i.e. } x \perp y\text{)}$$

*(c) If in addition, $\|x\| = 1$ for all $x \in S$, then S is an **orthonormal set**.*

Remark 3.17 *For $x \neq 0$, $\frac{x}{\|x\|}$ has norm 1, because $\left\|\frac{x}{\|x\|}\right\| = \left|\frac{1}{\|x\|}\right| \cdot \|x\| = \frac{1}{\|x\|} \cdot \|x\| = 1$.*

Theorem 3.18 ***(Pythagorean Theorem)** If $x \perp y$, then*

$$\|x+y\|^2 = \|x\|^2 + \|y\|^2.$$

Proof: From the proof of the triangle inequality,

$$\|x+y\|^2 = \|x\|^2 + \|y\|^2 + 2\langle x, y\rangle.$$

Once we note that $\langle x, y\rangle = 0$ because $x \perp y$, we are done.

■

3.3 Gram Schmidt Process

Proposition 3.19 ***(Recursion Step Gram Schmidt Process)** Let $(\mathcal{X}, \mathcal{F}, \langle\cdot,\cdot\rangle)$ be an inner product space, $\{y^1, \cdots, y^k\}$ a linearly independent set, and $\{v^1, \cdots, v^{k-1}\}$ an orthogonal set satisfying*

$$\mathrm{span}\{v^1, \cdots, v^{k-1}\} = \mathrm{span}\{y^1, \cdots, y^{k-1}\}. \tag{3.1}$$

Define

$$v^k = y^k - \sum_{j=1}^{k-1} \frac{\langle y^k, v^j\rangle}{\|v^j\|^2} \cdot v^j \tag{3.2}$$

where $\|v^j\|^2 = \langle v^j, v^j\rangle$. Then, $\{v^1, \cdots, v^k\}$ is orthogonal and

$$\mathrm{span}\{v^1, \cdots, v^k\} = \mathrm{span}\{y^1, \cdots, y^k\}. \tag{3.3}$$

Proof: We first note that from (3.1), $v^i \neq 0$ for $1 \leq i \leq k-1$. Next, the orthogonality of $\{v^1, \cdots, v^k\}$ is essentially by construction. To see this, we write

$$v^k = y^k - \sum_{i=1}^{k-1} a_{ki} v^i$$

and then check that $\langle v^k, v^j \rangle = 0$ for $1 \leq j \leq k-1$ if, and only if,

$$a_{kj} = \frac{\langle y^k, v^j \rangle}{\|v^j\|^2}.$$

Indeed,

$$\langle v^i, v^j \rangle = \begin{cases} 0 & j \neq i, 1 \leq i, j \leq k-1 \\ ||v^i||^2 & i = j, \end{cases}$$

and hence

$$0 = \langle v^k, v^j \rangle = \langle y^k, v^j \rangle - \sum_{i=1}^{k-1} a_{ki} \langle v^i, v^j \rangle = \langle y^k, v^j \rangle - a_{kj} \langle v^j, v^j \rangle = \langle y^k, v^j \rangle - a_{kj} ||v^j||^2.$$

We next show that $y^k \in \text{span}\{v^1, \cdots, v^k\}$ and $v^k \in \text{span}\{y^1, \cdots, y^k\}$.

From (3.2),

$$y^k = v^k + \sum_{j=1}^{k-1} \frac{\langle y^k, v^j \rangle}{\|v^j\|^2} \cdot v^j \implies y^k \in \text{span}\{v^1, \cdots, v^k\}.$$

Left to show: $v^k \in \text{span}\{y^1, \cdots, y^k\}$. By hypothesis,

$$v^j \in \text{span}\{y^1, \cdots, y^{k-1}\} \text{ for all } 1 \leqslant j \leqslant k-1,$$

so

$$\sum_{j=1}^{k-1} \left(\frac{\langle v^j, y^k \rangle}{\|v^j\|^2} \right) v^j \in \text{span}\{y^1, \cdots, y^{k-1}\} \subset \text{span}\{y^1, \cdots, y^k\}.$$

Clearly, $y^k \in \text{span}\{y^1, \cdots, y^k\}$. Putting these two facts together,

$$v^k = y^k - \sum_{j=1}^{k-1} \left(\frac{\langle v^j, y^k \rangle}{\|v^j\|^2} \right) v^j \in \text{span}\{y^1, \cdots, y^k\}$$

because $\text{span}\{y^1, \cdots, y^k\}$ is a subspace. ■

Definition 3.20 *The **Gram-Schmidt Process** consists of initialing* (3.1) *with* $v^1 = y^1$ *and then applying* (3.2) *recursively. When implementing the algorithm in code, it is quite easy to normalize the vectors as you go, as in the following pseudocode:*

```
# Given {y1, ..., yn} linearly independent
# Produce {v1, ..., vn} orthonormal such that
# span{v1, ... vk} = span{y1, ..., yk}
v1 = y1
v1=v1/norm(v1)
for k = 2 : n
vk = yk
    for j = 1 : k - 1
        vk = vk - < yk, vj >
    end
    vk = vk/norm(vk)
end
```

Example 3.21 *Given the following data in* $(\mathbb{R}^3, \mathbb{R})$*,*

$$\{y^1, y^2, y^3\} = \left\{ \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} \right\},$$

and inner product $< p, q >:= p^T q = \sum_{i=1}^{3} p_i q_i$*, apply Gram-Schmidt to produce an orthogonal basis. Normalize to produce an orthonormal basis.*

Solution:

$$v^1 = y^1 = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$$
$$\|v^1\|^2 = (v^1)^T v^1 = 2;$$

$$v^2 = y^2 - \frac{< v^1, y^2 >}{\|v^1\|^2} v^1$$
$$= \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} - \underbrace{\begin{bmatrix} 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}}_{3} \frac{1}{2} \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} -\frac{1}{2} \\ \frac{1}{2} \\ 3 \end{bmatrix}$$
$$\|v^2\|^2 = 9\frac{1}{2} = \frac{19}{2};$$

$$v^3 = y^3 - \frac{< v^1, y^3 >}{\|v^1\|^2} v^1 - \frac{< v^2, y^3 >}{\|v^2\|^2} v^2$$
$$= \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} - \underbrace{\begin{bmatrix} 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}}_{1} \frac{1}{2} \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} - \underbrace{\begin{bmatrix} -\frac{1}{2} & \frac{1}{2} & 3 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}}_{3\frac{1}{2}} \frac{1}{\frac{19}{2}} \begin{bmatrix} -\frac{1}{2} \\ \frac{1}{2} \\ 3 \end{bmatrix}$$
$$= \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} - \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \\ 0 \end{bmatrix} - \begin{bmatrix} -\frac{7}{38} \\ \frac{7}{38} \\ \frac{21}{19} \end{bmatrix} = \begin{bmatrix} -\frac{6}{19} \\ \frac{6}{19} \\ -\frac{2}{19} \end{bmatrix}.$$

Normalize to obtain Orthonormal Basis:often useful to do this, but never fun to do by hand.

$$\tilde{v}_1 = \frac{v^1}{\|v^1\|} = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \\ 0 \end{bmatrix}$$
$$\tilde{v}_2 = \frac{v^2}{\|v^2\|} = \begin{bmatrix} \frac{-1}{\sqrt{38}} \\ \frac{1}{\sqrt{38}} \\ 3\sqrt{\frac{2}{19}} \end{bmatrix}$$
$$\tilde{v}_3 = \frac{v^3}{\|v^3\|} = \frac{19}{\sqrt{76}} \begin{bmatrix} -\frac{6}{19} \\ \frac{6}{19} \\ -\frac{2}{19} \end{bmatrix}$$

■

Example 3.22 *Given the real vector space* $C[0,1] = \{f : [0,1] \to \mathbb{R} \mid f \text{ continuous}\}$, *inner product* $< f, g >:= \int_0^1 f(\tau)g(\tau)d\tau$, *and* $\{y^1, y^2, y^3\} = \{1, t, t^2\}$, *apply Gram-Schmidt to produce an orthogonal basis for* $\text{span}\{1, t, t^2\}$.

Solution:

$$
\begin{aligned}
v^1 &= y^1 = 1 \\
\|v^1\|^2 &= \int_0^1 (1)^2 d\tau = 1;
\end{aligned}
$$

$$
\begin{aligned}
v^2 &= y^2 - \frac{< v^1, y^2 >}{\|v^1\|^2} v^1 \\
&= t - \underbrace{\int_0^1 1 \cdot \tau d\tau}_{\frac{1}{2}} \cdot \frac{1}{1} \cdot 1 = t - \frac{1}{2} \\
\|v^2\|^2 &= \int_0^1 (\tau - \frac{1}{2})^2 d\tau = \frac{1}{12};
\end{aligned}
$$

$$
\begin{aligned}
v^3 &= y^3 - \frac{< v^1, y^3 >}{\|v^1\|^2} v^1 - \frac{< v^2, y^3 >}{\|v^2\|^2} v^2 \\
&= t^2 - \underbrace{\int_0^1 1 \cdot \tau^2 d\tau}_{\frac{1}{3}} \cdot \frac{1}{1} \cdot 1 - \underbrace{\int_0^1 (\tau - \frac{1}{2})\tau^2 d\tau}_{\frac{1}{12}} \left(\frac{1}{\frac{1}{12}}\right) \left(t - \frac{1}{2}\right) \\
&= t^2 - \frac{1}{3} - \left(t - \frac{1}{2}\right) \\
&= t^2 - t + \frac{1}{6}.
\end{aligned}
$$

■

Example 3.23 *Doing inner products on* $C[a,b]$ *in MATLAB*

```
>> clear *
>> syms t % declare to be a symbolic variable

>> % INT(S,a,b) is the definite integral of S with respect to
   % its symbolic variable from a to b. a and b are each
   % double or symbolic scalars.

>> y1=1+0*t; % Otherwise MATLAB will not treat y1 as a
             %trivial function of the symbolic variable
>> y2=t;
>> y3=t^2;

% Start the G-S Procedure. Here we assume C[0,1], that is
% C[a,b], with [a,b]=[0,1]

>> v1=y1
>> v2=y2-int(v1*y2,0,1)*v1/int(v1^2,0,1)
>> v3=y3-int(v1*y3,0,1)*v1/int(v1^2,0,1)-
int(v2*y3,0,1)*v2/int(v2^2,0,1)

% Next, normalize to length one

```

```
>> v1_tilde=v1/int(v1^2,0,1)^.5
>> v2_tilde=simplify( v2/int(v2^2,0,1)^.5 )
>> v3_tilde=simplyfy(v3/int(v3^2,0,1)^.5)
```

Output

```
v1=1
v2=t-1/2
v3=t^2+1/6-t

v1_tilde=1
v2_tilde=(t-1/2)*12^(1/2)
v3_tilde=(6*t^2+1-6*t)*5^(1/2)
```

Remark 3.24 ***(Round-off Errors Affect Classical Gram-Schmidt)*** *The classical Gram-Schmidt Process in Definition 3.20 is straightforward to understand, which is why it is taught in courses. Unfortunately, it behaves poorly under the round-off error that occurs in digital computations! Here is a standard example:*

$$u_1 = \begin{bmatrix} 1 \\ \varepsilon \\ 0 \\ 0 \end{bmatrix}, u_2 = \begin{bmatrix} 1 \\ 0 \\ \varepsilon \\ 0 \end{bmatrix}, u_3 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \varepsilon \end{bmatrix}, \varepsilon > 0$$

Let $\{e_1, e_2, e_3, e_4\}$ *be the standard basis vectors corresponding to the columns of the* 4×4 *identity matrix. We note that*

$$u_2 = u_1 + \varepsilon(e_3 - e_2)$$
$$u_3 = u_2 + \varepsilon(e_4 - e_3)$$

and thus, for $\epsilon \neq 0$,

$$\text{span}\{u_1, u_2\} = \text{span}\{u_1, (e_3 - e_2)\}$$
$$\text{span}\{u_1, u_2, u_3\} = \text{span}\{u_1, (e_3 - e_2), (e_4 - e_3)\}$$

Example 3.25 *Hence, Gram-Schmidt applied to* $\{u_1, u_2, u_3\}$ *and* $\{u_1, (e_3 - e_2), (e_4 - e_3)\}$ *should "theoretically" produce the same orthonormal vectors. To check this, we go to MATLAB, and for* $\varepsilon = 0.1$*, we do indeed get the same results. You can verify this yourself. However, with* $\varepsilon = 10^{-8}$*,*

$$Q_1 = \begin{bmatrix} 1.0000 & 0.0000 & 0.0000 \\ 0.0000 & -0.7071 & -0.7071 \\ 0.0000 & 0.7071 & 0.0000 \\ 0.0000 & 0.0000 & 0.7071 \end{bmatrix}$$
$$Q_2 = \begin{bmatrix} 1.0000 & 0.0000 & 0.0000 \\ 0.0000 & -0.7071 & -0.4082 \\ 0.0000 & 0.7071 & -0.4082 \\ 0.0000 & 0.0000 & 0.8165 \end{bmatrix}$$

where

$$Q_1 = \begin{bmatrix} \frac{v_1}{\|v_1\|} & \frac{v_2}{\|v_2\|} & \frac{v_3}{\|v_3\|} \end{bmatrix}$$

has been computed with Classical-Gram-Schmidt for $\{u_1, u_2, u_3\}$ *while*

$$Q_2 = \begin{bmatrix} \frac{v_1}{\|v_1\|} & \frac{v_2}{\|v_2\|} & \frac{v_3}{\|v_3\|} \end{bmatrix}$$

has been computed with Classical-Gram-Schmidt for $\{u_1, (e_3 - e_2), (e_4 - e_3)\}$*. Hence we do NOT obtain the same result!*

Definition 3.26 ***(Modified Gram-Schmidt)*** *has better numerical performance*
for $k = 1 : n$
$v_k = u_k$ *%copy over the vectors*

end

for $k = 1 : n$
  $v_k = \frac{v_k}{\|v_k\|}$
  for $j = (k+1) : n$
    $v_j = v_j - (v_j \bullet v_k)v_k$ *%Makes* v_j *orthogonal to* v_k
  end
end

At ***Step k=1****,* v_1 *is normalized to length one, and then* $v_2, \ldots, v_n$ *are redefined to be orthogonal to* v_1*. At* ***Step k=2:*** v_2 *is normalized to length one, and then* $v_3, \ldots, v_n$ *are redefined to be orthogonal to* v_2*. We note that they were already orthogonal to* v_1*. At* ***Step k:*** v_k *is normalized to length one, and then* $v_{k+1}, \ldots, v_n$ *are redefined to be orthogonal to* v_k*. We note that they were already orthogonal to* $v_1, \ldots, v_{k-1}$*.*

Example 3.27 *Hence, if Modified Gram-Schmidt is so great, when applied to* $\{u_1, u_2, u_3\}$ *and* $\{u_1, (e_3 - e_2), (e_4 - e_3)\}$*, it should produce the same orthonormal vectors and it does! To check this, we go to MATLAB for* $\varepsilon = 10^{-8}$ *and obtain*

$$Q_1 = \begin{bmatrix} 1.0000 & 0.0000 & 0.0000 \\ 0.0000 & -0.7071 & -0.7071 \\ 0.0000 & 0.7071 & 0.0000 \\ 0.0000 & 0.0000 & 0.7071 \end{bmatrix}$$

$$Q_2 = \begin{bmatrix} 1.0000 & 0.0000 & 0.0000 \\ 0.0000 & -0.7071 & -0.7071 \\ 0.0000 & 0.7071 & 0.0000 \\ 0.0000 & 0.0000 & 0.7071 \end{bmatrix}$$

where Q_1 *and* Q_2 *are defined above.* ***When one is equipped with the right Algorithm, the world is truly a marvelous place.***

Remark 3.28 *Just to be perfectly clear, with perfect arithmetic (no rounding errors), Classical Gram-Schmidt and Modified Gram-Schmidt are equivalent.*

3.4 Projection Theorem and the Normal Equations

Lemma 3.29 *(called the Pre-Projection Theorem in Luenberger) Let* $\mathcal{X}$ *be a finite-dimensional (real) inner product space,* M *be a subspace of* $\mathcal{X}$*, and* x *be an arbitrary point in* $\mathcal{X}$*.*

(a) If $\exists m_0 \in M$ *such that* $\|x - m_0\| \leq \|x - m\| \ \ \forall m \in M$*, then* m_0 *is unique.*

(b) A necessary and sufficient condition for m_0 *to be a minimizing vector in* M *is that the vector* $x - m_0$ *is orthogonal to* M*.*

Remarks:

(a') If $\exists m_0 \in M$ such that $\|x - m_0\| = d(x, M) = \inf_{m \in M} \|x - m\|$, then m_0 is unique. [equivalent to (a)]

(b') $\|x - m_0\| = d(x, M) \iff x - m_0 \perp M$. [equivalent to (b)]

Proof: We break the proof up into a series of claims.

Claim 3.30 *If* $m_0 \in M$ *satisfies* $\|x - m_0\| = d(x, M)$*, then* $x - m_0 \perp M$*.*

Proof: (By contrapositive) Assume $x - m_0 \not\perp M$. We will produce $m_1 \in M$ such that $\|x - m_1\| < \|x - m_0\|$. Indeed, suppose $x - m_0 \not\perp M$. Then, $\exists m \in M$ such that $\langle x - m_0, m \rangle \neq 0$. We know $m \neq 0$, and hence we define

- $\tilde{m} = \frac{m}{\|m\|} \in M$;
- $\delta := \langle x - m_0, \tilde{m} \rangle \neq 0$; and
- $m_1 = m_0 + \delta \tilde{m} \implies m_1 \in M$.

The intuition behind the definition of m_1 is that $x - m_1$ is "closer" to being perpendicular to M than is $x - m_0$, and hence it should follow that $||x - m_1|| < ||x - m_0||$. To prove the latter point, we do a few computations:

$$\begin{aligned}\|x - m_1\|^2 &= \|x - m_0 - \delta\tilde{m}\|^2 \\ &= \langle x - m_0 - \delta\tilde{m}, x - m_0 - \delta\tilde{m}\rangle \\ &= \langle x - m_0, x - m_0\rangle - \delta\underbrace{\langle x - m_0, \tilde{m}\rangle}_{\delta} - \delta\underbrace{\langle \tilde{m}, x - m_0\rangle}_{\delta} + \delta^2\underbrace{\langle \tilde{m}, \tilde{m}\rangle}_{=1} \\ &= \|x - m_0\|^2 - \delta^2 \\ &< \|x - m_0\|^2\end{aligned}$$

because $\delta^2 > 0$. Hence, $\|x - m_1\|^2 < \|x - m_0\|^2$ and therefore, $\|x - m_0\| \neq \inf_{m\in M} \|x - m\| := d(x, M)$. □

Claim 3.31 *If $x - m_0 \perp M$, then $\|x - m_0\| = d(x, M)$ and m_0 is unique.*

Proof: Recall the Pythagorean Theorem:

$$\|x + y\|^2 = \|x\|^2 + \|y\|^2 \text{ when } x \perp y$$

Let $m \in M$ be arbitrary and suppose $x - m_0 \perp M$. Then $x - m_0 \perp m_0 - m$, and thus

$$\begin{aligned}\|x - m\|^2 &= \|x - m_0 + \underbrace{m_0 - m}_{\in M}\|^2 \\ &= \|x - m_0\|^2 + \|m_0 - m\|^2.\end{aligned}$$

It follows that

$$\inf_{m\in M}\|x - m\|^2 = \inf_{m\in M}\left(\|x - m_0\|^2 + \|m_0 - m\|^2\right) = \|x - m_0\|^2 + \inf_{m\in M}\|m_0 - m\|^2 = \|x - m_0\|^2.$$

The unique minimizer is m_0 because $\|m_0 - m\|^2 = 0$ only for $m = m_0$. □

The two claims complete the proof.

■

Definition 3.32 *Let $(\mathcal{X}, \mathcal{F}, \langle\cdot,\cdot\rangle)$ be an inner product space, and $S \subset \mathcal{X}$ a* subset *(does not have to be a subspace).*

$$S^\perp := \{x \in \mathcal{X} | x \perp S\} = \{x \in \mathcal{X} | \langle x, y\rangle = 0 \text{ for all } y \in S\}$$

is called ***the orthogonal complement of*** *S.*

Exercise 3.33

- *Show that $S^\perp$ is always a subspace.*
- *If $M = \text{span}\{y^1, \ldots, y^k\}$, show that $(x \in M^\perp) \iff \langle x, y^i\rangle = 0,\ 1 \le i \le k$.*

Proposition 3.34 *Let $(\mathcal{X}, \mathcal{F}, \langle\cdot,\cdot\rangle)$ be a finite dimensional inner product space and M a subspace of $\mathcal{X}$. Then,*

$$\mathcal{X} = M \oplus M^\perp.$$

Remark 3.35 *Suppose that V and W are subspaces of $\mathcal{X}$. Then $V + W := \{x \in \mathcal{X} \mid x = v + w$, for some $v \in V, w \in W\}$. Because V and W are subspaces, $0 \in V \cap W$ (the zero vector is in their intersection). If that is the only vector in the intersection, meaning $V \cap W = \{0\}$, the zero subspace, then we write $V \oplus W$, and it is called the* ***direct sum*** *of V and W. What does the direct sum get you that an ordinary sum would not? You can show that $(x \in V \oplus W) \iff (\exists$ **unique** $v \in V, w \in W$ such that $x = v + w)$.*

Proof: If $x \in M \cap M^\perp$, then by the definition of $M^\perp$, $\langle x, x\rangle = 0$, which implies $x = 0$. Hence, $M \cap M^\perp = \{0\}$. Next, we need to show that $\mathcal{X} = M + M^\perp$, that is, every $X \in \mathcal{X}$ can be written as a sum of a vector in M and a vector in $M^\perp$.

Let $\{y^1, \ldots, y^k\}$ be a basis of M. By Corollary 2.35, it can be completed to a basis for $\mathcal{X}$, that is,

$$\mathcal{X} = \text{span}\{y^1, y^2, \ldots, y^k, y^{k+1}, \ldots, y^n\} \text{ and } \{y^1, y^2, \ldots, y^k, y^{k+1}, \ldots, y^n\} \text{ is linearly independent.}$$

We can then apply Gram-Schmidt to produce orthonormal vectors $\{v^1, \ldots, v^k, v^{k+1}, \ldots, v^n\}$ such that

$$\text{span}\{v^1, \ldots, v^k\} = \text{span}\{y^1, \ldots, y^k\} = M \text{ and } \text{span}\{v^1, \ldots, v^k, v^{k+1}, \ldots, v^n\} = \mathcal{X}.$$

An easy calculation gives

$$M^\perp = \text{span}\{v^{k+1}, \ldots, v^n\}.$$

Indeed, suppose $x = \alpha_1 v^1 + \cdots + \alpha_k v^k + \alpha_{k+1} v^{k+1} + \cdots + \alpha_n v^n$. Then $x \in M^\perp \iff x \perp M \iff \langle x, v^i \rangle = 0, 1 \leq i \leq k$. However,

$$\begin{aligned} \langle x, v^i \rangle &= \alpha_1 \langle v^1, v^i \rangle + \cdots + \alpha_i \langle v^i, v^i \rangle + \cdots + \alpha_n \langle v^n, v^i \rangle \\ &= \alpha_i \quad (\text{because } \langle v^j, v^i \rangle = 0,\ j \neq i, \text{ and } \langle v^i, v^i \rangle = 1) \end{aligned}$$

and therefore $x \perp M \iff \alpha_i = 0,\ 1 \leq i \leq k$. This yields $(x \in M^\perp) \iff (x = \alpha_{k+1} v^{k+1} + \cdots + \alpha_n v^n) \iff (x \in \text{span}\{v^{k+1}, \cdots, v^n\})$. Therefore,

$$M^\perp = \text{span}\{v^{k+1}, \ldots, v^n\}.$$

■

Theorem 3.36 ***(Classical Projection Theorem)*** *Let $(\mathcal{X}, \mathbb{R})$ be a finite dimensional real inner product space and M a subspace of $\mathcal{X}$. Then, $\forall x \in \mathcal{X}, \exists$ unique $\widehat{x} \in M$ such that*

$$\|x - \widehat{x}\| = d(x, M) := \inf_{m \in M} \|x - m\| = \min_{m \in M} \|x - m\|,$$

where we can write min *instead of* inf *because the infimum is achieved. Moreover, $\widehat{x} \in M$ is characterized by $x - \widehat{x} \perp M$.*

Proof: We only need to show that $\forall x \in \mathcal{X}$ there exists $\widehat{x} \in M$ such that $(x - \widehat{x}) \perp M$. This is because if such an $\widehat{x}$ exists, Lemma 3.29, the "Pre-projection Theorem," already shows that it is unique and $\|x - \widehat{x}\| = d(x, M)$. By Proposition 3.34, $\mathcal{X} = M \oplus M^\perp$. Therefore, there exist $\widehat{x} \in M$ and $m^\perp \in M^\perp$ such that

$$x = \widehat{x} + m^\perp.$$

Hence,

$$x - \widehat{x} = m^\perp \in M^\perp \implies (x - \widehat{x}) \perp M.$$

■

Remark 3.37 *You may have observed that $\mathcal{X} = M \oplus M^\perp$ also shows that $\widehat{x}$ is unique. While this is true, it is based on Proposition 3.34, which is true when $\mathcal{X}$ is a "complete" inner product space and M is a "closed" subspace, properties that are automatically satisfied when $\mathcal{X}$ is finite dimensional. We will touch on these more subtle properties later when we do some basic Real Analysis.*

Notation 3.38 $\widehat{x} = \arg\min d(x, M) = \arg\min_{m \in M} \|x - m\|$.

Our next goal is to turn Theorem 3.36 into a means to compute the best approximation value, $\widehat{x}$. By the Projection Theorem, $\widehat{x}$ exists and is characterized by $x - \widehat{x} \perp M$. We write

$$\widehat{x} = \alpha_1 y^1 + \alpha_2 y^2 + \cdots + \alpha_k y^k$$

and impose

$$(x - \widehat{x} \perp M) \iff (x - \widehat{x} \perp y^i,\ 1 \leq i \leq k) \iff (\langle x - \widehat{x}, y^i \rangle = 0,\ 1 \leq i \leq k) \iff (\langle \widehat{x}, y^i \rangle = \langle x, y^i \rangle\ 1 \leq i \leq k).$$

Then,

$$\begin{gathered} \langle \widehat{x}, y^i \rangle = \langle x, y^i \rangle\ 1 \leq i \leq k \\ \Updownarrow \\ \langle \alpha_1 y^1 + \alpha_2 y^2 + \cdots + \alpha_k y^k, y^i \rangle = \langle x, y^i \rangle\ 1 \leq i \leq k \\ \Updownarrow \\ \alpha_1 \langle y^1, y^i \rangle + \alpha_2 \langle y^2, y^i \rangle + \cdots + \alpha_k \langle y^k, y^i \rangle = \langle x, y^i \rangle\ 1 \leq i \leq k \end{gathered}$$

We now write these equations out in matrix form.

$$\begin{aligned} i=1 \quad & \alpha_1\langle y^1,y^1\rangle+\alpha_2\langle y^2,y^1\rangle+\cdots+\alpha_k\langle y^k,y^1\rangle=\langle x,y^1\rangle \\ i=2 \quad & \alpha_1\langle y^1,y^2\rangle+\alpha_2\langle y^2,y^2\rangle+\cdots+\alpha_k\langle y^k,y^2\rangle=\langle x,y^2\rangle \\ \vdots \quad & \qquad\qquad\qquad\qquad \vdots \\ i=k \quad & \alpha_1\langle y^1,y^k\rangle+\alpha_2\langle y^2,y^k\rangle+\cdots+\alpha_k\langle y^k,y^k\rangle=\langle x,y^k\rangle. \end{aligned} \tag{3.4}$$

Definition 3.39 *Equations 3.4 are called the **Normal Equations**. The **Gram Matrix** is the $k\times k$ matrix $G_{ij}:=\langle y^i,y^j\rangle$. The **Normal Equations** can also refer to*

$$G^\top\alpha=\beta \tag{3.5}$$

where,

$$\alpha:=\begin{bmatrix}\alpha_1\\ \alpha_2\\ \vdots\\ \alpha_k\end{bmatrix},\ \beta:=\begin{bmatrix}\langle x,y^1\rangle\\ \langle x,y^2\rangle\\ \vdots\\ \langle x,y^k\rangle\end{bmatrix}=:\begin{bmatrix}\beta_1\\ \beta_2\\ \vdots\\ \beta_k\end{bmatrix},\ \textit{and } G:=G(y^1,\cdots,y^k):=\begin{bmatrix}\langle y^1,y^1\rangle & \langle y^1,y^2\rangle & \cdots & \langle y^1,y^k\rangle\\ \langle y^2,y^1\rangle & \langle y^2,y^2\rangle & \cdots & \langle y^2,y^k\rangle\\ \vdots & \vdots & & \vdots\\ \langle y^k,y^1\rangle & \langle y^k,y^2\rangle & \cdots & \langle y^k,y^k\rangle\end{bmatrix}.$$

Remark 3.40 *Because we are assuming $\mathcal{F}=\mathbb{R}$, $G_{ij}=\langle y^i,y^j\rangle=\langle y^j,y^i\rangle=G_{ji}$, and we therefore have $G=G^\top$. We'll show below that $\det(G)\neq 0$ if, and only if, $\{y^1,y^2,\ldots,y^k\}$ is linearly independent. In this case,*

$$\widehat{x}=\alpha_1y^1+\alpha_2y^2+\cdots+\alpha_ky^k$$

with $G^\top\alpha=\beta$ is the best approximation to x by an element in $M:=\operatorname{span}\{y^1,y^2,\ldots,y^k\}$.

Proposition 3.41 ***(Invertibility of the Gram Matrix)*** *Let $g(y^1,\ldots,y^k):=\det G(y^1,\ldots,y^k)$ be the determinant of the Gram Matrix. Then $g(y^1,\ldots,y^k)\neq 0 \iff \{y^1,y^2,\ldots,y^k\}$ is linearly independent.*

Proof: From our construction of the normal equations, $G^\top\alpha=0$ if, and only if

$$\langle\alpha_1y^1+\alpha_2y^2+\cdots+\alpha_ky^k,y^i\rangle=0,\ 1\le i\le k.$$

This is equivalent to

$$\left(\alpha_1y^1+\alpha_2y^2+\cdots+\alpha_ky^k\right)\perp y^i=0,\ 1\le i\le k,$$

which is equivalent to

$$\left(\alpha_1y^1+\alpha_2y^2+\cdots+\alpha_ky^k\right)\perp \operatorname{span}\{y^1,\cdots,y^k\}=:M$$

and thus

$$\alpha_1y^1+\alpha_2y^2+\cdots+\alpha_ky^k\in M^\perp.$$

Because $\alpha_1y^1+\alpha_2y^2+\cdots+\alpha_ky^k\in M$, we have that

$$\alpha_1y^1+\alpha_2y^2+\cdots+\alpha_ky^k\in M\cap M^\perp=\{0\}$$

and therefore $\alpha_1y^1+\alpha_2y^2+\cdots+\alpha_ky^k=0$. By the linear independence of $\{y^1,y^2,\ldots,y^k\}$, we deduce that $\alpha_i=0, 1\le i\le k$. ■

Summary 3.42 (The normal equations provide a systematic solution of our best approximation problem). *Assume the set $\{y^1,\ldots,y^k\}$ is linearly independent and $M:=\operatorname{span}\{y^1,\cdots,y^k\}$. Then $\widehat{x}=\arg\min d(x,M)=\arg\min_{m\in M}\|x-m\|$ if, and only if,*

$$\begin{aligned} \widehat{x}&=\alpha_1y^1+\alpha_2y^2+\cdots+\alpha_ky^k\\ G^\top\alpha&=\beta\\ G_{ij}&=\langle y^i,y^j\rangle\\ \beta_i&=\langle x,y^i\rangle. \end{aligned} \tag{3.6}$$

What changes in each application of the normal equations is typically the inner product, $\langle\bullet,\bullet\rangle$.

Overdetermined Equations

Roughly speaking, a set of linear equations $Ax = b$ is **overdetermined** when A has more rows than equations. Typically, in this case, there is no value of x such that $Ax - b = 0$. Of course, if $b = 0$ then $x = 0$ is a solution, and more generally, there is a solution if, and only if, $b \in \text{col span}\{A\}$, that is, b can be written as a linear combination of the columns of A. When $b \notin \text{col span}\{A\}$, it makes sense to seek a "best approximate solution", which we'll define to be a value $\widehat{x}$ that minimizes the norm of the error in the solution, that is,

$$\widehat{x} := \arg\min_x \|Ax - b\|.$$

Is there a difference between being overdetermined and having no exact solutions? Yes. It's possible to be overdetermined and still have an exact solution when $b \in \text{col span}\{A\}$. If the columns of A are linearly independent, then

$$Ax = b \text{ is overdetermined} \iff b \notin \text{col span}\{A\}.$$

Example 3.43 (Overdetermined system of linear equations in $\mathbb{R}^n$) *Consider $A\alpha = b$, where $A = n \times m$ real matrix, $n \geq m$, rank$(A) = m$ (columns of A are linearly independent). From the dimension of A, we have that $\alpha \in \mathbb{R}^m, b \in \mathbb{R}^n$. Determine if there exists a "best approximate solution" using the Euclidean norm.*

Solution 1: The *Standard Problem Formulation and Solution* goes like this. We seek $\widehat{\alpha} \in \mathbb{R}^m$ such that

$$\|A\widehat{\alpha} - b\| = \min_{\alpha \in \mathbb{R}^m} \|A\alpha - b\|,$$

where the norm is defined on the error, $e := A\alpha - b \in \mathbb{R}^n$, namely $\|e\| = \sqrt{\sum_{i=1}^{n} (e_i)^2}$. Hence, the problem is

$$\widehat{\alpha} = \arg\min_{\alpha \in \mathbb{R}^m} \sqrt{(A\alpha - b)^\top (A\alpha - b)} = \arg\min_{\alpha \in \mathbb{R}^m} (A\alpha - b)^\top (A\alpha - b).$$

The problem can be solved by "completing the square" or by applying vector calculus to compute the gradient, namely,

$$\frac{\partial}{\partial \alpha} \left((A\alpha - b)^\top (A\alpha - b) \right) = A^\top \cdot (A\alpha - b),$$

setting it to zero, and solving for α. Once you convince yourself that $A^\top A$ is invertible, the solution is computed to be

$$A^\top \cdot (A\widehat{\alpha} - b) = 0 \iff (A^\top A)\widehat{\alpha} = A^\top b \iff \widehat{\alpha} = (A^\top A)^{-1} A^\top b.$$

Solution 2: *Our Problem Formulation and Solution:* The standard formulation puts all the emphasis on the α and misses that $A\alpha$ is a linear combination of the columns of A. The problem is calling for a linear combination of the columns of A that best approximates the vector b. We will therefore apply the normal equations (3.5), where the column span of A is the subspace M and b is the vector $x \in \mathcal{X}$.

Inner product space: $\mathcal{X} = \mathbb{R}^n$, $\mathcal{F} = \mathbb{R}$, $\langle x, y \rangle = x^T y = y^T x = \sum_{i=1}^{n} x_i y_i \implies \|x\|^2 = \langle x, x \rangle = \sum_{i=1}^{n} |x_i|^2$.

Partition $A = \begin{bmatrix} A_1 & A_2 & \cdots & A_m \end{bmatrix}$ and define $\alpha = \begin{bmatrix} \alpha_1 & \alpha_2 & \cdots & \alpha_m \end{bmatrix}^\top$, and note that

$$A\alpha = \alpha_1 A_1 + \alpha_2 A_2 + \cdots + \alpha_m A_m.$$

Seek: $\widehat{x} = A\widehat{\alpha} \in \text{span}\{A_1, A_2, \cdots, A_m\} =: M$ such that $\|\widehat{x} - b\| = d(b, M)$. By the Projection Theorem and the Normal Equations,

$$\widehat{x} = \widehat{\alpha}_1 A_1 + \widehat{\alpha}_2 A_2 + \cdots \widehat{\alpha}_m A_m$$

and $G^\top \widehat{\alpha} = \beta$, with

$$G_{ij} = \langle A_i, A_j \rangle = A_i^\top A_j = [A^\top A]_{ij} \text{ and } \beta_i = \langle A_i, b \rangle = A_i^\top b = [A^\top b]_i.$$

Hence, $G^\top = A^\top A$ and $\beta = A^\top b$. By Proposition 3.41, G is invertible because the columns of A are linearly independent. Therefore, we have that

$$\widehat{\alpha} = (A^\top A)^{-1} A^\top b.$$

■

Remark 3.44 *Once we note that*

$$A^\top = \begin{bmatrix} A_1^\top \\ A_2^\top \\ \vdots \\ A_m^\top \end{bmatrix} \text{ when } A = \begin{bmatrix} A_1 & A_2 & \cdots & A_m \end{bmatrix}$$

The standard "row times column" definition of matrix multiplication gives that $(A^\top A)_{ij} = A_i^\top A_j$. *Similarly,* $(A^\top b)_i = A_i^\top b$.

Definition 3.45 ***(Orthogonal Projection Operator)*** *Let* $\mathcal{X}$ *be a finite dimensional (real) inner product space and* M *a subspace of* $\mathcal{X}$. *For* $x \in \mathcal{X}$ *and* $\widehat{x} \in M$. *The Projection Theorem shows the TFAE:*

(a) $x - \widehat{x} \perp M$.

(b) $\exists m^\perp \in M^\perp$ *such that* $x = \widehat{x} + m^\perp$.

(c) $\|x - \widehat{x}\| = d(x, M) = \inf_{m \in M} \|x - m\|$.

A function $P\colon \mathcal{X} \to M$ *by* $P(x) = \widehat{x}$, *where* $\widehat{x}$ *satisfies any one of (a), (b), or (c), is called the* ***Orthogonal Projection*** *of* $\mathcal{X}$ *onto* M.

Exercise 3.46 *Key properties of the* orthogonal projection operator.

- *The orthogonal projection operator* $P\colon \mathcal{X} \to M$ *is a linear operator.*
- *Let* $\{v^1, \ldots, v^k\}$ *be an* orthonormal basis *for* M. *Then* $P(x) = \sum_{i=1}^{k} \langle x, v^i \rangle v^i$.

The following sets up the Projection Theorem for underdetermined linear equations.

Lemma 3.47 ***(Linear Varieties or Translates of Subspaces)*** *Let* $(\mathcal{X}, \mathbb{R}, < \cdot, \cdot >)$ *be a finite-dimensional inner product space. Let* $\{y_1, \cdots, y_p\}$ *be a linearly independent set in* $\mathcal{X}$ *and let* $c_1, \cdots, c_p$ *be real constants. Define*

$$V := \{x \in \mathcal{X} \mid < x, y_i >= c_i,\ 1 \le i \le p\}.$$

Then the following are true:

> ***Claim 3.48*** *There exists a unique* $x_0 \in \operatorname{span}\{y_1, \cdots, y_p\}$ *such that* $< x_0, y_i >= c_i,\ 1 \le i \le p$.
>
> ***Remark:*** *Another way of stating the Claim is that there exists a unique* $x_0 \in \mathcal{X}$ *such that*
>
> $$V \cap \operatorname{span}\{y_1, \cdots, y_p\} = \{x_0\}.$$
>
> ***Claim 3.49*** *Let* $M = (\operatorname{span}\{y_1, \cdots, y_p\})^\perp$. *Then* $V = x_0 + M$; *in other words,* $x \in V$ *if, and only if,* $(x - x_0) \perp \operatorname{span}\{y_1, \cdots, y_p\}$.
>
> ***Claim 3.50*** *There exists a unique* $v^* \in V$ *having minimum norm, and* v^* *is characterized by* $v^* \perp M$ *(just for emphasis, we note that the result does not say that* $v^* \perp V$*).*
>
> ***Remarks:*** *We note that* $v^* \perp M \iff v^* \in \operatorname{span}\{y_1, \cdots, y_p\}$ *because* $\mathcal{X} = M \oplus M^\perp$ *implies that*
>
> $$M^\perp := \left(\operatorname{span}\{y_1, \cdots, y_p\}^\perp\right)^\perp = \operatorname{span}\{y_1, \cdots, y_p\}.$$
>
> *We are using the standard induced norm,* $||x|| =< x, x >^{1/2}$. *There exists* v^* *having minimum norm means* $||v^*|| = \inf_{v \in V} ||v||$, *and thus* $v^* = \arg\min_{v \in V} ||v|| = \arg\min_{v \in V} ||v||^2$.

Proof: The three claims are proven in HW. ■

Theorem 3.51 *Let $(\mathcal{X}, \mathbb{R}, <\cdot,\cdot>)$ be a finite-dimensional inner product space. Let $\{y_1, \cdots, y_p\}$ be a linearly independent set in $\mathcal{X}$ and let $c_1, \ldots, c_p$ be real constants. Define $V = \{x \in \mathcal{X} \mid <x, y_i> = c_i,\ 1 \le i \le p\}$. Then there exists a unique $v^* \in V$ such that*

$$v^* = \arg\min_{v \in V} ||v||^2. \tag{3.7}$$

*Moreover, $v^* = \sum_{i=1}^{p} \beta_i y_i$, where the β_i's satisfy the **normal equations***

$$\begin{bmatrix} <y_1,y_1> & <y_2,y_1> & \cdots & <y_p,y_1> \\ <y_1,y_2> & <y_2,y_2> & \cdots & <y_p,y_2> \\ \vdots & \vdots & \ddots & \vdots \\ <y_1,y_p> & <y_2,y_p> & \cdots & <y_p,y_p> \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_p \end{bmatrix} \tag{3.8}$$

Proof: From Claim 3.50 and the remark that follows it, we have $v^* \in \text{span}\{y_1, \ldots, , y_p\}$ and thus we write $v^* = \beta_1 y_1 + \cdots + \beta_p y_p$. Then imposing that $v^* \in V$ immediately gives (3.8). ■

Remark 3.52 *Equation* (3.7) *is a special case of a **Quadratic Program**, typically called a **QP** for short. It has a quadratic cost, $\|v\|^2 = \langle v, v\rangle$, and a set of linearly independent equality constraints, namely $v \in V \iff \langle v, y_i\rangle = c_i$, $1 \le i \le p$.*

3.5 Relations between Symmetric and Orthogonal Matrices

The main goal of this section is to establish properties of real symmetric matrices. Throughout the section , we use the standard inner product on $\mathbb{C}^n$, namely, $\langle x, y\rangle := x^\top \overline{y}$, where $\overline{y}$ is the complex conjugate of the vector y.

Proposition 3.53 *If A is $n \times n$ and real, then its e-values and e-vectors occur in **complex conjugate pairs**.*

Proof: Let $\lambda \in \mathbb{C}$ and $(v \in \mathbb{C}^n, v \neq 0)$ satisfy $Av = \lambda v$. **To show:** $A\overline{v} = \overline{\lambda}\overline{v}$, that is, if λ is a e-value with e-vector v, then its complex conjugate $\overline{\lambda}$ is also an e-value with e-vector $\overline{v}$. The proof relies on the fact that if $z_1, z_2 \in \mathbb{C}$, then $\overline{z_1 \cdot z_2} = \overline{z}_1 \cdot \overline{z}_2$, and its natural extension to matrices and vectors that the reader can show.

- $\overline{A \cdot v} = \overline{A} \cdot \overline{v} = A \cdot \overline{v}$ (because A is real).
- $\overline{A \cdot v} = \overline{\lambda \cdot v} = \overline{\lambda} \cdot \overline{v}$ (because $Av = \lambda v$)

Hence, as we needed to show, $A\overline{v} = \overline{\lambda}\overline{v}$. ■

Definition 3.54 *A matrix A is **symmetric** if $A^\top = A$.*

Proposition 3.55 *If A is real (i.e., $\overline{A} = A$) and symmetric (i.e., $A^\top = A$), then $\forall\, x, y \in \mathbb{C}^n$, $\langle Ax, y\rangle = \langle x, Ay\rangle$.*

Proof: Using $A^\top = A$, we obtain $\langle Ax, y\rangle = x^\top A^\top \overline{y} = x^\top A\overline{y}$. Using $\overline{A} = A$, we obtain $\langle x, Ay\rangle = x^\top \overline{Ay} = x^\top \overline{A}\overline{y} = x^\top A\overline{y}$. Hence, $\langle Ax, y\rangle = \langle x, Ay\rangle = x^\top A\overline{y}$. ■

Proposition 3.56 *E-values of a real symmetric matrix A are real.*

Proof: To show $\lambda = \overline{\lambda}$, we apply Proposition 3.55 with $x = y = v$, where $v \neq 0$ satisfies $Av = \lambda v$.

$$\begin{gathered} \langle Av, v\rangle = \langle v, Av\rangle \\ \Updownarrow \\ \langle \lambda v, v\rangle = \langle v, \lambda v\rangle \\ \Updownarrow \\ \lambda\langle v, v\rangle = \overline{\lambda}\langle v, v\rangle \\ \Updownarrow \\ \lambda||v||^2 = \overline{\lambda}||v||^2 \\ \Updownarrow \\ \lambda = \overline{\lambda}, \end{gathered}$$

because $||v|| \neq 0$. ■

Remark 3.57 *We now know that when A is real and symmetric, any eigenvalue λ is real, and therefore we can assume the corresponding eigenvector is real. Indeed,*

$$\underbrace{(A-\lambda I)}_{\text{real}}v = 0.$$

Hence we have $0 \neq v \in \mathbb{R}^n$ and we can use the real inner product on $\mathbb{R}^n$, namely $\langle x, y\rangle = x^\top y$.

Proposition 3.58 *Let A be an $n \times n$ real symmetric matrix and let λ_1 and λ_2 be distinct (real) e-values. Then the corresponding (real) e-vectors are orthogonal.*

Proof: To show $\langle v^1, v^2\rangle = 0$, we apply Proposition 3.55 with $x = v^1$ and $y = v^2$.

$$\begin{aligned}
\langle Av^1, v^2\rangle &= \langle v^1, Av^2\rangle \\
\Updownarrow& \\
\langle \lambda_1 v^1, v^2\rangle &= \langle v^1, \lambda_2 v^2\rangle \\
\Updownarrow& \\
\lambda_1\langle v^1, v^2\rangle &= \lambda_2\langle v^1, v^2\rangle \\
\Updownarrow& \\
0 &= (\lambda_1 - \lambda_2)\langle v^1, v^2\rangle.
\end{aligned}$$

But because λ_1 and λ_2 are distinct, $(\lambda_1 - \lambda_2) \neq 0$ and hence $\langle v^1, v^2\rangle = 0$. ■

Proposition 3.59 *The e-vectors of an $n \times n$ real symmetric matrix can always be chosen to form an orthonormal basis of $\mathbb{R}^n$.*

Proof: Proposition 3.58 handles the case that the e-values of A are distinct. A HW assignment will treat the case of repeated e-values. ■

Remark 3.60 *Real symmetric matrices are special in that one can ALWAYS find a basis of $\mathbb{R}^n$ consisting of e-vectors. Recall that this is false for general (real) matrices as shown by the example $A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$.*

Definition 3.61 *An $n \times n$ real matrix Q is **orthogonal** if $Q^\top Q = I$. If we decompose $Q =: \begin{bmatrix} Q_1 & Q_2 & \cdots & Q_n \end{bmatrix}$ into its columns, then by the rules of matrix multiplication,*

$$\langle Q_i, Q_j\rangle = Q_i^\top Q_j =: [Q^\top Q]_{ij} = [I]_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j. \end{cases}$$

*Hence, Q is **orthogonal** when its columns form an orthonormal basis of $\mathbb{R}^n$.*

Remark 3.62 *Q square and $Q^\top Q = I \implies Q^{-1} = Q^\top$.*

Proposition 3.63 *Suppose A is an $n \times n$ real symmetric matrix. Then there exists an orthogonal matrix Q such that*

$$Q^\top AQ = \Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n).$$

Proof: By Proposition 3.59, there exists an orthonormal basis of $\mathbb{R}^n$ such that $Av^i = \lambda_i v^i$, $1 \leq i \leq n$. We define

$$Q := \begin{bmatrix} Q_1 & Q_2 & \cdots & Q_n \end{bmatrix} =: \begin{bmatrix} v^1 & v^2 & \cdots & v^n \end{bmatrix},$$

so that $Q^\top Q = I$ and $AQ = Q\Lambda$; see Theorem 2.57. Hence $\Lambda = Q^{-1}AQ = Q^\top AQ$.

■

Exercise 3.64 *Let Q be an $n \times n$ orthogonal matrix and consider $(\mathbb{R}^n, \mathbb{R}, \|\bullet\|)$ with the Euclidean norm. Then $\forall\, x \in \mathbb{R}^n$, $\|Qx\| = \|x\|$, that is, **orthogonal matrices are norm preserving**. (Hint: Work with $\|Qx\|^2$.)*

3.6 Quadratic Forms, Positive Definite Matrices, and Schur Complements

We are building toward estimation (or best approximation) problems where some measurements are "less certain" that others, and hence we need unequal weights on our error terms.

Remark 3.65 *(Useful Observation) Let A be $m \times n$ real matrix. Then both $A^\top A$ and $AA^\top$ are symmetric, and hence their eigenvalues are real.*

Claim 3.66 *The eigenvalues of $A^\top A$ and $AA^\top$ are non-negative (real numbers).*

Proof: We do the proof for $A^\top A$. Let $A^\top Av = \lambda v$ where $v \in \mathbb{R}^n$, $v \neq 0$, $\lambda \in \mathbb{R}$, $v \in \mathbb{R}^n$. To show: $\lambda \geq 0$.

Multiplying both sides of $A^\top Av = \lambda v$ by $v^\top$ on the left yields

$$\begin{aligned} v^\top A^\top Av &= v^\top \lambda v \\ &\Updownarrow \\ \langle Av, Av\rangle &= \lambda \langle v, v\rangle \\ &\Updownarrow \\ \|Av\|^2 &= \lambda \|v\|^2. \end{aligned}$$

Because $\|v\|^2 > 0$ and $\|Av\|^2 \geq 0$, it follows that $\lambda \geq 0$. ■

Definition 3.67 *Let M be an $n \times n$ real matrix and $x \in \mathbb{R}^n$. Then $x^\top Mx$ is called a **quadratic form**.*

Definition 3.68 *An $n \times n$ real matrix W is **skew symmetric** if $W^\top = -W$.*

Exercise 3.69 *If W is skew symmetric, then $x^\top Wx = 0$ for all $x \in \mathbb{R}^n$. (Hint: use the fact that a real number is equal to its transpose to show that $x^\top Wx = x^\top W^\top x = -x^\top Wx$.)*

Exercise 3.70 *Any real matrix M can be written as ,*

$$M = \underbrace{\frac{M + M^\top}{2}}_{\text{symmetric}} + \underbrace{\frac{M - M^\top}{2}}_{\text{skew symmetric}}.$$

Definition 3.71 $\frac{M+M^\top}{2}$ *is called the **symmetric part** of M.*

Exercise 3.72 *For any real square matrix M, $x^\top Mx = x^\top \left(\frac{M+M^\top}{2}\right) x$.*

Remark 3.73 *Consequently, when working with a quadratic form, one always assumes M is symmetric.*

Proposition 3.74 ***(E-value Bounds of Symmetric Matrices)** Let M be an $n \times n$ real symmetric matrix. Then $\forall\, x \in \mathbb{R}^n$,*

$$\lambda_{\min}\, x^\top x \leq x^\top Mx \leq \lambda_{\max}\, x^\top x,$$

where $\lambda_{\min} := \min\{\lambda_1, \ldots, \lambda_n\}$ and $\lambda_{\max} := \max\{\lambda_1, \ldots, \lambda_n\}$.

Proof: By Proposition 3.59, we can choose an orthonormal basis $v = \{v^1, v^2, \ldots, v^n\}$ of $\mathbb{R}^n$ consisting of eigenvectors of M. For $x \in \mathbb{R}^n$, we expand it in terms of the basis vectors as $x = \alpha_1 v^1 + \alpha_2 v^2 + \cdots + \alpha_n v^n$, and then using the fact that the basis is orthonormal, we have that

$$x^\top x = \sum_{i=1}^n \alpha_i^2.$$

Next, we use the fact that the v^i are e-vectors so that $Mx = \alpha_1\lambda_1 v^1 + \alpha_2\lambda_2 v^2 + \cdots + \alpha_n\lambda_n v^n$. Another straightforward calculation gives that

$$x^\top Mx = \sum_{i=1}^n \lambda_i \alpha_i^2.$$

It follows that

$$\min\{\lambda_1,\ldots,\lambda_n\}\left(\sum_{i=1}^n \alpha_i^2\right) \leq \sum_{i=1}^n \lambda_i\alpha_i^2 \leq \max\{\lambda_1,\ldots,\lambda_n\}\left(\sum_{i=1}^n \alpha_i^2\right)$$

and hence

$$\lambda_{\min}\, x^\top x \leq x^\top Mx \leq \lambda_{\max}\, x^\top x.$$

■

Definition 3.75 *A real symmetric matrix P is **positive definite** if* $(\forall\, x \in \mathbb{R}^n, x \neq 0) \implies (x^\top Px > 0)$.

Notation 3.76 *One writes $P > 0$ to indicated that P is **positive definite**.* **To be absolutely clear, $P > 0$ *does not mean* that all entries of P are positive!**

Theorem 3.77 ***(E-value Test for a Symmetric Matrix)*** *A symmetric real matrix P is* positive definite *if, and only if, all of its eigenvalues are strictly greater than zero.*

Proof: From Proposition 3.74, if $\lambda_{\min} > 0$, then for all $0 \neq x \in \mathbb{R}^n$, $x^\top Px > 0$ and hence P is positive definite. For the other direction, suppose that $0 \neq v \in \mathbb{R}^n$ satisfies $Pv = \lambda v$ and $\lambda \leq 0$. Then $v^\top Pv = \lambda v^\top v = \lambda\|v\|^2 \leq 0$, and hence P is not positive definite. ■.

Exercise 3.78 *Show* $P_a := \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} > 0$ *and* $P_b := \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$ *is not positive definite. (Hint: Compute their e-values.)*

Definition 3.79 *$P = P^\top$ is **positive semidefinite** if $x^\top Px \geq 0$ for all $x \in \mathbb{R}^n$.*

Exercise 3.80 *Show the following:*

- *$P \geq 0 \implies$ its diagonal entries are non-negative. (Hint: Try x with a single one and the remaining entries all zero.)*
- *$P > 0 \implies$ its diagonal entries are positive.*
- *The above conditions are necessary but not sufficient for P to be positive semi-definite and positive definite, respectively.*

Theorem 3.81 ***(E-value Test for** $P \geq 0$**)*** *P is* positive semidefinite *if, and only if, all of its eigenvalues are non-negative.*

Proof: Follow the proof of Theorem 3.77 *mutatis mutandis*, that is, follow the proof and make the necessary small adjustments. ■

Notation 3.82 *One writes $P \geq 0$ or $P \succcurlyeq 0$.*

Definition 3.83 *N is a **square root of a real symmetric matrix** P if $N^\top N = P$. We note that $N^\top N = \left(N^\top N\right)^\top \implies N^\top N$ is always symmetric.*

Remark 3.84 There are several notions of a square root of a matrix. We are using the simplest one. *From Wikipedia (https://en.wikipedia.org/wiki/Square_root_of_a_matrix) In mathematics, the square root of a matrix extends the notion of square root from numbers to matrices. A matrix B is said to be a square root of A if the matrix product BB is equal to A. Some authors use the name square root or the notation $A^{1/2}$ only for the specific case when A is positive semidefinite, to denote the unique matrix B that is positive semidefinite and such that $BB = B^\top B = A$ for real-valued matrices. Less frequently, the name square root may be used for any factorization of a positive semidefinite matrix A as $B^\top B = A$, as in the Cholesky factorization, even if $BB \neq A$.* We are doing the latter.

Theorem 3.85 *$P \geq 0$ if, and only if, $\exists\, N$ such that $N^\top N = P$.*

Proof:

1. Suppose $N^\top N = P$, and let $x \in \mathbb{R}^n$. Then,

$$x^\top Px = x^\top N^\top Nx = (Nx)^\top (Nx) = \|Nx\|^2 \geq 0.$$

and hence P is positive semi-definite.

2. Now suppose $P \geq 0$. To show: $\exists\, N$ such that $N^\top N = P$.

   Since P is real and symmetric, there exists an orthogonal matrix O such that

$$P = O^\top \Lambda O$$

   where $\Lambda = \operatorname{diag}(\lambda_1, \lambda_2, \cdots, \lambda_n)$. Since $P \geq 0$, by Theorem 3.81, $\lambda_i \geq 0,\ 1 \leq i \leq n$. Define

$$\Lambda^{1/2} := \operatorname{diag}(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \ldots, \sqrt{\lambda_n}),$$

   so that

$$(\Lambda^{1/2})^\top \Lambda^{1/2} = \Lambda^{1/2}\Lambda^{1/2} = \Lambda.$$

   Let $N = \Lambda^{1/2}O$, then

$$N^\top N = O^\top \left(\Lambda^{1/2}\right)^\top \Lambda^{1/2} O = O^\top \Lambda O = P.$$

   Therefore $N^\top N = P$. ■

Exercise 3.86 *For an $n \times n$ real symmetric matrix P and $x, y \in \mathbb{R}^n$, prove that*

$$(x+y)^\top P(x+y) = x^\top Px + y^\top Py + 2x^\top Py.$$

(Hint: $y^\top Px$ and $x^\top Py$ are scalars.)

Theorem 3.87 ***(Schur Complements)*** *Suppose that A, B and C are real matrices, $A = n \times n$ is symmetric, $B = n \times m$, and $C = m \times m$ is symmetric. Then for the symmetric matrix*

$$M = \begin{bmatrix} A & B \\ B^\top & C \end{bmatrix},$$

the following statements are equivalent:

(a) $M > 0$.

(b) $A > 0$, *and* $C - B^\top A^{-1} B > 0$.

(c) $C > 0$, *and* $A - BC^{-1}B^\top > 0$.

Proof: We will show $(a) \iff (b)$. The proof of $(a) \iff (c)$ is identical. First, let's show (a) $\implies$ (b). Suppose $M > 0$, Then for all $x \in \mathbb{R}^n$, $x \neq 0$,

$$\begin{bmatrix} x \\ 0 \end{bmatrix}^\top M \begin{bmatrix} x \\ 0 \end{bmatrix} > 0.$$

Expanding this out gives

$$0 < \begin{bmatrix} x \\ 0 \end{bmatrix}^\top \begin{bmatrix} A & B \\ B^\top & C \end{bmatrix} \begin{bmatrix} x \\ 0 \end{bmatrix} = \begin{bmatrix} x^\top & 0 \end{bmatrix} \begin{bmatrix} Ax \\ B^\top x \end{bmatrix} = x^\top Ax$$

and therefore A is positive definite. We will make a nice choice of x and y to show $C - B^\top A^{-1} B > 0$. Let $y \neq 0$ be otherwise arbitrary and suppose we choose x such that $Ax + By = 0$. This is motivated by zeroing the top row of M when multiplying by the vector $\begin{bmatrix} x^\top & y^\top \end{bmatrix}^\top$. Such a choice of x is always possible because we know that $A > 0$, which implies A is invertible and hence $x = -A^{-1}By$ satisfies $Ax + By = 0$. Using this pair of x and y,

$$\begin{aligned} 0 < \begin{bmatrix} x \\ y \end{bmatrix}^\top \begin{bmatrix} A & B \\ B^\top & C \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} &= \begin{bmatrix} -A^{-1}By \\ y \end{bmatrix}^\top \begin{bmatrix} A & B \\ B^\top & C \end{bmatrix} \begin{bmatrix} -A^{-1}By \\ y \end{bmatrix} \\ &= \begin{bmatrix} -y^\top B^\top A^{-1} & y^\top \end{bmatrix} \begin{bmatrix} 0 \\ -B^\top A^{-1}By + Cy \end{bmatrix} \\ &= y^\top Cy - y^\top B^\top A^{-1} By \\ &= y^\top (C - B^\top A^{-1} B) y. \end{aligned}$$

Hence, $(y \in \mathbb{R}^m, y \neq 0)$, implies $y^\top (C - B^\top A^{-1} B) y > 0$, and therefore $C - B^\top A^{-1} B > 0$.

To complete the proof, we show (b) $\implies$ (a). Hence, we suppose $A > 0$, $C - B^\top A^{-1} B > 0$ and seek to show $M > 0$. For an arbitrary $\begin{bmatrix} x \\ y \end{bmatrix}$, define $\bar{x} = x + A^{-1}By$ and note that $\begin{bmatrix} x \\ y \end{bmatrix} \neq \begin{bmatrix} 0 \\ 0 \end{bmatrix} \iff \begin{bmatrix} \bar{x} \\ y \end{bmatrix} \neq \begin{bmatrix} 0 \\ 0 \end{bmatrix}$. Substituting in and applying Exercise 3.86 yields

$$\begin{bmatrix} x \\ y \end{bmatrix}^\top M \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \bar{x} - A^{-1}By \\ y \end{bmatrix}^\top M \begin{bmatrix} \bar{x} - A^{-1}By \\ y \end{bmatrix}$$

$$= \underbrace{\begin{bmatrix} \bar{x} \\ 0 \end{bmatrix}^\top M \begin{bmatrix} \bar{x} \\ 0 \end{bmatrix}}_{\bar{x}^\top A \bar{x}} + \underbrace{\begin{bmatrix} -A^{-1}By \\ y \end{bmatrix}^\top M \begin{bmatrix} -A^{-1}By \\ y \end{bmatrix}}_{y^\top (C - B^\top A^{-1} B) y} + 2 \underbrace{\begin{bmatrix} \bar{x} \\ 0 \end{bmatrix}^\top M \begin{bmatrix} -A^{-1}By \\ y \end{bmatrix}}_{0}$$

$$= \bar{x}^\top A \bar{x} + y^\top (C - B^\top A^{-1} B) y > 0,$$

and thus M is positive definite. ■

Definition 3.88 *$C - B^\top A^{-1} B$ is the **Schur Complement of** A **in** M and $A - BC^{-1}B^\top$ is **the Schur Complement of** C **in** M.*

Example 3.89 *Find conditions for the matrix*

$$M = \begin{bmatrix} a & b \\ b & c \end{bmatrix}_{2 \times 2}$$

to be positive definite.

Solution: By the Schur-Complement Theorem, $M > 0$ if, and only if

$$\begin{array}{c} a > 0 \ \& \ c - ba^{-1}b > 0 \\ \Updownarrow \\ a > 0 \ \& \ ac - b^2 > 0 \\ \Updownarrow \\ a > 0 \ \& \ \det(M) > 0. \end{array}$$

■

Example 3.90 *Test whether the given matrices are positive definite or not.*

$$M_1 = \begin{bmatrix} 3 & -2 \\ -2 & 3 \end{bmatrix}, M_2 = \begin{bmatrix} 2 & 3 \\ 3 & 2 \end{bmatrix}, \text{ and } M_3 = \begin{bmatrix} \alpha & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 3 \end{bmatrix},$$

where $\alpha \in \mathbb{R}$.

Solution: For M_1, we can apply the result of Example 3.89 to determine that $a = 2 > 0$ and $\det(M_1) = 5 > 0$, and hence $M_1 > 0$. Doing the same for M_2, we have that $a = 2 > 0$ and $\det(M_2) = -5 < 0$ and hence M_2 is not positive definite.

The matrix M_3 is more interesting. We define

$$A = [\alpha], B = \begin{bmatrix} 1 & 1 \end{bmatrix}, \text{ and } C = \begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix}$$

and apply condition (c) of Theorem 3.87. From the result in Example 3.89 we quickly see that $C > 0$. Hence, we next form

$$A - BC^{-1}B^\top = \alpha - \begin{bmatrix} 1 & 1 \end{bmatrix} C^{-1} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \alpha - 0.6.$$

Hence, $M > 0$ for $\alpha > 0.6$.

What if we use the same decomposition of M and apply condition (b) of Theorem 3.87? Then we test

$$A > 0 \iff \alpha > 0$$
$$C - B^\top A^{-1} B > 0 \iff \begin{bmatrix} 2 - \frac{1}{\alpha} & 1 - \frac{1}{\alpha} \\ 1 - \frac{1}{\alpha} & 3 - \frac{1}{\alpha} \end{bmatrix} > 0.$$

It's more complicated to work out, but by applying Example 3.89 again, we end up with three conditions

$$\alpha > 0 \iff \alpha \in (0, \infty)$$
$$2 - \frac{1}{\alpha} > 0 \implies (\alpha < 0) \vee (\alpha > 1/2) \iff \alpha \in (-\infty, 0) \cup (1/2, \infty)$$
$$(5\alpha - 3)/\alpha > 0 \implies (\alpha < 0) \vee (\alpha > 3/5) \iff \alpha \in (-\infty, 0) \cup (3/5, \infty),$$

which must be simultaneously true[1]. All of these conditions are met for $\alpha > 0.6$, which agrees with our previous computation.

The reader is encouraged to try a different partition of the matrix M_3, such as

$$A = \begin{bmatrix} \alpha & 1 \\ 1 & 2 \end{bmatrix}, B = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \text{ and } C = [3].$$

The condition obtained for α will be the same. The question is, which partition of M_3 and which condition of Theorem 3.87 yield the easiest test (for a particular problem)?

■

3.7 Least Squares Problems

Our first step is to solve an extension of Example 3.43, where we now allow a more general quadratic error term.

Proposition 3.91 ***(Weighted Least Squares, Overdetermined Equations)****. Let S be an $n \times n$ positive definite matrix $(S > 0)$ and let the inner product on $\mathbb{R}^n$ be*

$$\langle x, y \rangle := x^\top S y,$$

so that $\|x\|^2 = \langle x, x \rangle = x^\top S x$. Consider the overdetermined equation $A\alpha = b$, where $A = n \times m, n \geq m, \operatorname{rank}(A) = m, \alpha \in \mathbb{R}^m$, and $b \in \mathbb{R}^n$. Then

$$\widehat{\alpha} = \underset{\alpha \in \mathbb{R}^m}{\arg\min} \|A\alpha - b\|^2 \iff (A^\top S A)\widehat{\alpha} = A^\top S b \iff \widehat{\alpha} = (A^\top S A)^{-1} A^\top S b.$$

Proof: The inner product space is $\mathcal{X} = \mathbb{R}^n, \mathcal{F} = \mathbb{R}, \langle x, y \rangle = x^\top S y$. We partition by columns as $A = \begin{bmatrix} A_1 & A_2 & \cdots & A_m \end{bmatrix}$. Then the normal equations give

$$\begin{aligned} \widehat{x} &= \widehat{\alpha}_1 A_1 + \widehat{\alpha}_2 A_2 + \cdots + \widehat{\alpha}_m A_m \\ G^\top \widehat{\alpha} &= \beta, \text{ with } G = G^\top \\ [G^\top]_{ij} &= [G]_{ij} = \langle A_i, A_j \rangle = A_i^\top S A_j = [A^\top S A]_{ij} \\ \beta_i &= \langle b, A_i \rangle = b^\top S A_i = A_i^\top S b = [A^\top S b]_i \\ &\Updownarrow \\ A^\top S A \widehat{\alpha} &= A^\top S b. \end{aligned}$$

Because $\operatorname{rank}(A) = m$ implies the columns of A are linearly independent, we know that the Gram matrix $A^\top S A$ is invertible. Hence,

$$\widehat{\alpha} = (A^\top S A)^{-1} A^\top S b.$$

■

[1] A great way to think about it is $M_3 > 0$ if, and only if, $\alpha \in S_1 \cap S_2 \cap S_3$, where $S_1 = (0, \infty)$, $S_2 = (-\infty, 0) \cup (1/2, \infty)$, and $S_3 = (-\infty, 0) \cup (3/5, \infty)$. Calculation of the indicated set intersections yields $S_1 \cap S_2 \cap S_3 = (3/5, \infty)$.

Proposition 3.92 ***(Recursive Least Squares)*** *Consider the model*

$$\textbf{\textit{Model:}} \begin{cases} y_i = C_i x + e_i,\ i = 1, 2, 3, \cdots \\ \\ C_i \in \mathbb{R}^{m \times n} \\ \quad i = \text{time index} \\ \quad x = \text{an unknown } \underline{\text{constant}} \text{ vector} \in \mathbb{R}^n \\ \quad y_i = \text{measurements} \in \mathbb{R}^m \\ \quad e_i = \text{model "mismatch"} \in \mathbb{R}^m \end{cases} \tag{3.9}$$

for generating the data stream $\{y_1, y_2, \ldots\}$. *Let* k_0 *be the smallest* $k \geq 1$ *such that* $\text{rank}([C_1^\top\ C_2^\top\ \cdots\ C_{k_0}^\top]) = n$. *Then, for all* $k \geq k_0$, *a solution to*

$$\begin{aligned} \widehat{x}_k :&= \arg\min_{x \in \mathbb{R}^n} \left(\sum_{i=1}^{k} (y_i - C_i x)^\top S_i (y_i - C_i x) \right) \\ &= \arg\min_{x \in \mathbb{R}^n} \left(\sum_{i=1}^{k} e_i^\top S_i e_i \right) \end{aligned}$$

where $S_i = m \times m$ *positive definite matrix (*$S_i > 0$ *for all time indices* i*) can be computed* ***recursively*** *by*

$$\widehat{x}_{k+1} = \widehat{x}_k + \underbrace{P_{k+1} C_{k+1}^\top S_{k+1}}_{\text{"Kalman gain"}} \underbrace{(y_{k+1} - C_{k+1} \widehat{x}_k)}_{\text{"Innovations"}}.$$

$$P_{k+1} = P_k - P_k C_{k+1}^\top \left[C_{k+1} P_k C_{k+1}^\top + S_{k+1}^{-1} \right]^{-1} C_{k+1} P_k$$

$$P_{k_0} := \left[\sum_{i=1}^{k_0} C_i^\top S_i C_i \right]^{-1}$$

Proof:

Step 1 is a Batch Solution that can be used for initialization: For $k \geq 1$, define

$$Y_k = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_k \end{bmatrix}, A_k = \begin{bmatrix} C_1 \\ C_2 \\ \vdots \\ C_k \end{bmatrix}, E_k = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_k \end{bmatrix}, \text{ and } R_k = \begin{bmatrix} S_1 & & \mathbf{0} & \\ & S_2 & & \\ & & \ddots & \\ \mathbf{0} & & & S_k \end{bmatrix} = \text{diag}(S_1, S_2, \cdots, S_k) > 0.$$

Then Y_k satisfies $Y_k = A_k x + E_k, k \geq 1$.

Claim 3.93 *Suppose* $\text{rank}(A_{k_0}) = n$, *the dimension of* x. *Then, for all* $k \geq k_0, \text{rank}(A_k) = n$.

Proof: For all $k \geq 1$, $\text{rank}(A_k) \leq n$, the number of columns of A_k. For $k > k_0$,

$$A_k = \begin{bmatrix} A_{k_0} \\ C_{k_0+1} \\ \vdots \\ C_k \end{bmatrix}$$

and hence $n = \text{rank}(A_{k_0}) \leq \text{rank}(A_k) \leq n$, and thus $\text{rank}(A_k) = n$ for all $k \geq k_0$. □

Next we note that $\|Y_k - A_k x\|^2 = (Y_k - A_k x)^\top R_k (Y_k - A_k x)$, and thus, by Proposition 3.91, for all $k \geq k_0$,

$$\widehat{x}_k := \arg\min_{x \in \mathbb{R}^n} \|E_k\|^2 = \arg\min_{x \in \mathbb{R}^n} \|Y_k - A_k x\|^2,$$

satisfies the Normal Equations, $(A_k^\top R_k A_k) \widehat{x}_k = A_k^\top R_k Y_k$. This yields what is called a **Batch Solution**

$$\boxed{\widehat{x}_k = (A_k^\top R_k A_k)^{-1} A_k^\top R_k Y_k},$$

because all of the measurements are processed together as a batch to best approximate x. The drawback is that A_k is a $km \times n$ matrix, and grows at each step. On a real robot, if you were collecting data at a KHz or more, you'd quickly fill your memory and forming $A_k^\top R_k A_k$ would take longer and longer at each step. The solution is to find a recursive means to compute $\widehat{x}_{k+1}$ in terms of $\widehat{x}_k$ and the new measurement y_{k+1}.

Step 2: Seeking a recursive solution. The normal equations at time $k \geq k_0$, $(A_k^\top R_k A_k)\widehat{x}_k = A_k^\top R_k Y_k$, are equivalent to

$$\left(\sum_{i=1}^{k} C_i^\top S_i C_i\right) \widehat{x}_k = \sum_{i=1}^{k} C_i^\top S_i y_i.$$

We define

$$Q_k = \sum_{i=1}^{k} C_i^\top S_i C_i$$

so that

$$Q_{k+1} = Q_k + C_{k+1}^\top S_{k+1} C_{k+1}.$$

And the normal equations at time $k+1$ become,

$$\underbrace{(\sum_{i=1}^{k+1} C_i^\top S_i C_i)}_{Q_{k+1}} \widehat{x}_{k+1} = \sum_{i=1}^{k+1} C_i^\top S_i y_i$$

or

$$Q_{k+1}\widehat{x}_{k+1} = \underbrace{\sum_{i=1}^{k} C_i^\top S_i y_i}_{Q_k \widehat{x}_k} + C_{k+1}^\top S_{k+1} y_{k+1}.$$

We now have a good start on recursion, namely,

$$\boxed{\begin{aligned} Q_{k+1} &= Q_k + C_{k+1}^\top S_{k+1} C_{k+1} \\ Q_{k+1}\widehat{x}_{k+1} &= Q_k \widehat{x}_k + C_{k+1}^\top S_{k+1} y_{k+1}. \end{aligned}}$$

The estimate at time $k+1$ is expressed as a linear combination of the estimate at time k and the latest measurement at time $k+1$.

Continuing,

$$\widehat{x}_{k+1} = Q_{k+1}^{-1} \left[Q_k \widehat{x}_k + C_{k+1}^\top S_{k+1} y_{k+1} \right].$$

Because $Q_k = Q_{k+1} - C_{k+1}^\top S_{k+1} C_{k+1}$, we have

$$\widehat{x}_{k+1} = \widehat{x}_k + \underbrace{Q_{k+1}^{-1} C_{k+1}^\top S_{k+1}}_{\text{Kalman gain}} \underbrace{(y_{k+1} - C_{k+1}\widehat{x}_k)}_{\text{Innovations}}.$$

The term $(y_{k+1} - C_{k+1}\widehat{x}_k)$ is called the **innovations** because it is effectively the "new information" provided by the measurement at time $k+1$. If there were no measurement, we could **predict** an estimate for y at time $k+1$ by $\widehat{y}_{k+1} := C_{k+1}\widehat{x}_k$ (recall that our model assumes that x is a constant vector, hence if we have an estimate of x at time step k, our best guess for it at time $k+1$, absent any further measurements, would simply be that it is unchanged from our previous estimate). The innovation is

$$(y_{k+1} - \widehat{y}_{k+1}) = (y_{k+1} - C_{k+1}\widehat{x}_k),$$

the difference between what we measure and what we predict. If our prediction does not change with y_{k+1}, then the measurement was not "innovative". The appellation "Kalman Gain" does not have any grounding at this point. But when we treat the Kalman filter a bit later in these notes, you'll see that the Kalman filter has a measurement update gain that looks just like the one in RLS (recursive least squares).

In a real-time implementation, computing the inverse of Q_{k+1} can be time consuming. An attractive alternative can be obtained by applying the **Matrix Inversion Lemma**,

$$(A+BCD)^{-1} = A^{-1} - A^{-1}B\left(DA^{-1}B + C^{-1}\right)^{-1} DA^{-1}$$

Now, following the substitution rule

$$A \leftrightarrow Q_k \quad B \leftrightarrow C_{k+1}^\top \quad C \leftrightarrow S_{k+1} \quad D \leftrightarrow C_{k+1},$$

yields, after some tedious calculations,

$$\begin{aligned} Q_{k+1}^{-1} &= \left(Q_k + C_k^\top S_{k+1} C_{k+1}\right)^{-1} \\ &= Q_k^{-1} - Q_k^{-1} C_{k+1}^\top \left[C_{k+1} Q_k^{-1} C_{k+1}^\top + S_{k+1}^{-1}\right]^{-1} C_{k+1} Q_k^{-1}, \end{aligned}$$

which is a recursion for Q_k^{-1}. Upon defining

$$P_k = Q_k^{-1},$$

we have

$$P_{k+1} = P_k - P_k C_{k+1}^\top \left[C_{k+1} P_k C_{k+1}^\top + S_{k+1}^{-1}\right]^{-1} C_{k+1} P_k$$

We note that we are now inverting a matrix that is $m \times m$, instead of one that is $n \times n$. Typically, $n >> m$ (means very much greater than), and thus the savings can be important.

■

Remark 3.94 ***(Recursive least squares (RLS) with a "forgetting factor")*** *is treated in HW #06. The forgetting factor allows you to exponentially "discount" older measurements. This is important if the paramter you are trying to estimate is slowly drifting.*

We consider $Ax = b$ and recall what we know about its solutions:

- $b \in \text{col span}\{A\} \iff$ a solution exists;
- the solution is unique if, and only if, the columns of A are linearly independent; and thus
- if there exists one solution and the columns of A are linearly dependent, then there exist an infinite number of solutions.

Underdetermined Equations

The columns of A will be linearly dependent when $Ax = b$ has fewer equations than unknowns. In other words, A is $n \times m$ and $m > n$; sometimes these are called wide matrices: more columns than rows. When dealing with an equation $Ax = b$ with fewer equations than unknowns, one says that it is **underdetermined**. Why? Because, to determine x uniquely, at a minimum, we need as many equations as unknowns.

Is there a difference between being underdetermined and having an infinite number of solutions? Yes. It's possible to be underdetermined and have no solution at all when $b \notin \text{col span}\{A\}$. If the rows of A are linearly independent, then

$$Ax = b \text{ is underdetermined} \iff Ax = b \text{ has an infinite number of solutions.}$$

The rows of A being linearly independent is equivalent to the columns of $A^\top$ being linearly independent.

When $Ax = b$ has an infinite number of solutions, is there a way that we can make one of them appear to be more interesting, more special, or just flat out "better" than all the other solutions? Is there a property that we could associate with each solution and optimize our choice of solution with respect to that property? The most common approach is to choose the solution with minimum norm!

Proposition 3.95 ***(Underdetermined Equations)*** *Consider the real finite dimensional inner product space $(\mathbb{R}^n, \mathbb{R}, \langle \bullet, \bullet \rangle)$ where $< x, z >:= x^\top S z$ and $S > 0$. If the rows of A are linearly independent, then*

$$\widehat{x} := \underset{Ax=b}{\arg\min} ||x|| = \underset{Ax=b}{\arg\min} ||x||^2$$

satisfies $\widehat{x} = S^{-1}A^\top \beta$, $AS^{-1}A^\top \beta = b$ or, equivalently, $\widehat{x} = S^{-1}A^\top \left(AS^{-1}A^\top\right)^{-1} b$.

Proof: The main idea is to express the constraint $Ax = b$ in terms of the inner product, which means to identify vectors $\{v^1, \ldots, v^p\}$ and constants $c_1, \ldots, c_p$ such that $Ax = b \iff \langle v^i, x\rangle = c_i$, $1 \leq i \leq p$, so that Theorem 3.51 is applicable. To identify the required vectors and constants, we partition A by rows, that is

$$A =: \begin{bmatrix} a_1 \\ \vdots \\ a_p \end{bmatrix},$$

and note that $Ax = b \iff AS^{-1}Sx = b$. Then, based upon the row by row interpretation of $AS^{-1}Sx = b$, we define $v_i \in \mathbb{R}^n$ by

$$(v_i)^\top := a_i S^{-1} \iff v_i = (a_i S^{-1})^\top = S^{-1} a_i^\top,$$

where we have used the fact that S is symmetric. It follows that

$$Ax = b \iff (a_i x = b_i),\ i = 1, ..., p \iff (a_i S^{-1} S x = b_i),\ i = 1, ..., p \iff \langle v_i,\ x\rangle = b_i,\ i = 1, ..., p.$$

Hence, by Theorem 3.51,

$$\widehat{x} = \sum_{i=1}^{p} \beta_i v_i = \begin{bmatrix} v_1 & v_2 & \cdots & v_p \end{bmatrix} \beta,$$

where $G^\top \beta = b$ and G is the Gram matrix. The last part is to show that

$$\begin{bmatrix} v_1 & v_2 & \cdots & v_p \end{bmatrix} = \begin{bmatrix} S^{-1}a_1^\top & S^{-1}a_2^\top & \cdots & S^{-1}a_p^\top \end{bmatrix} = S^{-1} \begin{bmatrix} a_1^\top & a_2^\top & \cdots & a_p^\top \end{bmatrix} = S^{-1}A^\top$$

and that $G^\top = G = AS^{-1}A^\top$ because

$$G_{ij} := \langle v_i,\ v_j\rangle := v_i^\top S v_j = (a_i S^{-1}) S (S^{-1} a_j^\top) = a_i S^{-1} a_j^\top = \left[AS^{-1}A^\top\right]_{ij}.$$

■

Examples of working with these results are included in the HW sets.

Chapter 4

Three Useful Matrix Factorizations

Learning Objectives

- Introduce the notion of factoring a matrix as a means of solving systems of linear equations
- See important applications of the work we have done in Chapters 2 and 3.

Outcomes

- Learn how to compute and use a QR Factorization.
- Understand that the theoretical definition of linear independence may not be adequate for engineering practice.
- Learn about the Singular Value Decomposition, a workhorse in numerical linear algebra, that addresses the above issue and much more.
- Learn the LU factorization, the LDLT (or Cholesky) Factorization, and their uses.

4.1 QR Factorization

Definition 4.1 *An $n \times m$ matrix R is **upper triangular** if $R_{ij} = 0$ for all $i > j$.*

Theorem 4.2 *(QR Decomposition or Factorization) Let A be a real $n \times m$ matrix with **linearly independent columns**. Then there exist an $n \times m$ matrix Q with **orthonormal columns** and an $m \times m$* upper triangular *matrix R such that*

$$A = QR.$$

Remark 4.3

1. $Q^\top Q = I_{m\times m}$
2. $R = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1(m-1)} & r_{1m} \\ 0 & r_{22} & \cdots & r_{2(m-1)} & r_{2m} \\ 0 & 0 & r_{33} & \cdots & r_{3m} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & 0 & r_{mm} \end{bmatrix}$
3. *Columns of A linearly independent $\iff$ R is invertible*

Proof: The proof is organized around the computation of the factorization by the *Gram-Schmidt Algorithm with Normalization*. Partition A into columns, $A = \begin{bmatrix} A_1 & A_2 & \cdots & A_m \end{bmatrix}$, $A_i \in \mathbb{R}^n$, and use the inner product $\langle x, y\rangle = x^\top y$. For $1 \le k \le n$, $\{A_1,\ A_2,\ \cdots,\ A_m\} \to \{v_1,\ v_2,\ \cdots,\ v_m\}$ by

for $k = 1 : m$
  $v^k = A_k$
  for $j = 1 : k-1$
    $v^k = v^k - \langle A_k, v^j\rangle v^j$
  end
  $v^k = \frac{v^k}{\|v^k\|}$
end

By construction, $Q := \begin{bmatrix} v^1 & v^2 & \cdots & v^m \end{bmatrix}$ has orthonormal columns, and hence $Q^\top Q = I_{m\times m}$ because $[Q^\top Q]_{ij} = \langle v^i, v^j\rangle = 1, i = j$ and zero otherwise.

What about R? By construction, $A_i \in \text{span}\{v^1, \cdots, v^i\}$, with $A_i = \langle A_1, v^1\rangle v^1 + \langle A_2, v^2\rangle v^2 + \cdots + \langle A_i, v^i\rangle v^i$. We define

$$R_i := \begin{bmatrix} \langle A_1, v^1\rangle \\ \vdots \\ \langle A_i, v^i\rangle \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

where $R_{ij} = 0$ for $i < j \le n$. The coefficients in R can be extracted directly from the Gram-Schmdit Algorithm; no extra computations are required. By construction, $A_i = QR_i$ and thus we have $A = QR$.

Note: $R_i = [A_i]_{\{v^1,\ldots,v^m\}}$, the representation of A_i in the basis $\{v^1, \ldots, v^m\}$.

■

Example 4.4 ***(QR Decomposition of Overdetermined Equations)*** *Suppose that $Ax = b$ is overdetermined with columns of A linearly*

independent. Write $A = QR$ and consider

$$\begin{aligned} A^\top A\widehat{x} &= A^\top b \\ &\Updownarrow \\ R^\top Q^\top QR\widehat{x} &= R^\top Q^\top b \\ &\Updownarrow \\ R^\top R\widehat{x} &= R^\top Q^\top b \qquad \text{(because } Q^\top Q = I\text{)} \\ &\Updownarrow \\ R\widehat{x} &= Q^\top b \qquad \text{(because } R \text{ is invertible)} \end{aligned}$$

Hence, we can solve for $\widehat{x}$ by back substitution using the triangular nature of R. For example, when $n = 3$

$$\begin{bmatrix} r_{11} & r_{12} & r_{13} \\ 0 & r_{22} & r_{23} \\ 0 & 0 & r_{33} \end{bmatrix} \widehat{x} = Q^\top b,$$

and therefore, $\widehat{x}_3$ to $\widehat{x}_1$ can be obtained easily without performing a matrix inverse.

Example 4.5 ***(QR Decomposition of Underdetermined Equations)*** *Suppose $Ax = b$ is underdetermined with rows of A linearly independent. For the inner product $\langle x, y\rangle = x^\top y,\ \widehat{x} = A^\top(AA^\top)^{-1}b$ is the value of x of smallest norm satisfying $Ax = b$.*

$A^\top$ has linearly independent columns, and hence we write $A^\top = QR,\ Q^\top Q = I,\ R$ is upper triangular and invertible. It follows that

$$\begin{aligned} \widehat{x} &= A^\top(AA^\top)^{-1}b \\ &\Updownarrow \\ \widehat{x} &= QR\left(R^\top Q^\top QR\right)^{-1}b \\ &\Updownarrow \\ \widehat{x} &= QR(R^\top R)^{-1}b \\ &\Updownarrow \\ \widehat{x} &= QRR^{-1}(R^\top)^{-1}b \\ &\Updownarrow \\ \widehat{x} &= Q(R^\top)^{-1}b. \end{aligned}$$

■

4.2 Singular Value Decomposition or SVD

The material here is inspired by a handout prepared by Prof. James Freudenberg, EECS, University of Michigan.

4.2.1 Motivation

In abstract linear algebra, a set of vectors is either linearly independent or not. There is nothing in between. For example, the set of vectors

$$\left\{ v_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, v_2 = \begin{bmatrix} 0.999 \\ 1 \end{bmatrix} \right\}$$

is linearly independent. In this case, one looks at the set of vectors and says, yes, BUT, the vectors are "almost" dependent because when one computes the determinant

$$\det \begin{bmatrix} 1 & 0.999 \\ 1 & 1 \end{bmatrix} = 0.001,$$

the result is pretty small, so it should be fine to call them dependent.

Well, what about the set

$$\left\{v_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, v_2 = \begin{bmatrix} 10^4 \\ 1 \end{bmatrix}\right\}?$$

When you form the matrix and check the determinant, you get

$$\det \begin{bmatrix} 1 & 10^4 \\ 0 & 1 \end{bmatrix} = 1,$$

which seems pretty far from zero. So are these vectors "adequately" linearly independent?

Maybe not! Let's note that

$$\det\left(\begin{bmatrix} 1 & 10^4 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 10^{-4} & 0 \end{bmatrix}\right) = \det \begin{bmatrix} 1 & 10^4 \\ 10^{-4} & 1 \end{bmatrix} = 0,$$

and hence it's possible to add a very small perturbation to one of the vectors and make the set linearly dependent! This cannot be good.

4.2.2 Definition and Main Theorem

Definition 4.6 *(Rectangular Diagonal Matrix) An $n \times m$ matrix Σ is a **Rectangular Diagonal Matrix** if*

$$\Sigma_{ij} = 0 \text{ for } i \neq j.$$

*The **diagonal** of Σ is the set of all Σ_{ii}, $1 \leq i \leq \min(n, m)$. An alternative and equivalent way to define a Rectangular Diagonal Matrix is*

(a) (tall matrix) $n > m$ $\quad \Sigma = \begin{bmatrix} \Sigma_d \\ 0 \end{bmatrix}$, where Σ_d is an $m \times m$ diagonal matrix.

(b) (wide matrix) $n < m$ $\quad \Sigma = \begin{bmatrix} \Sigma_d & 0 \end{bmatrix}$, where Σ_d is an $n \times n$ diagonal matrix.

*The **diagonal** of Σ is equal to the diagonal of Σ_d.*

Theorem 4.7 ***(SVD or Singular Value Decomposition)*** *Every $n \times m$ real matrix A can be factored as*

$$A = U \cdot \Sigma \cdot V^\top,$$

where U is an $n \times n$ orthogonal matrix, V is an $m \times m$ orthogonal matrix, Σ is an $n \times m$ rectangular diagonal matrix, and the diagonal of Σ,

$$\text{diag}(\Sigma) = [\sigma_1, \sigma_2, \cdots, \sigma_p],$$

*satisfies $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p \geq 0$, for $p := \min(n, m)$. Moreover, the columns of U are eigenvectors of $A \cdot A^\top$, the columns of V are eigenvectors of $A^\top \cdot A$, and $\{\sigma_1^2, \sigma_2^2, \ldots, \sigma_p^2\}$ are eigenvalues of both $A^\top \cdot A$ and $A \cdot A^\top$. The **Singular Values of A** are the elements $\{\sigma_1, \ldots, \sigma_p\}$ from the diagonal of Σ.*

Remark 4.8 *The entries of $\text{diag}(\Sigma)$ are called **singular values** of A.*

Proof: $A^\top A$ is $m \times m$, real, and symmetric. Hence, there exists a set of orthonormal eigenvectors $\{v^1, \ldots, v^m\}$ such that

$$A^\top A v^j = \lambda_j v^j.$$

Without loss of generality, we can assume that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m \geq 0$. If not, we simply re-order the v^i's to make it so. For $\lambda_j > 0$, say $1 \leq j \leq r$, we define

$$\sigma_j = \sqrt{\lambda_j}$$

and

$$q^j = \frac{1}{\sigma_j} A v^j \in \mathbb{R}^n$$

Claim 4.9 *For* $1 \leq i,\ j \leq r$, $\left(q^i\right)^\top q^j = \delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$. *That is, the vectors* $\{q^1, q^2, \ldots, q^r\}$ *are orthonormal.*

Proof of Claim:

$$\begin{aligned}
\left(q^i\right)^\top q^j &= \frac{1}{\sigma_i}\frac{1}{\sigma_j}\left(v^i\right)^\top A^\top A v^j \\
&= \frac{\lambda_j}{\sigma_i\sigma_j}\left(v^i\right)^\top v^j \\
&= \begin{cases} \frac{\lambda_i}{(\sigma_i)^2} & i = j \\ 0 & i \neq j \end{cases} \\
&= \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}
\end{aligned}$$

□

Claim 4.10 *The vectors* $\{q^1, q^2, \ldots, q^r\}$ *are eigenvectors of* $AA^\top$ *and the corresponding e-values are* $\{\lambda_1, \lambda_2, \ldots, \lambda_r\}$.

Proof of Claim: For $1 \leq i \leq r$, $\lambda_i > 0$ and

$$\begin{aligned}
AA^\top q^i :=& AA^\top\left(\frac{1}{\sigma_i}Av^i\right) \\
=& \frac{1}{\sigma_i}A\left(A^\top A\right)v^i \\
=& \frac{\lambda_i}{\sigma_i}Av^i \\
=& \lambda_i q^i,
\end{aligned}$$

and thus q^i is an e-vector of $AA^\top$ with e-value λ_i. The claim is also an immediate consequence of Lemma 2.63. □

From Fact 2.61, if $r < n$, then the remaining e-values of $AA^\top$ are all zero. Moreover, we can extend the q^i's to an orthonormal basis for $\mathbb{R}^n$ satisfying $AA^\top q^i = 0$, for $r+1 \leq i \leq n$. Define

$$U := \begin{bmatrix} q^1 & q^2 & \cdots & q^n \end{bmatrix} \text{ and } V := \begin{bmatrix} v^1 & v^2 & \cdots & v^m \end{bmatrix}.$$

Also, define $\Sigma = n \times m$ by

$$\Sigma_{ij} = \begin{cases} \sigma_i\delta_{ij} & 1 \leq i,\ j \leq r \\ 0 & \text{otherwise.} \end{cases}$$

Then, Σ is rectangular diagonal with

$$\operatorname{diag}(\Sigma) = [\sigma_1,\ \sigma_2,\ \cdots,\ \sigma_r,\ 0,\ \cdots,\ 0]$$

To complete the proof of the theorem, it is enough to show[1] that $U^\top AV = \Sigma$. We note that the ij element of this matrix is

$$(U^\top AV)_{ij} = q_i^\top Av^j$$

If $j > r$, then $A^\top Av^j = 0$, and thus $(q^i)^\top Av^j = 0$, as required. If $i > r$, then q^i was selected to be orthogonal to

$$\{q^1,\ \cdots,\ q^r\} = \{\frac{1}{\sigma_1}Av^1,\ \frac{1}{\sigma_2}Av^2,\ \cdots,\ \frac{1}{\sigma_r}Av^r\}$$

and thus $\left(q^i\right)^\top Av^j = 0$. Hence we now consider $1 \leq i,\ j \leq r$ and compute that

$$\begin{aligned}
\left(U^\top AV\right)_{ij} &= \frac{1}{\sigma_i}\left(v^i\right)^\top A^\top Av^j \\
&= \frac{\lambda_j}{\sigma_i}\left(v^i\right)^\top v^j \\
&= \sigma_i\delta_{ij}
\end{aligned}$$

[1] Because $U^\top U = I$ and $V^\top V = I$, it follows that $A = U\Sigma V^\top \iff U^\top AV = \Sigma$.

as required.

■

Remark 4.11 *Another way to write the SVD of A is*

$$A = \sigma_1 u_1 \cdot v_1^\top + \sigma_2 u_2 \cdot v_2^\top + \cdots + \sigma_p u_p \cdot v_p^\top,$$

where u_i and v_i are columns of U and V respectively.

$$U = \begin{bmatrix} u_1 & u_2 & \cdots & u_n \end{bmatrix} \text{ and } V = \begin{bmatrix} v_1 & v_2 & \cdots & v_m \end{bmatrix}. \tag{4.1}$$

This formula follows from Fact2.67, matrix multiplication based on the sum over columns times rows, where we note that the columns of V are the rows of $V^\top$.

Example 4.12 *Determine the SVD of A as well as its rank and nullity,*

$$A = \begin{bmatrix} 1 & 10^4 \\ 0 & 1 \end{bmatrix}.$$

Solution: Using the LinearAlgebra package in Julia, we find

$$U = \begin{bmatrix} 1.000e+00 & -1.000e-04 \\ 1.000e-04 & 1.000e+00 \end{bmatrix}$$
$$\Sigma = \begin{bmatrix} 1.000e+04 & 0.000e+00 \\ 0.000e+00 & 1.000e-04 \end{bmatrix}$$
$$V = \begin{bmatrix} 1.000e-04 & -1.000e+00 \\ 1.000e+00 & 1.000e-04 \end{bmatrix}$$

There are two non-zero singular values, and thus $r = 2$. It follows that $\mathrm{rank}(A) = 2$ and $\mathrm{nullity}(A) = 0$.

Information about the "near" linear dependence of the columns of A is in the diagonal matrix Σ. There are two singular values, $\sigma_1 = 10^4$ and $\sigma_2 = 10^{-4}$. Their ratio is 10^8, which is an indicator that these vectors are "nearly linearly dependent". "Numerically", one would say that $r = 1$ and hence $\mathrm{rank}(A) = r = 1$ and $\mathrm{nullity}(A) = 2 - r = 1$. ■

4.2.3 Numerical Linear Independence

Illustration: 5×5 matrix. For

$$A = \begin{bmatrix} -32.57514 & -3.89996 & -6.30185 & -5.67305 & -26.21851 \\ -36.21632 & -11.13521 & -38.80726 & -16.86330 & -1.42786 \\ -5.07732 & -21.86599 & -38.27045 & -36.61390 & -33.95078 \\ -36.51955 & -38.28404 & -19.40680 & -31.67486 & -37.34390 \\ -25.28365 & -38.57919 & -31.99765 & -38.36343 & -27.13790 \end{bmatrix},$$

and the Julia commands

```
using LinearAlgebra

A=[-32.57514 -3.89996 -6.30185 -5.67305 -26.21851;
-36.21632 -11.13521 -38.80726 -16.86330 -1.42786;
-5.07732 -21.86599 -38.27045 -36.61390 -33.95078;
-36.51955 -38.28404 -19.40680 -31.67486 -37.34390;
-25.28365 -38.57919 -31.99765 -38.36343 -27.13790 ]

(U ,Sigma, V) = svd(A)
```

one obtains

$$U = \begin{bmatrix} -2.475e-01 & -5.600e-01 & 4.131e-01 & 5.759e-01 & 3.504e-01 \\ -3.542e-01 & -5.207e-01 & -7.577e-01 & -1.106e-02 & -1.707e-01 \\ -4.641e-01 & 6.013e-01 & -1.679e-01 & 6.063e-01 & -1.652e-01 \\ -5.475e-01 & -1.183e-01 & 4.755e-01 & -3.314e-01 & -5.919e-01 \\ -5.460e-01 & 1.992e-01 & -2.983e-02 & -4.369e-01 & 6.859e-01 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} 1.325e+02 & 0.000e+00 & 0.000e+00 & 0.000e+00 & 0.000e+00 \\ 0.000e+00 & 3.771e+01 & 0.000e+00 & 0.000e+00 & 0.000e+00 \\ 0.000e+00 & 0.000e+00 & 3.342e+01 & 0.000e+00 & 0.000e+00 \\ 0.000e+00 & 0.000e+00 & 0.000e+00 & 1.934e+01 & 0.000e+00 \\ 0.000e+00 & 0.000e+00 & 0.000e+00 & 0.000e+00 & 7.916e-01 \end{bmatrix}$$

$$V = \begin{bmatrix} 4.307e-01 & 8.839e-01 & -5.303e-02 & 8.843e-02 & -1.503e-01 \\ 4.309e-01 & -2.207e-01 & -1.961e-01 & 7.322e-01 & 4.370e-01 \\ 4.617e-01 & -8.902e-02 & 7.467e-01 & -3.098e-01 & 3.539e-01 \\ 4.730e-01 & -3.701e-01 & 7.976e-02 & 1.023e-01 & -7.890e-01 \\ 4.380e-01 & -1.585e-01 & -6.283e-01 & -5.913e-01 & 1.968e-01 \end{bmatrix}$$

Because the **smallest singular value $\sigma_5 = 0.7916$ is less than 1% of the largest singular value $\sigma_1 = 132.5$**, in many cases, one would say that the numerical rank of A was 4 instead of 5.

This notion of numerical rank can be formalized by asking the following question: Suppose $\text{rank}(A) = r$. How far away is A from a matrix of rank strictly less than r?

The numerical rank of a matrix is based on the expansion in (4.11), which is repeated here for convenience,

$$A = U \cdot \Sigma \cdot V^\top = \sum_{i=1}^{p} \sigma_i u_i \cdot v_i^\top = \sigma_1 u_1 \cdot v_1^\top + \sigma_2 u_2 \cdot v_2^\top + \cdots + \sigma_p u_p \cdot v_p^\top,$$

where $p = \min\{m, n\}$, and once again, the singular values are ordered such that $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p \geq 0$. Each term $u_i \cdot v_i^\top$ is a rank-one matrix. The following will help you understand the expansion.

Exercises or Facts:

- $A \cdot A^\top = U \cdot \Sigma \cdot \Sigma^\top \cdot U^\top = \sum_{i=1}^{p} (\sigma_i)^2 \, u_i \cdot u_i^\top$
- $A^\top \cdot A = V \cdot \Sigma^\top \cdot \Sigma \cdot V^\top = \sum_{i=1}^{p} (\sigma_i)^2 \, v_i \cdot v_i^\top$
- $\left(u_i \cdot v_i^\top\right) \cdot v_j = \begin{cases} u_i & j = i \\ 0 & j \neq i \end{cases}$ and hence $\text{rank}(u_i \cdot v_i^\top) = 1$ and $\text{nullity}(u_i \cdot v_i^\top) = m - 1$
- $\left(u_i \cdot u_i^\top\right) \cdot u_j = \begin{cases} u_i & j = i \\ 0 & j \neq i \end{cases}$ and hence $\text{rank}(u_i \cdot u_i^\top) = 1$ and $\text{nullity}(u_i \cdot u_i^\top) = n - 1$
- $\left(v_i \cdot v_i^\top\right) \cdot v_j = \begin{cases} v_i & j = i \\ 0 & j \neq i \end{cases}$ and hence $\text{rank}(v_i \cdot v_i^\top) = 1$ and $\text{nullity}(v_i \cdot v_i^\top) = m - 1$
- $v_i \cdot v_i^\top$, and $u_i \cdot u_i^\top$ have e-values $\lambda_1 = 1$ distinct and $\lambda_2 = 0$ repeated $m - 1$ and $n - 1$ times, respectively.
- **Hint:** $\left(u_i \cdot v_i^\top\right) \cdot v_j = u_i \cdot \left(v_i^\top \cdot v_j\right) = \begin{cases} u_i & j = i \\ 0 & j \neq i \end{cases}$ because the $\{v_1, v_2, \ldots, v_m\}$ are orthonormal.

So far, we have only defined the norm of a vector. However, it is also useful to measure the "length" of matrices.

Definition 4.13 ***(Induced Matrix Norm)*** *Given an $n \times m$ real matrix A, the* ***matrix norm induced by the Euclidean vector norm*** *is given by:*

$$||A|| := \max_{x^\top x=1} ||Ax|| = \sqrt{\lambda_{\max}(A^\top A)}$$

where $\lambda_{\max}(A^\top A)$ denotes the largest eigenvalue of the matrix $A^\top A$. ***(Recall that the matrices of the form $A^\top A$ are at least positive semidefinite and hence their e-values are real and non-negative. Therefore, the square root exists.)***

Numerical Rank

Facts: Suppose that $\text{rank}(A) = r$, so that σ_r is the smallest non-zero singular value of A.

(i) If an $n \times m$ matrix E satisfies $||E|| < \sigma_r$, then $\text{rank}(A + E) \geq r$.

(ii) There exists an $n \times m$ matrix E with $||E|| = \sigma_r$ and $\text{rank}(A + E) < r$.

(iii) In fact, for $E = -\sigma_r u_r v_r^\top$, $\text{rank}(A + E) = r - 1$.

(iv) Moreover, for $E = -\sigma_r u_r v_r^\top - \sigma_{r-1} u_{r-1} v_{r-1}^\top$, $\text{rank}(A + E) = r - 2$.

Corollary: Suppose A is square and invertible. Then σ_r measures the distance from A to the nearest singular matrix.

Illustration Continued

```
u5=U[:,5]; v5=V[:,5]; sig5=Sigma[5]
E=-sig5*u5*v5'
# Induced Norm
M=E'*E
SquareRootEigs=(abs.(eigvals(E'*E))).^0.5
#
(U ,Sigma2, V) = svd(A+E)
```

$$E = \begin{bmatrix} 4.169e-02 & -1.212e-01 & -9.818e-02 & 2.189e-01 & -5.458e-02 \\ -2.031e-02 & 5.906e-02 & 4.784e-02 & -1.066e-01 & 2.659e-02 \\ -1.966e-02 & 5.716e-02 & 4.629e-02 & -1.032e-01 & 2.574e-02 \\ -7.041e-02 & 2.048e-01 & 1.658e-01 & -3.697e-01 & 9.220e-02 \\ 8.160e-02 & -2.373e-01 & -1.922e-01 & 4.284e-01 & -1.068e-01 \end{bmatrix}$$

$$\sqrt{\lambda_i(E^\top \cdot E)} = \begin{bmatrix} 7.376e-09 \\ 2.406e-09 \\ 1.977e-09 \\ 4.163e-09 \\ 7.916e-01 \end{bmatrix}$$

$$\Sigma_2 = \begin{bmatrix} 1.325e+02 & 0.000e+00 & 0.000e+00 & 0.000e+00 & 0.000e+00 \\ 0.000e+00 & 3.771e+01 & 0.000e+00 & 0.000e+00 & 0.000e+00 \\ 0.000e+00 & 0.000e+00 & 3.342e+01 & 0.000e+00 & 0.000e+00 \\ 0.000e+00 & 0.000e+00 & 0.000e+00 & 1.934e+01 & 0.000e+00 \\ 0.000e+00 & 0.000e+00 & 0.000e+00 & 0.000e+00 & 1.775e-15 \end{bmatrix}$$

We added a matrix with norm 0.7916 and made the (exact) rank drop from 4 to 5! How cool is that? This example shows that SVD can exactly measure how close a matrix is to being singular. We also see that $E^\top \cdot E$ has rank one: there is one non-zero e-value and

the rest are (essentially) zero as the theory promised.

Other Interesting and Useful Facts

(a) **Null space:** $\text{null}(A) := \{x \in \mathbb{R}^m \mid Ax = 0\}$

(b) **Range:** $\text{range}(A) := \{y \in \mathbb{R}^n \mid \text{ such that } \ y = Ax \ \text{ for some } \ x \in \mathbb{R}^m\}$

(c) **Fact:** Suppose $A = U \cdot \Sigma \cdot V^\top$. Then the columns of U corresponding to non-zero singular values are a basis for $\text{range}(A)$ and the columns of V corresponding to zero singular values are a basis for $\text{null}(A)$, viz

$$\begin{aligned} \text{range}(A) &:= \text{span}\{u_1, ..., u_r\}, \text{and} \\ \text{null}(A) &:= \text{span}\{v_{r+1}, ..., v_m\}. \end{aligned}$$

(d) The SVD can also be used to compute an "effective" range and an "effective" null space of a matrix.

(e) **Fact:** Suppose that $\sigma_1 \geq ... \geq \sigma_r > \delta \geq \sigma_{r+1} \geq ...\sigma_p \geq 0$, so that r is the "effective" or "numerical rank" of A. (Note the δ inserted between σ_r and σ_{r+1} to denote the break point.)

(f) **Fact:** Let $\text{range}_{\text{eff}}(A)$ and $\text{null}_{\text{eff}}(A)$ denote the effective range and effective null space of A, respectively. Then we can calculate bases for these subspaces by choosing appropriate singular vectors:

$$\begin{aligned} \text{range}_{\text{eff}}(A) &:= \text{span}\{u_1, ..., u_r\}, \text{and} \\ \text{null}_{\text{eff}}(A) &:= \text{span}\{v_{r+1}, ..., v_m\}. \end{aligned}$$

4.3 Lower Upper (LU) Factorization

This material comes from ROB 101 Computational Linear Algebra. The textbook and more can be found at https://github.com/michiganrobotics/rob101/tree/main/Fall%202021.

Definition 4.14 *An $n \times n$ matrix P consisting of only zeros and ones and satisfying $P^\top P = PP^\top = I$ is called a **permutation matrix**.*

Exercise 4.15 *A permutation matrix can be viewed in two ways: (a) as a permutation of the rows of an identity matrix; or (b) as a permutation of the columns of an identity matrix. Hence, it is common and useful to identify an $n \times n$ permutation matrix P with a list of indices $p = \{i_1, i_2, \ldots, i_n\}$ formed by permuting the list $\{1, 2, \ldots, n\}$. Show the following:*

- *Each row and each column of a permutation matrix has exactly one* 1.
- *Let $x \in \mathbb{R}^n$, I be the $n \times n$ identity matrix, and define $P := I[p, :]$, a row permutation of I. Then $Px = \begin{bmatrix} x_{i_1} & x_{i_2} & \cdots & x_{i_n} \end{bmatrix}^\top$. If $P := I[:, p]$, a column permutation of I, then $x^\top P = \begin{bmatrix} x_{i_1} & x_{i_2} & \cdots & x_{i_n} \end{bmatrix}$.*
- *Every $n \times n$ permutation matrix can be written as $P := I[p, :]$ and as $P := I[:, \widetilde{p}]$ for appropriate permutations p and $\widetilde{p}$ of the list $\{1, 2, \ldots, n\}$.*
- *Multiplying a matrix A on the left by a permutation matrix (of appropriate size) permutes the order of its rows, while multiplying it on the right permutes the order of its columns.*
- *The product of two permutation matrices is a permutation matrix.*

Definition 4.16 *A possibly rectangular matrix L is **lower triangular** if all entries above the diagonal are zero. A possibly rectangular matrix U is **upper triangular** if all entries below the diagonal are zero. Recall that the **diagonal** of an $n \times m$ matrix M consists of all entries of the form m_{ii}, $1 \leq i \leq \min\{n, m\}$. L is uni-lower triangular if its diagonal consists of all ones. By a slight abuse of terminology, we'll allow an empty matrix to be called uni-lower triangular because its diagonal, being empty, has no terms that violate the definition.*

Fact 4.17 *Assume the matrices in the following are non-empty.*

- *If M is square and either upper or lower triangular, then its determinant is given by the product of the terms on its diagonal.*
- *If the lower triangular matrix L is square and has non-zero determinant, then the equation $Lx = b$ can be solved by forward substitution; see the ROB 101 textbook.*
- *If the upper triangular matrix U is square and has non-zero determinant, then the equation $Ux = b$ can be solved by back substitution; see the ROB 101 textbook.*

As a lead in to LU Factorization, you can check that if L is a lower triangular matrix and U is an upper triangular matrix, then in general, their product $A := LU$ is neither. Can this process be reversed? That is, given a generic square matrix, can it be factored as the product of a lower-triangular matrix and an upper-triangular matrix? And if we can do such a factorization, would it be helpful?

Example 4.18 *The goal here is to show you the secret sauce that underlies a very nice method for constructing the required triangular matrices. We call it* **peeling the onion:** ***starting from the top left corner and working down the diagonal, it successively zeros out columns and rows of a matrix!*** *Consider the square matrix*

$$M = \begin{bmatrix} 1 & 4 & 5 \\ 2 & 9 & 17 \\ 3 & 18 & 58 \end{bmatrix}.$$

Our goal is to find a column vector C_1 and a row vector R_1 such that

$$M - C_1 \cdot R_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & * & * \\ 0 & * & * \end{bmatrix},$$

where $*$ denotes "don't care" in the sense that we do not care about their particular values. We want to zero out the first column and the first row of M. That means, C_1 and R_1 are chosen so that the first column and first row of their matrix product $C_1 \cdot R_1$ match the first column and first row of M. How can you do that?

We perform a ***special case of "peeling the onion" that works when the top left entry equals 1.0****. The general case will be treated later.*

We define C_1 and R_1 to be the first column of M and the first row of M, respectively, that is

$$C_1 = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \text{ and } R_1 = \begin{bmatrix} 1 & 4 & 5 \end{bmatrix}.$$

Then

$$C_1 \cdot R_1 = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \cdot \begin{bmatrix} 1 & 4 & 5 \end{bmatrix} = \begin{bmatrix} 1 & 4 & 5 \\ 2 & 8 & 10 \\ 3 & 12 & 15 \end{bmatrix},$$

and voilà,

$$M = \begin{bmatrix} 1 & 4 & 5 \\ 2 & 9 & 17 \\ 3 & 18 & 58 \end{bmatrix} \text{ and } C_1 \cdot R_1 = \begin{bmatrix} 1 & 4 & 5 \\ 2 & 8 & 10 \\ 3 & 12 & 15 \end{bmatrix}.$$

Consequently,

$$\begin{aligned} M - C_1 \cdot R_1 &= \begin{bmatrix} 1 & 4 & 5 \\ 2 & 9 & 17 \\ 3 & 18 & 58 \end{bmatrix} - \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \cdot \begin{bmatrix} 1 & 4 & 5 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 4 & 5 \\ 2 & 9 & 17 \\ 3 & 18 & 58 \end{bmatrix} - \begin{bmatrix} 1 & 4 & 5 \\ 2 & 8 & 10 \\ 3 & 12 & 15 \end{bmatrix} \\ &= \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 7 \\ 0 & 6 & 43 \end{bmatrix}. \end{aligned}$$

Oh! We have taken a 3×3 matrix and essentially made it into a 2×2 matrix!! Can we do this again? Let's try. We define C_2 and R_2 to be the second column and second row of $M - C_1 \cdot R_1$, that is

$$C_2 = \begin{bmatrix} 0 \\ 1 \\ 6 \end{bmatrix} \text{ and } R_2 = \begin{bmatrix} 0 & 1 & 7 \end{bmatrix}.$$

Then we compute that

$$\begin{bmatrix} 0 \\ 1 \\ 6 \end{bmatrix} \cdot \begin{bmatrix} 0 & 1 & 7 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 7 \\ 0 & 6 & 42 \end{bmatrix},$$

and we obtain

$$(M - C_1 \cdot R_1) = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 7 \\ 0 & 6 & 43 \end{bmatrix} \text{ and } C_2 \cdot R_2 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 7 \\ 0 & 6 & 42 \end{bmatrix}.$$

Consequently,

$$\begin{aligned} (M - C_1 \cdot R_1) - C_2 \cdot R_2 &= \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 7 \\ 0 & 6 & 43 \end{bmatrix} - \begin{bmatrix} 0 \\ 1 \\ 6 \end{bmatrix} \cdot \begin{bmatrix} 0 & 1 & 7 \end{bmatrix} \\ &= \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 7 \\ 0 & 6 & 43 \end{bmatrix} - \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 7 \\ 0 & 6 & 42 \end{bmatrix} \\ &= \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \end{aligned}$$

Oh! Now we are essentially down to a 1×1 matrix!! You might be seeing the pattern! We very quickly note that if we define C_3 and R_3 to be the third column and third row of $M - C_1 \cdot R_1 - C_2 \cdot R_2$,

$$C_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \text{ and } R_3 = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix},$$

then

$$C_3 \cdot R_3 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

and hence, $M - C_1 \cdot R_1 - C_2 \cdot R_2 - C_3 \cdot R_3 = 0_{3\times 3}$. We prefer to write this as

$$\boxed{M = C_1 \cdot R_1 + C_2 \cdot R_2 + C_3 \cdot R_3 = \underbrace{\begin{bmatrix} C_1 & C_2 & C_3 \end{bmatrix}}_{L} \cdot \underbrace{\begin{bmatrix} R_1 \\ R_2 \\ R_3 \end{bmatrix}}_{U}}.$$

Moreover,

- $L := \begin{bmatrix} C_1 & C_2 & C_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 3 & 6 & 1 \end{bmatrix}$ *is **uni-lower triangular**,*
- $U := \begin{bmatrix} R_1 \\ R_2 \\ R_3 \end{bmatrix} = \begin{bmatrix} 1 & 4 & 5 \\ 0 & 1 & 7 \\ 0 & 0 & 1 \end{bmatrix}$ *is **upper triangular**, and*
- *$M = L \cdot U$, the product of a lower triangular matrix and an upper triangular matrix.*

■

In the example, we arranged that the first non-zero entry of C_i was equal to one. This is a particularly special case chosen to illustrate that, at least in some cases, one can systematically factor a matrix. We now build toward the general case.

Definition 4.19 *A $n \times m$ matrix A is **left zeroed of order** $0 \leq k \leq \min\{n, m\}$ if it has the form*

$$A = \begin{bmatrix} 0_{k\times k} & 0_{k\times(m-k)} \\ 0_{(n-k)\times k} & \widetilde{A} \end{bmatrix}.$$

Remark 4.20 *When $k = 0$, the matrices $0_{k\times k}$, $0_{k\times(m-k)}$, and $0_{(n-k)\times k}$ are empty. Hence, every matrix is* left-zeroed of order *zero. When $k = \min\{n, m\}$, then $\widetilde{A}$ is the empty matrix and hence A is identically zero.*

Lemma 4.21 ***(Peeling the Onion)*** *Suppose that A is an $n \times m$ left-zeroed matrix of order $0 \leq k < \min\{n, m\}$. Then there exist a permutation matrix P, a column vector C, and a row vector R such that*

(a) $PA - CR$ is left-zeroed of order $(k + 1)$,

(b) C and R have zeros in their first k entries,

(c) the $(k + 1)$-st entry of C equals one (i.e., $[C]_{k+1} = 1.0$), and

(d) the first k rows of P are equal to the first k rows of the identity matrix.

Proof: We use proof by exhaustion to cover three cases. We let $[A]_{ij}$ and a_{ij} denote the ij entry of A, and a_i^{row} and a_j^{col} denote the i-th row and j-th column, respectively.

Case 1: Suppose $a_{k+1,k+1} \neq 0$. Then we define $P = I$, the identity matrix, $C := a_{k+1}^{\text{col}}/a_{k+1,k+1}$, and $R := a_{k+1}^{\text{row}}$. Note that the column C has been normalized by $a_{k+1,k+1}$, the first non-zero entry of C, and that $a_{k+1,k+1}$ is also the first non-zero entry of R. Moreover, by construction, the first k entries of C and R are zero. Based on these observations, we leave it to the reader to check that

$$[C \cdot R]_{ij} = [PA]_{ij}$$

for $i = k + 1$ and $j \in \{k + 1, \ldots, m\}$ and for $j = k + 1$, and $i \in \{k + 1, \ldots, n\}$. Hence, $PA - CR$ is left-zeroed of order $(k + 1)$. □

Case 2: Suppose $a_{k+1}^{\text{col}} = 0_{n\times 1}$. Then we define $P = I$, the identity matrix, the column vector C by

$$[C]_i := \begin{cases} 1 & i = (k + 1) \\ 0 & \text{otherwise} \end{cases},$$

and the row vector $R := a_{k+1}^{\text{row}}$. Note that as before, C has a one in its $(k+1)$-st entry, its first k entries are zero, and the first k entries of R are zero. Based on these observations, we leave it to the reader to check that

$$[C \cdot R]_{ij} = [PA]_{ij}$$

for $i = k + 1$ and $j \in \{k + 1, \ldots, m\}$ and for $j = k + 1$, and $i \in \{k + 1, \ldots, n\}$. Hence, $PA - CR$ is left-zeroed of order $(k + 1)$. □

Case 3: Suppose $a_{k+1,k+1} = 0$ and $a_{k+1}^{\text{col}} \neq 0_{n\times 1}$ so that there exists $k + 1 < \rho \leq n$ such that $a_{\rho,k+1} \neq 0$. Then we define p to be the unique permutation of $\{1, 2, \ldots, n\}$ such that

$$p_i = \begin{cases} \rho & i = k + 1 \\ k + 1 & i = \rho \\ i & i \notin \{k + 1, \rho\}. \end{cases}$$

Upon defining $P := I[p, :]$, the matrix $\check{A} := PA$ now satisfies the conditions of **Case 1** and its proof can be followed *mutatis mutandis* (i.e., by making the necessary simple changes). We thus leave the rest of the proof to the reader. □

■

Theorem 4.22 ***(LU Factorization)*** *Let A be an $n\times m$ real matrix and define $r = \min(n, m)$. There always exist an $n\times n$ permutation matrix P, an $n \times r$ uni-lower triangular matrix L, and an $r \times m$ upper triangular matrix U such that*

$$P \cdot A = L \cdot U.$$

Proof: We use proof by induction. At the base step $k = 0$, we set $P_0 := I$, and L_0, U_0 empty matrices.

At step $k \geq 0$, we assume that $A_k := P_k A - L_k U_k$ is a left-zeroed matrix of order k, L_k is an $n \times k$ uni-lower triangular matrix, and U_k is a $k \times m$ upper triangular matrix. By Lemma 4.21, there exist a permutation matrix P, column vector C, and row vector R such that

(a) $PA_k - CR$ is left-zeroed of order $(k+1)$,

(b) C and R have zeros in their first k entries,

(c) the $(k+1)$-st entry of C is one, and

(d) the first k rows of P are equal to the first k rows of the identity matrix.

Based on this we define

(i) $P_{k+1} := P \cdot P_k$;

(ii) $L_{k+1} := \begin{bmatrix} P \cdot L_k & C \end{bmatrix}$; and

(iii) $U_{k+1} := \begin{bmatrix} U_k \\ R \end{bmatrix}$.

Then

$$L_{k+1} \cdot U_{k+1} = \begin{bmatrix} P \cdot L_k & C \end{bmatrix} \cdot \begin{bmatrix} U_k \\ R \end{bmatrix} = P \cdot L_k \cdot U_k + C \cdot R,$$

and therefore

$$P_{k+1} \cdot A - L_{k+1} \cdot U_{k+1} = P \cdot P_k \cdot A - P \cdot L_k \cdot U_k - C \cdot R = P \cdot (P_k \cdot A - L_k \cdot U_k) - C \cdot R = P \cdot A_k - C \cdot R.$$

Hence, $P_{k+1} \cdot A - L_{k+1} \cdot U_{k+1}$ is left zeroed of order $(k+1)$.

Because L_k is an $n \times k$ uni-lower triangular and the first k rows of P are equal to the identity matrix, it follows that $P \cdot L_k$ is also uni-lower triangular. Because the first k entries of C are equal to zero and its $(k+1)$-st entry is one, it follows that $[P \cdot L_k \;\; C]$ is an $n \times (k+1)$ uni-lower triangular matrix. Because U_k is a $k \times m$ upper triangular matrix and the first k entries of R are equal to zero, it follows that $[U_k^\top \;\; R^\top]^\top$ is a $(k+1) \times m$ upper triangular matrix.

The algorithm stops at step $r = \min\{n, m\}$, producing the required matrices.

■

Solving $\mathrm{A}\mathbf{x} = \mathrm{b}$ via LU Factorization

We seek to solve the system of linear equations $Ax = b$, when A is a real square matrix. Suppose we factor $P \cdot A = L \cdot U$, where P is a permutation matrix, L is lower triangular and U is upper triangular. Would that even be helpful for solving linear equations?

Because $P^\top \cdot P = I$, $\det(P) = \pm 1$ and therefore P is always invertible. Hence,

$$Ax = b \iff P \cdot Ax = P \cdot b \iff L \cdot Ux = P \cdot b.$$

If we define $Ux = y$, then $L \cdot Ux = P \cdot b$ becomes two equations

$$Ly = P \cdot b \tag{4.2}$$
$$Ux = y. \tag{4.3}$$

Furthermore,

$$(P \cdot A = L \cdot U) \implies \det(A) = \pm \det(L) \det(U)$$

and A is invertible if, and only if, both L and U are invertible. Our solution strategy is therefore to solve (4.2) by forward substitution, and then, once we have y in hand, we solve (4.3) by back substitution to find x, the solution to $Ax = b$.

Example 4.23 *Use LU Factorization to solve the system of linear equations*

$$\underbrace{\begin{bmatrix} -2 & -4 & -6 \\ -2 & 1 & -4 \\ -2 & 11 & -4 \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}}_{x} = \underbrace{\begin{bmatrix} 2 \\ 3 \\ -7 \end{bmatrix}}_{b}. \tag{4.4}$$

Solution: We use the native LU function in Julia to compute $P \cdot A = L \cdot U$, with

$$\begin{aligned} P &= \begin{bmatrix} 1.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 1.0 \\ 0.0 & 1.0 & 0.0 \end{bmatrix} \\ L &= \begin{bmatrix} 1.000 & 0.000 & 0.000 \\ 1.000 & 1.000 & 0.000 \\ 1.000 & 0.333 & 1.000 \end{bmatrix} \\ U &= \begin{bmatrix} -2.000 & -4.000 & -6.000 \\ 0.000 & 15.000 & 2.000 \\ 0.000 & 0.000 & 1.333 \end{bmatrix}. \end{aligned} \tag{4.5}$$

Even though A admits an LU Factorization without row permutations, Julia inserts a permutation matrix. This is to improve the numerical accuracy on large problems. On our small problem, it's not really needed. Nevertheless, we'll use it to show that we obtain the same answer with essentially the same amount of work.

We first compute

$$Pb = \begin{bmatrix} 1.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 1.0 \\ 0.0 & 1.0 & 0.0 \end{bmatrix} \begin{bmatrix} 2.0 \\ 3.0 \\ -7.0 \end{bmatrix} = \begin{bmatrix} 2.0 \\ -7.0 \\ 3.0 \end{bmatrix}.$$

We then solve $Ly = Pb$ for the intermediate variable y, using forward substitution,

$$\underbrace{\begin{bmatrix} 1.000 & 0.000 & 0.000 \\ 1.000 & 1.000 & 0.000 \\ 1.000 & 0.333 & 1.000 \end{bmatrix}}_{L} \underbrace{\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}}_{y} = \underbrace{\begin{bmatrix} 2.0 \\ -7.0 \\ 3.0 \end{bmatrix}}_{Pb} \implies \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 2.0 \\ -9.0 \\ 4.0 \end{bmatrix}.$$

And finally, we use this result to solve $Ux = y$ for x, using back substitution,

$$\underbrace{\begin{bmatrix} -2.000 & -4.000 & -6.000 \\ 0.000 & 15.000 & 2.000 \\ 0.000 & 0.000 & 1.333 \end{bmatrix}}_{U} \underbrace{\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}}_{x} = \underbrace{\begin{bmatrix} 2.0 \\ -9.0 \\ 4.0 \end{bmatrix}}_{y} \implies \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -8.0 \\ -1.0 \\ 3.0 \end{bmatrix}.$$

■

4.4 LDLT or Cholesky Factorization (LU specialized for Positive Semi-definite Matrices)

Matrices that are positive semi-definite can be factored in a special form. We send the reader to the ROB 101 textbook for further details. To preserve the symmetric nature of positive semi-definite matrices when doing the factorization, it is necessary to use both row and column perturbations.

Enhanced LU Factorization and Rank of a Matrix

A real positive semi-definite matrix M always has an **LDLT Factorization (aka, Cholesky Factorization)**

$$P \cdot M \cdot P^\top = L \cdot D \cdot L^\top, \tag{4.6}$$

where

- P is a (row) permutation matrix;
- $P^\top$, the transpose of P, permutes the columns of M;
- L is uni-lower triangular and $L^\top$, the transpose of L, is therefore uni-upper triangular; and
- D is diagonal and has non-negative entries.

The terminology LDLT comes from L, D, L with the T being short for transpose. If $M = A^\top \cdot A$, then the number of non-zero entries on the diagonal of D is equal to $\mathrm{rank}(A)$.

Remark 4.24 *Recall that permutations in the LU Factorization only arise in Case 3 of Lemma 4.21 on "peeling the onion". Theorem 3.87 on Schur complements can be used to show that if Case 2 or Case 3 ever occurs, then M is not positive definite. Hence, for positive definite matrices, a simpler factorization is possible.*

Factorization of Positive Definite Matrices

A real positive definite matrix M always has an **LDLT Factorization (aka, Cholesky Factorization)** without requiring permutations,

$$M = L \cdot D \cdot L^\top, \tag{4.7}$$

where

- L is uni-lower triangular and $L^\top$, the transpose of L, is therefore uni-upper triangular; and
- D is diagonal and has positive entries.

Wikipedia https://en.wikipedia.org/wiki/Cholesky_decomposition provides additional information on Cholesky Factorization.

Chapter 5

Enough Probability and Estimation to Understand the Kalman Filter

Learning Objectives

- Learn enough probability that we gain a clear understanding of the Kalman Filter
- As foundations for the Kalman Filter, cover Best Linear Unbiased Estimation (BLUE) and Minimum Variance Estimation (MVE)

Outcomes

- Understand why probability is the most technical topic we have covered so far.
- Define (as best we can) a probability space.
- Cover random variables and random vectors.
- BLUE and MVE
- Understand the notion of conditional probability and how it relates to "fusing" measurements.
- Gaussian random vectors
- Our capstone topic in the Chapter is the Kalman filter, a recursive form of Minimum Variance Estimation.

5.1 Introduction

5.1.1 Intuition

Why use probability, much less even worry about what it means or how to use it properly? Let's back up first and ask why do we use mathematical models in Robotics, or in engineering?

When designing a system, it is common to combine mathematical models of individual components to predict the overall performance of the system, assuming known or hypothesized characteristics of the individual components. This saves us a lot time in terms of ordering the components, assembling, and then testing them. Followed by re-ordering, re-assembling, and re-testing everything before we have something that is close to a satisfactory system.

My lab focuses a lot of feedback control of bipedal robots. We design our controllers on the basis of mathematical models and then test the controllers on simulators that typically include terms that are more accurate than what we used in the control design model. We follow this process because in the end, we obtain a higher performing system (robot plus controller) in less time than if we went straight to the hardware and started "hacking" a controller together.

In summary, we use mathematical models to make predictions about how our robot would behave in an experiment, and we are motivated to do this because it leads to better results in less time, not to mention, very few things break when running a simulation, as opposed to when conducting actual experiments. When speaking of bipedal robots, you can imagine the models from Lagrangian mechanics that we use for the robot. What about the ground? The camera and LiDAR? the IMU?

Models of terrain must account for varying slope and texture, as well as the presence of holes and obstacles. At what frequency do changes in terrain characteristics occur and how do the characteristics of the terrain on which the robot is walking now depend on the terrain it has just traversed? Similarly, cameras are affected by the illumination of objects and by dust or smoke in the air. LiDAR returns are affected by surface texture and distance. The accelerometers in IMUs have biases (they read non-zero acceleration even when the robot is at rest, and these readings can vary with the temperature of the device). To date, engineers have found it mostly non-productive–if not absolutely impossible–to develop physics-based models for these effects. Instead, they have turned to descriptions based in "probability."

Probability has been approached through several lenses over the past few hundred years. In practice, the frequency interpretation is fairly widely adopted, wherein probabilities are numbers in the interval $[0, 1]$ that reflect the relative frequency of events, such as the relative occurrences of heads to tails in a coin or the relative frequency that a given pixel in a camera image will correspond to grass or a building, given that neighboring pixels have been identified as corresponding to a particular class of object. In the formal study of Statistics and Mathematics, the "frequentist" interpretation of probability has proven inadequate, leading to a formal definition of a probability space that parallels, in some sense, the formal definitions of fields, vector spaces, and normed spaces that we gave in earlier chapters. It is unfortunate that an equally careful development of probability theory is beyond the scope of ROB 501. We'll at least let you know where we are coming up short!

5.1.2 Suggested Online Material

- Very Elementary Review
  http://www.comp.nus.edu.sg/~cs5247/lecNotes/probability.pdf
- Medium Level Review: An Abridged Review of the Basic Theory of Probability
  https://people.math.wisc.edu/~anderson/431S14/ReviewSlidesV1.pdf
- Concise and amazingly clear
  http://webee.technion.ac.il/people/shimkin/Estimation09/ch2_Estimation.pdf
- Small Book on Probability, Meant as a Review
  https://www.cs.bham.ac.uk/~pxt/NIL/prob.stat.pdf
- Shorter, jumps straight into random variables
  https://studylib.net/doc/14227622/lecture-notes-1-1-probability-review-brief-review-of-basi...
- Starts with random vectors and moves into Gaussian or Normal random vectors
  https://www.probabilitycourse.com/chapter6/6_1_5_random_vectors.php

5.1.3 (Optional Read) Probability Spaces Provide a Means to Formalize the Theory of Probability

This section is meant to justify us taking a simplified (relaxed) approach to probability in ROB 501; basically, to do it right, you need to take Math 597 at Michigan. If you do not feel any particular need for a justification, then you can skip to the next section.

Definition 5.1 *$(\Omega, \mathscr{F}, P)$ is called a **probability space**.*

- *Ω is the sample space. Think of it as the set of all possible outcomes of an experiment.*
- *$E \subset \Omega$ is an event.*
- *$\mathscr{F}$ is the collection of allowed events[1]. It must at least contain $\emptyset$ and Ω. It is closed with respect to set complement, countable unions, and countable intersections[2]. Such sets are called sigma algebras https://en.wikipedia.org/wiki/%CE%A3-algebra.*

[1] Though it is too deep for ROB 501, there are subsets of the reals that are so complicated one cannot even define a reasonable notion of "probability" that agrees with how we would want to define a uniform probability on an interval, such as $[a, b]$.

[2] By De Morgan's laws, once a set is closed under set complements and countable unions, it is automatically closed under countable intersections; see https://en.wikipedia.org/wiki/De_Morgan.

- $P : \mathscr{F} \to [0,1]$ *is a probability measure. It has to satisfy a few basic operations*

  1. $P(\emptyset) = 0$ *and* $P(\Omega) = 1$.
  2. *For each* $E \in \mathscr{F}$, $0 \leq P(E) \leq 1$
  3. *If the sets* $E_1, E_2, \ldots$ *are disjoint (i.e.,* $E_i \cap E_j = \emptyset$ *for* $i \neq j$*), then*

$$P(\bigcup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} P(E_i).$$

  These are typically called the ***Axioms of Probability***.

Example 5.2 *Read reference* https://www.comp.nus.edu.sg/~cs5247/lecNotes/probability.pdf *on setting up the probability space for a (fair, or uniform) die. It defines* $\Omega = \{1,2,3,4,5,6\}$ *as the six faces of the die, but does not define* $\mathscr{F}$*, the collection of allowed events; hence, we'll do it. Because* Ω *is a finite set,* $\mathscr{F}$ *can be taken as the set of all possible subsets of* Ω*, namely*

$$\mathscr{F} = \{\emptyset, \{i_1\}_{i_1=1}^{6}, \{i_1, i_2\}_{\text{distinct}}, \ldots, \{i_1, \ldots, i_5\}_{\text{distinct}}, \Omega, \},$$

where $\{i_1, i_2\}_{\text{distinct}}$ *means all* $1 \leq i_1, i_2, \leq 6$*,* $i_1 \neq i_2$*. The die is fair or uniform when* $P(\{i\}) = P(\{j\})$ *for all* $i, j \in \Omega$*. The singletons* $\{i\}_{i=1}^{6}$ *as well as the set* $E := \{1,4,6\}$ *are allowed events (i.e.,*$\{2\} \in \mathscr{F}$ *and* $E \in \mathscr{F}$*). Because* E *is the disjoint union of the three singleton sets* $\{1\} \cup \{4\} \cup \{6\}$ *and the die is fair (uniform),*

$$P(E) = P(\{1\}) + P(\{4\}) + P(\{6\}) = 1/6 + 1/6 + 1/6 = 1/2.$$

■

Question 5.3 *The set of allowed events,* $\mathscr{F}$*, being the set of all subsets of* Ω *was already a bit awkward to write down explicitly for a die, but it was certainly doable. Is it always possible to write down* $\mathscr{F}$ *explicitly? And do we gain that much by doing it?*

Once we leave the simple settings of dice, balls in urns, etc., things become a lot more technically challenging. Take for example, $\Omega = S^1$, the circle, and instead of rolling a die, we spin a dial and check if it lands in a given subset of S^1. Let's identify S^1 with the interval $[0, 2\pi) \subset \mathbb{R}$ and suppose we want to define a "uniform probability measure" on it, by which we mean, if $a < b$ and $[a,b) \subset \Omega$, then $P([a,b)) = \frac{b-a}{2\pi}$. This seems to be a natural continuous extension of a fair (uniform) die, from six possible outcomes to a continuum of possible outcomes, where the notion of fairness is captured by two sets $[a_1, b_1) \subset \Omega$ and $[a_2, b_2) \subset \Omega$ having the same probability whenever, $b_1 - a_1 = b_2 - a_2$.

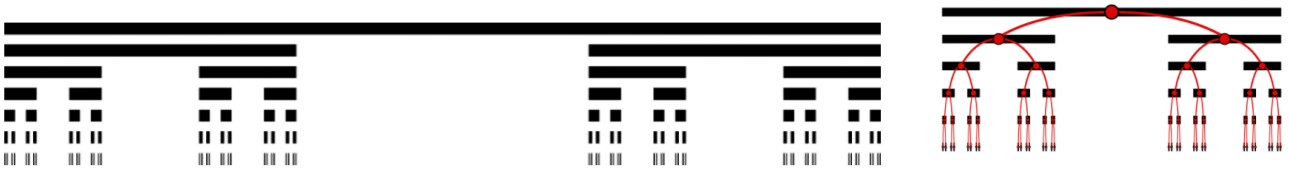
What is the set of allowable events $\mathscr{F}$ for this "uniform" probability measure? We already know it is "big" because there are an uncountable number of intervals of the form $[a,b)$ contained in $\Omega = [0, 2\pi)$. In the previous example, $\mathscr{F}$ contained all possible subsets of Ω. However, in the present case, it is impossible to assign a probability to every subset $E \subset \Omega$; said another way, there are subsets of Ω that must be disallowed events. The proper way to define the events $\mathscr{F}$ is to use the "Lebesgue measurable sets" in $[0, 2\pi)$, but to fully define Lebesgue measure takes at least a week in a course on measure theory https://en.wikipedia.org/wiki/Lebesgue_measure. Hence, in Engineering and Robotics, we mostly work with probability spaces in one of two ways

1. without carefully defining $\mathscr{F}$, the allowable events, or
2. taking $\mathscr{F}$ as the smallest sigma algebra generated by the half-open intervals $[a,b) \in \Omega$, the set we obtain by applying the operations of set complement and countable unions (called the Borel sigma algebra).

Furthermore, we place ourselves in contexts where we can integrate a density over a set to assign the probability of an event. In the above case, if $E = [0.1, 0.2) \cup \{0.5\} \cup (0.3, \pi]$, for example, we would compute the probability as

$$P(E) = \int_E \frac{1}{2\pi} dx = \int_{0.1}^{0.2} \frac{1}{2\pi} dx + \int_{0.5}^{0.5} \frac{1}{2\pi} dx + \int_{0.3}^{\pi} \frac{1}{2\pi} dx = \frac{0.1}{2\pi} + 0 + \frac{\pi - 0.3}{2\pi} = \frac{\pi - 0.2}{2\pi},$$

which seems easy enough. The trick is to only compute probabilities of sets (i.e., events) that are simple enough that a Riemann integral over the set (or at worst, a Riemann–Stieltjes integral) can be defined, thereby keeping us away from Lebesgue integration.



(a) (b)

Figure 5.1: Images of the Cantor set, thanks to Wikipedia. (a) The top line is the interval $[0, 1]$. The line below is what is left when one removes the (open) middle third, $(1/3, 2/3)$. The line below that is what is left when the next two (open) middle thirds, $(1/9, 2/9)$ and $(7/9, 8/9)$, are removed, etc. In the beginning, it's hard to believe that there is anything left over, but there is. In fact, the Cantor set can be placed in one to one correspondence with the original interval $[0, 1]$. (b) Shows various paths traced out in a ternary (base 3) expansion of numbers in the Cantor set; "each point in the Cantor set is uniquely located by a path through an infinitely deep binary tree, where the path turns left or right at each level according to which side of a deleted segment the point lies on. Representing each left turn with 0 and each right turn with 2 yields the ternary [expansion] for a point", from Wikipedia https://en.wikipedia.org/wiki/Cantor_set. Replacing the 2's with 1's yields a bijection from the Cantor set to the interval $[0, 1]$, which is cool and surprising, though otherwise irrelevant to ROB 501!

The Cantor set $C \subset [0, 1]$ is a famous set that is uncountable and has Lebesgue measure 0,

$$C = \left\{ x \in [0,1] \;\middle|\; x = \sum_{i=1}^{\infty} \frac{\epsilon_i}{3^i}, \epsilon_i \in \{0, 2\} \right\}.$$

The classical construction given in https://en.wikipedia.org/wiki/Cantor_set shows that it belongs to the Borel sigma algebra. It is impossible to define a uniform probability measure on $[0, 1]$ and compute the probability of the Cantor set by performing a Riemann–Stieltjes integral over C because the "index function"

$$I(x) = \begin{cases} 1 & x \in C \\ 0 & \text{otherwise} \end{cases}$$

has unbounded variation https://en.wikipedia.org/wiki/Bounded_variation. Said another way,

$$P(C) := \int_C dx = \int_0^1 I(x)\, dx = \text{ undefined as a Riemann or Riemann–Stieltjes integral.}$$

In any case, whether you are convinced or not, in the rest of this Chapter, we are obliged to be less careful than we have been in other parts of the book. When we discuss the probability of an event and compute it via an integral, we will simply assume that the integral exists within the usual theory of Riemann integration.

5.2 First Pass on Probability Basics

We assume that you may have skipped directly to here.

5.2.1 Densities and Random Variables

Definition 5.4 *$(\Omega, \mathscr{F}, P)$ is called a **probability space**.*

- *Ω is the sample space. Think of it as the set of all possible outcomes of an experiment.*
- *$E \subset \Omega$ is an event.*
- *$\mathscr{F}$ is the collection of allowed events[3]. It must at least contain $\emptyset$ and Ω. It is closed with respect to set complement, countable unions, and countable intersections[4]. Such sets are called sigma algebras https://en.wikipedia.org/wiki/%CE%A3-algebra.*

[3] Though it is too deep for ROB 501, there are subsets of the reals, for example, that are so complicated one cannot define a reasonable notion of probability that agrees with how we would want to define the probability of an interval, such as $[a, b]$.

[4] By De Morgan's laws, once a set is closed under set complements and countable unions, it is automatically closed under countable intersections; see https://en.wikipedia.org/wiki/De_Morgan.

- $P : \mathscr{F} \to [0,1]$ *is a probability measure. It has to satisfy a few basic operations*

  1. $P(\emptyset) = 0$ *and* $P(\Omega) = 1$.
  2. *For each* $E \in \mathscr{F}$, $0 \leq P(E) \leq 1$
  3. *If the sets* $E_1, E_2, \ldots$ *are disjoint (i.e.,* $E_i \cap E_j = \emptyset$ *for* $i \neq j$*), then*

$$P(\bigcup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} P(E_i).$$

These are typically called the ***Axioms of Probability***.

Shortly, we will define a *probability density*. On the one hand, a density can be viewed as allowing us to define a probability space on the range $\mathbb{R}$ of a random variable $X : \Omega \to \mathbb{R}$. On the other hand, it can be viewed as replacing all of the confusing probability space details with integrals over sets. It is fine to use the latter interpretation.

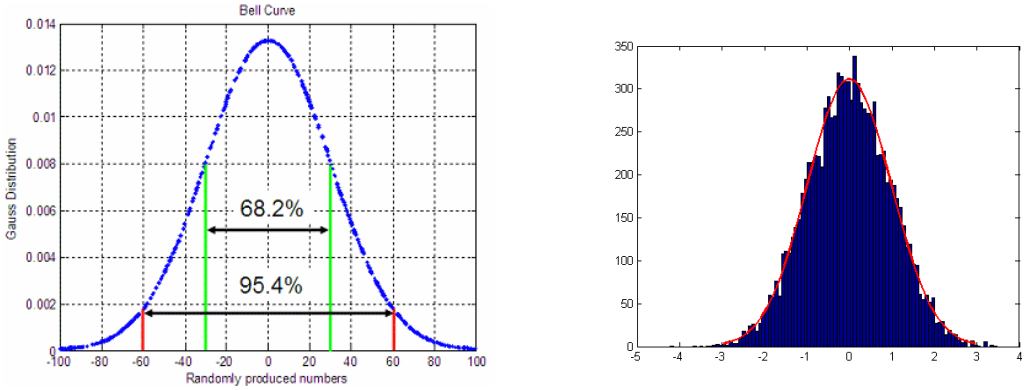


Figure 5.2: (Left) Normal distribution $N(\mu, \sigma)$ with $\mu = 0$ and $\sigma = 30$. (Right) How do you determine the density? You have to collect data! The figure shows a "fit" of a normal distribution to data.

Definition 5.5 *For* $E \subset \Omega$*, its* ***set complement*** *in* Ω *is often denoted in two ways,* $\sim E := E^c := \{x \in \Omega \mid x \notin E\}$*. Similarly, for* $A \subset \mathbb{R}$, $\sim A := A^c := \{x \in \mathbb{R} \mid x \notin A\}$.

Definition 5.6 *A function* $X : \Omega \to R$ *is a* ***random variable*** *if* $\forall\, x \in \mathbb{R}$*, the set* $\{\omega \in \Omega \,|X(\omega) \leq x\} \in \mathscr{F}$*, that is* $P(\{\omega \in \Omega \,|X(\omega) \leq x\})$ *is defined.*

With a random variable, we are typically interested in computing the probability that it takes values in a given set $A \subset \mathbb{R}$*, that is, we seek to determine* $P\{\omega \in \Omega \,|X(\omega) \in A\}$. With the "frequentist" interpretation, we are asking how "frequently" does X take values in the set A. This seems quite difficult to compute because we need to compute first, the inverse image of the set A under the function $X : \Omega \to \mathbb{R}$,

$$\{\omega \in \Omega \,|X(\omega) \in A\}, \tag{5.1}$$

and then secondly, compute the probability of this set using our probability space $(\Omega, \mathscr{F}, P)$. *The notion of a* density *allows us to circumvent the computation of* (5.1) *and work directly with the set* A *itself.*

Definition 5.7 *A (piecewise continuous*[5]*) function* $f : \mathbb{R} \to [0, \infty)$ *is a* ***probability density*** *if* $\int_{-\infty}^{\infty} f(x)dx = 1.0$

Example 5.8 *Some common densities include*

- ***Uniform density****: for* $a < b$*,*

$$f(x) = \begin{cases} \frac{1}{b-a} & x \in [a,b] \\ 0 & \textit{otherwise}. \end{cases}$$

[5]See https://en.wikipedia.org/wiki/Piecewise A square wave is a piecewise continuous function on $\mathbb{R}$ because one can decompose $\mathbb{R}$ into a countable disjoint union of half-open intervals on which the function is continuous

- ***Laplace density**: for $b > 0$, $\mu \in \mathbb{R}$, and $x \in (-\infty, \infty)$,*

$$f(x) = \frac{1}{2b} e^{-|x-\mu|/b}.$$

The parameter μ is the mean value, defined below.

- ***Gaussian or Normal density**: for $\sigma > 0$, $\mu \in \mathbb{R}$, and $x \in (-\infty, \infty)$,*

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

The parameter μ is the mean value and σ is the standard deviation; see below.

Definition 5.9 *A function $X : \Omega \to R$ is a **continuous random variable** with density $f : \mathbb{R} \to [0, \infty)$ if*

(a) it is a random variable, and

(b) $\forall\, x \in \mathbb{R}$, $P(\{\omega \in \Omega \,|X(\omega) \leq x\}) = \int_{-\infty}^{x} f(\bar{x})d\bar{x}$.

The lower bound of $-\infty$ is for convenience. It can be replaced with $\inf\{X(\Omega) \subset \mathbb{R}\}$.

Notation 5.10 *The notation $X \sim f$ is read as X is distributed with density f or that X is a random variable with density f.*

Remark 5.11 *Useful shorthand notation $\{X \leq x\} := \{\omega \in \Omega \,|X(\omega) \leq x\}$ and more generally, for $A \subset \mathbb{R}$, we define $\{X \in A\} := \{\omega \in \Omega \,|X(\omega) \in A\}$.* □

Remark 5.12 *Because $\mathscr{F}$ is closed under set complements, (countable) unions, and (countable) intersections, we can assign probabilities to*

(a) $\{X > x\} =\sim \{X \leq x\} = \{X \leq x\}^c$

(b) $\{x < X \leq y\} = \{X > x\} \cap \{X \leq y\}$

and compute them via

(a) $P(\{X > x\}) = \int_x^\infty f(x)dx$ and

(b) $P(\{x < X \leq y\}) = \int_x^y f(x)dx$, as long as $x \leq y$.

These integrals are well defined when f is piecewise continuous. □

Remark 5.13 *For how general of a set $A \subset \mathbb{R}$ can we compute $P(\{X \in A\})$? To understand this, we note that*

$$P(\{X \in A\}) := \int_A f(x)dx := \int_{-\infty}^{\infty} I_A(x)f(x)dx. \tag{5.2}$$

Because we are limiting ourselves to Riemann-Stiljes integrals, we need I_A to be sufficiently "nice" that the product $I_A(x)f(x)$ has bounded variation. A sufficient condition is that A can be expressed as a countable disjoint union of half-open intervals. This seems to be general enough for use in engineering. □

Definition 5.14 *Moments*

- *Mean: $\mu := \mathcal{E}\{X\} := \int_{-\infty}^{\infty} xf(x)dx$*
- *Variance: $\sigma^2 := \mathcal{E}\{(X-\mu)^2)\} := \int_{-\infty}^{\infty}(x-\mu)^2 f(x)dx$ (Var. for short)*
- *Standard Deviation: $\sigma := \sqrt{\sigma^2}$ (Std. Dev. for short)*

5.2.2 Random Vectors and Densities

Definition 5.15 *Let $(\Omega, \mathscr{F}, P)$ be a probability space. A function $X : \Omega \to \mathbb{R}^p$ is called a **random vector** if each component of $X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix}$ is a random variable, that is, $\forall\, 1 \leq i \leq p$, $X_i : \Omega \to \mathbb{R}$ is a random variable.*

Consequently, $\forall x \in \mathbb{R}^p$, the set $\{\omega \in \Omega \mid X(\omega) \leq x\} \in \mathscr{F}$ (i.e., it is an allowed event), where the inequality is understood **pointwise**, that is,

$$\{\omega \in \Omega \mid X(\omega) \leq x\} := \left\{ \omega \in \Omega \mid \begin{bmatrix} X_1(\omega) \\ X_2(\omega) \\ \vdots \\ X_p(\omega) \end{bmatrix} \leq \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix} \right\} := \left\{ \omega \in \Omega \;\mid\; \begin{bmatrix} X_1(\omega) \leq x_1 \\ X_2(\omega) \leq x_2 \\ \vdots \\ X_p(\omega) \leq x_p \end{bmatrix} \right\} = \bigcap_{i=1}^{p} \{\omega \in \Omega \mid X_i(\omega) \leq x_i\}.$$

Definition 5.16 *$X : \Omega \to \mathbb{R}^p$ is a **continuous random vector** if there exists a **density** $f_X : \mathbb{R}^p \to [0, \infty)$ such that,*

$$\forall\, x \in \mathbb{R}^P, \;\; P\big(\{X \leq x\}\big) = \int_{-\infty}^{x_p} ... \int_{-\infty}^{x_2} \int_{-\infty}^{x_1} f_X(\bar{x}_1, \bar{x}_2 ... \bar{x}_p) d\bar{x}_1 d\bar{x}_2 ... d\bar{x}_p.$$

More generally, for all $A \subset \mathbb{R}^p$ such that the indicator function I_A has bounded variation,

$$P\big(\{X \in A\}\big) = \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} I_A(\bar{x}_1, \bar{x}_2 ... \bar{x}_p) f_X(\bar{x}_1, \bar{x}_2 ... \bar{x}_p) d\bar{x}_1 d\bar{x}_2 ... d\bar{x}_p.$$

Notation 5.17 *The notation $X \sim f$ is read as X is distributed with density f or that X is a random vector with density f.*

Definition 5.18 *(Moments) Suppose $g : \mathbb{R}^p \to \mathbb{R}^k$*

$$\mathcal{E}\{g(X)\} := \int_{\mathbb{R}^p} g(x) f_X(x) dx := \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} g(x_1, ..., x_p) f_X(x_1, ..., x_p) dx_1 ... dx_p$$

- ***Mean or Expected Value***

$$\mu = \mathcal{E}\{X\} = \mathcal{E}\{ \begin{bmatrix} X_1 \\ \vdots \\ X_p \end{bmatrix} \} = \begin{bmatrix} \mathcal{E}\{X_1\} \\ \vdots \\ \mathcal{E}\{X_p\} \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_p \end{bmatrix}$$

- ***Covariance Matrices***

$$\Sigma := \operatorname{cov}(X) = \operatorname{cov}(X, X) = \mathcal{E}\{(X - \mu)(X - \mu)^T\}$$

  where $(X - \mu)$ is $p \times 1$, $(X - \mu)^T$ is $1 \times p$, $(X - \mu)(X - \mu)^T$ is $p \times p$

- ***Variance:*** $\operatorname{Var}(X) := \operatorname{tr} \Sigma := \sum_{i=1}^{p} \Sigma_{ii}$*, where Σ is the covariance of X.*

Exercise 5.19 *$\Sigma := \operatorname{cov}(X) = \operatorname{cov}(X, X)$ is a positive semi-definite matrix.*

Solution: For $v \in \mathbb{R}^p$, we need to show that $v^\top \Sigma v \geq 0$, where $\Sigma := \mathcal{E}\{(X - \mu) \cdot (X - \mu)^\top\}$.

$$\begin{aligned} v^\top \Sigma v :=& v^\top \mathcal{E}\{(X - \mu) \cdot (X - \mu)^\top\} v \\ =& \mathcal{E}\{v^\top (X - \mu) \cdot (X - \mu)^\top v\} \\ =& \mathcal{E}\{\big((X - \mu)^\top v\big)^\top \cdot \big((X - \mu)^\top v\big)\} \\ =& \mathcal{E}\{||(X - \mu)^\top v||^2\} \\ =& \int_{\mathbb{R}^p} ||(X - \mu)^\top v||^2 f_X(x) dx \\ \geq& 0 \end{aligned}$$

because the integral of a non-negative function over $\mathbb{R}^p$ is non-negative. ■

Definition 5.20 *If $\Sigma > 0$, then Σ^{-1} is called the **information matrix**. The interpretation is that "high variance" means "low information" and vice versa.*

5.3 Estimators

5.3.1 Best Linear Unbiased Estimator (BLUE)

Model: The measurement is $y = Cx + \varepsilon$, $y \in \mathbb{R}^m$, $x \in \mathbb{R}^n$, $\mathcal{E}\{\varepsilon\} = 0$, $\text{cov}\{\varepsilon, \varepsilon\} = \mathcal{E}\{\varepsilon\varepsilon^\top\} = Q > 0$, and the columns of C are linearly independent. *We assume no stochastic (random) model for the unknown x.* In other words, $x \in \mathbb{R}^n$ is an unknown deterministic vector. It is emphasized that ε and hence y are random vectors while x is not a random vector. Do not let the lowercase symbols being used for y, ϵ, and x distract you.

Seek: An estimate $\widehat{x}$ of x that satisfies:

1. Linear: $\widehat{x} = Ky$ for some $n \times m$ matrix K.
2. Unbiased for all $x \in \mathbb{R}^n$: $\mathcal{E}\{\widehat{x} - x\} = 0$ holds for all $x \in \mathbb{R}^n$. It needs to hold for all x because x can be an arbitrary value in $\mathbb{R}^n$.
3. Best: $\widehat{x}$ minimizes $\mathcal{E}\{(\widehat{x} - x)^\top(\widehat{x} - x)\} = \mathcal{E}\{\sum_{i=1}^{n} |\widehat{x}_i - x_i|^2\}$, the variance of $\widehat{x} - x$.

Claim 5.21 *A linear estimate $\widehat{x} = Ky$ is unbiased if, and only if $KC = I$.*

Proof:

$$
\begin{aligned}
&0 = \mathcal{E}\{\widehat{x} - x\}\ \forall x \in \mathbb{R}^n \\
&\quad \Updownarrow \\
&0 = \mathcal{E}\{Ky - x\}\ \forall x \in \mathbb{R}^n \\
&\quad \Updownarrow \\
&0 = \mathcal{E}\{K(Cx + \varepsilon) - x\}\ \forall x \in \mathbb{R}^n \\
&\quad \Updownarrow \\
&0 = \mathcal{E}\{(KC - I)x\} - \mathcal{E}\{K\varepsilon\}\ \forall x \in \mathbb{R}^n \\
&\quad \Updownarrow \\
&0 = (KC - I)x\ \forall x \in \mathbb{R}^n
\end{aligned}
$$

where we used $\mathcal{E}\{\varepsilon\} = 0$ by assumption and $\mathcal{E}\{(KC-I)x\} = (KC-I)x$ because x is deterministic. Finally, $0 = (KC-I)x\ \forall x \in \mathbb{R}^n \iff KC - I = 0_{n\times n}$ as can be seen by taking $x = e^i$, for $1 \le i \le n$. □

Aside: For $v, w \in \mathbb{R}^n$, $(v+w)^\top(v+w) = v^\top v + w^\top w + v^\top w + w^\top v = v^\top v + w^\top w + 2v^\top w$ because $v^\top w$ is a scalar.

Therefore, for an unbiased linear estimator,

$$
\begin{aligned}
\mathcal{E}\{(\widehat{x} - x)^\top(\widehat{x} - x)\} &= \mathcal{E}\{(KCx - x + K\varepsilon)^\top(KCx - x + K\varepsilon)\} \\
&= \mathcal{E}\{x^\top(KC - I)^\top(KC - I)x + 2(K\varepsilon)^\top(KC - I)x + \varepsilon^\top K^\top K\varepsilon\} \\
&= \mathcal{E}\{2(K\varepsilon)^\top(KC - I)x + \varepsilon^\top K^\top K\varepsilon\} \\
&= \mathcal{E}\{\varepsilon^\top K^\top K\varepsilon\}
\end{aligned}
$$

Moreover, by using the properties of the trace, we have

$$
\varepsilon^\top K^\top K\varepsilon = \text{tr}\left(\varepsilon^\top K^\top K\varepsilon\right) = \text{tr}\left(K\varepsilon\varepsilon^\top K^\top\right).
$$

and therefore,

$$
\begin{aligned}
\mathcal{E}\{(\widehat{x} - x)^\top(\widehat{x} - x)\} &= \text{tr}\,\mathcal{E}\{K\varepsilon\varepsilon^\top K^\top\} \\
&= \text{tr}(KQK^\top).
\end{aligned}
$$

Hence,

$$
\widehat{K} = \underset{KC=I}{\arg\min}\, \mathcal{E}\{(\widehat{x} - x)^\top(\widehat{x} - x)\} \iff \widehat{K} = \underset{KC=I}{\arg\min}\, \text{tr}(KQK^\top)
$$

A simplifying observation: If we partition K into rows via

$$K = \begin{bmatrix} k_1 \\ k_2 \\ \vdots \\ k_n \end{bmatrix}$$

then $K^\top = \begin{bmatrix} k_1^\top & \cdots & k_n^\top \end{bmatrix}$ and $\operatorname{tr}\left(\begin{bmatrix} k_1 \\ \vdots \\ k_n \end{bmatrix} Q \begin{bmatrix} k_1^\top & \cdots & k_n^\top \end{bmatrix} \right) = \sum_{i=1}^n k_i Q k_i^\top$. Moreover,

$$\begin{aligned} KC = I_{n\times n} &\iff C^\top K^\top = I_{n\times n} \\ &\iff C^\top \begin{bmatrix} k_1^\top & \cdots & k_n^\top \end{bmatrix} = \begin{bmatrix} e_1 & \cdots & e_n \end{bmatrix} \\ &\iff C^\top k_i^\top = e_i \quad 1 \le i \le n. \end{aligned}$$

Hence, we have n-separate optimization problems involving the column vectors $k_i^\top$, namely

$$\widehat{k}_i^\top = \underset{C^\top k_i^\top = e_i}{\arg\min}\ k_i Q k_i^\top,\ 1 \le i \le n.$$

From Proposition 3.95 for underdetermined equations, we have

$$\widehat{k}_i^\top = Q^{-1}C(C^\top Q^{-1} C)^{-1} e_i,$$

which yields

$$\widehat{K}^\top = \begin{bmatrix} \widehat{k}_1^\top & \cdots & \widehat{k}_n^\top \end{bmatrix} = Q^{-1}C(C^\top Q^{-1}C)^{-1}.$$

Therefore,

$$\widehat{K} = (C^\top Q^{-1} C)^{-1} C^\top Q^{-1}.$$

Theorem 5.22 ***(BLUE)*** *Let $x \in \mathbb{R}^n$, $y \in \mathbb{R}^m$, $y = Cx + \varepsilon$, $\mathcal{E}\{\varepsilon\} = 0$, $\mathcal{E}\{\varepsilon\varepsilon^\top\} =: Q > 0$, and $\operatorname{rank}(C) = n$. The Best Linear Unbiased Estimator (BLUE) is $\widehat{x} = \widehat{K}y$ where*

$$\widehat{K} = \left(C^\top Q^{-1} C\right)^{-1} C^\top Q^{-1}.$$

Moreover, the covariance of the error is

$$\mathcal{E}\{(\widehat{x} - x)\,(\widehat{x} - x)^\top\} = \left(C^\top Q^{-1} C\right)^{-1}.$$

Proof: The only thing left to show is the error covariance. From previous calculations,

$$\begin{aligned} \widehat{x} - x &= Ky - x \\ &= KCx + K\varepsilon - x \\ &= K\varepsilon \text{ (because } KC = I) \\ \therefore \mathcal{E}\{(\widehat{x} - x)(\widehat{x} - x)^\top\} &= \mathcal{E}\{(K\varepsilon)(K\varepsilon)^\top\} \\ &= \mathcal{E}\{K\varepsilon\varepsilon^\top K^\top\} \\ &= KQK^\top \end{aligned}$$

Hence,

$$\begin{aligned} \mathcal{E}\{(\widehat{x} - x)\,(\widehat{x} - x)^\top\} &= KQK^\top \\ &= \left(C^\top Q^{-1} C\right)^{-1} C^\top Q^{-1} Q Q^{-1} C \left(C^\top Q^{-1} C\right)^{-1} \\ &= \left(C^\top Q^{-1} C\right)^{-1} \left[C^\top Q^{-1} C\right] \left(C^\top Q^{-1} C\right)^{-1} \\ &= \left(C^\top Q^{-1} C\right)^{-1}. \end{aligned}$$

■

Remark 5.23 Comparing Weighted Least Squares to BLUE:

- *They are identical when the weighting matrix is taken as the inverse of the covariance matrix of the noise term: $W = Q^{-1}$. The inverse of the covariance matrix is called the information matrix. Hence, there is low information when the variance (or covariance) is large.*
- *Another way to say this, if you solve a least squares problem with weight matrix $W > 0$, you are implicitly assuming that your uncertainty in the measurements has zero mean and a covariance matrix of $Q = W^{-1}$.*
- *If you know the uncertainty has zero mean and a covariance matrix of Q, using $W = Q^{-1}$ makes a lot of sense. For simplicity, assume that Q is diagonal. A large entry in Q means high variance, which means the measurement is highly uncertain. Hence, the corresponding component of y should not be weighted very much in the optimization problem....and indeed, taking $W = Q^{-1}$ does just that because, the weight term W is small for large terms in Q.*
- *Weighted least squares was based on overdetermined systems of linear equations. To derive BLUE, we needed to understand underdetermined systems of linear equations. That's kind of cool!*

□

5.3.2 Minimum Variance Estimator (MVE)

Model: $y = Cx + \varepsilon, y \in \mathbb{R}^m, x \in \mathbb{R}^n$, and $\varepsilon \in \mathbb{R}^m$, with the following stochastic assumptions:

- Zero Means: $E\{x\} = 0, E\{\varepsilon\} = 0$.
- Covariances: $E\{\varepsilon\varepsilon^\top\} = Q, E\{xx^\top\} = P, E\{\varepsilon x^\top\} = 0$.

Remark 5.24 *$E\{\varepsilon x^\top\} = 0$ means that the state variables and noise terms are uncorrelated. Recall that uncorrelated does NOT imply independence, except for Gaussian random vectors.* □

Assumption: $Q \geq 0$, $P \geq 0$, and $CPC^\top + Q > 0$. (Will see why later.) We note that $Q > 0 \implies CPC^\top + Q > 0$ for all $m \times n$ matrices C. If $Q > 0$, we do not need to assume the columns of C are linearly independent. In fact, $C = 0_{m\times n}$ is possible.

Objective: We seek $\widehat{x}$ that minimizes the variance

$$E\{(\widehat{x} - x)^\top(\widehat{x} - x)\} = E\{\sum_{i=1}^{n}(\widehat{x}_i - x_i)^2\} = \sum_{i=1}^{n} E\{(\widehat{x}_i - x_i)^2\}.$$

Remark 5.25 *As for BLUE, we see that there are n separate optimization problems. We also see that a linear estimate $\widehat{x} = Ky$ would be automatically unbiased, because*

$$E\{\widehat{x} - x\} = E\{Ky - x\} = E\{KCx + K\varepsilon - x\} = (KC - I)E\{x\} + KE\{\varepsilon\} = 0,$$

without imposing that $KC = I$. □

Problem Formulation (the Unexpected Inner Product): We will pose this as a minimum norm problem in an inner product space of random variables. Suppose that

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \quad \text{and} \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_m \end{bmatrix}.$$

We recall that components of random vectors are random variables. Hence, x_i, $1 \leq i \leq n$ and ε_j, $1 \leq j \leq m$ are all random variables, and hence are **functions**. We define $\mathcal{F} = \mathbb{R}$, and $\mathcal{X} = span\{x_1, x_2, \ldots, x_n, \varepsilon_1, \varepsilon_2, \ldots, \varepsilon_m\}$, and for $z_1, z_2 \in \mathcal{X}$, we define their inner product by

$$< z_1, z_2 >:= E\{z_1 z_2\}.$$

We note that $\forall\ z \in \mathcal{X}$, $\mathcal{E}\{z\} = 0$ and $\text{var}(z) =< z, z >$. **Hence, this inner product space is designed to treat minimum variance optimization problems.**

Remark 5.26

$$E\{z_1 z_2\} = \begin{cases} P_{ij} & z_1 = x_i, z_2 = x_j \\ Q_{ij} & z_1 = \varepsilon_i, z_2 = \varepsilon_j \\ 0 & z_1 = x_i, z_2 = \varepsilon_j \\ 0 & z_1 = \varepsilon_i, z_2 = x_j. \end{cases}$$

□

Define:
$M = \text{span}\{y_1, y_2, \ldots, y_m\} \subset \mathcal{X}$ (M is the subspace spanned by the measurements),

$y_i = C_i x + \varepsilon_i = \sum_{j=1}^{n} C_{ij} x_j + \varepsilon_i, 1 \leq i \leq m$, ($i$-th row of y)

$\widehat{x}_i = \arg\min_{m \in M} \|x_i - m\|^2 = \arg\min_{m \in M} \langle x_i - m, x_i - m \rangle$

Fact: $\{y_1, y_2, \ldots, y_m\}$ is linearly independent if, and only if, $CPC^\top + Q$ is positive definite. This is proven below when we compute the Gram matrix. (Recall, $\{y_1, y_2, \ldots, y_m\}$ linearly independent if, and only if G is full rank, where $G_{ij} :=< y_i, y_j >$.)

Solution via the Normal Equations: By the normal equations,

$$\widehat{x}_i = \widehat{\alpha}_1 y_1 + \widehat{\alpha}_2 y_2 + \cdots + \widehat{\alpha}_m y_m$$

where $G^\top \widehat{\alpha} = \beta$ and

$$\begin{aligned} G_{ij} =< y_i, y_j >= E\{y_i y_j\} &= E\{[C_i x + \varepsilon_i][C_j x + \varepsilon_j]\} \\ &= E\{[C_i x + \varepsilon_i][C_j x + \varepsilon_j]^\top\} \\ &= E\{[C_i x + \varepsilon_i][x^\top {C_j}^\top + \varepsilon_j]\} \\ &= E\{C_i x x^\top C_j^\top\} + E\{C_i x \varepsilon_j\} + E\{\varepsilon_i x^\top C_j^\top\} + E\{\varepsilon_i \varepsilon_j\} \\ &= C_i E\{x x^\top\} C_j^\top + E\{\varepsilon_i \varepsilon_j\} \\ &= C_i P C_j^\top + Q_{ij} \\ &= [CPC^\top + Q]_{ij} \end{aligned}$$

where we have used the fact that x and ε are uncorrelated. We conclude that

$$G = CPC^\top + Q.$$

We now turn to computing β. Let's note that x_i, the i-th component of x, is equal to $x^\top e_i$, where e_i is the standard basis vector in $\mathbb{R}^n$.

$$\begin{aligned} \beta_j =< x_i, y_j > &= E\{x_i y_j\} \\ &= E\{x_i [C_j x + \varepsilon_j]\} \\ &= E\{x_i C_j x\} + E\{x_i \varepsilon_j\} \\ &= C_j E\{x x_i\} \\ &= C_j E\{x x^\top e_i\} \\ &= C_j E\{x x^\top\} e_i \\ &= C_j P e_i \\ &= C_j P_i \end{aligned}$$

where $P = \begin{bmatrix} P_1 & P_2 & \cdots & P_n \end{bmatrix}$, and hence P_i is the i-th column of P. Putting all of this together, we have

$$\begin{gathered} G^\top \widehat{\alpha} = \beta \\ \Updownarrow \\ [CPC^\top + Q]\widehat{\alpha} = CP_i \\ \Updownarrow \\ \widehat{\alpha} = [CPC^\top + Q]^{-1} CP_i \end{gathered}$$

$\widehat{x}_i = \widehat{\alpha}_1 y_1 + \widehat{\alpha}_2 y_2 + \cdots + \widehat{\alpha}_m y_m = \widehat{\alpha}^\top y =$ (row vector $\times$ column vector), where

$$\widehat{\alpha} = \begin{bmatrix} \widehat{\alpha}_1 \\ \vdots \\ \widehat{\alpha}_m \end{bmatrix}.$$

We now seek to identify a gain matrix K so that

$$\widehat{x} = Ky \iff \widehat{x}_i = K_i y, \text{where } K = \begin{bmatrix} K_1 \\ K_2 \\ \vdots \\ K_n \end{bmatrix};$$

that is, K_i is the i-th row of K.

$$\begin{aligned} K_i^\top = \widehat{\alpha}_i &= [CPC^\top + Q]^{-1} CP_i, 1 \le i \le n \\ &\Updownarrow \\ \begin{bmatrix} K_1^\top & \cdots & K_n^\top \end{bmatrix} &= [CPC^\top + Q]^{-1} CP \\ &\Updownarrow \\ K^\top &= [CPC^\top + Q]^{-1} CP \\ &\Updownarrow \\ K &= PC^\top [CPC^\top + Q]^{-1} \end{aligned}$$

$$\boxed{\widehat{x} = Ky = PC^\top [CPC^\top + Q]^{-1} y}$$

Theorem 5.27 ***(Minimum Variance Estimator)*** *Let $x \in \mathbb{R}^n$, $y \in \mathbb{R}^m$, $y = Cx + \varepsilon$, with the following stochastic assumptions:*

- *Zero Means: $E\{x\} = 0, E\{\varepsilon\} = 0$.*
- *Covariances: $E\{\varepsilon\varepsilon^\top\} = Q, E\{xx^\top\} = P, E\{\varepsilon x^\top\} = 0$.*

The Minimum Variance Estimator (MVE) is $\widehat{x} = \widehat{K} y$ where

$$\widehat{K} = PC^\top [CPC^\top + Q]^{-1}.$$

Moreover, the covariance of the error is

$$\mathcal{E}\{(\widehat{x} - x)(\widehat{x} - x)^\top\} = P - PC^\top [CPC^\top + Q]^{-1} CP.$$

Proof: The only thing left to show is the error covariance. From previous calculations, let's note that

$$\widehat{x} - x = Ky - x = KCx + K\varepsilon - x = (KC - I)x + K\varepsilon$$

and thus

$$(\widehat{x} - x)(\widehat{x} - x)^\top = (KC - I)xx^\top (KC - I)^\top + K\varepsilon\varepsilon^\top K^\top - (KC - I)x\varepsilon^\top K^\top - K\varepsilon x^\top (KC - I)^\top.$$

Taking expectations, and recalling that x and ε are uncorrelated, we have

$$\begin{aligned} E\{(\widehat{x} - x)(\widehat{x} - x)^\top\} &= (KC - I)P(KC - I)^\top + KQK^\top \\ &= KCPC^\top K^\top + P - 2PC^\top K^\top + KQK^\top \\ &= P + K[CPC^\top + Q]K^\top - 2PC^\top K^\top. \end{aligned}$$

Substituting with $K = PC^\top [CPC^\top + Q]^{-1}$ and simplifying yields the result. ■

Remark 5.28

1. $\text{cov}(\begin{bmatrix} x \\ y \end{bmatrix}) = \mathcal{E}\{ \begin{bmatrix} x \\ Cx+\varepsilon \end{bmatrix} \begin{bmatrix} x^\top & x^\top C^\top + \varepsilon^\top \end{bmatrix}\} = \begin{bmatrix} P & PC^\top \\ CP & CPC^\top + Q \end{bmatrix}$.
2. *Hence, we see that the covariance of the error $\widehat{x} - x$ is the Schur Complement of* $\text{cov}(x)$ *in* $\text{cov}(\begin{bmatrix} x \\ y \end{bmatrix})$.
3. *The term $PC^\top[CPC^\top + Q]^{-1}CP$ represents the "value" of the measurements. It is the reduction in the covariance of x given the measurement y.*
4. *If $Q > 0$ and $P > 0$, then from the Matrix Inversion Lemma*

$$\boxed{\widehat{x} = Ky = [C^\top Q^{-1}C + P^{-1}]^{-1}C^\top Q^{-1}y.}$$

   This form of the equation is useful for comparing BLUE vs MVE

5. ***BLUE vs MVE:***

   - ***BLUE:*** $\widehat{x} = [C^\top Q^{-1}C]^{-1}C^\top Q^{-1}y$
   - ***MVE:*** $\widehat{x} = [C^\top Q^{-1}C + P^{-1}]^{-1}C^\top Q^{-1}y$
   - *Hence, BLUE = MVE when* $P^{-1} = 0$.
   - $P^{-1} = 0$ *roughly means* $P = \infty I$, *that is infinite covariance in* x, *which in turn means we have no idea about how* x *is distributed.*
   - *For BLUE to exist, we need* $\dim(y) \geq \dim(x)$
   - *For MVE to exist, we can have* $\dim(y) < \dim(x)$ *as long as* $(CPC^\top + Q) > 0$.

□

Remark 5.29 *(Solution to MIL) We will show that if $Q > 0$ and $P > 0$, then*

$$PC^\top[CPC^\top + Q]^{-1} = [C^\top Q^{-1}C + P^{-1}]^{-1}C^\top Q^{-1}.$$

Recall the MIL: Suppose that A, B, C and D are compatible[6] *matrices. If A, C, and $(C^{-1} + DA^{-1}B)$ are each square and invertible, then $A + BCD$ is invertible and*

$$(A + BCD)^{-1} = A^{-1} - A^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1}.$$

We apply the MIL to $[C^\top Q^{-1}C + P^{-1}]^{-1}$, where we identify $A = P^{-1}, B = C^\top, C = Q^{-1}, D = C$. This yields

$$[C^\top Q^{-1}C + P^{-1}]^{-1} = P - PC^\top[Q + CPC^\top]^{-1}CP.$$

Hence,

$$\begin{aligned}
[C^\top Q^{-1}C + P^{-1}]^{-1}C^\top Q^{-1} &= PC^\top Q^{-1} - PC^\top[Q + CPC^\top]^{-1}CPC^\top Q^{-1} \\
&= PC^\top \left[I - [Q + CPC^\top]^{-1}CPC^\top\right] Q^{-1} \\
&= {\scriptstyle PC^\top\left[[Q+CPC^\top]^{-1}[Q+CPC^\top] - [Q+CPC^\top]^{-1}CPC^\top\right]Q^{-1}} \\
&= PC^\top[Q + CPC^\top]^{-1}\left[[Q + CPC^\top] - CPC^\top\right] Q^{-1} \\
&= PC^\top[Q + CPC^\top]^{-1}\left[Q + CPC^\top - CPC^\top\right] Q^{-1} \\
&= PC^\top[Q + CPC^\top]^{-1}\left[Q\right] Q^{-1} \\
&= PC^\top[Q + CPC^\top]^{-1}
\end{aligned}$$

□

[6] The sizes are such the matrix products and sum in $A + BCD$ make sense.

5.4 Second Pass on Probability Basics

5.4.1 Marginal Densities, Independence, and Correlation

Suppose the random vector $X : \Omega \to \mathbb{R}^p$ is partitioned into two components $X_1 : \Omega \to R^n$ and $X_2 : \Omega \to R^m$, with $p = n + m$, so that,

$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$$

Notation 5.30 *We denote the density of X by*

$$f_X(x) = f_{\begin{bmatrix} X_1 \\ X_2 \end{bmatrix}}(x_1, x_2) = f_{X_1X_2}(x_1, x_2)$$

*and it is called the **joint density** of X_1 and X_2. As before, we can define the mean and covariance.*

- *Mean is* $\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} = \mathcal{E}\{X\} = \mathcal{E}\{\begin{bmatrix} X_1 \\ X_2 \end{bmatrix}\} = \begin{bmatrix} \mathcal{E}\{X_1\} \\ \mathcal{E}\{X_2\} \end{bmatrix}$
- *Covariance is*

$$\begin{aligned} \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} &= \mathcal{E}\{\begin{bmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \end{bmatrix} \begin{bmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \end{bmatrix}^\top\} \\ &= \mathcal{E}\{\begin{bmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \end{bmatrix} \begin{bmatrix} (X_1 - \mu_1)^\top & (X_2 - \mu_2)^\top \end{bmatrix}\} \\ &= \mathcal{E}\{\begin{bmatrix} (X_1 - \mu_1)(X_1 - \mu_1)^\top & (X_1 - \mu_1)(X_2 - \mu_2)^\top \\ (X_1 - \mu_1)(X_2 - \mu_2)^\top & (X_2 - \mu_1)(X_2 - \mu_2)^\top \end{bmatrix} \end{aligned}$$

*where $\Sigma_{12} = \Sigma_{21}^\top = cov(X_1, X_2) = \mathcal{E}\{(X_1 - \mu_1)(X_2 - \mu_2)^\top\}$ is also called the **correlation** of X_1 and X_2.*

If $X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} : \Omega \to \mathbb{R}^{n+m}$ is a continuous random vector, then its components

$$X_1 : \Omega \to \mathbb{R}^n \text{ and } X_2 : \Omega \to \mathbb{R}^m$$

are also continuous random vectors and have densities, $f_{X_1}(x_1)$ and $f_{X_2}(x_2)$. These densities are given a special name.

Definition 5.31 *$f_{X_1}(x_1)$ and $f_{X_2}(x_2)$ are called the **marginal densities** of X_1 and X_2.*

Fact 5.32 *In general the marginal densities are a nightmare to compute.*

$$\begin{aligned} f_{X_1}(x_1) &:= \int_{-\infty}^{\infty} f_{X_1X_2}(x_1, x_2)dx_2 \\ &:= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{X_1X_2}(\underbrace{\bar{x}_1, \ldots, \bar{x}_n}_{x_1}, \underbrace{\bar{x}_{n+1}, \cdots, \bar{x}_{n+m}}_{x_2}) \underbrace{d\bar{x}_{n+1} \cdots d\bar{x}_{n+m}}_{dx_2} \end{aligned}$$

$$\begin{aligned} f_{X_2}(x_2) &:= \int_{-\infty}^{\infty} f_{X_1X_2}(x_1, x_2)dx_1 \\ &:= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{X_1X_2}(\underbrace{\bar{x}_1, \ldots, \bar{x}_n}_{x_1}, \underbrace{\bar{x}_{n+1}, \cdots, \bar{x}_{n+m}}_{x_2}) \underbrace{d\bar{x}_1 \cdots d\bar{x}_n}_{dx_1} \end{aligned}$$

For Normal Random Vectors, however, we can read the marginal densities directly from the joint density! We will not be doing any iterated integrals in ROB 501.

□

Definition 5.33 *Random vectors X_1 and X_2 are **independent** if their joint density factors*

$$f_X(x) = f_{X_1X_2}(x_1, x_2) = f_{X_1}(x_1)f_{X_2}(x_2).$$

*X_1 and X_2 are **uncorrelated** if their "cross covariance" or "correlation " is zero, that is,*

$$cov(X_1, X_2) := \mathcal{E}\{(X_1 - \mu_1)(X_2 - \mu_2)^\top\} = 0_{n\times m}.$$

Fact 5.34 *If random vectors X_1 and X_2 are independent, then they are also uncorrelated.* **The converse is in general false.** □

5.4.2 Conditional Probabilities

Definition 5.35 *Consider two events $A, B \in \mathscr{F}$, with $P(B) > 0$. The **conditional probability of** A **given** B is*

$$P(A \mid B) := \frac{P(A \bigcap B)}{P(B)}.$$

Remark 5.36 *Suppose A is the event that our robot is near a certain location and B is the event that our robot has been cited by a particular camera. These two events have individual probabilities $P(A)$ and $P(B)$. The **conditional probability of event** A **given that event** B **occurred is how we "fuse" the two pieces of information**,*

$$P(A \mid B) := \frac{P(A \bigcap B)}{P(B)}.$$

As an example, suppose we define a uniform probability across all floors of FRB, which has a total surface area of roughly 12,000 m^2 (roughly, 120,000 square feet) and let A be the event our robot is choosing a snack in the self-service section of the Eigen Cafe. Well, the area of the self-service section of the Eigen Cafe is roughly 8 m^2, and thus $P(A) \approx 6.66\ 10^{-4}$. We'll now let B be the event that our robot has been seen (i.e, measured to be) at the Eigen Cafe. The total area of the Eigen Cafe is approximately 30 m^2. Hence, $P(B) \approx 2.5\ 10^{-3}$. Because $A \subset B$, it follows that $A \cap B = A$ and hence

$$P(A \mid B) = \frac{P(A \bigcap B)}{P(B)} = \frac{P(A)}{P(B)} = \frac{6.66\ 10^{-4}}{2.5\ 10^{-3}} = 0.266.$$

We note that $P(A \mid B) = 0.266$ is a lot more certain than $P(A) \approx 6.66\ 10^{-4}$. Hence, as you can now imagine, conditional probabilities are very important in engineering. □

Remark 5.37 *Special cases:*

- $B \subset A$, $P(A \mid B) = \frac{P(A \bigcap B)}{P(B)} = \frac{P(B)}{P(B)} = 1$
- $A \subset B$, $P(A \mid B) = \frac{P(A \bigcap B)}{P(B)} = \frac{P(A)}{P(B)} \geq P(A)$

□

Definition 5.38 *Consider again our partitioned random vector $X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$. The **conditional density of** X_1 **given** $X_2 = x_2$ is*

$$f_{X_1|X_2}(x_1 \mid x_2) := \frac{f_{X_1X_2}(x_1, x_2)}{f_{X_2}(x_2)}.$$

Sometimes, for brevity, we simply write $f(x_1 \mid x_2)$.

Remark 5.39 ***Remarks on Conditional Random Vectors:***

- ***Very important:*** *X_1 given $X_2 = x_2$ is (still) a random vector. It's density is $f_{X_1|X_2}(x_1 \mid x_2)$*
- ***Conditional Mean:***

$$\begin{aligned}\mu_{X_1|X_2=x_2} &:= \mathcal{E}\{X_1 \mid X_2 = x_2\} \\ &:= \int_{-\infty}^{\infty} x_1 f_{X_1|X_2}(x_1 \mid x_2) dx_1\end{aligned}$$

$\mu_{X_1|X_2=x_2}$ is a function of x_2. Think of it as a function of the value read by your sensor.

- ***Conditional Covariance:***

$$\Sigma_{X_1|X_2=x_2} := \mathcal{E}\{(X_1 - \mu_{X_1|X_2=x_2})(X_1 - \mu_{X_1|X_2=x_2})^\top \mid X_2 = x_2\}$$

$$:= \int_{-\infty}^{\infty} (X_1 - \mu_{X_1|X_2=x_2})(X_1 - \mu_{X_1|X_2=x_2})^\top f_{X_1|X_2}(x_1 \mid x_2) dx_1$$

$\Sigma_{X_1|X_2=x_2}$ is a function of x_2. Think of it as a function of the value read by your sensor.

□

5.4.3 (Optional Read) Derivation of the Conditional Density Formula from the Conditional Distribution:

Definition 5.40 *Let $X : \Omega \to \mathbb{R}^p$ be a random vector. Then $F : \mathbb{R}^p \to [0,1]$ is a **cumulative probability distribution function** if $\forall\, x \in \mathbb{R}^p$, $F(x) = P(\{X \leq x\})$.*

Remark 5.41 *If $X \sim f$, then the cumulative distribution function and the density are related by $F(x) = \int_{-\infty}^{x} f(x)dx$ and $f(x) = \frac{\partial}{\partial x} F(x)$. By the definition of a density, the integral is well defined.* □

We define $A := \{X_1 \leq x_1\}$ and for $\epsilon > 0$, define $B_\epsilon := \{x_2 - \epsilon \leq X_2 \leq x_2 + \epsilon\}$ where $x_2 \pm \epsilon$ means adding or subtracting ϵ to each component of x_2. Then,

$$P(A \bigcap B_\epsilon) = \int_{-\infty}^{x_1} \int_{x_2-\epsilon}^{x_2+\epsilon} f_{X_1X_2}(\bar{x}_1, \bar{x}_2) d\bar{x}_2 d\bar{x}_1$$
$$P(B_\epsilon) = \int_{x_2-\epsilon}^{x_2+\epsilon} f_{X_2}(\bar{x}_2) d\bar{x}_2$$

Moreover, applying l'Hôpital's Rule,

$$\begin{aligned} F_{X_1|X_2}(x_1 \mid x_2) &= \lim_{\epsilon \to 0} \frac{P(A \bigcap B_\epsilon)}{P(B_\epsilon)} \\ &= \lim_{\epsilon \to 0} \frac{\int_{-\infty}^{x_1} \int_{x_2-\epsilon}^{x_2+\epsilon} f_{X_1X_2}(\bar{x}_1, \bar{x}_2) d\bar{x}_2 d\bar{x}_1}{\int_{x_2-\epsilon}^{x_2+\epsilon} f_{X_2}(\bar{x}_2) d\bar{x}_2} \\ &= \int_{-\infty}^{x_1} \lim_{\epsilon \to 0} \frac{\int_{x_2-\epsilon}^{x_2+\epsilon} f_{X_1X_2}(\bar{x}_1, \bar{x}_2) d\bar{x}_2}{\int_{x_2-\epsilon}^{x_2+\epsilon} f_{X_2}(\bar{x}_2) d\bar{x}_2} d\bar{x}_1 \\ &= \int_{-\infty}^{x_1} \frac{f_{X_1X_2}(\bar{x}_1, x_2)}{f_{X_2}(x_2)} d\bar{x}_1. \end{aligned}$$

Differentiating the distribution function with respect to x_1 gives the density and hence

$$(X_1 \mid X_2 = x_2) \sim f_{X_1|X_2}(x_1 \mid x_2) = \frac{f_{X_1X_2}(x_1, x_2)}{f_{X_2}(x_2)}.$$

Alternatively, one can differentiate with respect to x_1 first, and then take the limit as $\epsilon \to 0$,

$$f_{X_1|X_2}(x_1 \mid x_2) = \lim_{\epsilon \to 0} \frac{\int_{x_2-\epsilon}^{x_2+\epsilon} f_{X_1X_2}(x_1, \bar{x}_2) d\bar{x}_2}{\int_{x_2-\epsilon}^{x_2+\epsilon} f_{X_2}(\bar{x}_2) d\bar{x}_2} = \lim_{\epsilon \to 0} \frac{f_{X_1X_2}(x_1, x_2) \cdot 2\epsilon}{f_{X_2}(x_2) \cdot 2\epsilon} = \frac{f_{X_1X_2}(x_1, x_2)}{f_{X_2}(x_2)}.$$

5.5 Important Facts about Gaussian Random Vectors

Definition 5.42 *A **random variable** X is **normally distributed** with mean μ and variance $\sigma^2 > 0$ if it has density*

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

The standard deviation is $\sigma > 0$. The mean and variance satisfy

$$\mu := \mathcal{E}\{X\} := \int_{\mathbb{R}} x f_X(x) dx := \int_{-\infty}^{\infty} x f_X(x) dx$$

$$\sigma^2 := \mathcal{E}\{(X-\mu)^2\} := \int_{\mathbb{R}} (x-\mu)^2 f_X(x) dx := \int_{-\infty}^{\infty} (x-\mu)^2 f_X(x) dx.$$

You should be quite familiar with the "bell curve". X is also called a Gaussian random variable. We also say X has a *univariate normal distribution* or a *univariate Gaussian distribution* to emphasize that we are talking about a single random variable.

For the most part, we do not care too much about individual random variables. We are interested in collections of random variables and random vectors, and hence we are primarily concerned about *jointly distributed random variables*. If you take EECS 501, you can learn a tremendous amount of material about this subject. In the following, you will see a bare bones accounting of *multivariate normal random variables*.

Definition 5.43 *A finite collection of random variables $X_1, X_2, \cdots, X_p$, or equivalently, the random vector*

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix}$$

*has a (non-degenerate) **multivariate normal distribution** with mean μ and covariance $\Sigma > 0$ if the joint density is given by*

$$f_X(x) = \frac{1}{\sqrt{(2\pi)^p |\Sigma|}} e^{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)}.$$

In the above, $|\Sigma| = \det(\Sigma)$, which must be non-zero for the denominator to be well defined. This condition is what is meant by "non-degenerate". When $|\Sigma| = 0$, one can still define a multivariate normal distribution, but the "moment generating function" must be used. This is a technicality that we will skip. We note that

$$\begin{aligned}
\mu &:= \mathcal{E}\{X\} \in \mathbb{R}^p \\
\mu_i &:= \int_{\mathbb{R}^p} x_i \, f_X(x) dx := \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} x_i \, f_X(x_1, \cdots, x_p) dx_1 \cdots dx_p \\
\Sigma &:= \operatorname{cov}(X, X) := \mathcal{E}\{(X-\mu)(X-\mu)^\top\} \in \mathbb{R}^{p \times p} \\
\Sigma_{ij} &:= \mathcal{E}\{(X_i - \mu_i)(X_j - \mu_j)\} := \int_{\mathbb{R}^p} (x_i - \mu_i)(x_j - \mu_j) \, f_X(x) dx \\
x &= (x_1, x_2, \cdots, x_p) \text{ or } x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix} \text{ (depending on context).}
\end{aligned}$$

Fact 5.44 ***(Marginal Densities/Distributions)*** *Each random variable X_i has a **univariate normal distribution** with mean μ_i and variance Σ_{ii},*

$$f_{X_i}(x_i) = \frac{1}{\sqrt{2\pi\Sigma_{ii}}} e^{-\frac{(x_i-\mu_i)^2}{2\Sigma_{ii}}}.$$

Hence, no iterated integrals are required to compute the marginal densities. Why is this true? Because the univariate density depends on two terms, namely, $\mu_i := \mathcal{E}\{X_i\}$ and $\sigma_i^2 := \mathcal{E}\{(X_i - \mu_i)^2\} = \Sigma_{ii}$. □

We note the unfortunate lack of coordination in the notation: the standard deviation of X_i, which we typically denote by σ_i, is given by

$$\sigma_i = \sqrt{\Sigma_{ii}}.$$

I guess we will not be denoting the entries of Σ with lower case σ.

Fact 5.45 ***(Independence)*** *Gaussian random variables are very special in that they are independent if, and only if, they are uncorrelated. Hence, X_i and X_j are independent if, and only if, $\Sigma_{ij} = \Sigma_{ji} = 0$.*

□

Fact 5.46 ***(Linear Combinations)*** *Define a new random vector by $Y = AX + b$, with the rows of A linearly independent. Then Y is a Gaussian (normal) random vector with*

$$\begin{aligned}\mathcal{E}\{Y\} &= A\mu + b =: \mu_Y \\ \operatorname{cov}(Y,Y) &= \mathcal{E}\{(Y-\mu_Y)(Y-\mu_Y)^\top\} = A\Sigma A^\top =: \Sigma_{YY}.\end{aligned}$$

Indeed, $Y - \mu_Y = A(X - \mu)$. Hence,

$$\operatorname{cov}(Y,Y) = \mathcal{E}\{[A(X-\mu)][A(X-\mu)]^\top\} = A\mathcal{E}\{(X-\mu)(X-\mu)^\top\}A^\top = A\Sigma A^\top,$$

and $A\Sigma A^\top > 0$ when A has full row rank and $\Sigma > 0$. □

Remark 5.47 *Taking $b = 0$ and A to be a row vector with all zeros except a one in the i-th spot, that is $A = [0, \cdots, 1, \cdots, 0]$, recovers the **marginal** densities discussed above.* □

5.6 Conditioning with Gaussian Random Vectors:

Remark 5.48 *In various places, $X_{|Y}$ and $X|Y$ are each used to represent X conditioned on $Y = y$.*

In addition to looking at individual random variables making up a random vector, we can group the components to form two or more blocks of vectors as long as their sizes add up to p, the number of components in X. We abuse notation and write

$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \begin{matrix} \in \mathbb{R}^n \\ \in \mathbb{R}^m \end{matrix}$$

In books, you'll often see the blocks expressed in bold font, such as $\mathbf{X_1}$ and $\mathbf{X_2}$. We will NOT do this. Conformally with the partition of X into two blocks, we partition the mean and covariance as follows

$$\begin{aligned}\mu &=: \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \\ \Sigma &=: \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}.\end{aligned}$$

From our results on the Schur complement, we know that $\Sigma > 0$ if, and only if, $\Sigma_{22} > 0$ and $\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} > 0$.

Remark 5.49 *To be extra clear on the dimensions, we suppose $n + m = p$ and note that*

$$\begin{aligned}\mu_1 &= \mathcal{E}\{X_1\} \in \mathbb{R}^n \\ \mu_2 &= \mathcal{E}\{X_2\} \in \mathbb{R}^m \\ \Sigma_{11} &= \operatorname{cov}(X_1, X_1) \in \mathbb{R}^{n\times n} \\ \Sigma_{22} &= \operatorname{cov}(X_2, X_2) \in \mathbb{R}^{m\times m} \\ \Sigma_{12} &= \operatorname{cov}(X_1, X_2) \in \mathbb{R}^{n\times m} \\ \Sigma_{21} &= \operatorname{cov}(X_2, X_1) \in \mathbb{R}^{m\times n}.\end{aligned}$$

Furthermore, because $\Sigma = \Sigma^\top$, we have that

$$\Sigma_{11}^\top = \Sigma_{11}, \;\; \Sigma_{22}^\top = \Sigma_{22}, \text{ and } \; \Sigma_{12}^\top = \Sigma_{21}.$$

□

Fact 5.50 *Each vector X_i has a multivariate normal distribution with mean μ_i and covariance Σ_{ii}. This is also called the **marginal distribution** of X_i. If we know the mean and covariance for the composite vector X, it is very easy to read off the marginal distributions of its vector components. This can be established by Fact 5.46 after writing $X_1 = A_1X + b_1$ with $b_1 = 0_{n\times 1}$ and $A_1 = [I_{n\times n}, 0_{n\times m}]$ and $X_2 = A_2X + b_2$ with $b_2 = 0_{m\times 1}$ and $A_2 = [0_{m\times n}, I_{m\times m}]$.* □

Key Fact 1 *(Conditional Distributions of Gaussian Random Vectors:) Let X_1 and X_2 be as above, namely they are components of a larger vector X that has a multivariate normal distribution. Then the conditional distribution of X_1 given $X_2 = x_2$ has a multivariate normal distribution with*

$$\textit{Mean}: \quad \mu_{1|2} := \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2)$$

and

$$\textit{Covariance:} \ \Sigma_{1|2} := \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

In passing, we note that the mean depends on the value of x_2 while the covariance does not. □

To be extra clear on the meanings here, we note that

- $\mu_{1|2} = \mathcal{E}\{X_1 \mid X_2 = x_2\}$
- $\Sigma_{1|2} = \mathcal{E}\{(X_1 - \mu_{1|2})(X_1 - \mu_{1|2})^\top \mid X_2 = x_2\}$
- X_1 given $X_2 = x_2$ is a random vector. It has a multivariate normal distribution with the above mean vector and covariance matrix. Specifically, its density is

$$f_{X_1|X_2=x_2}(x_1) = \frac{f_{X_1,X_2}(x_1,x_2)}{f_{X_2}(x_2)} = \frac{1}{\sqrt{(2\pi)^n|\Sigma_{1|2}|}} e^{-\frac{1}{2}(x_1-\mu_{1|2})^\top \Sigma_{1|2}^{-1}(x_1-\mu_{1|2})},$$

  where it is emphasized that $\mu_{1|2}$ depends explicitly on x_2.

A proof of this can be found at the link below. The algebra is rather painful. If you are very ambitious, you can work out the special case where X_1 and X_2 are scalars. This will not be on any exam in ROB 501; see http://fourier.eng.hmc.edu/e161/lectures/gaussianprocess/node7.html; see also http://www.stats.ox.ac.uk/~steffen/teaching/bs2HT9/gauss.pdf.

Remark 5.51 *If X_1 and X_2 are uncorrelated, then $\mu_{1|2} = \mu_1$ and $\Sigma_{1|2} = \Sigma_{11}$. Similarly, let's suppose that Σ_{22} is large, specifically, $\Sigma_{22} = \rho I_{m\times m}$. Then, $\lim_{\rho\to\infty} \mu_{1|2} = \mu_1$ and $\lim_{\rho\to\infty} \Sigma_{1|2} = \Sigma_{11}$. The term $\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ measures the value of the "information gained by conditioning on X_2". As $\Sigma_{22}^{-1} \to 0$, the value of the added information tends to zero.*

□

Key Fact 2 *(Conditional Independence) Suppose we have 3 vectors X_1, X_2 and X_3 that are jointly normally distributed:*

$$X = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix}$$

and that X_1 and X_3 are each independent of X_2. We then have no special structure on the means,

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix}$$

but the covariance matrix has the form

$$\Sigma = \begin{bmatrix} \Sigma_{11} & 0 & \Sigma_{13} \\ 0 & \Sigma_{22} & 0 \\ \Sigma_{13}^\top & 0 & \Sigma_{33} \end{bmatrix}$$

*where $\Sigma_{12} = \Sigma_{21}^\top = \mathrm{cov}(X_1, X_2) = 0$ due to the independence of X_1 and X_2. Similarly for $\Sigma_{23} = \Sigma_{32}^\top = 0$. Because Σ is symmetric, $\Sigma_{31} = \Sigma_{13}^\top$. **Then X_1 and X_2 are conditionally independent given X_3.** Written a different way, the two normal random variables, $X_{1|X_3}$ (X_1 conditioned on knowing X_3) and $X_{2|X_3}$ (X_2 conditioned on knowing X_3) are independent.* □

To see why this is true, we partition Σ as

$$\Sigma = \left[\begin{array}{cc|c} \Sigma_{11} & 0 & \Sigma_{13} \\ 0 & \Sigma_{22} & 0 \\ \hline \Sigma_{13}^\top & 0 & \Sigma_{33} \end{array}\right].$$

We compute the covariance of X_1 and X_2 conditioned on X_3, that is

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \Big| X_3,$$

using the Schur complement from **Key Fact 1**

$$\begin{aligned} \operatorname{cov}(\begin{bmatrix} X_{1|X_3} \\ X_{2|X_3} \end{bmatrix}, \begin{bmatrix} X_{1|X_3} \\ X_{2|X_3} \end{bmatrix}) &= \begin{bmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{bmatrix} - \begin{bmatrix} \Sigma_{13} \\ 0 \end{bmatrix} \Sigma_{33}^{-1} \begin{bmatrix} \Sigma_{13}^\top & 0 \end{bmatrix} \\ &= \begin{bmatrix} \Sigma_{11} - \Sigma_{13}\Sigma_{33}^{-1}\Sigma_{13}^\top & 0 \\ 0 & \Sigma_{22} \end{bmatrix} \end{aligned}$$

Because the off-diagonal blocks are zero, the two random variables $X_{1|X_3}$ and $X_{2|X_3}$ are uncorrelated, and because they are normal, we conclude they are independent.

Once again, what we have seen is that if X_1 and X_2 are independent, and we also have X_2 is independent of X_3, then X_1 and X_2 remain independent when we condition them on X_3.

Key Fact 3 *(Covariance of a Sum of Independent Normal Random Variables) Let X_1 and X_2 be independent normal random vectors, with means μ_1 and μ_2, and covariances, Σ_{11} and Σ_{22}. Define Y as a "linear combination" of X_1 and X_2 via*

$$Y = AX_1 + BX_2$$

for appropriately sized matrices A and B. Then

$$\mu_Y = A\mu_1 + B\mu_2$$

and

$$\operatorname{cov}(Y, Y) = A\Sigma_{11}A^\top + B\Sigma_{22}B^\top.$$

□

To see why this is true, we first note that

$$\begin{aligned} (Y - \mu_Y)(Y - \mu_Y)^\top = &A(X_1 - \mu_1)(X_1 - \mu_1)^\top A^\top + B(X_2 - \mu_2)(X_2 - \mu_2)^\top B^\top \\ &+ A(X_1 - \mu_1)(X_2 - \mu_2)^\top B^\top + B(X_2 - \mu_2)(X_1 - \mu_1)^\top A^\top, \end{aligned}$$

and then note that when expectations are taken on each side, the independence of X_1 and X_2 gives

$$\mathcal{E}\{(X_1 - \mu_1)(X_2 - \mu_2)^\top\} = 0 \text{ and } \mathcal{E}\{(X_2 - \mu_2)(X_2 - \mu_2)^\top\} = 0.$$

Therefore,

$$\begin{aligned} \operatorname{cov}(Y, Y) &= \mathcal{E}\{(Y - \mu_Y)(Y - \mu_Y)^\top\} \\ &= A\mathcal{E}\{(X_1 - \mu_1)(X_1 - \mu_1)^\top\}A^\top + B\mathcal{E}\{(X_2 - \mu_2)(X_2 - \mu_2)^\top\}B^\top \\ &= A\Sigma_{11}A^\top + B\Sigma_{22}B^\top. \end{aligned}$$

Key Fact 4 *Suppose that X, Y and Z are jointly distributed random vectors with density f_{XYZ}. Then the conditional random vectors $(X|Z)\big|(Y|Z)$ and $X\big|\begin{bmatrix} Y \\ Z \end{bmatrix}$ have the same conditional densities, that is,*

$$(X|Z)\big|(Y|Z) \sim \frac{f_{(X|Z)(Y|Z)}}{f_{(Y|Z)}} = \frac{f_{XYZ}}{f_{YZ}} \sim X\Big|\begin{bmatrix} Y \\ Z \end{bmatrix}.$$

□

The above fact does not require the random vectors to be jointly normal. The proof goes like this,

$$(X|Z)\big|(Y|Z) \sim \frac{f_{(X|Z)(Y|Z)}}{f_{(Y|Z)}} = \frac{f_{\begin{bmatrix} X \\ Y \end{bmatrix}|Z}}{f_{Y|Z}} = \frac{\frac{f_{XYZ}}{f_Z}}{\frac{f_{YZ}}{f_Z}} = \frac{f_{XYZ}}{f_{YZ}} \sim X\Big|\begin{bmatrix} Y \\ Z \end{bmatrix}.$$

In the Kalman Filter, the conditional density on the left will give us a recursive implementation of the density on the right, in place of a batch update. You have to see it in action to believe it.

5.7 Discrete-time Kalman Filter

5.7.1 Model and Assumptions

Model: Linear time-varying discrete-time system with "white[7]" Gaussian noise

$$\begin{aligned} x_{k+1} &= A_k x_k + G_k w_k, \;\; x_0 \text{ initial condition} \\ y_k &= C_k x_k + v_k \end{aligned}$$

$x \in \mathbb{R}^n$, $w \in \mathbb{R}^p$, $y \in \mathbb{R}^m$, $v \in \mathbb{R}^m$. Moreover, the random vectors x_0, and, for $k \geq 0$, w_k, v_k are all independent[8] Gaussian (normal) random vectors.

Precise assumptions on the random vectors We'll denote $\delta_{kl} = 1 \iff k = l$ and $\delta_{kl} = 0, \;\; k \neq l$.

- For all $k \geq 0, \; l \geq 0$, x_0, w_k, v_l are jointly Gaussian.
- w_k is a 0-mean white noise process: $\mathcal{E}\{w_k\} = 0$, and $\operatorname{cov}(w_k, w_l) = R_k \delta_{kl}$
- v_k is a 0-mean white noise process: $\mathcal{E}\{v_k\} = 0$, and $\operatorname{cov}(v_k, v_l) = Q_k \delta_{kl}$
- Uncorrelated noise processes: $\operatorname{cov}(w_k, v_l) = 0$
- The initial condition x_0 is uncorrelated with all other noise sequences.
- We denote the mean and covariance of x_0 by

$$\bar{x}_0 = \mathcal{E}\{x_0\} \;\text{ and }\; P_0 = \operatorname{cov}(x_0) = \operatorname{cov}(x_0, x_0) = \mathcal{E}\{(x_0 - \bar{x}_0)(x_0 - \bar{x}_0)^\top\}$$

Short-hand notation for the noise modeling assumptions:

$$\operatorname{cov}\left(\begin{bmatrix} w_k \\ v_k \\ x_0 \end{bmatrix}, \begin{bmatrix} w_l \\ v_l \\ x_0 \end{bmatrix} \right) = \begin{bmatrix} R_k \delta_{kl} & 0 & 0 \\ 0 & Q_k \delta_{kl} & 0 \\ 0 & 0 & P_0 \end{bmatrix}, \;\; \delta_{kl} = \begin{cases} 1 & k = l \\ 0 & k \neq l \end{cases}$$

Lemma 5.52 *(Properties of x_k and y_k Coming from the Model)*

- *For all $k \geq 1$, x_k is a linear combination of x_0 and $w_0, \cdots, w_{k-1}$. In particular, x_k is uncorrelated with w_k.*
- *For all $k \geq 1$, y_k is a linear combination of x_0, $w_0, \cdots, w_{k-1}$, and $v_0, \cdots, v_k$. In particular, y_k is uncorrelated with w_k .*
- *For all $k \geq 0$, v_k is uncorrelated with x_k.*

The proof is by induction using the recursive nature of the discrete-time model. We skip it. The reader can easily fill it in.

In the next subsection, we give (one form of) the discrete-time Kalman Filter. After that, we provide the main elements of its derivation. There are many variations of the basic filter, all equivalent to the one we give, but some preferable over others for numerical reasons. Chapter 5.7.5 provides a version of the filter with the measurement update and prediction steps combined.

[7] Recall that in white light, all frequencies are present. When only certain frequency components are present, you get "colored" light, such as blue light or red light. The term "white" noise means that if you compute the power spectral density of the noise random process, it is a constant, meaning that all frequency components are equally represented, just as in white light.

[8] Recall that for normal random variables, uncorrelated and independent are the same thing. This is one of several special properties of Gaussian random variables.

5.7.2 Basic Kalman Filter

Definition of Terms:

$$\begin{aligned}\widehat{x}_{k|k} &:= \mathcal{E}\{x_k|y_0,\cdots,y_k\}\\ P_{k|k} &:= \mathcal{E}\{(x_k-\widehat{x}_{k|k})(x_k-\widehat{x}_{k|k})^\top|y_0,\cdots,y_k\}\end{aligned}$$

$$\begin{aligned}\widehat{x}_{k+1|k} &:= \mathcal{E}\{x_{k+1}|y_0,\cdots,y_k\}\\ P_{k+1|k} &:= \mathcal{E}\{(x_{k+1}-\widehat{x}_{k+1|k})(x_{k+1}-\widehat{x}_{k+1|k})^\top|y_0,\cdots,y_k\}\end{aligned}$$

Initial Conditions:

$$\widehat{x}_{0|-1} := \bar{x}_0 = \mathcal{E}\{x_0\}, \text{ and } P_{0|-1} := P_0 = \text{cov}(x_0)$$

For $k \geq 0$

Measurement Update Step:

$$\begin{aligned}K_k &= P_{k|k-1}C_k^\top\left(C_kP_{k|k-1}C_k^\top+Q_k\right)^{-1} \quad \text{(Kalman Gain)}\\ \widehat{x}_{k|k} &= \widehat{x}_{k|k-1}+K_k\left(y_k-C_k\widehat{x}_{k|k-1}\right)\\ P_{k|k} &= P_{k|k-1}-K_kC_kP_{k|k-1}\end{aligned}$$

Time Update or Prediction Step:

$$\begin{aligned}\widehat{x}_{k+1|k} &= A_k\widehat{x}_{k|k}\\ P_{k+1|k} &= A_kP_{k|k}A_k^\top+G_kR_kG_k^\top\end{aligned}$$

End of For Loop (Just stated this way to emphasize the recursive nature of the filter)

5.7.3 Preliminaries for the Derivation

All of the variables in the linear model, x_k, y_k, w_k and v_k, are multivariate normal random vectors. We seek to compute the density of the conditional random vector $x_k\big|(y_0,\cdots,y_k)$, that is, the state of the linear model conditioned on the collected measurements. Because the model is linear and the random vectors in the problem are jointly normal, the density is equivalent to the conditional mean of x_k and the conditional covariance, namely

$$\begin{aligned}\widehat{x}_{k|k} &:= \mathcal{E}\{x_k|y_0,\cdots,y_k\}\\ P_{k|k} &:= \mathcal{E}\{(x_k-\widehat{x}_{k|k})(x_k-\widehat{x}_{k|k})^\top|y_0,\cdots,y_k\}.\end{aligned}$$

We could apply the MVE from Theorem 5.27 and do a batch computation. This would rapidly become inefficient as the number of measurements grows over time. What we seek instead is a recursive form of minimum variance estimation, just as we created RLS, a recursive version of weighted least squares; see Proposition 3.91 and Proposition 3.92. The main difference here is that we seek to estimate more than the initial condition x_0. In fact, we seek to estimate x_k for $k \geq 0$.

Definition 5.53 ***(Measurements in the KF)*** *We collect all of the measurements up to and including time k*

$$Y_k = (y_k, y_{k-1},\cdots,y_0).$$

Strictly speaking, we should be stacking them up into a column vector as we have done for all of our estimation problems, but notationally, it is more convenient to write them in a row. Also, it is more convenient to put the most recent measurement at the head of the list. We note that $Y_k = (y_k, Y_{k-1})$.

Hence,

$$\begin{aligned}\widehat{x}_{k|k} :=& \mathcal{E}\{x_k|Y_k\}\\ P_{k|k} :=& \mathcal{E}\{(x_k-\widehat{x}_{k|k})(x_k-\widehat{x}_{k|k})^\top|Y_k\}\end{aligned}$$

mean and covariance of the conditional normal random vector $x_k|Y_k$

$$\begin{aligned}\widehat{x}_{k+1|k} :=& \mathcal{E}\{x_{k+1}|Y_k\}\\ P_{k+1|k} :=& \mathcal{E}\{(x_{k+1}-\widehat{x}_{k+1|k})(x_{k+1}-\widehat{x}_{k+1|k})^\top|Y_k\}\end{aligned}$$

mean and covariance of the conditional normal random vector $x_{k+1}|Y_k$

Remark 5.54 *We note that*

- *the conditional random vector $x_k|Y_k$ is distributed $N(\widehat{x}_{k|k}, P_{k|k})$, and*
- *the conditional random vector $x_{k+1}|Y_k$ is distributed $N(\widehat{x}_{k+1|k}, P_{k+1|k})$.*

5.7.4 Filter Derivation Using Induction and Properties of Conditional Distributions of Gaussian Random Vectors

Base step: The initial conditions of the filter at time $k = 0$, namely

$$\widehat{x}_{0|-1} := \bar{x}_0, \text{ and } P_{0|-1} := P_0$$

Induction step: At time $k \geq 0$, we suppose that $(\widehat{x}_{k|k-1}, P_{k|k-1})$ are known, and we derive $(\widehat{x}_{k|k}, P_{k|k})$ and $(\widehat{x}_{k+1|k}, P_{k+1|k})$.

Key idea of the development: We need to compute the distribution (or density) of the conditional random vector

$$x_k|Y_k = x_k|(y_k, Y_{k-1})$$

From **Key Fact** 4, we have that $X|(Y,Z) = X|Z \;\Big|\; Y|Z$. From this we obtain

$$x_k|Y_k = x_k|(y_k, Y_{k-1}) = x_k|Y_{k-1} \;\Big|\; y_k|Y_{k-1}, \tag{5.3}$$

where we have identified

$$x_k \leftrightarrow X, \;\; y_k \leftrightarrow Y, \text{ and } Y_{k-1} \leftrightarrow Z.$$

Hence, if we can compute the distribution (or density) of

$$\begin{bmatrix} x_k \\ y_k \end{bmatrix} \Big| Y_{k-1},$$

then we can apply **Key Fact** 1 to obtain (5.3). The following calculations are aimed at doing just this.

Measurement Update: We seek to derive the filter equations in Chapter 5.7.2. To begin, we have that $y_k = C_k x_k + v_k$. It follows by linearity that the conditional random variable $y_k|Y_{k-1}$ is equal to

$$y_k|Y_{k-1} = C_k\, x_k|Y_{k-1} + v_k|Y_{k-1}.$$

By the assumptions on the noise model, v_k is independent of both x_k and Y_{k-1}, and hence by **Key Fact** 2, the conditional random vector $v_k|Y_{k-1}$ is independent of the conditional random vector $x_k|Y_{k-1}$. Moreover, because v_k is independent of Y_{k-1}, we deduce that $v_k|Y_{k-1} = v_k$. Putting this together, we have that

$$y_k|Y_{k-1} = C_k\, x_k|Y_{k-1} + v_k,$$

and $x_k|Y_{k-1}$ and v_k are independent. Hence

$$\begin{aligned}\widehat{y}_{k|k-1} :=& \mathcal{E}\{y_k|Y_{k-1}\} \\ =& \mathcal{E}\{C_k x_k|Y_{k-1}\} + \mathcal{E}\{v_k\} \\ =& C_k \mathcal{E}\{x_k|Y_{k-1}\} + \mathcal{E}\{v_k\} \\ =& C_k \widehat{x}_{k|k-1} + 0 \\ =& C_k \widehat{x}_{k|k-1}.\end{aligned}$$

Moreover, the independence of $x_k|Y_{k-1}$ and v_k with **Key Fact** 3 yields

$$\mathrm{cov}(y_k|Y_{k-1}, y_k|Y_{k-1}) = C_k P_{k|k-1} C_k^\top + Q_k.$$

We use independence again to obtain

$$\mathrm{cov}(x_k|Y_{k-1}, y_k|Y_{k-1}) = \mathrm{cov}(x_k|Y_{k-1}, C_k\ x_k|Y_{k-1}) = P_{k|k-1} C_k^\top.$$

With this information, we conclude that the vector

$$\left[\begin{array}{c} x_k \\ y_k \end{array}\right] |Y_{k-1}$$

is jointly normally distributed, with mean and covariance

$$\left[\begin{array}{c} \widehat{x}_{k|k-1} \\ C_k \widehat{x}_{k|k-1} \end{array}\right], \left[\begin{array}{cc} P_{k|k-1} & P_{k|k-1} C_k^\top \\ C_k P_{k|k-1} & C_k P_{k|k-1} C_k^\top + Q_k \end{array}\right] \tag{5.4}$$

As discussed above, to compute the distribution of $(x_k|Y_k)$, we have from **Key Fact** 4

$$x_k|Y_k = x_k \Big| (y_k, Y_{k-1}) = x_k|Y_{k-1} \ \Big|\ y_k|Y_{k-1},$$

and thus applying **Key Fact** 1 to (5.4) we compute the mean and covariance of $x_k|Y_k = x_k|Y_{k-1} \ \Big|\ y_k|Y_{k-1}$ to be

$$\widehat{x}_{k|k} = \widehat{x}_{k|k-1} + P_{k|k-1} C_k^\top \left[C_k P_{k|k-1} C_k^\top + Q_k\right]^{-1} \left(y_k - C_k \widehat{x}_{k|k-1}\right)$$

$$P_{k|k} = P_{k|k-1} - P_{k|k-1} C_k^\top \left[C_k P_{k|k-1} C_k^\top + Q_k\right]^{-1} C_k P_{k|k-1}.$$

Remark 5.55 *We note that $P_{k|k}$ is the Schur complement $C_k P_{k|k-1} C_k^\top + Q_k$ in the covariance of*

$$\left[\begin{array}{c} x_k \\ y_k \end{array}\right] |Y_{k-1}$$

Prediction or Time Update: We seek to derive the remaining filter equations in Chapter 5.7.2. This time we use the state-variable model instead of the output model, namely

$$x_{k+1} = A_k x_k + G_k w_k,$$

and we are interested in the conditional random vector

$$x_{k+1}|Y_k = A_k x_k|Y_k + G_k w_k|Y_k.$$

Because x_k and Y_k are both independent of w_k, by **Key Fact** 2, $x_k|Y_k$ and $w_k|Y_k$ are also independent. It follows that

$$\begin{aligned}\widehat{x}_{k+1|k} &= \mathcal{E}\{x_{k+1}|Y_k\} \\ &= \mathcal{E}\{A_k x_k + G_k w_k|Y_k\} \\ &= A_k \mathcal{E}\{x_k|Y_k\} + G_k \mathcal{E}\{w_k|Y_k\} \\ &= A_k \widehat{x}_{k|k} + G_k \mathcal{E}\{w_k\} \\ &= A_k \widehat{x}_{k|k},\end{aligned}$$

where we have used $w_k|Y_k = w_k$, and $\mathcal{E}\{w_k\} = 0$. Next, we use **Key Fact** 3 and the conditional independence of the random vectors $x_k|Y_k$ and $w_k|Y_k$ to evaluate the covariance of $x_{k+1}|Y_k$ as

$$P_{k+1|k} = A_k P_{k|k} A_k^\top + G_k R_k G_k^\top.$$

That's the Proof Folks! The famous Kalman Filter can be derived using four Key Facts, where three of them depend on properties of conditional Gaussian random vectors and one is true for general conditional random vectors that have densities.

5.7.5 Combined Update Version of the Kalman Filter

Here, we will assume the model also has a *deterministic* input u_k, and thus

$$\begin{aligned} x_{k+1} &= A_k x_k + B_k u_k + G_k w_k \\ y_k &= C_k x_k + v_k, \end{aligned}$$

with the assumptions on the random vectors x_0, w_k and v_k the same as before.

Combined Filter: The measurement-update step and time-update step of the Kalman Filter can be combined into a single step. The algorithm becomes:

Initial Conditions:

$$\widehat{x}_{0|-1} := \bar{x}_0 = \mathcal{E}\{x_0\}, \text{ and } P_{0|-1} := P_0 = \text{cov}(x_0)$$

For $k \geq 0$

$$K_k = \left(P_{k|k-1} C_k^\top\right) \left[C_k P_{k|k-1} C_k^\top + Q_k\right]^{-1}$$

$$\widehat{x}_{k+1|k} = A_k \widehat{x}_{k|k-1} + B_k u_k + A_k K_k \left(y_k - C_k \widehat{x}_{k|k-1}\right)$$

$$P_{k+1|k} = A_k \left[P_{k|k-1} - K_k C_k P_{k|k-1}\right] A_k^\top + G_k R_k G_k^\top$$

End of For Loop

Remark 5.56 *You do not have to start at $k = 0$. In MATLAB, it is often easier to begin with $k = 1$. In that case, the initial conditions are*

$$\widehat{x}_{1|0} := \bar{x}_0 = \mathcal{E}\{x_0\}, \text{ and } P_{1|0} := P_0 = \text{cov}(x_0).$$

Remark 5.57 $K_k C_k P_{k|k-1} = \left(P_{k|k-1} C_k^\top\right) \left[C_k P_{k|k-1} C_k^\top + Q_k\right]^{-1} C_k P_{k|k-1}$ *is symmetric positive semi-definite and represents the value of the measurement in reducing the covariance of the state estimate, just as in the MVE.*

5.7.6 (Optional Read) Extended Kalman Filter or EKF

An important reason for developing the KF for time-varying linear systems is that it can then be applied to the linearization of a nonlinear system along an estimated trajectory, which gives the Extended Kalman Filter or EKF for short. Its objective is to estimate the state of a nonlinear model, perturbed by noise terms,

$$\begin{aligned} x_{k+1} &= f(x_k, u_k) + G w_k \\ y_k &= h(x_k) + v_k. \end{aligned}$$

For a nonlinear system, the EKF will not be an exact computation of the conditional mean of x_k given all of the measurements up to time k, but rather, **an approximate calculation of** $\widehat{x}_{k|k}$.

There are several other versions of "Extended Kalman Filters", many of which are treated here https://en.wikipedia.org/wiki/Kalman_filter. We particularly recommend you read about the **unscented Kalman filter**, which has a powerful means for approximating the effect on nonlinear functions on Gaussian noise, https://en.wikipedia.org/wiki/Kalman_filter#Unscented_Kalman_filter.

Definition of Terms: (Changes for EKF given in color)

$$\widehat{x}_{k|k} :\approx \mathcal{E}\{x_k|y_0,\cdots,y_k\}$$
$$P_{k|k} :\approx \mathcal{E}\{(x_k-\widehat{x}_{k|k})(x_k-\widehat{x}_{k|k})^\top|y_0,\cdots,y_k\}$$

$$\widehat{x}_{k+1|k} :\approx \mathcal{E}\{x_{k+1}|y_0,\cdots,y_k\}$$
$$P_{k+1|k} :\approx \mathcal{E}\{(x_{k+1}-\widehat{x}_{k+1|k})(x_{k+1}-\widehat{x}_{k+1|k})^\top|y_0,\cdots,y_k\}$$

Initial Conditions:

$$\widehat{x}_{0|-1} := \bar{x}_0 = \mathcal{E}\{x_0\}, \text{ and } P_{0|-1} := P_0 = \text{cov}(x_0)$$

For $k \geq 0$

Measurement Update Step:

$$C_k := \left.\frac{\partial h(x)}{\partial x}\right|_{\widehat{x}_{k|k-1}}$$
$$K_k = P_{k|k-1}C_k^\top\left(C_kP_{k|k-1}C_k^\top+Q_k\right)^{-1} \text{ (Kalman Gain)}$$
$$\widehat{x}_{k|k} = \widehat{x}_{k|k-1}+K_k\left(y_k-h(\widehat{x}_{k|k-1})\right)$$
$$P_{k|k} = P_{k|k-1}-K_kC_kP_{k|k-1}$$

Time Update or Prediction Step:

$$\widehat{x}_{k+1|k} = f(\widehat{x}_{k|k},u_k) \quad \text{(Use the nonlinear model to predict the next state)}$$
$$A_k := \left.\frac{\partial f(x,u_k)}{\partial x}\right|_{\widehat{x}_{k|k}} \quad \text{(Partial with respect to x only)}$$
$$P_{k+1|k} = A_kP_{k|k}A_k^\top+G_kR_kG_k^\top$$

End of For Loop

Yes, the EFK above is pretty much line for line the same as the standard KF. Recall our remarks at the beginning on alternative forms of the EKF.

Remark 5.58 *The EKF given above has been analyzed as a deterministic observer for a nonlinear discrete-time system. More information can be found in the journal paper* `http://ece.umich.edu/faculty/grizzle/papers/ekf.pdf` *or the conference version via DOI* ***10.23919/ACC.1992.4792775***

5.8 (Optional Read) Luenberger Observer

The true Luenberger Observer corresponds to Luenberger's "reduced-order" observer. What we are presenting was too simple for Luenberger to even write down. Nevertheless, the field of Dynamics and Control commonly refers to it as **The Luenberger Observer** and we will too!

The Luenberger observer is a "deterministic estimator". We consider the time-invariant case

$$x_{k+1} = Ax_k$$
$$y_k = Cx_k$$

where $x \in \mathbb{R}^n$, $y \in \mathbb{R}^p$, $A \in \mathbb{R}^{n\times n}$, and $C \in \mathbb{R}^{p\times n}$.

Question 1: (Observability) When can we reconstruct the initial condition (x_0) from the measurements $y_0, y_1, y_2, \ldots$?

$$\begin{aligned} y_0 &= Cx_0 \\ y_1 &= Cx_1 = CAx_0 \\ y_2 &= Cx_2 = CAx_1 = CA^2x_0 \\ &\vdots \\ y_k &= CA^kx_0 \end{aligned}$$

Rewriting the above in matrix form yields,

$$\begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_k \end{bmatrix} = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^k \end{bmatrix} x_0$$

We note that if $\text{rank} \begin{bmatrix} C \\ CA \\ \vdots \\ CA^k \end{bmatrix} = n$, then the null space consists of the zero vector and we can determine x_0 uniquely on the basis of the measurements.

The **Caley Hamilton Theorem** proves that $\text{rank} \begin{bmatrix} C \\ CA \\ \vdots \\ CA^k \end{bmatrix} = \text{rank} \begin{bmatrix} C \\ CA \\ \vdots \\ CA^n - 1 \end{bmatrix}$ for all $k \geq (n-1)$.

Fact 5.59 *(Kalman observability rank condition.) We can determine x_0 uniquely from the measurements if, and only if,* $\text{rank} \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix} = n$.

Question 2:(Full-State Luenberger Observer) Can we process the measurements dynamically (i.e. recursively) to "estimate" x_k?

Define

$$\hat{x}_{k+1} = A\hat{x}_k + L(y_k - C\hat{x}_k),$$

where the matrix L is to be chosen, and define the error to be $e_k := x_k - \hat{x}_k$. We seek conditions so that $e_k \to 0$ as $k \to \infty$, because, if $e_k \to 0$ because then $\hat{x}_k \to x_k$.

$$\begin{aligned} e_{k+1} &= x_{k+1} - \hat{x}_{k+1} \\ &= Ax_k - [A\hat{x}_k + L(y_k - C\hat{x}_k)] \\ &= A(x_k - \hat{x}_k) - LC(x_k - \hat{x}_k) \\ &= Ae_k - LCe_k \end{aligned}$$

$$\boxed{e_{k+1} = (A - LC)e_k}$$

Fact 5.60 *Let $e_0 \in \mathbb{R}^n$ and define $e_{k+1} = (A - LC)e_k$. The the sequence $e_k \to 0$ as $k \to \infty$ for all $e_0 \in \mathbb{R}^n$ if, and only if, $|\lambda_i(A - LC)| < 1$ for $i = 1, \ldots, n$.*

Fact 5.61 *A sufficient condition for the existence of $L : \mathbb{R}^p \to \mathbb{R}^n$ that places eigenvalues of $(A - LC)$ in the unit circle is:*

$$\text{rank} \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix} = n = dim(x)$$

.

Remark 5.62 *L is called the **Luenberger gain**. When the models in the the Kalman Filter for the state variables and the noise terms are time invariant, and the observability condition discussed above is met, the Kalman gain will also converge to a constant vector or matrix, K_{ss}, called the steady-state Kalman gain.*

1. *Reason to choose one gain over the other: Optimality of the estimate when you know the noise statistics.*
2. *Kalman Filter works for time-varying models A_k, C_k, G_k, etc.*

5.9 (Optional Read) Information Matrix of Gaussian Random Vectors

Remark 5.63 *We briefly discuss the "information" or "precision" matrix. You will likely encounter it in other courses, such as Mobile Robotics, or in papers. We first saw it in ROB 501 when we compared BLUE to Weighted Least Squares (they are the same when the weighting matrix W is chosen as the Information Matrix of the noise term). You will not need to know anything about the information matrix in the context of the Kalman Filter: when seeing the filter for the first time, you do not need to do every possible variation.* □

The Kalman filter can be written in many forms. One alternative form propagates the inverse of the covariance matrix instead of the covariance matrix. The inverse of the covariance matrix has two names: *information matrix* and *precision matrix*. We will use the first one:

$$\text{Information matrix: } \Lambda := \Sigma^{-1}$$

We decompose it just as we did with the covariance matrix.

$$\Lambda =: \begin{bmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{bmatrix}$$

The formula for inversion of block matrices gives

$$\begin{aligned}
\Lambda_{11} &= (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1} \\
\Lambda_{12} &= -(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}\Sigma_{12}\Sigma_{22}^{-1} \\
\Lambda_{21} &= \Lambda_{12}^{\top} \\
\Lambda_{22} &= \Sigma_{22}^{-1} + \Sigma_{22}^{-1}\Sigma_{21}(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}\Sigma_{12}\Sigma_{22}^{-1}
\end{aligned}$$

(See http://en.wikipedia.org/wiki/Matrix_inversion_lemma#Blockwise_inversion)

We also scale the mean by defining

$$\eta := \Lambda\mu$$

that is,

$$\begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} := \begin{bmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$$

Remark 5.64 *For a multivariate normal distribution, it is equivalent to know η and Λ or μ and Σ. We go back and forth between the two by matrix inversion and multiplication. One sometimes says that these are dual parameterizations for the normal distribution. We only mention the alternative parameterization with the information matrix because sometimes it is easier to use than the more standard mean and covariance representation.* □

Fact 5.65 *(Conditional Distributions Using the Information Matrix) The information matrix of the random variable X_1 given that $X_2 = x_2$ is*

$$\Lambda_{1|2} = \Lambda_{11}$$

and

$$\eta_{1|2} = \eta_1 - \Lambda_{12}x_2$$

In other words, if you have the information matrix handy, computing the conditional distribution is easier with it than with the covariance matrix. We note that if you want to go back to the standard representation, then

$$\Sigma_{1|2} = \Lambda_{1|2}^{-1}$$

and

$$\mu_{1|2} = \Lambda_{1|2}^{-1}\,\eta_{1|2}$$

□

Remark 5.66 *(Marginal Distributions Using the Information Matrix) Getting the marginal distributions from the information form of the distribution is more complicated. If you are interested, you can easily find it on the web or in most graduate level probability texts.* □

5.10 (Optional Read) Deriving MVE as we did BLUE

Example 5.67 *The Minimum Variance Estimator (MVE) can be derived by assuming a linear solution and coverting the problem to a deterministic optimization problem, in a similar manner to how we obtained the Best Linear Unbiased Estimator, BLUE. This multiple-choice problem requires that you develop that approach.*
Problem Data:

- $y = Cx + \epsilon$, $y \in \mathbb{R}^m$ *and* $x \in \mathbb{R}^n$.
- $E\{\epsilon\} = 0$ *and* $E\{x\} = 0$
- $E\{\epsilon\epsilon^T\} = Q$, $E\{xx^T\} = P$, $E\{\epsilon x^T\} = 0$, $E\{x\epsilon^T\} = 0$
- *For simplicity, we take* $Q > 0$ *and* $P > 0$, *which is a sufficient condition for* $(CPC^\top + Q) > 0$.
- *We seek to estimate* x *on the basis of* y *and the given statistical information about* x *so that we minimize the variance*

$$E\{||\hat{x} - x||^2\} = E\{\sum_{i=1}^{n} (\hat{x}_i - x_i)^2\} = \sum_{i=1}^{n} E\{(\hat{x}_i - x_i)^2\} \quad (*)$$

Given Fact: *From (*), we have* $\hat{x} = Ky$ *minimizes* $E\{||\hat{x} - x||^2\}$ *if, and only if, for* $1 \le i \le n$, $\hat{x}_i = k_i y$ *minimizes* $E\{(\hat{x}_i - x_i)^2\}$, *where*

$$K = \begin{bmatrix} k_1 \\ k_2 \\ \vdots \\ k_n \end{bmatrix}$$

Hints: *(1) If* M *is* $1 \times n$ *so that* $z = Mx$ *is a scalar, then* $z^2 = Mxx^\top M^\top$; *(2)* $x_i = e_i^\top x$, *where* $[e_1 \mid e_2 \mid \cdots \mid e_n] = I_{n\times n}$. *(Yes,* $[e_i]_j = 1 \iff i = j$, *and zero otherwise.)*

Your work starts here. Everything above is assumed given and true. *Determine which of the following statements are TRUE. At least one answer is FALSE and at least one answer is TRUE.*

(a) $\widehat{x} = Ky$ *is unbiased* [9] *if, and only if,* $KC = I$.

(b) Let $\widehat{x}_i = k_i y$. *Then*

$$E\{(\widehat{x}_i - x_i)^2\} = \begin{bmatrix} C^\top k_i^\top - e_i \\ k_i^\top \end{bmatrix}^\top \begin{bmatrix} P & 0 \\ 0 & Q \end{bmatrix} \begin{bmatrix} C^\top k_i^\top - e_i \\ k_i^\top \end{bmatrix}$$

where $[e_1 \mid e_2 \mid \cdots \mid e_n] = I_{n\times n}$.

(c) Let $\widehat{x}_i = k_i y$. *Then* $E\{(\widehat{x}_i - x_i)^2\} = k_i(CPC^\top + Q)k_i^\top$.

(d) Let $\widehat{k}_i := arg\ \min E\{(k_i(Cx + \epsilon) - x_i)^2\}$ *denote the gain that minimizes the variance for a linear estimator* $\widehat{x}_i = k_i y$. *Then* $\widehat{k}_i$ *can be obtained by minimizing the error of the over determined equation*

$$Ak_i^\top = b,$$

with

$$A = \begin{bmatrix} C^\top \\ I_{m\times m} \end{bmatrix}, \ b = \begin{bmatrix} e_i \\ 0_{m\times 1} \end{bmatrix},$$

and e_i *is defined as in part (b); the norm on* $\begin{bmatrix} \alpha \\ \beta \end{bmatrix} \in \mathbb{R}^{n+m}$ *is given by*

$$||\begin{bmatrix} \alpha \\ \beta \end{bmatrix}||^2 = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}^\top \begin{bmatrix} P & 0 \\ 0 & Q \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix}.$$

[9] Recall that an estimator is said to be unbiased when $\mathcal{E}\{\widehat{x}\} = \mathcal{E}\{x\}$

Solution: The answers are (b) and (d). We work the problem completely, then answer the questions.

Assume that $\widehat{x} = Ky$ for some $n \times m$ real matrix K. Then we compute

$$E\{\widehat{x}\} = E\{K(Cx + \epsilon)\} = KCE\{x\} + KE\{\epsilon\} = 0 = E\{x\},$$

because $E\{x\} = 0$ and $E\{\epsilon\} = 0$. In BLUE, x was deterministic, and thus $E\{x\} = x$ instead of zero gave us $E\{\widehat{x}\} = KCx$, which was the source of the constraint, $KC = I$. In MVE, we do not have this restriction.

We set $\widehat{x}_i = k_i y$ and note that $x_i = e_i^\top x$. We then compute

$$\begin{aligned}
(\widehat{x}_i - x_i)^2 &= (k_i[Cx + \epsilon] - x_i)^2 \\
&= \left(k_i[Cx + \epsilon] - e_i^\top x\right)^2 \\
&= \left([k_i C - e_i^\top]x + k_i\epsilon\right)^2 \\
&= \left([k_i C - e_i^\top]x + k_i\epsilon\right)\left([k_i C - e_i^\top]x + k_i\epsilon\right)^\top \\
&= [k_i C - e_i^\top]xx^\top[C^\top k_i^\top - e_i] + 2[k_i C - e_i^\top]x\epsilon^\top k_i^\top + k_i\epsilon\epsilon^\top k_i^\top \\
&= [C^\top k_i^\top - e_i]^\top xx^\top[C^\top k_i^\top - e_i] + 2[C^\top k_i^\top - e_i]^\top x\epsilon^\top k_i^\top + k_i\epsilon\epsilon^\top k_i^\top,
\end{aligned}$$

where we have used the fact that a scalar squared is equal to the scalar times its transpose.
We now use the given statistical data to conclude that

$$\begin{aligned}
E\{(\widehat{x}_i - x_i)^2\} &= [C^\top k_i^\top - e_i]^\top P\,[C^\top k_i^\top - e_i] + k_i Q k_i^\top \\
&= \begin{bmatrix} C^\top k_i^\top - e_i \\ k_i^\top \end{bmatrix}^\top \begin{bmatrix} P & 0 \\ 0 & Q \end{bmatrix} \begin{bmatrix} C^\top k_i^\top - e_i \\ k_i^\top \end{bmatrix}.
\end{aligned}$$

It follows that

$$\widehat{k}_i = \arg\,\min E\{(\widehat{x}_i - x_i)^2\} = \arg\,\min E\{(k_i(Cx + \epsilon) - x_i)^2\},$$

is also the solution to the over determined equation[10]

$$\widehat{k}_i^\top = \arg\,\min ||Ak_i^\top - b||^2$$

with

$$A = \begin{bmatrix} C^\top \\ I_{m\times m} \end{bmatrix},\ b = \begin{bmatrix} e_i \\ 0_{m\times 1} \end{bmatrix},$$

and the norm on $\mathbb{R}^{n+m}$ given by

$$||\begin{bmatrix} \alpha \\ \beta \end{bmatrix}||^2 = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}^\top \begin{bmatrix} P & 0 \\ 0 & Q \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix}.$$

Indeed, applying our "Magic Formula" gives

$$\begin{aligned}
\widehat{k}_i^\top &= \left(A^\top \begin{bmatrix} P & 0 \\ 0 & Q \end{bmatrix} A\right)^{-1} A^\top \begin{bmatrix} P & 0 \\ 0 & Q \end{bmatrix} b \\
&= (CPC^\top + Q)^{-1} CPe_i.
\end{aligned}$$

Therefore

$$\widehat{K}^\top = (CPC^\top + Q)^{-1} CP$$

and

$$\widehat{K} = PC^\top (CPC^\top + Q)^{-1},$$

the result from lecture.
The answers to the questions are therefore:

(a) False. This was given in lecture.

[10]The columns of A are linearly independent because the lower block is the identity matrix.

(b) True. This was a bit messy in terms of matrix algebra, but otherwise straightforward.

(c) False. Clear once (b) is known.

(d) True. If the answer were not given so that all you had to do was check it, then, yes, it would have been pretty tough to see it. As given, it was not so bad. Nevertheless, the solution developed in lecture, applying the Projection Theorem to a vector space of random variables, seems far superior to the method used here: (1) the method in lecture allowed a more general solution, namely $CPC^\top + Q \succ 0$ instead of $P \succ 0$ and $Q \succ 0$; (2) the Projection Theorem in lecture resulted in easier computations, and each step was well motivated through the calculation of the Gram matrix.

Chapter 6

Enough Real Analysis to Understand the Existence of Limits of Sequences as well as Extrema of Functions

Learning Objectives

- In Robotics Engineering, we are constantly faced with problems for which no closed form solution is possible. For such problems, we seek algorithms that compute solutions. In many cases, these algorithms do not terminate with an exact answer in a finite number of steps. Hence, we need to understand how to formulate and analyze the convergence of iterative processes.
- Another class of problems involve maximizing or minimizing real valued functions. It is important to understand sufficient conditions for extrema of real valued functions to exist.

Outcomes

- Understand the concepts of open and closed sets in normed spaces.
- Understand sequences and how to analyse if they have limits.
- Characterize open and closed sets in terms of distance (ie., using the norm directly) and in terms of convergent sequences.
- Understand in a rigorous manner what is a continuous function. While we saw the definition in Calculus, for most of us, it did not stick!
- Learn an interesting topological property called compactness and how it relates to the existence of extrema of functions.

6.1 Open and Closed Sets in Normed Spaces

We start with a few preliminaries. Let $(\mathcal{X},\ \mathbb{R},\ \|\bullet\|)$ be a real normed space. Recall from Chapter 3.1 that a norm is a function $\|\bullet\|:\ \mathcal{X}\to[0,\ +\infty)$ satisfying

(a) $\|x\|\geq 0$ and $\|x\|=0\iff x=0$

(b) $\|\alpha\cdot x\|=|\alpha|\cdot\|x\|$ for all $\alpha\in\mathbb{R}$, $x\in\mathcal{X}$

(c) $\|x+y\|\leq\|x\|+\|y\|$ for all $x,\ y\in\mathcal{X}$.

In addition, recall that

(d) For $x,\ y\in\mathcal{X}$, $d(x,\ y):=\|x-y\|$.

(e) For $x\in\mathcal{X}$, $S\subset\mathcal{X}$ a subset, $d(x,\ S):=\inf_{y\in S}\|x-y\|$.

(f) $A\subset B\iff A\cap(\sim B)=\emptyset$.

Definition 6.1 *Let $x_0\in\mathcal{X}$ and $a\in\mathbb{R}$, $a>0$. The* ***open ball of radius*** *a* ***center at*** *x_0 is*

$$B_a(x_0)=\{x\in\mathcal{X}\mid\|x-x_0\|<a\}.$$

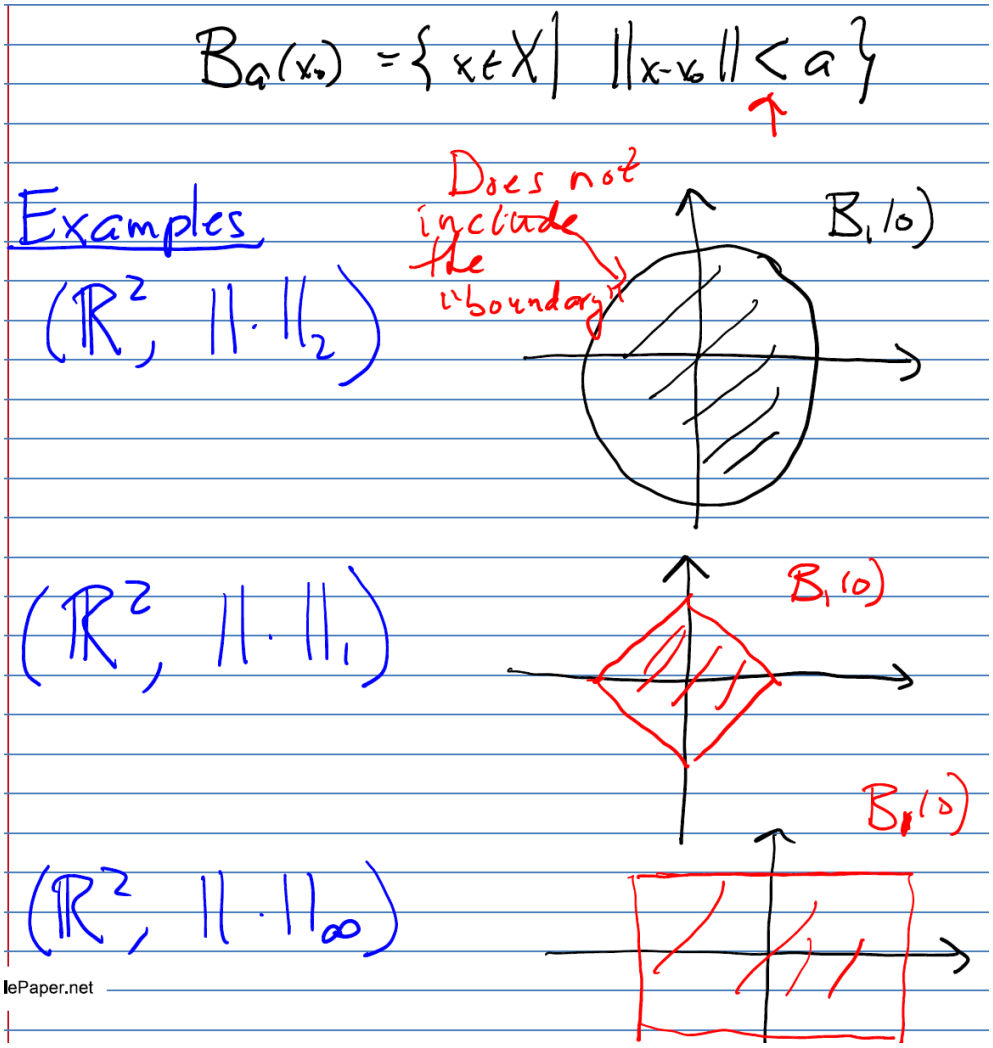


Figure 6.1: $(\mathbb{R}^2,\ \|\bullet\|)$ with various norms: (a) Euclidean norm $\|(x_1,\ x_2)\|_2=\sqrt{|x_1|^2+|x_2|^2}$; (b) One norm $\|(x_1,\ x_2)\|_1=|x_1|+|x_2|$; and (c) Max norm $\|(x_1,\ x_2)\|_\infty=\max_{1\leq i\leq 2}|x_i|$.

Lemma 6.2 *(Characterization of distance zero and greater than zero) Let $(\mathcal{X},\ \|\bullet\|)$ be a normed space, $x \in \mathcal{X}$, and $S \subset \mathcal{X}$. Then,*

$$\begin{aligned} d(x,\ S) = 0 &\iff \forall \epsilon > 0,\ \exists y \in S,\ \|x - y\| < \epsilon \ \ (\textit{definition of the infimum}) \\ &\iff \forall \epsilon > 0,\ B_\epsilon(x) \cap S \neq \emptyset \ \ (\textit{definition of an open ball of radius } \epsilon). \textit{ Moreover,} \end{aligned}$$

$$\begin{aligned} d(x,\ S) > 0 &\iff \exists \epsilon > 0,\ \forall y \in S,\ \|x - y\| \geq \epsilon \\ &\iff \exists \epsilon > 0 \text{ such that } B_\epsilon(x) \cap S = \emptyset \\ &\iff \exists \epsilon > 0 \text{ such that } B_\epsilon(x) \subset (\sim S). \end{aligned}$$

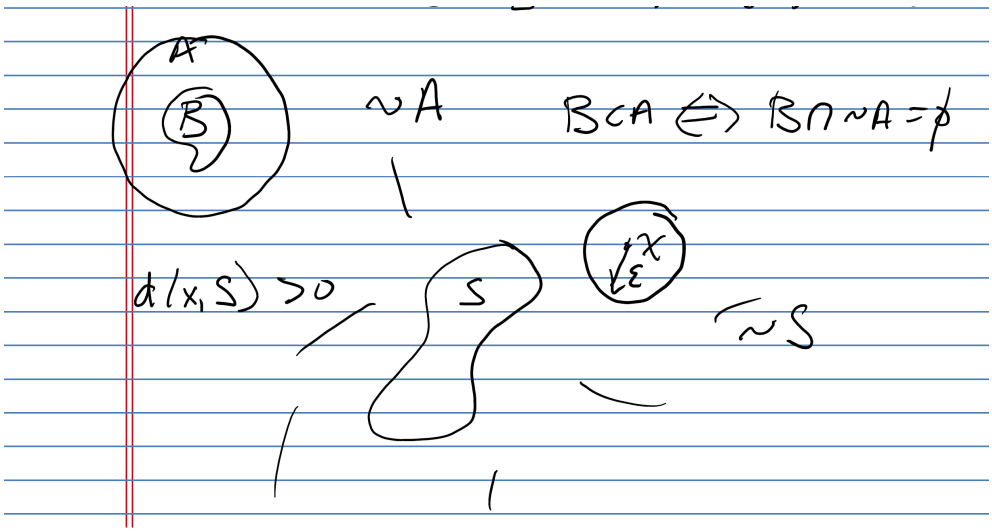


Figure 6.2: Why $(d(x, S) > 0) \iff (\exists\ \epsilon > 0$ such that $B_\epsilon(x) \subset (\sim S)) \iff (\exists\ \epsilon > 0$ such that $B_\epsilon(x) \cap S = \emptyset)$. We simply take $\epsilon = \frac{d(x,S)}{2} > 0$.

In the following, we assume $(\mathcal{X},\ \|\bullet\|)$ is given.

Definition 6.3 *Let $P \subset \mathcal{X}$, a subset of $\mathcal{X}$.*

*(a) A point $p \in P$ is **an interior point of** P if $\exists \epsilon > 0$ such that $B_\epsilon(p) \subset P$.*

*(b) The **interior of** P is $\mathring{P} := \{p \in P \mid p$ is an interior point$\}$.*

Remark 6.4

$$\begin{aligned} \mathring{P} &= \{p \in P \mid \exists \epsilon > 0, B_\epsilon(p) \subset P\} \\ &= \{p \in P \mid d(p, \sim P) > 0\} \\ &= \{x \in \mathcal{X} \mid d(x, \sim P) > 0\} \end{aligned}$$

because if $x \in (\sim P)$, then $d(x, \sim P) = 0$ and $\mathcal{X} = P \cup (\sim P)$. Hence,

$$\boxed{\mathring{P} = \{x \in \mathcal{X} \mid d(x, \sim P) > 0\}.}$$

Definition 6.5 *P is **open** if $\mathring{P} = P$.*

Remark 6.6 *Hence, P is open if, and only if, $P = \{x \in \mathcal{X} \mid d(x, \sim P) > 0\}$.*

Example 6.7 *Checking if sets are open or not.*

- *Is $P = (0,\ 1) \subset (\mathbb{R},\ \|\bullet\|)$ open? We note that $(x \in P \implies 0 < x < 1)$ and define $\epsilon = \min\{\frac{x}{2}, \frac{1-x}{2}\}$. Then $B_\epsilon(x) \subset P$ and hence P is open.*

- *Is $P = (0,\ 1) \subset (\mathbb{R},\ \|\bullet\|)$ open? We check it a second way. $\sim P = (-\infty, 0] \cup [1, \infty)$. We have $x \in P \iff 0 < x < 1$. Hence,*

$$\begin{aligned} d(x, (-\infty, 0]) &= x > 0 \\ d(x, [1, \infty)) &= 1 - x > 0 \\ d(x, \sim P) &= \min\{x, 1 - x\} > 0. \end{aligned}$$

  Hence, $x \in P \iff d(x, \sim P) > 0$, and thus P is open.

- *$P = [0,\ 1) \subset (\mathbb{R},\ |\bullet|)$ is not open because $0 \in P$, and $\forall \epsilon > 0$, $B_\epsilon(0) \cap (\sim P) \neq \emptyset$ or we can also check it's not open because $0 \in P$ and $d(0, \sim P) = 0$.*

Definition 6.8 *$P \subset \mathcal{X}$ a subset.*

1. *A point $x \in \mathcal{X}$ is a **closure point** of P if $\forall \epsilon > 0$, $\exists p \in P$ such that $\|x - p\| < \epsilon$, in other words, $d(x, P) = 0$.*
2. *The **closure of** P is $\overline{P} := \{x \in \mathcal{X} \mid x \text{ is a closure point}\}$*

Definition 6.9 *P is **closed** if $\overline{P} = P$.*

Example 6.10 *Consider $(\mathcal{X}, \|\bullet\|) = (\mathbb{R}, |\bullet|)$.*

1. *$P = [0,\ 1)$ is not closed because $1 \notin P$, and $d(1, P) = 0$.*
2. *$P = [0,\ 1]$ is closed because $x \notin P$ implies $d(x, P) = \max\{-x, x - 1\} > 0$.*
3. *$P = (0,\ 1) \implies \overline{P} = [0,\ 1]$ because $d(0, P) = 0$, $d(1, P) = 0$ and for $x \notin [0, 1], d(x, P) > 0$.*

Fact 6.11 *The rational numbers $\mathcal{Q} \subset \mathbb{R}$ are neither closed nor open. $\overline{\mathcal{Q}} = \mathbb{R}$.*

Theorem 6.12 ***(Characterization of Open and Closed Sets using Distance)** Let $(\mathcal{X}, \|\bullet\|)$ be a normed space and $P \subset \mathcal{X}$ a subset. Then P is open if, and only if, $\sim P$ is closed.*

$$\boxed{\begin{aligned} P \text{ is closed} &\iff \sim P \text{ is open.} \\ P \text{ is open} &\iff \sim P \text{ is closed.} \end{aligned}}$$

Proof: The proof can be given in one line.

$$\boxed{\underbrace{\sim P = \sim (\mathring{P})}_{P \text{ is open}} = \{x \in \mathcal{X} \mid d(x,\ \sim P) = 0\} = \underbrace{\overline{\sim P} = \sim P}_{\sim P \text{ is closed}}}$$

We'll unpack it for you.

$$\begin{aligned} P = \mathring{P} &\iff P = \{x \in \mathcal{X} \mid d(x, \sim P) > 0\} \\ &\iff P = \{x \in \mathcal{X} \mid \exists \epsilon > 0, B_\epsilon(x) \cap P = \emptyset\} \\ &\iff P = \sim \{x \in \mathcal{X} \mid \forall \epsilon > 0, B_\epsilon(x) \cap P \neq \emptyset\} \\ &\iff \sim P = \{x \in \mathcal{X} \mid \forall \epsilon > 0, B_\epsilon(x) \cap P \neq \emptyset\} \\ &\iff \sim P = \{x \in \mathcal{X} \mid d(x, \sim P) = 0\} \\ &\iff \sim P = \overline{\sim P} \end{aligned}$$

Hence, $P = \mathring{P} \iff \sim P = \overline{\sim P}$, so P is open if, and only if, $\sim P$ is closed. ■

Remark 6.13 *Can a set be both open and closed? Yes. Such sets are sometimes called **clopen**! If $(\mathcal{X}, \|\bullet\|)$ is a normed space, then $\mathcal{X}$ is both open and closed. By convention,the empty set $\emptyset$ is both open and closed (Why? For two reasons: (i) Because it does not violate the conditions to be open or closed. (ii) We want the set complement of an open set to be a closed set and vice versa).*

Exercise 6.14 *Show the following:*

(a) An arbitrary union of open sets is open.

(b) A arbitrary intersection of closed sets is closed.

(c) A finite intersection of open sets is open.

(d) A finite union of closed sets is closed.

Example 6.15 *Consider the real numbers as a normed space; that is, $\mathcal{X} = \mathbb{R}$ and define the norm to be $||x|| = |x|$, the standard absolute value. Is the infinite intersection of open sets*

$$\bigcap_{n=1}^{\infty}(-1-\frac{1}{n},1)$$

open?

Solution: No. Indeed, $\forall\ n \geq 1$, $[-1,1) \subset (-1-\frac{1}{n},1)$. Hence, by definition of the intersection, $[-1,1) \subset \bigcap_{n=1}^{\infty}(-1-\frac{1}{n},1)$. Moreover,

$$[-1,1) = \bigcap_{n=1}^{\infty}(-1-\frac{1}{n},1),$$

because if $x < -1$, then there exits $1 \leq K < \infty$ such that $x < -1-\frac{1}{K}$, which implies that $x \notin (-1-\frac{1}{K},1)$, and thus

$$x \notin \bigcap_{n=1}^{\infty}(-1-\frac{1}{n},1).$$

■

Exercise 6.16 *Show the following:*

(a) P is closed if, and only if, $\overline{P} \subset P$.

(b) P is open if, and only if, $P \subset \mathring{P}$.

Definition 6.17 *The **boundary** of $S \subset \mathcal{X}$ is $\partial S := \overline{S} \cap \overline{(\sim S)}$.*

Exercise 6.18 *Show that $\partial S = \overline{S} \setminus \mathring{S} := \{x \in \overline{S} \mid x \notin \mathring{S}\}$.*

6.2 Newton-Raphson Algorithm

As Roboticists and Engineers, we are often faced with problems for which closed-form solutions are unknown or do not exist. For such problems, we often seek iterative means to either compute a solution or to prove the existence of a solution.

Consider a function $f : \mathbb{R}^n \to \mathbb{R}^n$ that is continuously differentiable and suppose we seek a root $f(x^*) = 0$. Note that the domain and range are both $\mathbb{R}^n$ and thus this is the nonlinear equivalent of solving a square linear equation $Ax - b = 0$.

The **Newton-Raphson** Algorithm is a vector version of Newton's Algorithm; see Chapter 11 of the ROB 101 Textbook. Let $x_k \in \mathbb{R}^n$ be our current approximation of a root of the function f. We write the linear approximation of f about the point x_k as

$$f(x) \approx f(x_k) + \frac{\partial f(x_k)}{\partial x} \cdot (x - x_k). \tag{6.1}$$

We want to chose x_{k+1} so that $f(x_{k+1}) = 0$, but we cannot do that exactly. Based on the linear approximation in (6.1), we have that

$$f(x_{k+1}) \approx 0 \iff 0 \approx f(x_k) + \frac{\partial f(x_k)}{\partial x} \cdot (x_{k+1} - x_k). \tag{6.2}$$

If $\det\left(\frac{\partial f(x_k)}{\partial x}\right) \neq 0$, we can solve for x_{k+1}, giving us the standard form of the Newton-Raphson Algorithm,

$$\boxed{x_{k+1} = x_k - \left(\frac{\partial f(x_k)}{\partial x}\right)^{-1} f(x_k).}$$

The "hope" is that each iteration of the algorithm produces a better approximation x_k to a root of the function $f(x)$, where by a better approximation, we mean that if x^* is an unknown root of f, then in the limit as k gets sufficiently large, we can make $\|x^* - x_k\|$ arbitrarily small. We formalize these ideas in Chapter 6.3 with the notion of a converging sequence and in Chapter 6.5 on Contraction Mappings. For now, we'll simply see a numerical illustration of these ideas in action.

Alternative Form of the Newton-Raphson Algorithm

Based on (6.2), we define[a] $x_{k+1} := x_k + \Delta x_k$, where Δx_k is our update to x_k. We can then break the algorithm into two steps,

$$\left(\frac{\partial f(x_k)}{\partial x}\right)\Delta x_k = -f(x_k) \qquad (\text{solve for } \Delta x_k) \tag{6.3}$$

$$x_{k+1} = x_k + \Delta x_k \ \ (\text{use } \Delta x_k \text{ to update our estimate of the root}). \tag{6.4}$$

While for toy problems, we can use the matrix inverse to solve (6.3) for Δx_k, for larger problems, we recommend using an LU Factorization or a QR Factorization. Once (6.3) has been solved, x_{k+1} is updated in (6.4) and the process repeats.

A **damped Newton-Raphson Algorithm** is obtained by replacing (6.4) with

$$x_{k+1} = x_k + \epsilon \Delta x_k, \tag{6.5}$$

for some $\epsilon > 0$. The validity of the Newton-Raphson Algorithm rests upon:

- the function f being differentiable;
- the Jacobian $\frac{\partial f(x_k)}{\partial x}$ having a non-zero determinant at points generated by (6.3) and (6.4); and
- the linear equation $f_{\text{lin}}(x) = f(x_k) + \frac{\partial f(x_k)}{\partial x}(x - x_k)$ being a good approximation to the function.

[a] Note that $\Delta x_k = x_{k+1} - x_k$.

Example 6.19 *Find a root of* $F : \mathbb{R}^4 \to \mathbb{R}^4$ *near* $x_0 = \begin{bmatrix} -2.0 & 3.0 & \pi & -1.0 \end{bmatrix}$ *for*

$$F(x) = \begin{bmatrix} x_1 + 2x_2 - x_1(x_1 + 4x_2) - x_2(4x_1 + 10x_2) + 3 \\ 3x_1 + 4x_2 - x_1(x_1 + 4x_2) - x_2(4x_1 + 10x_2) + 4 \\ 0.5\cos(x_1) + x_3 - (\sin(x_3))^7 \\ -2(x_2)^2\sin(x_1) + (x_4)^3 \end{bmatrix}.$$

Solution: We programmed up (6.3) and (6.4) in Julia and used a symmetric difference approximation for the derivatives, with $h = 0.1$. Below are the first five results from the algorithm:

$$x_k = \begin{bmatrix} k=0 & k=1 & k=2 & k=3 & k=4 & k=5 \\ -2.0000 & -3.0435 & -2.4233 & -2.2702 & -2.2596 & -2.2596 \\ 3.0000 & 2.5435 & 1.9233 & 1.7702 & 1.7596 & 1.7596 \\ 3.1416 & 0.6817 & 0.4104 & 0.3251 & 0.3181 & 0.3181 \\ -1.0000 & -1.8580 & -2.0710 & -1.7652 & -1.6884 & -1.6846 \end{bmatrix}$$

and

$$f(x_k) = \begin{bmatrix} k=0 & k=1 & k=2 & k=3 & k=4 & k=5 \\ -39.0000 & -6.9839 & -1.1539 & -0.0703 & -0.0003 & -0.0000 \\ -36.0000 & -6.9839 & -1.1539 & -0.0703 & -0.0003 & -0.0000 \\ 2.9335 & 0.1447 & 0.0323 & 0.0028 & 0.0000 & -0.0000 \\ 15.3674 & -5.1471 & -4.0134 & -0.7044 & -0.0321 & -0.0001 \end{bmatrix}.$$

By iteration five, we have a good approximation of a root because $||f(x_5)|| \approx 10^{-4}$. To emphasize that as x_k evolves, so does the

Jacobian of f at x_k, we provide the Jacobians at the initial and final steps,

$$\frac{\partial f(x_0)}{\partial x} = \begin{bmatrix} -19.0000 & -42.0000 & 0.0000 & 0.0000 \\ -17.0000 & -40.0000 & 0.0000 & 0.0000 \\ 0.4539 & 0.0000 & 1.0000 & 0.0000 \\ 7.4782 & 10.9116 & 0.0000 & 3.0100 \end{bmatrix} \text{ and } \frac{\partial f(x_5)}{\partial x} = \begin{bmatrix} -8.5577 & -15.1155 & 0.0000 & 0.0000 \\ -6.5577 & -13.1155 & 0.0000 & 0.0000 \\ 0.3854 & 0.0000 & 0.9910 & 0.0000 \\ 3.9296 & 5.4337 & 0.0000 & 8.5616 \end{bmatrix}.$$

■

Hopefully, you now have in your mind that iteration is useful. We formalize the process of doing iterations through sequences of vectors in normed spaces.

6.3 Sequences

Once again, we let $(\mathcal{X}, \| \bullet \|)$ be a normed space.

Definition 6.20 *A set of vectors indexed by the non-negative integers is called a* ***sequence****. Common notion includes* (x_n) *or* $\{x_n\}$.

Definition 6.21 *A sequence of vectors* (x_n) ***converges*** *to* $x \in \mathcal{X}$ *if,* $\forall\, \epsilon > 0,\ \exists N(\epsilon) < \infty$ *such that,* $n \geq N \implies \|x_n - x\| < \epsilon$*, i.e.,* $n \geq N \rightarrow\ x_n \in B_\epsilon(x)$*. One writes*

$$\lim_{n\to\infty} x_n = x \text{ or } x_n \to x \text{ or } x_n \xrightarrow[n\to\infty]{} x.$$

Proposition 6.22 *Suppose* $x_n \to x$*. Then,*

(a) $\|x_n\| \to \|x\|$

(b) $\sup_n \|x_n\| < \infty$ *(The sequence is bounded.)*

(c) If $x_n \to y$ *then y = x. (Limits are unique.)*

Remark 6.23 *(Handy Inequality) For* $\overline{x},\ \overline{y} \in \mathcal{X}$,

$$\begin{aligned} \|\overline{x}\| &= \|\overline{x} - \overline{y} + \overline{y}\| \\ &\leq \|\overline{x} - \overline{y}\| + \|\overline{y}\| \\ &\Downarrow \\ \|\overline{x}\| - \|\overline{y}\| &\leq \|\overline{x} - \overline{y}\| \end{aligned}$$

The same argument shows that $\|\overline{y}\| - \|\overline{x}\| \leq \|\overline{x} - \overline{y}\|$*. Hence*

$$\boxed{|\ \|\overline{x}\| - \|\overline{y}\|\ | \leq \|\overline{x} - \overline{y}\|.}$$

□

Proof of the Proposition:

1. From Remark 6.23, $|\ \|x\| - \|x_n\|\ | \leq \|x - x_n\| \xrightarrow[n\to\infty]{} 0$.
2. Applying the definition of convergence of a sequence, we set $\epsilon = 1$ and deduce $\exists N(1) < \infty$ such that $n \geq N \implies \|x_n - x\| \leq 1$. Hence, $\forall n \geq N,\ \|x_n\| = \|x_n - x + x\| \leq \|x_n - x\| + \|x\| \leq 1 + \|x\|$. It follows that
$$\sup_k \|x_k\| \leq \max\{\underbrace{\|x_1\|, \|x_2\|, \cdots, \|x_{N-1}\|, 1 + \|x\|}_{\text{finite}}\} < \infty.$$
3. $\|x - y\| = \|x - x_n + x_n - y\| \leq \|x - x_n\| + \|x_n - y\| \xrightarrow[n\to\infty]{} 0 \implies x = y$.

■

Definition 6.24 *Let* $x \in \mathcal{X}$, $P \subset \mathcal{X}$ *a subset.*

1. x *is a* ***limit point*** *of* P *if there exists a non-trivial sequence of elements of* P *that converges to* x*. That is,* $\exists(x_n)$*, such that for all* $n \geq 1$, $x_n \in P$, $x_n \neq x$ *and* $\lim_{n\to\infty} x_n = x$*. Here, non-trivial means that you cannot create the sequence* $(x_n = x)$.

2. *If $x \in P$ is not a limit point of P, then x is called an* ***isolated point of*** P.

Proposition 6.25 ***(Characterization of Isolated Points)*** *x is an isolated point of P if, and only if, there exists $\epsilon > 0$ such that $B_\epsilon(x) \cap P = \{x\}$.*

Proof: Suppose there exists $\epsilon > 0$ such that $B_\epsilon(x) \cap P = \{x\}$, and let (x_n) be such that for all $n \geq 1$, $x_n \in P$ and $x_n \neq x$. Then, for all $n \geq 1$, $d(x_n, x) \geq \epsilon$ and hence $x_n \not\to x$. For the other direction, we suppose that $\forall \epsilon > 0$, $B_\epsilon(x) \cap P \neq \{x\}$. Since $x \in B_\epsilon(x) \cap P$, we deduce that for all $\epsilon = 1/n$, there exists $x_n \neq x$ and $x_n \in B_\epsilon(x) \cap P$. Hence x satisfies all conditions of a limit point, namely $x_n \in P$, $x_n \neq x$, and $\lim_{n\to\infty} x_n = x$. ■

Remark 6.26 *Letx be an isolated point of P. Then x is an element of P and the trivial sequence $(x_n = x) \to x$. Moreover, if (x_n) is a sequence of elements in P and $\lim_{n\to\infty} x_n = x$, then there exists $N < \infty$ such that, for all $n \geq N$, $x_n = x$. In other words, the "tail" of the sequence is a trivial sequence.*

Proposition 6.27 ***(Characterization of Set Closure using Limit Points and Isolated Points)*** *Let P_{iso} be the collection of all isolated points of P and let P_∞ be the collection of all limit points of P. Then*

$$\overline{P} = P_{\text{iso}} \cup P_\infty.$$

Proof:

1. Suppose x is a limit point or an isolated point. Then, $\exists (x_n)$ such that $x_n \in P$ and $x_n \to x$. Because $x_n \to x$, $\forall \epsilon > 0$, $\exists x_n \in P$ such that $\|x_n - x\| < \epsilon$, which implies that $d(x,\ P) = 0$. Hence $x \in \overline{P}$.
2. Suppose $x \in \overline{P}$. Then, $d(x, P) = 0$ and hence, for all $n \geq 1$, $B_{\frac{1}{n}}(x) \cap P \neq \emptyset$. Two cases are possible. For some $N < \infty$, and all $N \geq N$, $B_{\frac{1}{n}}(x) \cap P = \{x\}$, in which case, $x \in P_{\text{iso}}$. Otherwise, for $n \geq 1$, there exists $x_n \in B_{\frac{1}{n}}(x) \cap P$ such that $x_n \neq x$, in which case, the sequence (x_n) establishes that $x \in P_\infty$.

Corollary 6.28 *P is closed if, and only if, it contains its limit points, that is,*

$$\overline{P} = P \iff P_\infty \subset P.$$

Proof: By definition, $P_{\text{iso}} \subset P$. Hence, if $P_\infty \subset P$, then $P_{\text{iso}} \cup P_\infty \subset P$, which implies by Proposition 6.27 that $\overline{P} \subset P$, and hence P is closed. For the other direction, if P is closed, then Proposition 6.27 implies that $P_\infty \subset P$, and hence the proof is done. ■.

6.4 Cauchy Sequences and Completeness

In practice, the definition of a convergent sequence is hard to apply because, to check it, you must have a "guess" of what the limit actually is! This led the mathematician Augustin-Louis Cauchy to propose a a related property, that now bears his name https://en.wikipedia.org/wiki/Cauchy_sequence.

Definition 6.29 *A sequence (x_n) is* ***Cauchy*** *if $\forall\, \epsilon > 0\ \exists N(\epsilon) < \infty$, such that $n \geq N$ and $m \geq N \implies \|x_n - x_m\| < \epsilon$.*

Notation 6.30 *We'll denote (x_n) is Cauchy by $\|x_n - x_m\| \xrightarrow[n,\, m\, \to\infty]{} 0$*

The following captures the idea of terms in a sequence getting closer and closer together. The analysis will prepare us for the Contraction Mapping Theorem.

Lemma 6.31 *Let $0 \leq c < 1$ and let (a_n) be a sequence of real numbers satisfying, $\forall\, n \geq 1$,*

$$|a_{n+1} - a_n| \leq c|a_n - a_{n-1}|.$$

Then (a_n) is Cauchy in $(\mathbb{R}, |\bullet|)$.

Proof:

Step 1: $\forall\, n \geq 1$, $|a_{n+1} - a_n| \leq c^n |a_1 - a_0|$.

Pf: First observe that $|a_3 - a_2| \leq c|a_2 - a_1| \leq c^2|a_1 - a_0|$ and then complete the proof by induction.

□

Step 2: $\forall\, n \geq 1, k \geq 1$, $|a_{n+k} - a_n| \leq \frac{c^n}{1-c}|a_1 - a_0|$.

Pf:

$$\begin{aligned}
|a_{n+k} - a_n| &\leq |a_{n+k} - a_{n+k-1} + a_{n+k-1} - a_{n+k-2} + \cdots a_{n+1} - a_n| \\
&\leq |a_{n+k} - a_{n+k-1}| + |a_{n+k-1} - a_{n+k-2}| + \cdots |a_{n+1} - a_n| \\
&\leq c^{n+k-1}|a_1 - a_0| + c^{n+k-2}|a_1 - a_0| + \cdots + c^n|a_1 - a_0| \\
&\leq c^n \left(\sum_{i=0}^{k-1} c^i \right) |a_1 - a_0| \\
&\leq c^n \left(\sum_{i=0}^{\infty} c^i \right) |a_1 - a_0| \\
&\leq c^n \left(\frac{1}{1-c} \right) |a_1 - a_0| \\
&\leq \frac{c^n}{1-c} |a_1 - a_0|.
\end{aligned}$$

□

Step 3: (a_n) is Cauchy.

Pf: Consider m and n and without loss of generality, suppose $m \geq n$. If $m = n$, then $|a_m - a_n| = 0$. Thus, we can assume $m = n + k$ for some $k \geq 1$. Then

$$|a_m - a_n| = |a_{n+k} - a_n| \leq \frac{c^n}{1-c}|a_1 - a_0| \xrightarrow[n\to\infty,\ m\geq n]{0},$$

and thus (a_n) is Cauchy.

□

■

Proposition 6.32 *If $x_n \to x$, then (x_n) is Cauchy.*

Proof: If $x_n \to x$, then $\forall\, \epsilon > 0\ \exists\, N < \infty$ such that $\forall\, n \geq N$, $\|x_n - x\| < \frac{\epsilon}{2}$. Hence, if $m \geq N$

$$\begin{aligned}
\|x_n - x_m\| &= \|x_n - x + x - x_m\| \\
&\leq \|x_n - x\| + \|x - x_m\| \\
&< \frac{\epsilon}{2} + \frac{\epsilon}{2} \\
&< \epsilon \quad \text{for all } n,\ m \geq N
\end{aligned}$$

■

Unfortunately, not all Cauchy sequences are convergent. For a reason we will understand shortly, all counter examples are infinite dimensional.

Example 6.33 *Let $\mathcal{X} := \{f : [0,1] \to \mathbb{R} \mid \text{f is continuous}\}$ and equip it with the one-norm, $\|f\|_1 := \int_0^1 |f(\tau)|d\tau$. Define a sequence as follow, where each function piecewise is linear and $n \geq 2$,*

$$f_n(t) = \begin{cases} 0 & 0 \leq t \leq \frac{1}{2} - \frac{1}{n} \\ 1 + n(t - \frac{1}{2}) & \frac{1}{2} - \frac{1}{n} \leq t \leq \frac{1}{2} \\ 1 & t \geq \frac{1}{2}. \end{cases}$$

Show that the sequence is Cauchy and does not have a limit in $\mathcal{X}$.

Proof: You may want to check that each function is continuous at the breakpoints, $\frac{1}{2} - \frac{1}{n}$ and $\frac{1}{2}$. Moreover, by using the area under a triangle, you can show $\|f_n - f_m\|_1 = \frac{1}{2}\,|\frac{1}{n} - \frac{1}{m}| \xrightarrow[n,\ m\to\infty]{} 0$, and thus the sequence is Cauchy.

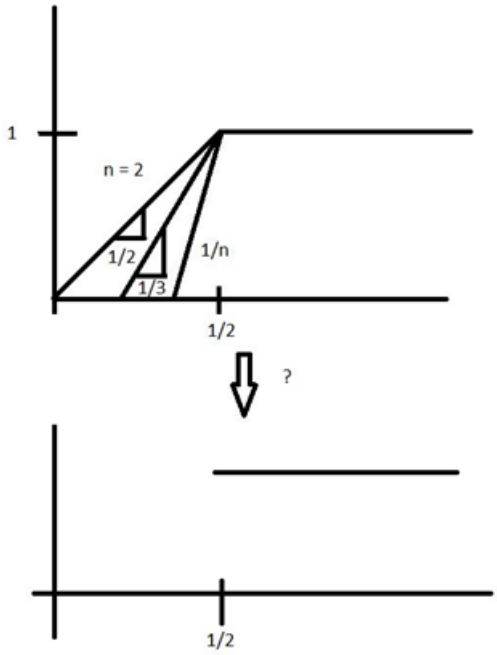


Figure 6.3: Visually, the sequence of continuous functions $(f_n(t))_{n=2}^{\infty}$ appears to be converging to a step function, which is discontinuous. If that is true, we then have a Cauchy sequence in $(\mathcal{X}, \| \bullet \|_1)$ that does not have a limit in the set $\mathcal{X}$.

How can we show that there does not exist any (continuous) function $f \in \mathcal{X}$ to which the sequence converges? We define

$$f_{\text{step}}(t) := \begin{cases} 0 & 0 \leq t < \frac{1}{2} \\ 1 & \frac{1}{2} \leq t \leq 1 \end{cases}$$

and note that f_{step} is discontinuous and hence $f_{\text{step}} \notin \mathcal{X}$. Define $\mathcal{Y} := \text{span}\{\mathcal{X}, f_{\text{step}}\} \subset \{f : [0,1] : \mathbb{R}\}$, the vector space of all functions from the interval $[0,1]$ to the real numbers. You can check that $\| \bullet \|_1$ is also a norm on $\mathcal{Y}$ and observe that $f \in \mathcal{Y}$ is continuous if, and only if, $f \in \mathcal{X}$. Finally, you can easily compute that

$$\|f_n - f_{\text{step}}\|_1 = \frac{1}{2n},$$

and hence, $f_n \to f_{\text{step}}$. By uniqueness of limits, there does not exist a continuous function in $\mathcal{Y}$ to which the sequence converges, and thus the sequence does not have a limit in $\mathcal{X}$.

■

The study of Cauchy sequences led mathematicians to wonder if it is possible to find normed spaces where all Cauchy sequences do have limits (within the given normed space), and moreover, if a normed space was "deficient" in the sense that it had Cauchy sequences without limit, could it be "enlarged" or "completed" to one where all Cauchy sequences do have limits. These are great questions!

Definition 6.34 *A normed space $(\mathcal{X}, \mathbb{R}, \| \cdot \|)$ is **complete** if every Cauchy Sequence in X has a limit in X. Such spaces are also called **Banach spaces**.*

The above definition can only be useful if a list of useful Banach spaces is known; see https://en.wikipedia.org/wiki/Banach_space#Examples_2. In EECS562, you will use $(C[0,\ T],\ \| \bullet \|_\infty)$, the set of continuous functions with the infinity (or max) norm. The sequence in Fig. 6.3 is not a Cauchy sequence in $(C[0,\ T],\ \| \bullet \|_\infty)$; you might want to check that.

Fact 6.35 *For $a < b$, both finite, $(C[a,\ b],\ \| \bullet \|_\infty)$ is complete where $C[a,\ b] = \{f : [a,\ b] \to \mathbb{R} \mid f$ continuous$\}$. We showed above that $(C[a,b], || \bullet ||_1)$ is not complete.*

Definition 6.36 *A subset S of a normed space is **complete** if every Cauchy Sequence in S has a limit in S.*

Remark 6.37 *S is complete implies that S is closed.*

Theorem 6.38 *Let $(\mathcal{X}, ||\bullet||)$ be a normed space. Then,*

(a) Every finite dimensional subspace is complete.

(b) Any closed subset of a complete set is also complete.

Fact 6.39 *Every normed space $(\mathcal{X}, ||\bullet||_X)$ has a **"completion"**. A bit loosely stated, this means there is a complete normed space $(\mathcal{Y}, ||\bullet||_Y)$ such that*

(a) $\mathcal{X} \subset \mathcal{Y}$ ($\mathcal{X}$ can naturally be viewed as a subset of $\mathcal{Y}$. The precise definition involves isometric isomorphisms.)

(b) $\forall x \in \mathcal{X}$, $||x||_Y = ||x||_X$.

(c) $\overline{\mathcal{X}} = \mathcal{Y}$ ($\overline{\mathcal{X}}$ is the closure of $\mathcal{X}$ in $\mathcal{Y}$. Hence, $\mathcal{X}$ fits "tightly" into $\mathcal{Y}$ in that sense that for any point in $y \in \mathcal{Y}$, $d(y, \mathcal{X}) = 0$).

(d) $\mathcal{Y}$ = $\mathcal{X} \cup \{$limit points of Cauchy sequences in $\mathcal{X}\}$

You might ask about the completion of $C[a, b]$ when the $||\bullet||_1$ is used? It turns out to be the set of Lebesgue integrable functions on $[0, 1]$. We alluded to Lebesgue back in Chapter 5.1.3.

6.5 Contraction Mapping Theorem

Definition 6.40 *Let $S \subset \mathcal{X}$ be a subset of a normed space $(\mathcal{X}, \|\bullet\|)$. A function $T : S \to S$ is a **contraction mapping** if, $\exists\, 0 \leq c < 1$ such that $\forall x, y \in S$,*

$$\|T(x) - T(y)\| \leq c\|x - y\|.$$

*A point $x^* \in S$ is a **fixed point** of T if $T(x^*) = x^*$.*

Theorem 6.41 ***(Contraction Mapping Theorem)*** *If T is a contraction mapping on a complete subset S of a normed space $(\mathcal{X}, \mathbb{R}, \|\bullet\|)$, then there exists a unique vector $x^* \in S$ such that $T(x^*) = x^*$. Moreover, for every initial point $x_0 \in S$, the sequence $x_{n+1} = T(x_n), n \geq 0$, is Cauchy, and $x_n \to x^*$.*

Proof: Let (x_n) be defined as in the statement of the theorem. Because T is a contraction mapping, there exists $0 \leq c < 1$ such that, for all $n \geq 1$,

$$\begin{aligned}\|x_{n+1} - x_n\| &= \|T(x_n) - T(x_{n-1})\| \\ &\leq c\|x_n - x_{n-1}\|\end{aligned}$$

Claim 6.42 *(x_n) is Cauchy and thus by the completeness of S, $\exists x^* \in S$ such that $x_n \to x^*$.*

Proof: We leave as an exercise to show by induction that $\|x_{n+1} - x_n\| \leq c^n\|x_1 - x_0\|$. Next, consider $\|x_m - x_n\|$, and without loss of generality, suppose $m = n + k,\ k > 0$. Then,

$$\begin{aligned}\|x_m - x_n\| &= \|x_{n+k} - x_n\| \\ &= \|x_{n+k} - x_{n+k-1} + x_{n+k-1} - \cdots + x_{n+1} - x_n\| \\ &\leq \|x_{n+k} - x_{n+k-1}\| + \cdots + \|x_{n+1} - x_n\| \\ &\leq \left(c^{n+k-1} + c^{n+k-2} + \cdots + c^n\right)\|x_1 - x_0\| \\ &= c^n\left(\sum_{i=0}^{k-1} c^i\right)\|x_1 - x_0\| \\ &\leq c^n\left(\sum_{i=0}^{\infty} c^i\right)\|x_1 - x_0\| \\ &= \frac{c^n}{1-c}\|x_1 - x_0\| \xrightarrow[\substack{n\to\infty \\ m>n}]{} 0\end{aligned}$$

where we used a simple fact about the geometric series for $\frac{1}{1-c}$. Therefore, (x_n) is Cauchy sequence in S, and by completeness, $\exists x^* \in S$ such that $x_n \to x^*$. □

Claim 6.43 *$x^* = T(x^*)$ and thus x^* is a fixed point of T.*

Proof: Let $n \geq 1$ be arbitary. Then,

$$\begin{aligned}\|x^* - T(x^*)\| &= \|x^* - x_n + x_n - T(x^*)\| \\ &= \|x^* - x_n + T(x_{n-1}) - T(x^*)\| \\ &\leq \|x^* - x_n\| + \|T(x_{n-1}) - T(x^*)\| \\ &\leq \|x^* - x_n\| + c\|x_{n-1} - x^*\| \xrightarrow[n\to\infty]{} 0.\end{aligned}$$

□

Claim 6.44 *x^* is unique.*

Proof: Suppose $y^* = T(y^*)$. Then,

$$\begin{aligned}\|x^* - y^*\| &= \|T(x^*) - T(y^*)\| \\ &\leq c\|x^* - y^*\|.\end{aligned}$$

The only non-negative real number γ that satisfies $\gamma \leq \gamma c$ for some $0 \leq c < 1$ is $\gamma = 0$. Hence, due to the property of norms, $0 = \|x^* - y^*\| \implies x^* = y^*$.

□
■

Remark 6.45 *The (local) convergence of the Newton-Raphson Algorithm is accomplished by identifying a closed ball in $\mathbb{R}^n$ on which the function*

$$\boxed{T(x) := x - \epsilon \left(\frac{\partial f}{\partial x}(x)\right)^{-1} (f(x) - y)}$$

is a contraction mapping. The estimate of a suitable value $0 \leq c < 1$ is based on a Lipschitz constant for the Jacobian. We check that a solution of $f(x) - y$ is a fixed point of $T(x)$. Indeed,

$$\begin{aligned} x^* &= T(x^*) \\ &\Updownarrow \\ x^* &= x^* - \epsilon \left(\frac{\partial f}{\partial x}(x^*)\right)^{-1} (f(x^*) - y) \\ &\Updownarrow \\ 0 &= -\epsilon \left(\frac{\partial f}{\partial x}(x^*)\right)^{-1} (f(x^*) - y) \\ &\Updownarrow \\ 0 &= (f(x^*) - y). \end{aligned}$$

6.6 Continuous Functions

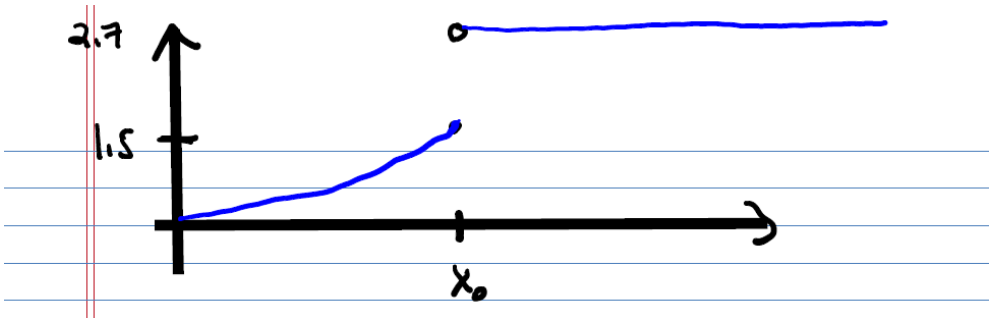


Figure 6.4: Let $\epsilon = 1.0$ Then, $\forall\ \delta > 0, \exists\ x \in B_\delta(x_0)$ such that $|f(x) - f(x_0)| \geq \epsilon$. Indeed, $x = x_0 + \delta_2$ works.

Definition 6.46 *Let $(\mathcal{X}, \|\bullet\|)$, and $(\mathcal{Y}, |||\bullet|||)$ be normed spaces.*

*(a) $f : \mathcal{X} \to \mathcal{Y}$ is **continuous at** $x_0 \in \mathcal{X}$ if $\forall \epsilon > 0,\ \exists \delta(\epsilon, x_0) > 0$ such that $\|x - x_0\| < \delta \implies |||f(x)||| < \epsilon$.*

*(b) f is **continuous** if it is continuous at x_0 for all $x_0 \in \mathcal{X}$.*

Remark 6.47 *It is also common to define continuity at a point as as $\forall\ \epsilon > 0,\ \exists \delta > 0$ such that $x \in B_\delta(x_0) \implies f(x) \in B_\epsilon(f(x_0))$. Though less common, it can also be defined as $\forall\ \epsilon > 0,\ \exists \delta > 0$ such that $f(B_\delta(x_0)) \subset B_\epsilon(f(x_0))$.*

Remark 6.48 ***(Discontinuous at a point)*** *Let's negate the definition of f continuous at x_0. f is **discontinuous** at $x_0 \in \mathcal{X}$ if $\exists\ \epsilon > 0$ such that, $\forall\ \delta > 0, \exists\ x \in \mathcal{X}$ such that $||x - x_0|| < \delta$ and $|||f(x) - f(x_0)||| \geq \epsilon$. This can also be stated as $\exists\ \epsilon > 0$ such that, $\forall\ \delta > 0$, $\exists\ x \in B_\delta(x_0)$ such that $f(x) \notin B_\epsilon(f(x_0))$. Finally, it can also be stated as $\exists\ \epsilon > 0$, $\forall \delta > 0$ such that $f(B_\delta(x_0)) \not\subset B_\epsilon(f(x_0))$.*

Theorem 6.49 ***(Characterization of Continuity at a Point via Sequences)*** *Let Let $(\mathcal{X}, \|\bullet\|)$, and $(\mathcal{Y}, |||\bullet|||)$ be normed spaces and $f : \mathcal{X} \to \mathcal{Y}$ a function.*

(a) If f is continuous at x_0 and the sequence (x_n) is a sequence in $\mathcal{X}$ that converges to x_0, then the sequence $(y_n := f(x_n))$ in $\mathcal{Y}$ converges to $f(x_0)$. [$y_n := f(x_n)$, $y_0 := f(x_0)$ implies $y_n \to y_0$ when f is continuous at x_0.]

(b) If f is discontinuous at x_0, then there exists a sequence (x_n) such that $x_n \to x_0$, and $f(x_n) \not\to f(x_0)$, that is, $f(x_n)$ does not converge to $f(x_0)$.

The proof is done in HW 10. The main point is, just as sequences can be used to completely characterize closed sets, they can also be used to completely characterize continuity at a point.

Corollary 6.50 *$f : \mathcal{X} \to \mathcal{Y}$ is continuous at x_0 if, and only if, every convergent sequence in $\mathcal{X}$ with limit x_0 is mapped by f to a convergent sequence in $\mathcal{Y}$ with limit $f(x_0)$. In other symbols, (f is continuous at x_0) $\iff (x_n \to x_0 \implies f(x_n) \to f(x_0))$.*

6.7 Compact Sets and the Existence of Extrema of Functions

Definition 6.51 *Let (x_n) be a sequence and $1 \leq n_1 < n_2 < n_3 < \cdots$ be an infinite set of strictly increasing integers. Then, (x_{n_i}) is called a **subsequence** of (x_n). We note in passing that $n_i \geq i$, $\forall\ i \geq 1$.*

Example 6.52 $n_i = 2i + 1$ or $n_i = 2^i$.

Lemma 6.53 *Suppose $x_n \to x$. Then every subsequence (x_{n_i}) of (x_n) converges to x.*

We leave the proof as an exercise.

Definition 6.54 *A set S is **bounded** if $\exists\ r < \infty$ such that $S \subset B_r(0)$.*

Exercise 6.55 *Show the following for $S \subset \mathcal{X}$:*

1. *S is bounded if, and only if, $\sup_{x \in S} \|x\| < \infty$.*
2. *Hence, S is unbounded if, and only if, $\sup_{x \in S} \|x\| = \infty$.*
3. *S is unbounded if, and only if, there exists a sequence (x_k) such that, for all $k \geq 1$, $x_k \in S$ and $||x_{k+1}|| \geq ||x_k|| + 1$.*

Lemma 6.56 *If S is unbounded, then it contains a sequence with no convergent subsequence.*

Proof: The sequence (x_n) constructed above in Exercise 6.55 has no convergent subsequence. Indeed, by Remark 6.23, if (x_{n_i}) is a subsequence of (x_n), then $\|x_{n_i} - x_{n_j}\| \geq |\ \|x_{n_i}\| - \|x_{n_j}\|\ | \geq |n_i - n_j|$, and thus is not Cauchy. Because it is not Cauchy, it cannot be convergent.

■

Definition 6.57 ***(Equivalent Norms)*** *Let $(\mathcal{X}, \mathbb{R})$ be a vector space. Two norms $||\cdot|| : \mathcal{X} \to [0, \infty)$ and $|||\cdot||| : \mathcal{X} \to [0, \infty)$ are **equivalent** if there exist positive constants K_1 and K_2 such that, for all $x \in \mathcal{X}$,*

$$K_1|||x||| \leq ||x|| \leq K_2|||x|||.$$

Remark 6.58 *It follows from the definition of equivalent norms that $\frac{1}{K_2}||x|| \leq |||x||| \leq \frac{1}{K_1}||x||$.*

-

We'll next develop a few bounds for norms on finite dimensional vector spaces, and use the bounds to relate convergence of a sequence of vectors in a finite dimensional normed space $(\mathcal{X}, ||\bullet||)$ to the convergence of the representation of the sequence with respect to a basis. Let $\{v\} := \{v^1, v^2, \ldots, v^n\}$ be a basis for $\mathcal{X}$ and define $M_i := \text{span}\{v^j \,|\, j \neq i\}$, the $(n-1)$-dimensional subspace spanned by all the basis vectors except the i-th one. Because M_i is finite dimensional, it is a complete and hence a closed subset of $\mathcal{X}$. By construction, $v^i \notin M_i$, and thus,

$$\delta_i := d(v^i, M_i) > 0.$$

Lemma 6.59 *Let $\{v\} = \{v^1, v^2, \ldots, v^n\}$, $\{M_1, M_2, \ldots, M_n\}$ and $\{\delta_1, \delta_2, \ldots, \delta_n\}$ be as above. Then for any vector $x = \alpha_1 v^1 + \alpha_2 v^2 + \cdots + \alpha_n v^n \in \mathcal{X}$,*

$$\kappa_* \left(\max_{1 \leq i \leq n} |\alpha_i| \right) \leq ||x|| \leq \kappa^* \left(\sum_{i=1}^{n} |\alpha_i| \right) \leq n\kappa^* \left(\max_{1 \leq i \leq n} |\alpha_i| \right), \tag{6.6}$$

where $\kappa_* := \min_{1 \leq i \leq n}\{\delta_i\} > 0$, $\kappa^* := \max_{1 \leq i \leq n}\{||v^i||\} < \infty$, and $\alpha := [x]_{\{v\}}$.

Proof: The right hand side of (6.6) follows from the triangle inequality for norms. The proof of the left hand side of (6.6) requires a few steps that we carefully enumerate and leave to the reader:

1. If $\alpha_i = 0$, then $d(\alpha_i v^i, M_i) = 0$.
2. If $\alpha_i \neq 0$, then $d(\alpha_i v^i, M_i) = |\alpha_i| d(v^i, M_i) = |\alpha_i| \delta_i$.
3. Hence, for all $\alpha_i \in \mathbb{R}$, $d(\alpha_i v^i, M_i) = |\alpha_i| \delta_i$
4. For arbitrary $m_i \in M_i$, $d(\alpha_i v^i + m_i, M_i) = d(\alpha_i v^i - m_i, M_i) = d(\alpha_i v^i, M_i)$.
5. For any vector $x = \alpha_1 v^1 + \alpha_2 v^2 + \cdots + \alpha_n v^n \in \mathcal{X}$, $d(x, M_i) = d(\alpha_i v^i, M_i)$.
6. Hence for any vector $x = \alpha_1 v^1 + \alpha_2 v^2 + \cdots + \alpha_n v^n \in \mathcal{X}$, and for all $1 \leq i \leq n$,

$$||x|| \geq \inf_{m \in M_i} ||x - m|| =: d(x, M_i) = |\alpha_i| \delta_i.$$

■

Corollary 6.60 ***(Equivalent Norms)*** *All norms on finite dimensional vector spaces are equivalent*[1].

Proof: Equation (6.6) shows that, once a basis is chosen, any norm on an n-dimensional normed space is equivalent to $(\mathbb{R}^n, ||\bullet||_{\max})$. We leave it to the reader to show that this is enough to prove the result. ■

Corollary 6.61 ***(Convergence of Components of Sequences in Finite-dimensional Spaces)*** *Let (x_k) be a sequence in a finite dimensional normed space $(\mathcal{X}, ||\bullet||)$ with basis $\{v^1, v^2, \ldots, v^n\}$. Let*

$$\alpha_k := [x_k]_{\{v\}} \in \mathbb{R}^n$$

be the representation of x_k with respect to the basis $\{v\}$ so that $x_k = \alpha_{k,1} v^1 + \alpha_{k,2} v^2 + \cdots + \alpha_{k,n} v^n$. Then (x_k) is a Cauchy sequence in $\mathcal{X}$ if, and only if, each sequence $(\alpha_{k,i})$ is a Cauchy sequence in $(\mathbb{R}, |\bullet|)$, $1 \leq i \leq n$.

Proof: Equation (6.6) reduces the proof to understanding that a sequence in $\mathbb{R}^n$ is Cauchy if, and only if, each of its real components is Cauchy. But with the max-norm, $(\mathbb{R}^n, ||\bullet||_{\max})$, this is quite easy to show because it selects the largest element, and if the largest element is getting small, then the rest of them are too. Have fun with it! ■

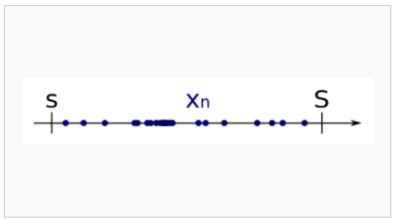
Theorem 6.62 ***(Bolzano-Weierstrass Theorem or the Sequential Compactness Theorem)*** *In a finite dimensional normed space $(\mathcal{X}, ||\bullet||)$, the following two properties are equivalent for a set $C \subset \mathcal{X}$.*

(a) C is closed and bounded;

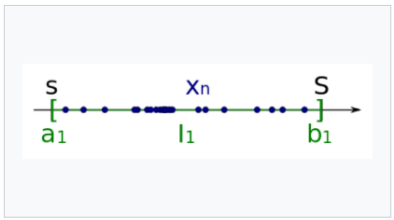
(b) Every sequence in C contains a convergent subsequence, that is, for every sequence (x_n) in C (i.e. $x_n \in C, \forall\, n \geq 1$), there exists $x_0 \in C$ and a subsequence (x_{n_i}) of (x_n) such that $x_{n_i} \to x_0$.

[1] It is an interesting fact that a vector space is finite dimensional if, and only if, all norms on it are equivalent.

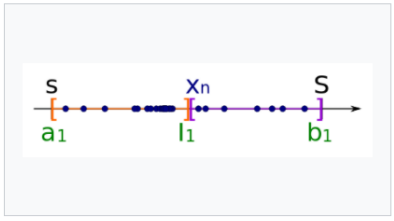
There is also an alternative proof of the Bolzano–Weierstrass theorem using nested intervals. We start with a bounded sequence (x_n):



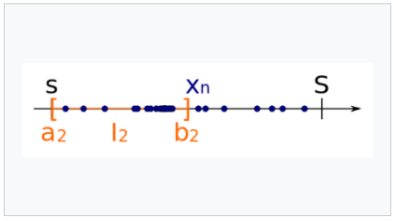
Because $(x_n)_{n\in\mathbb{N}}$ is bounded, this sequence has a lower bound s and an upper bound S.



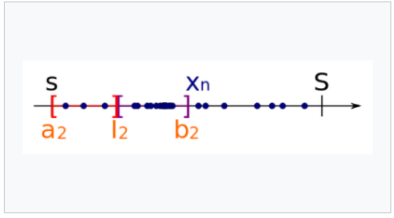
We take $I_1 = [s, S]$ as the first interval for the sequence of nested intervals.



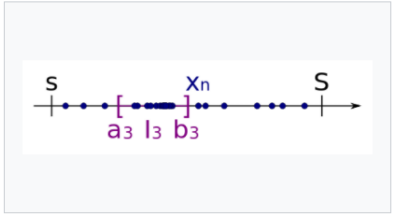
Then we split I_1 at the mid into two equally sized subintervals.



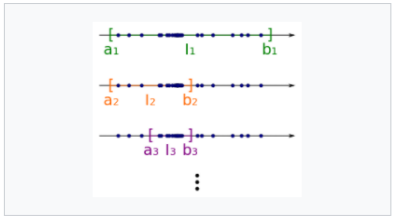
Because each sequence has infinitely many members, there must be (at least) one of these subintervals that contains infinitely many members of $(x_n)_{n\in\mathbb{N}}$. We take this subinterval as the second interval I_2 of the sequence of nested intervals.



Then we split I_2 again at the mid into two equally sized subintervals.



Again, one of these subintervals contains infinitely many members of $(x_n)_{n\in\mathbb{N}}$. We take this subinterval as the third subinterval I_3 of the sequence of nested intervals.



We continue this process infinitely many times. Thus we get a sequence of nested intervals.

Because we halve the length of an interval at each step, the limit of the interval's length is zero. Also, by the *nested intervals theorem*, which states that if each I_n is a closed and bounded interval, say

$I_n = [a_n, b_n]$

with

$a_n \leq b_n$

then under the assumption of nesting, the intersection of the I_n is not empty. Thus there is a number x that is in each interval I_n. Now we show, that x is an accumulation point of (x_n).

Take a neighbourhood U of x. Because the length of the intervals converges to zero, there is an interval I_N that is a subset of U. Because I_N contains by construction infinitely many members of (x_n) and $I_N \subseteq U$, also U contains infinitely many members of (x_n). This proves that x is an accumulation point of (x_n). Thus, there is a subsequence of (x_n) that converges to x.

Figure 6.5: Illustration of a sequntial comactness proof from Wikipedia https://en.wikipedia.org/wiki/Bolzano-Weierstrass_theorem.

Proof: We first show $\sim (a) \implies \sim (b)$. There are two cases, C is not closed or C is not bounded.

Suppose that C is not closed. Then by Corollary 6.28, there exists a limit point $x_0 \in \overline{C}$ such that $x_0 \notin C$. Hence, there exists a sequence (x_n) with $x_n \in C$ and $x_n \to x_0 \notin C$. By Lemma 6.53, all subsequences (x_{n_i}) of (x_n) satisfy $x_{n_i} \to x_0$. Hence, we have constructed a sequence of elements of C for which there is no subsequence with a limit in C.

Suppose next that C is unbounded. Then Exercise 6.55 produces a sequence of elements of C for which every subsequence is not Cauchy, and hence cannot have a limit in C. This completes the proof of $\sim (a) \implies \sim (b)$. Nothing we have done so far depends on $\mathcal{X}$ being finite dimensional.

We now turn to $(a) \implies (b)$. Let (x_n) be an arbitrary sequence built from elements of C. To show: it has a convergent subsequence with limit in C.

Case 1: (x_n) has only a finite number of distinct elements. Hence, at least one value is repeated an infinite number of times; let's call it $x_N \in C$. Because it is repeated an infinite number of times, we can choose a strictly increasing sequence $n_1 < n_2 < \cdots < n_i < \cdots$ such that $x_{n_i} = x_N$ for all $i \geq 1$. The subsequence (x_{n_i}) converges to $x_N \in C$ and hence we are done.

Case 2: (x_n) has an infinite number of distinct elements. Here, we will invoke that $\mathcal{X}$ is finite dimensional. Indeed, by Corollary 6.61, a subsequence of (x_n) will be convergent if, and only if, once it it is represented with respect to a basis, each of its components is convergent. Hence, it is enough to prove that every real sequence (a_n), with an infinite number of distinct elements, contained in a closed and bounded subset of $C_1 \subset \mathbb{R}$ has a convergent subsequence. Every bounded subset of $\mathbb{R}$ is contained within a closed interval of the form $[-N, N]$, for some integer $1 \geq N < \infty$, and hence we reduce ourselves to a set $C_1 \subset [-N, N]$ and C_1 contains an infinite number of (distinct) elements of the sequence (a_n). Because C_1 contains an infinite number of distinct elements of (a_n), some closed interval of the form $[n, n+1]$, for $|n| \leq N$ must contain an infinite number of elements of (a_n). Without loss of generality, we will assume the closed interval $[0, 1]$ contains an infinite number of elements of (a_n); the reader will see that our argument works *mutatis mutandis* for any other interval of length one.

Divide the interval $[0, 1]$ into two halves, $[0, 1] = [0, 1/2] \cup [1/2, 1]$. At least one of the halves must contain an infinite number of elements of (a_n). For the sake of argument, assume it is the right half. Now we divide in half $[1/2, 1] = [1/2, 3/4] \cup [3/4, 1]$ and note that at least one of the halves must contain an infinite number of elements of the sequence (a_n). For the sake of argument, assume the left half this time. We then divide in half $[1/2, 3/4] = [1/2, 5/8] \cup [5/8, 3/4]$ and note that at least one of the halves must contain an infinite number of elements of the sequence (a_n).

At the k-th step, we have a closed sub-interval of $I_k \subset [0, 1]$, with length I_k equalling $1/2^k$ and I_k contains an infinite number of distinct elements of (a_k). Hence, for all $k \geq 1$, there exists n_k such that $n_k > n_{k-1}$ and $a_{n_k} \in I_k$. Moreover, the sequence (a_{n_k}) is Cauchy because for all $i \geq k, j \geq k$, $|a_{n_i} - a_{n_j}| \leq 1/2^k$. Hence the subsequence (a_{n_k}) converges to a limit in C_1 and the proof is done.

■

Definition 6.63 *A set C satisfying (a) or (b) of Theorem 6.62 is said to be* ***compact****.*

.

Remark 6.64 *There are various definitions of compactness that are appropriate for specific settings. The one above is typically called sequential compactness. Because it is the only form of compactness we use in these notes, we will drop the term sequential and simply call such sets compact sets.*

Theorem 6.65 ***(Weierstrass Theorem)*** *If C is compact and $f : C \to \mathbb{R}$ is continuous, then f achieves its extreme values. That is,*

$$\exists x^* \in C, \text{ s.t. } f(x^*) = \sup_{x \in C} f(x)$$

and

$$\exists x_* \in C, \text{ s.t. } f(x_*) = \inf_{x \in C} f(x).$$

Proof: Let $f^* := \sup_{x \in C} f(x)$. To show $\exists x^* \in C$, such that $f(x^*) = f^*$.

Claim 6.66 *f^* is finite.*

Proof: We do a proof by contradiction and suppose not, that is, suppose $f^* = \infty$. Then, by definition of the supremum, for all $n \geq 1$, there exists $x_n \in C$ such that $f(x_n) \geq n$. Because C is compact, there exists $x_0 \in C$ and a subsequence (x_{n_i}) of (x_n) such that $x_{n_i} \to x_0$. Because f is assumed continuous,

$$f(x_0) = \lim_{i \to \infty} f(x_{n_i}) \geq \lim_{i \to \infty} n_i = \infty.$$

But because $f : C \to \mathbb{R}$, $f(x_0) \in \mathbb{R}$, which contradicts $f(x_0) = \infty$. Hence, it cannot be the case that $f^* = \infty$.

□

Continuing with the proof, we now know that f^* is finite. Once again, applying the definition of the supremum, we have that $\forall n > 0, \exists\, x_n \in C$ such that $|f^* - f(x_n)| < 1/n$. Because C is compact, there exists $x^* \in C$ and a subsequence (x_{n_i}) such that $x_{n_i} \to x^*$. Because f continuous, $f(x_{n_i}) \to f(x^*)$. We expect to show that $f(x^*) = f^*$. To do so, we note that, by the continuity[2] of f,

$$|f^* - f(x^*)| = \lim_{i \to \infty} |f^* - f(x_{n_i})| \leq \lim_{i \to \infty} \frac{1}{n_i} = 0.$$

And hence, $f^* = f(x^*)$.

The same proof works for the infimum.

■

[2]Technically, we are using the continuity of f implies that of $g(x) := |f^* - f(x)|$, which you are welcome to check. Otherwise, you can do the estimate as $|f^* - f(x^*)| = |f^* - f(x_{n_i}) + f(x_{n_i}) - f(x^*)| \leq |f^* - f(x_{n_i})| + |f(x_{n_i}) - f(x^*)| \leq \frac{1}{n_i} + |f(x_{n_i}) - f(x^*)| \xrightarrow[i \to \infty]{} 0$.

Chapter 7

Briefest of Remarks on Optimization

Learning Objectives

- Learn about "convexity", which defines a class of problems where local minima are also global minima
- Learn about two types of convex optimization problems that are fast enough to run in real time on many mobile platforms.

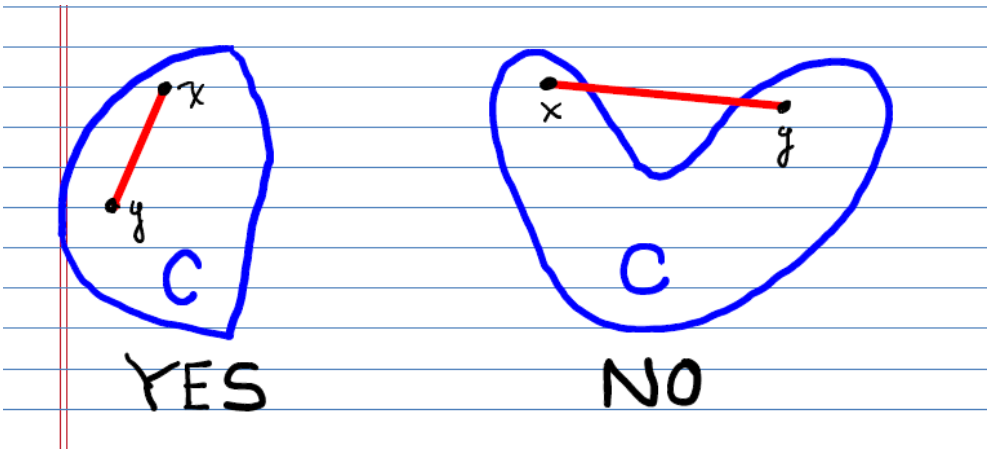
Outcomes

- Definition of convex sets and convex functions.
- Local minima of convex functions are also global minima.
- A Quadratic Program (QP) is an optimization problem with a quadratic cost and linear inequality and equality constraints.
- A Linear Program (LP) is an optimization problem with a linear cost and linear inequality and equality constraints.
- Through the concept of "slack variables", an optimization problem with either a one-norm or ∞-norm as the the cost and linear inequality and equality constraints can be turned into an LP.

7.1 Brief Remarks on Convex Sets and Convex Functions

Definition 7.1 *Let $(\mathcal{X}, \mathbb{R})$ be a real vector space. $C \subset V$ is **convex** if $\forall x, y \in C$ and $0 \leq \lambda \leq 1$, the **convex combination** $\lambda x + (1-\lambda)y \in C$.*

Remark 7.2



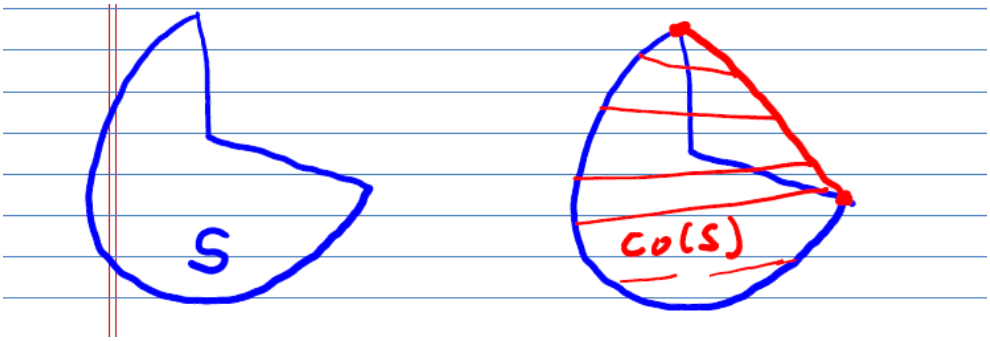
(a) For C to be convex, given any two points $x, y \in C$, then the line connecting x and y must also lie in C.

(b) Open and closed balls arsing from norms are always convex.

Definition 7.3 *The **convex hull** of a set $S \subset \mathcal{X}$ is*

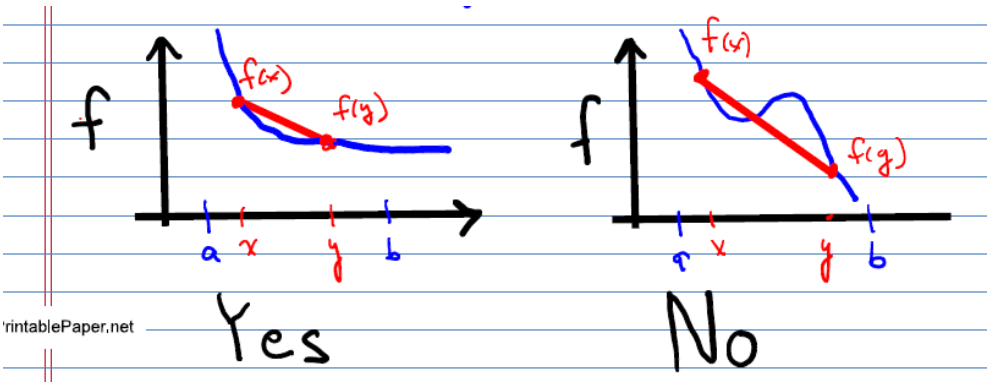
$$\text{co}(S) := \{\lambda x + (1-\lambda)y \mid 0 \leq \lambda \leq 1, x, y \in S\},$$

the set of all convex combinations of elements of S. It can also be defined as the smallest convex set that contains S.



Definition 7.4 *Suppose $C \subset \mathcal{X}$ is convex. A function $f : C \to \mathbb{R}$ is **convex** if $\forall x, y \in C, 0 \leq \lambda \leq 1$,*

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y).$$



Remark 7.5 *For a function to be convex, the line $\lambda x + (1-\lambda)y$ connecting x and y must lie at or above the graph of the function; it can never go below.*

Definition 7.6 *Suppose $(\mathcal{X}, \mathbb{R}, ||\bullet||)$ is a normed space, $D \subset \mathcal{X}$ a subset, and $f : D \to \mathbb{R}$ a function.*

*(a) $x^* \in D$ is a **local minimum** of f if $\exists \delta > 0$ such that $\forall x \in B_\delta(x^*)$, $f(x^*) \leq f(x)$.*

*(b) $x^* \in D$ is a **global minimum** if $\forall y \in D$, $f(x^*) \leq f(y)$.*

Theorem 7.7 ***(Local equals Global for Convex Functions)*** *If D and f are both convex, then any local minimum is also a global minimum.*

Proof: We show that if x is not a global minimum, then it cannot be a local minimum. Specifically, we prove the contrapositive statement: $((a) \implies (b)) \iff (\neg(b) \implies \neg(a))$, where

(a) $x \in D$ is a local minimum

(b) $x \in D$ is a global minimum.

$\neg$(b) $\exists y \in D$ such that $f(y) < f(x)$.

$\neg$(a) $\forall\, \delta > 0, \exists\, z \in B_\delta(x) \cap D$ such that $f(z) < f(x)$.

Claim 7.8 *If $f(y) < f(x)$, then $\forall 0 < \lambda \leq 1$, the vector $z := (1-\lambda)x + \lambda y$ satisfies $f(z) < f(x)$.*

Proof:

$$\begin{aligned} f(z) &= f((1-\lambda)x + \lambda y) \\ &\leq (1-\lambda)f(x) + \lambda f(y) \ \text{[convexity]} \\ &< (1-\lambda)f(x) + \lambda f(x) \ [f(x) > f(y)] \\ &= f(x). \end{aligned}$$

Hence, $f(z) < f(x)$. □

Claim 7.9 *$\forall \delta > 0$, $\exists 0 < \lambda < 1$ such that $z := (1-\lambda)x + \lambda y \in B_\delta(x) \cap D$.*

Proof:

$$\begin{aligned} ||z - x|| &= ||(1-\lambda)x + \lambda y - x|| \\ &= ||\lambda(y - x)|| \\ &= \lambda ||y - x|| \end{aligned}$$

Therefore, if $0 < \lambda < \max\{\frac{\delta}{||y-x||}, 1\}$, then $||z - x|| < \delta$, and hence $z \in B_\delta(x)$. Because D is convex, $z \in D$. Hence, $z \in B_\delta(x) \cap D$.

□

The two Claims establish $\neg(b) \implies \neg(a)$, and hence the proof is complete.

■

Fact 7.10

- *All norms $\| \bullet \| : \mathcal{X} \to [0, \infty)$ are convex.*
- *For all $1 \leq \beta < \infty$, $\| \bullet \|^\beta$ is convex. Hence, on $\mathbb{R}^n$, $\forall\, 1 \leq p < \infty$, $\Sigma_{i=1}^n |x_i|^p$ is convex.*
- *Let $r > 0$, $\| \bullet \|$ a norm, $B_r(x_0)$ is a convex set; special case: $B_1(0)$ convex set. (unit ball about the origin).*
- *Let C be an open, bounded and convex set, $0 \in C$. Then, $\exists \| \bullet \| : X \to [0, \infty)$ such that $C = \{x \in \mathcal{X} \mid ||x|| < 1\} = B_1(0)$. In other words, open unit balls are characterized by the fact that they are open, bounded, convex sets that contain the origin.*
- *K_1 and K_2 are convex, then $K_1 \cap K_2$ is convex. (by convention, the empty set is convex)*
- *Consider $(\mathbb{R}^n, \mathbb{R})$ and let A be a real $m \times n$ matrix and $b \in \mathbb{R}^m$. Then,*

- $K_1 = \{x \in \mathbb{R}^n | \; Ax \leq b\}$ *is also convex (linear inequality with* $Ax \leq b$ *interpreted row wise).*
- $K_2 = \{x \in \mathbb{R}^n | \; Ax = b\}$ *is convex (linear equality).*
- $K_3 = \{x \in \mathbb{R}^n | \; A_{eq}x = b_{eq}, \; A_{in}x \leq b_{in}\}$ *is convex as well by the intersection property.*

Remark 7.11 $\widetilde{A}x \geq \widetilde{b} \iff (-\widetilde{A})x \leq (-\widetilde{b})$.

Fact 7.12 ***(Not an Easy one to Prove)*** *Suppose* $(\mathcal{X}, \mathbb{R}, ||\bullet||)$ *is a finite dimensional normed space,* $C \subset \mathcal{X}$ *is convex, and* $f : C \to \mathbb{R}$ *is convex. Then* f *is continuous on* $\mathring{C}$.

Remark 7.13 f *can have jumps on the boundary of* C*, that is, on* $\partial C := \overline{C} \cap \overline{(\sim C)} = \overline{C} \setminus \mathring{C} := \{x \in \overline{C} \mid x \notin \mathring{C}\}$.

7.2 Remarks on Notation and Abuse of Notation

Let $(\mathcal{X}, \mathbb{R}, \| \bullet \|)$ be a real normed space, $S \subset \mathcal{X}$, and $f : S \to \mathbb{R}$. It is very common to write

$$x^* = \arg\min_{x \in S} f(x) \tag{7.1}$$

for the value of $x \in S$ that achieves the minimum of f over all possible elements of S; that is, $f(x^*) = \min_{x \in S} f(x)$. Now, the problem is, you should only write something like (7.1) when

1. There does exist a minimum value, and
2. it is unique.

If a minimum exists but is not unique, one should write

$$x^* \in \arg\min_{x \in S} f(x) \tag{7.2}$$

to indicate that x^* is one of a set of values that all minimize the function f over the set S. It is correct notation, but not commonly used. If you are not sure that a minimum exists, then definitely you should not use (7.1). Even worse, **something you should never do is write**

$$x^* = \arg \inf_{x \in S} f(x) \tag{7.3}$$

because it makes no sense! By the very definition of an infimum, there may be no value in S achieving the infimum.

In (7.1), f is called the **cost function** and S is the **constraint set.**

7.3 What is a Quadratic Program?

Example 3.43 and Proposition 3.95 dealt with least squares solutions of overdetermined systems of linear equations

$$\widehat{x} = \arg\min_x (Ax - b)^\top Q(Ax - b),$$

with Q a positive definite matrix, while Theorem 3.51 and Proposition 3.91 dealt with minimum norm squared solutions of underdetermined systems of linear equations

$$\widehat{x} := \arg\min_{Ax=b} x^\top Qx.$$

These are both quadratic optimization problems that admit closed-form solutions.

A **Quadratic Program** is a more general kind of quadratic optimization problem with constraints. The cost to be minimized in (7.1) is quadratic plus a linear term, meaning that $f : \mathbb{R}^m \to \mathbb{R}$ has the form

$$f(x) = \frac{1}{2}x^\top Qx + qx, \tag{7.4}$$

where Q is an $m \times m$ symmetric, positive **semi-definite** matrix, and q is a $1 \times m$ row vector. Moreover, instead of optimizing over all of $\mathbb{R}^m$ as in Example 3.43 and Proposition 3.95, or only over linear equality constrains, as in Theorem 3.51 and Proposition 3.91, we

are allowed to seek solutions that lie in a subset of $\mathbb{R}^m$ defined by **linear inequality** and **linear equality** constraints that are typically written in the form

$$A_{in}x \preceq b_{in} \tag{7.5}$$
$$A_{eq}x = b_{eq}. \tag{7.6}$$

Recall that the symbol $\preceq$ is a way to define "less than or equal to" for vectors; it means that each component of the vector on the left hand side is less than or equal to the corresponding component of the vector on the right hand side. As an example

$$\begin{bmatrix}3\\2\\4\end{bmatrix} \preceq \begin{bmatrix}4\\3\\4\end{bmatrix},$$

though

$$\begin{bmatrix}3\\2\\4\end{bmatrix} \not\preceq \begin{bmatrix}1\\3\\4\end{bmatrix};$$

and

$$\begin{bmatrix}3 & 1\\2 & 4\end{bmatrix}\begin{bmatrix}x_1\\x_2\end{bmatrix} \preceq \begin{bmatrix}0\\9\end{bmatrix},$$

means that x_1 and x_2 must satisfy

$$3x_1 + x_2 \leq 0$$
$$2x_1 + 4x_2 \leq 9.$$

What if you really wanted $2x_1 + 4x_2 \geq 9$? Then you need to remember that when you multiply both sides by a minus sign, the inequality sign flips. Hence,

$$\begin{matrix}3x_1 + x_2 \leq 0\\2x_1 + 4x_2 \geq 9\end{matrix} \iff \begin{matrix}3x_1 + x_2 \leq 0\\-2x_1 - 4x_2 \leq -9\end{matrix} \iff \begin{bmatrix}3 & 1\\-2 & -4\end{bmatrix}\begin{bmatrix}x_1\\x_2\end{bmatrix} \preceq \begin{bmatrix}0\\-9\end{bmatrix}.$$

In addition, most QP solvers allow one to specify lower and upper bounds on x of the form

$$lb \preceq x \preceq ub. \tag{7.7}$$

While such constraints could always be rewritten in the form of (7.5), using (7.7) is more convenient, intuitive, and less error prone. The inclusion of constraints allows for very interesting and practical optimization problems to be posed.

Useful Fact about QPs

We consider the QP

$$x^* = \begin{matrix}\arg\min\\x \in \mathbb{R}^m\end{matrix} \quad \frac{1}{2}x^\top Qx + qx \tag{7.8}$$
$$A_{in}x \preceq b_{in}$$
$$A_{eq}x = b_{eq}$$
$$lb \preceq\ x \preceq ub$$

and assume that Q is symmetric ($Q^\top = Q$) and **positive definite** ($x \neq 0 \implies x^\top Qx > 0$), and that the subset of $\mathbb{R}^m$ defined by the constraints is non empty, that is

$$S := \{x \in \mathbb{R}^m \mid A_{in}x \preceq b_{in},\ A_{eq}x = b_{eq},\ lb \preceq x \preceq ub\} \neq \emptyset. \tag{7.9}$$

Then x^* exists and is unique.

There are special purposes solvers available for QPs.

- https://www.ibm.com/docs/en/icos/20.1.0?topic=qp-optimizing-qps
- https://github.com/osqp/OSQP.jl

- https://web.stanford.edu/~boyd/papers/pdf/osqp.pdf
- https://stanford.edu/~boyd/software.html
- https://www.mathworks.com/help/optim/ug/quadprog.html
- https://www.mathworks.com/help/mpc/ug/qp-solver.html

How do QPs arise in Robotics? Here is one example. Consider the robot equations,

$$D(q)\ddot{q} + C(q,\dot{q})\dot{q} + G(q) = Bu$$

where $q \in \mathbb{R}^n$, $u \in \mathbb{R}^m$. The ground reaction forces for a bipedal robot can be expressed as

$$F = \Lambda_0(q,\dot{q}) + \Lambda_1(q)u = \begin{bmatrix} F^h \\ F^v \end{bmatrix}.$$

Suppose the commanded torque is $u = \gamma(q,\dot{q})$ and we need to respect bounds on the ground reaction forces, such as

$$F^v \geq 0.2 m_{total} g,$$

which means the normal force should be at least 20% of the total weight of the robot, and

$$|F^h| \leq 0.6 F^v,$$

which places the horizontal component of the ground reaction force in a friction cone with magnitude less than 60% of the total vertical force. Putting it all together:

$$\begin{bmatrix} F^v & \geq 0.2 m_{total} g \\ F^h & \leq 0.6 F^v \\ -F^h & \leq 0.6 F^v \end{bmatrix} \iff A_{in}(q)u \leq b_{in}(q,\dot{q}).$$

QP:

$$u^* = \text{argmin } u^\top u + p\, d^\top d$$

$$\begin{aligned} A_{in}(q)u &\leq b_{in}(q,\dot{q}) \\ u &= \gamma(q,\dot{q}) + d, \end{aligned}$$

where d is a relaxation parameter that allows the torque to deviate from its desired value in order to respect the constraints on the ground reaction forces. The parameter p is a scalar weight term.

7.4 What is a Linear Program and How can it be used to Minimize $||\bullet||_1$ and $||\bullet||_{\max}$?

Definition 7.14 *A **Linear Program** means to minimize a scalar-valued linear function subject to linear equality and inequality constraints. For $x \in \mathbb{R}^n$, and $f \in \mathbb{R}^n$*

$$\begin{aligned} \textit{minimize} \quad & f^\top x \\ \textit{subject to} \quad & A_{in}x \preceq b_{in} \\ & A_{eq}x = b_{eq} \end{aligned}$$

where $A_{in}x \preceq b_{in}$ means each row of $A_{in}x$ is less than or equal to the corresponding row of b_{in}. The only restrictions on A_{in} and A_{eq} are that the set

$$K = \{x \in \mathbb{R}^n \mid A_{in}x \preceq b_{in},\ A_{eq}x = b_{eq}\}$$

should be non-empty.

Remarkably, through a concept called **slack variables**, Linear Programs can be applied to cost functions that include the one-norm and the max-norm.

Linear Program for ℓ_1-norm: $||x||_1 = \sum_{i=1}^{n} |x_i|$

Suppose that A is an $m \times n$ real matrix. Minimize $||Ax - b||_1$ is equivalent to the following linear program on $\mathbb{R}^{n+m}$

$$\begin{array}{ll} \text{minimize} & f^\top X \\ \text{subject to} & A_{in} X \preceq b_{in} \end{array} \tag{7.10}$$

with $X = \begin{bmatrix} x \\ s \end{bmatrix}$ ($s \in \mathbb{R}^m$ are called slack variables)

$$f := \begin{bmatrix} 0_{1\times n} & \mathbf{1}_{1\times m} \end{bmatrix}, \; A_{in} := \begin{bmatrix} A & -I_{m\times m} \\ -A & -I_{m\times m} \end{bmatrix} \text{ and } b_{in} := \begin{bmatrix} b \\ -b \end{bmatrix}$$

If $\widehat{X} = [\widehat{x}^\top, \widehat{s}^\top]^\top$ is the solution of the linear programming problem, then $\widehat{x}$ solves the 1-norm optimization problem; that is

$$\widehat{x} \in \arg \min_{x \in \mathbb{R}^n} ||Ax - b||_1.$$

Let's see if we can understand why the above is true. Writing out the terms, equation (7.10) becomes

$$\begin{array}{ll} \underset{x,s}{\text{minimize}} & \displaystyle\sum_{i=1}^{m} s_i \\ \text{subject to} & Ax - s \preceq b \\ & -Ax - s \preceq -b. \end{array}$$

This is equivalent to

$$\begin{array}{ll} \underset{x,s}{\text{minimize}} & \displaystyle\sum_{i=1}^{m} s_i \\ \text{subject to} & -s \preceq b - Ax \\ & -s \preceq -(b - Ax), \end{array}$$

which is equivalent to

$$\begin{array}{ll} \underset{x,s}{\text{minimize}} & \displaystyle\sum_{i=1}^{m} s_i \\ \text{subject to} & -s \preceq b - Ax \\ & +s \succeq b - Ax, \end{array}$$

which is equivalent to

$$\begin{array}{ll} \underset{x,s}{\text{minimize}} & \displaystyle\sum_{i=1}^{m} s_i \\ \text{subject to} & -s \preceq b - Ax \preceq s. \end{array}$$

Because for real numbers s_i, y_i, the inequality $(-s_i \le y_i \le s_i) \iff (0 \le |y_i| \le s_i)$, we end up with

$$\begin{array}{ll} \underset{x,s}{\text{minimize}} & \displaystyle\sum_{i=1}^{m} s_i \\ \text{subject to} & 0 \le |b - Ax|_i \le s_i, \end{array}$$

which is equivalent to

$$\underset{x}{\text{minimize}} \ \sum_{i=1}^{m} |b - Ax|_i.$$

Whoever thought this up was pretty clever! It reduces a nonlinear problem to a linear problem. The max-norm is a bit simpler, requiring only a single slack variable.

Linear Program for ℓ_∞-norm: $||x||_\infty = \max_{1 \le i \le n} |x_i|$

Suppose that A is an $m \times n$ real matrix. Minimize $||Ax - b||_\infty$ is equivalent to the following linear program on $\mathbb{R}^{n+1}$

$$\begin{aligned} &\text{minimize} \ \ f^\top X \\ &\text{subject to} \ \ A_{in} X \preceq b_{in} \end{aligned}$$

with $X = \begin{bmatrix} x \\ s \end{bmatrix}$ ($s \in \mathbb{R}$ is called a slack variable)

$$f := \begin{bmatrix} 0_{1\times n} & 1 \end{bmatrix}, \ A_{in} := \begin{bmatrix} A & -\mathbf{1}_{m\times 1} \\ -A & -\mathbf{1}_{m\times 1} \end{bmatrix} \text{ and } b_{in} := \begin{bmatrix} b \\ -b \end{bmatrix}$$

If $\widehat{X} = [\widehat{x}^\top, \ \widehat{s}\,]^\top$ solves the linear programming problem, then $\widehat{x}$ solves the max-norm optimization problem; that is

$$\widehat{x} \in \arg \ \min_{x \in \mathbb{R}^n} ||Ax - b||_\infty.$$