EECS 127: Optimization Models in Engineering

Michael Pham

Fall 2024

Contents

Week 1
Introduction

1.1 Lecture – 8/29/2024

1.1.1 Motivating Examples

To begin with, we will provide the following motivating examples:

Example 1.1 (Oil Refinery). Suppose we have an oil refinery which produces jet fuel (at a profit of \$0.10/barrel) and gasoline (at a profit of \$0.20/barrel).

From here, we have the following constraints: first, the oil refinery has a capacity of at most 10,000 barrels. Furthermore, it has a contract stating that it has to produce at least 1000 barrels of jet fuel and 2000 barrels of gasoline.

Next, we know that the trucks has a capacity of 180,000 barrel-mile, with the jet fuel being 10 miles away, and gasoline being 30 miles.

Then, we can model it as follows. First, let us denote x_1 to represent jet fuel, and x_2 to represent gasoline. Then, we have the following:

$$\begin{aligned} \max_{x_1,x_2} \quad & 0.1x_1 + 0.2x_2 \\ \text{s.t.} \quad & x_1 + x_2 \leq 10000 \\ & x_1 \geq 1000 \\ & x_2 \geq 2000 \\ & 10x_1 + 30x_2 \leq 180000 \end{aligned}$$

And in this problem, we want to find the values of x_1 and x_2 which maximises the profit while staying within our constraints.

Example 1.2 (Knapsack Problem). Suppose we have a set of n items: $1, 2, \ldots, n$. Then, the item $i \in \{1, \ldots, n\}$ has corresponding weight w_i and value v_i.

Furthermore, our bag has a weight capacity of at most w. Then, we want to select a set of items from our set such that the total value v is maximised.

To model this problem, we can first denote each item by a variable x_i, where $i \in \{1, \ldots, n\}$. Next, we will have x_i be an indicator variable such that $x_i = 1$ if the i^{th} item is selected, and 0 otherwise.

The objective then is that we want to maximize $\sum_{i=1}^{n} v_i x_i$ while keeping to the constraints of $\sum_{i=1}^{n} w_i x_i \leq w$.

1.1.2 Standard Form of Optimization

Definition 1.3 (Standard Form of Optimization)**.** The standard form of optimization goes as follows:

$$\begin{aligned} \min_{f \in \mathbb{R}^n} \quad & f_0(\vec{x}) \\ \text{s.t.} \quad & f_i(\vec{x}) \leq 0 \quad i \in \{1, \ldots, m\} \\ & h_j(\vec{x}) = 0 \quad \forall j \in \{1, \ldots, p\} \end{aligned}$$

Here, we note that:

- $\vec{x} \in \mathbb{R}^n$ is our optimization variable.
- $f_0(\vec{x})$ is our objective function.
- $f_1, \ldots, f_m$ and $h_1, \ldots, h_p$ are functions $\mathbb{R}^n \to \mathbb{R}$.
- f_i and h_j are inequality and equality constraint functions respectively.

We note that if we are given a maximization problem, we can "convert" it to a minimization problem by simply looking at $\min(-f_0(\vec{x}))$ instead when given $\max f_0(\vec{x})$.

Definition 1.4 (Feasible Solutions)**.** We define $y \in \mathbb{R}^n$ to be a feasible solution/point if it satisfies all of our constraints. Otherwise, it is an infeasible solution.

We note that y doesn't necessarily have to minimize our objective function.

Definition 1.5 (Feasible Sets)**.** he feasible set – denoted either by Ω or X – is the set of all feasible solutions. That is, we have

$$X \coloneqq \{\vec{x} \in \mathbb{R}^n : f_1(\vec{x}) \leq 0, \ldots, f_m(\vec{x}) \leq 0\}.$$

Definition 1.6 (Global Minimum)**.** We denote x^* to be the "global minimum" of our feasible set if $f_0(x^*) \leq f_0(x)$ for all $x \in X$.

Definition 1.7 (Local Minimum)**.** A say that x^* is a local minimum if there exists some neighbourhood around x^* such that $f_0(x^*) \leq f_0(x)$ for all x in our neighbourhood.

We note here that if the radius of our "ball" is $\pm\infty$, then we in fact have a global minimum.

WEEK 2

SECOND WEEK WOES

2.1 Lecture – 9/3/2024

2.1.1 Review

Recall from the previous lecture that we denote x^* to be the optimal solution, and $f_0(x^*)$ to be the optimal objective value.

Then, the standard form would be the following: $\min_x f_0(x)$, such that $f_1(x) \leq 0, \ldots, f_m(x) \leq 0$.

Now, we had cases such as finding $\min x$; in this case, $x^* = -\inf$ and $f_0(x^*) = -\inf$. We say then that the optimal solution x^* doesn't exist, and the function is unbounded from below.

In the case of $\min e^x$, we have that $x^* = -\inf$. However, note that $f_0(x^*) = 0$. Thus, while the optimal solution x^* is not attainable, but the optimal objective value does exist.

And of course, the well-behaved case is when x^* is attainable, and $f_0(x^*)$ is fixed; they're both finite.

In the case where we have no feasible solution, recall that we call the problem is "infeasible."

> **Example 2.1.** Suppose we have $\min x$, such that $x^2 \leq -1$. We note that the optimization is infeasible, as we can't have $x^2 \leq -1$.
>
> We say then that the optimal objective value, $f_0(x^*)$, is equal to $+\infty$

Then, we have three cases for our optimal objective value:

1. $+\infty$; in this case, it's infeasible.
2. Finite; in this case, x^* *may* or may not be attainable.
3. $-\infty$; in this case, it's unbounded from below.

2.1.2 Categories of Optimization Problems

We say that optimization problems are either tractable or not-tractable.

> **Definition 2.2** (Tractable vs Not-Tractable)**.** A problem is "tractable" is an algorithm exists to solve the problem. On the other hand, not-tractable problems are ones where an algorithm doesn't exist to solve

the problem.

! Note that not-tractable problems may have an algorithm to solve it, but it may not be fast enough.

Example 2.3 (Binary Decisions). Let us suppose we have to make n decisions for a company to maximize their profits. For each decision, we have two options: yes or no.

Then, we can map the i^{th} decision to an indicator variable x_i, where $i \in \{1, \dots, n\}$.

Then, we have that

$$x_i = \begin{cases} +1 \text{ if yes.} \\ -1 \text{ if no.} \end{cases}$$

Then, we can write this as an optimization problem $\max f_0(x)$ (or $\min(-f_0(x))$), with the condition $x_i^2 = 1$. This can be written as

$$\begin{cases} x_i^2 - 1 \leq 0 \\ -x_i^2 + 1 \leq 0 \end{cases}$$

Then, the number of feasible points is equal to 2^n. Suppose we had $n = 500$; then, the number of evaluations of the objective function would be equal to 2^{500}.

Thus, the problem is not-tractable.

! Note that if problems can be solved within polynomial time, then it is tractable. Thus, in this class, we are looking at the class of "convex optimization" problems.

Interactable Functions and EECS 227B

Note that we have interactable problems; these are approximations. Suppose we wanted to minimize some function $\min f_0(x)$ that looks like the following:
To solve, it we can consider some smooth, underestimation of the actual function – we call this $g_0(x)$.
It is a much easier problem to look at than the original.
The goal then is to make our approximation of the global minimum as close as possible to the real value.

2.1.3 Linear Al Jabr

Definition 2.4 (Space). A space is a collection of objects of a certain type.

Example 2.5 ($\mathbb{R}^n$). For example, $\mathbb{R}^n$ is a collection of vectors with n elements.

We can also talk about things such as a space of matrices, polynomials, etc.

Example 2.6 ($\mathscr{P}$). We can talk about the collection of polynomials with one variable, $a_0 + a_1 t + \cdots + a_n t^n$.

Definition 2.7 (Subspace)**.** We define a subspace to be a non-empty set V in our space $\mathbb{R}^n$ with the following properties:

- The zero vector is contained in the subspace.
- For any two vectors $u, v \in V$, the linear combination $\alpha u + \beta v \in V$ as well.

Example 2.8 (Examples in $\mathbb{R}^2$)**.** Suppose we are working in $\mathbb{R}^2$. Then, consider the line $y = x$; we note that this is in fact a subspace of $\mathbb{R}^2$ as it satisfies both conditions.

On the other hand, if our line no longer crosses the origin, it wouldn't be a subspace as $(0, 0) \notin V$.

Definition 2.9 (Span)**.** We define the span of a set of vectors S to be the set of linear combinations of vectors in S.

Example 2.10. Let us consider m vectors: $x^{(1)}, x^{(2)}, \ldots, x^{(m)} \in \mathbb{R}^n$.

Then, we define $\operatorname{span}\left(x^{(1)}, x^{(2)}, \ldots, x^{(m)}\right) = a_1 x^{(1)} + \cdots + a_m x^{(m)}$. Note then that the span is in fact a subspace.

Definition 2.11 (Basis)**.** Let us consider a subspace V. Then, the basis is defined to be a linearly independent, spanning set of vectors $x_1, \ldots, x_n$ of V.

Definition 2.12 (Linear Independence)**.** We say that a set of vectors $x_1, \ldots, x_m$ are linearly independent if only the trivial solution satisfies the following:

$$a_1 x_1 + \cdots + a_m x_m = 0$$

where $a_1, \ldots, a_m$ are scalars.

Example 2.13. Suppose we have the following:

$$x_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$
$$x_2 = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$
$$x_3 = \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}$$

Then, we define $V = \operatorname{span}(x_1, x_2, x_3)$. Note that this isn't linearly independent; we note that $x_2 - 2x_1 = x_3$. Thus, they aren't a basis for V.

Instead, if we considered $V = \operatorname{span}(x_1, x_2)$, then this would be linearly independent and thus is a basis.

Similarly, if we considered $\operatorname{span}(x_1, x_3)$, this would also be a basis for V.

Furthermore, we can see that $\dim V = 2$.

Definition 2.14 (Affine Sets)**.** A set is $X \subseteq \mathbb{R}^n$ is an affine set if there exists a subspace V of $\mathbb{R}^n$ and a vector x_0 in $\mathbb{R}^n$ such that $X = x^0 + V$.

Example 2.15. Let us return to our example of $\mathbb{R}^2$ with $y = x$. Then, let us consider the other example which wasn't a vector space (the shifted line); we note then that, in fact, it is an affine set. It's just a shifted version of our subspace.

! Clearly, affine sets may not go through the origin.

Definition 2.16 (Norm)**.** We define the norm in $\mathbb{R}^n$ to be a function $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$ such that:

1. $\|x\| \geq 0, \forall x \in \mathbb{R}^n$.
2. $\|x\| = 0 \iff x = 0$.
3. $\|x + y\| \leq \|x\| + \|y\|, \forall x, y \in \mathbb{R}^n$.
4. $\|\alpha x\| = |\alpha| \|x\|, \forall \alpha \in \mathbb{R}, \forall x \in \mathbb{R}^n$.

Example 2.17 (Examples of Norms)**.** A norm on $\mathbb{R}^n$ can be defined as the absolute value:

$$\|x\|_1 = |x_1| + |x_2| + \cdots + |x_n|$$

This is called the L_1 norm.

We can also define the following:

$$\|x\|_2 = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}$$

This is called the L_2 norm.

Furthermore, we have:

$$\|x\|_\infty = \max\{|x_1|, \ldots, |x_n|\}$$

And the General L_p Norm, where $1 \leq p < \infty$:

$$\|x\|_p = (|x_1|^p + \cdots + |x_n|^p)^{\frac{1}{p}}$$

Finally, the zero "norm" is defined to simply be:

$$\begin{aligned}\|x\|_0 &= \lim_{p \to 0} (|x_1|^p + \cdots + |x_n|^p)^{\frac{1}{p}} \\ &= \operatorname{card} x \\ &= \#\text{ of non-zero elements of } x\end{aligned}$$

Definition 2.18 (Inner Products)**.** We define the inner product to be a function on the space X $\langle \cdot, \cdot \rangle : X \times X \to \mathbb{R}$ such that:

1. $\langle x, x \rangle \geq 0, \forall x \in X$.
2. $\langle x, x \rangle = 0 \iff x = 0$.
3. $\langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle, \forall x, y, z \in X$.
4. $\langle \alpha x, y \rangle = \alpha \langle x, y \rangle, \forall \alpha \in \mathbb{R}, \forall x, y \in X$.
5. $\langle x, y \rangle = \langle y, x \rangle, \forall x, y \in X$.

Example 2.19. Let us consider some inner product $\langle \cdot, \cdot \rangle$ on $\mathbb{R}^n$ as follows:

$$\langle x, y \rangle = \beta_1 x_1 y_1 + \cdots + \beta_n x_n y_n$$

where $\beta_1, \ldots, \beta_n > 0$. Note that if we let $\beta_1 = \cdots = \beta_n = 1$, then we get the "standard" $\langle \cdot, \cdot \rangle$.

Note we can also define $\langle x, y \rangle = \|x\|_2 \|y\|_2 \cos\theta$.

2.2 Lecture – 9/5/2024

2.2.1 Review

Recall from the previous lecture that we discussed the notions of subspaces, affine sets, etc.

Furthermore, we denoted the inner product $\langle x, y \rangle = x_1 y_1 + \cdots + x_n y_n = \|x\|_2 \|y\|_2 \cos(\theta)$

And if $\langle x, y \rangle = 0$, this means that the two vectors are thus orthogonal.

Example 2.20 (Applications of Linear Algebra)**.** Given two published articles by two news outlets (e.g. CNN and Fox News), we want to see the similarity of the two articles.

One way to do this is to create a simplified dictionary containing major key words, such as the following: {president, tax, california, ...}.

Now, we define $x \in \mathbb{R}^n$, where x_i is the number of times the word i appears in our first article. Similarly, we define $y \in \mathbb{R}^n$, where x_i is the number of times the word i appears in the other article.

Then, using the notion of inner product, we can first calculate $\cos\theta$; this gives us how similar these articles are. Namely, the closer $\cos\theta$ is to zero, the less similar the articles are to each other.

On the other hand, the closer the value is to one, the more similar the articles are to each other.

2.3 Discussion – 9/6/2024

2.3.1 Problem 1

You have $12,000 to invest at the beginning of the year, and three different funds from which to choose. The municipal bond fund has a 7% yearly return, the local bank's CDs have an 8% return, and a high-risk account has an expected 12% return.

To minimize the risk, you decide not to invest more than $2,000 in the high-risk account.

For tax reasons, you need to invest at least three times as much in the municipal bonds as in the bank CDs.

Problem 2.1. Assuming the year-end yields are as expected, formulate the optimization problem in standard form.

Solution. We let x_1, x_2, x_3 to denote the amount of money we invest in the bond, CD, and high-risk account respectively.

We want to maximize the following:

$$\max_{x_1,x_2,x_3} 1.07x_1 + 1.08x_2 + 0.12x_3$$

Furthermore, we have the following conditions:

$$\begin{aligned} x_1 + x_2 + x_3 &\leq 12000 \\ 3x_2 &\leq x_1 \\ x_3 &\leq 2000 \\ 0 &\leq x_1 \\ 0 &\leq x_2 \\ 0 &\leq x_3 \end{aligned}$$

Putting this all together, we have the following:

$$\begin{aligned} \min_{x_1,x_2,x_3} \quad & -1.07x_1 - 1.08x_2 - 0.12x_3 \\ \text{s.t.} \quad & x_1 + x_2 + x_3 - 12000 \leq 0 \\ & -x_1 + 3x_2 \leq 0 \\ & x_3 - 2000 \leq 0 \\ & -x_1 \leq 0 \\ & -x_2 \leq 0 \\ & -x_3 \leq 0 \end{aligned}$$

■

Problem 2.2. If instead we were to invest exactly three times as much in the bonds as in the CDs, how would the problem change?

Solution. In this case, it would be the same other than the fact that we need to add in an extra condition $3x_1 - x_2 \leq 0$. ■

! Note that we have the (hidden) constraint of the variables being non-negative...!

2.3.2 Problem 2

A slalom skier must pass through n parallel gates of known position (x_i, y_i) and width c_i with $i \in \{1, \ldots, n\}$

Problem 2.3. Write an optimization problem that minimizes the total length of the path in terms of the variables $\{(x_i, y_i, c_i)\}_{i=0}^{n+1}$.

Solution. We can minimize the distance between each gate using the $L2$ norm. We denote z_i to be the y position of the skier at each time frame i:

$$\min_{z_0,\dots,z_6} \sum_{i=1}^{6} \left\| \begin{bmatrix} x_i \\ z_i \end{bmatrix} - \begin{bmatrix} x_{i-1} \\ z_{i-1} \end{bmatrix} \right\|_2^2$$

And we have the following conditions:

$$\begin{aligned} y_i - \frac{c_i}{2} \le z_i \le y_i + \frac{c_i}{2} \\ z_0 = y_0 \\ z_6 = y_6 \end{aligned}$$

■

2.3.3 Problem 3

Problem 2.4. ...

Solution. Let x_1, x_2, x_3 denote the number of servings for corn, milk, and bread respectively. Thus, we want to minimize:

$$\min_{x_1,x_2,x_3} 0.15x_1 + 0.25x_2 + 0.05x_3$$

And we have the following conditions:

$$\begin{aligned} x_1 + x_2 + x_3 \le 10 \\ 2000 \le 70x_1 + 121x_2 + 65x_3 \le 2250 \\ 5000 \le 107x_1 + 500x_2 + 0x_3 \le 10000 \\ 45x_1 + 40x_2 + 60x_3 \le 1000 \\ x_1, x_2, x_3 \ge 0 \end{aligned}$$

■

2.3.4 Problem 4

Problem 2.5. ...

Solution. For each intersection j, let q_i represent the flow of traffic on each road segment coming in/out of our intersection.

Then, $q_i < 0$ denotes traffic going out, and $q_i > 0$ indicates traffic flowing in.

Then, for each intersection j, we can denote v_j to denote a vector as follows:

$$v_j = \begin{bmatrix} q_1 \\ \vdots \\ q_i \end{bmatrix}$$

Then, we can use a matrix with j rows and i columns such that each entry in our matrix determines whether there is flowing in or out of traffic due to a road q_i for the corresponding j^{th} intersection. ■

Problem 2.6. ...

Solution. To stay close to the historical data, we can use the $L2$ norm such that ■

WEEK 3

THIRD WEEK

3.1 Lecture – 9/10/2024

3.1.1 Projection Review

Example 3.1 (Formula for Projection on d-dimension Subspaces). Let S be a d dimensional subspace. Define a basis $x_1, \dots, x_d$ for S. Then, this means that $S = \text{span}\,(x_1, \dots, x_d)$.

Then, we want to project some vector $x \in \mathbb{R}^n$ onto the subspace S. Now, note that since the projection $x^* \in S$, it can be written as a linear combination of the basis vectors of S.

That is, we can write x^* as such:

$$x^* = \alpha_1 x_1 + \cdots + \alpha_d x_d$$

Recall from last time that $\langle x - x^*, y \rangle = 0$, for all $y \in S$.

Then, let us pick y to be x_j for $j \in \{1, \dots, d\}$. Then, we have:

$$\begin{aligned}
\langle x - x^*, y \rangle &= \langle x - \sum_{i=1}^{d} \alpha_i x_i, x_j \rangle \\
&= 0 \quad \text{for } j = 1, \dots, d \\
\langle \sum_{i=1}^{d} \alpha_i x_i, x_j \rangle &= \langle x, x_j \rangle \\
\sum_{i=1}^{d} \alpha_i \langle x_i, x_j \rangle &= \langle x, x_j \rangle
\end{aligned}$$

From here, we can expand the summation out for each j to get the following:

$$\begin{cases}
\alpha_1 \langle x_1, x_1 \rangle + \cdots + \alpha_d \langle x_d, x_1 \rangle = \langle x, x_1 \rangle \\
\vdots \\
\alpha_1 \langle x_1, x_d \rangle + \cdots + \alpha_d \langle x_d, x_d \rangle = \langle x, x_d \rangle.
\end{cases}$$

Now, we see that we have d equations with d unknown variables $\alpha_1, \dots, \alpha_d$. Then, if $x_1, \dots, x_d$ is orthonormal, then recall that $\langle x_i, x_j \rangle = 0$ for $i \neq j$.

Thus, in fact, we have the following:

$$\begin{aligned} a_j\langle x_j, x_j\rangle &= \langle x, x_j\rangle \quad \text{for } j = 1, \dots, d \\ a_j &= \langle x, x_j\rangle \end{aligned}$$

Now, going back to the projection, we call that $\prod_S(x) = x^* = \alpha v = \langle x, x_1\rangle x_1 + \cdots + \langle x, x_2\rangle x_2$.

3.1.2 Hyperplanes and Half-Spaces

Hyperplanes

Definition 3.2 (Hyperplane)**.** Let us denote a hyperplane as H. Then, this is an $n-1$ dimensional affine set which can be written as a set of vectors $z \in \mathbb{R}^n$ such that $a^\intercal z = b$, where a is non-zero.

That is, we have:

$$H = \{z \in \mathbb{R}^n : a^\intercal z = b\}.$$

Example 3.3 (Hyperplane in $\mathbb{R}^3$)**.** In $\mathbb{R}^3$, the hyperplane will be two dimensional. Now, let us pick two vectors $z_1, z_2 \in H$ such that $a^\intercal z_1 = b$ and $a^\intercal z_2 = b$.

Subtracting the two equations, we then get $a^\intercal (z_1 - z_2) = 0$. Thus, we conclude that $a \perp (z_1 - z_2)$ (that is, a is orthogonal to $z_1 - z_2$).

a has a special name: it is the "normal vector" of our hyperplane.

Problem 3.1. Find a formula for the projection onto H

Solution. ... ■

Half-Spaces

Definition 3.4 (Half-Spaces)**.** We say that a hyperplane H divides our space into two regions:

1. $H_- = \{x : a^\intercal x \leq b\}$
2. $H_+ = \{x : a^\intercal \geq b\}$.

3.1.3 Unconstrained Optimization

This idea of unconstrained optimization $\min_{x\in\mathbb{R}^n} f(x)$ shows up quite often. For example, it appears in least-squares.

Before we explore this notion, we must first review certain terms.

Gradient

Definition 3.5 (Gradient)**.** We define the gradient as follows:

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix}$$

We note that if $n = 1$, then it's the same as the derivative.

Example 3.6. Suppose we have $f(x) = \sin x_1 + 4x_1x_2 + x_2^2$. Then, the gradient is:

$$\nabla f(x) = \begin{bmatrix} \cos x_1 + 4x_2 \\ 4x_1 + 2x_2 \end{bmatrix}$$

Example 3.7. Suppose we have $f(x) = \|x\|_2^2$. Then, we see that $f(x) = x_1^2 + \cdots + x_n^2$. Then, we see that the gradient is simply:

$$\begin{aligned} \nabla f(x) &= \begin{bmatrix} 2x_1 \\ \vdots \\ 2x_n \end{bmatrix} \\ &= 2x \end{aligned}$$

Example 3.8. If $f(x) = \|x\|_2$, we note that the gradient doesn't exist if $x = 0$. And we claim that if $x \neq 0$, we have:

$$\nabla f = \frac{x}{\|x\|}.$$

Types of Functions

Definition 3.9 (Linear Functions)**.** We say that $f : \mathbb{R}^n \to \mathbb{R}$ is linear if $f(\alpha x + \beta y) = \alpha f(x) + \beta f(y)$, for all $\alpha, \beta \in \mathbb{R}$ and $x, y \in \mathbb{R}^n$.

Proposition 3.10. If $f(x)$ is a linear function, we claim that there exists a vector $a \in \mathbb{R}^n$ such that $f(x) = a^\intercal x$.

Definition 3.11 (Affine Functions)**.** We say that $f : \mathbb{R}^n \to \mathbb{R}$ is an affine function if $f(x) - f(0)$ is a linear function.

Proposition 3.12. We thus claim that there exists $a \in \mathbb{R}^n$ and $b \in \mathbb{R}$ such that $f(x) = a^\intercal x + b$, where $b = f(0)$.

Proposition 3.13. We note that if f is an affine function, then $\nabla f(x) = a$.

Chain Rule for Gradient

Theorem 3.14. Consider $f : \mathbb{R}^m \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}^m$.
Then, we define $\varphi(x) = f(g(x))$. The gradient $\nabla\varphi(x)$ will thus be:

$$\nabla\varphi(x) = \begin{bmatrix}\nabla g_1(x) & \cdots & \nabla g_m(x)\end{bmatrix} \times \nabla f(z)\big|_{z=g(x)}$$

Example 3.15. Let us look at:

$$\begin{bmatrix}x_1\\x_2\end{bmatrix} \to \begin{bmatrix}2x_1 + 5x_2 + 1\\-x_1 + 5x_2 - 5\end{bmatrix}$$

Then, $f(x) = f(x_1, x_2)$, and $\varphi = f(2x_1 + 5x_2 + 1, -x_1 + 5x_2 - 5)$. Then, we have the following:

$$\begin{aligned}\nabla\varphi(x) &= \begin{bmatrix}\nabla(2x_1 + 5x_2 + 1) & \nabla(-x_1 + 5x_2 - 5)\end{bmatrix} \times \nabla f(z)\\ &= \begin{bmatrix}2 & -1\\5 & 5\end{bmatrix} \times \nabla f(z)\big|_{z=g(x)}\end{aligned}$$

Taylor Series

Given a function $f : \mathbb{R}^n \to \mathbb{R}$ which is differentiable, at a point $x_0 \in \mathbb{R}^m$, we can approximate the function with an affine function in a neighbourhood of x_0.
Recall that $f(x) = f(x_0) + \nabla f^{\intercal}_{(x_0)}(x - x_0) + \zeta(x)$. And we note that

$$\lim_{x \to x_0} \frac{\zeta(x)}{\|x - x_0\|} = 0$$

That is, $f(x) \approx f(x_0) + \nabla f^{\intercal}_{(x_0)}(x - x_0)$. Thus, we see that it's an affine function in x.

Theorem 3.16. If x^* is a local minimization of $\min_{x\in\mathbb{R}^n} f(x)$, where $f(x)$ is differentiable at x^*, then the gradient $\nabla f(x^*) = 0$.

Proof. Since x^* is a local minimum, we note that $\exists r > 0$ such that $f(x^*) \le f(x)$, $\forall x : \|x - x^*\| \le r$.
Then, we see that $x^* + ty$ will be inside our neighbourhood for sufficiently small t. Then, we have:

$$\begin{aligned}f(x^* + ty) &\ge f(x^*)\\ 0 &\le \frac{f(x^* + ty) - f(x^*)}{t} \quad t > 0, \forall t : \text{small}\\ 0 &\le \lim_{t\to 0^+} \frac{f(x^* + ty) - f(x^*)}{t}\\ &= \frac{\partial f(x^* + ty)}{\partial t}\Big|_{t=0}\\ &= \frac{\partial f(x_1^* + ty_1, \ldots, x_n^* + ty_n)}{\partial t}\Big|t = 0\\ &= \sum_{i=1}^{n} \frac{\partial(x_i^* + ty_i)}{\partial t} \times \frac{\partial f(z)}{\partial x_i}\Big|_{z = x^* + ty \text{ with } t=0}\\ &= \sum_{i=1}^{n} y_i \frac{\partial f(x^*)}{\partial x_i}\\ &= y^{\intercal}\nabla f(x^*)\end{aligned}$$

Now, since y can be some arbitrary value that eventually gets scaled to be within our neighbourhood, let us set $y = -\nabla f(x^*)$. Then, we have:

$$\begin{aligned} 0 &\leq -\nabla f(x^*)^\intercal \nabla f(x^*) \\ &= -\left\|\nabla f(x^*)\right\|^2 \\ \implies \|\nabla f(x^*)\| = 0 &\implies \nabla f(x^*) = 0 \end{aligned}$$

■

Least-Squares

Let us consider $\min_{x\in\mathbb{R}^n} f(x)$, where $f(x) = \|Ax - b\|$, where A is an $n \times n$ matrix, and $b \in \mathbb{R}^m$.

Example 3.17 (Expressing Systems of Equations with Matrices). Suppose we ahve 3 equations and 2 unknowns:

$$\begin{aligned} x_1 - x_2 &= 1 \\ 2x_1 + 5x_2 &= -1 \\ -0.5x_1 + 3.5x_2 &= 10 \end{aligned}$$

We can express this as a matrix as follows:

$$\begin{bmatrix} 1 & -1 \\ 2 & 5 \\ -0.5 & 3.5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \\ 10 \end{bmatrix}$$

Example 3.18 (Fitting Power Law to Data). Suppose we have some system as follows:

$$x_1 \longrightarrow \boxed{\text{sys}} \longrightarrow y$$

Now, we say that $y = \alpha x_1^{a_1} \cdots x_n^{a_n}$. We have $\alpha, a_1, \ldots, a_n$ unknown variables.

Now, we have the following inputs to outputs:

$$\begin{Bmatrix} x_1 \\ \vdots \\ x_m \end{Bmatrix} \to \begin{Bmatrix} y_1 \\ \vdots \\ y_m \end{Bmatrix}$$

Now, we consider $\log(y) = \log(\alpha) + a_1 \log x_1 + \cdot + a_n \log x_n$. Then, we have:

$$\begin{bmatrix} 1 & \tilde{x}_1^{(1)} & \cdots & \tilde{x}_n^{(1)} \\ \vdots & & & \\ 1 & \tilde{x}_1^{(m)} & \cdots & \tilde{x}_n^{(m)} \end{bmatrix} \begin{bmatrix} \tilde{\alpha} \\ a_1 \\ \vdots \\ a_n \end{bmatrix}$$

3.2 Lecture – 9/12/2024

3.2.1 Examples

Example 3.19 (CAT Example)**.** To begin with, with can imagine an array of voxels. We denote x_i to be the density of the voxel i.

Then, we have the following equation:

$$I_1 = I_0 e^{-a_1 x_1 - \cdots - a_n x_n}.$$

Rearranging this, we get:

$$e^{-a_1 x_1 - \cdots - a_n x_n} = \frac{I_0}{I_1}$$
$$-a_1 x_1 - \cdots - a_n x_n = \log\left(\frac{I_0}{I_1}\right)$$

Now, recall that each a_i denotes the beam i. Then, we have: ...

3.2.2 Rank and Null Space

Consider $Ax = y$, where A, y are known, but x unknown. A is of dimension $m \times n$, x is $n \times 1$, and y is $m \times 1$.

Now, the solution set is defined as:

$$S = \{x \in \mathbb{R}^n : Ax = y\}.$$

Before we proceed, we need two notions:

- The range of A – denoted by $R(A)$ – is the set of Ax such that $x \in \mathbb{R}^n$. That is,

  $$R(A) = \{Ax : x \in \mathbb{R}^n\}.$$

  - Furthermore, we say that the dimension of $R(A)$ to be denoted by $\operatorname{rank}(A)$.
- The null-space $N(A)$ is defined as:

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

Theorem 3.20. The rank of a matrix A is equal to the number of linearly independent columns of A.

Corollary 3.21. The rank of a matrix A is equal to the number of linearly independent rows of A.

Thus, we have the following line inequalities:

$$\operatorname{rank}(A) \leq \text{maximum number of rows or columns}$$
$$0 \leq \operatorname{rank}(A) \leq \min(m, n)$$

Corollary 3.22. If $\operatorname{rank}(A) = 0$, then $A = 0$.

Theorem 3.23 (Fundamental Theorem of Linear Algebra)**.** We have the following two statements:

1. $N(A) \perp R(A^\intercal)$
2. $N(A) \oplus R(A^\intercal) = \mathbb{R}^n$.
3. $\dim N(A) + \operatorname{rank}(A) = n$.

Proof. First, we will show that $N(A) \perp R(A^\intercal)$.

Proof. To begin with, let us take $v \in N(A)$. Then, by definition, we have that $Av = 0$.

Now, let us take some $w \in R(A^\intercal)$. Then, we note that there exists some $u \in \mathbb{R}^m$ such that $A^\intercal u = w$.

Then, we observe that $\langle v, w\rangle = v^\intercal w = v^\intercal A^\intercal u = (Av)^\intercal u = 0$. □

Next, we prove the second statement. It says then that for any $h \in \mathbb{R}^n$, there exists some $v \in N(A)$ and $w \in \mathbb{R}(A^\intercal)$ such that $h = v + w$.

Proof. ... □

■

Now, let us go back to the equation $Ax = y$. We have a solution in each of the three cases:

1. $y \in R(A)$.
2. y is in the span of the columns of A.
3. $\operatorname{rank}(A) = \operatorname{rank}[Ay]$.

Example 3.24. First, let us pick an arbitrary solution $\bar{x}$: $A\bar{x} = y$. Then, we observe that $Ax = y$ reduces to $A(x - \bar{x}) = 0$.

Let $z = x - \bar{x}$. Then, we observe that $z \in N(A)$. Then, we observe that every solution x is equal to $\bar{x} + z$; i.e. it's the sum of $\bar{x}$ and a vector in the null space of A.

Thus, in fact, we can write $S = \bar{x} + N(A)$.

Remark 3.25. In fact, we note that the solution set S is actually an affine set. Though if the null space of $N(A)$ is zero, we note then that S has only one element; the solution is unique.

Now, in the case where we have infinitely many solutions, the question then arises: how do we pick the best one?

In this case, we want $\min_{x \in \mathbb{R}^n} \|x\|$, such that $Ax = y$.

Then, u^* is defined to be the projection of 0 onto S. Then, we project 0 onto $S = \bar{x} + N(A)$ (recall that S is an affine set!). Then, since $x^* \in S$, we have $x^* - \bar{x} \in N(A)$.

Thus, we have $\langle x^*, x^* - \bar{x}\rangle = 0$, for all $x \in S$. Thus, we conclude that x^* is orthogonal to the null space $N(A)$. And therefore, we note that $x^* \in R(A^\intercal)$.

WEEK 4

WEEK FOR SUFFERING

4.1 Discussion – 9/16/2024

Problem 4.1. Let $A \in \mathbb{R}^{3\times 2}$ be as follows:

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix}$$

Find the range of the matrix A, then express it in terms of a span of vectors.

Solution. To begin with, we let $\operatorname{range} A$ to be:

$$\operatorname{range} A = \left\{ \begin{bmatrix} 1 \\ 3 \\ 5 \end{bmatrix} x_1 + \begin{bmatrix} 2 \\ 4 \\ 6 \end{bmatrix} x_2 : x_1, x_2 \in \mathbb{R} \right\}$$

Then, we can rewrite this as:

$$\operatorname{range} A = \operatorname{span}\left(\begin{bmatrix} 1 \\ 3 \\ 5 \end{bmatrix}, \begin{bmatrix} 2 \\ 4 \\ 6 \end{bmatrix} \right)$$

■

Problem 4.2. Let $B \in \mathbb{R}^{3\times 3}$ defined to be:

$$B = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \\ 1 & 1 & -2 \end{bmatrix}$$

Find the null space of B, then find a basis for the null space and determine its dimension.

Solution. We want to find a vector x such that $Bx = 0$. Then, we observe:

$$\begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \\ 1 & 1 & -2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = 0$$
$$\begin{bmatrix} x_1 - x_3 \\ x_2 - x_3 \\ x_1 + x_2 - 2x_3 \end{bmatrix} = 0$$

Then, we see that $x_1 = x_2 = x_3$. So, we have that:

$$B \begin{bmatrix} x_3 \\ x_3 \\ x_3 \end{bmatrix} = 0$$

Then, we have that the null space $N(A)$ is equal to:

$$N(A) = \left\{ \alpha \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} : \alpha \in \mathbb{R} \right\}.$$

And we note that $v = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$ is a basis for $N(A)$, and thus has dimension 1. ■

Problem 4.3. Let $C \in \mathbb{R}^{4\times 3}$ be defined as:

$$C = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \\ 3 & 6 & 9 \\ 4 & 8 & 12 \end{bmatrix}$$

Find the rank of C, and verify it using different methods.

Solution. First, we observe that since all of the columns are scalar multiples of each other, we have that $\dim \operatorname{range} C = 1$.

We can also perform row-reduction to show that $\operatorname{rank} C = 1$. ■

Problem 4.4. Let $D \in \mathbb{R}^{4\times 4}$ be:

$$D = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Find the rank and nullity of D.

Solution. First, we see that $\operatorname{rank} D = 3$ since we have three linearly independent vectors. Then, we observe that $\operatorname{nullity} D = 4 - 3 = 1$.

Furthermore, we see that the null space $N(D)$ is:

$$\text{span}\left(\begin{bmatrix}0\\0\\0\\1\end{bmatrix}\right)$$

■

4.2 Lecture – 9/17/2024

When we have noise in our measurements, it will make our system inconsistent.

4.2.1 Least Squares, and Linear Regression

Example 4.1 (Noisy Measurements). Consider $m = 3$, $n = 2$. And we have the following:

$$\begin{cases} x_1 + x_2 = 2 \\ 2x_1 + 3x_2 = 5 \\ 4x_1 - x_2 = 3 \end{cases}$$

Now, we note that if let $x_1 = x_2 = 1$, the system of equations would be satisfied.

However, if instead we had:

$$x_1 + x_2 = 2 + \varepsilon$$

Where ε is some noise. In that case, to solve this system of equations, we would instead solve the least square $\min \|Ax - y\|_2$.

If the optimal objective is equal to zero, that is we have $\|Ax^* - y\|_2 = 0$, meaning that $Ax^* = y$. Otherwise, $Ax^* \neq y$.

From the example above, we see that $\|Ax - y\|$ is telling us how much we are violating $Ax = y$.

Example 4.2 (Linear Regression). Let's say we have a set of points

Then, we would like to find some line that is close to all of the points on our graph. The equation for this line then would be:

$$z_2 = az_1 + b.$$

We let $a = x_1$ and $b = x_2$. Let us assume that we have m different points.

So, we have $z_2^{(i)} - x_1 z_1^{(i)} - x_2$, for $i = 1, \ldots, m$. Note that this may not be equal to zero.

Then, we want to find:

$$\begin{aligned} \min_{x \in \mathbb{R}^2} \sum_{i=1}^{m} (x_1 z_1^{(i)} + x_2 - z_2^{(i)})^2 &= \sum_{i=1}^{m} \left(\begin{bmatrix} z_1^{(i)} & 1 \end{bmatrix} x - z_2^{(i)} \right)^2 \\ &= \min \|Ax - y\|_2^2 \end{aligned}$$

Where we have:

$$A = \begin{bmatrix} z_1^{(1)} & 1 \\ \vdots & \vdots \\ z_1^{(m)} & 1 \end{bmatrix}, \quad y = \begin{bmatrix} z_2^{(1)} \\ \vdots \\ z_2^{(m)} \end{bmatrix}$$

Example 4.3 (General Linear Regression). Let us consider having a general linear regression system, with input $a \in \mathbb{R}^n$ and output is $a^\intercal x \in \mathbb{R}$, where x is the vector of coefficients.

We apply inputs $a^{(1)}, \dots, a^{(m)} \in \mathbb{R}^n$. Then, we measure noisy outputs $y_1, \dots, y_m \in \mathbb{R}$.

In Least Squares, we minimize the sum of the residuals:

$$\min_{x \in \mathbb{R}^n} \sum_{i=1}^{m} \left((a^{(i)})^\intercal x - y_i \right)^2 = \min \|Ax - y\|_2^2$$

Where we define A, y to be:

$$A = \begin{bmatrix} (a^{(1)})^\intercal \\ \vdots \\ (a^{(m)})^\intercal \end{bmatrix}, \quad y = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix}$$

4.2.2 Non-Linear Regression

Example 4.4 (Non-Linear Regression). Suppose we want to find a quadratic function to fit points.

Now, we have points $(z_1^{(i)}), z_2^{(i)}$, with $i = 1, \dots, m$.

And we have the model $z_2 = az_1^2 + bz_1 + c$. We call $a = x_1$, $b = x_2$, and $c = x_3$.

Then, we want to minimize the following:

$$\min \sum_{i=1}^{m} \left((x_1 z_1^{(i)})^2 + x_2 z_1^{(i)} + x_3 - z_2^{(i)} \right)^2 = \min \|Ax - y\|^2$$

Where we let:

$$A = \begin{bmatrix} (z_1^{(1)})^2 & z_1^{(1)} & 1 \\ \vdots & \vdots & \vdots \\ (z_1^{(m)})^2 & z_1^{(m)} & 1 \end{bmatrix}, \quad y = \begin{bmatrix} z_2^{(1)} \\ \vdots \\ z_2^{(m)} \end{bmatrix}$$

4.2.3 Why the L2 Norm?

Now, the question arises on why we re minimizing the L2 Norm, and not any other function.

To answer this, let us suppose we are looking at the measured value $y =$ correct $\tilde{y} +$ sensor noise $\bar{y}$.

If $\bar{y}$ is Gaussian, then $\bar{y}_1, \dots, \bar{y}_m$ is i.i.d.

Then, we see that Least Squares becomes best estimation, which is maximum likelihood in statistics.

Then, we examine $\min_{x \in \mathbb{R}^n} \|Ax - y\|_2$, where $A \in \mathbb{R}^{m \times n}$.

Then, we look at the range $R(A) = \{Ax : x \in \mathbb{R}^n\}$, which is in fact a subspace.

Then, the projected point y^* exists and is unique as well. From previous lectures, we know that $y - y^* \perp R(A)$. Furthermore, we've seen before that $N(A^\intercal) \perp R(A)$, and that $N(A^\intercal) \oplus R(A) = \mathbb{R}^m$.

So, we see that $y - y^* \in N(A)$. That means then that $A^\intercal(y - y^*) = 0$. And since $y^* \in R(A)$, it follows then that there exists a point x^* such that $Ax^* = y^*$.

So, we have:

$$\begin{aligned} 0 = A^\intercal(y - y^*) = A^\intercal(y - Ax^*) = 0 \\ A^\intercal y = A^\intercal A x^* \end{aligned}$$

Now, we introduce the following theorem:

Theorem 4.5. Least Squares always has a solution, and is characterized as the set of solutions:

$$S = \{x^* : A^\intercal A x^* = A^\intercal y\}$$

If A turns out to be full column rank, we note then that $A^\intercal A$ will be invertible, and thus we can define $x^* = (A^\intercal A)^{-1} A^\intercal y = A^\intercal y$.

A Second Approach

Another way to find the formula is that $\min f(x) \to \nabla f(x) = 0$. But, note that $\|Ax - y\|_2$ is not differentiable at the origin. So, we replace it with $\|Ax - y\|_2^2$.

Recall that

$$\begin{aligned} \|Ax - y\|_2^2 &= \langle Ax - y, Ax - y \rangle \\ &= (Ax - y)^\intercal (Ax - y) \\ &= (x^\intercal A^\intercal - y^\intercal)(Ax - y) \\ &= x^\intercal p x + q^\intercal x + r \end{aligned}$$

Where we have $p = A^\intercal A \in \mathbb{R}^{n \times m}$, $q = -2A^\intercal y \in \mathbb{R}^n$, and $r = y^\intercal y \in \mathbb{R}$.

So, we have $x^\intercal p x + q^\intercal x + r$, a quadratic function. Suppose $n = 1$, so an example function we'd have is:

$$f(x) = 2x^2 + 5x - 1$$

If $n = 2$, we would have a function like:

$$f(x) = x_1^2 + 5x_2^2 - 6x_1x_2 + 7x_1 + 2x_2 - 5$$

To put it in standard form, we would do:

$$f(x) = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 1 & -3 \\ -3 & 5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 7 & 2 \end{bmatrix} x - 5$$

We always choose p to be symmetric. That is, $p = p^\intercal$.

Now, let us look at the (i, j) entry of p, $p_{i,j}$.

Then, we observe the following:

$$\begin{aligned} f(x) &= x^\intercal p x + q^\intercal x + r \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} p_{i,j} x_i x_j + \sum i = 1^n q_i x_i + r \\ &= \sum_{i=1}^{n} p_{i,i} x_i^2 + \sum_{i<j} (p_{i,j} + p_{j,i}) x_i x_j + \sum_{i=1}^{n} q_i x_i + r \end{aligned}$$

Then, we take the derivative to get:

$$\begin{aligned}\frac{\partial f(x)}{\partial x_i} &= 2p_{i,i}x_i + \sum_{i\neq j}(p_{i,j}+p_{j,i})x_j + q_i \\ &= 2p_{i,i}x_i + \sum_{i\neq j} 2p_{i,j}x_j + q_i \\ &= 2\begin{bmatrix}p_{i,1} & \cdots & p_{i,n}\end{bmatrix} + q_i\end{aligned}$$

Thus, we see that $\nabla f(x) = 2px + q$. Thus, we see that if x^* is a local minimum, we have $\nabla f(x^*) = 0$.

So, $\nabla f(x^*) = 0 \implies 2A^\intercal Ax^* + (-2A^\intercal y) = 0$.

Then, we have:

$$\begin{aligned}2A^\intercal Ax^* &= 2A^\intercal y \\ A^\intercal Ax^* &= A^\intercal y \\ x^* &= (A^\intercal A)^{-1}A^\intercal y\end{aligned}$$

4.2.4 Set of Solutions

Let us consider the set of solution

$$S = \{x^* : A^\intercal Ax^* = A^\intercal y\}$$

For this, we aren't making any assumptions on $A \in \mathbb{R}^{m\times n}$. In the case where $m \geq n$, then we can find the inverse of $A^\intercal A$, and thus have a unique solution. On the other hand, if $n > m$, then we have the case of an underdetermined system, and thus x^* may not be unique.

Eigenvalues, Singular Values, and Determinants

In the case of such an underdetermined system, we have to find the best solution. To solve such a problem, we need a few new concepts.

First, let us consider a matrix $A \in \mathbb{R}^{n\times n}$. Then, we denote its determinant to be $\det(A)$.

Definition 4.6 (Characteristic Polynomial)**.** Let $\lambda \in \mathbb{R}$. Then, we note that $P(\lambda) = \det(\lambda I_n - A)$, where I_n is the identity matrix in $R^{n\times n}$.

This polynomial is the characteristic polynomial of A.

We note that this polynomial has degree n, and thus has n zeroes. We can thus write it as:

$$P(\lambda) = (\lambda - \lambda_1)(\lambda - \lambda_2)\cdots(\lambda - \lambda_n).$$

These $\lambda_1, \ldots, \lambda_n$ are complex numbers which are roots of $P(\lambda)$.

Definition 4.7 (Eigenvalue)**.** We say that $\lambda_1, \ldots, \lambda_n$ are eigenvalues of A.

Then, this means that $\lambda \in \mathbb{C}$ if there exists some non-zero $U \in \mathbb{C}^n$ such that:

$$(\lambda I_n - A)u = 0$$

In other words, we have

$$AU = \lambda u$$

And we call u to be the eigenvector.

Example 4.8. Let us consider

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

Then, $\det(\lambda I - A) = (\lambda - 1)(\lambda - 1) = 0$. From the characteristic polynomial, we see that $\lambda_1 = \lambda_2 = 1$. These are the repeated eigenvalues of A.

Then, to find the eigenvector, we want to find:

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \lambda \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$$

Working through this yields us:

$$\begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \text{ or } \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

Example 4.9. Another example is to look at:

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

Then, we have: $\det(\lambda I - A) = (\lambda - 1)(\lambda - 1)$. Then, we have $\lambda_1 = \lambda_2 = 1$. Then, we see that:

$$Au = \lambda u \implies \begin{cases} u_1 + u_2 = u_1 \\ u_2 = u_2 \end{cases}$$

Then, we see that $u_2 = 0$, and u_1 is arbitrary. Thus, our eigenvector would be:

$$u = \begin{bmatrix} u_1 \\ 0 \end{bmatrix}$$

WEEK 5

WEEK FIVE

5.1 Discussion – 9/23/2024

PSD Matrices

Problem 5.1. Show that if $A \in \mathbb{S}_+^n$, then there existd some $P \in \mathbb{S}_+^n$ such that $A = P^2$.

Solution. We observe that if A is positive semidefinite, we note that it follows that A is symmetric. Then, we can diagonalize A as $A = U\Lambda U^\intercal$.

Then, with this in mind, we note that we can rewrite this as $A = (U\Lambda U^\intercal) = U\Lambda_1\Lambda_1 U^\intercal = U\Lambda_1^2 U^\intercal$.

This is because, since all entries of $\Lambda > 0$ and is a matrix with only non-zero entries along its diagonal, it follows then that we can simply let Λ_1 have diagonal entries which are square root of Λ. Then, we observe that:

$$\begin{aligned} U\Lambda_1^2 U^\intercal &= (U\Lambda_1 U^\intercal)(U\Lambda_1 U^\intercal) \\ &= PP \\ &= P^2 \end{aligned}$$

■

Problem 5.2. Show that for any matrix $Q \in \mathbb{R}^{m\times n}$, if $A = Q^\intercal Q$, then $A \in \mathbb{S}_+^n$.

Solution. Suppose that $Q \in \mathbb{R}^{m\times n}$, and $A = Q^\intercal Q$.

Then, we observe that:

$$\begin{aligned} x^\intercal Q^\intercal Q x &= v^\intercal v \\ &= |v|_2^2 \geq 0. \end{aligned}$$

■

Problem 5.3. Let $B \in \mathbb{R}^{m\times n}$ be an arbitrary matrix. Prove that non-zero eigenvalues of $BB^\intercal$ are the same as the non-zero eigenvalues of $B^\intercal B$.

Solution. Let us suppose some eigenvalue of B. Then, we have that $BB^\intercal x = \lambda x$, for some eigenvector x of B. Similarly, suppose we have $B^\intercal Bx = \mu x$.

Now, recall that since $B^\intercal B$ is PSD:

$$\begin{aligned} B^\intercal Bx &= \mu x \\ BB^\intercal Bx &= B\mu x \\ &= \mu Bx \\ BB^\intercal y &= \mu y \end{aligned}$$

Thus, we see that, indeed, the two share the same eigenvalue. ■

Gradient and Hessian

Problem 5.4. Consider the function

$$f(x) = e^{x_1} + e^{x_2} + \cdots + e^{x_n}.$$

Calculate its gradient, Hessian, and determine if the Hessian is a PSD.

Solution. First, we observe that the gradient is:

$$\nabla f(x) = \begin{bmatrix} e^{x_1} \\ e^{x_2} \\ \vdots \\ e^{x_n} \end{bmatrix}$$

Next, to check for its Hessian:

$$H_f = \begin{bmatrix} e^{x_1} & 0 & \cdots & 0 \\ 0 & e^{x_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & e^{x_n} \end{bmatrix}$$

To check that H_f is indeed PSD, we check that for any arbitrary vector x, and ■

WEEK 6

WEEK SIX SUFFERING

6.1 Discussion – 9/30/2024

Problem 6.1. Recall that SVD of a matrix A is $A = U\Sigma V^T$, where U is $n \times n$, Σ is $n \times m$, and $V^T = m \times m$. Note that $U^T U = I$, $V^T V = I$, and Σ is diagonal.

Then, find the SVD of the following matrix:

$$A = \begin{bmatrix} 3 & 0 \\ 4 & 5 \end{bmatrix}$$

Solution. Let us look at $A^T A$:

$$\begin{aligned} A^T A &= \begin{bmatrix} 3 & 4 \\ 0 & 5 \end{bmatrix} \begin{bmatrix} 3 & 0 \\ 4 & 5 \end{bmatrix} \\ &= \begin{bmatrix} 25 & 20 \\ 20 & 25 \end{bmatrix} \end{aligned}$$

Then, let us now look at $A^T A - \lambda I$ and find its determinant:

$$A^T A - \lambda I = \begin{bmatrix} 25 - \lambda & 20 \\ 20 & 25 - \lambda \end{bmatrix}$$

Then, we have that $(25 - \lambda)^2 - 20^2 = 0$. Then, we have $25^2 - 50\lambda + \lambda^2 - 20^2 = 0 \implies \lambda^2 - 50\lambda + 25^2 - 20^2$. So, we have $\lambda_1 = 5$ and $\lambda_2 = 45$.

For $\lambda_1 = 5$, we want $A^T A v = \lambda v$. That is, we see:

$$\begin{aligned} A^T A v &= \begin{bmatrix} 25v_1 + 20v_2 \\ 20v_1 + 25v_2 \end{bmatrix} \\ &= \begin{bmatrix} 5v_1 \\ 5v_2 \end{bmatrix} \end{aligned}$$

So, we have $v_1 = v_1$ and $v_2 = -v_1$ gives us an eigenvector $v = \begin{bmatrix} 1 & -1 \end{bmatrix}^T$

Then, for $\lambda = 45$, we see that:

$$\begin{bmatrix} 25v_1 + 20v_2 \\ 20v_1 + 25v_2 \end{bmatrix} = \begin{bmatrix} 45v_1 \\ 45v_2 \end{bmatrix}$$

Then, we see that $v_1 = v_1$ and $v_2 = v_1$. So, we have a second eigenvector $v = \begin{bmatrix} 1 & 1 \end{bmatrix}^T$. Note then that these correspond to columns of V. First, we note that Σ will be:

$$\Sigma = \begin{bmatrix} \sqrt{45} & 0 \\ 0 & \sqrt{5} \end{bmatrix}$$

Then, we note that our V would be:

$$V = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

To find U, we first look at Av_1 to get:

$$\begin{aligned} Av_1 &= \begin{bmatrix} u_1 & u_2 \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix} \begin{bmatrix} v_1^T \\ v_2^T \end{bmatrix} v_1 \\ &= \begin{bmatrix} u_1 & u_2 \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \\ &= \begin{bmatrix} u_1 & u_2 \end{bmatrix} \begin{bmatrix} \sigma_1 \\ 0 \end{bmatrix} \\ &= u_1 \sigma_1 \end{aligned}$$

■