Math 164 – Optimization

University of California, Los Angeles

Duc Vu

Spring 2021

This is math 164 – Optimization taught by Professor Li. We meet weekly on MWF from 3:00 pm to 3:50 pm for lecture. The main textbook used for the class is *An Introduction to Optimization* 4^{th} by *Chong* and *Zak*. Other course notes can be found at my github. Please let me know through my email if you spot any typos in the note.

Contents

List of Theorems

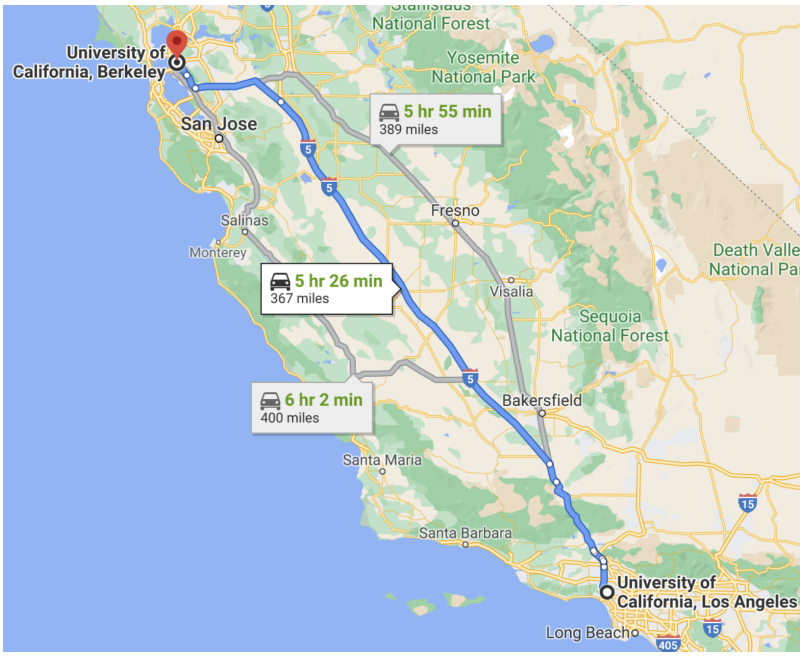
List of Definitions

§1 Lec 1: Mar 29, 2021

§1.1 Introduction

Question 1.1. Why Optimization?

- Find the fastest route from A to B.
- Possible constraints: avoid tolls?



Question 1.2. So what is optimization?

- Optimization is an important tool in decision science and in the analysis of artificial or physical systems.
- An optimization problem involves
  - An objective, which is a scalar, quantitative measure of the performance of the system under study.
  - examples of objectives include profit, time, energy, error, loss, cost, etc.
  - The objective depends on certain characteristics of the system, called variables or unknowns or parameters.
  - Often the variables are restricted, or constrained in some way.
  - The goal of solving an optimization problem is to find values of the variables that satisfy the constraints and optimize/minimize/maximize the objective.

In general, an optimization problem can be summarized as

Optimized	Objective(Variables)	Subject to	Constraints on variables

Applying the optimization framework to solve problems involves three steps:

1. Modeling: identifying objective, variables, and constraints for a given problem.
2. Solving: employing an optimization algorithm to find solutions, usually with the help of a computer.
3. Analyzing: recognizing whether the problem has been successfully solved using optimality conditions.

Mathematically speaking, optimization is the minimization or maximization of a (scalar valued) function subject to constraints on its variables.
We use the following notation

- x is a vector of variables/unknowns/parameters.
- $f(x)$ is the objective function, a scalar function of x that we want to maximize or minimize.
- $C_i(x)$ are constraint functions, which are scalar functions of x that define certain equations or inequalities that the unknown vector x must satisfy.

Using this notation, the optimization problem is

$$\text{minimize} \quad \underbrace{f(x)}_{\text{objective}} \quad \text{with} \quad \underbrace{x \in \mathbb{R}^n}_{\text{variables}} \quad \text{subject to} \quad \begin{cases} c_i(x) = 0, & i \in \underbrace{\mathcal{E}}_{\text{equality}} \\ c_i(x) \geq 0, & i \in \underbrace{\mathcal{I}}_{\text{inequality}} \end{cases}$$

The set of variables that satisfies all constraints, i.e.,

$$\Omega = \{x \in \mathbb{R}^n : c_i(x) = 0,\ i \in \mathcal{E},\ c_i(x) \geq 0,\ i \in I\}$$

is called the feasible region/set. So the optimization can also be written in an abstract manner as

$$\text{minimize } f(x) \text{ with } x \in \mathbb{R}^n \text{ subject to} \quad \underbrace{x \in \Omega}_{\text{feasible/constraint set}}$$

§1.2 Some Examples

Example 1.1

Consider the problem

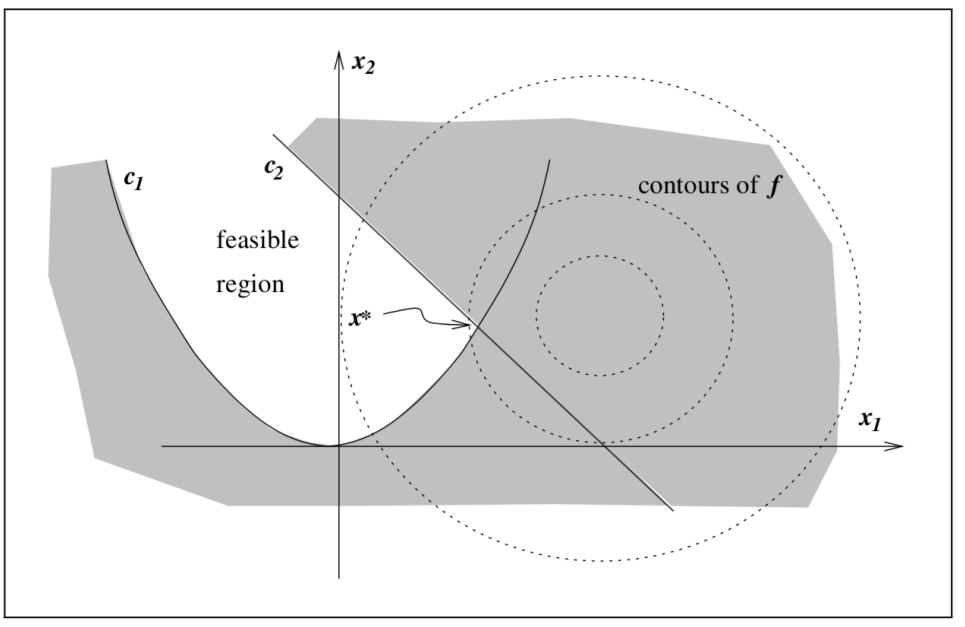
$$\begin{gathered}\text{minimize } (x_1-2)^2+(x_2-1)^2 \text{ subject to}\\ x_1^2-x_2\le 0\\ x_1+x_2\le 2\end{gathered}$$

We identify

- the optimization variable $x=\begin{bmatrix}x_1\\x_2\end{bmatrix}$
- the objective(cost) function $f(x)=(x_1-2)^2+(x_2-1)^2$
- the constraint $c(x)=\begin{bmatrix}c_1(x)\\c_2(x)\end{bmatrix}=\begin{bmatrix}x_2-x_1^2\\2-x_1-x_2\end{bmatrix}$, $\mathcal{I}=\{1,2\}$, $\mathcal{E}=\emptyset$.

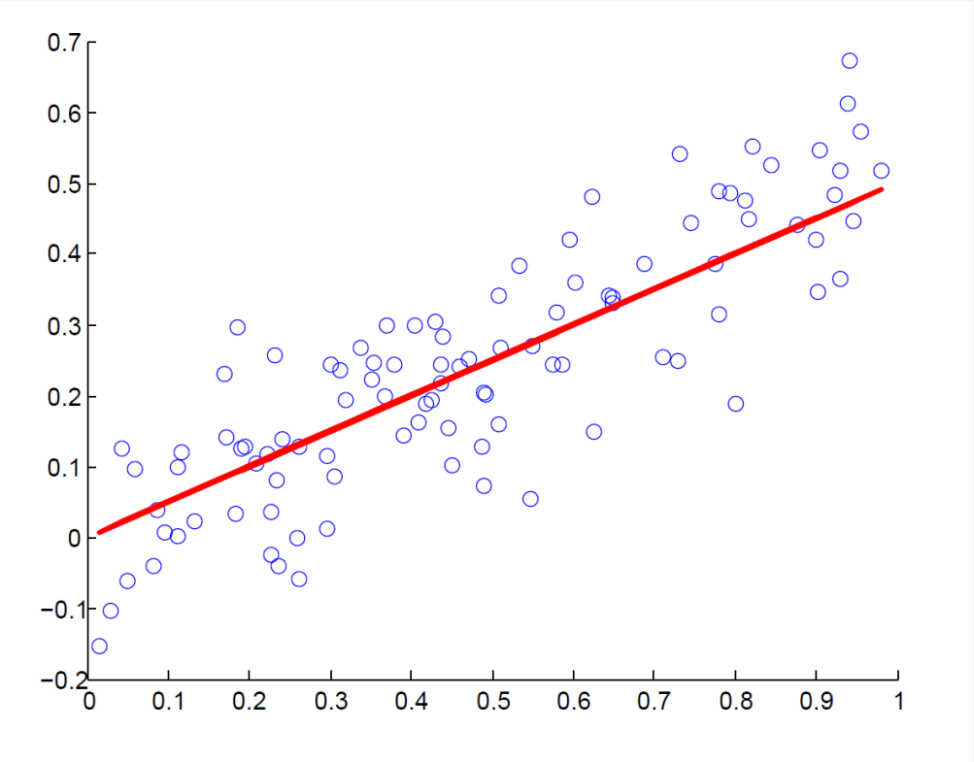
A lot of times we stack all equality constraints and/or inequality constraints into vector functions and write, e.g., $c(x)\ge 0$ meaning element-wise equality or inequality.

$$c(x)=\begin{bmatrix}c_1(x)\\c_2(x)\\c_3(x)\\\vdots\end{bmatrix}$$



Example 1.2 (Linear Regression)

Given a set of feature vectors $a_i \in \mathbb{R}^n$ and outcomes y_i, $i = 1, \ldots, N$, find weights x that predict the outcome accuracy $x^T a_i \approx y_i$.

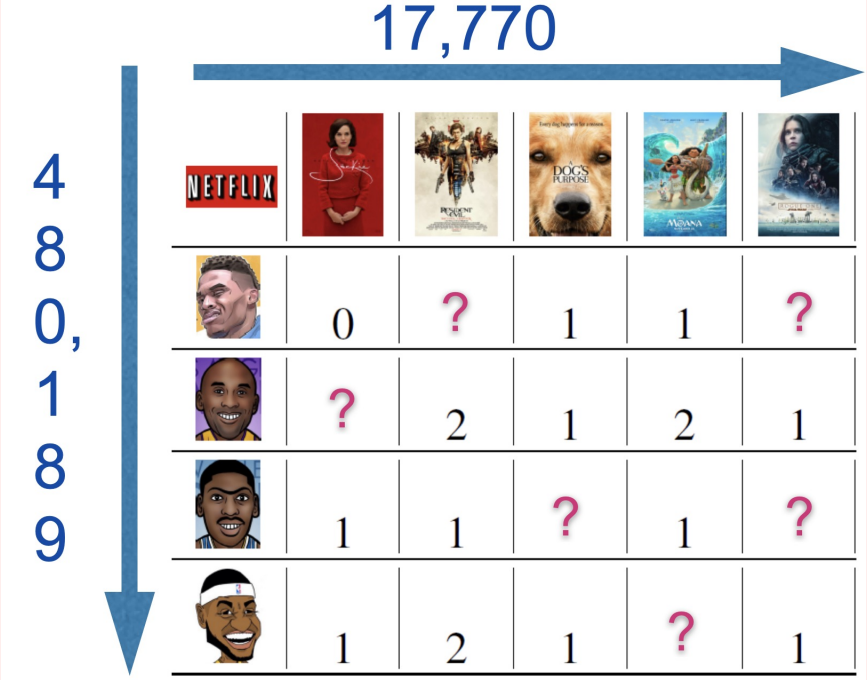


We can find the optimal x by solving the least squares problem.

$$\min_x \sum_{i=1}^{N} \left(y_i - x^T a_i\right)^2$$

Example 1.3

The Netflix prize: predict how a user will rate a movie.



- Some pattern exists: users do not assign ratings completely at random – if you like Godfather I, you'll probably like Godfather II.
- We have lots and lots of data: we know how a user has rated other movies, and we know how other users have rated this (and other) movies.
- Let y_{ij} denote the rate of user i for movie j.
- The Netflix price concerns finding a low-rank matrix X such that $x_{ij} = y_{ij}$ for observed (i, j), related to the following optimization problem

$$\min_X (\text{minimize}) \sum_{\text{observed}(i,j)} (x_{ij} - y_{ij})^2 \text{ subject to } \text{rank}(X) \leq r$$

or an alternative way is to

$$\min_X \text{ rank}(X) \text{ s.t. } x_{ij} = y_{ij} \text{ observed}(i, j)$$

§1.3 Classification of Optimizations

$$\begin{aligned} &\text{minimize } f(x) \\ &\text{subject to } c_i(x) = 0,\ i \in \mathcal{E} \\ &\qquad\qquad c_i(x) \geq 0, i \in \mathcal{I} \end{aligned}$$

- Unconstrained Optimization: $\mathcal{E} = \mathcal{I} = \emptyset$.
  - Many practical problems are unconstrained or the constraints can be safely discarded.
  - Unconstrained problems arise also as reformulation of constrained ones by replacing the constraints with penalization.
- Contrained Optimization: when constraints are essential for the problem.
- Linear Programming: when the objective function and all the constraints are linear function of x.
  - widely used in management, financial, and economic applications.
- Nonlinear Programming: at least some of the constraints or the objective are nonlinear functions.
  - tend to arise naturally in physical sciences, engineering, signal processing, and machine learning.
  - become more widely used in management and economic sciences as well.
- Global Optimization: aim at fining the global optimal solution, which is generally very challenging.
- Local Optimization: focuses instead on the computation and characterization of local solutions.
- Convex Optimization: the objective function is convex, the equality constraint functions are linear, and the inequality constraint functions are concave.

§2 Lec 2: Mar 31, 2021

§2.1 An Overview of Linear Algebra

Vector Spaces:

In "linear algebra", we denote a vector as a list of numbers,

$$\vec{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^n \text{ or } \mathbb{C}^n$$

The general definition of a vector space is as follows.

- A vector space S is composed of a set of elements (called "vectors") and members of a field R (called scalars).
  - Roughly, vectors are objects that can be added together and multiplied by numbers (namely, the scalars).
  - A field is a set of numbers for which addition and multiplication are will defined. We will typically use $R = \mathbb{R}$ or $R = \mathbb{C}$.
- In a vector space, there must be two rules defined for combining vectors and scalars.
  - The first operation is *vector addition*, which associates with any two vectors $\vec{x}, \vec{y} \in S$ the sum $\vec{x} + \vec{y}$ which also must belong to S.
  - The second operation is *scalar multiplication*, which associates with any vector $\vec{x} \in S$ and any scalar $a \in R$ the scalar multiple of $\vec{x}$ by a, denoted by $a\vec{x}$ or $a \cdot \vec{x}$, which must belong to S.
- The vector addition operation must obey four rules:
  1. Commutativity: $\vec{x} + \vec{y} = \vec{y} + \vec{x}\ \forall \vec{x}, \vec{y} \in S$.
  2. Associativity: $\vec{x} + (\vec{y} + \vec{z}) = (\vec{x} + \vec{y}) + \vec{z}\ \forall \vec{x}, \vec{y}, \vec{z} \in S$.
  3. There is a unique "zero vector" $\vec{0} \in S$ such that $\vec{x} + \vec{0} = \vec{x}\ \forall \vec{x} \in S$.
  4. For each $\vec{x} \in S$, there is a vector $-\vec{x} \in S$ such that $\vec{x} + (-\vec{x}) = \vec{0}$.
- The scalar multiplication operation must also obey four rules:
  1. Distributivity: $a\,(\vec{x} + \vec{y}) = a\vec{x} + a\vec{y}$ and $(a+b)\vec{x} = a\vec{x} + b\vec{x}\ \forall \vec{x}, \vec{y} \in S$ and $a, b \in R$.
  2. Associativity: $(ab)\vec{x} = a(b\vec{x})\ \forall \vec{x} \in S$ and $a, b \in R$.
  3. For the multiplicative identity of R (denoted by the scalar $1 \in R$), we have $1 \cdot \vec{x} = \vec{x} \in S$.
  4. For the additive identity of R (denoted by the scalar $0 \in R$), we have $0 \cdot \vec{x} = \vec{0} \in S$.

Linear Subspaces:

The concept of a linear subspace is useful for modeling, approximating signals, discussing the concept of bases,etc.

Definition 2.1 (Linear Subspace) — A nonempty subset T of a vector space S is called a subspace (or linear subspace) of S if

$$a\vec{x} + b\vec{y} \in T$$

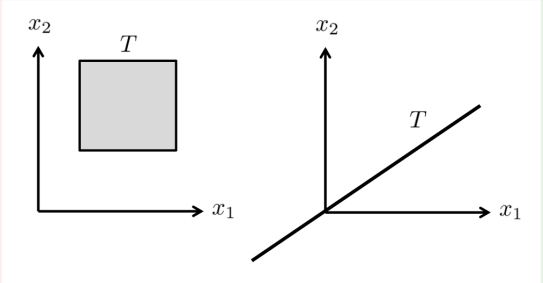
for all $\vec{x}, \vec{y} \in T$ and all $a, b \in R$.

Notes:

- Any linear subspace T must contain $\vec{0}$.
- Any vector space S is a linear subspace (of itself).
- Any linear subspace T meets all the properties of a vector space.

Example 2.2

Are either of these a subspace of $S = \mathbb{R}^2$



Left: No! Right: Yes!

Example 2.3

Which of the following are subspaces?

1. $S = \mathbb{R}^5$ and $T = \{\vec{x} \in \mathbb{R}^5 : x_4 = 0\}$ – Yes!
2. $S = \mathbb{R}^5$ and $T = \{\vec{x} \in \mathbb{R}^5 : x_4 = 1\}$ – No, add any two vectors in T and the sum will not belong to T.
3. $S = \mathbb{R}^5$ and T is the set of vectors in $\mathbb{R}^5$ with no more than 3 nonzero entries – No, can add certain vectors in T to get up to 5 nonzero elements.

Linear Combinations:

Linear combinations are used to build new vectors a weighted sum of other vectors.

Definition 2.4 (Linear Combination) — Let $M = \{\vec{v_1}, \vec{v_2}, \ldots, \vec{v_n}\}$ be a collection of vectors in a vector space S. (we will stick with finite collections at the moment).
A linear combination of vectors in M is a sum of the form

$$a_1\vec{v_1} + a_2\vec{v_2} + \ldots + a_n\vec{v_n}$$

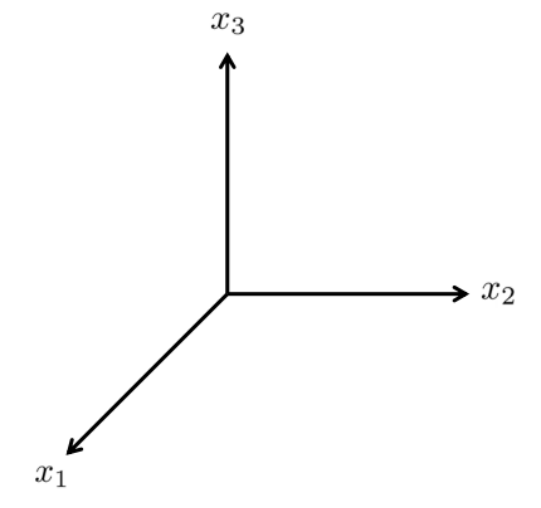
for some $a_1, a_2, \ldots, a_n \in R$. Since S is a vector space, this sum must belong to S.

Mentally, you might find it useful to replace "linear combination" with "weighted sum", although this is not standard terminology.

Definition 2.5 (Span) — Let $M = \{\vec{v_1}, \ldots, \vec{v_n}\}$ be a finite collection of vectors in a vector space S. The span of M, denoted by span(M) or span $\{\vec{v_1}, \ldots, \vec{v_n}\}$, is the set of all linear combinations of vectors in M.

Example 2.6

Consider the vectors $\vec{v_1} = \begin{bmatrix}1\\1\\0\end{bmatrix}$ and $\vec{v_2} = \begin{bmatrix}0\\1\\0\end{bmatrix}$. What is span $\{\vec{v_1}, \vec{v_2}\}$?



The entire (x_1, x_2)-plane.

When we are in $\mathbb{R}^n$ for some finite n, it is common to use matrix-vector notation as shorthand for linear combinations:

- Suppose $\vec{x} = c_1\vec{p_1} + \ldots + c_k\vec{p_k}$
- Then we may define the $n \times k$ matrix

$$A = \begin{bmatrix}\vec{p_1} & \vec{p_2} & \ldots & \vec{p_k}\end{bmatrix}$$

and the $k \times 1$ vector

$$\vec{c} = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_k \end{bmatrix}$$

and this allows us to write the $n \times 1$ vector $\vec{x}$ as $\vec{x} = A\vec{c}$.

Definition 2.7 (Linear Dependence) — A finite set of vectors $\vec{v_1}, \ldots, \vec{v_n}$ in a vector space S is said to be linearly dependent if there exists scalars $a_1, \ldots, a_n \in R$, not all equal to zero, such that

$$a_1\vec{v_1} + \ldots + a_n\vec{v_n} = \vec{0}$$

Definition 2.8 (Linear Independence) — A finite set of vectors $\vec{v_1}, \ldots, \vec{v_n}$ in a vector space S is said to be linearly independent if

$$a_1\vec{v_1} + \ldots + a_n\vec{v_n} = \vec{0}$$

only when all $a_k = 0$.

Every vector in the span of a linearly independent set of vectors has a unique expansion in terms of those vectors. This is formalized in the following lemma.

Lemma 2.9

Suppose $\vec{v_1}, \ldots, \vec{v_n}$ are linearly indep. and suppose

$$a_1\vec{v_1} + \ldots + a_n\vec{v_n} = b_1\vec{v_1} + \ldots + b_n\vec{v_n}$$

for some scalars $a_1, \ldots, a_n \in R$ and $b_1, \ldots, b_n \in R$. Then $a_k = b_k$ for $k = 1, 2, \ldots, n$.

Proof. Note that $\sum_{k=1}^{n} (a_k - b_k)\vec{v_k} = \vec{0}$. Since $\vec{v_1}, \ldots, \vec{v_n}$ are linearly indep., it must follow that $a_k - b_k = 0$ for all $k = 1, 2, \ldots, n$. □

Bases and Dimension:

Now that we know how to combine vectors (via linear combinations), let's think about important sets of vectors we'd be interested in combining.

Definition 2.10 (Basis) — A finite set of vectors $\vec{v_1}, \ldots, \vec{v_n}$ in a vector space S is said to form a basis for S if the following two conditions are satisfied:

1. $\vec{v_1}, \ldots, \vec{v_n}$ are linearly independent.
2. $\text{span}\{\vec{v_1}, \ldots, \vec{v_n}\} = S$.

A vector space wit a finite basis, as in the above definition, is said to be finite-dimensional. Any two bases for a finite-dimensional vector space contain the same number of elements. This leads to a meaningful definition of dimension.

Definition 2.11 (Dimension) —

- For a vector space S that can be spanned using a finite set of basis vectors, the dimension of S is the number of vectors required in any basis for S.
- For a vector space S that cannot be spanned using a finite set of basis vectors, the dimension of S is said to be infinite.

Example 2.12 (Bases for $S = \mathbb{R}^n$)

- The standard, or canonical, basis for $\mathbb{R}^n$ is given by:

$$\{\vec{v_1}, \vec{v_2}, \ldots, \vec{v_n}\} = \left\{ \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}, \ldots, \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix} \right\}$$

- Any set of n linearly indep. vectors in $\mathbb{R}^n$ forms a basis for $\mathbb{R}^n$.

§3 Lec 3: Apr 2, 2021

§3.1 Lec 2 (Cont'd)

Normed Linear Spaces:

A norm is a function used to measure the size of vectors in a vector space.

Definition 3.1 (Norm) — A norm $\|\cdot\|$ on a vector space S is a mapping $\|\cdot\| : S \to \mathbb{R}$ with the following properties:

1. $\|\vec{x}\| \geq 0$ for all $\vec{x} \in S$, and $\|\vec{x}\| = 0 \iff \vec{x} = \vec{0}$.
2. Triangle inequality: $\|\vec{x} + \vec{y}\| \leq \|\vec{x}\| + \|\vec{y}\|$ for all $\vec{x}, \vec{y} \in S$.
3. $\|a\vec{x}\| = |a| \cdot \|\vec{x}\|$ for any $a \in R$ and $\vec{x} \in S$.

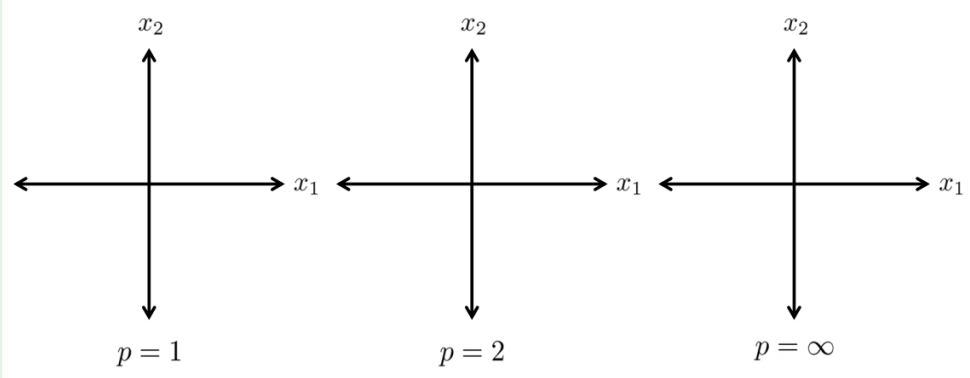
Definition 3.2 (Normed Linear Space) — A normed linear space is a vector space S together with a valid norm $\|\cdot\| : S \to \mathbb{R}$.

The l_p metrics for vectors in $\mathbb{R}^n$ extend naturally to l_p norms for these same spaces:

- l_1 norm: $\|\vec{x}\|_1 = \sum_{i=1}^{n} |x_i|$
- l_2 ("Euclidean") norm:
$$\|\vec{x}\|_2 = \left(\sum_{i=1}^{n} |x_i|^2 \right)^{\frac{1}{2}}$$
- l_p norm for $1 \leq p < \infty$:
$$\|\vec{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}}$$
- l_∞ norm:
$$\|\vec{x}\|_\infty = \max_{i=1,\ldots,n} |x_i|$$
- Bonus – l_0 norm (not really a norm – not satisfied the definition of a norm)
$$\|\vec{x}\|_0 = \#\text{ of nonzeros in } \vec{x}$$

Example 3.3

The "unit ball" in a normed linear space is the set of all vectors in S having norm less than or equal to 1. Suppose $S = \mathbb{R}^2$ and draw the l_p balls for $p = 1, 2, \infty$.



- $p = 1$: $\|\vec{x}\|_1 = |x_1| + |x_2| \leq 1$ – diamond.
- $p = 2$: $\|\vec{x}\|_2 = x_1^2 + x_2^2 \leq 1$ – circle.
- $p = \infty$: $\|\vec{x}\|_\infty = \max\{|x_1|, |x_2|\}$ – square.

Inner Product Spaces:

An inner product is a function used to compare two vectors in a vector space. The concept of an inner product will give us additional geometric structure beyond what is available in general normed linear spaces. In particular, using an inner product we can define a meaningful measure of the angle between two vectors, discuss orthonormal bases and orthogonal projections, etc.

Definition 3.4 (Inner Product) — An inner product $\langle \cdot, \cdot \rangle$ on a vector space S is a mapping $\langle \cdot, \cdot \rangle : S \times S \to R$ with the following properties:

1. $\langle \vec{x}, \vec{y} \rangle = \langle \vec{y}, \vec{x} \rangle^*$ for all $\vec{x}, \vec{y} \in S$.
2. For any $\vec{x}, \vec{y}, \vec{z} \in S$ and any $a, b \in R$, $\langle a\vec{x} + b\vec{y}, \vec{z} \rangle = a\langle \vec{x}, \vec{z} \rangle + b\langle \vec{y}, \vec{z} \rangle$.
3. For any $\vec{x} \in S$, $\langle \vec{x}, \vec{x} \rangle$ is real-valued and non-negative, and $\langle \vec{x}, \vec{x} \rangle = 0$ iff $\vec{x} = \vec{0}$.

Definition 3.5 (Inner Product Space) — An inner product space is a vector space S together with a valid inner product $\langle \cdot, \cdot \rangle : S \times S \to R$.

Definition 3.6 (Orthogonality) — Two vectors $\vec{x}$ and $\vec{y}$ in an inner product space S are said to be orthogonal if $\langle \vec{x}, \vec{y} \rangle = 0$.

Example 3.7

When $S = \mathbb{R}^n$, the standard inner product between two vectors $\vec{x}, \vec{y} \in S$ is given by

$$\langle \vec{x}, \vec{y} \rangle = \sum_{i=1}^{n} x_i y_i = \vec{y}^\top \vec{x}$$

The standard inner product on $\mathbb{R}^n$ is also known as the dot product.
$S = \mathbb{C}^n : \langle \vec{x}, \vec{y} \rangle = \vec{y}^H \vec{x} = \sum_{i=1}^{n} x_i y_i^*$ where H denotes conjugate transpose/hermitian.

Before we get to the connection between inner product and angles, it is worth noting that inner products can actually be used to measure the length of vectors (and thus distances between vectors as well).
In particular, any valid inner product induces a valid norm by

$$\|\vec{x}\| = \sqrt{\langle \vec{x}, \vec{x} \rangle}$$

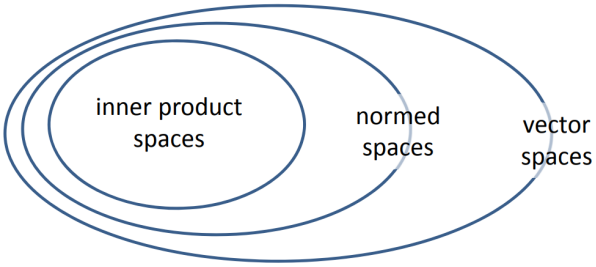
Example 3.8

When $S = \mathbb{R}^n$, the standard inner product induces the following norm:

$$\|\vec{x}\| = \sqrt{\langle \vec{x}, \vec{x} \rangle} = \sqrt{\sum_{i=1}^{n} x_i^2}$$

We can recognize this as the l_2 norm.

Other l_p norm (for $p \neq 2$)cannot be induced by inner products. Because every valid inner product induces a valid norm, every inner product space is also a normed linear space. But, not every normed linear space is also an inner product space:



Recall the definition of a basis in a generic, finite-dimensional vector space. In inner product spaces, a particularly useful class of bases are orthogonal bases.

Definition 3.9 (Orthogonal Basis) — A finite sets of non-zero vectors $\vec{v_1}, \ldots, \vec{v_n}$ in an inner product space S is said to form an orthogonal basis for S if the following two conditions are satisfied:

1. $\langle \vec{v_k}, \vec{v_l} \rangle = 0$ for all $k \neq l$ (note that this implies the vectors are linearly indep.)
2. $\text{span}\{\vec{v_1}, \ldots, \vec{v_n}\} = S$

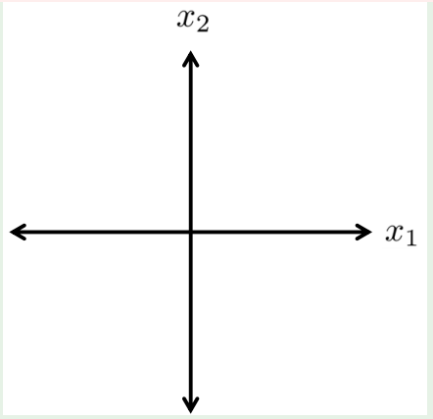
Definition 3.10 (Orthonormal Basis) — An orthogonal basis is called orthonormal basis or orthobasis if every basis vector $\vec{v_k}$ has unit norm (i.e., $\|\vec{v_k} = 1\|$) according to the induced norm in the inner product space.

Example 3.11

Using the standard inner product,

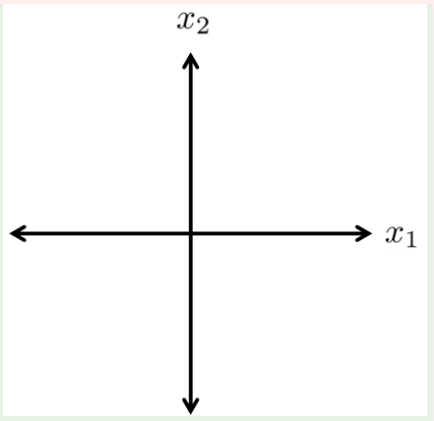
$$\vec{v_1} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \text{ and } \vec{v_2} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

form an orthobasis for $\mathbb{R}^2$.



Example 3.12 (Rotation of 45 degree from the last example)
Another possible orthobasis in $\mathbb{R}^2$ is given by

$$\vec{v_1} = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix} \text{ and } \vec{v_2} = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{bmatrix}$$



§3.2 Linear Operators

Operators are transformations that map vectors in some vector space to vectors in some other (possible different) vector space.

Definition 3.13 (Linear Operator) — Suppose X and Y are vector spaces. We say the operator $A : X \to Y$ is a linear operator if

$$A(\alpha_1 \vec{x}_1 + \alpha_2 \vec{x}_2) = \alpha_1 A\vec{x}_1 + \alpha_2 A\vec{x}_2$$

for all $\alpha_1, \alpha_2 \in R$ and $\vec{x}_1, \vec{x}_2 \in X$.

Fact 3.1. If $X = \mathbb{R}^n$ and $Y = \mathbb{R}^m$, any linear operator from X to Y can be represented as multiplication by an $m \times n$ matrix.

Therefore, a particularly interesting class of linear operators for us will simply be matrices.

§3.3 Operator Norms

Roughly speaking, operator norms help us talk about the "gain" of a system.

Definition 3.14 (Operator Norm) — Let X and Y be normed linear spaces with corresponding norms $\|\cdot\|_X$ and $\|\cdot\|_Y$ and suppose $A : X \to Y$ is linear operator. The operator norm $\|A\|$ is defined as

$$\|A\| := \sup_{\vec{0} \neq \vec{x} \in X} \frac{\|A\vec{x}\|_Y}{\|\vec{x}\|_X}$$

This is equivalent to

$$\|A\| := \sup_{\vec{x} \in X,\, \|\vec{x}\|_X = 1} \|A\vec{x}\|_Y$$

§4 | Lec 4: Apr 5, 2021

§4.1 Operator Norms (Cont'd)

Any operator norm as defined in the last lecture will satisfy the following properties

1. $\|A\| \geq 0$ with equality iff $A = 0$
2. $\|\alpha A\| = |\alpha|\|A\|$ for all $\alpha \in R$
3. $\|A + B\| \leq \|A\| + \|B\|$ for all linear operators A and B between the vector spaces X and Y.
   - Examining these first three properties, we see that $\|A\|$ is a valid norm on the vector space of linear operators!
4. $\|A\vec{x}\|_Y \leq \|A\|\|\vec{x}\|_X$ for all $\vec{x} \in X$
   - Therefore, the operator norm helps us bound how much an operator can "amplify" a signal.
5. $\|AB\| \leq \|A\|\|B\|$
6. If $X = Y$ and if $\|A\| < 1$, then we can write
$$\sum_{i=0}^{\infty} A^i = (I - A)^{-1}$$
just as for a scalar $a \in \mathbb{R}$ with $|a| < 1$, we can write $\sum_{i=0}^{\infty} a_i = \frac{1}{1-a}$

Let's restrict our attention to the special case where $X = \mathbb{R}^n$ and $Y = \mathbb{R}^m$. For these choices of X and Y, recall that any linear operator $A : X \to Y$ can be represented as multiplication by an $m \times n$ matrix. In such a case, the operator norm $\|A\|$ is also called a matrix norm.
When X and Y are both equipped with the l_p norm for $p \in [1, \infty]$, we can write

$$\|A\|_p \coloneqq \sup_{\vec{x} \in X,\, \|\vec{x}\|_p = 1} \|A\vec{x}\|_p$$

We can relate $\|A\|_p$ to certain properties of the matrix A:

- In the case $p = \infty$, letting $\vec{y} = A\vec{x}$, we have
$$\|A\|_\infty \coloneqq \sup_{\|\vec{x}\|_\infty = 1} \|\underbrace{A\vec{x}}_{\vec{y}}\|_\infty = \sup_{\|\vec{x}\|_\infty = 1} \left\| \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} \right\|_\infty$$
Note that $|y_i| = \left|\sum_{j=1}^{n} a_{ij} x_j\right|$. Overall $\vec{x}$ with $\|\vec{x}\|_\infty = 1$, the largest $|y_i|$ for a given $i = 1, 2, \ldots, m$ is achieved by taking $x_j = \text{sign}(a_{ij})$ for $j = 1, 2, \ldots, n$, and for this choice of i and $\vec{x}$, we will have
$$|y_i| = \sum_{j=1}^{n} |a_{ij}| = \text{absolute sum of row i of } A$$

Thus,

$$\|A\|_\infty = \max_{i=1,2,\ldots,m} \sum_{j=1}^{n} |a_{ij}| = \text{maximum absolute row sum of } A$$

- Similarly, in the case $p = 1$, we have

$$\|A\|_1 = \max_{j=1,2,\ldots,n} = \sum_{i=1}^{m} |a_{ij}| = \text{maximum absolute column sum of } A$$

- When $p = 2$, $\|A\|_2$ is also referred to as the spectral norm of A. We can understand $\|A\|_2$ geometrically: the operator A maps the l_2 unit ball in $\mathbb{R}^n$ to an ellipsoid in $\mathbb{R}^m$. The length of the major axis of the ellipsoid is equal to $\|A\|_2$. We can also write

$$\|A\|_2 = \sqrt{\lambda_{\max}(A^\top A)} = \sigma_{\max}(A)$$

where $\lambda_{\max}$ denotes the largest eigenvalue of a matrix and $\sigma_{\max}$ denotes the largest singular value of a matrix. If the matrix A happens to be symmetric (i.e. if $A = A^\top$) then we can also write

$$\|A\|_2 = \max_i |\lambda_i(A)|$$

There is also a special type of "matrix norm" that does not actually follow the definition of an operator norm,

$$\|A\| \coloneqq \sup_{\vec{x} \in X,\, \|\vec{x}\|_X = 1} \|A\vec{x}\|_Y$$

In particular, the Frobenius form of a matrix A is defined to be

$$\|A\|_F \coloneqq \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2} = \sqrt{\text{trace}(A^H A)}$$

Note that Frobenius form $\|A\|_F$ is not an operator norm, because $A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$.

$$\|A\|_F = \sqrt{2} > 1 = \sup_{\|x\|_X = 1} \|Ax\|_Y = \|x\|_X = 1$$

$$\|A\|_F = \sqrt{1+1} = \sqrt{2}$$

$$\|A\|_2 = \sup_{\|x\|_2 = 1} \|Ax\|_2 = 1$$

How about $A = \begin{bmatrix} 1 & 1 \\ 2 & 0 \end{bmatrix}$?

$$\|A\|_F = \sqrt{1+1+4+0} = \sqrt{6}$$

§4.2 Inverse Operator

Definition 4.1 (Invertibility) — A linear operator $A : X \to Y$ between two vector spaces X and Y is said to be invertible if there exists an operator $A^{-1} : Y \to X$ s.t.

- $AA^{-1} = I$, i.e., $AA^{-1}\vec{y} = \vec{y}$ for all $\vec{y} \in Y$ and
- $A^{-1}A = I$, i.e., $A^{-1}A\vec{x} = \vec{x}$ for all $\vec{x} \in X$.

In such a case, A^{-1} is referred to as the inverse of A.

Lemma 4.2

If A is an invertible linear operator, A^{-1} is itself a linear operator.

Invertibility is a topic of interest when we want to find an exact solution to a linear equation. Let us again restrict our attention to the special case where $X = \mathbb{R}^n$ and $Y = \mathbb{R}^m$. For these choices of X and Y, recall that any linear operator $A : X \to Y$ can be represented as multiplication by an $m \times n$ matrix.

Fact 4.1. An $m \times n$ matrix A cannot be invertible unless it is square.

Not all square matrices are invertible. An invertible matrix is also known as the nonsingular matrix.

Proposition 4.3

If A is a square matrix, the following statements are all equivalent

- A is invertible
- A is nonsingular
- $\det(A) \neq 0$
- $A\vec{x} = \vec{0} \iff \vec{x} = \vec{0}$
- The rows of A are linearly indep.
- The columns of A are linearly indep.
- $\dim(\mathcal{N}(A)) = 0$, i.e., $\mathcal{N} = \left\{\vec{0}\right\}$
- $\dim(\mathcal{R}(A)) = n$.
- A is full rank
- All eigenvalues of A are nonzero
- The matrix $A^\top A$ is positive definite.
- $A^\top$ is invertible.

§4.3 Adjoint Operators

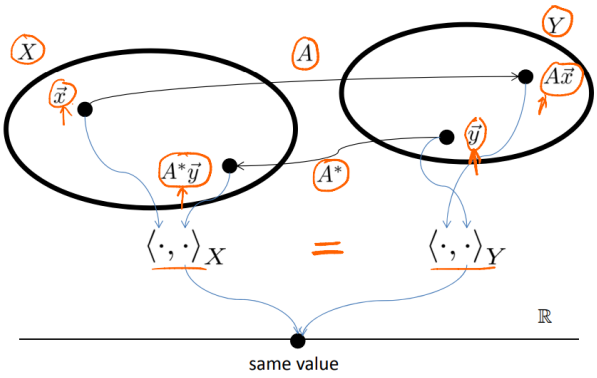
Definition 4.4 (Adjoint of an Operator) — Let $A : X \to Y$ be a bounded ($\|A\| < \infty$) linear operator between two inner product spaces X and Y. The adjoint of A, denoted $A^* : Y \to X$ is the unique operator such that

$$\langle A\vec{x}, \vec{y}\rangle_Y = \langle \vec{x}, A^*\vec{y}\rangle_X$$

for all $\vec{x} \in X$ and $\vec{y} \in Y$.

Definition 4.5 (Self-Adjoint Operator) — An operator $A : X \to X$ is said to be self-adjoint if $A = A^*$.

An illustration:



Lemma 4.6

If A is bounded linear operator with adjoint A^* then A^* is itself a bounded linear operator, and

$$\|A^*\| = \|A\|$$

Lemma 4.7

If A is bounded linear operator with adjoint A^* then $(A^*)^* = A$

Lemma 4.8

If A is an invertible bounded linear operator with adjoint A^* and bounded inverse A^{-1}, then

$$(A^{-1})^* = (A^*)^{-1}$$

In the special case where $X = \mathbb{R}^n$ and $Y = \mathbb{R}^m$, recall that any linear operator $A : X \to Y$ can be represented as multiplication by an $m \times n$ matrix.

Question 4.1. What is the adjoint of a matrix?

Using the standard inner product on X and Y, the adjoint of A is the unique operator s.t.

$$\vec{y}^H A\vec{x} = \langle A\vec{x}, \vec{y}\rangle = \langle \vec{x}, A^*\vec{y}\rangle = \vec{y}^H (A^*)^H \vec{x}$$

for all $\vec{x}$ and $\vec{y}$. This requires that $A = (A^*)^H$ which is satisfied by taking

$$A^* = A^H$$

Therefore, the adjoint of a matrix is simply its conjugate transpose (not its inverse). Self-adjoint matrices satisfy $A = A^H$. These are also known as symmetric (if real), conjugate symmetric or Hermitian (if complex).

Theorem 4.9

Let A be a real-valued $m \times n$ matrix. For a fixed $\vec{y} \in \mathbb{R}^m$, the vector $\vec{x} \in \mathbb{R}^n$ is a minimizer of $\|\vec{y} - A\vec{x}\|_2 \iff$

$$A^\top A\vec{x} = A^\top \vec{y}$$

If $A^\top A$ is invertible, then the unique minimizer of $\|\vec{y} - A\vec{x}\|_2$ is given by

$$\vec{x} = (A^\top A)^{-1} A^\top \vec{y}$$

The same theorem holds if $A, \vec{x}$ and $\vec{y}$ are all complex-valued.

§5 Lec 5: Apr 7, 2021

§5.1 Fundamental Subspaces of Linear Operators

Definition 5.1 (Range) — Let $A : X \to Y$ be a linear operator between two vector spaces X and Y. The range or range space of A, denoted by $\mathcal{R}(A)$, is defined to be

$$\mathcal{R}(A) \coloneqq \{\vec{y} \in Y : A\vec{x} = \vec{y} \text{ for some } \vec{x} \in X\}$$

The range space is a linear subspace of Y.

Definition 5.2 (Nullspace) — Let $A : X \to Y$ a linear operator between two vector spaces X and Y. The nullspace of A, denoted by $\mathcal{N}(A)$, is defined to be

$$\mathcal{N}(A) \coloneqq \left\{\vec{x} \in X : A\vec{x} = \vec{0}\right\}$$

The nullspace is a linear subspace of X.

Again, we consider the case where $X = \mathbb{R}^n$ and $Y = \mathbb{R}^m$ (or where $X = \mathbb{C}^n$ and $Y = \mathbb{C}^m$) and $A : X \to Y$ can be represented as multiplication by an $m \times n$ matrix.
When A is a matrix, $\mathcal{R}(A)$ is just the span of the columns of A:

$$\mathcal{R}(A) = \text{colspan}(A)$$

For any $m \times n$ matrix A,

$$\text{rank}(A) \leq \min\{m, n\}$$

We say that A is full rank if $\text{rank}(A) = \min\{m, n\}$, otherwise we call it rank deficient.
We can related the rank of A to the dimensions of the four fundamental subspaces of A:

- $\dim(\mathcal{R}(A)) = \dim(\text{colspan}(A)) = \text{rank}(A)$
- $\dim(\mathcal{N}(A)) = n - \text{rank}(A)$
- $\dim(\mathcal{R}(A^*)) = \dim(\text{rowspan}(A)) = \text{rank}(A)$
- $\dim(\mathcal{N}(A^*)) = m - \text{rank}(A)$
- $\dim(\mathcal{R}(A)) + \dim(\mathcal{N}(A)) = n = \#$ columns of A.

For two matrices A and B, we have

- $\text{rank}(AB) \leq \min\{\text{rank}(A), \text{rank}(B)\}$
- $\text{rank}(A + B) \leq \text{rank}(A) + \text{rank}(B)$

§5.2 Projection Operators

Definition 5.3 (Projection Operator) — A linear operator $P : X \to X$ from a vector space X into itself is called a projection or a projection operator if

$$P^2 = P$$

i.e., $P(P(\vec{x})) = P(\vec{x})$ for all $\vec{x} \in X$.

For $P^2 = P$, P is called idempotent operator.

Definition 5.4 (Orthgonal Projection Operator) — A projection operator P in an inner product space X is called an orthogonal projection or an orthogonal projection operator if

$$\mathcal{R}(P) \perp \mathcal{N}(P)$$

i.e., if $\langle \vec{x}, \vec{y} \rangle = 0$ for all $\vec{x} \in \mathcal{R}(P)$ and $\vec{y} \in \mathcal{N}(P)$.

We notice

$$\vec{x} = \underbrace{P\vec{x}}_{\in \mathcal{R}(P)} + \underbrace{(I - P)\vec{x}}_{\in \mathcal{N}(P)}$$

If P is an orthogonal projection operator,

$$\langle P\vec{x}, (I - P)\vec{x} \rangle = 0$$

Lemma 5.5

A bounded linear operator $P : X \to X$ on an inner product space X is an orthogonal projection iff

1. $P^2 = P$ and
2. $P = P^*$

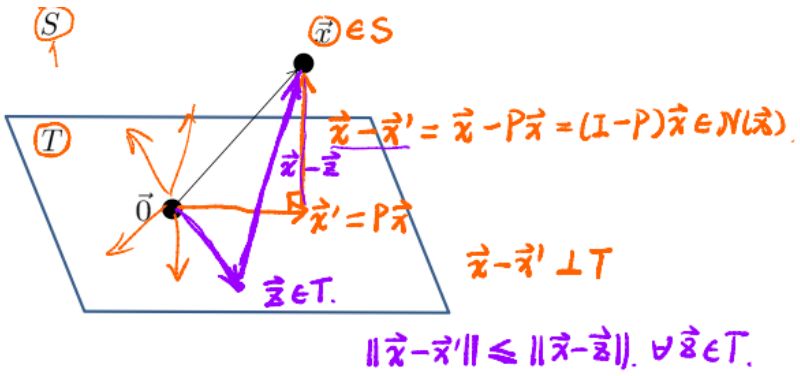
Theorem 5.6

Suppose S is an inner product space and suppose T is a linear subspace of S. For a given vector $\vec{x} \in S$, there is a unique vector $\vec{x}' \in T$ such that $\|\vec{x} - \vec{x}'\| \leq \|\vec{x} - \vec{z}\|$ for all $\vec{z} \in T$. Furthermore, this minimizer has the property that

$$\vec{x} - \vec{x}' \perp T$$

i.e. $\langle \vec{x} - \vec{x}', \vec{y} \rangle = 0$ for all $\vec{y} \in T$.

The minimizing vector $\vec{x}'$ is referred to as the orthogonal projection of $\vec{x}$ onto T. In other words, $\vec{x}' = P\vec{x}$, where P is an orthogonal projection operator with $\mathcal{R}(P) = T$.



For an $n \times n$ matrix P,

- P is a projection if $P^2 = P$
- P is orthogonal projection if $P^2 = P$ and $P^\top = P$

Example 5.7

Consider the operator $P : \mathbb{R}^2 \to \mathbb{R}^2$ defined as

$$P\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = \begin{bmatrix} x_1 \\ x_1 \end{bmatrix}$$

We can express P as the 2×2 matrix $P = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} x_1 \\ x_1 \end{bmatrix}$ Consider:

- Does $P^2 = P$? Yes!
- Does $P^\top = P$? No!
- What is $\mathcal{R}(P)$? the line $x_2 = x_1$
- What is $\mathcal{N}(P)$? x_2-axis
- Is $\mathcal{R}(P) \perp \mathcal{N}(P)$? No!

Example 5.8

Consider the operator $P : \mathbb{R}^3 \to \mathbb{R}^3$ defined as

$$P\left(\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}\right) = \begin{bmatrix} x_1 \\ x_2 \\ 0 \end{bmatrix}$$

where

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

Consider:

- Does $P^2 = P$? Yes!
- Does $P^\top = P$? Yes!
- What is $\mathcal{R}(P)$? the plane $x_1 - x_2$
- What is $\mathcal{N}(P)$? x_3-axis
- Is $\mathcal{R}(P) \perp \mathcal{N}(P)$? Yes!

Consider a set of m linearly indep. vectors $\vec{v_1}, \vec{v_2}, \ldots, \vec{v_m} \in \mathbb{R}^n$ or $\mathbb{C}^n$. We can construct an orthogonal projection matrix P onto the subspace $T = \text{span}\{\vec{v_1}, \vec{v_2}, \ldots, \vec{v_m}\}$ as follows:

1. Construct an $n \times m$ matrix

   $$A = \begin{bmatrix} \vec{v_1} & \vec{v_2} & \ldots & \vec{v_m} \end{bmatrix}$$

   Note that colspan(A) = T.

2. Let

   $$P = A\left(A^\top A\right)^{-1} A^\top = AA^\dagger$$

Example 5.9

Consider the vectors

$$\vec{v_1} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \text{ and } \vec{v_2} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

The orthogonal projection can be constructed as follows

$$P = AA^\dagger = A(A^\top A)^{-1}A^\top = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

§6 Lec 6: Apr 9, 2021

§6.1 Motivating Examples

Consider a 2×2 matrix

$$A = \begin{bmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{bmatrix}$$

For what nonzero vectors $\vec{x} \in S$ (eigenvectors) and scalar $\lambda \in \mathbb{C}$ (eigenvalue), do we have that $A\vec{x} = \lambda x$?

- We know that $\vec{x}$ and λ must satisfy $(A - \lambda I)\vec{x} = \vec{0}$.
- Thus, $\vec{x}$ must be in $\mathcal{N}(A - \lambda I)$.
- Thus, $A - \lambda I$ must have a nontrivial nullspace.
- Thus, $A - \lambda I$ must be singular.
- We can solve for λ s.t. $\det(A - \lambda I) = 0$:

  $$\det(A - \lambda I) = \det\left(\begin{bmatrix} 1-\lambda & \frac{1}{2} \\ \frac{1}{2} & 1-\lambda \end{bmatrix}\right) = (1-\lambda)^2 - \frac{1}{4} = 0$$

  which equals 0 for $\lambda = 1.5$ or $\lambda = 0.5$
- Now we know the eigenvalues. What are the corresponding eigenvectors?
- For $\lambda = 1.5$, we need $A\vec{x} = 1.5\vec{x}$.

  $$\begin{bmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1.5x_1 \\ 1.5x_2 \end{bmatrix}$$

  which requires $x_1 = x_2$. To have unit norm, we can choose

  $$\vec{x} = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$$

  but notice that any rescaling of this $\vec{x}$ is also an eigenvector.
- For $\lambda = 0.5$, we need $A\vec{x} = 0.5\vec{x}$.

  $$\begin{bmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0.5x_1 \\ 0.5x_2 \end{bmatrix}$$

  This requires $x_1 = -x_2$. To have unit norm, we can choose

  $$\vec{x} = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{bmatrix}$$

  but notice that any rescaling of this $\vec{x}$ is also an eigenvector.
- We say that the eigenvectors of A are

  $$\vec{v}_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix} \text{ and } \vec{v}_2 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{bmatrix}$$

  with the understanding that any rescaling of either one is also an eigenvector.

§6.2 Eigenvalues and Eigenvectors

Big picture:

- first find the eigenvalues
  - Suppose A is an $n \times n$ matrix.
  - We want to know: for what values of $\lambda \in \mathbb{C}$ does there exist a non-zero $\vec{x} \in \mathbb{C}^n$ s.t.
    $$A\vec{x} = \lambda\vec{x}?$$
  - We know that for such a $(\lambda, \vec{x})$ pair to exist, we must have
    $$(A - \lambda I)\,\vec{x} = \vec{0}$$
    and therefore, $\vec{x} \in \mathcal{N}(A - \lambda I)$.
  - For such a non-zero $\vec{x}$ to exist in the nullspace of $A - \lambda I$, we need
    $$\dim\left(\mathcal{N}(A - \lambda I)\right) > 0$$
    Since
    $$\dim\left(\mathcal{R}(A - \lambda I)\right) + \dim\left(\mathcal{N}(A - \lambda I)\right) = n = \#\text{ columns of } A - \lambda I$$
    which means we require
    $$\dim\left(\mathcal{R}(A - \lambda I)\right) < n$$
    so $A - \lambda I$ cannot be full rank. We also know this requires
    $$\det\left(A - \lambda I\right) = 0$$
  - Hence, by finding all λ s.t. $\det\left(A - \lambda I\right) = 0$, we get the eigenvalues of A.
- the find the eigenvectors
  - Suppose λ is an eigenvalue of A, which has size $n \times n$.
  - Then every $\vec{x} \in \mathcal{N}(A - \lambda I)$ is considered an eigenvector of A, corresponding to the eigenvalue λ.
  - Because $\mathcal{N}(A - \lambda I)$ is a linear subspace of $\mathbb{C}^n$, we conventionally just specify enough vectors to span this subspace, i.e., a basis for $\mathcal{N}(A - \lambda I)$.

Let's work through an example.

Example 6.1

Let

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

Solving $\det(A - \lambda I) = 0$, we want

$$\det(A - \lambda I) = (2 - \lambda)\left(\left(\frac{3}{2} - \lambda\right)^2 - \frac{1}{4}\right) = 0$$

which is satisfied if $\lambda = 2, 1$.

- For $\lambda = 1$ (with multiplicity one)
  - We want to find all $\vec{x} \in \mathbb{C}^3$ s.t.

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

  - This requires that $x_1 = 0$ and that $\frac{x_2}{2} + \frac{x_3}{2} = 0$. We have two linear equations, which imply that $\mathcal{N}(A - \lambda I)$ is a line in $\mathbb{C}^3$.
  - So we can pick $\vec{v}_1 = \begin{bmatrix} 0 \\ \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{bmatrix}$ as our first eigenvector.
- For $\lambda = 2$ (with multiplicity two)
  - We want to find $\vec{x} \in \mathbb{C}^3$ s.t.

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

  - This requires that $\frac{-x_2}{2} + \frac{x_3}{2} = 0 \implies x_2 = x_3$. We have one linear equation, which implies that $\mathcal{N}(A - \lambda I)$ is a plane in $\mathbb{C}^3$.
  - So we can pick two linearly indep. vectors from this nullspace, e.g.,

$$\vec{v}_2 = \begin{bmatrix} 0 \\ \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix} \text{ and } \vec{v}_3 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

  - Thus, for this particular matrix A, we have two distinct eigenvalues but three linearly indep. eigenvectors.

In the case where the eigenvalues of the matrix are distinct, we have an important result:

Theorem 6.2

Eigenvectors corresponding to distinct eigenvalues are linearly independent.

Proof. Left for readers to figure out on your own :) (just kidding, I guess I am just lazy). □

Eigenvectors corresponding to repeated eigenvalues could be linearly indep.

Example 6.3

Consider

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \lambda_1 = \lambda_2 = 1, \vec{v}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \vec{v}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

in which case $A - \lambda I$ has a two-dimensional nullspace.

Example 6.4

Consider:

$$A = \begin{bmatrix} 4 & 2 \\ 0 & 4 \end{bmatrix}, \lambda_1 = \lambda_2 = 4, \vec{v}_1 = \vec{v}_2 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

in which case $A - \lambda I$ has a one-dimensional nullspace.

§7 | Lec 7: Apr 12, 2021

§7.1 Diagonalization

If an $n \times n$ matrix A happens to have n linearly indep. eigenvectors, then it can be written (or "diagonalized") as

$$A = T\Lambda T^{-1}$$

where

- T is an $n \times n$ invertible matrix
- Λ is an $n \times n$ diagonal matrix

Construction:

- Let $\lambda_1, \lambda_2, \ldots, \lambda_n$ be the eigenvalues of A (not necessarily distinct)
- Let $\vec{v_1}, \vec{v_2}, \ldots, \vec{v_n}$ be the corresponding eigenvectors (not necessarily linearly indep.)
- For all $i = 1, 2, \ldots, n$, we know that $A\vec{v_i} = \lambda_i \vec{v_i}$.
- We can stack these n equations in the form of a matrix equation:

$$A \begin{bmatrix} \vec{v_1} & \vec{v_2} & \ldots & \vec{v_n} \end{bmatrix} = \begin{bmatrix} \lambda_1\vec{v_1} & \lambda_2\vec{v_2} & \ldots & \lambda_n\vec{v_n} \end{bmatrix}$$

  that is,

$$A \underbrace{\begin{bmatrix} \vec{v_1} & \vec{v_2} & \ldots & \vec{v_n} \end{bmatrix}}_{T} = \underbrace{\begin{bmatrix} \vec{v_1} & \vec{v_2} & \ldots & \vec{v_n} \end{bmatrix}}_{T} \underbrace{\begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix}}_{\Lambda}$$

- Because the $\vec{v_1}, \vec{v_2}, \ldots, \vec{v_n}$ are linearly indep., then T must be invertible. Then

$$A = T\Lambda T^{-1}$$

Lemma 7.1 (Eigenvalues of a Hermitian Matrix)
If $A = A^H$, then all eigenvalues of A are real-valued (even if A has complex entries).

Proof. Let λ be an eigenvalue of A and let $\vec{x}$ be an eigenvector corresponding to λ. Then

$$\langle A\vec{x}, \vec{x}\rangle = \langle \lambda\vec{x}, \vec{x}\rangle = \lambda\langle \vec{x}, \vec{x}\rangle$$

But also,

$$\langle \vec{x}, A\vec{x}\rangle = \langle \vec{x}, \lambda\vec{x}\rangle = \lambda^*\langle \vec{x}, \vec{x}\rangle$$

Since $A = A^H$, then $\langle A\vec{x}, \vec{x}\rangle$ must equal $\langle \vec{x}, A\vec{x}\rangle$, and so this implies that $\lambda = \lambda^*$. Thus, λ is real. □

This lemma does not mean that all the eigenvalues must be distinct (only that they must be real). So what can we say about the eigenvectors? Will they be linearly indep.?

Lemma 7.2 (Eigenvectors of a Hermitian Matrix)
If $A = A^H$, then there exists a set of n orthonormal eigenvectors $\vec{v_1}, \vec{v_2}, \ldots, \vec{v_n}$ s.t.

$$A\vec{v_i} = \lambda_i \vec{v_i}$$

for all $i = 1, 2, \ldots, n$.

This result holds even if there are repeated eigenvalues, but it uses the assumption that $A = A^H$.
Let A be an $n \times n$ matrix and suppose $A = A^H$. Then choosing an orthonormal set of eigenvectors $\vec{v_1}, \vec{v_2}, \ldots, \vec{v_n}$ and letting $T = [\vec{v_1}, \vec{v_2}, \ldots, \vec{v_n}]$ as before, we have

$$A = T\Lambda T^{-1}$$

However, since the $\{\vec{v_i}\}$ are orthonormal, then T s unitary. Therefore, $T^{-1} = T^H$ and so

$$A = T\Lambda T^H$$

Note: If A is real, it is possible to choose T real and have $A = T\Lambda T^\top$.

Example 7.3
Let

$$A = A^H = \begin{bmatrix} \frac{3}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{3}{2} \end{bmatrix}$$

Then $\lambda_1 = 2$ and $\lambda_2 = 1$, both of which are real since $A = A^H$. We can derive

$$\vec{v_1} = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix} \text{ and } \vec{v_2} = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{bmatrix}$$

Thus, $A = T\Lambda T^H$, where

$$T = \begin{bmatrix} \vec{v_1} & \vec{v_2} \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}, \quad \Lambda = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$$

§7.2 Positive Definite Matrices

Let A be an $n \times n$, Hermitian, symmetric matrix. Recall that we say A is postiive definite if

$$\vec{x}^H A \vec{x} > 0, \quad A \succ 0$$

holds for all non-zero $\vec{x} \in \mathbb{R}^n$ (or $\mathbb{C}^n$). Similarly, we say that A is positive semi-definite if

$$\vec{x}^H A \vec{x} \geq 0, \quad A \succeq 0$$

holds for all non-zero $\vec{x} \in \mathbb{R}^n$ (or $\mathbb{C}^n$). Such matrices are called symmetric, positive (semi-) definite.
If $A = A^H$, we already know the eigenvalues of A are real. Furthermore, if A is positive definite, then all eigenvalues of A are positive.

Proof. Let $\vec{v}$ be an eigenvector of A and let λ be the corresponding eigenvalue. Assume $\vec{v} \neq \vec{0}$. Then, because A is positive definite,

$$\vec{v}^H \underbrace{A\vec{v}}_{\lambda\vec{v}} > 0$$

Substituting,

$$\vec{v}^H(\lambda\vec{v}) > 0 \implies \lambda\vec{v}^H\vec{v} > 0 \implies \lambda > 0$$

because $\|\vec{v}\| > 0$. Similarly, if A is positive semi-definite, then all eigenvalues of A are non-negative. □

Positive definite matrices can be used to define variations on the standard l_2 inner product. In particular, suppose A is a symmetric, positive definite matrix. Then

$$\langle \vec{x}, \vec{y} \rangle_A := \vec{y}^H A\vec{x}$$

defines a valid inner product on $\mathbb{C}^n$. Consequently,

$$\|\vec{x}\|_A = \sqrt{\langle \vec{x}, \vec{x} \rangle_A} = \sqrt{\vec{x}^H A\vec{x}}$$

defines a valid induced norm on $\mathbb{C}^n$.
Consider the optimization problems

$$\max_{\vec{x}\in\mathbb{C}^n} \frac{\|\vec{x}\|_A^2}{\|\vec{x}\|_2^2} = \max_{\substack{\vec{x}\in\mathbb{C}^n \\ \|\vec{x}\|_2=1}} \vec{x}^H A\vec{x}$$

and

$$\min_{\vec{x}\in\mathbb{C}^n} \frac{\|\vec{x}\|_A^2}{\|\vec{x}\|_2^2} = \min_{\substack{\vec{x}\in\mathbb{C}^n \\ \|\vec{x}\|_2=1}} \vec{x}^H A\vec{x}$$

The maximum value of the first problem is given by $\lambda_{\max}(A)$ and occurs when $\vec{x}$ equals the corresponding eigenvector of A. Similarly, the minimum value of the second problem is given by $\lambda_{\min}(A)$ and occurs when $\vec{x}$ equals the corresponding eigenvector of A.

Proof. Refer to the lecture note. □

Recall the 2-norm (spectral norm) of a matrix A:

$$\|A\|_2 = \sup_{\substack{\vec{x}\in\mathbb{C}^n \\ \|\vec{x}\|_2=1}} \|A\vec{x}\|_2$$

Note that

$$\|A\vec{x}\|_2 = \sqrt{(A\vec{x})^H A\vec{x}} = \sqrt{\vec{x}^H A^H A\vec{x}} = \|\vec{x}\|_{A^H A}$$

For any matrix A, it turns out that $A^H A$ is positive semi-definite. Thus it follows that

$$\|A\|_2 = \sup_{\|\vec{x}\|_2=1} \|\vec{x}\|_{A^H A} = \sqrt{\lambda_{\max}(A^H A)}$$

Now, consider the special case where $A = A^H$. In this case,

$$\|\vec{x}\|_{A^H A} = \|\vec{x}\|_{A^2}$$

Also, since $A = T\Lambda T^H$ with T unitary, then $A^2 = T\Lambda^2 T^H$, and so $(\lambda_i(A))^2 = \lambda_i(A^2)$. Thus, when $A = A^H$ we have

$$\|A\|_2 = \sup_{\|\vec{x}\|_2=1} \|\vec{x}\|_{A^2} = \sqrt{\lambda_{\max}(A^2)} = \max_i |\lambda_i(A)|$$

Similarly, it follows that

$$\|A^{-1}\|_2 = \frac{1}{\min_i |\lambda_i(A)|}$$

Notice

$$A = T\Lambda T^H \implies A^{-1} = T\Lambda^{-1}T^H \implies \lambda_i(A^{-1}) = \frac{1}{\lambda_i(A)}$$

$$\|A^{-1}\|_2 = \max_i \left|\lambda_i(A^{-1})\right| = \max_i \frac{1}{|\lambda_i(A)|} = \frac{1}{\min_i |\lambda_i(A)|}$$

§8 Lec 8: Apr 14, 2021

§8.1 Some Properties of Eigenvalues

Let A be an $n \times n$ matrix. Then,

- $\det(A) = \prod_{i=1}^{n} \lambda_i(A)$
- $\text{trace}(A) = \sum_{i=1}^{n} \lambda_i(A)$

Rank:

- If A is not full rank, at least one of its eigenvalues must equal 0.
- If $A = A^H$, we have seen that we can write $\underbrace{A = U\Lambda U^H}_{T\Lambda T^H}$ for some unitary U.
  - $\text{rank}(A) = \#$ nonzero eigenvalues of A.
  - Writing $U = \begin{bmatrix} \vec{u_1} & \vec{u_2} & \dots & \vec{u_n} \end{bmatrix}$, we have

$$\begin{aligned} A &= U\Lambda U^H \\ &= \begin{bmatrix} u_1 & \dots & u_n \end{bmatrix} \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix} \begin{bmatrix} u_1^H \\ \vdots \\ u_n^H \end{bmatrix} \\ &= \sum_{i=1}^{n} \lambda_i \vec{u_i} \vec{u_i}^H \end{aligned}$$

If A is not full rank, some of the terms in this summation equal 0.

§8.2 Singular Value Decomposition

We've seen that a square Hermitian matrix A can be factored as

$$A = U\Lambda U^H$$

where U is orthonormal and Λ is diagonal.
The SVD allows us to generalize this type of factorization to any matrix, even those that are not square. Let A be an $m \times n$ matrix, with real or complex entries. Then A can be factored as

$$A = U\Sigma V^H \leftarrow \text{ SVD}$$

where

- U is $m \times n$ and orthonormal.
- Σ is $m \times n$ and diagonal.
- V is $n \times n$ and orthonormal.

Let $p = \min(m, n)$. Then $m \times n$ diagonal matrix Σ has the form

$$\begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & 0 \\ & & \ddots & \\ 0 & & & \sigma_p \end{bmatrix} \text{ if } m \leq n$$

or

$$\begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & 0 \\ & & \ddots & \\ 0 & & & \sigma_p \\ & & & \end{bmatrix} \text{ if } m \geq n$$

Terminology: The elements $\sigma_1, \sigma_2, \ldots, \sigma_p$ are known as the singular values of A. For any matrix A, the singular values are always real and non-negative. Thus, it is customary to order them as follows:

$$\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_p \geq 0$$

The columns of U are known as the left singular vectors. The columns of V are known as the right singular vectors.

The SVD an eigenvalue decomposition are closely related. For any $m \times n$ matrix A with SVD given by $A = U\Sigma V^H$, consider the matrix

$$A^H A = (U\Sigma V^H)^H \left(U\Sigma V^H\right) = V\Sigma^H \underbrace{U^H U}_{=I} \Sigma V^H = V \underbrace{\Sigma^H \Sigma}_{=\Sigma^2} V^H = V\Lambda V^H$$

Let $\Lambda = \Sigma^H \Sigma = \Sigma^2$. Then Λ is an $n \times n$ diagonal matrix with the following entries along the main diagonal

$$\lambda_i = \begin{cases} \sigma_i^2, & i \leq p \\ 0, & i > p \end{cases}$$

Thus, we see that the singular values of A are the square root of the eigenvalues of $A^H A$, and the right singular vectors of A are the eigenvectors of $A^H A$. Similar statements can be made for AA^H, since

$$AA^H = \left(U\Sigma V^H\right)\left(U\Sigma V^H\right)^H = U\Sigma V^H V\Sigma^H U^H = U\Sigma\Sigma^H U^H = U\Sigma^2 U^H = U\Lambda U^H$$

Thus, we see that the singular values of A are the square roots of the eigenvalues of AA^H, and the left singular vectors of A are the eigenvectors of AA^H.

An $m \times n$ matrix A can have at most $p = \min(m, n)$ nonzero singular values. Suppose that a matrix A has fewer than p nonzero singular values. In other words, suppose for some $r < p$ that

$$\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > 0$$

but that

$$\sigma_{r+1} = \sigma_{r+2} = \ldots = \sigma_p = 0$$

That is, A has only r nonzero singular values. Then, we can write

$$\Sigma = \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix}$$

where Σ_1 is $r \times r$ and diagonal (with non-zeros on the diagonal, and thus it is invertible), and Σ_2 is $(m-r) \times (n-r)$ and all zeros.
We can similarly partition

$$U = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \text{ and } V^\top = \begin{bmatrix} V_1^\top \\ V_2^\top \end{bmatrix}$$

and note that

$$U_1^\top U_1 = I, \quad V_1^\top V_1 = I, \quad U_1^\top U_2 = 0, \quad V_1^\top V_2 = 0$$

This allows us to write the "compact" or "reduced" form of the SVD

$$A = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix} \begin{bmatrix} V_1^H \\ V_2^H \end{bmatrix} = U_1 \Sigma_1 V_1^H$$

where U_1 is $m \times r$ with orthonormal columns, Σ_1 is diagonal with positive real entries along the diagonal, and V_1 is $n \times r$ with orthonormal columns. Equivalently, we can write

$$A = \sum_{i=1}^{r} \sigma_i \vec{u_i} \vec{v_i}^\top$$

which decomposes A as a sum of r rank-1 matrices.
The SVD reveals the rank of the matrix A as

$$\text{rank}(A) = r = \#\text{ of nonzero singular values of } A$$

The SVD also reveals the four fundamental subspaces of A

- $\mathcal{R}(A) = \text{colspan}(U_1)$
- $\mathcal{N}(A) = \text{colspan}(V_2)$
- $\mathcal{R}(A^\top) = \text{colspan}(V_1)$
- $\mathcal{N}(A^\top) = \text{colspan}(U_2)$

For example, suppose $\vec{x} \in \text{colspan}(V_2)$. Then $V_1^\top \vec{x} = \vec{0}$ because the columns of V are orthonormal (and so $V_1^\top V_2 = 0$). So $A\vec{x} = \vec{0}$.

§8.3 Gradient, Hessian, Jacobian, and Chain Rule

- Let $f : \mathbb{R}^n \to \mathbb{R}$ be a real-valued function of n variables that is continuously differentiable. The gradient of f at x, denoted by, $\nabla f(x)$, is

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix} \in \mathbb{R}^n$$

  Here $\partial f / \partial f x_i$ represents the partial derivative of f with respect to x_i.

- A gradient with respect to only a subset of the unknowns can be expressed by means of subscript on the symbol ∇. For example, $\nabla_x f(x, z)$ denotes the gradient with respect to x while holding z constant.

- If f is twice continuously differentiable, the matrix of second-order partial derivatives of f is known as the Hessian, and is defined as

$$\nabla^2 f(x) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

- The Hessian is a symmetric matrix since $\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}$
- When f is a vector-valued function, that is, $f : \mathbb{R}^n \to \mathbb{R}^m$, we define $\nabla f(x)$ to be the $n \times m$ matrix whose i^{th} columns is $\nabla f_i(x)$, that is

$$\nabla f(x) = \begin{bmatrix} | & | & & | \\ \nabla f_1(x) & \nabla f_2(x) & \dots & \nabla f_m(x) \\ | & | & & | \end{bmatrix}$$

- The rows of the gradient are indexed by variable components, while the columns by function components, and the $(i,j)^{\text{th}}$ entry is $[\nabla(x)]_{ij} = \frac{\partial f_j(x)}{\partial x_i}$.
- Often it's easier to work with the transpose of this matrix called Jacobian and usually denoted by $Df(x), J_f(x)$ or $\frac{\partial(f_1,\dots,f_m)}{\partial(x_1,\dots,x_n)}$

$$Df(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} & \dots & \frac{\partial f}{\partial x_n} \end{bmatrix} = (\nabla f(x))^\top \in \mathbb{R}^{m \times n}$$

- The rows of Jacobian are indexed by function components, while the columns by variable components, and the $(i,j)^{\text{th}}$ entry is $[Df(x)]_{ij} = \frac{\partial f_i(x)}{\partial x_j}$.

Question 8.1. Is $\nabla^2 f(x) = \nabla(\nabla f(x))$ or $= D(\nabla f(x))$

Ans: Both! For $f : \mathbb{R}^n \to \mathbb{R}$

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix} \triangleq g(x) \quad g : \mathbb{R}^n \to \mathbb{R}^n$$

$$\nabla(\nabla f(x)) = \nabla g(x) = \begin{bmatrix} | & | & & | \\ \nabla g_1(x) & \nabla g_2(x) & \dots & \nabla g_n(x) \\ | & | & & | \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_1} \\ \vdots & & \vdots \\ \frac{\partial^2 f}{\partial x_1 \partial x_n} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_n} \end{bmatrix} = \nabla^2 f(x)$$

$$D(\nabla f(x)) = Dg(x) = \begin{bmatrix} \frac{\partial g}{\partial x_1} & \dots & \frac{\partial g}{\partial x_n} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \vdots & & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_n} \end{bmatrix} = \nabla^2 f(x)$$

§9 Lec 9: Apr 16, 2021

§9.1 Lec 8 (Cont'd)

Chain Rule: Let $g : \mathbb{R}^n \to \mathbb{R}^m$ and $f : \mathbb{R}^m \to \mathbb{R}^p$, and their composition $h = f \circ g : \mathbb{R}^n \to \mathbb{R}^p$ be defined via $h(x) = f(g(x))$ for $x \in \mathbb{R}^n$. Then

$$\begin{aligned}\nabla h(x) &= \begin{bmatrix}\dots & \nabla h_i(x) & \dots\end{bmatrix} = \begin{bmatrix}\dots & \sum_{j=1}^m \frac{\partial f_i}{\partial g_j}\nabla g_j(x) & \dots\end{bmatrix} \\ &= \begin{bmatrix}\dots & \nabla g(x)\nabla f_i(g) & \dots\end{bmatrix} \\ &= \nabla g(x)\nabla f(g(x))\end{aligned}$$

Or simply

$$\nabla(f \circ g) = \nabla g \nabla f(g) \implies (\nabla(f \circ g))^\top = (\nabla g \nabla f(g))^\top = (\nabla f(g))^\top (\nabla g)^\top$$

Or in terms of Jacobian,

$$J_{f\circ g}(x) = J_f(g(x))\, J_g(x),\ D(f \circ g) = Df Dg$$

Example 9.1

Calculate the gradient and Hessian for $f(x) = x_1^3 + 3x_1x_2^2$

- Gradient:
$$\nabla f(x) = \begin{bmatrix}3x_1^2 + 3x_2^2 \\ 6x_1x_2\end{bmatrix}$$
- Hessian:
$$\nabla^2 f(x) = \begin{bmatrix}6x_1 & 6x_2 \\ 6x_2 & 6x_1\end{bmatrix}$$

Example 9.2

Calculate the gradient and Jacobian for $f(x) = Ax + b$ for $A \in \mathbb{R}^{m \times n}$.

Notice $f : \mathbb{R}^n \to \mathbb{R}^m$, denote

$$A = \begin{bmatrix} — & a_1^T & — \\ — & a_2^T & — \\ & \vdots & \\ — & a_m^T & — \end{bmatrix}, \; f(x) = Ax + b = \begin{bmatrix} a_1^\top x + b_1 \\ \vdots \\ a_m^\top x + b_m \end{bmatrix}$$

- Gradient:

$$g(x) \triangleq a^\top x = \sum_{i=1}^{n} \alpha_i \gamma_i = a_1\gamma_1 + a_2\gamma_2 + \dots$$

$$\nabla g(x) = \begin{bmatrix} \frac{\partial g}{\partial x_1} \\ \vdots \\ \frac{\partial g}{\partial x_n} \end{bmatrix} = \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} = a$$

$$\nabla f(x) = \begin{bmatrix} \nabla f_1(x) & \dots & \nabla f_m(x) \end{bmatrix} = \begin{bmatrix} a_1 & \dots & a_m \end{bmatrix} = A^\top$$

- Jacobian:

$$Df(x) = [\nabla f(x)]^\top = A$$

§10 Lec 10: Apr 19, 2021

§10.1 Lec 9 (Cont'd)

Example 10.1

How about $f(A) = Ax + b$?

Hint:

1. Vectorize A
2. Connect it to Ex 2.

Have:

$$A = \begin{bmatrix} | & & | \\ a_1 & \dots & a_n \\ | & & | \end{bmatrix}, \operatorname{vec}(A) = \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix}$$

$$f(\operatorname{vec}(A)) = f(A) = Ax + b = \begin{bmatrix} a_1 & \dots & a_n \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} + b$$

$$= \sum_{i=1}^{n} x_i a_i + b = \sum_{i=1}^{n} (x_i I) a_i + b$$

$$= \begin{bmatrix} x_1 I & \dots & x_n I \end{bmatrix} \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} + b$$

So

$$Df(A) = \begin{bmatrix} x_1 I & \dots & x_n I \end{bmatrix} = x^\top \otimes I$$

where $\otimes$ denotes the Kronecker product and

$$\underbrace{A}_{m \times n} \otimes \underbrace{B}_{p \times q} = \begin{bmatrix} a_{11}B & a_{12}B & \dots & a_{1n}B \\ a_{21}B & a_{22}B & \dots & a_{2n}B \\ \vdots & \vdots & & \vdots \\ a_{m1}B & a_{m2}B & \dots & a_{mn}B \end{bmatrix} = \underbrace{C}_{mp \times nq}$$

So

$$\nabla f(A) = \begin{bmatrix} x_1 I \\ \vdots \\ x_n I \end{bmatrix} = x \otimes I$$

Example 10.2

Calculate the gradient and Hessian for $f(x) = \frac{1}{2}\|Ax - b\|_2^2$

Hint: Chain rule

$$\underbrace{g(x) \triangleq Ax - b}_{g:\mathbb{R}^n \to \mathbb{R}^n}, \quad \underbrace{h(g) = \frac{1}{2}\|g\|_2^2}_{h:\mathbb{R}^m \to \mathbb{R}} \implies f(x) = h\left(g(x)\right)$$

$$\nabla f(x) = \nabla g(x) \nabla h\left(g(x)\right) = A^\top g(x) = A^\top (Ax - b)$$

$$\nabla^2 f(x) = D\left(\nabla f(x)\right) = D\left(A^\top (Ax - b)\right) = D\left(A^\top Ax\right) = A^\top A$$

§10.2 Taylor's Theorem

The foundational result for many algorithms in smooth nonlinear optimization is Taylor's theorem. Taylor's theorem shows how smooth functions can be approximated locally by low-order (linear or quadratic) functions. The next iterate of many iterative algorithms can be obtained by minimizing a local approximation of the objective function around the previous iterate. Therefore, the convergence property of these algorithms on the accuracy of this approximation.

Given a continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ and $x, p \in \mathbb{R}^n$, we have

$$\text{mean-value version: } f(x+p) = f(x) + \underbrace{\nabla f(x+tp)^\top}_{\text{slope in scalar case}} p \text{ for some } t \in (0,1)$$

$$\text{integral version: } f(x+p) = f(x) + \int_0^1 \nabla f(x+tp)^\top p \, dt$$

If f is twice continuously differentiable then

$$\nabla f(x+p) = \nabla f(x) + \int_0^1 \nabla^2 f(x+tp) p \, dt$$

$$f(x+p) = f(x) + \nabla f(x)^\top p + \frac{1}{2} p^\top \nabla^2 f(x+tp) p \text{ for some } t \in (0,1) \qquad (\star)$$

$\star$ is also known as **Taylor's Expansion Theorem**. We can use it to compute gradient and Hessian.

Example 10.3

Compute $\nabla f(x)$ and $\nabla^2 f(x)$ with Taylor's Expansion Theorem where $f(x) = \frac{1}{2}\|Ax - b\|_2^2$ for $A \in \mathbb{R}^{m\times n}$.

Have

$$\begin{aligned} f(x) &= \frac{1}{2}(Ax-b)^\top (Ax-b) \\ &= \frac{1}{2}x^\top A^\top Ax - \frac{1}{2}b^\top Ax - \frac{1}{2}x^\top A^\top b + \frac{1}{2}b^\top b \\ &= \underbrace{\frac{1}{2}x^\top A^\top Ax}_{g(x)} - \underbrace{b^\top Ax}_{\nabla(b^\top Ax)} + \frac{1}{2}b^\top b \\ g(x+p) &= \frac{1}{2}(x+p)^\top A^\top A(x+p) \\ &= \underbrace{\frac{1}{2}x^\top A^\top Ax}_{g(x)} + \underbrace{\frac{1}{2}p^\top A^\top Ax + \frac{1}{2}x^\top A^\top Ap}_{=x^\top A^\top Ap} + \frac{1}{2}p^\top \underbrace{A^\top A}_{\nabla^2 g(x)} p \end{aligned}$$

Then

$$\begin{aligned} \nabla f(x) &= A^\top Ax - A^\top b = A^\top(Ax-b) \\ \nabla^2 f(x) &= A^\top A \end{aligned}$$

§11 | Lec 11: Apr 21, 2021

§11.1 Taylor's Theorem (Cont'd)

A brief intro to norm: a norm ξ is a function from a vector space to the non-negative real numbers, that satisfies the following three properties for any $x, y \in \mathbb{R}^n$.

- Non-negative: $\xi(x) \geq 0$ and $\xi(x) = 0 \iff x = 0$.
- Absolutely homogeneous: $\xi(ax) = |a|\xi(x)$ for any $a \in \mathbb{R}$.
- Triangle inequality: $\xi(x + y) \leq \xi(x) + \xi(y)$

Common examples: the l_p norm ($p \geq 1$) in $\mathbb{R}^n$ is $\|x\|_p = (\sum_{i=1}^n |x_i|^p)^{\frac{1}{p}}$.
When $p = 2$, we have

$$\|x\|_2 = \sqrt{\sum_{i=1}^n |x_i|^2}$$

which for convenience is often written as $\|x\|$. It defines the Euclidean distance of a vector and is by far the most commonly used norm in $\mathbb{R}^n$.

Fact 11.1. $\|x\| = \sqrt{x^\top x} = \sqrt{\langle x, x\rangle}$

Cauchy-Schwarz Inequality:

$$|\langle x, y\rangle| \leq \|x\|_2 \cdot \|y\|_2 \qquad (x, y \in \mathbb{R}^n)$$

A crucial quantity in optimization is the Lipschitz constant L for the gradient of f, which is defined to satisfy

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\| \leftarrow l_2\text{-norm}$$

for all x, y in the domain of f.
Given f with ∇f uniformly Lipschitz continuous with constant L, we have for any x, y in the domain of f that

$$f(y) \leq \underbrace{f(x) + \nabla f(x)^\top (y - x) + \frac{L}{2}\|y - x\|_2^2}_{\triangleq Q(y)}$$

where " $\leq$ " is sometimes known as as the quadratic upper bound.

Proof. Denote $p = y - x$. Using the integral-version of Taylor's theorem,

$$\begin{aligned} f(y) = f(x + p) &= f(x) + \int_0^1 \nabla f(x + tp)^\top p \, dt \\ &= f(x) + \int_0^1 \left(\nabla f(x + tp) - \nabla f(x) + \nabla f(x)\right)^\top p \, dt \\ &= f(x) + \nabla f(x)^\top (y - x) + \int_0^1 \left(\nabla f(x + tp) - \nabla f(x)\right)^\top p \, dt \end{aligned}$$

$$\begin{aligned} \left|\left(\nabla f(x + t(y - x)) - \nabla f(x)\right)^\top (y - x)\right| &\leq \|\nabla f(x + t(y - x)) - \nabla f(x)\|_2 \cdot \|y - x\|_2 \\ &\leq Lt\|y - x\|_2^2 \end{aligned}$$

Finally,

$$\begin{aligned} f(y) &\leq f(x) + \nabla f(x)^\top (y-x) + \int_0^1 Lt\|y-x\|_2^2 \, dt \\ &= f(x) + \nabla f(x)^\top (y-x) + \frac{1}{2}\|y-x\|_2^2 \end{aligned}$$

□

This means that f can be upper bounded by a quadratic function whose value at x is equal to $f(x)$. If f is twice continuously differentiable, then ∇f is Lipschitz continuous with Lipschitz constant $L \iff$ for x we have all the eigenvalues of $\nabla^2 f(x)$ are between $-L$ and L.

$$-L \leq \lambda_{\min}\left(\nabla^2 f(x)\right) \leq \lambda_{\max}\left(\nabla^2 f(x)\right) \leq L$$

§11.2 Solution and Optimality Conditions

Let $f : \mathcal{D} \to \mathbb{R}$ where $\mathcal{D} \subseteq \mathbb{R}^n$. We define different notions of minimizers.

- $x^* \in \mathcal{D}$ is a local minimizer of f if there is a neighborhood $\mathcal{N}$ of x^* s.t. $f(x) \geq f(x^*)$ for all $x \in \mathcal{N} \cap \mathcal{D}$.
- $x^* \in \mathcal{D}$ is a global minimizer of f if $f(x) \geq f(x^*)$ for all $x \in \mathcal{D}$.
- $x^* \in \mathcal{D}$ is a strict local minimizer if it is a local minimizer and in addition $f(x) > f(x^*)$ for all $x \in \mathcal{N}$ with $x \neq x^*$.

Definition 11.1 (Critical/Stationary Point) — A point x^* with $\nabla f(x^*) = 0$ is called a critical point or a stationary point.

Theorem 11.2 (Necessary Conditions for Smooth Unconstrained Optimization)

There are two necessary conditions:

1. Suppose that f is continuously differentiable. Then if x^* is a local minimizer of the unconstraint optimization $\min_x f(x)$, then $\nabla f(x^*) = 0$.
2. Suppose that f is twice continuously differentiable. Then if x^* is a local minimizer of the unconstraint optimization $\min_x f(x)$, then $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ is positive semi-definite, i.e. $\nabla^2 f(x^*) \succeq 0$

(1) is called the first-order necessary condition and (2) is called the second-order necessary condition.

A matrix A is positive semi-definite if

1. A is symmetric.
2. eigs(A) are non-negative.
3. For all nonzero $x \in \mathbb{R}^n$, $x^\top A x \geq 0$.

§12 | Lec 12: Apr 23, 2021

§12.1 Solution and Optimality Conditions (Cont'd)

Proof. 1. Assume x^* is a local min but $\nabla f(x^*) \neq 0$. Consider $x(\alpha) = x^* - \alpha \nabla f(x^*)$, $\alpha > 0$

$$\begin{aligned} f\left(x(\alpha)\right) &= f(x^* - \alpha \nabla f(x^*)), \quad \alpha > 0 \\ \text{Taylor's(mean-value)} &= f(x^*) + \nabla f\left(x^* - \alpha t \nabla f(x^*)\right)^\top \left(-\alpha \nabla f(x^*)\right), \text{ for some } t \in (0,1) \\ &= f(x^*) - \alpha \nabla f\left(x^* - \alpha t \nabla f(x^*)^\top \nabla f(x)\right) \end{aligned}$$

Have

$$\lim_{\alpha \to 0} \nabla f\left(x^* - \alpha t \nabla f(x^*)\right)^\top \nabla f(x) = \|\nabla f(x^*)\|_2^2 > 0$$

By definition of limit, there exists α_1 s.t. for all $\alpha \in [0, \alpha_1)$, we have

$$\begin{aligned} &\nabla f\left(x^* - \alpha t \nabla f(x^*)\right)^\top \nabla f(x)^* > \frac{1}{2}\|\nabla f(x^*)\|_2^2 \\ \implies &f(x(\alpha)) < f(x^*) - \underbrace{\frac{1}{2}\alpha \|\nabla f(x^*)\|_2^2}_{>0} < f(x^*), \forall \alpha \in [0, \alpha_1) \end{aligned}$$

2. By (1), we have $\nabla f(x^*) = 0$. Suppose $\nabla^2 f(x^*) \not\succeq 0$. Then $\exists v \neq 0$, s.t. $v^\top \nabla^2 f(x^*) v = \lambda < 0$. Consider $x(\alpha) = x^* + \alpha v$

$$\begin{aligned} f\left(x(\alpha)\right) &= f(x^* = \alpha v) \\ \text{Taylor's Thm} &= f(x^*) + \nabla f(x^*)^\top (\alpha v) + \frac{1}{2}\alpha^2 v^\top \nabla^2 f(x^* + \alpha t v) v, \text{ for some } t \in (0,1) \\ &= f(x^*) = \frac{1}{2}\alpha^2 v^\top \nabla^2 f\left(x^* + \alpha t v\right) v, \text{ for some } t \in (0,1) \end{aligned}$$

Notice

$$\lim_{\alpha \to 0} v^\top \nabla^2 f\left(x^* + \alpha t v\right) v = v^\top \nabla^2 f(x^*) v = \lambda < 0$$

By definition of limit, there exists α_1 s.t. for all $\alpha \in [0, \alpha_1)$, we have

$$v^\top \nabla^2 f\left(x^* + \alpha t v\right) v < \frac{\lambda}{2} \implies f(x(\alpha)) < f(x^*), \forall \alpha \in [0, \alpha_1)$$ □

Theorem 12.1 (Sufficient Condition for Smooth Unconstrained Optimization)
Suppose that f is twice continuously differentiable and for some x^*, we have $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ is positive definite. Then x^* is a strict local minimizer of the unconstraint optimization $\min_x f(x)$.

Proof. Have

$$\begin{aligned}
\nabla^2 f(x^*) \succ 0 \implies \nabla^2 f(x^* + tp) \succ 0 \text{ for any } \|p\| \le r,\ t \in (0,1) \\
f(x^* + p) = f(x^*) + \nabla f(x^*)^\top p + \frac{1}{2} p^\top \nabla^2 f(x^* + tp) p, \text{ for some } t \in (0,1) \\
= f(x^*)) + \frac{1}{2} \underbrace{p^\top \nabla^2 f(x^* + tp) p}_{>0}, \text{ for some } t \in (0,1) \\
> f(x^*), \quad \text{for any } \|p\| \le r
\end{aligned}$$

Choose $\mathcal{N} = \{x^* + p : \|p\| \le r\}$. □

Definition 12.2 (Local Maximizer) — If $\nabla f(x^*) = 0$ and all the eigenvalues of $\nabla^2 f(x^*)$ are negative, x^* is a local maximizer.

Definition 12.3 (Saddle Point) — If $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ has both positive and negative eigenvalues, x^* is a saddle points.

§12.2 Midterm

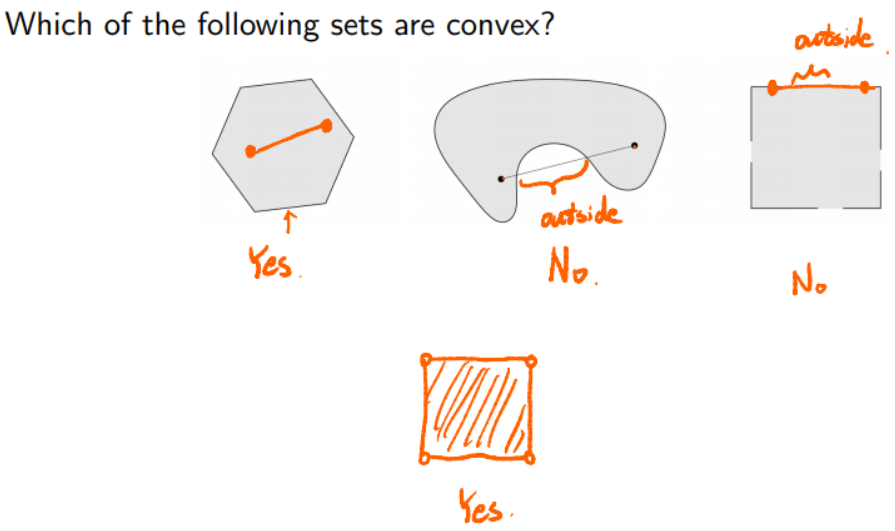
- Date: Next Friday, Apr 30, 2021 at 9:00 am
- Deadline: May 1, 2021 at 9:00 am
- Materials covered: Lec 1, Lec 2.1 – Lec 2.4 on CCLE
- Review:
  - Notes/Examples on slides
  - Questions in HW1 & HW2
- Open note/book.

§12.3 Convexity

Convex functions take a central role in optimization – they are the class of functions that are guaranteed to find global minimizer within a reasonable amount of time.

Definition 12.4 (Convex Set) — A set $\omega \subseteq \mathbb{R}^n$ is convex if for any $x, y \in \omega$ and any $\alpha \in [0,1]$, one has $(1-\alpha)x + \alpha y \in \omega$.

Geometrically, for all pairs of points in a convex set, the line segment between them is also contained in the set.



Example 12.5

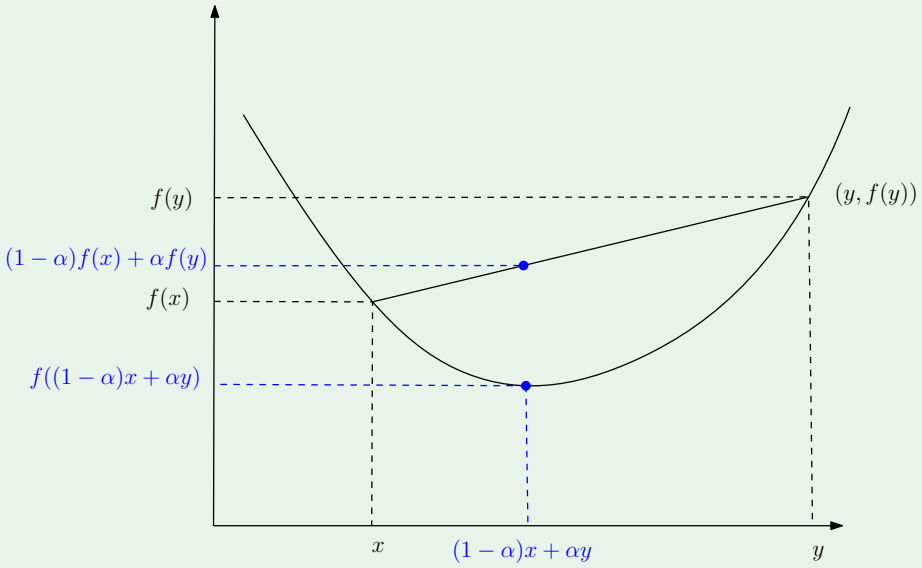
Examples of convex sets:

- Hyperplane: $\{x : a^\top x = b\}$ with $a \neq 0 \in \mathbb{R}^n$ and $b \in \mathbb{R}$
- Halfspace: $\{x : a^\top x \leq b\}$ with $a \neq 0 \in \mathbb{R}^n$ and $b \in \mathbb{R}$
- Norm ball: $\{x : \|x - x_c\| \leq r\}$
- Non-negative Orthant: $\mathbb{R}^n_+ = \{x : x \geq 0\}$
- Positive semi-definite cone: $S^n_+ = \{X : X \text{ is symmetric}, X \succeq 0\}$

Definition 12.6 (Convex Function) — A function $f : \omega \to \mathbb{R}$ where $\omega \subseteq \mathbb{R}^n$ is a convex function if its domain ω is a convex set and for all $x, y \in \omega$ and all $\alpha \in [0, 1]$, one has

$$f\left((1-\alpha)x + \alpha y\right) \leq (1-\alpha)f(x) + \alpha f(y)$$

Geometrically, for convex function, the line segment connecting $(x, f(x))$ and $(y, f(y))$ lie above the graph of the function f.

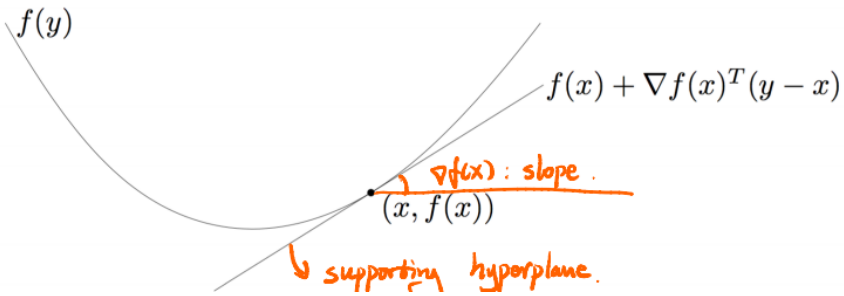


A function f is concave if $-f$ is convex.

A continuously differentiable function f is convex iff its domain is convex and for all $x, y \in \text{dom}(f)$

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x)$$

A twice continuously differentiable function f is convex iff its domain is convex and its Hessian is positive semi-definite, that is, $\nabla^2 f(x) \succeq 0$ for all $x \in \text{dom}(f)$.



§13 Lec 13: Apr 26, 2021

§13.1 Strongly Convex Functions

When $f : \omega \to \mathbb{R}$ is continuously differentiable, we call f is strongly convex with modulus of convexity m if ω is convex and there exists $m > 0$ s.t.

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{m}{2}\|y - x\|_2^2$$

Just as the Lipschitz constant of the gradient provides an upper bound on the eigenvalues of the Hessian, the strongly convex parameter provides a lower bound for twice continuously differentiable functions.
More precisely, suppose f is twice continuously differentiable, then f has modulus of convexity $m \iff \lambda_{\min}\left(\nabla^2 f(x)\right) \geq m > 0$ for all x.

Exercise 13.1. Answer whether or when the following functions are (strongly) convex. Hint:

1. Check domain of f is convex.
2. Compute $\nabla^2 f(x)$
3. If $\lambda_{\min}\left(\nabla^2 f(x)\right) \geq m > 0 \implies$ strongly convex.
4. $\lambda_{\min}\left(\nabla^2 f(x)\right) \geq 0 \implies$ convex.

- x^2, $x \in \mathbb{R}$
  1. Domain is convex
  2. $\nabla f(x) = 2x \implies \nabla^2 f(x) = 2 \implies$ strongly convex
- x^4, $x \in \mathbb{R}$
  1. Domain is convex
  2. $\nabla f(x) = 4x^3$, $\nabla^2 f(x) = 12x^2 \geq 0 \implies$ convex
- e^{ax} on $\mathbb{R}$
  1. Domain is convex
  2. $\nabla f(x) = ae^{ax}$, $\nabla^2 f(x) = a^2 e^{ax} \geq 0 \implies$ convex
- $-\log(x)$ on $\underbrace{\mathbb{R}_{++}}_{x>0}$
  1. Domain is convex
  2. $\nabla f(x) = -\frac{1}{x}$, $\nabla^2 f(x) = \frac{1}{x^2} \geq 0 \implies$ convex
- $f(x) = \|x\|_2^2$, $x \in \mathbb{R}^n$
  1. Domain is convex
  2. $\nabla f(x) = 2x$, $\nabla^2 f(x) = 2I \implies$ strongly convex

- $f(x) = a^\top x + b$, $x \in \mathbb{R}^n$
  1. Domain is convex
  2. $\nabla f(x) = a$, $\nabla^2 f(x) = 0 \implies$ convex and concave
- $f(x) = \frac{1}{2} x^\top P x + q^\top x + r$, $x \in \mathbb{R}^n$ and $P = P^\top$
  1. Domain is convex
  2. $\nabla f(x) = Px + q$, $\nabla^2 f(x) = P$. If $P \succeq 0 \implies$ convex. If $P \succ 0 \implies$ strongly convex. If $P \preceq 0 \implies$ concave. If $P \prec 0 \implies$ strongly concave

Theorem 13.1

Suppose f is continuously differentiable and convex. Then if $\nabla f(x^\star) = 0$, then $x^\star$ is a global minimizer of $\min_x f(x)$. When, in addition, f is strongly convex, then x^* is the unique global minimizer.

Proof. Take $x = x^\star$,

$$f(y) \geq f(x^\star) + \underbrace{\nabla f(x^\star)^\top}_{=0}(y - x^\star) = f(x^\star), \quad \forall y$$

for convex f, and

$$f(y) \geq f(x^\star) + \underbrace{\nabla f(x^\star)^\top}_{=0}(y - x^\star) + \frac{m}{2}\|y - x^\star\|_2^2 = f(x^\star) + \underbrace{\frac{m}{2}\|y - x^\star\|_2^2}_{>0 \text{ if } y \neq x^\star} > f(x^\star)$$

for strongly convex f and $y \neq x^\star$. □

§14 | Lec 14: Apr 28, 2021

§14.1 Examples of Finding Global Minimizers

Example 14.1

$f : \mathbb{R} \to \mathbb{R}$, $f(x) = (x-3)^2$

1. Is $f(x)$ strongly convex?
2. Find the global minimizers of $f(x)$. Is it unique?

- $\text{dom}(f) = \mathbb{R}$ is a convex set
- $\nabla f(x) = 2(x-3)$, $\nabla^2 f(x) = 2 \implies f(x)$ is a strongly convex function.
  $\nabla f(x) = 0 \implies x = 3$ is the unique global minimizer of $f(x)$.

Example 14.2

$f : \mathbb{R}^n \to \mathbb{R}$, $f(x) = \|x - x^\star\|_2^2$, $x^\star \in \mathbb{R}^n$ is given.

1. Is $f(x)$ (strongly) convex?
2. Find the global minimizers of $f(x)$. Is it unique?

- $\text{dom}(f) = \mathbb{R}^n$ is a convex set
- Consider:

$$\begin{aligned} f(x+d) &= \|x + d - x^\star\|_2^2 = \|(x - x^\star) + d\|_2^2 \\ &= \langle (x - x^\star) + d, (x - x^\star) + d \rangle \\ &= \langle x - x^\star, x - x^\star \rangle + \langle x - x^\star, d \rangle + \langle d, x - x^\star \rangle + \langle d, d \rangle \\ &= \underbrace{\|x - x^\star\|_2^2}_{f(x)} + \underbrace{\langle 2(x - x^\star)}_{\nabla f(x)}, d\rangle + \frac{1}{2} d^\top (\underbrace{2I}_{\nabla^2 f(x)}) d \end{aligned}$$

$\implies f(x)$ is a strongly convex function, so $\nabla f(x) = 2(x - x^*) = 0 \implies x = x^*$ is the unique global minimizer.

Example 14.3

$f : \mathbb{R}^d \to \mathbb{R}$. Given $x_1, \ldots, x_n \in \mathbb{R}^d$. Find the global minimizer of

$$f(x) = \sum_{k=1}^{n} \|x - x_k\|_2^2$$

Hint: Show $f(x)$ is a convex function, so any critical point is a global minimizer.

- $\text{dom}(f = \mathbb{R}^d)$ is a convex set.
- Consider

$$\begin{aligned} f(x+d) &= \sum_{k=1}^{n} \|x + d - x_k\|_2^2 = \sum_{k=1}^{n} \|(x - x_k) + d\|_2^2 \\ &= \sum_{k=1}^{n} \left(\|x - x_k\|_2^2 + \langle 2(x - x_k), d \rangle + \frac{1}{2} d^\top (2I) d \right) \\ &= \underbrace{\sum_{k=1}^{n} \|x - x_k\|_2^2}_{f(x)} + \underbrace{\langle 2 \sum_{k=1}^{n} (x - x_k), d \rangle}_{\nabla f(x)} + \frac{1}{2} d^\top \underbrace{(2nI)}_{\nabla^2 f(x)} d \end{aligned}$$

  $\implies$ $f(x)$ is a strongly convex function.

$$\nabla f(x) = 2nx - 2 \sum_{k=1}^{n} x_k = 0 \implies x = \frac{1}{n} \sum_{k=1}^{n} x_k$$

  is the unique global minimizer of $f(x)$.

§15 Midterm: Apr 30, 2021 – :D

Figure 1: Better study now!

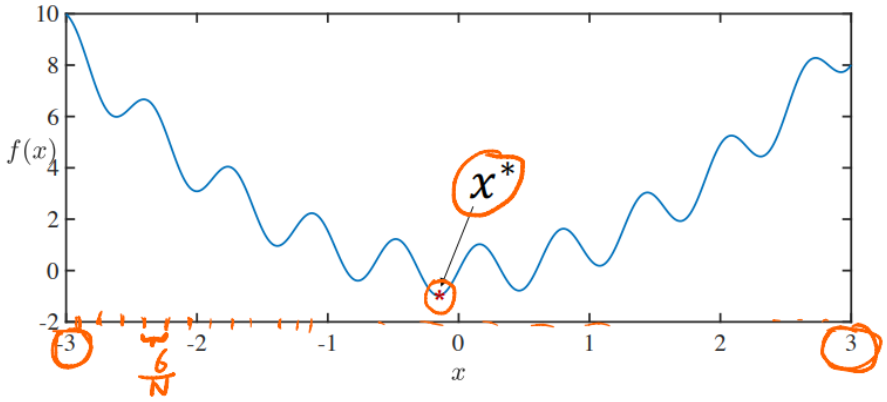
§16 Lec 15: May 3, 2021

§16.1 Gradient Descent Methods

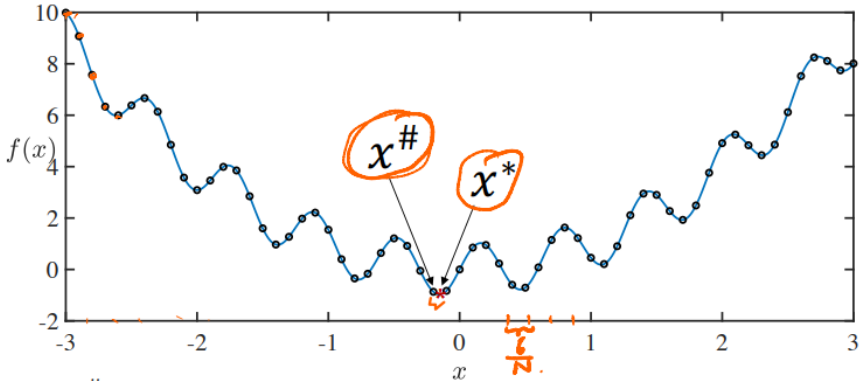
We will focus on the problem of unconstrained optimization

$$\text{minimize}_{x\in\mathbb{R}^n} f(x)$$

We will first consider first-order algorithms such as steepest/gradient descent. Then, we switch to second-order methods such as Newton method. All algorithms in unconstrained optimization require the user to supply a starting point x_0. Beginning at x_0, optimization algorithms generate a sequence of iterates $\{x_k\}_{k=0}^{\infty}$ that terminates when either no more progress can be made or when it seems that an approximate solution has been found. In deciding how to proceed from one iterate x_k to the next, the algorithms use information about the function at x_k, and possibly information from earlier states. They typically use this information to find a new iterate x_{k+1} with a lower function value.



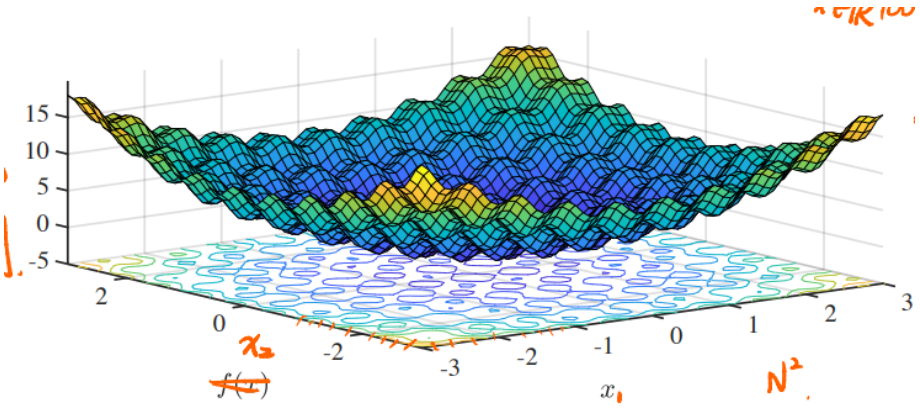
Consider finding the global minimizer x^*. A native way is to uniformly sample the x-axis N points, and find the minimizer among these N points.



Let $x^\#$ be the minimizer among the N points. Then,

$$\left|x^\# - x^*\right| \leq \frac{6}{N}$$

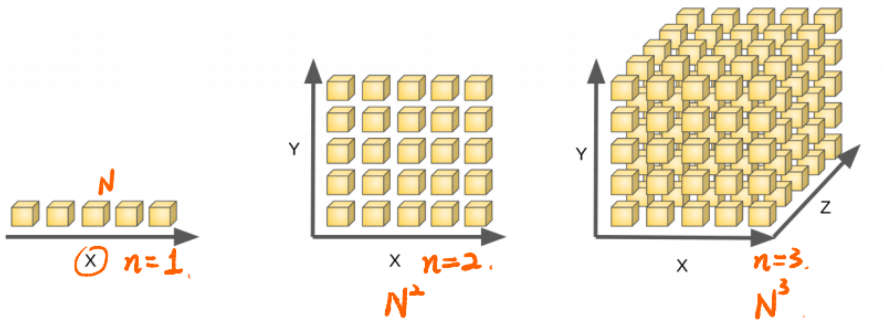
Also, $f(x^{\#}) - f(x^*) \leq \frac{6}{N}L$, where L is the Lipschitzness of f, which captures the smoothness of f. This also implies $f(x^{\#}) - f(x^*) \leq \epsilon$ if $N = \frac{6L}{\epsilon}$. This brute-force approach finds very good solution. Then why do we need other optimization algorithms?



Curse of dimensionality: When x is $\mathbb{R}^n$, then to guarantee

$$f(x^{\#}) - f(x^*) \leq \epsilon$$

we need $N = O\left(\frac{1}{\epsilon^n}\right)$



n	1	10	100
$N(\epsilon = 0.1)$	10	10^{10}	10^{100}

Nonpractical!

$N=10$. $O(10)$. $N = O\left(\frac{1}{\epsilon^{100}}\right) = O\left(\frac{1}{\epsilon^{10}}\right) = O(10^{10})$

We call d a descent direction for f at x if $f(x+d) < f(x)$ for all $t > 0$ sufficiently small. For any continuously differentiable function, any d s.t. $\nabla f(x)^\top d < 0$ is a descent direction.

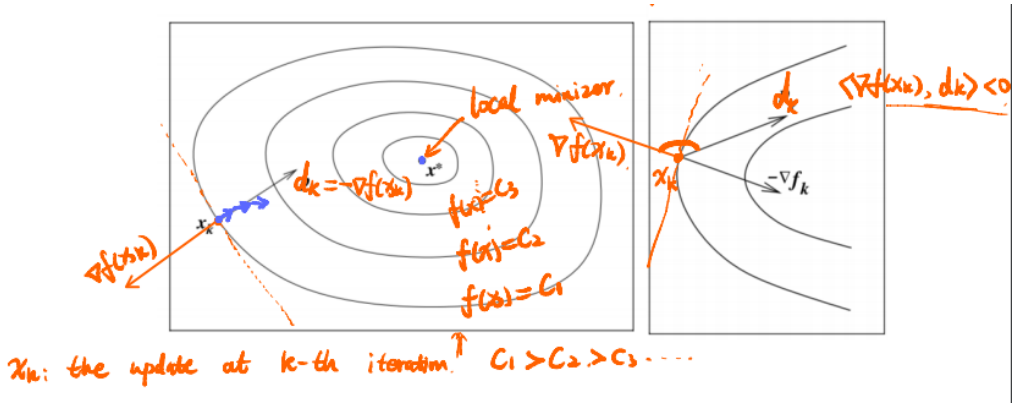
Proof. Continuity of ∇f ensures the existence of $\bar{t}$ s.t.

$$\nabla f(x + td)^\top d < 0 \qquad \forall t \in [0, \bar{t}]$$

Thus, Taylor's theorem implies that

$$f(x + td) = f(x) + t\nabla f(x + \gamma + d)^\top d < f(x) \quad \text{for some } \gamma \in (0, 1) \qquad \square$$

When t is sufficiently small, the amount of decrease is approximately $t\nabla f(x)^\top d$. Among all directions with unit norm, the minimum of $\nabla f(x)^\top d < 0$ is achieved when $d = -\frac{\nabla f(x)}{\|\nabla f(x)\|_2}$. Thus, $d = -\nabla f(x)$ is called the direction of steepest descent.



The simplest methods for optimization

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k), \qquad k = 0, 1, 2, \ldots$$

for some length $\alpha_k > 0$.
At each step, there is either some $\alpha_k > 0$ s.t. the function value is decreased, or $\nabla f(x_k) = 0$, at which we have found a local minimum, or a global one if f is convex. This algorithm is called gradient descent or the steepest descent. Large step-size risks taking a step that increases the function value, while too small step-size risks making too little progress in each iteration. Short-step variant: for functions with Lipschitz gradient with Lipschitz constant L, choose a constant step-size

$$x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k), \qquad k = 0, 1, 2, \ldots$$

This iteration scheme can also be obtained by minimizing the quadratic upper bound with respect to y :

$$f(y) \leq \underbrace{f(x_k) + \nabla f(x_k)^\top (y - x_k) + \frac{L}{2}\|y - x_k\|_2^2}_{\triangleq G(y)}$$

Plugging in the iteration $y = x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$ yields

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|_2^2$$

This so called descent lemma/inequality or sufficient decrease condition quantifies the amount of decrease we obtain and is one of the foundational inequalities in the analysis of optimization algorithms.

§17 | Lec 16: May 5, 2021

§17.1 Gradient Descent Methods (Cont'd)

f is bounded below, i.e., $f(x) \geq \overline{f}$ for all x. Adding the descent inequalities for $k = 0$ to $T-1$ yields

$$\overline{f} \leq f(X_T) \leq f(x_0) - \frac{1}{2L}\sum_{k=0}^{T-1} \|\nabla f(x_k)\|_2^2$$

Since $f(x_T) \geq \overline{f}$, we have

$$T \cdot \min_{0\leq k\leq T-1} \|\nabla f(x_k)\|_2^2 \leq \sum_{k=0}^{T-1} \|\nabla f(x_k)\|_2^2 \leq 2L\left(f(x_0) - \overline{f}\right) < \infty$$

implying

$$\lim_{T\to\infty} \|\nabla f(x_T)\|_2 = 0 \text{ and } \min_{0\leq k\leq T-1} \|\nabla f(x_k)\|_2 \leq \sqrt{\frac{2L(f(x_0)-\overline{f})}{T}} \leq \epsilon$$

So after $T \geq \frac{2L(f(x_0)-\overline{f})}{\epsilon^2}$ steps, we can find a point whose gradient norm is less than ϵ.
Now suppose that f is convex, smooth with Lipschitz L gradients, and has a minimizer x^* with $f^\star = f(x^\star)$. Convexity implies

$$f(x^\star) \geq f(x_k) + \nabla f(x_k)^\top (x^\star - x_k)$$

Substituting this into the descent inequality gives

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|_2^2 \leq f(x^\star) + \nabla f(x_k)^\top (x_k - x^\star) - \frac{1}{2L}\|\nabla f(x_k)\|_2^2 \\ &= f(x^\star) + \frac{L}{2}\left(\|x_k - x^\star\|_2^2 - \|x_k - x^\star - \frac{1}{L}\nabla f(x_k)\|_2^2\right) \\ &= f(x^\star) + \frac{L}{2}\left(\|x_k - x^\star\|_2^2 - \|x_{k+1} - x^\star\|_2^2\right) \end{aligned}$$

Summing over $k = 0$ to $T-1$ gives

$$\sum_{k=0}^{T-1} \left(f(x_{k+1}) - f^\star\right) \leq \frac{L}{2}\left(\|x^0 - x^\star\|_2^2 - \|x_T - x^\star\|_2^2\right) \leq \frac{L}{2}\|x_0 - x^\star\|_2^2$$

Since $f(x_k)$ is decreasing, one has

$$f(X_T) - f^\star \leq \frac{L}{2T}\|x_0 - x^\star\|_2^2$$

To find a solution with $f(x_T) - f^\star \leq \epsilon$, we need $T \geq \frac{L\|x_0 - x^\star\|_2^2}{2\epsilon}$ iterations.
m-strongly convex functions satisfy the Polyak-Lojasiewicz (PL) inequality:

$$\frac{1}{2}\|\nabla f(x)\|_2^2 \geq m\left(f(x) - f^\star\right)$$

Proof. m-strongly convexity implies

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{m}{2}\|y - x\|^2 \coloneqq q(y)$$

This implies

$$\min_y f(y) \geq \min_y q(y) = q\left(x - \frac{1}{m}\nabla f(x)\right)$$

Therefore,

$$f^\star \geq f(x) - \frac{1}{m}\|\nabla f(x)\|^2 + \frac{1}{2m}\|\nabla f(x)\|^2 = f(x) - \frac{1}{2m}\|\nabla f(x)\|^2 \qquad \square$$

§18 Lec 17: May 7, 2021

§18.1 Gradient Descent Methods (Cont'd)

This combines with the descent lemma gives

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|_2^2 \leq f(x_k) - \frac{m}{L}\left(f(x_k) - f^\star\right)$$

Subtracting $f^\star$ from both sides gives the recursion

$$f(x_{k+1}) - f^\star \leq \left(1 - \frac{m}{L}\right)\left(f(x_k) - f^\star\right) \leq \left(1 - \frac{m}{L}\right)^2\left(f(x_{k-1}) - f^\star\right) \leq \ldots \leq \left(1 - \frac{m}{L}\right)^{k+1}\left(f(x_0) - f^\star\right)$$

The function values converge linearly to the optimum

$$f\left(x^\top\right) - f^\star \leq \left(1 - \frac{m}{L}\right)^\top \left(f(x_0) - f^\star\right)$$

To find a solution with $f(x_T) - f^\star \leq \epsilon$, we need $T \geq \frac{1}{\log\left(\frac{1}{1-\frac{m}{L}}\right)} \log\left(\frac{f(x_0)-f^\star}{\epsilon}\right)$ iterations.

Denote by $\{\tau_k\}$ the sequence of positive scalar quantities of interest with $\tau_k \to 0$. Examples include $\tau_k = f(x_k) - f^\star$, $\|\nabla f(x_k)\|_2$ or $\|x_k - x^\star\|_2$. We say that $\{\tau_k\}$ has a linear rate of a convergence if

$$\lim_{k\to\infty} \frac{\tau_{k+1}}{\tau_k} = \phi, \qquad \text{for some } \phi \in (0,1)$$

or equivalently,

$$\tau_{k+1} \leq \phi\tau_k \leq \ldots \leq \phi^{k+1}\tau_0, \qquad \text{for some } \phi \in (0,1)$$

Example 18.1

Consider

$$(\tau_k) = \left\{1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}, \ldots, \frac{1}{2^k}, \ldots\right\}$$

Note that $\tau_{k+1} = \frac{1}{2}\tau_k$.

We say that $\{\tau_k\}$ has a sublinear rate of convergence if

$$\lim_{k\to\infty} \frac{\tau_{k+1}}{\tau_k} = 1$$

Example 18.2

Consider

$$\tau_k \leq \frac{A}{k+B}$$

for some scalars $A > 0$ and $B \geq 0$, denoted by $\tau_k = \mathcal{O}\left(\frac{1}{k}\right)$

$$(\tau_k) = \left\{1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots, \frac{1}{k+1}, \ldots\right\}$$

when

$$\tau_k \leq \frac{A}{\sqrt{k+B}}$$

denoted by $\tau_k = \mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$.

$$(\tau_k) = \left\{1, \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{4}}, \ldots, \frac{1}{\sqrt{k+1}}, \ldots\right\}$$

We say that $\{\tau_k\}$ has a superlinear rate of convergence if

$$\lim_{k\to\infty} \frac{\tau_{k+1}}{\tau_k} = 0$$

Example 18.3

This includes convergence with order q for $q > 1$ when

$$\lim_{k\to\infty} \frac{\tau_{k+1}}{\tau_k^q} \leq M$$

for some $M > 0$. In particular, $q = 2$ is called quadratic convergence

$$(\tau_k) = \left\{\frac{1}{2}, \frac{1}{4}, \frac{1}{16}, \frac{1}{256}, \frac{1}{65,536}, \ldots, \frac{1}{2^{2^k}}, \ldots\right\}$$

Denote the target value of τ_k by $\epsilon > 0$. We can obtain expression for $k = k(\epsilon)$ for the number of iterations required to guarantee $\tau_k \leq \epsilon$ as done previously. The expression $k(\epsilon)$ is usually called (iteration) complexity.

§18.2 An Example

Example 18.4

$f : \mathbb{R}^2 \to \mathbb{R}$, $f(x) = x^\top Ax$, $A = \begin{bmatrix} 1 & 0 \\ 0 & 100 \end{bmatrix}$

a) Find the global minimizer $x^\star$of $f(x)$

$$f(x+d) = (x+d)^\top A(x+d) = x^\top Ax + d^\top Ax + x^\top Ad + d^\top Ad$$
$$= x^\top Ax + \langle \underbrace{2Ax}_{=\nabla f(x)}, d\rangle + \frac{1}{2} d^\top \underbrace{(2A)}_{\nabla^2 f(x)} d$$

So

$$\implies \nabla f(x) = 2Ax, \quad \forall \nabla^2 f(x) = 2A = \begin{bmatrix} 2 & 0 \\ 0 & 200 \end{bmatrix}$$

$\implies \lambda_1\left(\nabla^2 f(x)\right) = 2$, $\lambda_2\left(\nabla^2 f(x)\right) = 200 \implies f(x)$ is strongly convex with $m = 2$

$\implies$ Its critical point $x^\star$ is the unique global minimizer of $f(x)$

$$\implies \nabla f(x^\star) = 2Ax^\star = \vec{0} \implies Ax^\star = \vec{0}$$

Since A is invertible, $x^\star = A^{-1}\vec{0} = \vec{0} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$

b) What is the first iteration of gradient descent method with the step-size chosen as one over the Lipschitz constant of $\nabla f(x)$ and starting point as $x_0 = \begin{bmatrix} 100 \\ 100 \end{bmatrix}$?

Hint: $\nabla^2 f(x) \leq LI$,i.e., $\left(-L \leq \lambda_{\min}(\nabla^2 f(x)) \leq \lambda_{\max}\left(\nabla^2 f(x)\right) \leq L\right)$

$\implies L = 200$

$$x_1 = x_0 - \frac{1}{L}\nabla f(x_0) = x_0 - \frac{1}{200} \cdot 2 \begin{bmatrix} 1 & 0 \\ 0 & 100 \end{bmatrix} x_0 = \left(\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} \frac{1}{100} & 0 \\ 0 & 1 \end{bmatrix} \right) x_0$$
$$= \begin{bmatrix} 0.99 & 0 \\ 0 & 0 \end{bmatrix} x_0$$

§19 Lec 18: May 10, 2021

§19.1 An Example (Cont'd)

Example 19.1 (Cont'd from Lec 17) c) What is the closed-form expression of x_k in the k-th iteration of gradient descent method for any positive integer k? ($\alpha_k = \frac{1}{L}$)

$$x_k = x_{k-1} - \frac{1}{L}\nabla f(x_{k-1}) = x_{k-1} - \frac{1}{200}\cdot 2\begin{bmatrix}1 & 0\\ 0 & 100\end{bmatrix}x_{k-1} = \left(\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix} - \begin{bmatrix}\frac{1}{100} & 0\\ 0 & 1\end{bmatrix}\right)x_{k-1}$$

$$= \begin{bmatrix}.99 & 0\\ 0 & 0\end{bmatrix}x_{k-1} = \begin{bmatrix}.99 & 0\\ 0 & 0\end{bmatrix}^2 x_{k-2} = \ldots = \begin{bmatrix}.99 & 0\\ 0 & 0\end{bmatrix}^k x_0$$

c) After how many iterations, we have $\|x_k - x^\star\|_2 < \frac{1}{100}$?

$$\|x_k - x^\star\|_2 = \|x_k\|_2 = \|\begin{bmatrix}.99^k & 0\\ 0 & 0\end{bmatrix}\begin{bmatrix}100\\ 100\end{bmatrix}\|_2 = \|\begin{bmatrix}.99^k\cdot 100\\ 0\end{bmatrix}\|_2 = .99^k\cdot 100 < \frac{1}{100}$$

We need $.99^k < 10^{-4}$ which means $k\log(0.99) < -4$, so $k > -\frac{4}{\log(0.99)} \approx 916.4$.
So after $k = 917$ iterations, we have $\|x_k - x^\star\|_2 < \frac{1}{100}$

c) What's the convergence rate of the sequence $\{\|x_k - x^\star\|_2\}$? (sublinear/linear/quadratic)

$$\|x_k - x^\star\|_2 = .99^k\cdot 100 = .99\left(.99^{k-1}\cdot 100\right) = .99\underbrace{\|x_{k-1} - x^\star\|_2}_{\tau_{k-1}}$$

so $\tau_k = 0.99\tau_{k-1} \implies$ linear.

§19.2 Newton's Method

Newton's method uses the search direction given by

$$d_k^N = -\nabla^2 f(x_k)^{-1}\nabla f(x_k)$$

Newton's direction is derived from the second-order Taylor series approximation to $f(x_k+d)$, which is

$$f(x_k + d) \approx f(x_k) + \nabla f(x_k)^\top d + \frac{1}{2}d^\top\nabla^2 f(x_k)d := m_k(d)$$

When the Hessian is positive definite, the Taylor series approximate is minimized by setting

$$d = d_k^N = -\nabla^2 f(x_k)^{-1}\nabla f(x_k)$$

The Newton direction is reliable when the difference between the true function $f(x_k + d)$ and its quadratic model $m_k(d)$ is not too large.

$$\nabla_d m_k(d) = \nabla f(x_k) + \nabla^2 f(x_k)d = 0 \implies d = -\nabla^2 f(x_k)^{-1}\nabla f(x_k)$$

The Newton's direction is usually not computed by taking the inverse of the Hessian, but rather solving the system of linear equations

$$\nabla^2 f(x_k)d_k^N = -\nabla f(x_k)$$

When the Hessian is positive definite, Newton's direction is a descent direction because

$$\nabla f(x_k)^\top d_K^N = \left(-\nabla^2 f(x_k) d_k^N\right)^\top d_k^N = -{d_k^N}^\top \nabla^2 f(x_k) d_k^N \leq -{d_k^N}^\top \left(\lambda_{\min}\left(\nabla^2 f(x_k)\right) I\right) d_k^N$$

There is a step length of 1 associated with the Newton direction. Most implementations of Newton's method use the unit step $\alpha = 1$ where possible and adjust it only when it does not produce a satisfactory reduction in the value of f.
Methods that use the Newton direction have a fast rate of local convergence, typically quadratic. More formally, we have the following theorem

Theorem 19.2

Suppose f is twice differentiable and Hessian is ρ-Lipschitz continuous, i.e.,

$$\|\nabla^2 f(x) - \nabla^2 f(y)\|_2 \leq \rho \|x - y\|_2$$

around a solution $x^\star$ where the Hessian is positive definite. Then

1. If the starting point is close to $x^\star$, the sequence of iterates converges to $x^\star$.
2. The rate of convergence is quadratic.

§20 | Lec 19: May 12, 2021

§20.1 Newton's Method (Cont'd)

Proof. By definition of $k+1$-th step of Newton method, we get

$$\begin{aligned}
x_{k+1} - x^\star &= x_k + d_k^N - x^\star \\
&= x_k - x^\star + d_k^N \\
&= (x_k - x^\star) - \nabla^2 f(x_k)^{-1} \nabla f(x_k) \\
&= \nabla^2 f(x_k)^{-1} \nabla^2 f(x_k)(x_k - x^\star) - \nabla^2 f(x_k)^{-1} \left(\nabla f(x_k) - \underbrace{\nabla f(x^\star)}_{=0} \right) \\
&= \nabla^2 f(x_k)^{-1} \left[\nabla^2 f(x_k)(x_k - x^\star) - (\nabla f(x_k) - \nabla f(x^\star)) \right]
\end{aligned}$$

Using integral version of Taylor's theorem

$$\nabla f(x_k) - \nabla f(x^\star) = \int_0^1 \nabla^2 f\left(x^\star + t(x_k - x^\star)\right)(x_k - x^\star)\, dt$$

Therefore, we note that

$$\nabla^2 f(x_k)(x_k - x^\star) - (\nabla f(x_k) - \nabla f(x^\star)) = \nabla^2 f(x_k)(x_k - x^\star) - \int_0^1 \nabla^2 f\left(x^\star + t(x_k - x^\star)\right)(x_k - x^\star)\, dt$$

which is equal to

$$\int_0^1 \nabla^2 f(x_k)(x_k - x^\star)\, dt - \int_0^1 \nabla^2 f\left(x^\star + t(x_k - x^\star)\right)(x_k - x^\star)\, dt$$

Note that

$$\|Ax\|_2 \le \|A\|_2 \cdot \|x\|_2$$

$$\begin{aligned}
&\|\nabla^2 f(x_k)(x_k - x^\star) - (\nabla f(x_k) - \nabla f(x^\star))\|_2 \\
&= \| \int_0^1 \left[\nabla^2 f(x_k)(x_k - x^\star) - \nabla^2 f\left(x^\star + t(x_k - x^\star)\right)(x_k - x^\star) \right] dt \|_2 \\
&= \| \int_0^1 \left[\nabla^2 f(x_k) - \nabla^2 f\left(x^\star + t(x_k - x^\star)\right) \right](x_k - x^\star)\, dt \|_2 \\
&\le \int_0^1 \| \left[\nabla^2 f(x_k) - \nabla^2 f\left(x^\star + t(x_k - x^\star)\right) \right](x_k - x^\star) \|_2\, dt \\
&\le \int_0^1 \| \left[\nabla^2 f(x_k) - \nabla^2 f\left(x^\star + t(x_k - x^\star)\right) \right] \|_2 \|x_k - x^\star\|_2\, dt \\
&\le \int_0^1 \rho \|x_k - (x^\star + t(x_k - x^\star))\|_2 \|x_k - x^\star\|_2\, dt \\
&= \int_0^1 \rho \|x_k - x^\star\|_2^2 (1-t)\, dt = \frac{\rho}{2} \|x_k - x^\star\|_2^2
\end{aligned}$$

Now, we have

$$\begin{cases} \|x_{k+1} - x^\star\|_2 = \|\nabla^2 f(x_k)^{-1} \left[\nabla^2 f(x_k)(x_k - x^\star) - (\nabla f(x_k) - \nabla f(x^\star))\right] \|_2 \\ \|\left[\nabla^2 f(x_k)(x_k - x^\star) - (\nabla f(x_k) - \nabla f(x^\star))\right] \|_2 \leq \frac{\rho}{2}\|x_k - x^\star\|_2^2 \end{cases}$$

This implies that

$$\|x_{k+1} - x^\star\|_2 \leq \|\nabla^2 f(x_k)^{-1}\|_2 \cdot \frac{\rho}{2}\|x_k - x^\star\|_2^2 \tag{1}$$

Finally note when $x_k \to x^\star$, we have

$$\|\nabla^2 f(x_k)^{-1}\|_2 = \|\nabla^2 f(x^\star)^{-1}\|_2$$

Therefore, by the definition of continuity, there exists a small number ϵ s.t. for any

$$\|x_k - x^\star\|_2 \leq \epsilon$$

We have

$$\|\nabla^2 f(x_k)^{-1}\|_2 \leq 2\|\nabla^2 f(x^\star)^{-1}\|_2 \tag{2}$$

Then

$$\|x_{k+1} - x^\star\| \leq \rho\|\nabla^2 f(x^\star)^{-1}\|_2\|x_k - x^\star\|_2^2 = \mathcal{O}\left(\|x_k - x^\star\|_2^2\right)$$

Plugging (2) into (1)

$$\underbrace{\|x_{k+1} - x^\star\|_2}_{\tau_{k+1}} \leq 2\|\nabla^2 f(x^\star)^{-1}\|_2 \cdot \frac{\rho}{2}\|x_k - x^\star\|_2^2 = \rho\|\nabla^2 f(x^\star)^{-1}\|_2 \cdot \|x_k - x^\star\|_2^2$$

$$= \mathcal{O}\left(\underbrace{\|x_k - x^\star\|_2^2}_{\tau_k^2}\right)$$

□

When initialized away from a second-order optimal point, the Hessian matrix might not be positive definite, and the Newton direction defined by

$$\nabla^2 f(x_k) d_k^N = -\nabla f(x_k)$$

may not be a search direction. We can overcome this difficulty y replacing the Hessian matrix with a **positive definite approximation**.
Eigenvalue modification: assume the eigenvalue decomposition of the Hessian is available. To approximate an indefinite Hessian with a positive definite matrix, one option is to replace all negative eigenvalues with a small positive number δ.
The simplest idea of modifying the Hessian is to find a scalar $\tau > 0$ s.t. $\nabla^2 f(x_k) + \tau I$ is sufficiently positive definite.

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k) = x_k - \frac{1}{L}\nabla f(x_k)$$

Gradient descent can be derived from the quadratic upper-bound of $f(x_k + d)$

$$f(x_k + d) \leq f(x_k) + \nabla f(x_k)^\top d + \frac{L}{2}\|d\|_2^2 \coloneqq q_k(d)$$

The minimize $q_k(d)$ gives $d = -\frac{1}{L}\nabla f(x_k)$.
Newton's method is derived from exactly the second-order Taylor series approximation to $f(x_k + d)$

$$f(x_k + d) \leq f(x_k) + \nabla f(x_k)^\top d + \frac{1}{2} d^\top \left[\nabla^2 f(x_k)\right] d \coloneqq m_k(d)$$

Then minimize $m_k(d)$ gives $d = -\nabla^2 f(x_k)^{-1} \nabla f(x_k)$. We can view gradient descent is obtained from second-order Taylor series by replacing $\nabla^2 f(x_k)$ by LI while Newton's method uses exactly $\nabla^2 f(x_k)$. Therefore, Newton's method is more accurate.

§21 Lec 20: May 14, 2021

§21.1 Gradient Descent v.s. Newton's Method

Example 21.1

Consider minimizing $f(x,y) = 100x^2 + y^2$ with starting point $\begin{bmatrix} x_0 \\ y_0 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and $L = \lambda_{\max}(\nabla^2 f) = \lambda_{\max}\left(\begin{bmatrix} 200 & 0 \\ 0 & 2 \end{bmatrix}\right) = 200$.

We have

$$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{bmatrix} = \begin{bmatrix} 200x \\ 2y \end{bmatrix}$$

$$\nabla^2 f = \begin{bmatrix} 200 & 0 \\ 0 & 2 \end{bmatrix}$$

So $\lambda_1 = 200$, $\lambda_2 = 2$, and $L = 200 \implies f(x,y)$ is strongly convex

$$\nabla f = \begin{bmatrix} 200x \\ 2y \end{bmatrix} = 0$$

$$\implies \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \text{ is the unique global minimizer}$$

- Gradient Descent: $x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$

$$\begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} - \frac{1}{200}\begin{bmatrix} 200 \\ 2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0.99 \end{bmatrix}$$

$$\begin{bmatrix} x_2 \\ y_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0.99 \end{bmatrix} - \frac{1}{200}\begin{bmatrix} 0 \\ 2 \cdot 0.99 \end{bmatrix} = \begin{bmatrix} 0 \\ 0.99^2 \end{bmatrix}$$

$$\vdots$$

- Newton's Method: $x_{k+1} = x_k - \nabla^2 f(x_k)^{-1}\nabla f(x_k)$

$$\begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} - \begin{bmatrix} 200 & 0 \\ 0 & 2 \end{bmatrix}^{-1}\begin{bmatrix} 200 \\ 2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Conclusion: Newton's method needs much fewer iterations than the gradient descent method. However, gradient descent method still runs much faster than Newton's method since it does not need to compute the second-order derivative information.

§21.2 Subgradient Methods

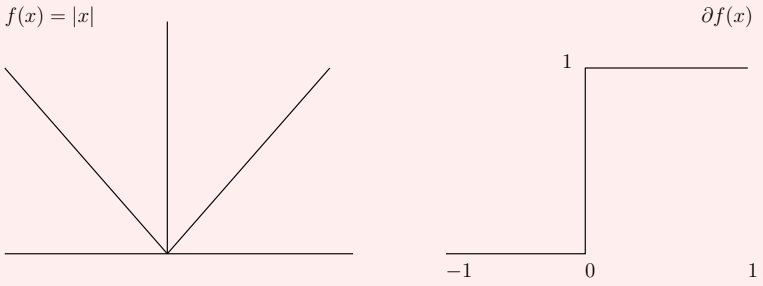
We first consider the simple case where the function $f : \mathbb{R}^n \to \mathbb{R}$ is convex. We say g is a subgradient of f at x if

$$f(z) \geq \underbrace{f(x) + g^\top (z - x)}_{\text{linear underestimate of } f}, \qquad \forall z$$

g is not unique. The set of all subgradients of f at x is called subdiffernetial of f at x, denoted by $\partial f(x) \coloneqq \{f(z) \geq f(x) + g^\top (z - x) \, \forall z\}$.

Example 21.2

Consider: $f(x) = |x|$, $x \in \mathbb{R}$.



So,

$$\partial f(x) = \begin{cases} \{-1\}, & x < 0 \\ [-1, 1], & x = 0 \\ \{1\}, & x > 0 \end{cases}$$

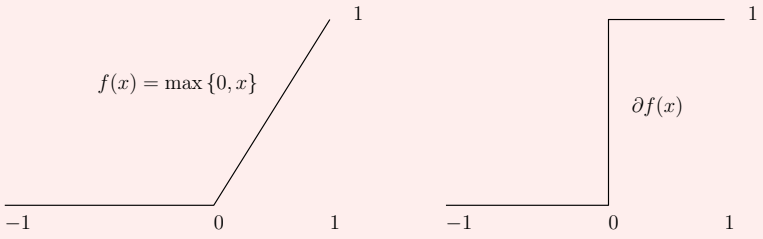
Subgradient of $|x|$ at $x = 0$:

$$|z| = f(z) \geq f(0) + g(z - 0), \qquad \forall z \in \mathbb{R}$$
$$\implies \text{find } g \ni |z| \geq gz, \qquad \forall z \in \mathbb{R}$$
$$\left.\begin{array}{l} z > 0, \quad z \geq gz \implies g \leq 1 \\ z < 0, \quad -z \geq gz \implies -1 \leq g \end{array}\right\} \implies -1 \leq g \leq 1$$

If $-1 \leq g \leq 1$, we have $|z| \geq gz$, $\forall z \in \mathbb{R}$.

Example 21.3

Consider: Rectified Linear Unit (ReLU) $f(x) = \max\{0, x\}$, the mostly used nonlinear function in deep learning



So,

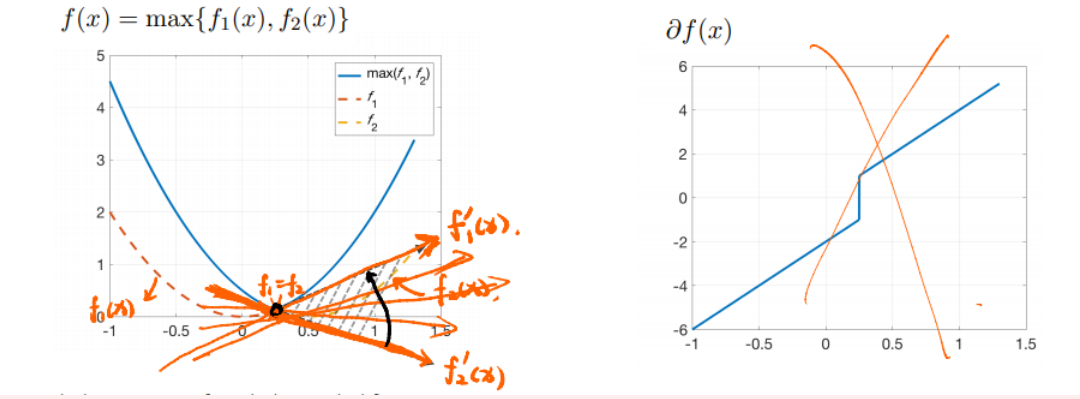
$$\partial f(x) = \begin{cases} \{0\}, & x < 0 \\ [0, 1], & x = 0 \\ \{1\}, & x > 0 \end{cases}$$

§22 Lec 21: May 17, 2021

§22.1 Subgradient Methods (Cont'd)

Example 22.1

Consider $f(x) = \max\{f_1(x), f_2(x)\}$



$f(x) = \max\{f_1(x), f_2(x)\}$ where f_1 and f_2 are differentiable

$$\partial f(x) = \begin{cases} f_2'(x), & f_1(x) < f_2(x) \\ [f_2'(x), f_1'(x)], & f_1(x) = f_2(x) \\ f_1'(x), & f_1(x) > f_2(x) \end{cases}$$

Basic Rules for Subdifferential:

- **Scaling**: $\partial(af) = a\partial f$, $a \geq 0$
- **Summation**: $\partial(f_1 + f_2) = \partial f_1 + \partial f_2$
- **Affine Transformation**: if $h(x) = f(Ax + b)$, then

$$\partial h(x) = A^\top \partial f(Ax + b)$$

Consider:

$$f(x) = \|x\|_1 = \sum_{i=1}^n |x_i| = \sum_{i=1}^n f_i(x)$$

and

$$\partial f(x) = \sum_{i=1}^{n} \partial f_i(x), \quad \partial_{x_i}|x_i| = \begin{cases} \text{sign}(x_i), & x_i \neq 0 \\ [-1,1], & x_i = 0 \end{cases}$$

$$\partial f_i(x) = \begin{bmatrix} \partial_{x_1}|x_i| \\ \partial_{x_2}|x_i| \\ \vdots \\ \partial_{x_n}|x_i| \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \partial_{x_i}|x_i| \\ \vdots \\ 0 \end{bmatrix}$$

$$= \partial_{x_i}|x_i| e_i$$

$$= \begin{cases} \text{sign}(x_i) e_i, & x_i \neq 0 \\ [-1,1]\, e_i, & x_i = 0 \end{cases} \quad = \partial f_i(x)$$

So,

$$[\partial f(x)]_i = \begin{cases} \text{sign}(x_i), & x_i \neq 0 \\ [-1,1], & x_i = 0 \end{cases}$$

$$\partial f(x) = \widetilde{\text{sign}}(x) \in \mathbb{R}^n$$

$$\left[\widetilde{\text{sign}}(x)\right]_i = \begin{cases} \text{sign}(x_i), & x_i \neq 0 \\ [-1,1], & x_i = 0 \end{cases}$$

Example 22.2

Consider:

$$x = \begin{bmatrix} 1 \\ 3 \\ -4 \\ 0 \end{bmatrix}$$

Then

$$\widetilde{\text{sign}}(x) = \begin{bmatrix} 1 \\ 1 \\ -1 \\ [-1,1] \end{bmatrix}$$

Example 22.3

Consider $h(x) = \|Ax+b\|_1$. Letting $f(x) = \|x\|_1$ and $A = [a_1, \dots, a_m]$, and in addition

$$Ax + b = \begin{bmatrix} a_1^\top x + b_1 \\ \vdots \\ a_m^\top x + b_m \end{bmatrix} = \begin{cases} \text{sign } (a_i^\top x + b_i), & a_i^\top x + b_i \neq 0 \\ [-1, 1], & a_i^\top x + b_i = 0 \end{cases}$$

and

$$\partial h(x) = A^\top \partial f\,(Ax + b)$$

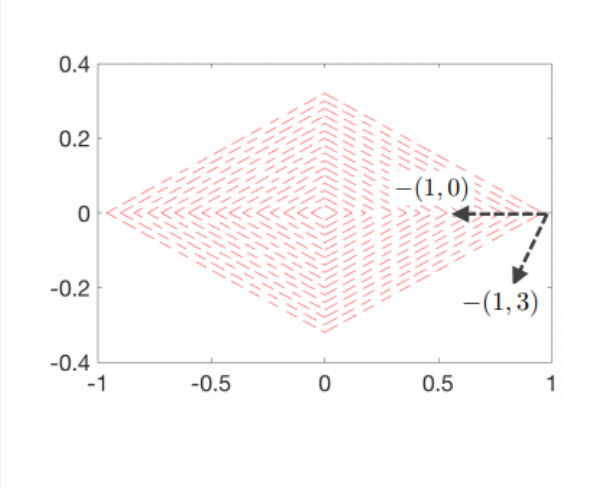
Consider minimize $f(x)$ which may not be continuously differentiable at some points. To mimic the gradient descent, subgradient methods involve the update

$$x_{k+1} = x_k - \alpha_k g_k, \qquad g_k \in \partial f(x_k)$$

for $k = 0, 1, \dots$. Unlike gradient, the negative subgradient is not necessarily a descent direction. This also makes it more difficult to select appropriate step sizes α_k for the subgradient methods.

Example 22.4

Consider $f(x) = |x_1| + 3|x_2|$



We have

$$\partial f(x) = \begin{bmatrix} \partial x_1 f(x) \\ \partial x_2 f(x) \end{bmatrix} = \begin{bmatrix} \widetilde{\text{sign}}(x_1) \\ 3\widetilde{\text{sign}}(x_2) \end{bmatrix}$$

At $x = (1, 0)$,

$$\begin{aligned} \partial f(1,0) &= \begin{bmatrix} 1 \\ [-3, 3] \end{bmatrix} \\ &= \left\{ \begin{bmatrix} 1 \\ t \end{bmatrix}, \quad t \in [-3, 3] \right\} \end{aligned}$$

Example 22.5 (Cont'd from above)

At $x = (1, 0)$,

- $g_1 = (1, 0) \in \partial f(x)$, and $-g_1$ is a descent direction.
- $g_2 = (1, 3) \in \partial f(x)$, and $-g_2$ is **not** a descent direction.

Reason: lack of continuity – one can change direction significantly without violating validity of subgradient.

§23 Lec 22: May 19, 2021

§23.1 Subgradient Methods (Cont'd)

Since $\{f(x_k)\}$ is not necessarily monotone, we will keep track of the best

$$f_k^\star = \min_{1\le i\le k} f(x_i)$$

We also denote $f^\star := \min_x f(x)$, the optimal objective value. We cannot analyze all non-smooth functions. A nice and widely encountered class to start with is Lipschitz functions, i.e., we assume f is Lipschitz

$$|f(x) - f(z)| \le L_f \|x - z\|_2 \qquad \forall x, z \in \mathrm{dom}(f)$$

Property: $\|g\| \le L_f$ for all subgradients $g \in \partial f(x)$ for all x.

Proof. From definition of subgradient,

$$f(z) \ge f(x) + g^\top (z - x) \qquad \forall z, x$$

which implies

$$\langle g, z - x\rangle = g^\top (z - x) \le f(z) - f(x)$$

Choose $z = x + g \implies \|g\|_2^2 = \langle g, g\rangle \le f(x + g) - f(x) \le L_f \|g\|_2$

$$\implies \|g\|_2 \le L_f \qquad \forall g \in \partial f(x), \quad \forall x$$ □

Subgradient methods with constant step size may converge to non-optimal point

Example 23.1

Consider $f(x) = |x|$, $x_0 = 0.1$, and step size $\alpha_k = \alpha = 0.08$. Then,

$$x_1 = x_0 - \alpha \widetilde{\mathrm{sign}}(x_0) = 0.1 - 0.08 \times 1 = 0.02$$
$$x_2 = x_1 - \alpha \widetilde{\mathrm{sign}}(x_1) = 0.02 - 0.08 \times 1 = -0.06$$
$$x_3 = x_2 - \alpha \widetilde{\mathrm{sign}}(x_2) = -0.06 - 0.08 \times (-1) = 0.02$$

Thus, the sequence of iterates $\{x_k\}$ never converge, and neither of its limit points is the optimal solution of $|x|$.

How to select appropriate step size? Polyak's step size, diminishing step sizes, etc ... We'd like to optimize $\|x_{k+1} - x^\star\|_2$ but don't have access to $x^\star$.

Lemma 23.2

Subgradient update rule $x_{k+1} = x_k - \alpha_k g_k$ obeys

$$\|x_{k+1} - x^\star\|_2^2 \le \|x_k - x^\star\|_2^2 - 2\alpha_k \left(f(x_k) - f^\star\right) + \alpha_k^2 \|g_k\|_2^2$$

Proof. We have

$$\begin{aligned}\|x_{k+1}-x^\star\|_2^2 &= \|x_k-\alpha_k g_k-x^\star\|_2^2 = \|(x_k-x^\star)-\alpha_k g_k\|_2^2\\ &= \|x_k-x^\star\|_2^2 - 2\alpha_k\langle x_k-x^\star, g_k\rangle + \alpha_k^2\|g_k\|_2^2\\ &\le \|x_k-x^\star\|_2^2 - 2\alpha_k\left(f(x_k)-f^\star\right) + \alpha_k^2\|g_k\|_2^2\end{aligned}$$

where the last line uses subgradient inequality

$$\begin{aligned}f(x_k)-f(x^\star) &\ge \langle x_k-x^\star, g\rangle\\ f(x_k)\ge f(x^\star)+g^\top(x_k-x^\star) \implies f(x_k)-f(x^\star) &\ge g^\top(x_k-x^\star)\end{aligned}$$

□

Majorizing function suggests Polyak's stepsize

$$\alpha_k = \frac{f(x_k)-f^\star}{\|g_k\|_2^2}$$

which leads to error reduction

$$\|x_{k+1}-x^\star\|_2^2 \le \|x_k-x^\star\|_2^2 - \frac{\left(f(x_k)-f^\star\right)^2}{\|g_k\|_2^2} \le \|x_k-x^\star\|_2^2$$

It's useful if $f^\star$ is known

$$\begin{aligned}h\left(\alpha_k=\frac{f(x_k)-f^\star}{\|g_k\|_2^2}\right) &= \|x_k-x^\star\|_2^2 - \frac{2\left(f(x_k)-f^\star\right)^2}{\|g_k\|_2^2} + \frac{\left(f(x_k)-f^\star\right)^2}{\|g_k\|_2^2}\\ &= \|x_k-x^\star\|_2^2 - \frac{\left(f(x_k)-f^\star\right)^2}{\|g_k\|_2^2}\end{aligned}$$

Estimation error is monotonically decreasing with Poylak's stepsize.

Theorem 23.3 (Convergence of Subgradient Method with Polyak's Stepsize)

Suppose f is convex and L_f-Lipschitz continuous. Then subgradient method with Polyak's stepsize rule obeys

$$f_k^\star - f^\star \le \frac{L_f\|x_0-x^\star\|_2}{\sqrt{k+1}}$$

Sublinear convergence rate $\mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$

Proof. We have

$$\begin{aligned}\left(f(x_k)-f^\star\right)^2 &\le \left(\|x_k-x^\star\|_2^2 - \|x_{k+1}-x^\star\|_2^2\right)\|g_k\|_2^2\\ &\le \left(\|x_k-x^\star\|_2^2 - \|x_{k+1}-x^\star\|_2^2\right)L_f^2\end{aligned}$$

Applying iterations (from 0-th to k-th) and summing up yield

$$\begin{aligned}\sum_{i=0}^{k}\left(f(x_i)-f^\star\right)^2 &\le \left(\|x_0-x^\star\|_2^2 - \|x_{k+1}-x^\star\|_2^2\right)L_f^2\\ \implies (k+1)\left(f_k^\star-f^\star\right)^2 &\le \|x_0-x^\star\|_2^2 L_f^2\end{aligned}$$

which completes the proof. □

§24 Lec 23: May 21, 2021

§24.1 Subgradient Methods (Cont'd)

Theorem 24.1 (Subgradient Method for Convex and Lipschitz Functions)
Suppose f is convex and L_f-Lipschitz continuous. Then subgradient method

$$x_{k+1} = x_k - \alpha_k g_k, \qquad g_k \in \partial f(x_k),\ k = 0, 1, 2, \dots$$

obeys

$$f_k^\star - f^\star \leq \frac{\|x_0 - x^\star\|_2^2 + L_f^2 \sum_{i=0}^k \alpha_i^2}{2\sum_{i=0}^k \alpha_i}$$

For constant step size $\alpha_k = \alpha$,

$$\lim_{k\to\infty} f_k^\star - f^\star \leq \frac{L_f^2 \alpha}{2}$$

i.e., may converge to a non-optimal point.
Diminishing step size obeys $\sum_k \alpha_k^2 < \infty$ and $\sum_k \alpha_k \to \infty$

$$\lim_{k\to\infty} f_k^\star - f^\star = 0,$$

i.e., converge to an optimal point.
The optimal choice here is $\alpha_k = \frac{1}{\sqrt{k}}$

$$f_k^\star - f^\star = \mathcal{O}\left(\frac{\|x_0 - x^\star\|_2^2 + L_f^2 \log k}{\sqrt{k}}\right)$$

i.e., matches the results with the Polyak's stepsize and attains ϵ-accuracy within about $\mathcal{O}\left(\frac{1}{\epsilon^2}\right)$ iterations.

Proof. Applying the lemma in lecture 22 recursively gives

$$\|x_{k+1} - x^\star\|_2^2 \leq \|x_0 - x^\star\|_2^2 - 2\sum_{i=0}^k \alpha_i \left(f(x_i) - f^\star\right) + \sum_{i=0}^k \alpha_i^2 \|g_i\|_2^2$$

Rearranging these terms, we have

$$\begin{aligned} 2\sum_{i=0}^k \alpha_i \left(f(x_i) - f^\star\right) &\leq \|x_0 - x^\star\|_2^2 - \|x_{k+1} - x^\star\|_2^2 + \sum_{i=0}^k \alpha_i^2 \|g_i\|_2^2 \\ &\leq \|x_0 - x^\star\|_2^2 + L_f^2 \sum_{i=0}^k \alpha_i^2 \\ \implies f_k^\star - f^\star &\leq \frac{\|x_0 - x^\star\|_2^2 + L_f^2 \sum_{i=0}^k \alpha_i^2}{2\sum_{i=0}^k \alpha_i} \end{aligned}$$

□

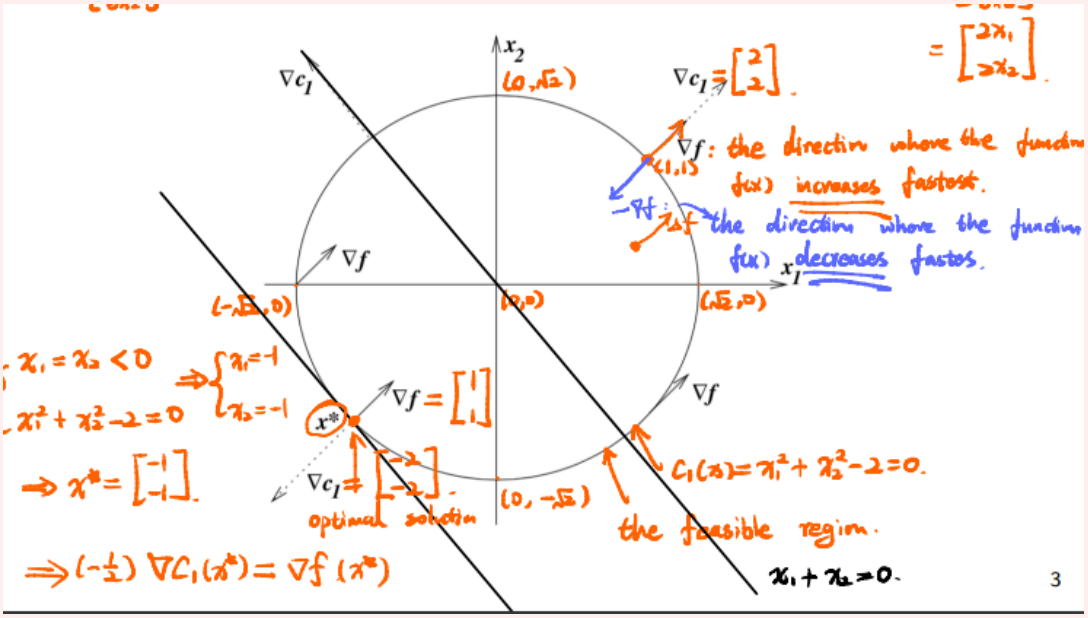
§24.2 Theory of Constrained Optimization

Overview:

$$\min_{x\in\mathbb{R}^n} f(x) \text{ subject to } \begin{cases} c_i(x) = 0, & i \in \mathcal{E} \\ c_i(x) \geq 0, & i \in \mathcal{I} \end{cases}$$

Example 24.2 (A Single Equality Constraint)

$\min_{x_1,x_2} x_1 + x_2$ subject to $x_1^2 + x_2^2 - 2 = 0$



At the solution $x^\star$, the constraint gradient $\nabla c_1(x^\star)$ is parallel to $\nabla f(x^\star)$

$$\nabla f(x^\star) = \lambda_1^\star \nabla c_1(x^\star) \quad \text{with} \quad \lambda_1^\star = -\frac{1}{2}$$

Lagrangian function:

$$\mathcal{L}(x, \lambda_1) = f(x) - \lambda_1 c_1(x)$$

The condition becomes: at $x^\star$, there is a scalar $\lambda_1^\star$ s.t.

$$\nabla_x \mathcal{L}(x^\star, \lambda_1^\star) = 0$$

The condition is only necessary, but not sufficient.