CO 367

272284444

September 2023

Contents

1 Review

1.1 Calculus

Definition 1.1 – Norm
$\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$ is a mapping from $\mathbb{R}^n$ to $\mathbb{R}$ such that

- For every $\mathbf{x} \in \mathbb{R}^n$, $\|\mathbf{x}\| \geq 0$
- $\|\mathbf{x}\| = 0$ if and only if $\mathbf{x} = 0$.
- For every $\alpha \in \mathbb{R}$ and for every $\mathbf{x} \in \mathbb{R}^n$, $\|\alpha\mathbf{x}\| = |\alpha|\|\mathbf{x}\|$
- (Triangle inequality) For every $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$,

$$\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$$

Two common norms, ℓ_1 and ℓ_2 norms:

1. ℓ_1: $\|\mathbf{x}\|_1 = \sum_{i=1}^n |x_i| = |x_1| + |x_2| + \cdots + |x_n|$
2. ℓ_2: $\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^n x_i^2} = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}$

Proposition 1.1 – Cauchy-Schwarz Inequality
For every $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$

$$|\mathbf{x}^T\mathbf{y}| = \langle \mathbf{x}, \mathbf{y} \rangle \leq \|\mathbf{x}\|_2 \|\mathbf{y}\|_2$$

with equality if and only if $\mathbf{x} = \alpha\mathbf{y}$ for some $\alpha \in \mathbb{R}$.

Definition 1.2 – Neighborhood/Open ball
Given $\delta > 0$, $\bar{\mathbf{x}} \in \mathbb{R}^n$, the open ball $B_\delta(\bar{\mathbf{x}}) = \{\mathbf{x} \in \mathbb{R}^n \mid \|\mathbf{x} - \bar{\mathbf{x}}\|_2 < \delta\}$

Definition 1.3 – Sequence Convergence in $\mathbb{R}^n$
We say that a sequence $\{\mathbf{x}_k\} \subseteq \mathbb{R}^n$ converges to $\mathbf{x}^* \in \mathbb{R}^n$ and write

$$\lim_{k \to \infty} \mathbf{x}_k = x^*$$

if for every $\epsilon > 0$ there exists $N \in \mathbb{N}$ such that for every $k \geq N$, $\|\mathbf{x}_k - \mathbf{x}^*\| < \epsilon$.

Definition 1.4 – Limit Point
If $\{\mathbf{x}_k\}$ has a subsequence that converges to $\mathbf{x}^*$, then $\mathbf{x}^*$ is called a limit point of $\{\mathbf{x}_k\}$.
Given a set $E \subseteq \mathbb{R}^n$, if there exists a sequence $\{\mathbf{x}_k\} \subseteq E$ that converges to $\mathbf{x}^*$, then $\mathbf{x}^*$ is called a limit point of E.

Definition 1.5 – Closed Set
A set $E \subseteq \mathbb{R}^n$ is closed if it contains all of its limit points.
That is, for every sequence $\{\mathbf{x}_k\} \subseteq E$ with

$$\lim_{k \to \infty} \mathbf{x}_k = \mathbf{x}^*$$

if $\mathbf{x}^* \in E$, then E is closed.

Definition 1.6 – Interior Point
A point $\mathbf{x} \in E$ is an interior point of E is there is a neighborhood/open ball of $\mathbf{x}$ that is contained in E.

Definition 1.7 – Open Set
A set E is open if all of its elements are interior points.
So, for any point you pick in E, you can find a small neighborhood around that point which is entirely contained in E (there are no boundary points in E).

Theorem 1.1 – Properties of Open/Closed Sets
The followings hold:

- A set is closed (open) if its complement, $\mathbb{R}^n \setminus E$, is open (closed).
- Union of finitely many closed sets is closed
- Intersection of (finitely or infinitely many) closed sets is closed.
- Intersection of finitely many open sets is open
- Union of (finitely or infinitely many) open sets is open.

Definition 1.8 – Bounded
A set $E \subset \mathbb{R}^n$ is bounded if it can be contained in a ball of finite radius. That is, there exists a neighborhood, $B_\delta(\mathbf{x})$, such that $E \subseteq B_\delta(\mathbf{x})$.

Definition 1.9 – Compact
A set $E \subset \mathbb{R}^n$ is compact if it is closed and bounded.

Definition 1.10 – Lipschitz Continuous/Contraction
Let $E \subseteq \mathbb{R}^n$. Let $f : E \to \mathbb{R}^m$ be a function. We say f is Lipschitz continuous on E if there exists a constant $L > 0$ such that for all $\mathbf{x}, \mathbf{y} \in E$,

$$\|f(\mathbf{x}) - f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|$$

Theorem 1.2 – Continuity of Functions
Let $E \subseteq \mathbb{R}^n$, $f, g : E \to \mathbb{R}^m$ and $\alpha \in \mathbb{R}$. If f and g are continuous at $\mathbf{x}_0$, then

1. $f + g, fg, \alpha f$ are continuous at x_0
2. $\frac{f}{g}$ is continuous at $\mathbf{x}_0$ provided that $g(\mathbf{x}_0) \neq 0$.

Definition 1.11 – Formal Definition of Derivative
The **derivative** of f at a is defined as

$$f'(a) = \lim_{x \to a} \frac{f(x) - f(a)}{x - a}$$

if the limit exists.

An alternate definition is

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

Theorem 1.3 – Extreme Value Theorem (EVT)
If $E \subset \mathbb{R}^n$ is a compact set (closed and bounded) and $f : E \to \mathbb{R}$ is a continuous function, then f attains a maximum and minimum on E. That is, there exists points $\mathbf{x}_{\text{min}}, \mathbf{x}_{\text{max}} \in E$ such that for all $\mathbf{x} \in E$,

$$f(x) \leq f(\mathbf{x}_{\text{max}}) \quad \text{and} \quad f(\mathbf{x}_{\text{min}}) \leq f(\mathbf{x})$$

Theorem 1.4 – Continuously Differentiable
f is continuously differentiable at $\mathbf{x}_0$ if all partial derivatives exist and are continuous in a neighborhood of $\mathbf{x}_0$. We say f is continuously differentiable, $f \in C^1$, if its partial derivatives are continuous everywhere.

f is twice differentiable on E if $\nabla^2 f(x)$ exists for all $\mathbf{x} \in E$. If each entry of the Hessian $\nabla^2 f(\mathbf{x})$ is continuous, we say f is twice differentiable on E, $f \in C^2$.

Theorem 1.5
If $f : E \to \mathbb{R}$ is twice differentiable, then the Hessian is symmetric.

Theorem 1.6 – Mean Value Theorem (MVT)
Let $f : \mathbb{R} \to \mathbb{R}$ be continuous on the closed interval $[a, b]$ and differentiable on the open interval (a, b). Then there exists $c \in (a, b)$ such that

$$f(b) - f(a) = f'(c)(b - a)$$

Lemma 1.1 – Directional Derivative
Let $f : \mathbb{R}^n \to \mathbb{R}, \bar{x}, d \in \mathbb{R}^n$ where d is the direction. We define

$$\phi(\epsilon) = f(\bar{x} + \epsilon d) : \mathbb{R} \to \mathbb{R}$$

the value of the function f at a point that is displaced from $\bar{x}$ by a distance of ϵ in the direction d. Then the **directional derivative**, denoted $f'(x; d)$ of f at x at the direction d is

$$f'(x; d) = \phi'(0) = \lim_{\epsilon \to 0} \frac{f(x + \epsilon d) - f(x)}{\epsilon} = \nabla f(x)^T d$$

Definition 1.12 – Directional Derivative (Different notation)
The directional derivative of f at a point $\bar{x}, \in \mathbb{R}^n$ in the direction d is

$$f'(\bar{x}; d) = \frac{d}{ds} f(\bar{x} + sd) \bigg|_{s=0}$$

Theorem 1.7
If f is differentiable at $\bar{x}$, then

$$f'(\bar{x}; d) = \nabla f(\bar{x})^T d$$

Proof. We only prove in the case where $\bar{x} = (a, b) \in \mathbb{R}^2$.

$$\begin{aligned}
f'(\bar{x}; d) &= \frac{d}{ds} f(\bar{x} + sd)\Big|_{s=0} \\
&= \frac{d}{ds} f(\underbrace{a + sd_1}_{x}, \underbrace{b + sd_2}_{y})\Big|_{s=0} \\
&= \left[\frac{\partial f}{\partial x}\frac{dx}{ds} + \frac{\partial f}{\partial y}\frac{dy}{ds}\right]_{s=0} && \text{Chain rule} \\
&= \left[\frac{\partial}{\partial x} f(a + sd_1, b + sd_2) \cdot d_1 + \frac{\partial}{\partial y} f(a + sd_1, b + sd_2) \cdot d_2\right]_{s=0} \\
&= \frac{\partial f}{\partial x}(a, b) \cdot d_1 + \frac{\partial f}{\partial y}(a, b) \cdot d_2 \\
&= \left(\frac{\partial f}{\partial x}(a, b), \frac{\partial f}{\partial y}(a, b)\right) \cdot (d_1, d_2) \\
&= \nabla f(a, b) \cdot d
\end{aligned}$$

□

Problem 1.1

Let $f(x, y, z) = x^2 z + y^3 z^2 - xyz$ with $d = \begin{pmatrix} -1 \\ 0 \\ 3 \end{pmatrix}$ Then the **directional derivative** in the direction d is

$$\nabla f(x, y, z)^T d = \begin{pmatrix} 2xz - yz \\ 3y^2 z^2 - xz \\ x^2 + 2y^3 z - xy \end{pmatrix}^T \begin{pmatrix} -1 \\ 0 \\ 3 \end{pmatrix} = -2xz + yz + 3x^2 + 6y^3 z - 3xy$$

Theorem 1.8 – Taylor's Theorem on the real line

Let $f : (a, b) \to \mathbb{R}$, and $\bar{x}, x \in (a, b)$. Then the Taylor's series centered at $\bar{x}$ (approximation near $\bar{x}$) is

$$f(x) = f(\bar{x}) + f'(\bar{x})(x - \bar{x}) + \frac{f''(z)}{2}(x - \bar{x})^2$$

where z is between x and $\bar{x}$ that gives the largest value of $f''(z)$. The term $\frac{f''(z)}{2}(x - \bar{x})^2$ is the **error term**

Equivalently,

$$f(\bar{x} + \Delta x) = \underbrace{f(x) + f'(x)\Delta x}_{\text{Linear approximation}} + o(|\Delta x|)(\text{little O})$$

This formula emphasizes its use in approximating changes in f for small changes in x, denoted Δx. $o(|\Delta x|)$, the error term, means that the error goes to 0 faster than $|\Delta x|$ as Δx goes to 0. Therefore, this is saying that the linear approximation becomes more and more accurate for smaller Δx.

Theorem 1.9 – Multivariate Taylor

Consider a C^2-smooth function $f : U \to \mathbb{R}$ on an open set $U \subset \mathbb{R}^n$. If $\bar{x}$ and x are such that the segment $[\bar{x}, x] := \{\bar{x} + t(x - \bar{x}) : t \in [0, 1]\}$ is contained in U, then the Taylor series expansion of f centered around $\bar{x}$ is

$$f(x) = f(\bar{x}) + \langle \nabla f(\bar{x}), x - \bar{x} \rangle + \frac{1}{2} \langle \nabla^2 f(z)(x - \bar{x}), (x - \bar{x}) \rangle$$

where z is between x and $\bar{x}$.

Theorem 1.10 – Taylor's Theorem
Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable and let $\mathbf{d} \in \mathbb{R}^n$ be a direction vector. Then

$$f(\mathbf{x} + \mathbf{d}) = f(\mathbf{x}) + \underbrace{\nabla f(\mathbf{x} + t\mathbf{d})^T \mathbf{d}}_{\text{directional derivative}}$$

for some $t \in (0, 1)$. Moreover, if f is twice differentiable, then

$$f(\mathbf{x} + \mathbf{d}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^T \mathbf{d} + \frac{1}{2} \underbrace{\mathbf{d}^T \nabla^2 f(\mathbf{x} + t\mathbf{d}) \mathbf{d}}_{\text{second order directional derivative}}$$

for some $t \in (0, 1)$.

This theorem provides a way to approximate the value of the function f at a point $\mathbf{x} + \mathbf{d}$ based on the value and derivatives of f at or near the point $\mathbf{x}$.

Alternatively, we can write

$$f(x + d) = f(x) + \langle \nabla f(x), d \rangle + o(\|d\|)$$

and

$$f(x + d) = f(x) + \langle \nabla f(x), d \rangle + \frac{1}{2} \langle \nabla^2 f(x) d, d \rangle + o(\|d\|^2)$$

for small d.

Theorem 1.11 – Taylor's Thoerem (alternative)
Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable at $\mathbf{x}^*$. Then $\forall x \in \mathbb{R}^n$

$$f(\mathbf{x}) = \underbrace{f(\mathbf{x}^*) + \nabla f(\mathbf{x}^*)^T (\mathbf{x} - \mathbf{x}^*)}_{\text{linear approximation of } f \text{ at } \mathbf{x}^*} + o(\|\mathbf{x} - \mathbf{x}^*\|)$$

where $o(\|\mathbf{x} - \mathbf{x}^*\|)$ is the error term that goes to 0 faster than $\|\mathbf{x} - \mathbf{x}^*\|$ as $\mathbf{x} \to \mathbf{x}^*$.
Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable at $\mathbf{x}^*$. Then $\forall x \in \mathbb{R}^n$

$$f(\mathbf{x}) = \underbrace{f(\mathbf{x}^*) + \nabla f(\mathbf{x}^*)^T (\mathbf{x} - \mathbf{x}^*) + \frac{1}{2} (\mathbf{x} - \mathbf{x}^*)^T \nabla^2 f(\mathbf{x}^*) (\mathbf{x} - \mathbf{x}^*)}_{\text{quadratic approximation of } f \text{ at } \mathbf{x}^*} + o(\|\mathbf{x} - \mathbf{x}^*\|^2)$$

where $o(\|\mathbf{x} - \mathbf{x}^*\|^2)$ is the error term that goes to 0 faster than $\|\mathbf{x} - \mathbf{x}^*\|^2$ as $\mathbf{x} \to \mathbf{x}^*$.

Problem 1.2 – Lagrange Multiplier Example
Maximize $f(x, y) = x^2 y$ subject to $g(x, y) = x^2 + y^2 = 1$. By using Lagrange multipliers, we know that the maximizer, (x^*, y^*) satisfies

$$\nabla f(x^*, y^*) = \lambda \nabla g(x^*, y^*)$$

We have

$$\nabla g(x,y) = \begin{bmatrix} 2x \\ 2y \end{bmatrix}$$
$$\nabla f(x,y) = \begin{bmatrix} 2xy \\ x^2 \end{bmatrix}$$

Then, using Lagrange multipliers, we solve

$$\begin{bmatrix} 2xy \\ x^2 \end{bmatrix} = \lambda \begin{bmatrix} 2x \\ 2y \end{bmatrix}$$

Since we also have the constraint $x^2 + y^2 = 1$, we solve the system of equations

$$\begin{aligned} 2xy &= 2\lambda x \\ x^2 &= 2\lambda y \\ x^2 + y^2 &= 1 \end{aligned}$$

Solving it, we get

$$(x,y) = \left(\pm\sqrt{\frac{2}{3}}, \pm\sqrt{\frac{1}{3}}\right), \quad \lambda = y$$

Testing each point, we get that $\left(\sqrt{\frac{2}{3}}, \sqrt{\frac{1}{3}}\right)$ is the maximizer of f.

1.2 Linear Algebra

Definition 1.13 – Characteristic Polynomial
The characteristic polynomial of $A \in \mathbb{R}^{n\times n}$ is

$$p(\lambda) = \det(A - \lambda I)$$

The degree of $p(\lambda)$ is n, and the leading term is $(-1)^n \lambda^n$.
The eigenvalues of A are the roots of the characteristic polynomial.

Definition 1.14 – Eigenvectors and Eigenvalues
$0 \neq v \in \mathbb{R}^n$ is an **eigenvector** of A if there exists $\lambda \in \mathbb{R}$ such that $Av = \lambda v$. The number λ is called an **eigenvalue** of A.

Theorem 1.12 – Finding Eigenvectors and Eigenvalues
Let A be a matrix.

1. Set up the characteristic equation. We find

$$\det(A - \lambda I) = 0$$

2. Solve for λ. These are the eigenvalues.
3. Plug eigenvalues $\lambda_1, \ldots, \lambda_n$ into $(A - \lambda I)v = 0$ and solve for v. These are the eigenvectors.

Definition 1.15 – Quadratic Form

Let A be a symmetric matrix and $x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$. The **quadratic form** Q of the matrix A is defined as

$$Q = x^T A x$$

Problem 1.3 – Example

Consider the matrix $A = \begin{bmatrix} 5 & -5 \\ -5 & 1 \end{bmatrix}$. The quadratic form of A is

$$Q(x) = 5x_1^2 - 10x_1x_2 + x_2^2$$

Definition 1.16 – Classification of Quadratic Forms

Let $Q = x^T A x$ be a quadratic form of a matrix A. Then A is

1. positive definite if $Q(x) > 0$ for all non-zero vectors x, and $Q(x) = 0$ if and only if $x = 0$. Or all eigenvalues of A are positive. Denoted by $A \succ 0$.
2. positive semidefinite if $Q(x) \geq 0$ for all vectors x, with $Q(x) = 0$ occurring for some non-zero vectors x. Or all eigenvalues of A are non-negative. Denoted by $A \succeq 0$.
3. negative definite if $Q(x) < 0$ for all non-zero vectors x, and $Q(x) = 0$ if and only if $x = 0$. Or all eigenvalues of A are negative. Denoted by $A \prec 0$.
4. negative semidefinite if $Q(x) \leq 0$ for all vectors x, with $Q(x) = 0$ occuring for some non-zero vectors x. Or all eigenvalues of A are non-negative. Denoted by $A \preceq 0$.
5. indefinite if $Q(x)$ can be positive or negative. Or there are positive and negative eigenvalues for A.

Theorem 1.13 – Orthogonal Spectral Decomposition

Let $A \in \mathbb{S}^n$. Then A has an **orthogonal spectral decomposition**

$$A = \sum_i \lambda_i u_i u_i^T = UDU^T$$

where U is orthogonal with the orthogonal eigenvectors u_i as columns and D is a diagonal matrix with real eigenvalues on the diagonal.

Proposition 1.2

Let $A \in \mathbb{S}^n$. The following are equivalent (Positive definite):

1. $A \succ 0$.
2. All the eigenvalues of A are in $\mathbb{R}^n_{++}$, the interior of the nonnegative orthant.
3. A has a real symmetric positive definite square root, $A = SS, S \in \mathbb{S}^n_{++}$.
4. A has a lower triangular factorization, a Cholesky factorization, $A = LL^T$ and L has positive diagonal elements.
5. All principal minors of A are positive.

6. All leading principal minors of A are positive.

And the following are equivalent (Positive semidefinite):

1. $A \succeq 0$.
2. All the eigenvalues of A are in $\mathbb{R}^n_+$, the nonnegative orthant.
3. A has a real symmetric square root, $A = SS, S \in \mathbb{S}^n$.
4. A has a lower triangular factorization, a Cholesky factorization, $A = LL^T$.
5. All principal minors of A are nonnegative.

Definition 1.17 – Principal Submatrices
Let

$$A = \begin{bmatrix} 1 & 1 & 2 & 7 \\ 1 & 1 & 4 & 6 \\ 2 & 4 & 7 & 8 \\ 7 & 6 & 8 & 1 \end{bmatrix}, \quad I = \{1, 3\}, \quad A[I] = \begin{bmatrix} 1 & 2 \\ 2 & 7 \end{bmatrix}$$

Then $A[I]$ is a **principal submatrix** of A.

Definition 1.18 – Principal Minors
Let $A \in \mathbb{S}^n$, where $\mathbb{S}^n$ is the set of all symmetric $n \times n$ matrices.

1. $\det(A[I])$ is called the **principal minor** of A.
2. If $I = \{1, \dots, k\}$ then $\det(A[I])$ is called the **leading principal minor** of A.

Proposition 1.3 – Characterizing Positive Definiteness with Principal Minors
Let $A \in \mathbb{S}^n$. Then

1. $A \succeq 0 \iff \det(A[I]) \geq 0$ for all principal minors $\det(A[I])$.
2. $A \succ 0 \iff \det(A[I]) > 0$ for all **leading** principal minors $\det(A[I])$.

Theorem 1.14
Let $A \in \mathbb{R}^{n \times n}$ be symmetric. Then the following are equivalent:

1. A is positive semidefinite (definite).
2. All eigenvalues of A are nonnegative (positive).
3. A can be factored as $A = BB^T$ where B is an $n \times p$ matrix for some p. (Cholesky factorization)

Definition 1.19 – Diagonally Dominant
A matrix A is diagonally dominant if for every i,

$$|a_{ii}| \geq \sum_{j \neq i} |a_{ij}|$$

So each diagonal element is greater than or equal to the sum of the absolute values of the other elements in the same row.

It is called strictly diagonally dominant if for every i,

$$|a_{ii}| > \sum_{j \neq i} |a_{ij}|$$

Proposition 1.4
Let $A \in \mathbb{R}^{n \times n}$ be a symmetric, diagonally dominant matrix whose diagonal entries are nonnegative, then A is positive semidefinite.

Let $A \in \mathbb{R}^{n \times n}$ be a symmetric, strictly diagonally dominant matrix whose diagonal entries are positive, then A is positive definite.

Note that the converse is not true.

Definition 1.20 – Four Fundamental Subspaces
Let A be a $m \times n$ matrix.

- The range space of A is defined as $\text{Range}(A) = \{Ax : x \in \mathbb{R}^n\} \subseteq \mathbb{R}^m$.
- The range space of A^T is defined as $\text{Range}(A^T) = \{A^T y : y \in \mathbb{R}^m\} \subseteq \mathbb{R}^n$.
- The null space of A is defined as $\text{Null}(A) = \{x \in \mathbb{R}^n : Ax = 0\} \subseteq \mathbb{R}^n$.
- The null space of A^T is defined as $\text{Null}(A^T) = \{y \in \mathbb{R}^m : A^T y = 0\} \subseteq \mathbb{R}^m$.

Theorem 1.15 – Rank Nullity Theorem
Let A be an $m \times n$ matrix. Then

$$\text{rank}(A) + \dim(\text{Null}(A)) = n$$

Recall that rank of a matrix is the number of pivots in the reduced row echelon form of the matrix. The dimension of a subspace is the number of linearly independent vectors that span the subspace.

2 Introduction

2.1 Lecture 1-Preliminaries

Definition 2.1 – Big O and little o
Big O is basically the rate of growth of that function. A function $f(n)$ is of order 1, or $O(1)$ if there exists some non zero constant c such that

$$\frac{f(n)}{c} \to 1$$

as $n \to \infty$.
Little o is the upper bound of the rate of growth of that function. Therefore, a function $f(n)$ is of order 1, or $o(1)$ if for all constants $c > 0$,

$$\frac{f(n)}{c} \to 0$$

as $n \to \infty$.

Definition 2.2 – Differentiability Based on Big o and Little o
If f is differentiable at $x = a$, then

$$f(a + h) = f(a) + f'(a)h + o(h)$$

Conversely, if there exists constants A and B such that

$$f(a+h) = A + Bh + o(h)$$

then f is differentiable at $x = a$. Moreover, $A = f(a)$ and $B = f'(a)$.

Definition 2.3 – Product Rule
If f, g are differentiable at $x = a$, then

$$f(a+h) = f(a) + f'(a)h + o(h), \quad g(a+h) = g(a) + g'(a)h + o(h)$$

Then

$$\begin{aligned} p(a+h) &= f(a+h)g(a+h) \\ &= f(a)g(a) + [f(a)g'(a) + g(a)f(a)]h + o(h) \end{aligned}$$

Then by above theorem, $p = fg$ is differentiable at $x = a$, and $p'(a) = f(a)g'(a) + g(a)f'(a)$.

Definition 2.4 – Chain Rule
WIP

Definition 2.5 – Inner Product Space
Let $x \in \mathbb{R}^n$, represented as:

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

The inner product space is defined as:

$$\langle x, y \rangle = x^T y = \sum_{i=1}^{n} x_i y_i \quad \text{(dot product)}$$

The angle between vectors x and y is given by $\cos(\theta) = \frac{\langle x,y \rangle}{\|x\|}$.
With corresponding norm to be the Euclidean Norm

Definition 2.6 – map
Suppose the map $f : \mathbb{R}^n - > \mathbb{R}$.

Definition 2.7 – differ
We define f to be in C^1, C^2 on an open set $D \subseteq \mathbb{R}^n$, denoted $f \in C^1(D), C^2(D)$, respectively, if the partial first $\frac{\partial f(x)}{\partial x_i}$ and second $\frac{\partial^2 f(x)}{\partial x_i \partial x_j}$ derivatives exist and are continuous for all i, j, respectively. We then get the gradient vector in $\mathbb{R}^n$ and the $n \times n$ symmetric Hessian matrix, respectively denoted as:

$$\nabla f(x) = \left(\frac{\partial f(x)}{\partial x_i} \right) \in \mathbb{R}^n, \quad \nabla^2 f(x) = \left[\frac{\partial^2 f(x)}{\partial x_i \partial x_j} \right] \in \mathbb{S}^n.$$

Here, $\mathbb{S}^n$ is the vector space of $n \times n$ symmetric matrices.

Definition 2.8 – General Nonlinear opt. function NLO
The general problem of nonlinear optimization, denoted NLO, is defined as follows: Given C²-smooth functions $f, g_i, h_j : D \subseteq \mathbb{R}^n \to \mathbb{R}$ for $i = 1, \ldots, m$ and $j = 1, \ldots, p$, where D is an open subset of $\mathbb{R}^n$, the objective is to find the optimal value p^* and an optimum x^* of NLO, represented as:

$$p^* := \min f(x) \text{s.t.} \quad g_i(x) \leq 0, \quad \forall i = 1, \ldots, m h_j(x) = 0, \quad \forall j = 1, \ldots, p x \in D$$

If f, g_i, h_i are all **affine** function and D=R^2, then we have an LP

Definition 2.9 – affine

$$f(x) = Ax + b \tag{1}$$

where b≠0

Definition 2.10 – Types of Minimality
Consider $f : \mathbb{R}^n \to \mathbb{R}$ and $D \subset \mathbb{R}^n$. Then $\bar{x} \in D$ is:

- a *global minimizer* for f on D if $f(\bar{x}) \leq f(x)$ for all $x \in D$.
- a *strict global minimizer* for f on D if $f(\bar{x}) < f(x)$ for all $x \in D$ where $x \neq \bar{x}$.
- a *local minimizer* for f on D if there exists $\delta > 0$ such that $f(\bar{x}) \leq f(x)$ for all $x \in D \cap B_\delta(\bar{x})$.
- a *strict local minimizer* for f on D if there exists $\delta > 0$ such that $f(\bar{x}) < f(x)$ for all $x \in D \cap B_\delta(\bar{x})$ where $x \neq \bar{x}$.

Definition 2.11 – Linear Approximation
Suppose f is a function that is differentiable on an interval I containing the point a. The **linear approximation** to f at a is the linear function

$$L(x) = f(a) + f'(a)(x - a)$$

for $x \in I$.

Definition 2.12 – Quadratic Approximation
Similar as above, the **quadratic approximation** to f at a is the quadratic function

$$Q(x) = f(a) + f'(a)(x - a) + \frac{1}{2} f''(a)(x - a)^2$$

for $x \in I$.

2.2 Lecture 2

Definition 2.13 – General NLO/NLP
A **Non-linear Optimization Problem** (NLP) is of the following form:

$$\underbrace{p^*}_{\text{Optimal Value}} = \min \underbrace{f(x)}_{\text{Objective function}}$$

s.t.

$$g(x) = (g_i(x)) \leq 0 \in \mathbb{R}^m$$
$$h(x) = (h_j(x)) = 0 \in \mathbb{R}^p$$

Problem 2.1 – Example

$$\min (x_1 - 2)^2 + (x_2 - 1)^2$$

s.t.

$$x_1^2 - x_2 \leq 0 \qquad (g_1(x) \leq 0)$$
$$x_1 + x_2 - 2 \leq 0 \qquad (g_2(x) \leq 0)$$

Definition 2.14 – Contour
For $\alpha \in \mathbb{R}$, the **contour** of a function f is

$$C_\alpha = \{x \in \mathbb{R}^n : f(x) = \alpha\}$$

Definition 2.15 – Feasible Set
The **feasible set** is

$$F = \{x \in \mathbb{R}^n : g(x) \leq 0, h(x) = 0, x \in D\}$$

(Is D the domain??)

Definition 2.16 – Gradient
The **gradient** of f is

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

For the optimal solution x^*, we have

$$\alpha \nabla f(x^*) = \lambda_1 \nabla g_1(x^*) + \lambda_2 \nabla g_2(x^*)$$

for some $\alpha, \lambda_1, \lambda_2 \in \mathbb{R}$.
We will see later that we can choose $\alpha = 1$ and we need $\lambda_1 \geq 0, \lambda_2 \geq 0$.

Problem 2.2 – Max-cut Problem
Given a weighted graph $G = (\underbrace{V}_{\text{vertices}}, \underbrace{E}_{\text{edges}}, \underbrace{w}_{\text{weight}})$, a **cut** is $U \subseteq V, U \neq \emptyset$. The objective function

$$\max \quad \frac{1}{2} \sum_{\substack{i \in U, j \notin U \\ (i,j) \in E}} w_{i,j}$$

maximizes the sum of edges in a cut.
Formulating as an NLP, we introduce variables $x_i \in \{\pm 1\}, i = 1, \dots, n$. Then the Max-cut problem (MC) is as

follows:

$$\max \quad \frac{1}{2}\sum_{ij\in E} w_{ij}(1-x_ix_j)$$

Why 1/2 s.t.

$$x_i \in \{\pm 1\} \quad (\text{equivalent to } x_i^2 = 1) \quad \forall i = 1,\dots,n$$

This works because

$$1 - x_ix_j = \begin{cases} 0 & \text{if } x_i = x_j \quad (i,j \text{ in the same set, } U \text{ or } U^c) \\ 2 & \text{otherwise} \end{cases}$$

MC is a **quadratically constrained quadratic program** (QQP) since each constraint $x_i \in \{-1,1\}$ is equivalent to the quadratic constraint $x_i^2 = 1$. Note that MC is an NP-hard problem.

3 Unconstrained Optimization

3.1 Lecture 2

Problem 3.1 – Simplest Case - No Constraints
Let $\Omega \subseteq \mathbb{R}^n$ be an open set. Assume f is sufficiently smooth (differentiable) then the NLP with no constraints is

$$\min_{x\in\Omega} \quad f(x)$$

Definition 3.1 – Secant Line
A secant line is a line that connects two points on a function.

Theorem 3.1 – Chain Rule (2 dimensions)
Let $f:\mathbb{R}^2 \to \mathbb{R}$, and two other functions $x(t):\mathbb{R}\to\mathbb{R}$ and $y(t):\mathbb{R}\to\mathbb{R}$. Let $\phi(t) = f(x(t), y(t))$. The chain rule then states

$$\frac{d\phi}{dt} = \frac{\partial f}{\partial x}\frac{dx}{dt} + \frac{\partial f}{\partial y}\frac{dy}{dt}$$

Problem 3.2 – Example of Chain Rule in 2 Dimensions
Let $f(x,y) = x^2 + y^2$. We want to find the rate of change of f along a curve defined by $x(t) = t$ and $y(t) = 2t$. The partial derivatives of f are:

$$\frac{\partial f}{\partial x} = 2x, \quad \frac{\partial f}{\partial y} = 2y$$

The derivatives of $x(t)$ and $y(t)$ are

$$\frac{dx}{dt} = 1, \quad \frac{dy}{dt} = 2$$

Then we get

$$\begin{aligned} \frac{d}{dt}f(x(t),y(t)) &= \frac{\partial f}{\partial x}\frac{dx}{dt} + \frac{\partial f}{\partial y}\frac{dy}{dt} \\ &= 2x\cdot 1 + 2y \cdot 2 \\ &= 2(t)\cdot 1 + 2(2t)\cdot 2 \\ &= 10t \end{aligned}$$

Corollary 3.1
Let $f : (a, b) \to \mathbb{R}$

1. (Necessity)If $\bar{x}$ is a **local minimizer** of f on (a, b), then $f'(\bar{x}) = 0$ and $f''(\bar{x}) \geq 0$.
2. (sufficient)If $f(\bar{x}) = 0$, $f''(\bar{x}) > 0$ then $\bar{x}$ is a **strict local minimizer** of f.

Proof.

Since $\bar{x}$ is the local minimizer of f, given small $|\Delta x|$,

$$f(\bar{x} + \Delta x) - f(\bar{x}) \geq 0$$

$$f(\bar{x} + \Delta x) = f(\bar{x}) + f'(\bar{x}) * \Delta x + o(|\Delta x|)$$

$$f'(\bar{x}) * \Delta x + o(|\Delta x|) \geq 0$$

$$\frac{f'(\bar{x})}{|\Delta x|} * \Delta x + \frac{o(|\Delta x|)}{|\Delta x|} \geq 0$$

Since this holds for all positive and negative Δx. Thus $f'(x) = 0$ For the second order proof, we used taylor thm to construct the contradiction ☐

Definition 3.2 – Hessian

The **Hessian** of f at $x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$ is the matrix

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

Lemma 3.1
Let $v \in \mathbb{R}^n$. Then

$$v = 0 \iff \langle v, d \rangle = 0, \quad \forall d \in \mathbb{R}^n$$

3.2 Lecture 3

Definition 3.3 – Matrix Norm

$$\|Q\| = max_{\|x\|=1} \|Qx\| = \text{Largest singular value of A}$$

Theorem 3.2 – Second Order Optimality Conditions (Min)
Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable at an open set D. Then

1. Necessary conditions: If $\bar{x}$ is a local minimizer for f on D, then

$$\nabla f(\bar{x}) = 0 \quad \text{and} \quad \nabla^2 f(\bar{x}) \succeq 0$$

2. Sufficient conditions: If $\nabla f(\bar{x}) = 0$, $\nabla^2 f(\bar{x}) \succ 0$ is positive definite then $\bar{x}$ is a strict local minimizer

of f on D.

Proof of Necessary conditions. Assume $\bar{x}$ is a local minimizer of f, we show that $\nabla f(\bar{x}) = 0$ and $\nabla^2 f(\bar{x}) \succeq 0$.
By multivariate Taylor Theorem, we have

$$f(\bar{x} + d) = f(\bar{x}) + \nabla f(\bar{x})^T d + o(\|d\|)$$

Rearranging, then dividing both sides by $\|d\|$ gives

$$\frac{f(\bar{x} + d) - f(\bar{x})}{\|d\|} = \nabla f(\bar{x})^T \frac{d}{\|d\|} + \frac{o(\|d\|)}{\|d\|}$$

Since $\bar{x}$ is a local minimizer, $f(\bar{x} + d) \geq f(\bar{x})$, thus $\frac{f(\bar{x}+d)-f(\bar{x})}{\|d\|} \geq 0$, so $\nabla f(\bar{x})^T \frac{d}{\|d\|} \geq 0$ as well. Now, $\frac{d}{\|d\|}$ is a unit vector with arbitrary d, so $\nabla f(\bar{x}) v \geq 0$ for all unit vectors v.
Consider $+v$ and $-v$, we have $\nabla f(\bar{x}) v \geq 0$ and $\nabla f(\bar{x})(-v) \geq 0$, so $\nabla f(\bar{x}) v = 0$ for all v. Thus, $\nabla f(\bar{x}) = 0$.

We now show that $\nabla^2 f(\bar{x}) \succeq 0$. By multivariate Taylor Theorem, we have

$$f(\bar{x} + d) = f(\bar{x}) + \nabla f(\bar{x})^T d + \frac{1}{2} d^T \nabla^2 f(\bar{x})^T d + o(\|d\|^2)$$

Rearranging and dividing both sides by $\|d\|^2$ gives

$$\frac{f(\bar{x} + d) - f(\bar{x}) - \nabla f(\bar{x})^T d}{\|d\|^2} = \frac{1}{2} \frac{d^T \nabla^2 f(\bar{x})^T d}{\|d\|^2} + \frac{o(\|d\|^2)}{\|d\|^2}$$

Since $\nabla f(\bar{x}) = 0$ from first part, we get

$$\frac{f(\bar{x} + d) - f(\bar{x})}{\|d\|^2} = \frac{1}{2} \frac{d^T \nabla^2 f(\bar{x})^T d}{\|d\|^2} + \frac{o(\|d\|^2)}{\|d\|^2}$$

Similar to first part, $\frac{d^T \nabla^2 f(\bar{x})^T d}{\|d\|^2} \geq 0$. Since $\|d\|^2 \geq 0$, we have $d^T \nabla^2 f(\bar{x})^T d \geq 0$. The Hessian is symmetric, so $\nabla^2 f(x) \succeq 0$. □

Theorem 3.3 – Second Order Optimality Conditions (Max)
Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable at an open set D. Then

1. Necessary conditions: If $\bar{x}$ is a local maximizer for f on D, then

$$\nabla f(\bar{x}) = 0 \quad \text{and} \quad \nabla^2 f(\bar{x}) \preceq 0$$

2. Sufficient conditions: If $\nabla f(\bar{x}) = 0$, $\nabla^2 f(\bar{x}) \prec 0$ is positive definite then $\bar{x}$ is a strict local maximizer of f on D.

Proof. 1. updated later after I confirmed some details with the professor □

Theorem 3.4
Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable and let $\bar{x}$ be a critical point of f. Then

- $\bar{x}$ is a global minimizer if $\nabla^2 f(\mathbf{x}) \succeq 0$ for all $\mathbf{x} \in \mathbb{R}^n$.
- $\bar{x}$ is a strict global minimizer if $\nabla^2 f(\mathbf{x}) \succ 0$ for all $\mathbf{x} \in \mathbb{R}^n$.
- $\bar{x}$ is a global maximizer if $\nabla^2 f(\mathbf{x}) \preceq 0$ for all $\mathbf{x} \in \mathbb{R}^n$.
- $\bar{x}$ is a strict global maximizer if $\nabla^2 f(\mathbf{x}) \prec 0$ for all $\mathbf{x} \in \mathbb{R}^n$.

Definition 3.4 – Critical/Stationary Points
A point $\bar{x} \in U$ is a critical point of a function $f : U \to \mathbb{R}$ if $\nabla f(\bar{x})$ exists and satisfies $\nabla f(\bar{x}) = 0$.

Problem 3.3 – Algorithm to Find Local Minimizer
Given $f : \mathbb{R} \to \mathbb{R}$ and $f'(\bar{x}) \neq 0$, then $x_{new} = \bar{x} - (\text{step}) * f'(\bar{x})$.
The idea is that if $f'(\bar{x}) > 0$, then we know that the function is increasing at $\bar{x}$, so we want to move to the left to obtain the minimum. Similarly, if $f'(\bar{x}) < 0$, then we know that the function is decreasing at $\bar{x}$, so we want to move to the right to obtain the minimum.

Problem 3.4
Given f:$\mathbb{R}^n \implies \mathbb{R}, \phi(\epsilon) = f(\bar{x} + \epsilon d)$
using tylar expansion $f(\bar{x} + \epsilon d)$= $f(\bar{x})$+ϵ $\nabla f(\bar{x})^T d$ +$o(\|\epsilon\|)$ shouldnt be ϵd? or d is the unit vector
let $d = -\nabla f(\bar{x})$, f($\bar{x}$)-$\epsilon\|f(\bar{x})\|^2$+$o(\epsilon)$. $< f(x)$ (if $\nabla f(\bar{x}) \neq 0$)
i.e test nec condition
If $\nabla f(\bar{x}) \neq 0$, then $x_{new} = \bar{x} + \epsilon(-\nabla f(\bar{(x)}))$ Move to the deapest direction

Definition 3.5 – Cauchy's method of steepest descant
https://www.math.usm.edu/lambers/mat419/lecture10.pdf $x_0 \in \mathbb{R}^n$.

$$Is \nabla f(x_k) \approxeq 0?\text{IF yes Stop}$$

.

$$O.W, \text{ find a Stop}\alpha > 0$$

$$x_{k+1} = x_k - \alpha \nabla f(x_k)$$

$$repeat$$

Problem 3.5 – Example of finding global and local minimizers
Find global and local minimizers of $f(x, y) = x^3 - 12xy + 8y^3$.

We first find the gradient and the Hessian:

$$\nabla f(x, y) = \begin{pmatrix} 3x^2 - 12y \\ -12x + 24y^2 \end{pmatrix}$$

$$\nabla^2 f(x, y) = \begin{bmatrix} -6x & -12 \\ -12 & 48y \end{bmatrix}$$

We can find the critical points when we solve for $\nabla f(x, y) = 0$. Solving it, we get solutions $(0, 0)$ or $(2, 1)$.
The Hessian at $(0, 0)$ is

$$\nabla^2 f(0, 0) = \begin{bmatrix} 0 & -12 \\ -12 & 0 \end{bmatrix}$$

The eigenvalues of $\nabla^2 f(0, 0)$ are $-12, 12$. Therefore it is indefinite. So $(0, 0)$ is a saddle point.
The Hessian at $(2, 1)$ is

$$\nabla^2 f(2, 1) = \begin{bmatrix} -12 & -12 \\ -12 & 48 \end{bmatrix}$$

Checking all leading principal minors, we see that they are all positive. So $\nabla^2 f(2,1)$ is positive definite. So $(2,1)$ is a local minimizer.

3.3 Lecture 4

Definition 3.6 – Principal Submatrices
Let

$$A = \begin{bmatrix} 1 & 1 & 2 & 7 \\ 1 & 1 & 4 & 6 \\ 2 & 4 & 7 & 8 \\ 7 & 6 & 8 & 1 \end{bmatrix}, \quad I = \{1,3\}, \quad A[I] = \begin{bmatrix} 1 & 2 \\ 2 & 7 \end{bmatrix}$$

Then $A[I]$ is a **principal submatrix** of A.

Definition 3.7 – Principal Minors
Let $A \in \mathbb{S}^n$, where $\mathbb{S}^n$ is the set of all symmetric $n \times n$ matrices.

1. $\det(A[I])$ is called the **principal minor** of A.
2. If $I = \{1, \ldots, k\}$ then $\det(A[I])$ is called the **leading principal minor** of A.

Proposition 3.1 – Characterizing Positive Definiteness with Principal Minors
Let $A \in \mathbb{S}^n$. Then

1. $A \succeq 0 \iff \det(A[I]) \geq 0$ for all principal minors $\det(A[I])$.
2. $A \succ 0 \iff \det(A[I]) > 0$ for all principal minors $\det(A[I])$.
3. $A \succ 0 \iff \det(A[I]) > 0$ for all **leading** principal minors $\det(A[I])$.

Definition 3.8 – Eigenvectors and Eigenvalues
$0 \neq v \in \mathbb{R}^n$ is an **eigenvector** of A if there exists $\lambda \in \mathbb{R}$ such that $Av = \lambda v$. The number λ is called an **eigenvalue** of A.

Theorem 3.5 – Finding Eigenvectors and Eigenvalues
Let A be a matrix.

1. Set up the characteristic equation. We find

$$\det(A - \lambda I) = 0$$

2. Solve for λ. These are the eigenvalues.
3. Plug eigenvalues $\lambda_1, \ldots, \lambda_n$ into $(A - \lambda I)v = 0$ and solve for v. These are the eigenvectors.

Theorem 3.6 – Orthogonal Spectral Decomposition
Let $A \in \mathbb{S}^n$. Then A has an **orthogonal spectral decomposition**

$$A = \sum_i \lambda_i u_i u_i^T = UDU^T$$

where U is orthogonal with the orthogonal eigenvectors u_i as columns and D is a diagonal matrix with real eigenvalues on the diagonal.

Corollary 3.2
Let $A \in \mathbb{S}^n$. Then

1. $A \succeq 0$ (positive semidefinite) iff all eigenvalues of A are nonnegative.
2. $A \succ 0$ (positive definite) iff all eigenvalues of A are positive.

Proposition 3.2
Let $A \in \mathbb{S}^n$. The following are equivalent (Positive definite):

1. $A \succ 0$.
2. All the eigenvalues of A are in $\mathbb{R}^n_{++}$, the interior of the nonnegative orthant.
3. A has a real symmetric positive definite square root, $A = SS, S \in \mathbb{S}^n_{++}$.
4. A has a lower triangular factorization, a Cholesky factorization, $A = LL^T$ and L has positive diagonal elements.
5. All principal minors of A are positive.
6. All leading principal minors of A are positive.

And the following are equivalent (Positive semidefinite):

1. $A \succeq 0$.
2. All the eigenvalues of A are in $\mathbb{R}^n_{+}$, the nonnegative orthant.
3. A has a real symmetric square root, $A = SS, S \in \mathbb{S}^n$.
4. A has a lower triangular factorization, a Cholesky factorization, $A = LL^T$.
5. All principal minors of A are nonnegative.

Theorem 3.7
$A \in S^n$ is positive definite if and only if it can be factored as $A = LL^T$, where L is a lower triangular matrix with positive diagonal entries.

Proof of $\Leftarrow$. Let $A = LL^T$ where L is a lower triangular matrix with positive diagonal entries. Then

$$\begin{aligned} x^T A x &= x^T L L^T x \\ &= y^T y && \text{where } y = L^T x \\ &= \|y\|^2 \geq 0 \end{aligned}$$

$\|y\|^2 = 0$ iff $y = 0$. Since L is lower triangular with positive diagonal entries, it is invertible, thus $y = Lx = 0$ iff $x = 0$. Thus $x^T A x > 0$ for all $x \neq 0$, so A is positive definite. □

Problem 3.6 – Motivation
When can we guarantee that global minimizers of $f : \mathbb{R}^n \to \mathbb{R}$ exist?
For example, the real valued function on $\mathbb{R}$ $f(x) = e^x$ is bounded below by 0 but has no minimizers. The

minimum value is 0 but is not attained.

Proposition 3.3 – Weierstrass Extreme Value Theorem
If $f : \mathbb{R}^n \to \mathbb{R}$ is continuous, and if $D \subset \mathbb{R}^n$ is a compact set (closed and bounded), then f is attains its global maximum and minimum on D.

Problem 3.7 – Example of a continuous function that does not attain its minimum
Let $f : \mathbb{R} \to \mathbb{R}$ be defined by $f(x) = e^x$. Then f is continuous, but $D = \mathbb{R}$ is not compact (closed but not bounded). We can see that f does not have any global minimizer.

Definition 3.9 – Coercive function
A continuous function $f : \mathbb{R}^n \to \mathbb{R}$ is **coercive** if for any sequence x_i with $\|x_i\| \to \infty$, it must be the case that $f(x_i) \to +\infty$. In other words,

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

Here are some examples:

1. $f_1(x) = x^2$ is coercive.
2. $g(x) = x$ is not coercive (because as $x \to -\infty$, $g(x) \to -\infty \neq \infty$).
3. $h(x) = e^x$ is not coercive.

Proposition 3.4 – Coercive Functions and Minimizers
A coercive function $f : \mathbb{R}^n \to \mathbb{R}$ has a global minimizer.

Definition 3.10 – Level Sets
Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function and let $\alpha \in \mathbb{R}$. An α-level set of f is defined by

$$L_\alpha = \{x \in \mathbb{R}^n : f(x) = \alpha\}$$

That is, all points x such that $f(x) = \alpha$.

- When $n = 2$, we call this a level curve.
- When $n = 3$, we call this a level surface.
- When $n > 3$, we call this a level hypersurface.

Definition 3.11 – Sub-level set
Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function and let $\alpha \in \mathbb{R}$. An α-sublevel set of f is defined by

$$S_\alpha(f) = \{x \in \mathbb{R}^n : f(x) \leq \alpha\}$$

That is, all points x below the line $f(x) = \alpha$.

Problem 3.8
Can we find minimizer for $q : \mathbb{R}^{\ltimes} \to \mathbb{R}$. $q(x) = \frac{1}{2}x^TQx + b^Tx + \alpha(Q = Q^T)$
Let $p^* = inf q(x)$

- When is it finite?
- When is it attained?

Proposition 3.5 – bdd blow Needtoconfirm
$q(x)$ is bdd below iff $Q \succeq 0$ and b$\in$ Range(Q) and 0=$\nabla q(x) = Qx - b$ so $x^* = Q^{-1}b$ is Orthogonal (why) so always attained

Proof. Since Q is positive semidef, we can use spectural decomposition, $Q = UDU^T, U^TU = I$. $x^TQx = x^TUDU^Tx$ which is also a quadratic form, let $y = U^Tx, x = Uy$
Thus for q(x), we sub $x = Uy$, then we get

$$\begin{aligned} q(x) &= \frac{1}{2}(Uy)^TQ(Uy) + b^TUy + \alpha \\ &= \frac{1}{2}y^TDy + (U^Tb)^Ty + \alpha \text{ Let } \bar{b} = U^Tb \\ &= \sum \lambda_i y_i^2 + \bar{b}_i y_i + \alpha \end{aligned}$$

Which is a seperable problem. $p_i^* = min\, \frac{1}{2}\lambda_i y_i^2 + \bar{b}_i y_i$
p_i^* is finite iff $\lambda_i \geq 0$, $\lambda_i = 0 \implies \bar{b_i} = 0 Why$. $\bar{b_i} = u_i^Tb$ which is the ith eigenvector. □

Proposition 3.6
If f is coercive, f maps R^n to R and cts, then there exists $\bar{x} \in argmin\, f(x)$

Proof. Omit □

Theorem 3.8 – q(x) is coercive iff Q is positive definite

Proof. Omit □

4 Linear Least Squares & Solving Linear Systems

4.1 Lecture 5

4.1.1 Linear Least Squares

Problem 4.1 – Motivation For Least Squares
Suppose we have a series of observed values from an experiment:

$$\{(t_1, s_1), (t_2, s_2), \ldots, (t_m, s_m)\}$$

where t_i is the time and s_i is the observed value at time t_i. We want to find a polynomial function

$$p(t) = x_0 + x_1 t + \cdots + x_n t^n$$

that fits the data. So we want to find coefficients $x_0, \ldots x_n$ such that $p(t_i) \approx s_i$ for all i. More formally, we want to minimize the absolute value of the error of each term. The error (ℓ_1 norm) is defined as

$$|e_i| = |p(t_i) - s_i|$$

This can be formulated into a ℓ_1 norm minimization problem:

$$\min \left\{ \sum_{i=1}^{m} |p(t_i) - s_i| : (x_0, \ldots, x_n) \in \mathbb{R}^{n+1} \right\}$$

This is a non-differentiable optimization problem since we have absolute values which make it not smooth. So we can be reformualte it as a linear program:

$$\min \quad \sum_{i=1}^{m} \lambda_i$$

s.t.

$$\begin{aligned} s_i - p(t_i) &\leq \lambda_i && \text{for all } i = 1, \ldots, m \\ p(t_i) - s_i &\leq \lambda_i && \text{for all } i = 1, \ldots, m \end{aligned}$$

This minimization problem is called **compressive sensing**.

Definition 4.1 – Vandermonde Matrix
Let

$$A = \begin{bmatrix} 1 & t_1 & t_1^2 & \cdots & t_1^n \\ 1 & t_2 & t_2^2 & \cdots & t_2^n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & t_m & t_m^2 & \cdots & t_m^n \end{bmatrix} \in \mathbb{R}^{m \times (n+1)}, \quad b = \begin{pmatrix} s_1 \\ s_2 \\ \vdots \\ s_m \end{pmatrix}, \quad x = \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_n \end{pmatrix}$$

A is called a **Vandermonde matrix**.

Theorem 4.1

The Vandermonde Matrix

$$A = \begin{bmatrix} 1 & t_1 & t_1^2 & \cdots & t_1^n \\ 1 & t_2 & t_2^2 & \cdots & t_2^n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & t_m & t_m^2 & \cdots & t_m^n \end{bmatrix} \in \mathbb{R}^{m \times (n+1)}, \quad b = \begin{pmatrix} s_1 \\ s_2 \\ \vdots \\ s_m \end{pmatrix}, \quad x = \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_n \end{pmatrix}$$

is full column rank if $n + 1 \leq m$ and the points t_i are distinct.

Definition 4.2 – ℓ_1 and ℓ_2 Norm
The ℓ_1 norm of a vector x is defined to be

$$\|x\| = \sum |x_i|$$

The ℓ_2 norm of a vector x is defined to be

$$\|x\| = \sqrt{\sum x_i^2}$$

Problem 4.2 – Linear Least Squares Problem
Recall our ℓ_1 norm minimization problem:

$$\min \left\{ \sum_{i=1}^{m} |p(t_i) - s_i| : (x_0, \ldots, x_n) \in \mathbb{R}^{n+1} \right\}$$

We can instead use ℓ_2 norm defined as $\|e\|_2 = \sqrt{\sum e_i^2}$. So our ℓ_2 minimization problem is

$$\min \left\{ \sum_{i=1}^{m} (p(t_i) - s_i)^2 : (x_0, \ldots, x_n) \in \mathbb{R}^{n+1} \right\}$$

where $p(t) = x_0 + x_1 t + \cdots + x_n t^n$. Using the Vandermonde matrix, we can rewrite our problem to be

$$\min \frac{1}{2} \|Ax - b\|^2$$

where

$$A = \begin{bmatrix} 1 & t_1 & t_1^2 & \cdots & t_1^n \\ 1 & t_2 & t_2^2 & \cdots & t_2^n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & t_m & t_m^2 & \cdots & t_m^n \end{bmatrix} \in \mathbb{R}^{m \times (n+1)}, \quad b = \begin{pmatrix} s_1 \\ s_2 \\ \vdots \\ s_m \end{pmatrix}, \quad x = \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_n \end{pmatrix}$$

The objective function is $g(x) = \frac{1}{2}\|Ax - b\|^2$. Let's first expand $g(x)$:

$$\begin{aligned} g(x) &= \frac{1}{2}\|Ax - b\|^2 \\ &= \frac{1}{2}(Ax - b)^T (Ax - b) \\ &= \frac{1}{2}(Ax)^T Ax - (Ax)^T b + \frac{1}{2}\|b\|^2 \\ &= \frac{1}{2} x^T A^T Ax - x^T A^T b + \frac{1}{2}\|b\|^2 \end{aligned}$$

Then, applying the gradient on both sides, we have

$$\nabla g(x) = A^T Ax - A^T b$$

To find the critical points, we solve for $\nabla g(x) = 0$. So the critical points are x^* that satisfy the equation

$$A^T Ax = A^T b$$

This is also called a **normal equation**.
also something about the condition number, i dont really understand.

Theorem 4.2 – Matrix Calculus

$$\nabla(Ax) = A^T$$
$$\nabla(x^T A) = A$$
$$\nabla(x^T x) = 2x$$
$$\nabla(x^T Ax) = (A + A^T)x$$

Definition 4.3 – Frechet Derivative
Let $h : U \to W$, where U is an open subset of V, and V, W are finite dimensional vector spaces. The function h is **Frechet differentiable** at $x \in U$ if there exists a linear transformation $A : V \to W$ such that

$$\lim_{d \to 0} \frac{\|h(x + d) - h(x) - Ad\|}{\|d\|} = 0$$

idk

4.1.2 Best Linear Least Squares

Suppose $\bar{x}$ is a solution to the normal equation

$$A^T Ax = A^T b$$

When A is not full column rank, the solution $\bar{x}$ is not unique. Therefore, the solution set to $A^T Ax = A^T b$ is

$$S = \{\bar{x} + y : y \in \text{null}(A)\}$$

Then, an algorithm might choose $x \in S$ with very large norm. This is not the best solution.

Definition 4.4 – Best Linear Least Square
The best linear least square is the least squares solution of minimal norm. It is the solution to the problem

$$\min_x \{\|x\| : x \in \text{argmin}\|Ax - b\|\}$$

4.2 Lecture 6

Goal: Solving normal equation/non linear case

Definition 4.5 – Singular Values of a Matrix
The singular values of a matrix A are the square roots of the eigenvalues of the matrix $A^T A$. They are always non-negative real numbers.
The number of non-zero singular values of a matrix equals the rank of that matrix.

Definition 4.6 – Condition Number of a Matrix
Suppose $A \in \mathbb{R}^{m \times n}, m > n$ is full column rank. The condition number of the matrix A, $\text{cond}(A)$, is the ratio of the largest to smallest nonzero singular values of A. Let $\sigma_{\max}$ be the largest singular value and $\sigma_{\min}$ be the smallest singular value. Then

$$\text{cond}(A) = \frac{\sigma_{\max}}{\sigma_{\min}}$$

Definition 4.7 – SVD Decomposition
Let A be an $m \times n$ matrix. Then A can be factored into

$$\underbrace{A}_{m\times n} = \underbrace{U}_{m\times m}\underbrace{\Sigma}_{m\times n}\underbrace{V^T}_{n\times n}$$

where

- U is an $m \times m$ orthogonal matrix consisting of eigenvectors of AA^T
- V^T is the transpose of an $n \times n$ matrix containing the eigenvectors of $A^T A$
- Σ is a diagonal matrix with $r = \text{rank}(A)$ positive eigenvalues of AA^T (Singular values of A) on the diagonal.

there is a section on piazza posted lecture notes that shows why using SVD decomposition to solve normal equation is a bad idea. Not sure if i should include here

Definition 4.8 – Orthogonal Matrix
A matrix Q is orthogonal if $Q^T Q = I$.

Definition 4.9 – Orthonormal Columns
A matrix Q has orthonormal columns if each column vector is a unit vector (norm is 1), and any two distinct columns are orthogonal (inner product is 0).

Problem 4.3

Applicable Model for ”best” data fitting

$$(s_i, p_i), \quad i = 1, ..., m$$

We got linear model and we need to solve $A^TAx = A^Tb$ normal equation

Today: Solve normal equation (whenever case)

i.e., this is from optimality condition for

$$\min \frac{1}{2}\|Ax - b\|_2^2 = f(x)$$

$$\text{, s.t.} A \in M_{m,n}$$

$$x \in \mathbb{R}^n$$

$$\text{stationary: } \nabla f(\bar{x}) = A^T(A\bar{x} - b) = 0$$

An alternate view is min $\|y - b\|$

, s.t. $y \in \text{Range}(A) \Rightarrow$ a projection problem

i.e., project b onto Range of A

$$\text{Aside: use SVD of } A = U\Sigma V^T : \Sigma = \begin{bmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_r \\ & 0 & \end{bmatrix}, \quad r = \text{rank}(A)\, \sigma_i > 0$$

SVD can be used for a ”rank revealing” decomposition.

$$U = \begin{bmatrix} U_1 & U_2 \end{bmatrix}$$

Where:

$$U_1 : \text{m x r orthogonal columns span R(A)}$$
$$U_2 : \text{m x (m-r) spanning } \mathcal{N}(A^T)$$

Properties:

$$U^TU = I \quad \text{and} \quad I = U_1^TU_1 = (U_1^TU_1)^T = \mathcal{C}(U_1U_1^T) = U_1U_1^T$$

This means $U_1U_1^T$ is symmetric idempotent, i.e., orthogonal projection.
To solve the projection minimizing $\|y - b\|$ where $y \in \text{Range}(\mathcal{R}(A))$:

$$\text{optimal } y^* = U_1U_1^Tb$$

Defining the Moore-Penrose generalized inverse:

$$A^+ = V\Sigma^+U^T$$

where

$$\Sigma^+ = \begin{bmatrix} \frac{1}{\sigma_1} & \dots & 0 \\ 0 & \dots & 0 \end{bmatrix}$$

and

$$\gamma_i = \frac{b_i}{\sigma_i} \quad \text{if } \sigma_i \neq 0$$

Finally, to find x such that $Ax = y^* = U_1U_1^Tb = AA^+b$: If A has full column rank (i.e., $\mathcal{R}(A^T) = \mathbb{R}^n$), then x is unique. Otherwise $\bar{x} + v, v \in N(A)$ is a solution for all v in N(A) Choosing a large solution. This solution of min is called ”Best least square solution”
we can write it as bilevel problem
The solution of minimum norm is called the *”best least square solution”*. This can be represented as a ”bilevel problem”:

$$\min_{\hat{x}} \|\hat{x}\|_2 \quad \text{for values when different}$$

subject to:

$$x \in \arg\min_z \frac{1}{2} \|Ax - b\|_2^2$$

The bLSS (best least square solution) is the unique LSS (least square solution) in $\mathcal{R}(A^T)$.

Definition 4.10 – Condition number
Given:

$$A = U\Sigma V^T \quad \text{(SVD of } A\text{)}$$

Where Σ is a diagonal matrix with:

$$\Sigma = \begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_r \end{pmatrix} \quad \text{and} \quad \sigma_r \geq 0$$

The condition number of matrix A is given by:

$$\text{Cond}(A) = \frac{\sigma_1}{\sigma_r}$$

If A is not of full rank, then:

$$\text{Cond}(A) = \infty \implies \text{Cond}(A^T) = \text{Cond}(A)$$

We also have:

$$A^T A = V\Sigma^T U^T U\Sigma V^T = V\Sigma^2 V^T$$

With:

$$\Sigma^2 = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_r^2 \end{pmatrix}$$

Thus:

$$\text{Cond}(A^T A) = \frac{\sigma_1^2}{\sigma_r^2} = Cond(A)^2$$

For the solution B:

$$Bx^*(True\ sol) = b, B\hat{x}\,(Numerical) = b + \delta b$$

Where:

$$\frac{\|x^* - \hat{x}\|}{\|x^*\|} \leq \text{Cond}(B)\frac{\delta\|b\|}{\|b\|}$$

To avoid solving the normal equation with sequares condition number, we should use QR

Definition 4.11 – QR Factorization
For any $m \times n$ matrix A, there there exists an $m \times m$ orthogonal matrix Q ($QQ^T = I$) and an $m \times n$ upper triangular matrix R ($R_{i,j} = 0, \forall i < j$) satisfying $A = QR$. Moreover, if the columns of A are linearly independent then we can get

$$\begin{aligned} A &= QR \\ &= Q \begin{bmatrix} R_1 \\ 0 \end{bmatrix} \\ &= \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \begin{bmatrix} R_1 \\ 0 \end{bmatrix} \\ &= Q_1 R_1 \end{aligned}$$

where

- R_1 is an invertible $n \times n$ upper triangular matrix
- 0 is an $(m-n) \times n$ zero matrix
- Q_1 is an $m \times n$ matrix with orthonormal columns
- Q_2 is an $m \times (m-n)$ matrix with orthonormal columns

Theorem 4.3 – QR Factorization on Normal Equation
Assuming that the columns of A are linearly independent, then the normal equation $A^T Ax = A^T b$ can be solved by applying QR factorization to A:

$$\begin{aligned}
(A^T A)x &= A^T b \\
((Q_1 R_1)^T Q_1 R_1)x &= (Q_1 R_1)^T b \\
(R_1^T Q_1^T Q_1 R_1)x &= R_1^T Q_1^T b \\
R_1^T R_1 x &= R_1^T Q_1^T b && \text{Since } Q_1 \text{ is orthogonal} \\
R_1 x &= Q_1^T b && \text{Since } R_1 \text{ is invertible}
\end{aligned}$$

Definition 4.12 – Methods of Solving General Linear Systems
Suppose we are given a linear system $Bx = b$, and we know that this system has a solution, i.e. $b \in \text{range}(B)$. There are 3 important algorithms/factorizations used to find x:

- Gaussian Elimination (LU factorization) ($PB = LU$)
- QR factorization
- SVD, singular value decomposition

Problem 4.4 – Solving Large Positive Definite Systems
Suppose we have a linear system, $Ax = b$, with A postive definite. If x^* is a solution, then $Ax^* - b = 0$. Then this is equivalent to minimizing the function

$$f(x) = \frac{1}{2}\|A_x - b\|^2, \nabla f(x) = Ax - b = 0$$

Dont understand this and the part after as well. You will have to add more notes here. Link to notes HERE

Definition 4.13 – Steepest Descent Direction in First order model
The steepest descent direction in the first order model is d=-$\frac{\nabla f(x_c)}{|x_c|}$

Theorem 4.4 – Conjugate Gradient Method
The first search direction is the negative gradient,

$$v_0 = -\nabla q(x_0)$$

with $q = f$. At the kth iteration:

$$v_{k+1} = -\nabla q(x_k) + \beta_k v_k$$

where β_k is chosen to ensure $\langle Av_{k+1}, v_k \rangle = 0$. This guarantees that the directions are A-conjugate wtf is

A conjugate $u, v\ conjugate, if\ u^T Av = 0$. We then set

$$x_{k+1} = x_k + \alpha_{k+1} v_{k+1}$$

where α_{k+1} is chosen from an exact line search (what is line search).

Problem 4.5

$$\begin{aligned} \min \quad & f(x) \\ & x \in D \subset \mathbb{R} \end{aligned}$$

- 1^{st} & 2^{nd} order optimality condition both necessary and sufficient
- 1^{st}, 2^{nd} order model

4.3 Lecture 7

Definition 4.14 – Nonlinear Least Square
Suppose we have $F : \mathbb{R}^n \to \mathbb{R}^m$ where

$$F(x) = \begin{pmatrix} f_1(x) \\ f_2(x) \\ \vdots \\ f_m(x) \end{pmatrix}$$

Then the nonlinear least squares problem is

$$\min\{h(x)\}$$

where

$$h(x) = \frac{1}{2}\|F(x)\|^2 = \frac{1}{2}\langle F(x), F(x)\rangle = \frac{1}{2}\sum_{i=1}^{m} f_i^2(x)$$

Definition 4.15 – Jacobian Matrix
Let F be defined as above, then the Jacobian matrix is $J(x) = F'(x)$ where

$$F'(x) = \begin{bmatrix} \nabla f_1(x)^T \\ \nabla f_2(x)^T \\ \vdots \\ \nabla f_m(x)^T \end{bmatrix}$$

Problem 4.6 – Solving Nonlinear Least Squares
For the nonlinear least squares problem defined above, we consider the special case $m = n$. Then we can consider the problem as solving the square system of nonlinear equations $F(x) = 0$. Recall that for current approximation x_c,

$$0 = F(x_c + d) \approx F(x_c) + \underbrace{F'(x_c)}_{\text{Jacobian}} \underbrace{d}_{\text{search direction}}$$

So we solve

$$F'(x_c)d = -F(x_c)$$

which is called the Newton equation. Then we can take a step in the search/Newton direction d to get a new approxiamtion $x_{c+1} = x_c + \alpha d$ for appropriate step length α.
In general, $\nabla h(x_c) = \sum_{i=1}^m f_i(x_c) \nabla f_i(x_c) = J(x_c)^T F(x_c)$

Definition 4.16 – Second Order Model for Nonlinear Problem
$h(x + \nabla x) \sim h()$ upload later

Definition 4.17 – Argmin and argmax
argmin $f(x)$ is the set of all minimizers of $f(x)$, similarly argmax $f(x)$ is the set of all maximizers of $f(x)$.

Theorem 4.5 – Lagrange Multipliers for Equality Constraints
Suppose that $f : C \to \mathbb{R}$ and $h : C \to \mathbb{R}^m$ are sufficiently smooth functions on the open set $C \subseteq \mathbb{R}^n$. Let $x \in C$ be a local minimum of f subject to the constraints $h(x) = 0, x \in C$. In addition, assume the following regularity condition at x (constraint qualification at x)

$$h'(x) \text{ is onto}$$

Then there exists $\lambda \in \mathbb{R}^m$ (a Lagrange multiplier vector) such that

$$0 = \nabla f(x) + \langle \lambda, h'(x) \rangle$$

Equivalently, the gradient $\nabla f(x)$ is in the range of $h'(x)^T$.

Problem 4.7 – DNE of multiplier
Recall:

$$\begin{aligned}
&\text{minimize } f(x) \\
&\text{subject to } f_i(x) = 0, \quad i = 1, \dots, m \\
&L(x, \lambda) = f(x) - \sum_{i=1}^m \lambda_i f_i(x) \quad \text{(Lagrangian multiplier)}
\end{aligned}$$

If x^* is a local minimum and

$$\{\nabla f_i(x^*)\} \text{ is a linearly independent set, then } \nabla L(x^*, \lambda) = \nabla f(x^*) + \sum_{i=1}^m \lambda_i \nabla h_i(x^*) = 0 \text{ for some } \lambda \in \mathbb{R}$$

Example:

$$\begin{aligned}
f(x) &= x \\
h(x) &= x^2 = 0 \\
L(x, \lambda) &= x + \lambda x^2, \quad \text{with constraint } x^2 = 0 \\
0 &= \nabla L(x, \lambda) = 1 + 2\lambda x = 1 + 0. \text{ No } \lambda \text{ exist}
\end{aligned}$$

5 Iterative Methods for Unconstrained Optimization

5.1 Lecture 8-10

Before, we have iterative methods to solve linear and linear least squares system.

Goal: Solve more general problems of minimization.
An iterative algorithm is a procedure that produces an (infinite) sequence of points $\{x_k\}$, in $\mathbb{R}^n$ such that our sequence converges to a critical point of f or to a point that satisfies the second order necessary conditions of optimality.

5.1.1 Line Search Strategy

Most Line Search Algorithms have the following procedure:
Choose a starting point x_0. For $k = 0, 1, 2, \ldots$

1. Choose a search direction d_k.
2. Choose a step length $\alpha_k > 0$ (that satisfies $f(x_k + \alpha_k d_k) < f(x_k)$)
3. Update $x_{k+1} = x_k + \alpha_k d_k$
4. Stop if a stopping criterion is satisfied.

Definition 5.1 – Line Search Strategy
At each iteration k, we choose a vector (search direction) $d_k \neq 0 \in \mathbb{R}^n$, then choose $\alpha_k > 0$ (step length) that approximately solves

$$\alpha_k \in \operatorname{argmin} f(x_k + \alpha d_k)$$

Then we update $x_{k+1} = x_k + \alpha_k v_k$.

5.1.1.1 Finding Descent Direction

Definition 5.2 – Descent Direction
Let $x \in U \subseteq \mathbb{R}^n$ with U being an open set. Then $d \in \mathbb{R}^n$ is a descent direction for f at x if there exists $\bar{\alpha} > 0$ such that

$$x + \alpha d \in U, f(x + \alpha d) < f(x)$$

for all $0 < \alpha < \bar{\alpha}$.

So $f(x_{k+1}) < f(x_k)$.

Lemma 5.1
Let f be sufficiently smooth on an open set U and let $\bar{x} \in U$. Let $d \in \mathbb{R}^n$ satisfy

$$\langle d, \nabla f(\bar{x}) \rangle = \nabla f(\bar{x})^T d < 0$$

Then d is a descent direction for f at $\bar{x}$.

(i.e., the directional derivative of f at $\bar{x}$ in the direction of d is negative, so we know that it is the descent direction.)

Proof. From the hypothesis we have

$$f(x + td) = f(x) + t\nabla f(x) \cdot d + o(t).$$

Therefore we deduce

$$\frac{f(x + td) - f(x)}{t} = \langle d, \nabla f(x) \rangle + \frac{o(t)}{t} < 0, \text{ for all sufficiently small } t > 0.$$

Moreover, since U is open, we have $x + td \in U$ for all sufficiently small $t > 0$. □

Problem 5.1 – Example(direction of steepest descent)
We saw

$$\min \nabla f(x)^t \cdot d \quad \text{s.t.} \quad ||d|| = 1$$

get

$$d = -\frac{\nabla f(x)^T}{||\nabla f(x)||}$$

Cauchy direction of steepest descent
Alternatively,

$$-\nabla f(x)^T \cdot d \leq ||d|| \cdot ||\nabla f(x)^T|| = ||d|| \cdot f(x)^T$$

but

$$d = -\nabla f(x)^T \text{ gives equality.}$$

So C.S yields s.t direction
If $B \succ 0$, **then**

$$d = -B^{-1}\nabla f(x)^T \text{ in a "deflated gradient" direction.}$$

Proposition 5.1 – Existence of B
Suppose$\nabla f(\bar{x})^T d < 0$ Then there exists$B \succ 0$that satisfies d=$-B^{-1}\nabla f(\bar{x})$

Proof. Ignore it for nowtext □

5.1.2 Second order model: Newton Type

Model $mq(x) = f(x_c + \nabla f(x_c)^T \Delta x + \frac{1}{2}\Delta x^T B_c \Delta x$ where $B_c \simeq \nabla^2 f(x_c)$

Lemma 5.2 – TFAE
1. $mq(\Delta x)$ is bdd below
2. $B_c \succeq 0$ and $\nabla f(x_c) \in range(B_c)$
3. $\Delta \bar{x} = -B_c^+ \nabla f(x_c)$ is a global minimum of mq

Proof. 2 to 1:
Suppose 2 holds, we have $B_c \succeq 0$ and $\exists \bar{\Delta x}, B_c \bar{\Delta x} = -\nabla f(x_c))$. This is also the stationary condition. Again, since B_c is psd, there exists a psd square root such that $ss = B_c$.
Then lets complete the square

$$mq(\Delta x) =$$

confirm

□

The main difference between line search methods is the choice of the descent direction

Problem 5.2 – Example
If $\nabla f(x) \neq 0$ then $d = -\nabla f(x)$ is a descent direction since

$$d^T \nabla f(x) = -\nabla f(x)^T \nabla f(x) = -\|\nabla f(x)\|_2^2 < 0$$

Another option is letting $d = -H^{-1}\nabla f(x)$ where $H \succ 0$ and $\nabla f(x) \neq 0$ since

$$d^T \nabla f(x) = -\nabla f(x)^T H^{-1} \nabla f(x) < 0$$

since $H \succ 0$, we have $H^{-1} \succ 0$, so $\nabla f(x)^T H^{-1} \nabla f(x) > 0$.

Theorem 5.1 – Existence of a Descent Direction
Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable and let $\bar{x} \in \mathbb{R}^n$. Assume that d is a descent direction of f at $\bar{x}$, so

$$\nabla f(\bar{x})^T d < 0$$

Then there exists $\delta > 0$ such that for every $\alpha \in (0, \delta]$,

$$f(\bar{x} + \alpha d) < f(\bar{x})$$

.

Proof. By Taylor's theorem, for all $x \in \mathbb{R}^n$,

$$f(x) = f(\bar{x}) + \nabla f(\bar{x})^T (x - \bar{x}) + o(\|x - \bar{x}\|)$$

Then

$$f(\bar{x} + \alpha d) = f(\bar{x}) + \nabla f(\bar{x})^T (\bar{x} + \alpha d - \bar{x}) + o(\|\bar{x} + \alpha d - \bar{x}\|)$$

Simplifying gives

$$f(\bar{x} + \alpha d) = f(\bar{x}) + \alpha \underbrace{\nabla f(\bar{x})^T d}_{<0} + o(\|\alpha d\|)$$

Rearranging and dividing both sides by $\alpha\|d\|$, we get

$$\frac{f(\bar{x} + \alpha d) - f(\bar{x})}{\alpha\|d\|} = \frac{\nabla f(\bar{x})^T d}{\|d\|} + \frac{o(\|\alpha d\|)}{\alpha\|d\|}$$

Taking the limit as $\alpha \to 0$, we get

$$\lim_{\alpha \to 0} \frac{f(\bar{x} + \alpha d) - f(\bar{x})}{\alpha\|d\|} = \frac{\nabla f(\bar{x})^T d}{\|d\|}$$

Since $\nabla f(\bar{x})^T d < 0$, we know that the limit is negative. Then there exists $\delta > 0$ such that for all $\alpha \in (0, \delta]$, we have

$$\frac{f(\bar{x} + \alpha d) - f(\bar{x})}{\alpha\|d\|} < 0$$

□

5.1.2.1 Finding Step Size

The process of finding the step size α_k is called line search. Some common choices for step size are as follows:

1. $\alpha_k = \alpha$ (a constant) for every k
2. Exact line search: $\alpha_k = \text{argmin}_{\alpha \geq 0} f(x_k + \alpha d_k)$
3. Inexact line search: A step size α_k that achieves a sufficient decrease in f. One popular inexact line search conditions are called the Wolfe conditions.

Problem 5.3 – Exact Line Search
For exact line search, we define a new function

$$\phi(\alpha) = f(x_k + \alpha d_k)$$

We can differentiate w.r.t. α using the chain rule to obtain

$$\phi'(\alpha) = \nabla f(x_k + \alpha d_k)^T d_k$$

Then we set $\phi'(\alpha) = 0$ to obtain the critical point of ϕ. Then we can use the second derivative test to determine if the critical point is a local minimizer. If it is, then we have found the optimal step size α_k.

Definition 5.3 – Wolfe Conditions
Let $0 < c_1 < c_2 < 1$. The Wolfe conditions are

1. Sufficient decrease condition:
$$f(x_k + \alpha d_k) \leq f(x_k) + c_1 \alpha \nabla f(x_k)^T d_k$$
or equivalently
$$\phi(\alpha_k) \leq \phi(0) + c_1 \alpha \phi'(0)$$
2. Curvature condition:
$$\nabla f(x_k + \alpha d_k)^T d_k \geq c_2 \nabla f(x_k)^T d_k$$
or equivalently
$$\phi'(\alpha_k) \geq c_2 \phi'(0)$$

Theorem 5.2 – Existence of Step Lengths that Satisfy Wolfe Conditions
Assume $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable. Let d_k be a descent direction at x_k and assume f is bounded below along the ray $\{x_k + \alpha d_k : \alpha \geq 0\}$. Then if $0 < c_1 < c_2 < 1$ there exist intervals of step lengths satisfying the Wolfe conditions.

Proof. Omitted □

5.1.2.2 Steepest Descent Method

The steepest descent method is easy to implement as it requires the calculation of the gradient but not second derivatives. The descent direction at each iteration is chosen as
$$d_k = -\nabla f(x_k)$$

Lemma 5.3
Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable and let $\bar{x} \in \mathbb{R}^n$ such that $\nabla f(\bar{x}) \neq 0$. Then the optimal solution of the following minimization problem
$$\min_{d \in \mathbb{R}^n} \{\nabla f(\bar{x})^T d : \|d\|_2 = 1\}$$
is
$$d^* = \frac{-\nabla f(\bar{x})}{\|\nabla f(\bar{x})\|_2}$$

Proof. By Cauchy-Schwarz inequality, we have
$$\nabla f(\bar{x})^T d \geq -\|\nabla f(\bar{x})\|_2 \|d\|_2 = -\|\nabla f(\bar{x})\|_2$$
The smallest value of d is $d = \frac{-\nabla f(\bar{x})}{\|\nabla f(\bar{x})\|_2}$ since if we plug in d into the inequality, we get
$$\nabla f(\bar{x})^T \frac{-\nabla f(\bar{x})}{\|\nabla f(\bar{x})\|_2} = \frac{-\|\nabla f(\bar{x})\|_2^2}{\|\nabla f(\bar{x})\|_2} = -\|\nabla f(\bar{x})\|_2$$
So $d^* = \frac{-\nabla f(\bar{x})}{\|\nabla f(\bar{x})\|_2}$ is a minimizer of this problem. □

Theorem 5.3
Let $\{x_k\}$ be the sequence generated by the method of steepest descent method where the step sizes are chosen by the exact line search. Then for every $k \geq 0, k \in \mathbb{N}$

$$(x_{k+2} - x_{k+1})^T (x_{k+1} - x_k) = 0$$

5.1.2.3 Backtracking Line Search

Definition 5.4 – Backtrack Inexact Line Search
We make use of the Wolfe Conditions.

1. Initialize $\alpha > 0$, for example, choose $\alpha = 1, c \in (0, 1)$
2. if $(\phi(\alpha) > \phi(0) + c\alpha\phi'(0))$
   (a) while $(\phi(\alpha) > \phi(0) + c\alpha\phi'(0))$
      i. $\alpha = \alpha/2$
3. else if $(\phi(2\alpha) \leq \phi(0) + 2c\alpha\phi'(0))))$
   (a) while $(\phi(2\alpha) \leq \phi(0) + 2c\alpha\phi'(0))$
      i. $\alpha = 2\alpha$
4. Output $= \alpha$.

5.1.2.4 Newton's Method

In Newton's Method, the search direction is given by

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

The Newton step is defined only when $\nabla^2 f(x_k)$ is positive definite.

Definition 5.5 – Newton's Method
Input: Initial point: x_0, Tolerance: $\epsilon > 0$

1. Solve the linear system of equations $\nabla^2 f(x_k) d_k = -\nabla f(x_k)$ for d_k, so $d_k = -[\nabla^2 f(x_k)]^{-1} \nabla f(x_k)$
2. Update $x_{k+1} = x_k + \alpha_k d_k$ where $\alpha_k = 1$ for Newton's method
3. If $\|\nabla f(x_{k+1})\|_2 \leq \epsilon$, stop. Otherwise, go to step 1.

Definition 5.6 – Operator Norm
The operator norm of a matrix $Q \in \mathbb{R}^{m \times n}$ is defined by

$$\|Q\|_2 = \max\{\|Qx\_2 : x \in \mathbb{R}^n, \|x\|_2 = 1\}$$

This definition satisfies the properties of the norm. For every $A \in \mathbb{R}^{m \times n}$,

- $\|A\|_2 = 0 \iff A = 0$
- $\|\alpha A\|_2 = |\alpha| \|A\|_2$ for every $\alpha \in \mathbb{R}$
- $\|A + B\|_2 \leq \|A\|_2 + \|B\|_2$

Theorem 5.4
For every $A \in \mathbb{R}^{m\times n}$, $x \in \mathbb{R}^n$, we have

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

5.1.2.5 Quasi-Newton Methods

Since computing the Hessian matrix is expensive, we can use an approximation of the Hessian matrix instead. We now set the search direction as

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

where the symmetric and positive definite matrix B_k is updated at every iteration by a quasi-Newton updating formula.

There are many quasi-Newton methods, but we will only discuss the BFGS method.
Consider the quadratic model of the objective function:

$$m_k(d) = f(x_k) + \nabla f(x_k)^T d + \frac{1}{2} d^T B_k d$$

where B_k is a positive definite matrix and will be updated at every iteration. The minimizer of this quadratic model is

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

It is used as the search direction and the new iterate is given by

$$x_{k+1} = x_k + \alpha_k d_k$$

where α_k is the step length, chosen so that it satisfies the Wolfe conditions. In this method, we match the gradient of m_{k+1} to the gradient of f at x_k and x_{k+1}. That is, we require

$$\nabla m_{k+1}(0) = \nabla f(x_{k+1})$$

$$\nabla m_{k+1}(-\alpha_k d_k) = \nabla f(x_k)$$

From the second equation above, we get

$$\nabla_{k+1}(-\alpha_k d_k) = \nabla f(x_{k+1}) - \alpha_k B_{k+1} d_k = \nabla f(x_k)$$

Which implies

$$\nabla (x_{k+1}) - \nabla f(x_k) = B_{k+1}(x_{k+1} - x_k)$$

This is the secant equation. We can use this equation to update B_{k+1}.

5.2 Lecture 11

5.2.1 Convergence of Line Search Methods

Definition 5.7 – Key Quantity
The key quantity is the angle between the search direction and the gradient. It is

$$\cos\theta_k = \frac{\langle \nabla f(x_k), v_k \rangle}{\|\nabla f(x_k)\| \|v_k\|}$$

We will use this together with Wolfe conditions to prove convergence.

Theorem 5.5 – Convergence of Line Search Methods
Let $x_0 \in \mathbb{R}^n \to \mathbb{R}$ be C^1-smooth, and f be bounded below. Let $0 < c_1 < c_2 < 1$ and suppose that there exists a Lipschitz constant $L > 0$ such that

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$$

on an open set containing the level set $\{z : f(z) \leq f(x_0)\}$.
Let $\{x_k\}$ be generated using descent directions v_k and step lengths that satisfy the Wolfe conditions. Then we have

$$\sum_{k=1}^{\infty} \cos^2 \theta_k \|\nabla f(x_k)\|^2 < \infty$$

Proof. In course notes. □

Corollary 5.1
If there exists $\delta > 0$ satisfying $\cos^2 \theta_k \geq \delta > 0$ for all k, then

$$\lim_{k \to \infty} \nabla f(x_k) = 0$$

5.2.2 Convergence Rate of Steepest Descent

For exact line search, we have

$$x_{k+1} = x_k - \left(\frac{\|\nabla f(x_k)\|^2}{\langle A\nabla f(x_k), \nabla f(x_k) \rangle} \right) \nabla f(x_k)$$

If we plot out our iterations, we can see that there is a zig-zagging behaviour and the directions $x_{k+1} - x_k$ and $x_k - x_{k-1}$ are orthogonal.

Theorem 5.6 – Zig Zagging of Steepest Descent
For the general steepest descent method with an exact step-size, we have

$$(x_{k+1} - x_k)^T (x_k - x_{k-1}) = 0$$

So they are orthogonal.

Theorem 5.7 – Rate of Convergence of Steepest Descent
Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is C^2-smooth and the iterates $\{x_k\}$ generated by steepest descent with exact line search converge to x^*, that is, $\lim_{k \to \infty} x_k = x^*$. Suppose that $\nabla^2 f(x^*)$ is positive definite. Then for all large k, we have

$$f(x_{k+1}) - f(x^*) \leq K(f(x_k) - f(x^*))$$

with

$$0 \leq K = \left(\frac{\lambda_1 - \lambda_n}{\lambda_1 + \lambda_n} \right)^2 < 1$$

where λ_1 and λ_n are the largest and smallest eigenvalues of $\nabla^2 f(x^*)$.

5.2.3 Convergence Rate of Newton's Method

Definition 5.8 – Comparing Algorithms

upload later

Definition 5.9 – p-th Order Convergence
If the sequence of residuals/errors $r_k \to 0$ and

$$\lim_{k\to\infty} \frac{\|r_{k+1}\|}{\|r_k\|^p} \to r > 0$$

then the order of convergence is called p-th order.
If $p = 1, r < 1$, then this is called linear convergence. (For s.d, linear , r $\cong r$ (cond(Hessian at x^*)))
If $p = 2$, then this is called quadratic convergence.
If

$$\lim_{k\to\infty} \frac{\|r_{k+1}\|}{\|r_k\|} \to 0$$

then this is called superlinear convergence.

Theorem 5.8 – Convergence of Newton's Method
Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is C^2-smooth, $\nabla^2 f(x^*)$ is positive definite, $\nabla f(x^*) = 0$ (also minor condition: Hessian is locally Lipschitz continuous, which in particular holds if f is C^3-smooth). Consider the iterates

$$\begin{aligned} x_{k+1} &= x_k + t_k v_k \\ &= x_k + t_k(\nabla^2 f(x_k))^{-1}(-\nabla f(x_k)) \end{aligned}$$

where t_k is the step size chosen to satisfy the Wolfe conditions (with $c_1 < \frac{1}{2}$). Then if the starting point x_0 is sufficiently close to x^*, then we have:

1. $t_k = 1$ satisfies the Wolfe conditions (pure/free Newton step).
2. x_k converges to x^*.
3. If we choose $t_k = 1$, then we have quadratic convergence locally of the iterates, that is for some $r \geq 0$ and some x_0 close enough to the strict local minimum x^*, we have

$$\|x_{k+1} - x^*\| \leq r\|x_k - x^*\|^2$$

$$\|\nabla f(x_{k+1})\| \leq r\|\nabla f(x_k)\|^2$$

Problem 5.4
$d_{SC} = -\nabla f(x_k), d_N = -\nabla^2 f(x_k)^{-1}\nabla f(x_k)$

levenberg marquardt algorithm $-(\nabla^2 f(x_k) + \lambda I)^{-1}\nabla f(x_k)$ $\lambda > 0$

Problem 5.5
This leads to trust region method , follows by 5.8.1 Trust Region Methods Outline. note: Find d_k is trust region subproblem

5.2.4 Trust Region Strategy

Problem 5.6 – Trust Rigion Strategy
Construct a model for f at x_k, $f(x_k + \Delta x) \sim mq(\Delta x) = f(x_k) + \Delta x^T \nabla f(x_k) + \frac{1}{2}\Delta x^T B_k \Delta x$, where

$B_k \simeq \nabla^2 f(x_k)$. COnstruct region where we trust the model. Region can be a ball, eg. $\Omega_k = \{\Delta x | ||\Delta x|| \leq \delta_k^2\}$

The difference in trust region strategy and line search strategy is that in line search strategy, we first choose the direction, then choose step size. In trust region, we first choose the step size, then we choose the direction.

Definition 5.10 – Trust Region Strategy
In each iteration, we construct a model of f. That is, in each step we consider $m_k : \mathbb{R}^n \to \mathbb{R}$ that is a simple function that approximates f well on some simple set Ω_k (the trust region) around our current approximation x_k. Then we find the new approximation

$$\hat{x} = \text{argmin}\ m_k(x) \quad x \in \Omega_k$$

A common model is the quadratic model

$$m_k(x) = f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2} \langle x - x_k, B_k(x - x_k) \rangle$$

where $B_k \approx \nabla^2 f(x_k)$ approximates the Hessian. If the values $f(\hat{x})$ and $m_k(\hat{x})$ are close, then we declare $x_{k+1} = \hat{x}$. Otherwise, we shrink the size of the trust region Ω_k and repeat the process.
Usually Ω_k is a ball, ellipsoid, or a box around x_k.
The main points are how to choose a model function m_k, and how to choose a trust region Ω_k.

Theorem 5.9 – Trust Region Methods Outline
Outline:

1. Given $f : \mathbb{R}^n \to \mathbb{R}$, an iterate x_k, a local model function

$$m_k(d) = f(x_k) + \langle \nabla f(x_k), d \rangle + \frac{1}{2} \langle B_k d, d \rangle$$

   and a trust region radius $\nabla_k > 0$

2. Find $d_k \in \text{argmin}\{m_k(d) : \|d\| \leq \nabla_k\}$

3. Use the relative decrease (actual/predicted decrease)

$$\rho_k = \frac{f(x_k) - f(x_k + d_k)}{m_k(0) - m_k(d_k)}$$

   to gauge the strength of the model in order to obtain the new trust region radius Δ_{k+1} and to decide whether to accept the new point as $x_{k+1} = x_k + d_k$ or use $x_{k+1} = x_k$.

6 Convex set and function

6.1 Lecture 11-13

Definition 6.1 – Line Segment
Let $x, y \in \mathbb{R}^n$. The line segment joining x and y is the set

$$[x, y] = \{\lambda x + (1 - \lambda) y : \lambda \in [0, 1]\}$$

Definition 6.2 – Convex set

Set S in R^n in a convex set if for every $x, y \in S$ and for every $0 \leq \lambda \leq 1$

$$\lambda x + (1 - \lambda) y \in S$$

That is, the line segment $[x, y]$ is contained in S.

Problem 6.1 – Example of Convex Sets
The following are convex sets:

1. The empty set $\emptyset$, any single point $\{x_0\}$, and $\mathbb{R}^n$
2. Lines in $\mathbb{R}^n$. For example, a line through the point x_0 in the direction d

$$L = \{x_0 + \alpha d : \alpha \in \mathbb{R}\}$$

3. A halfspace in $\mathbb{R}^n$: $H = \{x : a^T x \leq c\}$
4. An open halfspace in $\mathbb{R}^n$: $H^o = \{x : a^T x < c\}$
5. A polyhedron in $\mathbb{R}^n$: $P = \{x : Ax \leq b\}$
6. An open and closed ball in $\mathbb{R}^n$
7. The set of $n \times n$ positive semi definite and positive definite matrices:

$$S_+^n = \{A \in \mathbb{R}^{n \times n} : A \succeq 0\}, \quad S_{++}^n = \{A \in \mathbb{R}^{n \times n} : A \succ 0\}$$

Definition 6.3 – Convex combination
Let S be a convex set and let $x_1, \dots x_n \in S$. If $\lambda_1, \lambda_2, \dots, \lambda_n$ are nonnegative numbers such that

$$\sum_{i=1}^{n} \lambda_i = 1$$

then

$$\sum_{i=1}^{n} \lambda_i x_i$$

is called a convex combination and is contained in S.

Lemma 6.1
S is a convex set iff S contains all its convex combination

Definition 6.4
Given $S \in R^n$, the convex hull of S, conv S is the smallest convex set containing S and is equal to the set of all convex combinations of points in S.

Lemma 6.2
Suppose S is a compact set, then conv S is a compact set.

Definition 6.5 – Affine Set
Let $S \subseteq \mathbb{R}^n$ be a subspace, and $a \in \mathbb{R}^n$. Then $L = a + S$ is an affine set (i.e. it is a translation of a subspace).
Moreover, $\dim L = \dim S$. In addition, let $T \subseteq \mathbb{R}^n$. Then the affine hull, aff T, is the smallest affine set containing T. If T is a convex set, then $\dim(T) = \dim(\text{aff}\, T)$.

Definition 6.6 – Convex Cone
$K \subseteq \mathbb{R}^n$ is a convex cone if the Minkowski sums satisfy

1. $\lambda K \subseteq K$ for all $\lambda \geq 0$
2. $K + K \subseteq K$

Definition 6.7 – Hyperplane
For $a \in \mathbb{R}^n, b \in \mathbb{R}$, their hyperplane is defined as

$$H = \{x \in \mathbb{R}^n : a^T x = b\}$$

Definition 6.8 – Halfspace
For $a \in \mathbb{R}^n, b \in \mathbb{R}$, their halfspace is defined as

$$H = \{x \in \mathbb{R}^n : a^T x \leq b\}$$

Definition 6.9 – Polytope
A polytope(polyhedron set) is a finite intersection of halfspaces.

Definition 6.10 – Face
Let C be a convex set. The convex set $F \subseteq C$ is a face of C, denoted as $F \trianglelefteq C$ if

$$x, y \in C, \quad z \in (x, y) \cap F \implies x, y \in F$$

That is, for any two points $x, y \in C$, if a point z lies in the open line segment between x and y (denoted (x, y)) also lies in F, then $x, y \in F$.
If this is true, then F is a face of C.

Theorem 6.1
If C is a convex cone, then the faces of C are convex cones.

Proof. Exercise. □

Definition 6.11 – Non-negative Polar (Dual) Cone
The non-negative polar (dual) cone of a set C is

$$C^+ = C^* = \{\phi : \langle \phi, c \rangle \geq 0, \forall c \in C\}$$

Definition 6.12 – Conjugate Face
Let F be a face of the convex cone K denoted $F \trianglelefteq K$. The conjugate face of F is

$$F^* = F^\perp \cap K^*$$

Definition 6.13 – Exposed Face
Let F be a face of the convex cone K, $F \trianglelefteq K$. Then F is an exposed face if there exists $\phi \in K^*$ such that

$$F = K \cap \phi^\perp$$

where $\phi^\perp = \{\phi\}^\perp$.

Definition 6.14 – Extreme Point
Let $C \subseteq \mathbb{R}^n$. The point $x \in C$ is an extreme point of C if

$$x \in [y, z], \quad y, z \in C \implies x = y = z$$

That is, there is no nontrivial convex combination with points in C.

Theorem 6.2
$\{z\}$ is an extreme point of C iff $\{z\} = F$ is a face of C of dimension 0.

6.2 Lecture 14

6.2.1 Review of Linear Programming

Review of basic solutions, basic feasible solutions, simplex etc. Basic feasible solutions are extreme points.

Definition 6.15 – Column Sub-matrix
Suppose we have a matrix

$$\begin{array}{c} \begin{matrix} 1 & 2 & 3 & 4 & 5 \end{matrix} \\ \begin{pmatrix} 1 & 2 & -1 & 1 & -1 \\ 0 & 1 & 0 & 1 & -1 \\ 0 & 0 & 1 & 1 & -1 \end{pmatrix} \end{array}$$

with columns labelled $1, 2, 3, 4, 5$. Let B be a subset of column indices, so $B \subseteq \{1, 2, 3, 4, 5\}$. Then A_B is a column sub-matrix of A indexed by set B.
For example, if $B = \{1, 2, 3\}$, then A_B is

$$\begin{pmatrix} 1 & 2 & -1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Using this notation, A_j denotes column j of A.

Definition 6.16 – Basis
Let B be a subset of column indices. B is a basis if

1. A_B is a square matrix
2. A_B is non-singular (columns are independent)

Let A be a matrix with independent rows, then B is a basis if and only if B is a maximal set of independent columns of A.

Theorem 6.3
The maximum number of independent columns is equal to the maximum number of independent rows.

Definition 6.17 – Basic and Non-basic Variables
Let B be a basis for A

- if $j \in B$ then x_j is a basic variable
- if $j \notin B$ then x_j is a non-basic variable

For example, for basis $B = \{1, 2, 4\}$, then x_1, x_2, x_4 are basic variables and x_3, x_5 are non-basic variables.

Definition 6.18 – Basic Solution
x is a basic solution for basis B if

1. $Ax = b$ and
2. $x_j = 0$ whenever $j \notin B$.

For example, suppose we have

$$A = \begin{array}{c} \begin{matrix} 1 & 2 & 3 & 4 & 5 \end{matrix} \\ \begin{pmatrix} 1 & 2 & -1 & 1 & -1 \\ 0 & 1 & 0 & 1 & -1 \\ 0 & 0 & 1 & 1 & -1 \end{pmatrix} \end{array}$$

$$B = \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix}$$

Then $x = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \end{pmatrix}$ is a basic solution for basis $B = \{1, 2, 3\}$ since $Ax = b$ and $x_4 = x_5 = 0$.

Problem 6.2
Suppose we have the system of linear equations

$$\begin{pmatrix} 1 & 0 & 1 & -1 \\ 0 & 1 & 1 & 1 \end{pmatrix} x = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$$

What is the basic solution x for basis $B = \{1, 4\}$? We have

$$\begin{aligned}
\begin{pmatrix} 2 \\ 2 \end{pmatrix} &= \begin{pmatrix} 1 & 0 & 1 & -1 \\ 0 & 1 & 1 & 1 \end{pmatrix} x \\
&= x_1 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + x_2 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + x_3 \begin{pmatrix} 1 \\ 1 \end{pmatrix} + x_4 \begin{pmatrix} -1 \\ 1 \end{pmatrix} \\
&= x_1 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + x_4 \begin{pmatrix} -1 \\ 1 \end{pmatrix} && \text{Since } x \text{ is a basic solution, } x_2 = x_3 = 0 \\
&= \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_4 \end{pmatrix}
\end{aligned}$$

Since A_B is a square matrix and non-singular, it must have an inverse, therefore

$$\begin{pmatrix} x_1 \\ x_4 \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 2 \\ 2 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 2 \\ 2 \end{pmatrix} = \begin{pmatrix} 4 \\ 2 \end{pmatrix}$$

Therefore the basic solution is $x = \begin{pmatrix} 4 \\ 0 \\ 0 \\ 2 \end{pmatrix}$.

Notice that when we are given a basis, then there is only 1 solution. This is a theorem.

Theorem 6.4
Consider $Ax = b$ and a basis B of A. Then there exists a unique basic solution x for B.

Proof. We have

$$\begin{aligned}
b &= Ax \\
&= \sum_j A_j x_j \\
&= \sum_{j \in B} A_j x_j + \sum_{j \notin B} A_j x_j \\
&= \sum_{j \in B} A_j x_j && \text{Since } x_j = 0 \text{ for all } j \notin B \\
&= A_B x_B
\end{aligned}$$

Now, since B is a basis, it implies A_B is invertible, so A_B^{-1} exists. Hence, $x_B = A_B^{-1} b$. □

Definition 6.19 – Basic Solution
Consider $Ax = b$ with independent rows. Vector x is a basic solution if it is a basic solution for some basis B.

Problem 6.3
For the system of linear equations

$$\begin{pmatrix} 3 & 2 & 1 & 4 & 1 \\ -1 & 1 & 0 & 2 & 1 \end{pmatrix} x = \begin{pmatrix} 6 \\ 3 \end{pmatrix}$$

is $x = \begin{pmatrix} 0 \\ 0 \\ 3 \\ 0 \\ 3 \end{pmatrix}$ basic?

We just have to give a basis B and show that x is basic for the basis B. We let $B = \{3, 5\}$ and we have $Ax = b$, and $x_1 = x_2 = x_4 = 0$. Therefore, x is basic for basis B, and so x is a basic solution.

Now is $x = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \\ 0 \end{pmatrix}$ basic?

No, we will prove it. Suppose x is basic for basis B. By definition, since $x_2 \neq 0$ and $x_4 \neq 0$, we have $2, 4 \in B$. Thus

$$A_{\{2,4\}} = \begin{pmatrix} 2 & 4 \\ 1 & 2 \end{pmatrix}$$

is a column submatrix of A_B. But the columns of $A_{\{2,4\}}$ are dependent, so A_B is singular and B is not a basis, a contradiction.

Definition 6.20 – Feasible Basic Solution
A basic solution x of $Ax = b$ is feasible if $x \geq 0$, i.e., if it is feasible for the problem P.

Definition 6.21 – Simplex Algorithm
Suppose we have a basic solution to the LP problem in canonical form with basis B. We can find a better feasible solution by:

1. Pick $k \notin B$ such that $c_k > 0$.
2. Set $x_k = t \geq 0$ as large as possible
3. Keep all other non-basic variables at 0
4. Choose basic variables such that $Ax = b$ holds

Here is an example:
Suppose we have the following LP problem:
max

$$\begin{pmatrix} 0 & 1 & 3 & 0 \end{pmatrix} x$$

s.t.

$$\begin{pmatrix} 1 & 1 & 2 & 0 \\ 0 & 1 & 1 & 1 \end{pmatrix} x = \begin{pmatrix} 2 \\ 5 \end{pmatrix}$$

$$x_1, x_2, x_3, x_4 \geq 0$$

Our LP is in canonical form for basis $B = \{1, 4\}$ and $(2, 0, 0, 5)^T$ is a basic solution. We will pick $k \notin B$ such that $c_k > 0$. So we pick $k = 2$, and set $x_2 = t \geq 0$, while keeping other non-basic variables to 0. So $x_3 = 0$.

Then we choose basic variables such that $Ax = b$ holds. So we have

$$\begin{aligned}\begin{pmatrix}2\\5\end{pmatrix} &= \begin{pmatrix}1 & 1 & 2 & 0\\0 & 1 & 1 & 1\end{pmatrix} x\\ &= x_1\begin{pmatrix}1\\0\end{pmatrix} + x_2\begin{pmatrix}1\\1\end{pmatrix} + x_3\begin{pmatrix}2\\1\end{pmatrix} + x_4\begin{pmatrix}0\\1\end{pmatrix}\\ &= \begin{pmatrix}x_1\\0\end{pmatrix} + t\begin{pmatrix}1\\1\end{pmatrix} + 0\begin{pmatrix}2\\1\end{pmatrix} + \begin{pmatrix}0\\x_4\end{pmatrix}\\ &= t\begin{pmatrix}1\\1\end{pmatrix} + \begin{pmatrix}x_1\\x_4\end{pmatrix}\end{aligned}$$

Rearranging gives

$$\begin{pmatrix}x_1\\x_4\end{pmatrix} = \begin{pmatrix}2\\5\end{pmatrix} - t\begin{pmatrix}1\\1\end{pmatrix}$$

This can be written in general as

$$x_B = b - tA_k$$

Since we want $x_1, x_4 \geq 0$, we have

$$\begin{aligned}x_1 = 2 - t \geq 0 \Rightarrow t \leq 2\\ x_4 = 5 - t \geq 0 \Rightarrow t \leq 5\end{aligned}$$

So $t \leq 2$ and we choose $t = 2$. Therefore, our new feasible solution is

$$x = \begin{pmatrix}2-t\\t\\0\\5-t\end{pmatrix} = \begin{pmatrix}0\\2\\0\\3\end{pmatrix}$$

Notice that our new solution is a basic solution for basis $B = \{2, 4\}$, we can say that 1 left the basis and 2 entered the basis. Rewriting our LP to canonical form with basis $B = \{2, 4\}$, we have
max

$$\begin{pmatrix}-1 & 0 & 1 & 0\end{pmatrix} x$$

s.t.

$$\begin{pmatrix}1 & 1 & 2 & 0\\-1 & 0 & -1 & 1\end{pmatrix} x = \begin{pmatrix}2\\3\end{pmatrix}$$
$$x_1, x_2, x_3, x_4 \geq 0$$

Again, we choose $k \notin B$ such that $c_k > 0$ and set $x_k = t$. So we choose $k = 3$, so $x_3 = t \geq 0$. We then pick

$$\begin{gathered}x_B = b - tA_k\\ \begin{pmatrix}x_2\\x_4\end{pmatrix} = \begin{pmatrix}2\\3\end{pmatrix} - t\begin{pmatrix}2\\-1\end{pmatrix}\end{gathered}$$

From here, we pick $t = 1$, so $x_2 = 0$, thus 2 leaves the basis. The new basis is now $B = \{3, 4\}$ with canonical form
max

$$\begin{pmatrix}-1.5 & -0.5 & 0 & 0\end{pmatrix} x$$

s.t.

$$\begin{pmatrix} 0.5 & 0.5 & 1 & 0 \\ -0.5 & 0.5 & 0 & 1 \end{pmatrix} x = \begin{pmatrix} 1 \\ 4 \end{pmatrix}$$
$$x_1, x_2, x_3, x_4 \geq 0$$

The basic solution is $x = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 4 \end{pmatrix}$. Since there is no $k \notin B$ such that $c_k > 0$, we have reached the optimal solution.
However, the Simplex algorithm is not guaranteed to terminate.

Definition 6.22 – Feasible Region
For an optimization problem, the feasible region is the set of all feasible solutions.

6.2.2 Operations that Preserve Convexity

- If Q_i are convex, then $\bigcap_i Q_i$ is convex (can be infinite).
- For an affine map $f : C \to D$, $f(x) = Ax + b$, where C, D are convex. The images $f(C)$, and preimages $f^{-1}(D)$ are convex sets.
- Projections of convex sets $P(C)$ are convex.
- Scaling of convex sets are convex, αC
- Rotation of convex sets are convex, $QC, Q^T Q = I$.
- Translation of convex sets are convex, $C + x$.

6.2.3 Hyperplane Separation and Support Theorems

Theorem 6.5 – Hyperplane Separation Theorem
Let $C, D \subset \mathbb{R}^n$ be convex sets and $C \cap D = \emptyset$. Then there exists $a \neq 0, \alpha \in \mathbb{R}$ such that

$$a^T x \leq \alpha \leq a^T y$$

for all $x \in C, y \in D$.

In other words, this theorem guarantees that if you have two sets that do not intersect and are convex, there is a hyperplane (which is a generalization of a flat surface to n-dimensions; for instance, a line in 2D or a plane in 3D) that separates the two sets. This hyperplane can be thought of as a decision boundary, which one can use to distinguish points belonging to one set from those belonging to the other.

Proof. See studocu course notes □

Definition 6.23 – Minkowski Sum and Difference
Let $C, D \subseteq \mathbb{R}^n$ be two sets. The Minkowski sum of C and D is defined as

$$C + D = \{x + y : x \in C, y \in D\}$$

The Minkowski difference of C and D is defined as

$$C - D = \{x - y : x \in C, y \in D\}$$

Lemma 6.3
Let $C \subset \mathbb{R}^n$ be a closed convex set and let $d \notin C$. Let x^* be the nearest point in C to d. Then the hyperplane

$$H = \{x : \langle v, x\rangle = \langle v, d\rangle - \|v\|^2\}$$

strictly separates C from d. That is, we have

$$\langle v, x\rangle \leq \langle v, d\rangle - \|v\|^2 < \langle v, d\rangle$$

for all $x \in C$.

Definition 6.24 – Supporting Hyperplane
$H = \{x : v^T x = b\}$ is a supporting hyperplane to a convex set Q at $\bar{x} \in Q$, if

$$x^T v \leq b, \quad \forall x \in Q$$

and the equation $x^T v = b$ holds.

Theorem 6.6 – Existence of Supporting Hyperplane
Suppose that $S \subseteq \mathbb{R}^n$ is a convex set, and $x_0 \in \text{bd}(S)$, the boundary of S. Then there exists a supporting hyperplane containing x_0.

Theorem 6.7 – Krein-Milman Theorem
Let K be a compact convex set. Then

$$K = \text{conv}(\text{ext}(K))$$

where $\text{ext}(K)$ is the set of extreme points of K.

Theorem 6.8 – Caratheodory Theorem
Let $S \subseteq \mathbb{R}^n$ and let $x \in \text{conv}(S)$. Then there exists $x_i \in S, i = 1, \ldots, n+1$ such that $x \in \text{conv}\{x_1, \ldots, x_{n+1}\}$. In other words, we need at most $n + 1$ points for the convex combination.

6.3 Lecture 15

6.3.1 Convex Functions

Definition 6.25 – Epigraph
Let $f : \mathbb{R}^n \to \mathbb{R}$. Then the epigraph of f is defined as

$$\text{epi}(f) = \{(x, r) : f(x) \leq r\}$$

f is convex if and only if $\text{epi}(f)$ is a convex set.

Definition 6.26 – Convex Functions and Strictly Convex Functions/Zero Order Characterization Convex Function
Let $f : C \to \mathbb{R}$ where $C \subseteq \mathbb{R}^n$ is a convex set. We say that f is a convex function if for every $x, y \in C$ and for every $\lambda \in [0, 1]$,

$$f((1-\lambda)x + \lambda y) \leq (1-\lambda)f(x) + \lambda f(y)$$

We say that f is strictly convex if for every $x, y \in C, x \neq y$ and for every $\lambda \in [0, 1]$,

$$f((1-\lambda)x + \lambda y) < (1-\lambda)f(x) + \lambda f(y)$$

If the above is an equality, then f is an affine function.

Definition 6.27 – Concave Functions
The function f is called concave if $-f$ is convex.

Theorem 6.9 – Affine Function
Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is both concave and convex. Then f is affine.

Definition 6.28 – Unit Simplex
The unit simplex is defined as

$$\Delta_k = \{\lambda \in \mathbb{R}^n_+ : e^T \lambda = 1\}$$

where $e = (1, \ldots, 1)^T \in \mathbb{R}^n$.

Theorem 6.10 – Jensen's Inequality
Let $C \subseteq \mathbb{R}^n$ be a convex set and $f : C \to \mathbb{R}$ be a convex function. Then for any $x_i \in C, i = 1, \ldots, k$ and $\lambda \in \Delta_k$, we have

$$f\left(\sum_{i}^{k} \lambda_{i=1} x_i\right) \leq \sum_{i=1}^{k} \lambda_i f(x_i)$$

Theorem 6.11
Let C be an open convex set in $\mathbb{R}^n$ and $f : C \to \mathbb{R}$ is a convex function. Then f is continuous.

Theorem 6.12 – First Order Condition
Assume that $f(x)$ is continuously differentiable on a convex set $C \subseteq \mathbb{R}^n$. Then the function $f(x)$ is

1. convex if and only if for all $x, y \in C$,

$$f(x) + \nabla f(x)^T (y - x) \leq f(y)$$

2. strictly convex if and only if for all $x, y \in C$ with $x \neq y$,

$$f(x) + \nabla f(x)^T (y - x) < f(y)$$

Proof. See past course notes ∎

Theorem 6.13 – Second Order Condition
Assume that $f(x)$ is twice continuously differentiable on an open convex set $C \subseteq \mathbb{R}^n$. Then the function $f(x)$ is convex if and only if $\nabla^2 f(x)$ is positive semidefinite for every $x \in C$.

Proof. See past course notes $\square$

Corollary 6.1 – Critical Points of Convex Functions
Let f be convex on convex set C, $\bar{x} \in C$. Then

$$\nabla f(\bar{x}) = 0 \iff f(\bar{x}) \leq f(y)$$

for all $y \in C$ (So $\bar{x}$ is a global minimizer of f on C).
So, $\bar{x}$ is a critical point of f if and only if it is a global minimizer of f on C.

Lemma 6.4
Let $C \subseteq \mathbb{R}^n$, C convex. Let $f : C \to \mathbb{R}$ be a C^1-smooth convex function on C. Then for $x, y \in C$,

$$f(y) < f(x) \implies \underbrace{f'(x; y - x)}_{\nabla f(x)^T(y-x)} < 0$$

Theorem 6.14
Let $C \subseteq \mathbb{R}^n$ be a convex set. Then any local minimizer (in C) of a convex function $f : C \to \mathbb{R}$ is also a global minimizer of f on C. Furthermore, any local minimizer of a strictly convex function $f : C \to \mathbb{R}$ is the unique global minimizer of f on C.

Lemma 6.5 – Global Maximizer of Convex Functions
Let $f : C \to \mathbb{R}$, f convex, $C \subseteq \mathbb{R}^n$ convex. Let $\bar{x}, \bar{y} \in C$ with

$$\bar{x} \in \text{argmax}_{x \in C} f(x), \quad f(\bar{x}) > f(\bar{y})$$

then $\bar{x} \notin \text{int}(C)$ (Must be on boundary).
So a global maximum of a convex function on a convex set must be on the boundary of the set.

6.3.2 Operation that Preserve Convexity

Course notes

7 Constrained Optimization

7.1 Lectures 16-17

7.1.1 Review

Recall that constrained optimization problems can be formulated as

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & g_i(x) \leq 0, \quad i \in \{1, \dots, m\} \\ & h_j(x) = 0, \quad j \in \{1, \dots, p\} \end{aligned}$$

$f : \mathbb{R}^n \to \mathbb{R}$ is the objective function, $x \in \mathbb{R}^n$ is the optimization variable. The inequalities $g_i(x) \leq 0$ are called inequality constraints, the equations $h_j(x) = 0$ are called the equality constraints. If there are no constraints, i.e. $m = p = 0$ then we have an unconstrained optimization problem.

A special case of constrained optimization is linear programming:

$$\begin{aligned} \min \quad & c^T x \\ \text{s.t.} \quad & Ax = b \in \mathbb{R}^n \\ & x \geq 0 \end{aligned}$$

In this case, the equality constraints are

$$h_i(x) = A_i x - b_i$$

where A_i is the ith row vector and the inequality constraints are

$$g_i(x) = -x_i$$

For an LP, the feasible region is a polyhedron. Recall that a polyhedron is a set of the form

$$P = \{x \in \mathbb{R}^n : Ax \leq b\}$$

where A is a $m \times n$ matrix and $b \in \mathbb{R}^m$.

7.1.2 Linear Programming, Lagrange Multipliers, and Implicit Function Theorem

Example 7.1 – Simplex Method
Consider the following LP in standard form:

$$\begin{aligned} \min \quad & c^T x \\ \text{s.t.} \quad & Ax = b \in \mathbb{R}^n \\ & x \geq 0 \end{aligned}$$

where A has full row rank (constraints are in linear independence (LICQ)).
Let $x = \begin{pmatrix} x_B \\ x_N \end{pmatrix}$ denote a basic feasible solution, BFS. So x is feasible and $x_N = 0$.
We call x_B the basic variables and x_N the non-basic variables.
We let $A = \begin{pmatrix} B & N \end{pmatrix}$ where B, the basis matrix, is an $m \times m$ invertible matrix. W.L.O.G. we have assumed the first m variables to be the basic variables, so the basis is $B = \{1, \dots, m\}$.
We split the constraints $Ax = b$ into two parts,

$$Bx_B + Nx_N = b$$

We solve our equation w.r.t. x_B to get

$$x_B = B^{-1}b - B^{-1}Nx_N$$

Similarly, we can split our objective function $c^T x$ into two parts,

$$c_B^T x_B + c_N^T x_N$$

Substituting $x_B = B^{-1}b - B^{-1}Nx_N$ into objective function gives

$$c_B^T x_B + c_N^T x_N = c_B^T(B^{-1}b - B^{-1}Nx_N) + c_N^T x_N$$

Then let $y^T = c_B^T B^{-1}$ denote the dual variable estimates. The objective function becomes

$$\begin{aligned} c^T x &= y^T b - y^T N x_N + c_N^T x_N \\ &= y^T b + (c_N^T - y^T N)x_N \end{aligned}$$

We can see that the objective values decreases if we find a reduced cost $c_N^T - y^T N < 0$ and increase the corresponding nonbasic variable x_j, i.e. j enters the basis. Review Simplex method if needed.

Example 7.2
Example of Lagrange multipliers and implicit function thoerem here that I do not understand.

Theorem 7.1 – Lagrange Multipliers for Equality Constraints
Suppose that $f : C \to \mathbb{R}$ and $h : C \to \mathbb{R}^m$ are sufficiently smooth functions on the open set $C \subseteq \mathbb{R}^n$. Let $x \in C$ be a local minimum of f subject to the constraints $h(x) = 0, x \in C$. In addition, assume the following regularity condition at x (constraint qualification at x)

$$h'(x) \text{ is onto}$$

Then there exists $\lambda \in \mathbb{R}^m$ (a Lagrange multiplier vector) such that

$$0 = \nabla f(x) + \langle \lambda, h'(x) \rangle$$

Equivalently, the gradient $\nabla f(x)$ is in the range of $h'(x)^T$.

7.2 Optimality Conditions and Geometry

Consider the following LP:

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & x \in Q \end{aligned}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is C^1 or C^2 smooth and Q is a closed set.
We say that $\bar{x}$ is a local minimizer if there exists $\epsilon > 0$ such that

$$\forall x \in Q \cap B_\epsilon(\bar{x}), \quad f(x) \geq f(\bar{x})$$

where $B_\epsilon(\bar{x})$ is a ball of radius ϵ centered at $\bar{x}$. For an NLP, the feasible set is

$$Q = \{x : g_i(x) \leq 0, h_j(x) = 0, i \in I, j \in E\}$$

where I, E are index sets for the inequality and equality constraints respectively.
For a convex program, the feasible set is a convex set.

Example 7.3

If we have the LP

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & c_1(x_1, x_2) = x_1^2 + x_2^2 = 0 \end{aligned}$$

Let x^* be a local optimal solution. Then $\nabla f(x^*) \in \text{span}(\nabla c_i)$ by Lagrange Multiplier since

$$\nabla f(x^*) + \lambda \nabla c_1(x^*) = 0$$

and rearranging gives

$$\nabla f(x^*) = -\lambda \nabla c_1(x^*)$$

Example 7.4
Consider the LP

$$\begin{aligned} \min \quad & f(x) = x_1 + x_2 \\ \text{s.t.} \quad & c_1(x_1, x_2) = 1 - x_1^2 - x_2^2 \geq 0 \end{aligned}$$

Let $\bar{x}$ be a local optimal solution, then

1. If $\bar{x}$ satisfies $c_1(\bar{x}) > 1$, then $\nabla f(\bar{x}) = 0$.
2. If $\bar{x}$ satisfies $c_1(\bar{x}) = 1$, then $\nabla f(\bar{x}) \in \underbrace{\text{cone}\{\nabla c_1(\bar{x})\}}_{-\lambda \nabla c_1(\bar{x}), \lambda \leq 0}$.

Definition 7.1 – Active Set
Let $\bar{x}$ be a feasible solution of an NLP. Then the set

$$A(\bar{x}) = \{i \in I : g_i(\bar{x}) = 0\}$$

is called the active set at the feasible point $\bar{x}$. The inequality constraint $i \in I$ is said to be active at a feasible point $\bar{x}$ if $g_i(\bar{x}) = 0$ and inactive if $g_i(\bar{x}) < 0$.

Theorem 7.2
Let $\bar{x}$ be a local minimizer of

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & g_i(x) \leq 0, \quad i \in I \end{aligned}$$

where $f, g_1, \ldots, g_m \in C^1$. Then there does not exist vector d such that

$$\nabla f(\bar{x})^T d < 0 \quad \text{and} \quad \nabla g_i(\bar{x})^T d < 0, \quad \forall i \in A(\bar{x})$$

where $A(\bar{x}) = \{i \in I : g_i(\bar{x}) = 0\}$ is the active set at $\bar{x}$.

Proof. Suppose such d exists. Then by Taylor's Theorem, there exists $\alpha' > 0$ such that for every $\alpha \in (0, \alpha')$

$$\begin{aligned} f(\bar{x} + \alpha d) &< f(\bar{x}) \\ g_i(\bar{x} + \alpha d) &< g_i(\bar{x}) = 0 \quad \forall i \in A(\bar{x}) \end{aligned}$$ and

which contradicts the fact that $\bar{x}$ is a local minimizer. ☐

Lemma 7.1
Let $A \in \mathbb{R}^{m\times n}$. Then exactly one of the following statements holds:

1. There exists $x \in \mathbb{R}^n$ such that $Ax < 0$
2. There exists $y \in \mathbb{R}^m$ such that $A^T y = 0$ and $y \geq 0$

Theorem 7.3
Let $\bar{x}$ be a local minimizer of

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & g_i(x) \leq 0, \quad i \in I \end{aligned}$$

where $f, g_1, \ldots, g_m \in C^1$. Assume that the set $\{\nabla g_i(\bar{x}) : i \in A(\bar{x})\}$ is linearly independent. Then there exists $\lambda_1, \ldots, \lambda_m \geq 0$ such that

$$\begin{aligned} \nabla f(\bar{x}) + \sum_{i=1}^{m} \lambda_i \nabla g_i(\bar{x}) &= 0 \\ \lambda_i g_i(\bar{x}) &= 0 \quad \forall i \in I \end{aligned}$$

Definition 7.2 – LICQ
Given $\bar{x}$ and $A(\bar{x})$, we say that the linear independence constraint qualification (LICQ) holds if the set of active constraint gradients

$$\{\nabla g_i(\bar{x}) : i \in A(\bar{x})\} \cup \{\nabla h_j(\bar{x}) : j \in E\}$$

is linearly independent.

Definition 7.3 – Lagrangian Function for NLP
We define the Lagrangian function for an NLP as

$$\begin{aligned} \mathcal{L}(x, \lambda, \mu) &= f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) + \sum_{j=1}^{p} \mu_j h_j(x) \\ &= f(x) + \lambda^T g_I(x) + \mu^T h_E(x) \end{aligned}$$

Theorem 7.4 – First Order KTT Necessary Condition
Suppose x^* is a local minimizer of an NLP and $f, g_1, \ldots, g_m, h_1, \ldots, h_p \in C^1$ and LICQ holds at x^*. Then there exists Lagrangian multipliers $\lambda^* \in \mathbb{R}^m, \mu^* \in \mathbb{R}^p$ such that

$$\begin{aligned} \nabla_x \mathcal{L}(x^*, \lambda^*, \mu^*) = 0, &\quad \lambda^* \geq 0 && \text{Dual feasibility} \\ g_i(x^*) \leq 0, &\quad i \in I, \quad h_j(x^*) = 0, \quad j \in E && \text{Primal feasibility} \\ \lambda_i^* g_i(x^*) = 0, &\quad i \in I && \text{Complementary slackness} \end{aligned}$$

We call the dual feasibility, primal feasibility and complementary slackness conditions as KTT conditions, and call the point satisfying these conditions as KTT point.

Definition 7.4 – Constraint Qualification
A condition on the constraints is called a constrained qualification (CQ) at a feasible point $\bar{x}$:
if $\bar{x}$ is a local minimizer then KTT conditions hold at $\bar{x}$.

Definition 7.5 – Cone
We say that the set K is a cone if for every $x \in K$ and for every $\alpha > 0$, we have $\alpha x \in K$.

Definition 7.6 – Tangent Vector
Let $\bar{x} \in Q \subseteq \mathbb{R}^n$. A vector $v \in \mathbb{R}^n$ is tangent to Q at $\bar{x} \in Q$ if there exists sequences $\{x_k\} \subseteq Q$ and $\{t_k\} \subset R_+$ such that

$$x_k \to \bar{x}, \qquad \underbrace{t_k \downarrow 0}_{t_k > 0, t_k \to 0, \text{non-increasing}} \qquad , \quad \frac{1}{t_k}(x_k - \bar{x}) \to v$$

Definition 7.7 – Tangent Cone
We denote the set of all tangent vectors to Q at $\bar{x}$ as the tangent cone $T_Q(\bar{x})$.

Definition 7.8 – Linearized Feasible Directions
Given an NLP with a feasible set F, the set of linearized feasible directions at $\bar{x} \in F$ is defined as

$$\mathcal{L}_F(\bar{x}) = \{d : d^T \nabla g_i(\bar{x}) \leq 0, \forall i \in A(\bar{x}) \cap I, \quad d^T \nabla h_j(\bar{x}) = 0, \forall j \in E\}$$

8 Matrix Calculus

Youtube Series on Matrix Calculus: here

8.1 Kronecker Product

Given two matrices $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{p\times q}$, the Kronecker product of A and B is defined as

$$A \otimes B = \begin{bmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{bmatrix} \in \mathbb{R}^{mp\times nq}$$

For example,

$$\begin{bmatrix} 2 & 3 & 1 \\ 5 & 2 & 3 \\ 2 & 6 & 4 \end{bmatrix} \otimes \begin{bmatrix} 1 & 7 \end{bmatrix} = \begin{bmatrix} 2 & 14 & 3 & 21 & 1 & 7 \\ 5 & 35 & 2 & 14 & 3 & 21 \\ 2 & 14 & 6 & 42 & 4 & 28 \end{bmatrix}$$

8.2 Differentiating w.r.t a Scalar

Suppose we have y scalar, $\mathbf{y}$ vector, and Y matrix

$$y = \sin(x^2) \quad \mathbf{y} = \begin{pmatrix} x \\ \cos x \\ 2x^2 \end{pmatrix} \quad Y = \begin{bmatrix} x^2+1 & \cos x \\ \sin x & x-1 \end{bmatrix}$$

Differentiating them w.r.t. a scalar x is just element-wise differentiation.

$$\begin{aligned} \frac{\partial y}{\partial x} &= \cos(x^2)2x \\ \frac{\partial \mathbf{y}}{\partial x} &= \begin{pmatrix} \frac{\partial}{\partial x}x \\ \frac{\partial}{\partial x}\cos x \\ \frac{\partial}{\partial x}2x^2 \end{pmatrix} = \begin{pmatrix} 1 \\ -\sin x \\ 4x \end{pmatrix} \\ \frac{\partial Y}{\partial x} &= \begin{bmatrix} \frac{\partial}{\partial x}x^2+1 & \frac{\partial}{\partial x}\cos x \\ \frac{\partial}{\partial x}\sin x & \frac{\partial}{\partial x}x-1 \end{bmatrix} = \begin{bmatrix} 2x & -\sin x \\ \cos x & 1 \end{bmatrix} \end{aligned}$$

8.3 Differentiating w.r.t a Vector

Definition 8.1 – Layout Notation
When we differentiate we have to follow one of two layout notations: either Numerator layout notation or Denominator layout notation.

Denominator layout notation first takes the transpose of the numerator before differentiating. For example if we do $\frac{\partial y}{\partial x}$, we will transpose y first.

Numerator layout notation first takes the transpose of the denominator before differentiating. For example if we do $\frac{\partial y}{\partial x}$, we will transpose x first. So we are always differentiating w.r.t. a row vector. So if $\mathbf{x} = (x_1\ x_2\ x_3\ \cdots)^T$ then

$$\frac{\partial}{\partial \mathbf{x}} = \begin{pmatrix} \frac{\partial}{\partial x_1} & \frac{\partial}{\partial x_2} & \cdots \end{pmatrix}$$

Below we will use numerator layout notation.

Suppose we have y scalar, $\mathbf{y}$ vector, and Y matrix

$$y = \sin(x+yz), \quad \mathbf{y} = \begin{pmatrix} e^{xyz} \\ x^2z \\ yx \end{pmatrix}, \quad Y = \begin{bmatrix} x^2yz & xy^2z \\ xyz^2 & \ln(xyz) \end{bmatrix}$$

We will take the derivatives w.r.t. $\mathbf{r} = (x\ y\ z)^T$.

$$
\begin{aligned}
\frac{\partial y}{\partial \mathbf{r}} &= \frac{\partial}{\partial \mathbf{r}} \otimes y && \text{Treat this as Kronecker product} \\
&= \begin{pmatrix} \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \end{pmatrix} \otimes \sin(x + yz) \\
&= \begin{pmatrix} \cos(x+yz) & z\cos(x+yz) & y\cos(x+yz) \end{pmatrix} \\
\frac{\partial \mathbf{y}}{\partial \mathbf{r}} &= \frac{\partial}{\partial \mathbf{r}} \otimes \mathbf{y} \\
&= \left(\frac{\partial}{\partial x} \begin{pmatrix} e^{xyz} \\ x^2 z \\ yx \end{pmatrix} \ \frac{\partial}{\partial y} \begin{pmatrix} e^{xyz} \\ x^2 z \\ yx \end{pmatrix} \ \frac{\partial}{\partial z} \begin{pmatrix} e^{xyz} \\ x^2 z \\ yx \end{pmatrix} \right) \\
&= \begin{bmatrix} yze^{xyz} & xze^{xyz} & xye^{xyz} \\ 2xz & 0 & x^2 \\ y & x & 0 \end{bmatrix} && \text{also called the Jacobian} \\
\frac{\partial Y}{\partial \mathbf{r}} &= \frac{\partial}{\partial \mathbf{r}} \otimes Y \\
&= \left(\frac{\partial}{\partial x} \begin{bmatrix} x^2yz & xy^2z \\ xyz^2 & \ln(xyz) \end{bmatrix} \ \frac{\partial}{\partial y} \begin{bmatrix} x^2yz & xy^2z \\ xyz^2 & \ln(xyz) \end{bmatrix} \ \frac{\partial}{\partial z} \begin{bmatrix} x^2yz & xy^2z \\ xyz^2 & \ln(xyz) \end{bmatrix} \right) \\
&= \begin{bmatrix} 2xyz & y^2z & x^2z & 2xyz & x^2y & xy^2 \\ yz^2 & \frac{yz}{xyz} & xz^2 & \frac{1}{y} & 2xyz & \frac{1}{2} \end{bmatrix}
\end{aligned}
$$

8.4 Differentiating w.r.t a Matrix

Suppose we have y scalar, $\mathbf{y}$ vector, and Y matrix

$$
y = 4x_3 + 3x_2 + 2x_1 + x_0, \quad \mathbf{y} = \begin{pmatrix} e^{x_0 x_1} \\ e^{x_2 x_3} \end{pmatrix}, \quad Y = \begin{bmatrix} \sin(x_0 + 2x_1) & 2x_1 + x_3 \\ 2x_0 + x_2 & \cos(2x_2 + x_3) \end{bmatrix}
$$

We differentiate w.r.t. matrix $X = \begin{bmatrix} x_0 & x_1 \\ x_2 & x_3 \end{bmatrix}$. So

$$
\frac{\partial}{\partial X} = \begin{bmatrix} \frac{\partial}{\partial x_0} & \frac{\partial}{\partial x_2} \\ \frac{\partial}{\partial x_1} & \frac{\partial}{\partial x_3} \end{bmatrix}
$$

Differentiating gives

$$
\begin{aligned}
\frac{\partial y}{\partial X} &= \frac{\partial}{\partial X} \otimes y \\
&= \begin{bmatrix} 1 & 3 \\ 2 & 4 \end{bmatrix} \\
\frac{\partial \mathbf{y}}{\partial X} &= \frac{\partial}{\partial X} \otimes \mathbf{y} \\
&= \begin{bmatrix} x_1 e^{x_0 x_1} & 0 & 0 & x_3 e^{x_2 x_3} \\ x_0 e^{x_0 x_1} & 0 & 0 & x_2 e^{x_2 x_3} \end{bmatrix} \\
\frac{\partial Y}{\partial X} &= \frac{\partial}{\partial X} \otimes Y \\
&= \begin{bmatrix} \cos(x_0 + 2x_1) & 0 & 0 & 0 \\ 2 & 0 & 1 & -2\sin(2x_2 + x_3) \\ 2\cos(x_0 + 2x_1) & 2 & 0 & 1 \\ 0 & 0 & 0 & -\sin(2x_2 + x_3) \end{bmatrix}
\end{aligned}
$$