Nonlinear Optimization

CO 367 (Fall 2018) with Professor Laurent Poirrier

University of Waterloo

Scribe : Saiyue Lyu

Contents

List of Definitions

Chapter 1

Introduction

Mathematical Optimization (formally math programming)

Find a best soln to the model of a problem

Application :

- Operation Research
    1) Scheduling and Planning
    2) Supply Chain Management
    3) Vehicle Routing
    4) Power Grid Optimization
- Statistics and Machine Learning
    1) Curve Fitting
    2) Classification, Clustering, SVM ...
    3) Deep Learning
- Finance
- Optimal Control
- Biology – Protein Folding

Optimization

$$\begin{aligned} (OPT)\underset{X}{\text{minimize}} \quad & f(x) && \text{objective function} \\ \text{subject to} \quad & g_i(x) \leq 0,\ \forall i = 1, \cdots, m && \text{constraints} \\ & x \in \mathbb{R}^n. \end{aligned}$$

Remark

1) $\max f(x) = -\min -f(x)$

2) $\{x \in \mathbb{R}^n, g(x) \geq 0\} = \{x \in \mathbb{R}^n, -g(x) \leq 0\}$

3) $\{x \in \mathbb{R}^n, g(x) \leq b\} = \{x \in \mathbb{R}^n, g(x) - b \leq 0\}$

1.1 Classification of Solns

Definition 1.1.1 (Open ball & Closure)

The open ball of radius δ around $\bar{x}$ is $B_\delta(\bar{x}) = \{x \in \mathbb{R}^n, ||x - \bar{x}|| < \delta\}$

The closure of $B_\delta(\bar{x})$ is $\overline{B_\delta}(\bar{x}) = \{x \in \mathbb{R}^n, ||x - \bar{x}|| \leq \delta\}$

Definition 1.1.2 (Global & Local Minimizer)

Consider $f : D \to \mathbb{R}$. the point $x^* \in D$ is

- a global minimizer for f on D if $f(x^*) \leq f(x), \forall x \in D$
- a strict global minimizer for f on D if $f(x^*) < f(x), \forall x \in D, x \neq x^*$
- a local minimizer for f on D if $\exists \delta > 0, f(x^*) \leq f(x), \forall x \in B_\delta(x^*) \cap D$
- a strict local minimizer for f on D if $\exists \delta > 0, f(x^*) < f(x), \forall x \in B_\delta(x^*) \cap D, x \neq x^*$

1.2 Classification of Problems

1. If $f(x) = 0, \forall x \in \mathbb{R}^n$, then (OPT) is a feasible problem
2. If we have $m = 0$ constraints, then (OPT) is an unconstrained optimization problem.

1.3 Classification of Problems – Types of functions involved

Why do we care?

In the absence of hypothesis on f and g, (OPT) is unsovlable.

Remark

"Black box" optimization framework.

All we have is an "oracle" that can compute values of $f(x)$ for any x (and possibly some derivatives)

Example 1.3.1
Consider $h(x) = \begin{cases} 0, & \text{if } x \in \mathbb{Z}^n \\ 1, & \text{otherwise} \end{cases}$

$$\begin{aligned} \underset{X}{\text{minimize}} \quad & f(x) \\ \text{subject to} \quad & g_i(x) \leq 0, \ \forall i = 1, \cdots, m \\ & h(x) \leq 0, \\ & x \in \mathbb{R}^n. \end{aligned}$$

In other word, we want $x \in \mathbb{Z}^n$, where $\mathbb{Z}^n$ is a lattice

Definition 1.3.1 (discrete optimization problem)
When the constraints of (OPT) restrict solns to a lattice, then (OPT) is called a discrete optimization problem

Definition 1.3.2 (Continuous Function)
A function $f : D \to \mathbb{R}$ is continuous over D if $\forall \epsilon > 0, \exists \delta > 0$ such that $|x-y| < \delta \Leftarrow |f(x)-f(y)| < \epsilon, \forall x, y \in D$

Definition 1.3.3 (C^k-smooth)
A function $f : D \to \mathbb{R}$, $D \subset \mathbb{R}^n$ is open, then f is C^k-smooth over D $\big($i.e. $f \in C^k(D)\big)$ if all its $\leq k$-th derivatives are continuous over D

Example 1.3.2
$h(x) = \begin{cases} 1, & \text{if } x \geq 2 \\ -1, & \text{if } x < 2 \end{cases}$ is discontinuous

$g(x) = |x - 2|$ is continuous and C^0 smooth

$f(x) = \begin{cases} \frac{1}{2}(x-2)^2, & \text{if } x \geq 2 \\ \frac{1}{2}(2-x)^2, & \text{if } x < 2 \end{cases}$ is continuous and C^1 smooth

Definition 1.3.4 (Gradient)
Let $f \in C^1(D)$ for some $D \subset \mathbb{R}^n$. Its Gradient $\nabla f \in C^0(D) : D \to \mathbb{R}^n$ is given by

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

Definition 1.3.5 (Hessian)
Let $f \in C^2(D)$ for some $D \subset \mathbb{R}^n$. Its Hessian $\nabla^2 f \in C^1(D) : D \to \mathbb{R}^{n \times n}$ is given by

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

Remark
If f and g are linear functions, then (OPT) is a linear programming problem.

Chapter 2

Linear Algebra

2.1 Vector and Matrix Norm

Definition 2.1.1 (Norm)

A norm $||\cdot||$ on $\mathbb{R}^n$ assigns a scalar $||x||$ to every $x \in \mathbb{R}^n$ such that

1) $||x|| \geq 0, \forall x \in \mathbb{R}^n$

2) $||c \cdot x|| = |c| \cdot ||x|| \, \forall c \in \mathbb{R}, x \in \mathbb{R}^n$

3) $||x|| = 0 \iff x = 0$

4) $||x + y|| \leq ||x|| + ||y||$

Remark

$$
\begin{aligned}
&L^k Norm && ||x||_k = (\sum_{i=1}^{n} |x_i|^k)^{1/k} \\
&Manhattan Norm && ||x||_1 = \sum |x_i| \\
&Euclidean Norm && ||x||_2 = \sqrt{\sum x_i^2} \\
&Infinite Norm && ||x||_\infty = \max |x_i|
\end{aligned}
$$

Theorem 2.1.1 (Schwartz Inequality)

$\forall x, y \in \mathbb{R}^n, |x^T y| \leq ||x||_2 \cdot ||y||_2$, the equality holds when $x = \lambda y$ for some $\lambda \in \mathbb{R}$

Theorem 2.1.2 (Pythagorean Thm)

If $x, y \in \mathbb{R}^n$ are orthogonal, then $||x + y||_2^2 = ||x||_2^2 + ||y||_2^2$

Definition 2.1.2 (Induced Norm)

Given a vector norm $||\cdot||$, the induced matrix norm associates a scalar $||A||$ to all $A \in \mathbb{R}^{n \times n}$ with $||A|| = \max_{||x||=1} ||Ax||$

Proposition 2.1.1

$||A||_2 = \max_{||x||_2=1} ||Ax||_2 = \max_{||x||_2=||y||_2=1} |y^T Ax|$

Proof
Apply Schwartz Inequality to $|y^T Ax|$

Proposition 2.1.2

$||A||_2 = ||A^T||_2$

Proof
Swap x and y in the above Proposition 2.1.1

Proposition 2.1.3

Let $A \in \mathbb{R}^{n\times n}$, TFAE:

1) A is nonsingular

2) A^T is nonsingular

3) $\forall x \in \mathbb{R}^n \setminus \{0\}, Ax \neq 0$

4) $\forall b \in \mathbb{R}^n, \exists x \in \mathbb{R}^n$ unique such that $Ax = b$

5) Columns of A are linear independent

6) Rows of A are linearly independent

7) $\exists B \in \mathbb{R}^{n\times n}$ unique such that $AB = I = BA$, where B is the inverse of A

8) $\forall A, B \in \mathbb{R}^{n\times n}, (AB)^{-1} = B^{-1}A^{-1}$ if B^{-1} exists

2.2 Eigenvalues

Definition 2.2.1 (Eigenvalue & Eigenvector)

The characteristic polynomial $\phi : \mathbb{R} \to \mathbb{R}$ of $A \in \mathbb{R}^{n\times n}$ is $\phi(\lambda) = \det(A - \lambda I)$. It has n (possibly complex or repeated) roots, which are the eigenvalues of A. Given an eigenvalue λ of A, $x \in \mathbb{R}^n$ is the corresponding eigenvector of A if $Ax = \lambda x$

Proposition 2.2.1

Given $A \in \mathbb{R}^{n\times n}$

1) λ is an eigenvalue $\iff \exists$ a corresponding eigenvector

2) A is singular $\iff$ it has a zero eigenvalue

3) If A is triangular, then its eigenvector are its diagonal entries

4) If $S \in \mathbb{R}^{n\times n}$ is nonsingular and $B = SAS^{-1}$, then A, B have the same eigenvalues

5) If the eigenvalues of A are $\lambda_1, \cdots, \lambda_n$, then

- the eigenvalues of $A + cI$ are $\lambda_1 + c, \cdots, \lambda_n + c$
- the eigenvalues of A^k are $\lambda_1^k, \cdots, \lambda_n^k, k \in \mathbb{R}$
- the eigenvalues of A^{-1} are $\frac{1}{\lambda_1}, \cdots, \frac{1}{\lambda_n}$

- the eigenvalues of A^T are $\lambda_1, \cdots, \lambda_n$

Definition 2.2.2 (Spectral Radius)
The spectral radius of $\rho(A)$ of $A \in \mathbb{R}^{n\times n}$ is $\max\limits_{\lambda \text{ is eigenvalue}} |\lambda|$

Proposition 2.2.2
For any induced norm $||\cdot||, \rho(A) \leq ||A^k||^{1/k}$ for $k = 1, 2, \cdots$

Proof
Trick : $||A^k|| = \max\limits_{||y||=1} ||A^k y|| = \max\limits_{y\neq 0} \frac{1}{||y||}||A^k y||$

In particular, let λ be any eigenvalue of A, x be the corresponding eigenvector

$$\begin{aligned} ||A^k|| \geq \frac{1}{||x||}||A^k \cdot x|| &= \frac{1}{||x||}||A \cdots A \cdot x|| \\ &= \frac{1}{||x||}||\lambda^k \cdot x|| \\ &= |\lambda^k| \end{aligned}$$

So for any eigenvalue $\lambda, ||A^k|| \geq |\lambda|^k$

Therefore $||A^k||^{1/k} \geq |\lambda|, \forall \lambda$, thus $||A^k||^{1/k} \geq \rho(A)$

Proposition 2.2.3
For any induced norm $||\cdot||, \lim\limits_{k\to\infty} ||A^k||^{1/k} = \rho(A)$

Proof
Too long, omitted

2.3 Symmetric Matrices

Proposition 2.3.1
Let $A \in \mathbb{R}^{n\times n}$ be symmetric, then

1) Its eigenvalues are all $\mathbb{R}$eal

2) Its eigenvetors are n mutually orthogonal $\mathbb{R}$eal nonzero vectors

3) If the n eigenvectors $x_1, \cdots, x_n \in \mathbb{R}^n$ are normalized such that $||x||_2 = 1$ with corresponding eigenvalues $\lambda_1, \cdots, \lambda_n$, then $A = \sum_{i=1}^n \lambda_i x_i x_i^T$

Proof
Easy

Proposition 2.3.2
Let $A \in \mathbb{R}^{n\times n}$ be symmetric, then $||A||_2 = \rho(A)$

Proof
We already know $\rho(A) \leq ||A^k||^{1/k}$, in particular, $\rho(A) \leq ||A||_2$

It remains to show that $\rho(A) \geq ||A||_2$

Because the eigenvectors $x_i, i = 1, \cdots, n$ of A can be assumed ,utually orthogonal

Then we can write any $y \in \mathbb{R}^n$ as $y = \sum_{i=1}^n \beta_i x_i$ for some $\beta_i \in \mathbb{R}$

By Pythagorean Thm, $||y||_2^2 = \sum \beta_i^2 ||x_i||_2^2$

Now $Ay = A \sum \beta_i x_i = \sum \beta_i A x_i = \sum \beta_i \lambda_i x_i$

Since all x_i are mutually orthogonal, by Pthahorean Thm again, have

$$\begin{aligned} ||Ay||_2^2 &= \sum \beta_i^2 \lambda_i^2 ||x_i||_2^2 \\ &\leq \sum \beta_i^2 \rho(A)^2 ||x_i||_2^2 \\ &= \rho(A)^2 ||y||_2^2 \end{aligned}$$

By which we get, $||Ay||_2 \leq \rho(A) ||y||_2$

Also by the definition, we have

$$\begin{aligned} ||A||_2 &= \max_{y \neq 0} \frac{1}{||y||_2} ||Ay||_2 \\ &\leq \max_{y \neq 0} \frac{1}{||y||_2} \cdot \rho(A) ||y||_2 \\ &= \rho(A) \end{aligned}$$

Proposition 2.3.3

Let $A \in \mathbb{R}^{n \times n}$ be symmetric with eigenvalues $\lambda_1, \cdots, \lambda_n \in \mathbb{R}$

Then $\forall y \in \mathbb{R}^n, \lambda_1 ||y||_2^2 \leq y^T A y \leq \lambda_n ||y||_2^2$

Proof

Again, write $y = \sum \beta_i x_i$ for some $\beta_i \in \mathbb{R}$ with x_i are the orthogonal eigenvectors

On the one hand,

$$\begin{aligned} y^T A y &= (\sum \beta_i x_i)^T (\sum \beta_i \lambda_i x_i) \\ &= \sum \beta_i^2 \lambda_i x_i^T x_i \text{ as } x_i x_j = 0 \text{ if } i \neq j \\ &= \sum \beta_i^2 \lambda_i ||x_i||_2^2 \end{aligned}$$

WLOG, we can assume $||x_i||_2 = 1$, then we have

$$y^T A y = \sum \beta_i^2 \lambda_i$$

On the other hand,

$$||y||_2^2 = \sum \beta_i^2 ||x_i||_2^2 = \sum \beta_i^2$$

Clearly, we have

$$\begin{array}{ccccc} \lambda_1 \sum \beta_i & \leq & \sum \beta_i^2 \lambda_i & \leq & \lambda_n \sum \beta_i^2 \\ \lambda_1 ||y||_2^2 & \leq & y^T A y & \leq & \lambda_n ||y||_2^2 \end{array}$$

Remark
Why can we assume $||x_i||_1 = 1$?

As x_i being the eigenvectors of A are defined up to scalar $Ax = \lambda x$, we have

$$A(\frac{1}{||x||}x) = \lambda(\frac{1}{||x||}x)$$

Proposition 2.3.4

Let $A \in \mathbb{R}^{n\times n}$ be symmetric, then $||A^k||_2 = ||A||_2^k, \forall k = 1, 2, \cdots$

Proof
Since $A^T = A$, then $(A^k)^T = (A\cdots A)^T = A^T \cdots A^T = A \cdots A = A^k$

Since A^k is symmetric, then $||A^k||_2 = \rho(A^k)$

We know that the eigenvalues of A^k are $\lambda_1^k, \cdots, \lambda_n^k$

Thus $\rho(A^k) = \big(\rho(A)\big)^k$

We know $\rho(A) = ||A||_2$, therefore $||A||_2^k = ||A^k||_2$

Proposition 2.3.5

Let $A \in \mathbb{R}^{n\times n}$ (not necessary symmetric), then $||A||_2^2 = ||A^TA||_2 = ||AA^T||_2$

Proof
On the one hand,

$||Ax||_2^2 = (Ax)^T(Ax) = x^T(A^TAx) \leq ||x||_2 \cdot ||A^TAx||_2 \leq ||x||_2 \cdot ||AA^T||_2 \cdot ||x||_2$

$||A||_2^2 = \max_{x\in\mathbb{R}^n} \dfrac{1}{||x||_2^2} \cdot ||Ax||_2^2 \leq ||A^TA||_2$

On the other hand,

$$\begin{aligned}
||A^TA|| &= \max_{||x||=||y||=1} |y^TA^TAx| \\
&\leq \max_{||y||=1,||x||=1} ||y^TA^T||_2 \cdot ||Ax||_2 \text{ by CS ineq} \\
&= \big(\max_{||y||=1} ||y^TA^T||_2\big)\big(\max_{||x||=1} ||Ax||_2\big) \\
&= ||A||_2 \cdot ||A||_2 \\
&= ||A||_2^2
\end{aligned}$$

Combine these two things, we get $||A||_2^2 = ||A^TA||_2$

For the other equality, swap A and A^T in the proof and use $||A||_2 = ||A^T||_2$

Proposition 2.3.6

$||A^{-1}||_2$ is $\frac{1}{|\lambda_1|}$, where λ_1 is the smallest magnitude eigenvalue of A

Proof
We know $||A^{-1}||_2 = \rho(A^{-1})$, and the eigenvalues of A^{-1} are the inverse of the eigenvalues of A

2.4 Positive Definite Matrices

Definition 2.4.1 (Positive Definite & Positive Semidefinite)

A symmetric matrix $A \in \mathbb{R}^{n\times n}$ is positive definite if $x^TAx > 0, \forall x \in \mathbb{R}^n, x \neq 0$, it is positive semidefinite if $x^TAx \geq 0, \forall x \in \mathbb{R}^n$

Remark
pd. for symmetric positive definite, psd. for symmetric positive semidefinite

Proposition 2.4.1

For any $A \in \mathbb{R}^{n\times n}$(possibly non square), A^TA is psd. Then the matrix A^TA is pd iff A has full column rank (i.e. $rank(A) = n$; which implies $m \geq n$)

Proof
1) A^TA is square and symmetric (immediate)

2) A^TA is psd. : $\forall x \in \mathbb{R}^n, x^T(A^TA)x = (Ax)^T(Ax) = ||Ax||_2^2 \geq 0$

3) pd. iff $rank(A) = n$:

$$
\begin{aligned}
& x^TA^TAx > 0, \forall x \in \mathbb{R}^n, x \neq 0 \\
\iff & ||Ax||_2^2 > 0 \\
\iff & ||Ax||_2 > 0 \\
\iff & Ax \neq 0, \forall x \in \mathbb{R}^n, x \neq 0 \\
\iff & \text{ the columns of } A \text{ are linearly independent} \\
\iff & rank(A) = n
\end{aligned}
$$

Corollary 2.4.1

If $A \in \mathbb{R}^{n\times n}$ is square, then A^TA is pd. iff A is nonsingular

Proposition 2.4.2

A square symmetric matrix is psd. (rsp pd.) iff all its eigenvalues are ≥ 0 (rsp > 0)

Proof
We will prove the statement for psd/≥ 0, the proof is similar for pd/(¿0)

($\Rightarrow$) Let λ be an eigenvalue of A psd. and let x be the corresponding nonzero eigenvector

Then $x^tAx \geq 0$, so $x^T\lambda x = \lambda||x||_2^2 \geq 0$

Thus $\lambda \geq 0$

($\Leftarrow$) Let $\lambda_1, \cdots, \lambda_n$ be the eigenvalues of A and let $x_1, \cdots, x_n$ be the n (nonzero, real, mutually orthogonal) eigenvectors.

For any $y \in \mathbb{R}^n$, we can write

$$y = \sum \beta_i x_i \text{ for some } \beta_i \in \mathbb{R}$$

Then we have

$$\begin{aligned} y^T A y &= (\sum \beta_i x_i)^T \cdot A \cdot (\sum \beta_i x_i) \\ &= (\sum \beta_i x_i)^T (\sum \beta_i A x_i) \\ &= (\sum \beta_i x_i)^T (\sum \beta_i \lambda_i x_i) \\ &= \sum \beta_i^2 \lambda_i ||x_i||_2^2 \text{ as } x_i \text{ are orthogonal} \\ &\geq 0 \end{aligned}$$

Proposition 2.4.3

The inverse of a pd. matrix is pd.

Proof

Let $\lambda_1, \cdots, \lambda_n > 0$ be the eigenvalues of A pd.

Then the eigenvalues of A^{-1} are $\frac{1}{\lambda_1}, \cdots, \frac{1}{\lambda_n}$

Chapter 3

Convexity

3.1 Basic Intro

Definition 3.1.1 (Convex Set)
A set $C \subset \mathbb{R}^n$ is convex if $\lambda x + (1-\lambda)y \in C, \forall x, y \in C, \forall 0 \leq \lambda \leq 1$

Example 3.1.1
The set of two disjoint sets is nonconvex

A "donut" is nonconvex

Definition 3.1.2 (Convex Function)
Let $D \subset \mathbb{R}^n$ be a convex set, a function $f : D \to \mathbb{R}$ is said to be convex if $f\big(\lambda x + (1-\lambda)y\big) \leq \lambda f(x) + (1-\lambda)f(y), \forall x, y \in D, \forall 0 \leq \lambda \leq 1$

A function is said to be strict convex if a strict inequality ($<$) holds as well

Example 3.1.2
$y = x^2$ is a convex function, $\qquad y = -x^2$ is non-convex (concave)

Proposition 3.1.1
1) For any collection of $\{C_i : i \in I\}$ of convex sets, their intersection $\cap_{i \in I} C_i$ is convex

2) The vector (Minkowski) sum $\{x + y : x \in C_1, y \in C_2\}$ of two convex sets C_1, C_2 is convex

3) The image of a convex set under a linear transformation is a convex set

Definition 3.1.3 (Level Set & Epigraph)
Let $f : D \to \mathbb{R}$ be a function with D convex,

The level sets of f are $\{x \in D : f(x) \leq \alpha\}$ for all $\alpha \in \mathbb{R}$ (sometimes " $<$ ")

The epigraph of f is s subset of $\mathbb{R}^{n+1}$ given by $epi(f) = \{(x, \alpha), x \in D, \alpha \in \mathbb{R}, f(x) \leq \alpha\}$

Proposition 3.1.2
1) If $f : D \to \mathbb{R}$ is convex, then its level sets are convex as well

2) $f : D \to \mathbb{R}$ is convex iff its epigraph is a convex set

Note : The converse of 1) is not true ! For example, $f(x) = \sqrt{|x|}$

The level sets of f is $\sqrt{|x|} \leq \alpha \iff |x| \leq \alpha^2 \iff -\alpha^2 \leq x \leq \alpha^2$

However, f is not convex !

Proposition 3.1.3

1) Any linear function is convex (but not strictly convex)

2) If f is a convex function, then $g(x) = \lambda f(x)$ is convex for all $\lambda \geq 0$

3) The sum of two convex functions is a convex function

4) The maximum of two convex functions is a convex function (does not work for minimum)

Proposition 3.1.4

Any vector norm is convex (this is useful as optimize convex function is usually possible)

Proof

Let $f(x) = ||x||$, then $\forall x, y \in \mathbb{R}^n, 0 \leq \lambda \leq 1$, have

$$\begin{aligned} f\big(\lambda x + (1-\lambda)y\big) &= ||\lambda x + (1-\lambda)y|| \\ &\leq ||\lambda x|| + ||(1-\lambda)y|| \\ &= \lambda \cdot ||x|| + (1-\lambda) \cdot ||y|| \\ &= \lambda f(x) + (1-\lambda) f(y) \end{aligned}$$

3.2 Taloy's Thms

Theorem 3.2.1 (Talor's Thm For Uni-variate Functions)

$f(x+h) = \sum_{i=0}^{k} \frac{h^i}{i!} d_i(f)$, where $d_i(f)$ is the i-th derivative of f and $\phi(x) = \frac{h^{k+1}}{(k+1)!} d_{i+1} f(x+\lambda h), 0 \leq \lambda \leq 1$ is the residual function. In particular, $\lim_{h\to 0} \frac{\phi(h)}{h^k} = 0$

Theorem 3.2.2 (Talor's Thm For Multivariate Functions – 1st order ($k = 1$))

$f(x+h) = f(x) + h^T \nabla f(x) + \phi(h)$, where $\phi(h) = \frac{1}{2} h^T \nabla^2 f(x+\lambda h) h, 0 \leq \lambda \leq 1$ with $\lim_{h\to 0} \frac{\phi(h)}{||h||} = 0$

Theorem 3.2.3 (Talor's Thm For Multivariate Functions – 2nd order ($k = 2$))

$f(x+h) = f(x) + h^T \nabla f(x) + \frac{1}{2} h^T \nabla^2 f(x) h + \phi(h)$ with $\lim_{h\to 0} \frac{\phi(h)}{||h||^2} = 0$

Theorem 3.2.4 (Mean Value Thm)

Let $f : D \to \mathbb{R}, D \subset \mathbb{R}, f \in C^1(D)$, then $\forall x, y \in D, \exists z \in [x, y]$ such that $f(y) = f(x) + \nabla f(z)(y-x)$

Proof

By 0-th order Talor Expansion

Definition 3.2.1 (Directional Derivative)

The directional Derivative of f in the direction of y is $\nabla_y f(x) = \lim_{\alpha\to 0} \frac{f(x+\alpha y) - f(x)}{\alpha}$

In particular, $\nabla_{e_i} f(x) = \frac{\partial f}{\partial x_i}(x)$ and $\nabla f = \big(\nabla_{e_1} f(x) \cdots \nabla_{e_n} f(x)\big)^T$

The "direction" draws out the function

Theorem 3.2.5

Let $f \in C^1$, then $\nabla_h f = h^T \nabla f$

Proof

$$
\begin{aligned}
\nabla_h f &= \lim_{\alpha\to 0} \frac{f(x+\alpha h) - f(x)}{\alpha} \\
&= \lim_{\alpha\to 0} \frac{f(x) + \alpha h^T \nabla f(x) + \phi(\alpha h) - f(x)}{\alpha} \\
&= \lim_{\alpha\to 0} \frac{\alpha h^T \nabla f(x) + \phi(\alpha h)}{\alpha} \\
&= \lim_{\alpha\to 0} \frac{\alpha h^T \nabla f(x)}{\alpha} + \lim_{\alpha\to 0} \frac{\phi(\alpha h)}{\alpha} \\
&= h^T \nabla f(x) + \lim_{\alpha\to 0} \frac{\phi(\alpha h)}{\alpha} \\
&= h^T \nabla f(x) \text{ by definition of residual above}
\end{aligned}
$$

Proposition 3.2.1

Let $D \subset \mathbb{R}^n$ be convex and $f : D \to \mathbb{R}$ be differentiable over D. Then f is convex iff $f(z) \geq f(x) + (z-x)^T \nabla f(x), \forall x, z \in D$

Proof

($\Longrightarrow$) As D is convex, then $x + (z-x)\alpha = \alpha z + (1-\alpha)x \in D, \forall 0 \leq \alpha \leq 1$

$$
\lim_{\alpha\to 0} \frac{f\big(x+\alpha(z-x)\big) - f(x)}{\alpha} = \nabla_{z-x} f(x) = (z-x)^T f(x)
$$

By convexity of f, $\forall 0 \leq \alpha \leq 1$

$$
\begin{aligned}
f\big(x+\alpha(z-x)\big) &\leq \alpha f(z) + (1-\alpha) f(x) \\
f\big(x+\alpha(z-x)\big) - f(x) &\leq \alpha f(z) - \alpha f(x) \\
\frac{f\big(x+\alpha(z-x)\big) - f(x)}{\alpha} &\leq f(z) - f(x)
\end{aligned}
$$

Taking the $\lim_{\alpha\to 0}$

$$
(z-x)^T \nabla f(x) \leq f(z) - f(x)
$$

($\Longleftarrow$) If $f(z) \geq f(x) + (z-x)^T \nabla f(x), \forall x, z \in D$

Let $a, b \in D$ be any points in the domain of f, let $c := \alpha a + (1-\alpha) b$

$$
f(a) \geq f(c) + (a-c)^T \nabla f(x) \tag{3.1}
$$
$$
f(b) \geq f(c) + (b-c)^T \nabla f(x) \tag{3.2}
$$

Multiply (3.1) by α and (3.2) by $(1-\alpha)$, then add them together, we get:

$$\begin{aligned}\alpha f(a)+(1-\alpha)f(b) &\geq \alpha\big(f(c)+(a-c)^T\nabla f(c)\big)+(1-\alpha)\big(f(c)+(b-c)^T\nabla f(c)\big)\\ &\geq f(c)+\alpha(a-c)^T\nabla f(c)+(1-\alpha)(b-c)^T\nabla f(c)\\ &\geq f(c)+(\alpha a-\alpha c+b-\alpha b-c+\alpha c)^T\nabla f(c)\\ &\geq f(c)+(\alpha a+b-\alpha b-c)^T\nabla f(c)\\ &\geq f(c)\\ &\geq f\big(\alpha a+(1-\alpha)b\big)\end{aligned}$$

Hence f is convec over D

Proposition 3.2.2

Let $f:\mathbb{R}^n\to\mathbb{R}, f\in C^2(D)$, then

(1) If $\nabla^2 f(x), \forall x\in D$ is p.s.d., then f is convex over D

(2) If $\nabla^2 f(x), \forall x\in D$ is p.d., then f is strict convex over D

(3) If $D=\mathbb{R}^n$ and f is convex over $\mathbb{R}^n$, then $\nabla^2 f(x), \forall x\in D$ is p.s.d.

Proof

(1) $\forall x,y\in D$, by 1st order Taylor

$$f(y)=f(x)+(y-x)^T\nabla f(x)+\frac{1}{2}(y-x)^T\nabla^2 f\big(x+\alpha(y-x)\big)(y-x), 0\leq\alpha\leq 1$$

(2) Similar to (1) with $y\neq x$, strict inequality

(3) Suppose for contradiction that $\exists x,z\in\mathbb{R}^n$ such that $z^T\nabla^2 f(x)z<0$

Since $\nabla^2 f(x)$ is continuous, we can find a z small enough that $z^T\nabla^2 f(x+\alpha z)z<0, \forall 0\leq\alpha\leq 1$

By Taylor

$$\begin{aligned}f(x+z) &= x(x)+z^T\nabla f(x)+\frac{1}{2}z^T\nabla^2 f(x+\beta z)z, 0\leq\beta\leq 1\\ &< f(x)+z^T\nabla f(x)\end{aligned}$$

Which contradicts convexity

Chapter 4

Optimality Conditions

Definition 4.0.1 (Critical/Stationary Points)

All x such that $\nabla f(x) = 0$ are called critical or stationary points

All local minimizers are critical points, but the converse is not always true

Remark

$\nabla^2 f$ is symmetric since

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}$$

Theorem 4.0.1 (First Order Necessary Conditions For Optimality)

Let $f : \mathbb{R}^n \to \mathbb{R}$ be C^1-smooth. If x^* is a local minimizer, then $\nabla f(x^*) = 0$

Proof

Let $B_\delta(x^*)$ be such that $f(x^*) \leq f(x), \forall x \in B_\delta(x^*)$

$\forall i, \forall |h| < \delta, f(x^* + h \cdot e_i) - f(x^*) \geq 0$

Hence $\frac{f(x^*+h \cdot e_i) - f(x^*)}{h} \geq 0$ if $h > 0$

$\frac{f(x^*+h \cdot e_i) - f(x^*)}{h} \leq 0$ if $h < 0$

Since $f \in C^1$, then $\lim_{h \to 0} \frac{f(x^*+h \cdot e_i)}{h}$ exists

If both $\geq 0, \leq 0$ hold, then $= 0$ hold

Hence $\frac{\partial f}{\partial x_i}(x^*) = 0, \forall i$

Therefore $\nabla f(x^*) = 0$

Theorem 4.0.2 (Second Order Necessary Conditions For Local Optimality)

Let $f : \mathbb{R}^n \to \mathbb{R}$ be C^2-smooth. If x^* is a local minimizer, then $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ is p.s.d.

Proof

Let $z \in \mathbb{R}^n \backslash \{0\}$, we need to prove $z^T \nabla^2 f(x^*) z \geq 0$

Let $B_\delta(x^*)$ be such that $f(x^*) \leq f(x), \forall x \in B_\delta(x^*)$

Let $y := h \cdot \frac{z}{||z||}$ with $0 < h < \delta$, then we have

$$f(x^*+y) - f(x^*) \geq 0$$
$$f(x^*) + y^T\nabla f(x^*) + \frac{1}{2}y^T\nabla^2 f(x^*)y + \phi(x) - f(x^*) \geq 0 \text{ where } \lim_{y\to 0, y\neq 0}\frac{\phi(y)}{||y||} = 0$$

By 1st order condition, we have $y^T\nabla f(x^*) = 0$, hence we have

$$\frac{1}{2}\frac{h^2}{||z||^2}z^T\nabla^2 f(x^*)z + \phi(h\frac{z}{||z||}) \geq 0$$
$$z^T\nabla^2 f(x^*)z + 2||z||^2\frac{1}{h^2}\phi(h\frac{z}{||z||}) \geq 0$$

Take the limit when $h \to 0$, by Talor, we have

$$\lim_{h\to 0, h\neq 0}\frac{\phi(h\cdot\frac{z}{||z||})}{h^2} = 0$$

Therefore we have $z^T\nabla^2 f(x^*)z \geq 0$

Theorem 4.0.3 (Second Order Sufficient Conditions For Local Optimality)

Let $f : \mathbb{R}^n \to \mathbb{R} \in C^2\big(B_\delta(x^*)\big), x^* \in \mathbb{R}, \delta > 0$. If $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ is p.d., then x^* is a strict local minimizer

Proof

By Talor 2nd order, $\forall h \in B_\delta(x^*)$

$$f(x^*+h) = f(x^*) + h^T\nabla f(x^*) + \tfrac{1}{2}h^T\nabla^2 f(x^*)h + \phi(x) \text{ where } \lim_{h\to 0, h\neq 0}\tfrac{\phi(h)}{||h||} = 0$$

Let $0 < \lambda_1 < \cdots < \lambda_n$ be the positive eigenvalues of $\nabla^2 f(x^*)$

By the definition of limit

$$\exists r > 0 : \forall h \in B_r(x^*), |\frac{\phi(x)}{||h||^2}| \leq \frac{\lambda_1}{4} \iff |\phi(x)| \leq ||h||^2\frac{\lambda_1}{4}$$

Remember that,

$$||y||^2\cdot\lambda_1 \leq y^T\nabla^2 f(x^*)y \leq ||y||^2\cdot\lambda_n$$

Also by assumption, $\nabla f(x^*) = 0$, then we have

$$\begin{aligned} f(x^*+h) &= f(x^*) + \frac{1}{2}h^T\nabla^2 f(x^*)h + \phi(h) \\ &\geq f(x^*) + \frac{1}{2}||h||^2\lambda_1 - ||h||^2\frac{\lambda_1}{4} \\ &= f(x^*) + \frac{1}{4}||h||^2\cdot\lambda_1 \\ &> f(x^*) \text{ for all } h \in B_r(x^*)\backslash\{0\} \end{aligned}$$

Therefore x^* is a strict local minimizer over $B_r(x^*)$

4.1 Summary For Necessary And Sufficient Optimality Conditions

$$\begin{cases} \nabla f(x^*) = 0 \\ \nabla^2 f(x^*) \ p.d. \end{cases}$$

$\xrightarrow{(1)}$

$$x^* \text{ is a strict local minimizer}$$

$\xrightarrow{(2)}$

$$x^* \text{ is a local minimizer}$$

$\xrightarrow{(3)}$

$$\begin{cases} \nabla f(x^*) = 0 \\ \nabla^2 f(x^*) \ \ p.s.d. \end{cases}$$

The converses of (1), (2), (3) are all false!!!

Counterexamples:

(1) $f(x) = x^4$ at $x^* = 0$

(2) $f(x) = 1$ at $x^* = 0$

(3) $f(x) = x^3$ at $x^* = 0$

Theorem 4.1.1
Let $C \subset \mathbb{R}^n$ be a convex set, and $f : C \to \mathbb{R}$ be a convex function. A local minimizer of f is also a global minimizer. If f is strictly convex, then there is at most one global minimizer

Proof
Suppose x^* is a local minimizer, and y^* is a global minimizer with $f(y^*) \le f(x^*)$

By convexity of f, have

$$\begin{aligned} f\big(\alpha y^* + (1-\alpha)x^*\big) &\le \alpha \cdot f(y^*) + (1-\alpha) \cdot f(x^*) \\ &= f(x^*) + \alpha \cdot \big(f(y^*) - f(x^*)\big) \\ &< f(x^*), \forall 0 \le \alpha \le 1 \end{aligned}$$

Thus, $\forall r > 0, \exists z \ne x^*$ such that $||z - x^*|| < r$ and $f(z) < f(x^*)$

For instance, $z = \alpha \cdot y^* + (1-\alpha) \cdot x^*$ with $\alpha = \frac{r}{2||y^* - x^*||}$

Thus x^* is not a local minimizer, which is a contradiction

Therefore $f(y^*) \ge f(x^*)$

4.2 P.S.D

Theorem 4.2.1 (Spectral Decomposition of Symmetric P.S.D.)
$\forall A \in \mathbb{R}^{n \times n}$ symmetric, $\exists D, Q \in \mathbb{R}^{n \times n}$ such that

(1) D is Diagonal, its diagonal entries are eigenvalues of A

(2) Q is orthogonal, i.e. $Q^{-1} = Q^T$

(3) $A = QDQ^T$

Proof
Let $\lambda_1, \cdots, \lambda_n$ be the eigenvalues of A and $x_1, \cdots, x_n$ be their corresponding eigenvectors

Then $\forall i = 1, \cdots, n$, $Ax_i = \lambda_i x_i$, thus

$$A \cdot \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}^T = \begin{bmatrix} \lambda_1 x_1 \\ \lambda_2 x_2 \\ \vdots \\ \lambda_n x_n \end{bmatrix}^T = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}^T \cdot \begin{bmatrix} \lambda_1 & 0 & 0 & \dots & 0 \\ 0 & \lambda_2 & 0 & \dots & 0 \\ 0 & 0 & \lambda_3 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \lambda_n \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}^T \cdot diag(\lambda_1, \cdots, \lambda_n)$$

As A is symmetric, these x_i's are mutually orthogonal, i.e. $x_i x_j = 0$ when $i \neq j$

WLOG, assume $||x_i|| = 1$

Let $Q = \begin{bmatrix} x_1 & x_2 & \dots & x_n \end{bmatrix}$, then $Q^T Q = I$

Let $D = diag(\lambda_1, \cdots, \lambda_n)$

We get $AQ = QD$, thus $A = QDQ^{-1} = QDQ^T$

Theorem 4.2.2 (Cholestic Decomposition)

Let $A \in \mathbb{R}^{n \times n}$ be symmetric, then

A is p.s.d. $\iff \exists G \in \mathbb{R}^{n \times n}$ such that $A = GG^T$

Proof
($\Longrightarrow$) Assume $A = QDQ^T$ by the previous thm, where $Q^T = Q^{-1}$ and D is diagonal

Denote $\sqrt{D} = diag(\sqrt{D_{11}}, \cdots, \sqrt{D_{nn}})$

Let $G = Q \cdot \sqrt{D}$

Then $GG^T = Q\sqrt{D}(Q\sqrt{D})^T = Q\sqrt{D}\sqrt{D}^T Q^T = QDQ^T = A$

($\Longleftarrow$) Assume $A = GG^T$

Note that $\forall M \in \mathbb{R}^{n \times n}, M^T M$ is p.s.d.

Let $M = G^T$

Observations :

(1) If $\sqrt{D} = \begin{bmatrix} d & 0 \\ 0 & 0 \end{bmatrix}$, then $G = Q\sqrt{D} = \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix} \cdot \begin{bmatrix} d & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} Q_{11}d & 0 \\ Q_{21}d & 0 \end{bmatrix}$

So $\bar{G} = \begin{bmatrix} Q_{11}d \\ Q_{21}d \end{bmatrix}$ satisfies $\bar{G}\bar{G}^T = A$

(2) If A is p.d., then $\sqrt{D}$ is invertible

Since Q is always invertible, we get $G + Q\sqrt{D} \in \mathbb{R}^{n \times n}$ is also invertible

Definition 4.2.1 (Bounded Set & Closed Set & Compact Set)
(1) A set $S \subset \mathbb{R}^n$ is bounded if $S \subset B_\delta(0)$ for some δ finite

(2) A set $S \subset \mathbb{R}^n$ is closed if for any sequence $x_1, x_2, \cdots \in S$ such that $\lim_{i\to\infty} x_i$ exists, then $\lim_{i\to\infty} x_i \in S$

(3) A set is compact if it is bounded and closed

Theorem 4.2.3 (Existence of A Global Minimizer)
If $S \subset \mathbb{R}^n$ is nonempty and compact and $f : S \to \mathbb{R}$ is continuous, then $\exists y, z \in S$ such that $f(y) \le f(x) \le f(z), \forall x \in S$

Theorem 4.2.4 (Continuous Leads To Closed Level Set)
If f is continuous, then its level sets are closed

Proof
Let $S = \{x \in \mathbb{R}^n : f(x) \le \alpha\}$ be any level set

For any sequence $x_1, x_2, \cdots \in S$, we have $f(x_i) \le \alpha$. then by the continuity of f

$$f(\lim_{i\to\infty} x_i) = \lim_{i\to\infty} f(x_i) \le \alpha$$

Thus $\lim_{i\to\infty} x_i \in S$

Theorem 4.2.5 (Continuous And Bounded Level Set Gives Global Minimizer)
If $f : \mathbb{R}^n \to \mathbb{R}$ is continuous and has at least one bounded nonempty level set, then f has a global minimizer

Proof
Let α be such that $S = \{x \in \mathbb{R}^n : f(x) \le \alpha\}$ is bounded nonempty

By Thm 4.2.4, S is closed, thus compact

By Thm 4.2.3, f has a "global" minimizer over S :

$$\exists y \in S : f(y) \le f(x), \forall x \in S$$

Consider all points $x \in \mathbb{R}^n \backslash \{S\}$, we have $f(x) > \alpha \ge f(y)$

Thus $f(y) \le f(x), \forall x \in \mathbb{R}^n$

Example 4.2.1 (Functions without global minimizers)
Want to show each level set is either unbounded or empty

(1) $f(x) = 2x$, pick any α, see that the level set is unbounded

(2) $f(x) = e^x$, if $\alpha = 2$, then $S = \{x \in \mathbb{R} : x \le \ln 2\}$, if $\alpha < 0$, then $S = \emptyset$

Definition 4.2.2 (Coercive Function)
A function $f : \mathbb{R}^n \to \mathbb{R}$ is coercive if all its level sets are bounded

Note : If f is coercive, unless $f(x) = \pm\infty, \forall x$, then it has a global minimizer.

Theorem 4.2.6 (Equivalence of Coercive)
Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuous function, TFAE:

(1) f is coercive

(2) $\forall r \in \mathbb{R}, \exists\, m > 0$ such that $||x|| \geq m \Rightarrow f(x) \geq r$ (If we want f above r, x has to be $m-$far away from origin

Proof
(2) $\Rightarrow$ (1) Consider $S = \{x \in \mathbb{R}^n : f(x) \leq \alpha\}$

By (2), let $r = \alpha + 1$, we have

$$\exists\, m > 0 \,:\, ||x|| \geq m \Rightarrow f(x) \geq \alpha + 1$$

So $S \subset B_m(0)$, i.e. S is bounded. The reasoning holds for all α

(1) $\Rightarrow$ (2) For any given r, consider $T = \{x \in \mathbb{R}^n : f(x) \leq r\}$ bounded by assumption

Hence $\exists\, \delta > 0$ such that $T \subset B_\delta(0)$

For all x such that $||x|| \geq \delta + 1$, must have $x \notin T$, thus $f(x) > r$

Letting $m = \delta + 1$ and we are done

Example 4.2.2
Let $A \in \mathbb{R}^{m\times n}$ be of rank n, then $f(x) = ||Ax - b||$ with $b \in \mathbb{R}^m$ is coercive

Trick $f(x) = ||Ax - b|| \geq ||Ax|| - ||b||$ by the triangle inequality

Note $A^T A$ is P.S.D, in fact, p.d. because A is full rank (Proposition 2.4.1)

$$\begin{aligned} f(x) \geq ||Ax|| - ||b|| &= \sqrt{(Ax)^T(Ax)} - ||b|| \\ &= \sqrt{x^T(A^TA)x} - ||b|| \\ &\geq \sqrt{\lambda_1 ||x||^2} - ||b|| \text{ by Proposition 2.3.3} \\ &\geq \sqrt{\lambda_1}||x|| - b \end{aligned}$$

So given any $r > 0$, we have $f(x) \geq r$ when ever $||x|| \geq \frac{r+||b||}{\sqrt{\lambda_1}}$

Chapter 5

Unconstrained Quadratic Optimization

5.1 Quadratic Functions

Definition 5.1.1 (Quadratic Function)

A quadratic function takes the form $q(x) = x^T Ax + b^T x + c$ for any $A \in \mathbb{R}^{n\times n}, b \in \mathbb{R}^n, c \in \mathbb{R}$, where $x^T Ax = \sum_{i,j} A_{ij}x_i x_j$

We can assume WLOG that A is symmetric.

Theorem 5.1.1 (Generalization When A is not symmetric in the quadratic form)

Let $A \in \mathbb{R}^{n\times n}$ and let G be the symmetric part of A, i.e. $G := (A + A^T)/2$. Then

(1) G is symmetric and (2) $q(x) = x^T Gx + bx + c, \forall x \in \mathbb{R}^n$

Proof

(1) Let's compute the transpose of G

$$G^T = (\frac{A + A^T}{2})^T = \frac{1}{2}(A + A^T) = G$$

(2) Observe that $x^T Ax$ is scalar so $(x^T Ax)^T = x^T A^T x$, then

$$x^T Ax = \frac{x^T Ax}{2} + \frac{x^T Ax}{2} = \frac{x^T Ax}{2} + \frac{x^T A^T x}{2} = \frac{1}{2}x^T(A + A^T)x = x^T Gx$$

Definition 5.1.2 (Range & Kernel)

The range (or column space) of $A \in \mathbb{R}^{m\times n}$ is $Range(A) = \{Ax : x \in \mathbb{R}^n\}$

The kernel (or null space)of $A \in \mathbb{R}^{m\times n}$ is $Null(A) = \{x : Ax = 0, x \in \mathbb{R}^n\}$

Theorem 5.1.2 (Relation of Range and Null)

Let $C \in \mathbb{R}^{m\times n}$. If $y \in Range(C^T)$ and $z \in Null(C)$, then $y^T z = 0$

Proof

Since $y \in Range(C^T)$, then $\exists x \in \mathbb{R}^m$ such that $y = C^T x$, hence

$$y^T z = (C^T x)^T z = x^T \underbrace{Cz}_{0 \text{ since } z\in Null(C)} = 0$$

Theorem 5.1.3 (Decomposition of any vector (follows the fundamental thm of linear algebra))
Let $C \in \mathbb{R}^{m\times n}$ For any $\omega \in \mathbb{R}^n$, there exists $y \in Range(C^T)$ and $z \in Null(C)$ unique such that $\omega = y + z$

Proof
Let $\omega = y + z + b$, where $y \in Range(C^T), z \in Null(C), b \in Range(C^T)^\perp \cap Null(C)^\perp$

The decomposition is unique since $Range(C^T) \perp Null(C) \perp b$

Consider $Cb \in \mathbb{R}^m$, then $C^T(Cb) \in Range(C^T)$

Hence $b \perp C^T(Cb)$ as $b \in Range(C^T)^\perp$

Thus $0 = b^T\big(C^T(Cb)\big) = (Cb)^T(Cb) = ||Cb||_2^2$

So $Cb = 0$, then $b \in Null(C)$

We get $b \in Null(C) \cap Null(C)^\perp$, therefore $b = 0$

Derivative of $q(x)$

(1) $\frac{\partial}{\partial x_k} b^T x = b_k, \nabla b^T x = b$

(2) $\frac{\partial}{\partial x_k} x^T A x = \frac{\partial}{\partial x_k}\sum_{i,j} A_{ij}x_i x_j = \frac{\partial}{\partial x_k}(\sum_{j\neq k} A_{kj}x_k x_j + \sum_{i\neq k} A_{ik}x_i x_k + A_{kk}x_k^2)$

$$\begin{aligned}
\text{(as } A \text{ is symmetric)} &= \tfrac{\partial}{\partial x_k}(\textstyle\sum_{j\neq k} A_{kj}x_k x_j + \sum_{i\neq k} A_{ki}x_i x_k + A_{kk}x_k^2)\\
&= \tfrac{\partial}{\partial x_k}(2\textstyle\sum_{j\neq k} A_{kj}x_k x_j + A_{kk}x_k^2)\\
&= 2\textstyle\sum_{j\neq k} A_{kj}x_j + 2A_{kk}x_k = 2\sum_j A_{kj}x_j\\
&= k\text{th row of } 2Ax
\end{aligned}$$

$\nabla x^T A x = 2Ax$

(3) $\nabla^2 b^T x = 0$

(4) $\frac{\partial^2}{\partial x_k \partial x_l} x^T A x = \frac{\partial}{\partial x_l}(2\sum_j A_{kj}x_j) = 2A_{kl}$, $\nabla^2 x^T A x = 2A$

Theorem 5.1.4
Given $A \in \mathbb{R}^{n\times n}$ be symmetric, $b \in \mathbb{R}^n, c \in \mathbb{R}$, let $q(x) = x^T A x + bx + c$

(1) If A is p.d., then $q(x)$ has a unique global minimizer $x^* = -\frac{1}{2}A^{-1}b$

(2) If A is p.s.d. and $b \in Range(A)$, then $q(x)$ has a global minimizer

(3) Otherwise, $q(x)$ has no global minimizer, i.e. $q(x) \to -\infty$ for some $||x|| \to +\infty$

Proof
Necessary Conditions:

$$x^* \text{ local minimizer} \implies \begin{cases}\nabla q(x^*) = 0\\ \nabla^2 q(x^*)\ p.s.d.\end{cases} \implies \begin{cases}2Ax^* + b = 0\\ 2A, i.e. A \quad p.s.d.\end{cases}$$

(1) Assume A is p.d., then all eigenvalues > 0, thus A^{-1} exists

There is a unique critical point (i.e. point where $\nabla q = 0$) $x^* = -\frac{1}{2}A^{-1}b$

It is a local minimizer since $\nabla^2 q(x^*) = A$ is p.d. (see sufficient conditions)

Note that for any $h \in \mathbb{R}^n$

$$x^{*T}Ah = (x^{*T}Ah)^T = h^TA^Tx^* = h^TAx^* \tag{5.1}$$

Hence we have

$$\begin{aligned} q(x^*+h) &= (x^*+h)^TA(x^*+h) + b^T(x^*+h) + c \\ &= x^{*T}Ax^* + x^{*T}Ah + h^TAx^* + h^TAh + b^Tx^* + b^Th + c \\ &= (x^{*T}Ax^* + b^Tx^* + c) + (x^{*T}Ah + h^TAx^*) + h^TAh + b^Th \\ &= q(x^*) + 2h^TAx^* + h^TAh + b^Th \text{ by (5.1)} \\ &= q(x^*) + 2h^TA(-\frac{1}{2}A^{-1}b) + h^TAh + b^Th \\ &= q(x^*) - h^Tb + h^TAh + b^Th \\ &= q(x^*) + h^TAh \\ &\geq q(x^*) \end{aligned}$$

(2) $b \in Range(A) \implies -\frac{1}{2}b \in Range(A)$, so $Ax^* = -\frac{1}{2}b$ for some x^*

Hence x^* satisfies $\nabla g(x^*) = 2Ax^* + b = 0$

Then same proof as (1), $q(x^*+h) \geq q(x^*), \forall h \in \mathbb{R}^n$

(3.1) Assume A is p.s.d but $b \notin Range(A)$

We try to find a direction z that q goes to $-\infty$

Write $b = y + z$ uniquely with $y \in Range(A^T) = Range(A), z \in Null(A), z \neq 0$ since $b \notin Range(A)$

For any $\lambda \in \mathbb{R}$

$$\begin{aligned} q(\lambda z) &= \lambda^2 z^T \underbrace{Az}_{=0} + \lambda b^T z + c \\ &= \lambda (y+z)^T z + c \\ &= \lambda \underbrace{y^T z}_{=0} + \lambda z^T z + c \\ &= \lambda \underbrace{||z||_2^2}_{>0 \text{ since } z\neq 0} + c \end{aligned}$$

For $\lambda \to -\infty$, we get $q(\lambda z) \to -\infty$

(3.2) Assume A is not p.s.d., then $\exists v \in \mathbb{R}^n, v^TAv < 0$

Still we want to find a direction

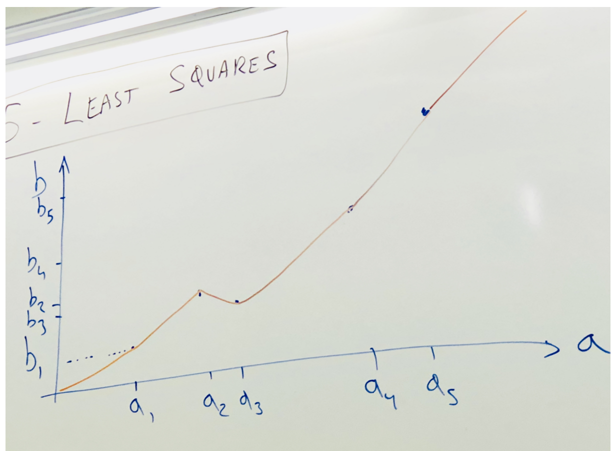
Let $\omega \in \mathbb{R}^n$ with $\omega = \begin{cases} v & \text{if } b^Tv \geq 0 \\ -v & \text{if } b^Tv < 0 \end{cases}$, we have $\omega^TA\omega < 0$ and $b^T\omega \geq 0$

For any $\lambda \in \mathbb{R}, q(\lambda\omega) = \lambda^2 \underbrace{\omega^T A\omega}_{<0} + \lambda \underbrace{b^T\omega}_{\geq 0} + c$

Take $\lambda \to -\infty$, we get $q(\lambda\omega) \to -\infty$

Chapter 6

Least Squares Problem



Given $a_1, \cdots, a_m \in \mathbb{R}^k, b_1, \cdots, b_m \in \mathbb{R}$, find a function $h : \mathbb{R}^k \to \mathbb{R}$ such that $h(a_i) \approx b_i, \forall i$

Least Squares : Minimize $\sum_i \big(h(a_i) - b_i\big)^2$

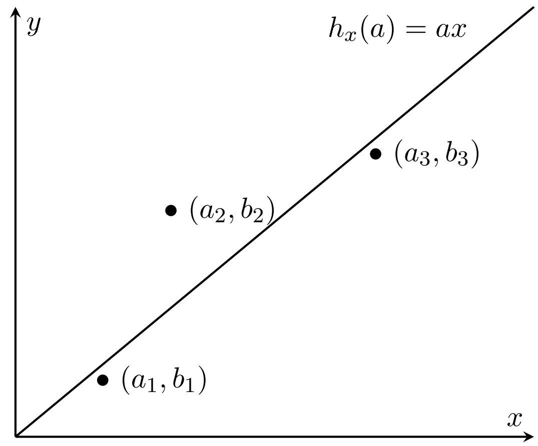
Goal : Determine the best h among a family of functions, parametrized by $x \in \mathbb{R}^n$:

$$\min_{x \in \mathbb{R}^n} \sum_i \big(h_x(a_i) - b_i\big)^2$$

Let $f(x) = \sum_i \big(h_x(a_i) - b_i\big)^2, \min_{x \in \mathbb{R}^n} f(x)$

6.1 Linear Least Squares

$h_x(a_i) = x_1 a_{i1} + x_2 a_{i2} + \cdots + x_k a_{ik} = a_i^T x$ in 1 dimension :



Note : How to get a hyperplane (or line) that does not contain the origin?

Let $n = k+1, a_{i,k+1} = 1, \forall i$, then $h_x(a_i) = x_1 a_{i1} + \cdots + x_k a_{ik} + x_{k+1}$

$$\begin{aligned} f(x) &= \sum_i (a_i^T x - b_i)^2 = (Ax-b)^2 = (Ax-b)^T(Ax-b) = ||Ax-b||_2^2 \\ &= x^T A^T Ax - x^T A^T b - b^T Ax + b^T b \\ &= x^T(A^T A)x - (2A^T b)^T x + b^T b \end{aligned}$$

Thus $f(x)$ is a quadratic function

If $rank(A) = n$, we have seen that $||Ax - b||_2$ is coercive, so it has a global minimizer.

If $rank(A) = n$ and $A^T A$ is p.d., then $f(x)$ has a global minimizer

$$x^* = -\tfrac{1}{2}(A^T A)^{-1}(-2A^T b) = (A^T A)^{-1} A^T b$$

6.2 Nonlinear Least Squares

Let $g : \mathbb{R}^n \to \mathbb{R}^m$ with $g_i(x) = h_x(a_i) - b$, we have $f(x) = \sum_i \big(g_i(x)\big)^2 = g(x)^T g(x)$

Definition 6.2.1 (Jacobian Matrix)
The Jacobian matrix of g is given by $J(x) = \begin{bmatrix} \nabla g_1(x)^T \\ \vdots \\ \nabla g_m(x)^T \end{bmatrix} = \begin{bmatrix} \frac{\partial}{\partial x_1} g_1(x) & \cdots & \frac{\partial}{\partial x_n} g_1(x) \\ \vdots & \ddots & \vdots \\ \frac{\partial}{\partial x_1} g_m(x) & \cdots & \frac{\partial}{\partial x_n} g_m(x) \end{bmatrix}$

$$\begin{aligned} \frac{\partial}{\partial x_k} f(x) &= \frac{\partial}{\partial x_k} \sum_i \big(g_i(x)\big)^2 \\ &= \sum_i 2 g_i(x) \frac{\partial}{\partial x_k} g_i(x) \\ &= 2 e_k^T J(x)^T g_i(x) \end{aligned}$$

Thus $\nabla f(x) = 2J(x)^T g(x)$ (think about Chain Rule)

Remark
If $g_i(x^*) = 0$, then $\nabla f(x^*) = 0$ and x^* is a global minimizer

$$
\begin{aligned}
\frac{\partial^2}{\partial x_k \partial x_l} g(x) &= \frac{\partial}{\partial x_l} 2\big(\sum_i g_i(x) \frac{\partial}{\partial x_k} g_i(x)\big) \\
&= 2\sum_i \big(\frac{\partial}{\partial x_l} g_i(x) \frac{\partial}{\partial x_k} g_i(x) + g_i(x) \frac{\partial^2}{\partial x_k \partial x_l} g_i(x)\big) \\
\nabla^2 f(x) &= \big(2\sum_i \underbrace{g_i(x)\nabla^2 g_i(x)}_{\text{not necessary p.d.}}\big) + 2\underbrace{J(x)^T J(x)}_{\text{p.s.d}}
\end{aligned}
$$

Chapter 7

Descent Algorithms

General Framework

$$\begin{aligned} &\text{Choose } x^0 \in \mathbb{R}^n \\ &\quad \text{for } k = 0, 1, 2, \cdots \\ &\qquad \text{Choose a search direction } p^k \in \mathbb{R}^n \\ &\qquad \text{Choose a step length } \alpha^k > 0 \\ &\qquad \text{Let } x^{k+1} = x^k + \alpha^k p^k \end{aligned}$$

Remark
α^k is not α to the power k, same for p^k, x^k. Also the objective function $f(x^{k+1})$ should be much smaller than $f(x^k)$ and x^k converges as fast as possible

Steepest Descent $p^k = -\nabla f(x^k)$

Lemma 7.0.1 (From Limit to Bound)

Let $\lim_{\epsilon \to 0, \epsilon > 0} \frac{\phi(\epsilon h)}{\epsilon} = 0$ for any $K > 0$, there exists ϵ small enough such that $|\phi(\epsilon h)| \leq \epsilon K$

Proof
For any $K > 0$, there exists $\gamma > 0$ such that $|\frac{\phi(\epsilon h)}{\epsilon} - 0| \leq K, \forall 0 < \epsilon\gamma$, i.e.

$$|\phi(\epsilon h)| \leq \epsilon K, \forall 0 < \epsilon < \gamma$$

Thus ϵ sufficiently small is $\epsilon \leq \gamma$

Theorem 7.0.1

Let $f \in C^1\big(B_t(x^k)\big), t > 0$ and $\nabla f(x^k) \neq 0$. Consider the optimization problem, for some $0 < \epsilon < t, \min\{f(x^k + \epsilon p) : ||p||_2 = 1\}$. Let p^* be a minimizer, then $\lim_{\epsilon \to 0} p^*_\epsilon = -\frac{\nabla f(x^k)}{||\nabla f(x^k)||}$

Proof

Let $x = x^k, p = -\frac{\nabla f(x)}{||\nabla f(x)||_2}$, hence $\nabla f(x) = -p||\nabla f(x)||$

Let $u \in \mathbb{R}^n$ with $||u||_2 = 1, u \neq p$, so $||u-p|| > \delta > 0$, hence

$$
\begin{aligned}
(u-p)^T(u-p) &> \delta^2 \\
u^Tu - 2u^Tp + p^Tp &> \delta^2 \\
2 - 2u^Tp &> \delta^2 \\
u^Tp &< 1 - \frac{\delta^2}{2}
\end{aligned}
$$

First use Taylor to write $f(x+\epsilon u)$

$$
\begin{aligned}
f(x+\epsilon u) &= f(x) + \epsilon u^T \nabla f(x) + \phi(\epsilon u), \text{ with } \lim_{\epsilon \to 0} \frac{\phi(\epsilon u)}{\epsilon} = 0 \\
&= f(x) - \epsilon ||\nabla f(x)|| u^T p + \phi(\epsilon u) \\
&\geq f(x) - \epsilon ||\nabla f(x)|| (1 - \frac{\delta^2}{2}) + \phi(\epsilon u)
\end{aligned}
$$

Now we want to get rid of $\phi(\epsilon u)$. For ϵ small enough, by Lemma 7.0.1, we have

$$
\begin{aligned}
|\phi(\epsilon u)| &\leq \epsilon(||\nabla f(x)|| \frac{\delta^2}{4}) \\
\phi(\epsilon u) &\geq -\epsilon ||\nabla f(x)|| \frac{\delta^2}{4}
\end{aligned}
$$

Hence we have a lower bound

$$
\begin{aligned}
f(x+\epsilon u) &\geq f(x) - \epsilon ||\nabla f(x)|| (1 - \frac{\delta^2}{2}) - \epsilon ||\nabla f(x)|| \frac{\delta^2}{4} \\
&= f(x) - \epsilon ||\nabla f(x)|| + \epsilon \frac{\delta^2}{4} ||\nabla f(x)||
\end{aligned}
$$

Then use Taylor to write $f(x+\epsilon p)$

$$
\begin{aligned}
f(x+\epsilon p) &= f(x) + \epsilon p^T \nabla f(x) + \phi(\epsilon p), \text{ with } \lim_{\epsilon \to 0} \frac{\phi(\epsilon p)}{\epsilon} = 0 \\
&= f(x) - \epsilon ||\nabla f(x)|| p^T p + \phi(\epsilon p) \\
&= f(x) - \epsilon ||\nabla f(x)|| + \phi(\epsilon p)
\end{aligned}
$$

Again, for ϵ small enough, combined with the lower bound, we choose our magic upper bound

$$
\begin{aligned}
|\phi(\epsilon p)| &\leq \epsilon \frac{\delta^2}{5} ||\nabla f(x)|| \\
\phi(\epsilon p) &\leq \epsilon \frac{\delta^2}{5} ||\nabla f(x)||
\end{aligned}
$$

Hence we have a upper bound

$$
f(x+\epsilon p) \leq f(x) - \epsilon ||\nabla f(x)|| + \epsilon \frac{\delta^2}{5} ||\nabla f(x)||
$$

Using two bounds, we have

$$f(x+\epsilon p) \leq f(x) - \epsilon||\nabla f(x)|| + \epsilon\frac{\delta^2}{5}||\nabla f(x)|| \leq f(x) - \epsilon||\nabla f(x)|| + \epsilon\frac{\delta^2}{4}||\nabla f(x)|| \leq f(x+\epsilon u)$$

Therefore $f(x+\epsilon p)$ is the minimizer

Definition 7.0.1 (Descent Direction)
p^k is a descent direction if $f(x^k+\epsilon p^k) < f(x^k)$, for all ϵ small enough

Theorem 7.0.2
Let x^k be such that $\nabla f(x^k) \neq 0$, if $(p^k)^T\nabla f(x^k) < 0$, then p^k is a descent direction

Proof
Let $p = p^k$, WLOG, $||p|| = 1$, by Taylor, have

$$f(x^k+\epsilon p) = f(x) + \epsilon p^T\nabla f(x^k) + \phi(\epsilon p), \text{ with } \lim_{\epsilon\to 0}\frac{\phi(\epsilon p)}{\epsilon} = 0$$

For ϵ small enough, $|\phi(\epsilon p)| \leq \epsilon|\frac{1}{2}p^T\nabla f(x^k)| = -\epsilon\frac{1}{2}p^T\nabla f(x^k)$

Hence we have

$$f(x^k+\epsilon p) \leq f(x^k) + \frac{1}{2}\epsilon p^T\nabla f(x) < f(x^k)$$

7.1 Line Search

Once p^k is chosen, determine α^k such that $x^{k+1} = x^k + \alpha^k + p^k$

• Exact Line Search : $\alpha^k = argmin_{\alpha\geq 0}\{f(x^k+\alpha p^k)\}$

We define $\psi(\alpha) = f(x^k+\alpha p^k)$

Note :

$\psi(0) = f(x^k)$ and once α^k is chosen, $\psi(\alpha^k) = f(x^{k+1})$

$\psi'(\alpha) = \frac{d}{d\alpha}\psi(\alpha) = \nabla f(x^k+\alpha p^k)^T p^k$ (directional derivative)

$\psi'(0) = \nabla f(x^k)^T p^k < 0$ since we assume p^k is a descent direction

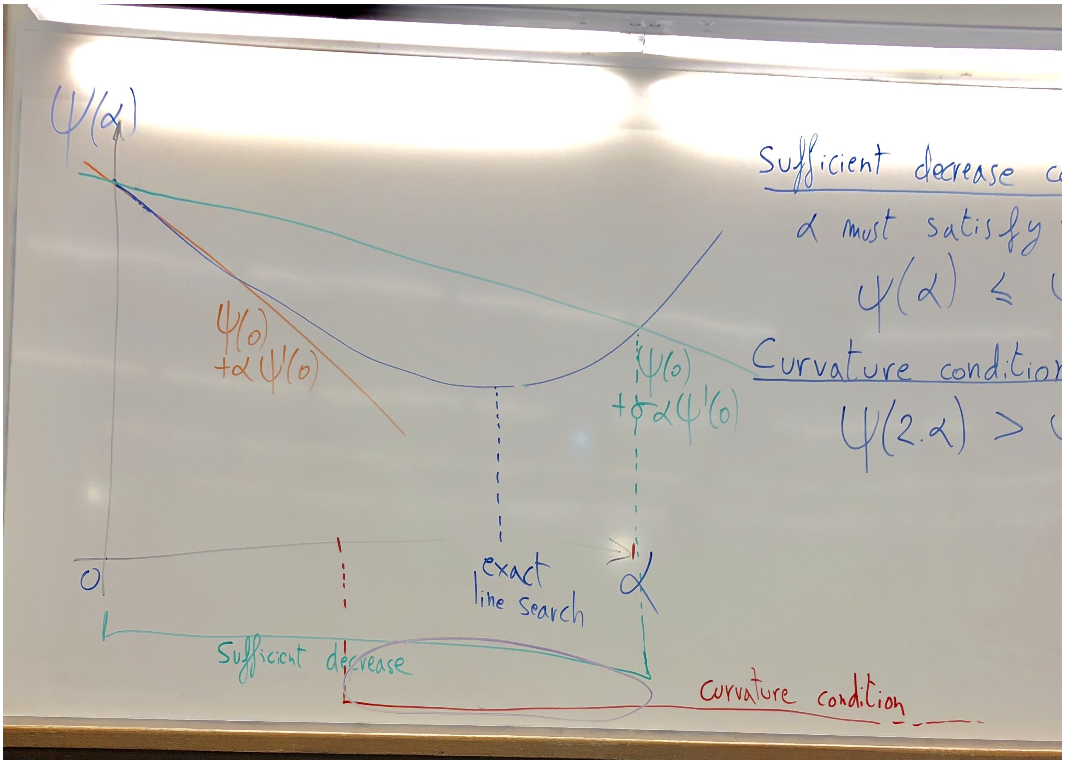
Sufficient Decrease Condition : Fix $0 < \sigma < \frac{1}{2}$, α must satisfy that :

$$\psi(\alpha) \leq \psi(0) + \sigma\alpha\psi'(0)$$

Curvature Condition (Scveral Variants)

$$\psi(2\cdot\alpha) > \psi(0) + \sigma 2\alpha\psi'(0)$$

Armijo ("backtrack") Inexact Linea Search



Let $\alpha := 1$

If α fails sufficient decrease

While α fails sufficient decrease

$\alpha := \alpha/2$

Else α fails curvature condition

While α fails curvature condition

$\alpha := \alpha \cdot 2$

Theorem 7.1.1

Let $f \in C^1, \nabla f(x^k) \neq 0$ and let p^k be a descent direction, then

Either the Armijo Algorithm terminates and α satisfies both conditions

Or $\alpha \to +\infty$ and f is unbounded below ($f(x^k) \to -\infty$)

Proof

• If the first loop terminates, α satisfies sufficient decrease and 2α fails it, i.e. α satisfies curvature condition

We need to show that the first loop terminates:

$$\psi(\alpha) = \psi(0) + \alpha \cdot \psi'(0) + \phi(\alpha)$$

For α sufficient small, by Lemma 7.0.1, (note $\psi'(0) < 0$), have

$$|\phi(\alpha)| \leq \alpha|\frac{1}{2}\psi'(0)|$$
$$\phi(\alpha) \leq -\alpha\frac{1}{2}\psi'(0)$$

Thus $\psi(\alpha) \leq \psi(0) + \frac{1}{2}\psi'(0) \cdot \alpha$

• If the second loop terminates, α satisfies curvature condition, $\frac{\alpha}{2}$ fails it, i.e. α satisfies sufficient decrease

• If the second loop does not terminate,

$$\psi(2^j) \leq \psi(0) + 2^j\sigma \underbrace{\psi'(0)}_{<0}, \forall j \in \mathbb{Z}^+$$

Thus $\psi(2^j) \to -\infty$ for $j \to +\infty$

• If we did not go in either loop, $\alpha = 1$ satisfies both conditions

Definition 7.1.1 (Lipschitz Continuous)
A function $f : \mathbb{R}^n \to \mathbb{R}$ is Lipschitz continuous with constant L if $|f(y) - f(x)| \leq L \cdot ||y - x||, \forall x, y \in \mathbb{R}^n$

Note : Lipschitz continuous implies C^0 continuous, but the converse is not true.

Theorem 7.1.2
Let $f \in C^1\big(B_\delta(0)\big)$, f is Lipschitz continuous with constant L on $B_\delta(0)$ if and only if $||\nabla f(x)|| \leq L, \forall x \in B_\delta(0)$

Theorem 7.1.3 (Zoutendijk's Thm)
Let $f : \mathbb{R}^n \to \mathbb{R}$ with $f \in C^1(\mathbb{R}^n)$. If

(1) ∇f is Lipschitz continuous

(2) $\forall k$, p^k is a descent direction with $\nabla f(x^k)^T p^k \leq -\mu||\nabla f(x^k)||_2 \cdot ||p^k||_2$ for some $0 < \mu \leq 1$. This means if $\mu = 1$, then p^k would be the steepest direction (think of vector dot product, μ would be a cosine of an angle)

(3) $\forall k$, α^k satisfies both decrease and curvature condition

Then either (a) $\lim_{k\to\infty} f(x^k) = -\infty$, or (b) $\lim_{k\to\infty} \nabla f(x^k) = 0$

Proof
We will prove there is no other situation (c), i.e. if (b) does not happen, then (a) does

Note that (b) $\lim_{k\to\infty} \nabla f(x^k) = 0$ can be stated in the following way

$$\forall \epsilon > 0, \exists K \geq 0 : \forall k \geq K, ||\nabla f(x^k)|| < \epsilon$$

If instead (b) does not happen, i.e. $\lim_{k\to\infty} \nabla f(x^k)$ does not exists or not equal to 0, then

$$\exists \epsilon > 0, \forall K > 0, , \exists k \geq K : ||\nabla f(x^k)|| \geq \epsilon \tag{7.1}$$

In the rest of the proof, we will show that (7.1) implies $f(x^{k+1}) \leq f(x^k) - \delta$ for some constant $\delta > 0$, thus $f(x^k) \to -\infty$

We need to find a upper bound of $\psi(\alpha^k)$, note that $\psi'(0) < 0$, we must find a lower bound of α^k

Now consider for some $0 < \sigma < \frac{1}{2}$, the curvature condition gives:

$$\psi(2\alpha^k) > \psi(0) + 2\alpha^k \sigma \psi'(0)$$

And the mean value thm

$$\exists\, 0 \leq \gamma \leq 2\alpha^k \; \psi(2\alpha^k) = \psi(0) + 2\alpha^k \psi'(\gamma)$$

Together we have

$$\begin{aligned} \psi'(\gamma) &> \sigma \psi'(0) \\ \nabla f(x^k + \gamma p^k)^T p^k &> \sigma \nabla f(x^k)^T p^k \end{aligned} \tag{7.2}$$

Then consider Lipschitz gives

$$||\nabla f(x^k + \gamma p^k) - \nabla f(x^k)|| \leq L \cdot \gamma ||p^k||$$

And the CS inequality gives

$$\begin{aligned} [\nabla f(x^k + \gamma p^k) - \nabla f(x^k)]^T p^k &\leq ||\nabla f(x^k + \gamma p^k) - \nabla f(x^k)|| \cdot ||p^k|| \\ &\leq L \cdot \gamma ||p^k||^2 \\ \nabla f(x^k + \gamma p^k)^T p^k &\leq \nabla f(x^k)^T p^k + L \cdot \gamma ||p^k||^2 \end{aligned} \tag{7.3}$$

Combined (7.2) and (7.3) together, have

$$\begin{aligned} \nabla f(x^k)^T p^k + L \cdot \gamma ||p^k||^2 &> \sigma \nabla f(x^k)^T p^k \\ L \cdot \gamma ||p^k||^2 &> (\sigma - 1) \nabla f(x^k)^T p^k \\ L \cdot \gamma ||p^k||^2 &> (\sigma - 1) \psi'(0) \\ \gamma &> \frac{(1-\sigma)(-\psi'(0))}{L \cdot ||p^k||^2} \end{aligned}$$

Recall that $0 \leq \gamma \leq 2\alpha^k$, thus $\alpha^k \geq \gamma/2$, hence **finally we get an lower bound for** α^k

$$\alpha^k > \frac{(1-\sigma)(-\psi'(0))}{2L \cdot ||p^k||^2}$$

Now we show the sufficient decrease

$$\begin{aligned} \psi(\alpha^k) &\leq \psi(0) + \sigma \alpha^k \psi'(0) \\ &\leq \psi(0) + \sigma \frac{(1-\sigma)(-\psi'(0))}{2L \cdot ||p^k||^2} \psi'(0) \quad \text{note } \psi'(0) < 0 \\ &= \psi(0) - \frac{\sigma(1-\sigma)}{2L} \cdot \left(\frac{\psi'(0)}{||p^k||^2}\right)^2 \end{aligned}$$

By hypothesis

$$\big(\psi'(0)\big)^2 = \big(\nabla f(x^k)^T p^k\big)^2 \geq \mu^2 \cdot ||\nabla f(x^k)||^2 \cdot ||p^k||^2$$

Hence we have

$$\psi(\alpha^k) \leq \psi(0) - \frac{\sigma(1-\sigma)}{2L} \cdot \mu^2 \cdot ||\nabla f(x^k)||^2$$

By (7.1), $||\nabla f(x^k)|| \geq \epsilon$, so

$$\psi(\alpha^k) \leq \psi(0) - \frac{\sigma(1-\sigma)\mu^2\epsilon^2}{2L}$$
$$f(x^k + \alpha^k p^k) \leq f(x^k) - \frac{\sigma(1-\sigma)\mu^2\epsilon^2}{2L}$$

- We now have a complete algorithm:

  Start at an arbitrary $x^0 \in \mathbb{R}^n$

  For $k = 1, 2, \cdots$

  Choose p^k such that $\nabla f(x^k)^T p^k \leq -\mu \cdot ||\nabla f(x^k)|| \cdot ||p^k||$, for some $0 < \mu \leq 1$

  for example $p^k := -\nabla f(x^k)$, the steepest descent

  Choose α^k with Armijo inexact line search

  Let $x^{k+1} := x^k + \alpha^k p^k$

  If $(f(x^{k+1} < -M)$ or $(||\nabla f(x^{k+1})|| \leq \epsilon)$

  STOP

7.2 Convergence of Descent Algorithms

Definition 7.2.1 (Converge Degree)

A sequence $s^0, s^1, \cdots$ converges with degree d to 0 if $|s^{k+1}| \leq C \cdot |s^k|^d$. Convergence is said to be linear if $d = 1$ and quadratic if $d = 2$

Definition 7.2.2 (Strongly Convex)

A function $f \in C^1(\mathbb{R}^n)$ is strongly convex if $\big(\nabla f(y) - \nabla f(x)\big)^T (y - x) \geq l \cdot ||y - x||^2, \forall x, y \in \mathbb{R}^n$ for some $l > 0$

Lemma 7.2.1

If f is strongly convex, then $||\nabla f(y) - \nabla f(x)||^2 \geq l \cdot |f(y) - f(x)|, \forall x, ty \in \mathbb{R}^n$

Lemma 7.2.2 (Assignment 1 Q2)

Let $f \in C^2(\mathbb{R}^n)$, f is strongly convex if and only if $\big(\nabla^2 f(x) - l \cdot I\big)$ is p.s.d. for all $x \in \mathbb{R}^n$

Lemma 7.2.3

Any strongly convex function is strictly convex

Theorem 7.2.1

Assume the same condition as Zoutendijk's Thm, if in addition, f is strongly convex, then $f(x^k)$ converges linearly to a local(global) minimizer $f(x^*)$

Proof

From the previous proof we have

$$f(^{k+1}) \leq f(x^k) - \frac{\sigma(1-\sigma)\mu^2}{2L}||\nabla f(x^k)||^2$$
$$f(^{k+1}) - f(x^*) \leq f(x^k) - f(x^*) - \frac{\sigma(1-\sigma)\mu^2}{2L}||\nabla f(x^k)||^2$$

By Lemma 7.2.1, $||\nabla f(x^k) - \underbrace{\nabla f(x^*)}_{0}||^2 \geq l \cdot |f(x^k) - f(x^*)|$

I.e. we get an lower bound $||\nabla f(x^k)||^2 \geq l \cdot \big(f(x^k) - f(x^*)\big)$, hence

$$\begin{aligned} f(x^{k+1}) - f(x^*) &\leq f(x^k) - f(x^*) - \frac{\sigma(1-\sigma)\mu^2}{2L}|| \cdot l \cdot \big(f(x^k) - f(x^*)\big) \\ &\leq \big(f(x^k) - f(x^*)\big)(1 - \underbrace{\frac{\sigma(1-\sigma)\mu^2 l}{2L})}_{>0,<1} \end{aligned}$$

Thus the sequence $\big(f(x^k) - f(x^*)\big)$ converges linearly to 0

Theorem 7.2.2

For a strongly convex quadratic function, the steepest descent method with exact line search has $||x^k - x^*||$ converges linearly to 0. Also this bound is tight, i.e. cannot converge with $d > 1$

Thm 7.2.1 shows that in many cases, the sequence converges linearly and Thm 7.2.2 shows that not many can converge over linealy, this leads to the next section

7.3 Newton Step

Consider a quadratic approximation of f at x^k :

$$f(x^k + h) \approx q(h) = f(x^k) + h^T \nabla f(x^k) + \frac{1}{2} h^T \nabla^2 f(x) h$$

If(and only if) $\nabla f^2(x^k)$ is p.d., $q(h)$ has a unique minimizer $h = -[\nabla^2 f(x^k)]^{-1} \nabla f(x^k)$

The newton step is given by taking $p^k = -[\nabla^2 f(x^k)]^{-1} \nabla f(x^k)$, note clearly it only works if the $\nabla^2(f^k)$ is p.d.

Definition 7.3.1 (Linearly & Quadratic Convergence)

A sequence $s^0, s^1, \cdots$ converges linearly to zero if $|s^{k+1}| \leq C \cdot |s^k|$ for some $0 < C < 1$, the sequence converges quadratically if $|s^{k+1}| \leq C \cdot |s^k|^2$ for some $C > 0$

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

Lemma 7.3.1
Let $F : \mathbb{R}^n \to \mathbb{R}^{m\times m}$ be continuous over $B_r(x_0)$ for some $x_0 \in \mathbb{R}^n$ such that $F(x_0)$ is nonsingular. Then there exists $R > 0$ such that $F(x)$ is invertible for all $x \in B_R(x_0)$ and $F(x)^{-1}$ is continuous over $B_R(x_0)$

Proof
Given any matrix $A \in \mathbb{R}^{m\times m}$, $det(A)$ is a polynomial in all entries of A

Since $det\big(F(x_0)\big) \neq 0$, then there exists $R > 0$ such that $det\big(F(x)\big) \neq 0$ for all $x \in B_R(x_0)$

Still given $A \in \mathbb{R}^{m\times m}$, $(A^{-1})_{ij} = \frac{P}{det(A)}$ where P is a polynomial in entries of A

Thus $\big(F(x)^{-1}\big)$ is a polynomial in $F(x)$ divided by another nonzero polynomial, hence it is continuous

Theorem 7.3.1
Let $f : \mathbb{R}^n \to \mathbb{R}$ be such that $f \in C^2(\big(B_r(x^*)\big)$, if

(1) $\nabla^2 f$ is Lipschitz continuous over $B_r(x^*)$, i.e.

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

(2) $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ is p.d. (2nd order sufficient conditions for local optimality)

(3) $\nabla^2 f(x)^{-1}|| \leq 2||\nabla^2 f(x^*)^{-1}||$ for all $x \in B_r(x^*)$ (By lemma, there exists r sufficiently small such that this is satisfied)

(4) $r \leq \frac{1}{2L||\nabla^2 f(x^*)^{-1}||}$

Then Newton's Method converges quadratically to x^* if $x^0 \in B_r(x^*)$

Proof
Assume for induction that $x^k \in B_r(x^*)$, we will show that $x^{k+1} \in B_r(x^*)$

$$
\begin{aligned}
x^{k+1} - x^* &= x^k - \nabla^2 f(x^k)^{-1} \nabla f(x^k) - x^* \\
&= \nabla^2 f(x^k)^{-1} \cdot \big(\nabla^2 f(x^k)(x^k - x^*) - (\nabla f(x^k) - \overbrace{\nabla f(x^*)}^{=0})\big) \\
&= \nabla^2 f(x^k)^{-1} \cdot \Big(\int_0^1 \underbrace{\nabla^2 f(x^k)(x^k - x^*)dt}_{\text{does not vary with } t} - \int_0^1 \underbrace{\nabla^2 f(x^* + t(x^k - x^*))(x^k - x^*)dt}_{\substack{\text{directional derivative of } \nabla f\Big) \\ \text{in direction } (x^k - x^*) \\ \text{integrated from } x^* \text{ to } x^k}} \\
&= \underbrace{\nabla^2 f(x^k)^{-1}}_{(a)} \cdot \underbrace{\int_0^1 \big(\nabla^2 f(x^k) - \nabla^2 f(x^* + t(x^k - x^*))\big)(x^k - x^*)dt}_{(b)}
\end{aligned}
$$

$$
\begin{aligned}
&(a)\ : ||\nabla^2 f(x^k)^{-1}|| \leq 2 \cdot ||\nabla^2 f(x^*)^{-1}|| \text{ by } (3) \\
&(b)\ :\ || \int_0^1 \big(\nabla^2 f(x^k) - \nabla^2 f(x^* + t(x^k - x^*))\big)(x^k - x^*)dt|| \\
&\quad \leq \int_0^1 ||(\nabla^2 f(x^k) - \nabla^2 f(x^* + t(x^k - x^*))|| \cdot ||x^k - x^*||dt \\
&\quad \leq \int_0^1 L \cdot ||x^k - x^* - t(x^k - x^*)|| \cdot ||x^k - x^*||dt \\
&\quad \leq L \cdot \int_0^1 ||(x^k - x^*)(1-t)|| \cdot ||x^k - x^*||dt \\
&\quad = L \cdot ||x^k - x^*||^2 \int_0^1 (1-t)dt = \frac{L \cdot ||x^k - x^*||^2}{2}
\end{aligned}
$$

Therefore have

$$||x^{k+1} - x^*|| \leq 2 \cdot ||\nabla^2 f(x^*)^{-1}|| \cdot \frac{L \cdot ||x^k - x^*||^2}{2}$$

By induction, we have $||x^k - x^*|| \leq r$ and by (4) $r \leq \frac{1}{2L||\nabla^2 f(x^*)^{-1}||}$, hence have

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

So the convergence (if any) is quadratic

And since $||x^k - x^*|| \leq r$, we have $||x^{k+1} - x^*|| \leq \frac{1}{2}||x^k - x^*||$

Therefore we have the convergence

Chapter 8

Trust Region Methods

Algorithm

Choose x^0 arbitrarily

Let $\delta^0 = 1$

For $k = 0, 1, \cdots$

Let $q(x)$ be a quadratic approximation of f that is accurate around x^k

$x^{TEST} := argmin\{q(x) : ||x - x^k| \leq \delta^k\}$

$\rho := \frac{f(x^k)-f(x^{TEST})}{q(x^k)-q(x^{TEST})}$ // The ratio of decrease

If $\rho \geq 1/8$

$x^{k+1} = x^{TEST}$

Else $x^{k+1} = x^k$

If $\rho \leq 1/4$

$\delta^{k+1} = \delta^k/2$

Else if $\rho \geq 3/4$ and $||x^{TEST} - x^k|| = \delta^k$

$\delta^{k+1} = 2 \cdot \delta^k$

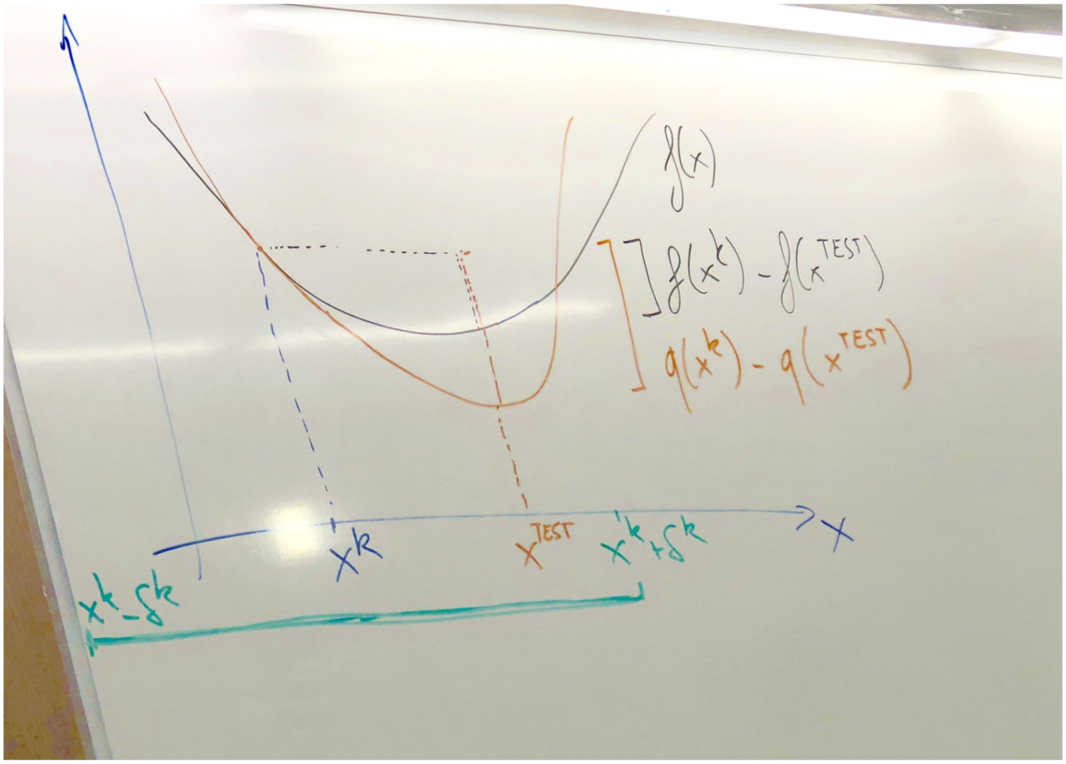
Else $\delta^{k+1} = \delta^k$

Note :

- There are other possible choices, but we consider

$$q(x) = f(x^k) + (x - x^k)^T \nabla f(x^k) + \frac{1}{2}(x - x^k)^T \nabla^2 f(x^k)(x - x^k)$$

- ρ is the ratio $\frac{\text{decrease in } f}{\text{decrease in } q}$ from x^k to x^{TEST}. The decrease in q is guaranteed ≥ 0 since $q(x^k)$ is considered in the agrmin set.
- If $x^{TEST} = x^k$, then the 2nd order sufficient conditions are satisfied. $\rightarrow$ STOP



- δ^k is the **Trust Region Radius**. We consider q is a "good" approximation of f in $B_{\delta^k}(x^k)$. If ρ is small, the approximation is bad, and we decrease δ^k

Theorem 8.0.1
Let $f \in C^2(\mathbb{R}^n)$ and assume that $\nabla^2 f$ is Lipschitz continuous in a ball that contains the level set of x^0. Then for the trust region method

(1) Either $x^k \to -\infty$ or $\nabla f(x^k) \to 0$ (similar as descent method)

(2) If $x^k \to x^*$, then x^* satisfies 1st and 2nd order necessary condition for local optimality

(3) If $x^k \to x^*$ and x^* satisfies the 1st and 2nd sufficient conditions for local optimality, then for k large enough, $||x^{TEST} - x^k|| \leq \delta^k$, the step is **Newton's Step**, so the convergence is quadratic.

8.1 The Trust Region Subproblem (TRS)

$$argmin\{ \overbrace{f(x^k)}^{constant} + (x - x^k)^T \nabla f(x^k) + \frac{1}{2}(x - x^k)^T \nabla^2 f(x^k)(x - x^k) \ : \ ||x - x^k|| \leq 1\}$$

For simplicity, let $\tilde{x} = \frac{x - x^k}{\delta^k}$, we get

$$agrmin\{\left(\delta^k \nabla f(x^k)\right)^T \cdot \tilde{x} + \tilde{x}^T\left((\delta^k)^2 \frac{1}{2} \nabla^2 f(x^k)\right)\tilde{x} \ : \ ||\tilde{x}|| \leq 1\}$$

$$= argmin\{x^T A x + b^T x : ||x|| \leq 1\} \text{ where } A = \frac{(\delta^k)^2}{2}\nabla^2 f(x^k), b = \delta^k \nabla f(x^k)$$

How do we solve (TRS)?

$$\begin{aligned} &min \;\; x^T Ax + b^T x \\ &s.t. \quad ||x| \leq 1 \end{aligned}$$

If A is p.d., then we can compute $\hat{x} = -\frac{1}{2}A^{-1}b$

CASE 1 A is p.d. and $||\hat{x}|| = || - \frac{1}{2}A^{-1}b|| \leq 1$. Then $\hat{x}$ is optimal for (TRS)

CASE 2 A is not p.d. or $||\hat{x}|| > 1$. Let $\hat{x}(\lambda) = -\frac{1}{2}(A + \lambda I)^{-1}b$

Note :

• $(A + \lambda I)$ shifts all the eigenvalue, so at some point, all the eigenvalue would be positive thus the inverse $(A + \lambda I)^{-1}$ is well-defined.

• Let $\lambda_1 \leq \cdots \leq \lambda_n$ be the eigenvalues of A, $\hat{x}$ is defined for all $\lambda > -\lambda_1$

• $\hat{x}(0)$ would be optimal in **CASE 1**

• $\hat{x}(\lambda)$ would be a global minimizer for $x^T(A + \lambda I)x + b^T x$

Theorem 8.1.1
$||\hat{x}(\lambda)||$ is a decreasing function of λ over $(-\lambda_1, -\infty)$. Moreover $\lim_{\lambda \to \infty} ||\hat{x}(\lambda)|| = 0$

Proof
Let $A = QDQ^T$ where Q is orthogonal and $D = diag(\lambda_1, \cdots, \lambda_n)$

Observe that for all $z \in \mathbb{R}^n$, $||Qz|| = ||z||$ since $||Qz||^2 = (Qz)^T(Qz) = z^T Q^T Qz = z^T z = ||z||^2$ or intuitively Qz is a rotation of z as Q is orthogonal. Thus

$$
\begin{aligned}
\hat{x}(\lambda) &= -\frac{1}{2}(QDQ^T + \lambda I)^{-1}b \\
&= -\frac{1}{2}(QDQ^T + \lambda QIQ^T)^{-1}b \\
&= -\frac{1}{2}[Q(D + \lambda I)Q^T]^{-1}b \\
&= -\frac{1}{2}Q^{T^{-1}}(D + \lambda I)^{-1}Q^{-1}b \\
&= -\frac{1}{2}Q(D + \lambda I)^{-1}Q^T b \\
||\hat{x}(\lambda)|| &= \frac{1}{2}||Q(D + \lambda I)^{-1}Q^T b|| \\
&= \frac{1}{2}||(D + \lambda I)^{-1}\underbrace{Q^T b}_{c}|| \quad \text{by } ||Qz|| = ||z|| \\
&= \frac{1}{2}||(D + \lambda I)^{-1}c|| \\
&= \frac{1}{2}||(\begin{bmatrix} \lambda_1 & 0 \cdots 0 & & \\ 0 & \lambda_2 & \cdots 0 & \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots \lambda_n & \end{bmatrix} + \lambda I)^{-1} \cdot c|| \\
&= \frac{1}{2}||\begin{bmatrix} \frac{1}{\lambda_1+\lambda} & 0 \cdots 0 & & \\ 0 & \frac{1}{\lambda_2+\lambda} & \cdots 0 & \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots \frac{1}{\lambda_n+\lambda} & \end{bmatrix} \cdot c|| \\
&= \frac{1}{2}||\begin{bmatrix} \frac{c_1}{\lambda_1+\lambda} \\ \frac{c_2}{\lambda_2+\lambda} \\ \vdots \\ \frac{c_n}{\lambda_n+\lambda} \end{bmatrix}|| \\
&= \frac{1}{2}\sqrt{\sum_i \underbrace{(\frac{1}{\lambda_i + \lambda})^2}_{decreasing\ for\ \lambda > -\lambda_i} \cdot \underbrace{c_i^2}_{constant\ \geq 0}} \\
\lim_{\lambda \to \infty} ||\hat{x}(\lambda)|| &= 0
\end{aligned}
$$

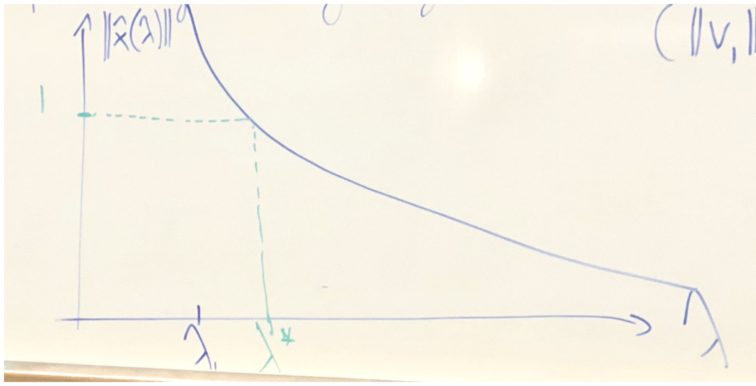
CASE 2a : $c_1 \neq 0, \lambda_1 \neq 0$ **Note :** $c_1 = (Q^T b)_1 = v_1^T b = v_1^T \nabla f(x^k)\delta^k$, where v_1 is the eigenvector of $\nabla^2 f(x^k)$ corresponding to λ_1 ($||v_1|| = 1$), so $c_1 \neq 0$ means $\nabla f(x^k)$ is not orthogonal to v_1

In this case, $\lim_{\lambda \to \lambda_1} ||\hat{x}(\lambda)|| = +\infty$

Then $\exists\, \lambda^*$ such that $||\hat{x}(\lambda^*)|| = 1$

Lemma 8.1.1

If $\hat{x}$ is a global minimizer for (TRS) in **CASE 2a**, then $||\hat{x}|| = 1$



Proof
If $||\hat{x}|| < 1$, then $\exists B_\delta(\hat{x}) \subseteq B_1(0)$

Since $\hat{x}$ is a global minimizer, it is also a local minimizer for $x^TAx + b^Tx$

If $\lambda_1 < 0$, A is not p.s.d. and x^TAx+b^Tx has no local minimizers. $\lambda_1 = 0$ excluded by hypothesis of **CASE 2a**

If $\lambda_1 > 0$, A is p.d. and $x^TAx + b^Tx$ has a unique local (and global) minimizer, but by **CASE 2a** hypothesis, $||\hat{x}|| > 1$

Theorem 8.1.2

$\hat{x}(\lambda^*)$ is a global minimizer for (TRS) in **CASE 2a**

Proof
Recall that $\hat{x}(\lambda^*)$ is a global minimizer for $x^T(A + \lambda^* I)x + b^Tx$

If we restrict to $||x|| = 1$, and we have $||\hat{x}(\lambda^*)|| = 1$

$$\begin{aligned}
\hat{x}(\lambda^*) &= argmin\{x^T(A + \lambda^* I)x + b^Tx \,:\, ||x|| = 1\} \\
&= argmin\{x^TAx + \lambda^* \underbrace{x^Tx}_{||x||^2=1} + b^Tx \,:\, ||x|| = 1\} \\
&= argmin\{x^TAx + b^Tx + \underbrace{\lambda^*}_{constant} \,:\, ||x|| = 1\} \\
&= argmin\{x^TAx + b^Tx \,:\, ||x|| = 1\}
\end{aligned}$$

By Lemma 8.1.1, $\hat{x}(\lambda^*) = argmin\{x^TAx + b^Tx \,:\, ||x|| \leq 1\}$

CASE 2b : either $c_1 = 0$ or $\lambda_1 = 0$

Theorem 8.1.3
A global minimizer for (TRS) in **CASE 2b** is given by

$$\hat{x} = \sum_{i:\lambda_i \neq \lambda_1} \frac{v_i^T b}{\lambda_i - \lambda_1} + \tau v_1$$

where v_i is the eigenvector of A corresponding to λ_1, $||v_i|| = 1$ and τ is chosen such that $||\hat{x}|| = 1$

Proof
Nocedal-Weight page 84 "the hard case"

Chapter 9

Optimality Conditions For Constrained Optimization

9.1 KKT Points

Definition 9.1.1 (Local Minimizer for Constrained OPT & Feasible Improving Direction)
Consider

$$\begin{aligned} &min \;\; f(x) \\ &s.t. \;\; x \in G \subseteq \mathbb{R}^n \end{aligned}$$

the point $\hat{x}$ is a local minimizer if $\hat{x} \in G$ and there exists $\epsilon > 0$ such that for all $x \in B_\epsilon(\hat{x}) \cap G$, we must have $f(x) \geq f(\hat{x})$

Note :

• The above definition does not require $\big(B_\epsilon(\hat{x}) \cap G\big) \backslash \{\hat{x}\} \neq \emptyset$, i.e. $\hat{x}$ could be the only point, in which case it is the local minimizer

• Equivalently, $\nexists\, d \in B_\epsilon(0) \,:\, \hat{x} + d \in G$ and $f(\hat{x} + d) < f(\hat{x})$. Such a d would be called a feasible improving direction (or step)

Informally, consider

$$\begin{aligned} &min \;\; f(x) \\ &s.t. \;\; h(x) = 0 \end{aligned}$$

Let $\bar{x} \in \mathbb{R}^n$ such that $h(\bar{x}) = 0$. Is there any improving direction d at $\bar{x}$?

If d is small, $h(\bar{x} + d) \approx h(\bar{x}) + d^T \nabla h(\bar{x}) = d^T \nabla h(\bar{x})$

• d "feasible" : we want $d^T \nabla h(\bar{x}) = 0$

• d "improving" : we want $d^T \nabla f(\bar{x}) < 0$

Take an arbitrary such vector $d \perp \nabla h(\bar{x})$.

If $d^T \nabla f(\bar{x}) < 0$, then we are done

If $d^T \nabla f(\bar{x}) > 0$, we can take $(-d)$: have $(-d)^T \nabla h(\bar{x}) = 0$ and $(-d)^T \nabla f(\bar{x}) < 0$

If $d^T \nabla f(\bar{x}) = 0$, we need another direction d

When are there **no** feasible improving directions?

When, for all $d \in \mathbb{R}^n$ such that $d^T \nabla h(\bar{x}) = 0$, we have $d^T \nabla f(\bar{x}) = 0$

When all directions orthogonal to $\nabla h(\bar{x})$ are also orthogonal to $\nabla f(\bar{x})$

I.e. when $\nabla h(\bar{x})$ is parallel to $\nabla f(\bar{x})$

Such $\bar{x}$ is called **Karush-Kuhn-Tucker (KKT) Point**

Example 9.1.1

$$\begin{aligned} min \;\; & x_1 + x_2 \\ s.t. \;\; & x_1^2 + x_2^2 - 2 = 0 \end{aligned}$$

where h is a convex function, and the feasible region is the circle of radius $\sqrt{2}$ (just the boundary not include the inside part), which is not convex as there is hole in it.

How can we change h such that the feasible region $h(x) = 0$ is also convex? We must need h is a linear function

Note that we want $h(x) = 0$ instead of $h(x) \leq 0$, so the thm about convex function and convex set does not work here.

$\nabla f(x) = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, $\nabla h(x) = \begin{bmatrix} 2x_1 \\ 2x_2 \end{bmatrix}$, KKT points : $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and $\begin{bmatrix} -1 \\ -1 \end{bmatrix}$ such that $x_1 = x_2$

Informally, consider

$$\begin{aligned} min \;\; & f(x) \\ s.t. \;\; & g(x) \leq 0 \end{aligned}$$

Let $\bar{x}$ be such that $g(\bar{x}) \leq 0$

CASE 1 : $g(\bar{x}) < 0$

For all $||d||$ sufficiently small (thus in the feasible region), $g(\bar{x} + d) < 0$

We want $d^T \nabla f(\bar{x}) < 0$, which exists iff $\nabla f(\bar{x}) \neq 0$

CASE 2 : $g(\bar{x}) = 0$

d "feasible" : $g(\bar{x} + d) \approx g(\bar{x}) + d^T \nabla g(\bar{x}) = d^T \nabla g(\bar{x})$, so want $d^T \nabla g(\bar{x}) \leq 0$

d "improving" : $d^T \nabla g(\bar{x}) < 0$

When are there **no** feasible improving directions?

CASE $g(\bar{x}) < 0$ **:** we want $\nabla f(\bar{x}) = 0$

CASE $g(\bar{x}) = 0$ **:** for all d such that $d^T \nabla g(\bar{x}) \leq 0$, we have $d^T \nabla f(\bar{x}) \geq 0$

Lemma 9.1.1

Let $a, b \in \mathbb{R}^n$, TFAE:

(1) for all $d \in \mathbb{R}^n$, $d^T a \leq 0 \implies d^T b \geq 0$ (think of vector multiplication with angle)

(2) $b = -\lambda a$ for some $\lambda \geq 0$

CASE $g(\bar{x}) < 0$ **:** we want $\nabla f(\bar{x}) = 0$

CASE $g(\bar{x}) = 0$ **:** we want $\nabla f(\bar{x}) = -\lambda \nabla g(\bar{x})$ for some $\lambda \geq 0$

$$\text{KKT points :} \begin{cases} \nabla f(\bar{x}) = -\lambda \nabla g(\bar{x}) \\ \lambda \geq 0 \\ \lambda \nabla g(\bar{x}) = 0 \end{cases}$$

Given $min\{f(x) : g(x) \leq 0\}$, KKT at y are :

CASE $g(y) < 0$: $\nabla f(y) = 0$

CASE $g(y) = 0$: $\nabla f(y) = -\lambda \nabla g(y)$ for some positive λ

Example 9.1.2

$$\begin{aligned} &min\ x_1 + x_2 \\ &s.t.\ x_1^2 + x_2^2 - 2 \leq 0 \end{aligned}$$

CASE 1 $x_1^2 + x_2^2 - 2 < 0$, $\nabla f(y) = [1, 1]^T = 0$ never holds

CASE 2 $x_1^2 + x_2^2 - 2 = 0$

$$\begin{aligned} \nabla f(y) = [1, 1]^T &= -\lambda g(y) \\ &= -\lambda [2y_1, 2y_2]^T = -\lambda/2 \cdot y \end{aligned}$$

Hence $y = [-1, -1]$ is the only KKT point

9.2 Nonlinear Problem (NLP)

Definition 9.2.1 (NLP)

$$\begin{aligned} min\ \ & f(x) \\ s.t.\ \ & g_i(x) \leq 0 \forall i \in \{1, \cdots, m\} \\ & h_i(x) = 0 \forall i \in \{1, \cdot, p\} \end{aligned}$$

Definition 9.2.2 (Linearized Feasible Direction & the Cone $L_{(NLP)}$)

Let y be feasible for (NLP), a linearized feasible direction is a vector $d \in \mathbb{R}^n$ such that

(1) $\forall i \in \{1, \cdots, m\}$ if $g_i(x) = 0$, then $d^T \nabla g_i(y) = 0$

(2) $\forall i \in \{1, \cdots, p\}$ have $d^T \nabla h_i(y) = 0$

The **Cone of Linearized feasible directions** at y is the set of all such directions, denoted as $L_{(NLP)}(y)$

Definition 9.2.3 (KKT Points)

Let $y \in \mathbb{R}^n$, y is a KKT point if it satisfies the KKT conditions :

(1) y is feasible for (NLP)

(2) $\forall d \in L_{(NLP)}(y),\ d^T \nabla f(y) \geq 0$

Theorem 9.2.1 (Farkas' Lemma)

Given $A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m$, $\{Ax = b : x \geq 0\}$ is feasible iff $\{A^T y \geq 0 : b^T y < 0\}$ is infeasible

Proof

($\Rightarrow$) Let $\bar{x}$ be such that $A\bar{x} = b, \bar{x} \geq 0$, i.e. a feasible solution

Then $\forall y \in \mathbb{R}^m$, if $A^T y \geq 0$, have $x^T A^T y \geq x^T 0 \geq 0$

Also $x^T A^T y \geq 0$ gives $(Ax)^T y \geq 0$, i.e. $b^T y \geq 0$

Thus $\{A^T y \geq 0 : b^T y < 0\}$ is infeasible

($\Leftarrow$) Consider the **Primal-Dual** pair

(P) $= min\{0^T x : Ax = b, x \geq 0\}$

(D) $= max\{b^T y : A^T y \leq 0\}$

Note that (P) cannot be unbounded as $0^T x$ is always 0

Note that (D) cannot be infeasible as $y = 0$ is a feasible solution

Using contrapositive, have

$$
\begin{aligned}
&\{Ax = b : x \geq 0\} \text{ being infeasible} \\
\Rightarrow &(P) \text{ is infeasible} \\
\Rightarrow &(D) \text{ is unbounded} \\
\Rightarrow &\exists\, d \in \mathbb{R}^m : A^T d \leq 0 \text{ and } b^T d > 0 \\
&\text{Let } y = -d, \text{ so } A^T y \geq 0 \text{ and } b^T y < 0 \\
\Rightarrow &\{A^T y \geq 0 : b^T y < 0\} \text{ is feasible}
\end{aligned}
$$

Theorem 9.2.2

Let $A \in \mathbb{R}^{m \times n}, B \in \mathbb{R}^{m \times p}, b \in \mathbb{R}^m$, then $\{Ax + Bw = b : x \geq 0\}$ is feasible iff $\{A^T y \geq 0 : B^T y = 0, b^T y < 0\}$ is infeasible

Proof

$$
\begin{aligned}
&\{Ax + Bw = b : x \geq 0\} \text{ is feasible} \\
\Rightarrow&\{Ax + Bw^+ - Bw^- = b : x, w^+, w^- \geq 0\} \text{ is feasible} \\
\Rightarrow&\{\begin{bmatrix} A & B & -B \end{bmatrix} \cdot \begin{bmatrix} x \\ w^+ \\ w^- \end{bmatrix} = b : \begin{bmatrix} x \\ w^+ \\ w^- \end{bmatrix} \geq 0\} \text{ is feasible} \\
\Rightarrow&\{\begin{bmatrix} A & B & -B \end{bmatrix}^T y \geq 0 : b^T y \geq 0\} \text{ is infeasible} \\
\Rightarrow&\{A^T y \geq 0 : B^T y \geq 0, -B^T y \geq 0, b^T y \geq 0\} \text{ is infeasible} \\
\Rightarrow&\{A^t y \geq 0, b^T y = 0, b^T y \geq 0\} \text{ is infeasible}
\end{aligned}
$$

Variant of Fakas' Lemma : $\begin{cases} Ax + Bw & = b \\ x & \geq 0 \end{cases}$ feasible $\Longleftrightarrow \begin{cases} A^T y & \geq 0 \\ B^T y & = 0 \\ b^T y & < 0 \end{cases}$ is infeasible

KKT conditions at $\bar{x}$ feasible for (NLP)

For all d such that $\begin{cases} d^T \nabla g_i(\bar{x}) \leq 0 & \text{for all } i = 1, \cdots, m \text{ with } g_i(\bar{x}) = 0 \\ d^T \nabla h_i(\bar{x}) = 0 & \text{for all } i = 1, \cdots, p \end{cases}$ and we have these two conditions $\Rightarrow \; d^T \nabla f(\bar{x}) \geq 0$

Using $A \wedge B \Rightarrow C$ is equivalent to $\neg(A \wedge B \wedge \neg C)$, i.e. the system

$$
\begin{cases} -\nabla g_i(\bar{x})^T d \geq 0 & \text{for all } i \,:\, g_i(\bar{x}) = 0 \\ \nabla h_i(\bar{x})^T d = 0 & \text{for all } i \\ \nabla f(\bar{x})^T d < 0 \end{cases} \quad \text{is infeasible}
$$

By Farkas's Lemma, it is equivalent to $\exists \, \lambda \in \mathbb{R}^m, \lambda \geq 0, \mu \in \mathbb{R}^l$ such that

$$
- \sum_{i: g_i(\bar{x})=0} \lambda_i \nabla g_i(\bar{x}) + \sum_i \mu_i \nabla h_i(\bar{x}) = \nabla f(\bar{x})
$$

Theorem 9.2.3 (KKT Gradient Equation or Complementary Equation)
Given (NLP), a feasible point $\bar{x}$ is a KKT point iff

$\exists \, \lambda \in \mathbb{R}^m, \lambda \geq 0, \mu \in \mathbb{R}^l$ such that $\begin{cases} - \sum_{i: g_i(\bar{x})=0} \lambda_i \nabla g_i(\bar{x}) + \sum_i \mu_i \nabla h_i(\bar{x}) & = \nabla f(\bar{x}) \\ \lambda_i \cdot g_i(\bar{x}) & = 0 \end{cases}$

Example 9.2.1
consider the system $\begin{cases} min & c^T x \\ s.t. & Ax = b \\ & x \geq 0 \end{cases}$ or equivalently $\begin{cases} min & c^T x \\ s.t. & -Ix \leq 0 \\ & Ax - b = 0 \end{cases}$

We have $g_i(x) = -x_i$, $h_i(x) = A^{i^T} - b_i$, $\nabla g(x) = -e_i$, $\nabla h_i(x) = A^{i^T}$

KKT conditions : $\exists\, \lambda \geq 0\,, \mu$ such that $\begin{cases} \sum_i \lambda_i e_i + \sum_i \mu_i A^{i^T} = c \\ \lambda_i \cdot (-\bar{x}_i) = 0 \end{cases} \iff \begin{cases} \lambda I + A^T \mu = c & (1) \\ \bar{x}^T \lambda = 0 & (2) \\ \lambda \geq 0 & (3) \end{cases}$

(1) gives $\lambda = c - A^T \mu \geq 0$, i.e. the system is equivalent to

$$\begin{cases} A^T \mu \leq c & \Leftarrow \text{ (dual feasibility)} \\ (c - A^T \mu)^T \bar{x} = 0 & \Leftarrow \text{ (complementary slackness)} \end{cases}$$

Let $\Omega = \{x \in \mathbb{R}^n : g_i(x) \leq 0\, \forall i\,,\, h_i(x) = 0\, \forall i\}$ (feasible region of (NLP))

Definition 9.2.4 (Feasible Arc)
A feasible arc at x in the direction of d is a function $\phi : [0, c] \to \mathbb{R}^n$ for some $c > 0$ s.t.

(1) $\phi(0) = x$

(2) $\phi \in C^1([0, c])$

(3) $\phi'(0) = d$

(4) $\phi(t) \in \Omega$, for all $t \in [0, c]$

Definition 9.2.5 (Tangent Cone)
Given a point $x \in \mathbb{R}^n$, the tangent cone to Ω at x is $T_\Omega(x) = \{d \in \mathbb{R}^n : \exists$ feasible arc at x with direction $d\}$

Example 9.2.2
$\Omega = \{x \in \mathbb{R}^2 : ||x|| \leq 1\}$ and $x = [-1, 0]^T$, then $T_\Omega(x) = \{[d_1, d_2]^T : d_1 \geq 0\}$

Lemma 9.2.1
Let $\phi : \mathbb{R} \to \mathbb{R}^n$ and $f : \mathbb{R}^n \to \mathbb{R}$ with $\phi_i \in C^1(\mathbb{R})$ for all i and $f \in C^1(\mathbb{R}^n)$, then $\big(\frac{d}{dt} f(\phi(t))\big)(t_0) = \nabla f(\phi(t_0))^T (\frac{d}{dt}\phi)(t_0)$

Proof
By the Chain Rule, given functions a, b :

$$a(b(x))'(x_0) = a'(b(x_0))b'(x_0)$$

Also by the Chain Rule, given $a(y_1, y_2)$

$$a\big(b(x), c(x)\big)'(x_0) = \frac{\partial}{\partial y_1} a\big(b(x_0), c(x_0)\big) b'(x_0) + \frac{\partial}{\partial y_2} a\big(b(x_0), c(x_0)\big) b'(x_0)$$

Therefore

$$\big(\frac{d}{dt} f(\phi(y))\big)(t_0) = \big(\frac{d}{dx_1} f(\phi(t_0))\big)\big(\frac{d}{dt}\phi\big)(t_0) + \cdots + \big(\frac{d}{dx_n} f(\phi(t_0))\big)\big(\frac{d}{dt}\phi\big)(t_0) = \nabla f(t_0)^T \phi'(t_0)$$

Theorem 9.2.4
Let x be feasible for (NLP) and assume $L_{(NLP)} = T_\Omega$, then if x is a local minimizer of (NLP), then it is a KKT point.

Proof
By definition of a KKT point, we want a feasible x and all $d \in L_{(NLP)}(x)$ such that $d^T \nabla f(x) \geq 0$

Let $d \in L_{(NLP)}(x)$, then $d \in T_\Omega(x)$, so $\exists \phi$ feasible arc at x with direction d

Let $\gamma(t) = f(\phi(t))$ for $t \geq 0$

$$\gamma'(0) = \lim_{t \to 0, t>0} \frac{\gamma(t) - \gamma(0)}{t}$$

by definition of γ, have $\gamma(0) = f(\phi(0)) = f(x)$

Since x is a local minimizer, $\gamma(t) - \gamma(0) = f(\phi(t)) - f(x) \geq 0$, thus $\gamma'(0) \geq 0$

By the above lemma, $\gamma'(0) = \nabla f(\phi(0))^T \phi'(0) = \nabla f(x)^T d$

Therefore $\nabla f(x)^T d \geq 0$

Example 9.2.3 (when minimizer is not a KKT point)

$$\begin{aligned} min\ & x_1 + x_2 \\ s.t.\ & -x_2 \leq 0 \\ & -x_1^3 + x_2 \leq 0 \end{aligned}$$

Minimizer is $x^* = [0,0]^T$

Let $f(x) = x_1 + x_2, \nabla f(x) = [1,1]^T, \nabla f(x^*) = [1,1]^T$

And $g_1(x) = -x_2, \nabla g_1(x) = [0,-1]^T, \nabla g_1(x^*) = [0,-1]^T$

Also $g_2(x) = -x_1^3 + x_2, \nabla g_2(x) = [-3x_1^2, 1]^T, \nabla g_2(x^*) = [0,1]^T$

KKT gradient equation system gives

$$\begin{aligned} \nabla f(x^*) = -\lambda_1 g_1(x^*) - \lambda_2 g_2(x^*) \\ \lambda_1 g_1(x^*) = 0 \\ \lambda_2 g_2(x^*) = 0 \end{aligned}$$

Which is $-\lambda_1[0,-1] - \lambda_2[0,1] = [1,1]$, which is infeasible, so x^* cannot be a KKT point

Remark
In the example,

$$\begin{aligned} T_\Omega(x^*) = \{x \in \mathbb{R}^2 : x_1 \geq 0, x_2 = 0\} \\ L_{(NLP)}(x^*) = \{x \in \mathbb{R}^2 : x_2 = 0\} \end{aligned}$$

These two cones are distinct

The tangent cone only care about the feasible region

The cone of linearized feasible directions cares about the gradients of the specific problem

Example 9.2.4

$$
\begin{aligned}
min\ & x_1 + x_2 \\
s.t.\ & -x_2 \leq 0 \\
& -x_1^3 + x_2 \leq 0 \\
& -x_1 \leq 0
\end{aligned}
$$

$g_3(x) = -x_1, \nabla g_3(x) = [-1, 0]^T, \nabla g(x^*) = [-1, 0]^T$

The gradient equation gives $\nabla f(x^*) = -\lambda_1[0, -1] - \lambda_2[0, 1] - \lambda_3[1, 0] = [1, 1]$, which is feasible

Therefore x^* is a KKT point

Theorem 9.2.5

$\forall x \in \Omega, T_\Omega \subseteq L_{(NLP)}(x)$

9.3 Constrained Optimization

Given NLP

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

$T_\Omega(x) = \{d \in \mathbb{R}^n : \exists \text{ feasible arc } \phi : \phi'(0) = d\}$

$L_{NLP}(x) = \{d \in \mathbb{R}^n : \nabla g_i(x)^T d \leq 0, \forall i : g_i(x) = 0, \nabla h_j(x)^T d = 0, \forall j\}$

KKT Point : $x \in \Omega$ is a KKT point if $\nabla f(x)^T d \geq 0, \forall d \in L_{NLP}(x)$, iff $\exists\, \lambda_i, \mu_i$ where all these holds:

$$
\begin{aligned}
-\sum \lambda_i \nabla g_i(x) + \sum \mu_i h_i(x) &= \nabla f(x) \\
\lambda_i &\geq 0 \\
\lambda_i g_i(x) &= 0, \forall i
\end{aligned}
$$

Theorem 9.3.1

Let $x \in \Omega$ such that $T_\Omega(x) = L_{NLP}(x)$, if x is a local minimizer, then x is a KKT point

Definition 9.3.1 (Constraint Qualification)

A constrained qualification (CQ) is a condition on the feasible set of NLP s.t. $T_\Omega(x) = L_{NLP}(x)$

Theorem 9.3.2

Let $x \in \Omega$, then $T_\Omega(x) \subseteq L_{NLP}(x)$

Proof
Let $x \in \Omega$ and $d \in T_\Omega(x)$

There exists $c > 0$ and $\phi : [0, c]$ such that

$$\phi(0) = x$$
$$\phi \text{ is } C^0 \text{ smooth and } \phi'(0) = d$$
$$\phi(t) \in \Omega, \forall t \in [0, c]$$

We want $d \in L_{NLP}(x)$ such that

$$\nabla g_i(x)^T d = 0, \forall i \text{ such that} g_i(x) = 0$$
$$\nabla h_j(x)^T d = 0, \forall j$$

Suppose $\exists\, i : g_i(x) = 0$, consider Taylor expansion $g_i \circ \phi$ at 0 in the direction $t \in [0, c]$

Define a function $o(t)$ where $\lim_{t\to 0} \frac{o(t)}{t} = 0$

Note that $g_i\big(\phi(t)\big) \leq 0$ and that $g_i\big(\phi(0)\big) = g_i(x) = 0$, hence have

$$\begin{aligned}
g_i\big(\phi(t)\big) &= g_i\big(\phi(0)\big) + g_i'\big(\phi(0)\big)t + o(t) \\
0 &\geq g_i'\big(\phi(0)\big)t + o(t) \\
&= \nabla g_i\big(\phi(0)\big)^T \phi'(0)t + o(t) \\
&= \nabla g_i\big(\phi(0)\big)^T dt + o(t) \\
\text{(Divide both sides by } t\text{) } 0 &\geq \nabla g_i\big(\phi(0)\big)^T d + \frac{o(t)}{t} \\
\text{(Taking the limit of both sides) } 0 &\geq \nabla g_i\big(\phi(0)\big)^T d + \lim_{t\to 0} \frac{o(t)}{t} \\
&= \nabla g_i\big(\phi(0)\big)^T d
\end{aligned}$$

Exercise : do for $h_j(x)$

Definition 9.3.2 (Linear Independence CQ (LICQ))
The LICQ holds at $x \in \Omega$ if the set $\{\nabla g_i(x) : g_i(x) = 0\} \cup \{\nabla h_j(x) : \forall j\}$ is linear independent

Theorem 9.3.3
Let $x \in \Omega$, if x satisfies LICQ, then $T_\Omega(x) = L_{NLP}(x)$

Proof
read up on it

Remark
$h(x) = 0 \iff h(x) \leq 0, -h(x) \leq 0$

Example 9.3.1
$\min\{x_1 + x_2 : -x_2 \leq 0, -x_1^3 + x_2 \leq 0\}$ with $x^* = [0, 0]$

Does the LICQ hold at x^*?

$\nabla g_1(x) = [0, -1]$, $\nabla g_2(x) = [-3x_1^2, 1]$ at $x^* : \{[0, -1], [0, 1]\}$, not linearly independent

Definition 9.3.3 (Linear Programming CQ (LPCQ))

The LPCQ holds at $x \in \Omega$ if all the tight constraints are affine (of form $ax - b$)

Theorem 9.3.4

Let $x \in \Omega$, if the LPCQ holds at x, then $T_\Omega(x) = L_{NLP}(x)$

Proof

Let $x \in \Omega$ be such that LPCQ holds

Then by definition of LPCQ, have

$$\begin{aligned}
&g_i(x) < 0, i = 1, \cdots, k \text{ for some } k \\
&g_i(x) = 0, i = k+1, \cdots, m \implies g_i(x) = a_i^T x - b_i \\
&h_j(x) = 0, \forall j \implies h_j(x) = a_j^T x - b_j
\end{aligned}$$

Let $d \in L_{NLP}(x)$, we want to prove $d \in T_\Omega$ by definition of $L_{NLP}(x)$, we have

$$\begin{aligned}
&0 \geq \nabla g_i(x)^T d = a_i^T d, i = k+1, \cdots, m \\
&0 = \nabla h_j(x)^T d = a_j^T d, \forall j
\end{aligned}$$

Consider $\phi(t) = x + td$, we have $\phi(0) = x$ smooth with $\phi'(t) = d, \phi(t) \in \Omega$, have

$$\begin{aligned}
\forall j, h_j(x+td) = a_j^T(x+td) - b_j = a_j^T x - b_j + a_j^T td = a_j^T x - b_j = 0 \\
\forall i = k+1, \cdots, m, g_i(x+td) = a_j^T(x+td) - b_i = a_i^T x - b_i + a_i^T td = g_i(x) + t\nabla g_i(x)^T d \geq 0 \\
\forall i = 1, \cdots, k, g_i(x+td) \leq 0, \forall t \in [0, \epsilon_i), \epsilon_i > 0, \text{ by continuity of } g_i
\end{aligned}$$

Then ϕ is a feasible arc $\forall t \leq min\{\epsilon_i\}$, so $d \in T_\Omega(x)$

Theorem 9.3.5

Let $x \in \Omega$ such that a CQ holds at x, if x is a local minimizer, then x is also a KKT point

Review

LICQ The set $\{\nabla gi(x) : g_i(x) = 0\} \cup \{\nabla h_j(x), \forall i\}$ is linearly independent

LPCQ The tight constraints at x are all affine.

x is a KKT point if $\exists \lambda_i, \mu_j$ such that

(1) $\nabla f(x) = -\sum \lambda_i \nabla g_i(x) + \sum \mu_j \nabla h_j(x)$

(2) $\lambda_i \geq 0$

(3) $\lambda_i g_i(x) = 0$

9.4 Constraint Qualifications

Example 9.4.1

$$\begin{aligned} &min \;\; x^TAx \;\; , x \in \mathbb{R}^n, A \in \mathbb{R}^{n\times n} \; symmetric \\ &s.t. \; ||x|| = 1 \end{aligned}$$

note this is a continuous function on a compact constrained set, but the norm function is not continuously differentiable, how do we fix this? we can fix the constrained to be $||x||^2 = 1$, which is a equivalent one. Now check the constrained qualification before do KKT point

We have **LICQ** as the set $\{2x\}$ is linearly independent when $x \neq 0$, then every minimizer will satisfies the KKT, i.e. KKT satisfies at $\bar{x}$ if $\exists\, \mu$ such that

$$(1) 2A\bar{x} = \mu(2\bar{x}) \iff A\bar{x} = \mu\bar{x}\,,\text{ at } \bar{x} \text{ the objective value is } \bar{x}^TA\bar{x} = \bar{x}^T(\mu\bar{x}) = \mu$$

9.5 Convex NLP

$$\begin{aligned} &min \; f(x) \\ &\;s.t. \; g_i(x) \leq 0, \forall i \\ &\qquad h_j(x) = 0, \forall j \\ &\qquad f\,,\, g_i \text{ are convex} \\ &\qquad h_j \;\text{ is affine} \end{aligned}$$

Definition 9.5.1 (Slater CQ or Strict Feasibility)

The Slater CQ holds for (Convex Program) if $\exists\, \bar{x} \in \Omega$ s.t. $g_i(\bar{x}) < 0, \forall i$

Theorem 9.5.1

If the slater CQ holds for (CP), then $T_\Omega(x) = L_{NLP}(x)$ for all $x \in \Omega$

Theorem 9.5.2

Let x be a KKT point for (CP), then x is a global minimizer of (CP).

Proof

Let $y \in \Omega, y \neq x$, we want to show that $f(x) \leq f(y)$

Since x is a KKT point, it follows that $\nabla f(x)^T d \geq 0\, \forall d \in L_{NLP}(x)$

Now we show that $d := (y - x) \in L_{NLP}(x)$

Recall that if $c \in C^1$ and convex function, then

$$c(\hat{x}) \geq c(\bar{x} + \nabla c(\bar{x})^T(\hat{x} - \bar{x})\forall \hat{x}, \bar{x} \tag{9.1}$$

Suppose there is a tight constraint i be such that $g_i(x) = 0$, then by (9.1), have

$$\begin{aligned} \underbrace{g_i(y)}_{\leq 0} &\geq \underbrace{g_i(x)}_{=0} + \nabla g_i(x)^T(y - x) \\ 0 &\geq \nabla g_i(x)^T(y - x) \end{aligned}$$

Suppose there is a affine constraint $j \in \{1, \cdots, p\}$, we need $\nabla h_j(x)^T(y-x) = 0$

Since h_j is affine, $h_j(x) = a_j^T x + b_j$, hence

$$h(x) = 0 \iff a_j^T x + b_j = 0 \tag{9.2}$$
$$h(y) = 0 \iff a_j^T y + b_j = 0 \tag{9.3}$$

Subtract (9.2) and (9.3), we have $a_j^T(y-x) = 0$, i.e. $\nabla h_j(x)^T(y-x) = 0$

We have shown that $d := (y-x) \in L_{NLP}(x)$

By the KKT conditions $\nabla f(x)^T d \geq 0, \forall d \in L_{NLP}(x)$, which gives $\nabla f(x)^T(y-x) \geq 0$

Since f is convex by (9.1), we have

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

What if $g_i(x) < 0$? Things can be really good or really bad

Consider $g_i(x) < 0$ is an open set (for example interval), if f is linear, then we can never obtain an optimal soln, thus there is no KKT point

if f is quadratic and obtain the minimal value at z in the interior of the open set, then $\nabla f(z) = 0$, we have $0 = \nabla f(x) = -\sum \lambda_i \nabla g_i(x) + \sum \mu_j \nabla h_j(x)$, we can just choose $\lambda_i = \mu_j = 0$

Corollary 9.5.1

Suppose the slater CQ holds for (CP), then x is a global minimizer iff x is a KKT point

Remark

All the results for convex optimization also hold when the function are not C^1

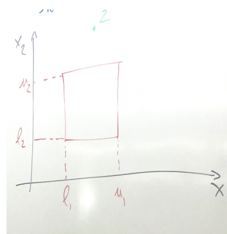
Example 9.5.1

$$\begin{aligned} min \ \ & ||x - x^0||_2^2 \quad \text{this is called projection} \\ s.t. \ \ & l_i \leq x_i \leq u_i, \forall i \ l_i < u_i \quad \text{this is called Box constraint} \end{aligned}$$

note the objective function is a strict convex quadratic function as the hessian $\nabla^2 f(x) = 2I$ is positive definite. And the constraints $x_i - u_i \leq 0$ and $-x_i + l_i \leq 0$ are affine thus convex. So this is a (CP). We can choose $x_i = (l_i + u_i)/2$ as a slater point.

The KKT conditions are

$$\begin{aligned} 2(x - x^0) &= -\sum \lambda_i^+ e_i + \sum \lambda_i^-(-e_i) \\ x &= \frac{1}{2}(x^0 - \sum(\lambda_i^+ + \lambda_i^-)e_i) \end{aligned}$$



Example 9.5.2 (Projection onto a box)
Given $l, u, z \in \mathbb{R}^n$ with $l < u$, solve $min \, ||x - z||^2$ subject to $l \leq x \leq u$. (how to construct the point and prove it is the optimal)

$$f(x) = (x-z)^T(x-z) = x^T I z - 2z^T x + z^T z$$
$$g_i^l(x) = l_i - x_i \, , \, g_i^u(x) = x_i - u_i$$
$$\nabla f(x) = 2x - 2z$$
$$\nabla g_i^l(x) = -e_i \, , \, \nabla g_i^u(x) = e_i$$

LICQ : For all i, at most one of $g_i^l(x), g_i^u(x)$ is zero

Thus the gradient of the active constraints at any feasible x give a subset of the columns of an identity matrix $\Rightarrow$ linearly independent

LPCQ : All $g_i(x)$ are linear (affine)

Slater : Slater point : $\frac{l+u}{2}$

Hence KKT is necessary

f is convex and feasible region Ω is convex, thus KKT is sufficient

Feasibility : $l \leq x \leq u$

KKT eqn : $-\sum_i \lambda_i^l(-e_i) - \sum_i \lambda_i^u e_i = 2(x-z) \iff \lambda^l - \lambda^u = 2(x-z), \lambda_i^l, \lambda_i^u \geq 0$

Complementarity : $(x_i - l_i)\lambda_i^l = 0 \, , \, (u_i - x_i)\lambda_i^u = 0$

For $i = 1, \cdots, n$

CASE 1 : $z_i < l_i$

For any $x \in \Omega, z_i < x_i$, for $\lambda_i^l - \lambda_i^u = 2(x_i - z_i) > 0$, we need $\lambda_i^l > 0$

Since $(x_i - l_i)\lambda_i^l = 0$, we have $x_i = l_i$

CASE 2 : $z_i > u_i$

For any $x \in \Omega, z_i > x_i$, so $\lambda_i^u > 0$, thus $x_i = u_i$

CASE 3 : $l_i \leq z_i \leq u_i$

If $z_i < x_i$, by **CASE 1** , we have $\lambda_i^l = 0 \Leftarrow x_i = l_i \leq z_i < x_i$, Contradiction!

Similarly if $z_i > x_i$, $\lambda_i^u > 0 \Leftarrow x_i = u_i \geq z_i > x_i$, Contradiction!

Therefore $x_i = z_i$

Algorithm : For all i, $x_i = median(l_i, u_i, z_i)$

Chapter 10

Algorithms For Constrained Optimization

10.1 Equality-Constrained Optimization

$$min\,\{f(x) : h_i(x) = 0, \forall i = 1, \cdots, n\}$$

Quadratic Penalty Method

Choose $x^0, \rho > 0$

For $k = 0, 1, 2, \cdots$

$$x^{k+1} = argmin\{g_\rho(x)\}, \text{ where } g_\rho(x) = f(x) + \rho \sum_{i=1}^{n} \big(h_i(x)\big)^2$$

(initialize unconstrained method at x^k)

$\rho = C \cdot \rho,$ where $C > 1$

Note that for a large ρ, to minimize the objective function g_ρ, it forces $h_i(x)$ to be really small

But when ρ is too big, such as $\frac{1}{\epsilon}$, we will have a problem, this is why we just say ρ is a large number but not directly given a really large number

Example 10.1.1

$$\begin{aligned} &min\ (x_1 - 1)^2 + (x_2 - 1)^2 \\ &s.t.\ x_1 + x_2 = 4 \end{aligned}$$

The level set is a flat ellipsoid around $x_1 = x_2 = 2$ on the line $x_1 + x_2 = 4$, with a really large ρ, the algorithm will give any point in this ellipsoid, finally converging to the minimizer point, i.e. the soln will become unstable as ρ getting large.

Theorem 10.1.1
Let $f, h_1, \cdots, h_n \in C^1(\mathbb{R}^n)$ and let $g(x) = || \begin{bmatrix} h_1(x) \\ \ddots \\ h_n(x) \end{bmatrix} ||^2 = \sum_i \big(h_i(x)\big)^2$

Suppose $x^k \to x^*$ and $\nabla h_1(x^*), \cdots, h_n(x^*)$ are linearly independent, then

Either (1) $\nabla g(x^*) = 0$ and $g(x^*) > 0$ Or (2) x^* is a KKT point

Quadratic Penalty Method :

$$g_\rho(x) = f(x) + \rho \sum_{i=1}^{n} \left(h_i(x)\right)^2$$

Drawbacks :

For a large ρ, the unconstrained problem is bad numerically.

By design, ρ has to be large as when $h_i(x^k) = 0$, $\nabla\left(h_i(x)\right)^2$ becomes small.

Exact Penalty Method :

$$g_\rho(x) = f(x) + \rho \sum_{i=1}^{n} |h_i(x)|$$

Advantages : When $h_i(x^k) \approx 0$, $\nabla |h_i(x^k)|$ is constant

Drawbacks : $|h_i(x)|$ is not differentiable

Augmented Lagrangian Penalty Method :

Lagrangian Relaxation :

$$L(x, \mu) = f(x) - \sum_{i=1}^{n} \mu_i \cdot h_i(x)$$

Theorem 10.1.2

KKT points $\bar{x}$ of (NLP) with multipliers $\bar{\mu}$ coincide with stationary points $(\bar{x}, \bar{\mu})$ of L

Proof

KKT conditions for (NLP) :

Feasibility : $h_i(x) = 0, \forall i = 1, \cdots, n$

Gradient Equation : $\exists\, \bar{\mu}\ :\ \sum_i \bar{\mu}_i \nabla h_i(\bar{x}) = \nabla f(\bar{x})$

Stationary Point of L : $\nabla L(\bar{x}, \bar{\mu}) = 0$, since

$$\nabla L(\bar{x}, \bar{\mu}) = \begin{bmatrix} \nabla_{\bar{x}} L(\bar{x}, \bar{\mu}) \\ \nabla_{\bar{\mu}} L(\bar{x}, \bar{\mu}) \end{bmatrix} = \begin{bmatrix} \nabla f(\bar{x}) - \sum_i \bar{\mu}_i \nabla h_i(\bar{x}) \\ h_1(\bar{x}) \\ \vdots \\ h_n(\bar{x}) \end{bmatrix} = 0$$

Remark

$\bar{x}$ is a KKT point for (NLP) iff $\exists\, \bar{\mu}\ :\ \nabla L(\bar{x}, \bar{\mu}) = 0$

Important : The above does not imply that $(\bar{x}, \bar{\mu})$ is a local minimizer for L, however, for any KKT point $\bar{x}$, $\exists\, \bar{\mu}$ such that $\bar{x}$ is a local minimizer for $min\left\{L(\bar{x}, \mu)\ :\ \mu = \bar{\mu}\right\}$

Finding $\bar{\mu}$ By Augmented Lagrangian Method

$$\begin{aligned}
&\text{Choose } x^0, \mu^0, \rho > 0 \\
&\quad \text{For } k = 0, 1, 2, \cdots \\
&\qquad x^{k+1} = argmin\{L_A(x, \mu^k)\} \ \# \text{ argmin of } x \\
&\qquad \text{where } L_A(x,\mu) = f(x) - \sum_{i=1}^{n} \mu_i \cdot h_i(x) + \rho \sum_{i=1}^{n} \big(h_i(x)\big)^2 \\
&\qquad \text{(initialize unconstrained method at } x^k) \\
&\qquad \mu_i^{k+1} = \mu_i^k - 2\rho \cdot h_i(x^{k+1}) \qquad (10.1) \\
&\qquad \rho = C \cdot \rho, \ \text{where } C > 1
\end{aligned}$$

Why (10.1)? At x^{k+1} :

$$0 = \nabla_x L_A(x^{k+1}, \mu^k) = \nabla f(x^{k+1}) - \sum_{i=1}^{n} \mu_i^k \cdot \nabla h_i(x^{k+1}) + 2\rho \sum_{i=1}^{n} h_i(x^{k+1}) \cdot \nabla h_i(x^{k+1})$$

$$\begin{aligned}
\nabla f(x^{k+1}) &= \sum_{i=1}^{n} \mu_i^k \cdot \nabla h_i(x^{k+1}) + 2\rho \sum_{i=1}^{n} h_i(x^{k+1}) \cdot \nabla h_i(x^{k+1}) \\
&= \sum_{i=1}^{n} \big(\mu_i^k - 2\rho \cdot h_i(x^{k+1})\big) \cdot \nabla h_i(x^{k+1})
\end{aligned}$$

Setting μ_i^{k+1} to $\big(\mu_i^k - 2\rho \cdot h_i(x^{k+1})\big)$ lets us satisfy the gradient equation at x^{k+1}

Keep in mind we still miss feasibility of x^{k+1}, so x^{k+1} is not necessarily KKT

Advantages of Augmented Lagrangian Method :

$L_A(x, \mu)$ is differentiable, in practice, will usually converge before ρ grows too large

10.2 Inequality Constrained Optimization

Focus on the closed convex cone

Definition 10.2.1 (Closed Convex Cone)
K is a closed convex cone if it is closed, convex, nonempty and $x \in K, \lambda \geq 0 \Rightarrow (\lambda x) \in K$

The three most important cone are :

$$\begin{aligned}
\mathbb{R}_+^n &= \{x \in \mathbb{R}^n, x \geq 0\} \\
\mathbb{C}_2^{n+1} &= \{(y, x) \in \mathbb{R} \times \mathbb{R}^n \ : \ y \geq ||x||_2\} \\
\mathcal{S}_+^n &= \{X \in \mathbb{R}^{n \times n} : \ X \ \text{ is p.s.d}\}
\end{aligned}$$

Conic Programming

$$\begin{aligned}
&min \ c^T x \\
&s.t. \ Ax = b
\end{aligned}$$

Definition 10.2.2 (Interior and Boundary)
$int(K) = \{x \in K : \exists\, \delta > 0 \text{ such that } x \in K,\ B_\delta(x) \subseteq K,\ K \text{ is a closed convex cone}\}$

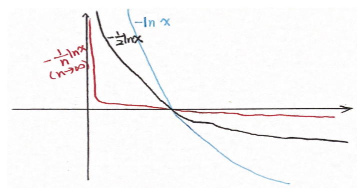
$boundary(K) = K \backslash int(K)$

Definition 10.2.3 (Barrier Function)
A barrier function is a convex function $\phi : int(K) \to \mathbb{R}$ such that $\lim_{x \to b} \phi(x) = +\infty$ for any $b \in boundary(K)$

The standard boundary functions are

- $\mathbb{R}^n_+ : \phi(x) = -\sum_i \log x_i$
- $\mathbb{C}^{n+1}_2 : \phi(x) = -\log(y^2 - ||x||^2_2)$
- $\mathcal{S}^n_+ : \phi(x) = -\log(det\, X)$

For $\mathbb{R}^n_+$, we want the penalty function of 0 makes all values equal to 0 and equal to $+\infty$ at 0, for $\mathbb{C}^{n+1}_2$, $y^2 - ||x||^2_2$ stands for x in the graph, and for $\mathcal{S}^2_+$, all the eigenvalues are ≥ 0, so $det\, X \geq 0$, consider when one of the eigenvalues goes to 0, $det\, X$ stands for x.



Primal Interior Point Method

$$
\begin{aligned}
&\text{Choose } x^0 \in\ int(K) : Ax^0 = b,\ \rho^0 > 0 \\
&\quad \text{For } k = 0, 1, \cdots \\
&\qquad x^{k+1} \simeq argmin\{g_\rho(x) : Ax = b\} \\
&\qquad \text{where } g_\rho(x) = c^T x + \rho^k \phi(x) \\
&\qquad \text{initialize at } x^k \\
&\qquad \rho^{k+1} = C \cdot \rho^k \text{ with } C < 1
\end{aligned}
\tag{10.2}
$$

The LP Case $(K = \mathbb{R}^n_+)$

We take a quadratic approximation of $g_\rho(x)$ for (10.2) at x^k :

$$
\begin{aligned}
g_\rho(x) &= c^T x + \rho^k \phi(x) \\
&\simeq c^T x \rho^k \phi(x^k) + \rho(x - x^k)^T \nabla \phi(x^k) + \frac{\rho^k}{2}(x - x^k)^T \nabla^2 \phi(x^k)(x - x^k)
\end{aligned}
$$

Let $h = x - x^k$, then (10.2) becomes :

$$\begin{aligned}
&min \quad \overbrace{c^T x^k}^{constant} + c^T h + \overbrace{\rho^k \phi(x^k)}^{constant} + \rho^k h^T \nabla h(x^k) + \frac{\rho^k}{2} h^T \nabla^2 \phi(x^k) h \\
&s.t. \ A(x^k + h) = b \\
&\quad \phi(x) = -\sum_i \log x_i \\
&\quad \nabla\phi(x) = \begin{bmatrix} -1/x_1 \\ \vdots \\ -1/x_n \end{bmatrix} \\
&\quad \nabla^2\phi(x) = diag(\frac{1}{x_1^2}, \cdots, \frac{1}{x_n^2})
\end{aligned}$$

Which is

$$\begin{aligned}
&min \ (c - \rho^k \begin{bmatrix} 1/x_1^k \\ \vdots \\ 1/x_n^k \end{bmatrix})^T h + \frac{\rho^k}{2} h^T \cdot diag(1/(x_1^k)^2, \cdots, 1/(x_n^k)^2) h \\
&s.t. \ Ah = b - Ax^k = 0 \text{ since } x^k \text{ satisfies } Ax^k = b
\end{aligned}$$

KKT conditions are : $\begin{cases} \text{gradient eq } : \sum_i \mu_i \nabla\gamma(h) = \nabla f(h) \\ \text{feasibility } : \mu_i \gamma_i(h) = 0 \end{cases}$

$$\Rightarrow \begin{cases} A^T \mu = (c - \rho^k \begin{bmatrix} 1/x_1^k \\ \vdots \\ 1/x_n^k \end{bmatrix}) + \rho^k \, diag(\frac{1}{(x_1^k)^2}, \cdots, \frac{1}{(x_n^k)^2}) h \\ Ah = 0 \end{cases}$$

$$\Rightarrow -\begin{bmatrix} -\rho^k \, diag(\frac{1}{(x_1^k)^2}, \cdots, \frac{1}{(x_n^k)^2}) & A^T \\ A & 0 \end{bmatrix} \cdot \begin{bmatrix} h \\ \mu \end{bmatrix} = -\begin{bmatrix} (c - \rho^k \begin{bmatrix} 1/x_1^k \\ \vdots \\ 1/x_n^k \end{bmatrix}) \\ 0 \end{bmatrix}$$

Note that the matrix is symmetric and easy to solve it numerically

Theorem 10.2.1

For $\mathbb{R}_+^n, \mathbb{C}_2^{n+1}, \mathbb{S}_+^n$, the primal interior point method satisfies $||x^k - x^*|| \leq \epsilon$ after $k = p(E, \epsilon)$ iterations, where

- p is a polynomial
- E is the encoding size of the problem

$\Rightarrow$ **Polynomial Time Algorithm**

Given p closed convex cones $K_1, \cdots, K_p$, we have that $K_1 \times \cdots \times K_p = K$ is also a convex cone. In practice, we can solve

$$\begin{aligned} min \;\; & c^T x \\ s.t. \;\; & Ax = b \\ & x \in K \end{aligned}$$

Example 10.2.1 (Euclidean Facility Location Problem)

Given $b_1, \cdots, b_k \in \mathbb{R}^n$, find $x \in \mathbb{R}^n$ that minimizes $\sum ||b_i - x||_2$ (note each b_i is a distinct vector, not ith element)

This problem is non-differentiable and $f(x) = \sum ||b_i - x||_2$ are not Lipschitz-continuous, reformulating :

$$\begin{aligned} min \;\; & \sum t_i \\ s.t. \;\; & t_i \geq ||b_i - x||_2 \, , i = 1, \cdots, k \, , \, (t_i, b_i - x) \in \mathbb{C}_2^{n+1} \end{aligned}$$

Reformulating again :

$$\begin{aligned} min \;\; & \sum t_i \\ s.t. \;\; & t_i \geq ||y_i||_2 \, , \, i = 1, \cdots, k \, , \, (t_i, y_i) \in \mathbb{C}_2^{n+1} \\ & y_i = b_i - x \, , \, i = 1, \cdots, k \end{aligned}$$

Once again

$$\begin{aligned} min \;\; & \sum t_i \\ s.t. \;\; & \mu \geq ||x||_2 \\ & t_i \geq ||y_i||_2 \\ & y_i = b_i - x \end{aligned}$$

Finally

$$\begin{aligned} min \;\; & 1^T t \\ s.t. \;\; & (\mu, x, t_1, y_1, \cdots, t_k, y_k) \in \mathbb{C}_2^{n+1} \times \cdots \times \mathbb{C}_2^{n+1} \\ & y_i = b_i - x \end{aligned}$$

10.3 Review For Duality

Definition 10.3.1 (Dual Cone)

The dual cone of a closed convex cone $K \subseteq \mathbb{R}^n$ is $K^* = \{s \in \mathbb{R}^n : s^T x \geq 0 \, , \forall x \in K\}$

Theorem 10.3.1

(1) K^* is a closed convex cone

(2) $(K^*)^* = k$

(3) $(\mathbb{R}^n_+)^* = \mathbb{R}^n_+ \, , \, (\mathbb{C}_2^{n+1})^* = \mathbb{C}_2^{n+1} \, , \, (\mathbb{S}^n_+)^* = \mathbb{S}^n_+$, these cones are self-dual

$$\begin{array}{ll} \textit{primal } (P) & \textit{dual } (D) \\ \min\ c^T x & \max\ b^T y \\ \quad Ax = b & \quad c - A^T y \in K \\ \quad x \in K & \end{array}$$

Theorem 10.3.2 (Weak Duality)
If x is feasible for (P) and y is feasible for (D), then $c^T x \geq b^T y$

Proof
Let $s = c - A^T y \in K^*$, then clearly $c = A^T y + s$, have

$$\begin{aligned} c^T x &= (A^T y + x)^T x \\ &= y^T Ax + s^T x \\ &= y^T b + s^T x \\ &\geq y^T b = b^T y \end{aligned}$$

Note that $s^T x \geq 0$ by definition of the dual cone.

Theorem 10.3.3 (Strong Duality)
If (P) has a Slater point, i.e. $\exists x \in int(K) : Ax = b$ and x^* is optimal for (P), then $\exists y^*$ optimal for (D) where $c^T x^* = b^T y^*$

Recall that at each iteration, we can solve $min\{g_\rho(x) : Ax = b\}$, where $g_\rho = f(x) + \rho \cdot \phi(x)$, solutions to this problem for fixed ρ are central points. Together, taking all $\rho > 0$, they create the central path.

In the case of LP, $(K = \mathbb{R}^n_+)$, $g_\rho = f(x) - \rho \sum \log(x_i)$, and we can assume that there is a slater point for any nontrivial case, so KKT conditions are sufficient for global optimality.

KKT conditions :

$$\begin{aligned} A^T \mu &= \nabla g_\rho(x) \\ Ax &= b \end{aligned}$$

Reformulate as

$$\begin{aligned} &A^T \mu = \nabla f(x) + \rho[-\frac{1}{x_1}, \cdots, -\frac{1}{x_n}]^T \\ &Ax = b \end{aligned}$$

Dual LP :

$$
\begin{aligned}
&max\ b^T y \\
&\quad A^T y \leq c
\end{aligned}
$$

Is equivalent to

$$
\begin{aligned}
-min\ & -b^T y \\
& A^T y + s = c \\
& s \geq 0
\end{aligned}
$$

Adding a barrier :

$$
\begin{aligned}
-min\ & -b^T y - \rho \sum \log(s_i) \quad \#\text{ denote this as } F(y,s) \\
& A^T y + s = c
\end{aligned}
$$

KKT conditions for modified dual

$$
\begin{aligned}
[A, I]^T \gamma &= [\nabla_y F(y,s), \nabla_s F(y,a)]^T \\
&= [-b_1, \cdots, -b_n, -\frac{\rho}{s_1}, \cdots, -\frac{\rho}{s_n}]^T \\
A^T y + s &= c
\end{aligned}
$$

Reformulate as

$$
\begin{aligned}
A\gamma &= -b \\
\gamma &= [-\frac{\rho}{s_1}, \cdots, -\frac{\rho}{s_n}]^T \\
A^T y + s &= c
\end{aligned}
$$

By identifying $\mu = y$ and $\gamma = -x$, we get

$$
\begin{aligned}
A^T y &= c + \rho[-\frac{1}{x_1}, \cdots, -\frac{1}{x_n}]^T \\
Ax &= b \\
\gamma &= [-\frac{\rho}{s_1}, \cdots, -\frac{\rho}{s_n}]^T \\
A^T y + s = c &\Rightarrow\ s = \rho[-\frac{1}{x_1}, \cdots, -\frac{1}{x_n}]^T
\end{aligned}
$$

Primal-Dual Interior Point Method : solve the system

$$
\begin{aligned}
A^T x &= b \\
A^T y + s &= c \\
x_i s_i &= \rho, \forall i = 1, \cdots, n
\end{aligned}
$$

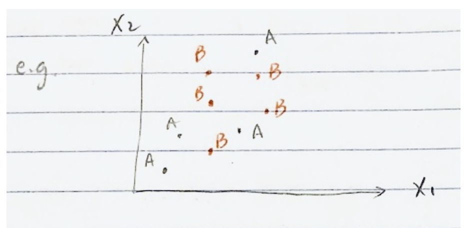
using Newton's method (variant)

Chapter 11

Introduction to Neural Networks

Machine Learning : Classification/Labelling Problem

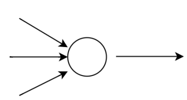
We are given N input vectors in $[0,1]^n$ that are already labelled into categories (the "training set"), can an algorithm assign "good" (accurate) labels to more vectors?



11.1 Neural Networks (NN)

A trained NN provides a function $F : \mathbb{R}^n \to \mathbb{R}^k$. If $x \in \mathbb{R}^n$ is an input vector, $j^* = argmax_j\{F(x)\}$.
(1) Given NN, how is $F(x)$ computed? (2) How to get an NN that is a good classifier?

For a single neuron (one neuron):



output = σ_1 (a linear combination of inputs)

Typical choices for $\sigma_1(x) : \mathbb{R} \to \mathbb{R}$ is

sigmoid function

$$\sigma_1(x) = \frac{1}{1+e^{-x}}$$

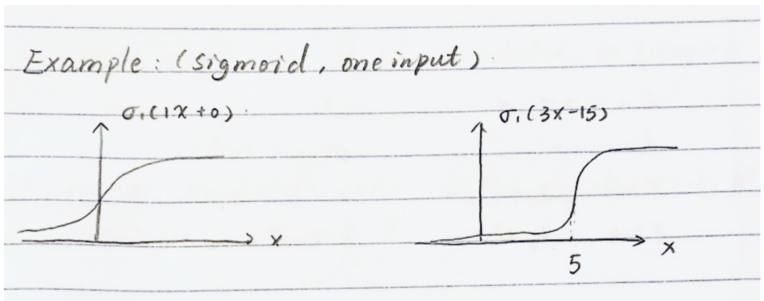
or Rectified linear unit (ReLU) :

$$\sigma_1(x) = \begin{cases} 0 & x \leq 0 \\ x & x > 0 \end{cases}$$

Example 11.1.1

$\sigma_1(1x + 0)$, normal sigmoid function

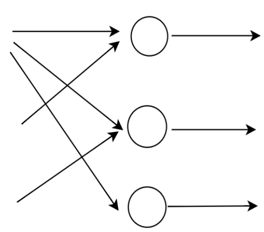
$\sigma_1(3x + 15)$, shifts points to $x = 5$, transition is much sharper



Definition 11.1.1 (Weight & Bias)

In $\sigma_1(w^t x + b_1), w \in \mathbb{R}^k_l$ is the weight and $b_1 \in \mathbb{R}^1$ is the bias

For a layer of neurons :



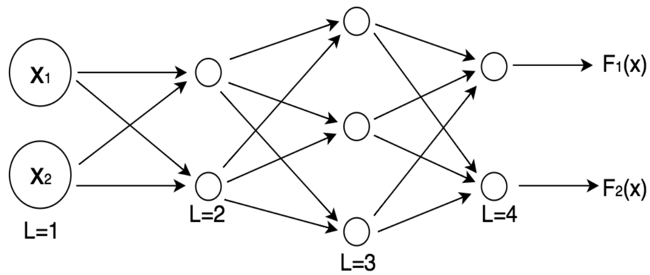
All k_l neurons in layer l have the same inputs $x \in \mathbb{R}^{n_l}$. Together, their output is in $\mathbb{R}^{k_l}$. The output of a layer l is $\sigma(w \cdot x + b), W \in \mathbb{R}^{k_l}$. For $w \in \mathbb{R}^{k_l \times n_l}, b \in \mathbb{R}^{k_l}$,

we define $\sigma : \mathbb{R}^{k_l} \to \mathbb{R}^{k_l}$ as

$$\sigma(x) = \begin{bmatrix} \sigma_1(x_1) \\ \vdots \\ \sigma_{k_l}(x_{k_l}) \end{bmatrix}$$

For a neural network :

$l = 1$	output x
$l = 2$	output $\sigma(w^2 x + b^2)$
$l = 3$	output $\sigma\big(w^3 \cdot \sigma(w^2 x + b^2) + b^3\big)$
$l = 4$	output $\sigma\Big(w^4 \cdot \sigma\big(w^3 \cdot \sigma(w^2 x + b^2) + b^3\big) + b^4\Big)$



A deep neural networks means the number L of layers is large. A hidden layer is a layer l with $l \neq 1$ and $l \neq L$

Definition 11.1.2 (Training)
Training is the process of finding w^l, b^l for $l = 2, \cdots, L$ that give a "good" neural network (give a accurate classifier)

Definition 11.1.3 (Cost Function)
A cost function is a function of the weights and biases that has a "low" value when the neural network gives a "good" classification of the training data.

Typical Cost Function : Quadratic cost function

$$cost(w^2, \cdots, w^L, b^2, \cdots, b^L) = \frac{1}{N}\sum_{j=1}^{n}\frac{1}{2}||y(x^j) - F(x^j)||_2^2$$

where $y(x^j) = e_k$ if x^j is labelled to category k

Training is to find :

$$\min_{w^l, b^l, l=2,\cdots,L} \frac{1}{N}\sum_{j=1}^{n}\frac{1}{2}||y(x^j) - F(x^j)||_2^2$$

11.2 Gradient Descent For NN

Let's define the parameter vector $p \in \mathbb{R}^p$ containing all entries of w^l, b^l for $l = 2, \cdots, L$. Typically, no line search. Instead, we find a constant step size η, called **the learning rate**. eta is one of many hyperparameters (constant chosen heuristically because they work)

Consider the Training

$$\min_{w^l, b^l, l=2,\cdots,L} \frac{1}{N}\sum_{j=1}^{n}\frac{1}{2}||y(x^j) - F(x^j)||_2^2$$

The gradient is

$$\frac{1}{N}\sum_{j=1}^{n}\nabla(\frac{1}{2}||y(x^j) - F(x^j)||_2^2)$$

where ∇ is w.r.t. $w^2, b^2, \cdots, w^L, b^L$

Note L is relatively large, so consider the **Stochastic Gradient Descent**, the gradient is

$$\frac{1}{|S|}\sum_{j\in S}\nabla(\frac{1}{2}||y(x^j)-F(x^j)||_2^2)$$
$$\text{where } S\subseteq\{1,\cdots,N\}$$

• Single-Sample ($|S|=1$) or Mini-batch ($|S|>1$)

• Either done with repetitions (at each iteration, choose a random S)

• Or done without repetitions : $\{1,\cdots,N\}$ is partitioned into disjoint subsets $S^1, S^2,\cdots$ and iterations cycle through these subsets

11.3 Backpropagation

Problem : For a given $j\in\{1,\cdots,N\}$, compute

$$\frac{\partial}{\partial w_{ik}^l}\frac{1}{2}||y(x^j)-F(x^j)||_2^2, \forall l,i,k$$
$$\frac{\partial}{\partial b_{ik}^l}\frac{1}{2}||y(x^j)-F(x^j)||_2^2, \forall l,i,k$$

Let's denote

$$\begin{aligned} y &= y(x^j) \\ a^l &= \text{ output of layer } l \\ a^L &= \text{ output of last layer} \\ z^l &= w^l a^{l-1} b^l \text{ \#weighted input of layer } l \end{aligned}$$

Thus,

$$a^l=\sigma(z^l)$$

We define

$$\delta_i^l := \frac{\partial}{\partial z_i^l}\frac{1}{2}||y-a^l||^2$$

Lemma 11.3.1 (Last layer)
$\delta_i^L=\sigma'(z_i^L)\cdot(a_i^L-y_i)$ # note the derivative of σ

Proof

$$\delta_i^L = \frac{\partial}{\partial z_i^L}\frac{1}{2}||y - a^L||^2$$
$$= \frac{\partial}{\partial a_i^L}\frac{1}{2}||y - a^L||^2 \cdot \frac{\partial a_i^L}{\partial z_i^L} \text{ (Chain Rule)}$$
$$\frac{\partial}{\partial a_i^L}\frac{1}{2}||y - a^L||^2 = \frac{\partial}{\partial a_i^L}\frac{1}{2}\sum_k (y_k - a_k^L)^2$$
$$= \sum_k \frac{\partial}{\partial a_i^L}\frac{1}{2}(y_k - a_k^L)^2$$
$$= -(y_i - a_i^L)$$

$$a_i^L = \sigma(z_i^L)$$
$$\frac{\partial a_i^L}{\partial z_i^L} = \sigma'(z_i^L)$$

Together, get

$$\delta_i^L = -(y_i - a_i^L) \cdot \sigma'(z_i^L)$$
$$= \sigma'(z_i^L) \cdot (a_i^L - y_i)$$

Lemma 11.3.2 (Other smaller layer)

$\delta_i^l = \sigma'(z_i^l)[(w^{l+1})^T \delta^{l+1}]_i$

Proof

$$\delta_i^l = \frac{\partial}{\partial z_i^l}\frac{1}{2}||y - a^l||^2$$
$$= \frac{\partial}{\partial z_i^l}\frac{1}{2}\sum_k (y_k - a_k^l)^2$$
$$= \sum_k \frac{\partial}{\partial z_i^l}\frac{1}{2}(y_k - a_k^l)^2$$
$$= \sum_k \frac{\partial}{\partial z_i^{l+1}}\frac{1}{2}(y_k - a_k^l)^2 \cdot \frac{\partial z_i^{l+1}}{\partial z_i^l} \text{ (Chain Rule)}$$
$$z_k^{l+1} = \sum_s w_{ks}^{l+1}\sigma(z_s^l) + b_k^{l+1}$$
$$\Rightarrow \quad \frac{\partial z_i^{l+1}}{\partial z_i^l} = w_{ki}^{l+1}\sigma'(z_i^l)$$

Together, get

$$\delta_i^l = \sum_k \delta_k^{l+1} \cdot w_{ki}^{l+1}\sigma'(z_i^l)$$

Lemma 11.3.3
$\frac{\partial}{\partial b_i^l}\frac{1}{2}||y-a^L||^2=\delta_i^l$

Proof

$$\frac{\partial}{\partial b_i^l}\frac{1}{2}||y-a^L||^2=\underbrace{\frac{\partial}{\partial z_i^l}\frac{1}{2}||y-a^L||^2}_{\delta_i^l}\cdot\frac{\partial z_i^l}{\partial b_i^l}\ \text{(Chain Rule)}$$

$$z_i^l=\left(w^l(\sigma(z^{l-1})\right)_i+b_i^l$$

$$\frac{\partial z_i^l}{\partial b_i^l}=1$$

$$\text{So}\quad \frac{\partial}{\partial b_i^l}\frac{1}{2}||y-a^L||^2=\delta_i^l$$

Lemma 11.3.4
$\frac{\partial}{\partial w_{sk}}\frac{1}{2}||y-a^L||_2^2=\delta_s^l\cdot a_k^{l-1}$

Proof

$$\begin{aligned}\frac{\partial}{\partial w_{sk}}\frac{1}{2}||y-a^L||_2^2&=\sum_i\frac{\partial}{\partial w_{sk}}\frac{1}{2}(y_i-a_i^L)^2\\&=\sum_i\underbrace{\frac{\partial}{\partial z_i^l}\frac{1}{2}(y_i-a_i^L)^2}_{\delta_i^l}\cdot\frac{\partial z_i^l}{\partial w_{sk}}\end{aligned}$$

$$\begin{aligned}z_i^l&=\left(w^l\sigma(z^{l+1})\right)_i+b_i^l\\&=[\sum_k w_{ik}^l\underbrace{\sigma(z_k^{l-1}}_{a_k^{l-1}}]+b_i^l\\&=[\sum_k w_{ik}^l a^{l-1}]+b_i^l\end{aligned}$$

$$\text{So}\quad\frac{\partial z_i^l}{\partial w_{sk}}=\begin{cases}0, & \forall s\neq i\\ a_k^{l-1}, & s=i\end{cases}$$

$$\frac{\partial}{\partial w_{sk}}\frac{1}{2}||y-a^L||_2^2=\sum_i\delta_i^l\cdot\frac{\partial z_i^l}{\partial w_{sk}}=\delta_s^l a_k^{l-1}$$

11.4 Summary

(1) $\delta_i^L=\sigma'(z_i^L)\cdot(a_i^L-y_i)$

(2) $\delta_i^l=\sigma'(z_i^l)[(w^{l+1})^T\cdot\delta^{l+1}]_i,\ \forall l=2,\cdots,L-1$

(3) $\frac{\partial}{\partial b_i}\frac{1}{2}||y-a^L||^2=\delta_i^l$

(4) $\frac{\partial}{\partial w_{sk}} \frac{1}{2} ||y - a^L||^2 = \delta_s^l \cdot a_k^{l+1}$

Chapter 12

Course Summary for Final

Chapter 2 : psd/pd matrices

Chapter 3 : Convexity, Strong Convexity (Chapter 7)

Chapter 4 : Constrained Optimization, show a function is coercive

$$\nabla f(x) = 0, \nabla^2 f(x)\, p.d. \Rightarrow x \text{ strict local min} \Rightarrow x \text{ local min} \Rightarrow \nabla^2 f(x)\, p.s.d$$

Chapter 5 : Quadratic Optimization, Newton's Method

Chapter 6 : Least Square Problem (Direct application of Chapter 5)

Chapter 7 : Descent Algorithms, Newton's Method Convergence,
Steepest Direction(= opposite of gradient)

Chapter 8 : Trust Region Methods, Trust Region Subproblem

Chapter 9 : Constrained Optimization, KKT conditions

Chapter 10 : Constrained Optimization Algorithms, Conic Optimization, Looking for Dual

Chapter 11 : Neural Network, no proofs on final, maybe some T/F