# Study Notes of Numerical Optimization

Pei Zhong

Update on November 3, 2023

# Contents

# Update progress

- writing ch3… 2023.10.20

# Preface

all the codes in the notes are completed by python. You can get the codes in the following website:

- -

# Chapter 1

# Preliminary Knowledge

The goal of this chapter is to refresh your memory on some topics in linear algebra and multivariable calculus that will be relevant to the following content. You can use this as a reference throughout the semester.

## 1.1 Matrix Differentiation

Definition 1.1 vector

A vector is a matrix with only one column. Thus, all vectors are inherently column vectors.

**Remark.** We follows the convention of the gradient being a column vector, while the derivative is a row vector.

Definition 1.2 Jacobian Matrix

Let $y = f(x)$ where $y$ is an $m$-element vector, and $x$ is an $n$-element vector. The symbol

$$\frac{\partial y}{\partial x} = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \dots & \frac{\partial y_1}{\partial x_n} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \dots & \frac{\partial y_2}{\partial x_n} \\ \dots & \dots & \dots & \dots \\ \frac{\partial y_m}{\partial x_1} & \frac{\partial y_m}{\partial x_2} & \dots & \frac{\partial y_m}{\partial x_n} \end{pmatrix} \tag{1.1}$$

will denote the $m \times n$ matrix of first-order partial derivatives of the function from $x$ to $y$. Such a matrix is called the Jacobian martix of the function $f$. ($J_{ij} = \frac{\partial y_i}{\partial x_j}$)

**Remark.** If $x$ is actually a scalar in (1.1), then the resulting Jacobian matrix is a $m \times 1$ matrix; that is, a single column(a vector). On the other hand, if $y$ is actually a scalar in (1.1) , then the resulting Jacobian matrix is a $1 \times n$ matrix; that is, a single row (the transpose of a vector).

**Remark.** We follows the convention of $\frac{\partial y^T}{\partial x} = (\frac{\partial y}{\partial x})^T$

Proposition 1.1

# Part I

# Convexity & Unconstrained Optimization

# Chapter 2

# Unconstrained Optimization and Optimality Conditions

## 2.1 Reference

- lecture notes from princeton university

# Chapter 3

# Convex Optimization

*"The great watershed in Optimization is not between linearity and nonlinearity, but convexity and nonconvexity. "*

## 3.1 General Convex Optimization Problems

Definition 3.1 (Convex Set)

a set $\Omega \subset \mathbb{R}^n$ is convex if

$$\forall x, y \in \Omega, t \in [0,1] : x + t(y - x) \in \Omega. \tag{3.1}$$

**Remark.** $x + t(y-x) \in \Omega \iff ty + (1-t)x \in \Omega$. As $x, y$ are arbitrary points, we can say $\Omega$ is convex if $\forall x, y \in \Omega, t \in [0,1] : tx + (1-t)y \in \Omega$.

**Remark.** This definition is equivalent to saying that all connecting lines lie inside set.

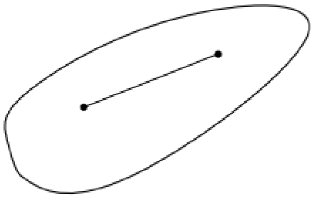

Figure 3.1: an example of a convex set

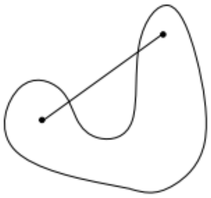

Figure 3.2: an example of a non convex set

Definition 3.2 (Convex Function)

a function $f : \Omega \to \mathbb{R}$ is convex, if $\Omega$ is convex and if

$$\forall x, y \in \Omega, t \in [0,1] : f(x + t(y-x)) \leq f(x) + t(f(y) - f(x)). \tag{3.2}$$

**Remark.** $f(x+t(y-x)) \leq f(x)+t(f(y)-f(x)) \iff f(ty+(1-t)x) \leq tf(y)+(1-t)f(x)$. As $x, y$ are arbitrary points, we can say $f$ is convex function if $\forall x, y \in \Omega, t \in [0,1] : f(tx+(1-t)y) \leq tf(x)+(1-t)f(y)$.

**Remark.** This definition is equivalent to saying that all secants are above graph.

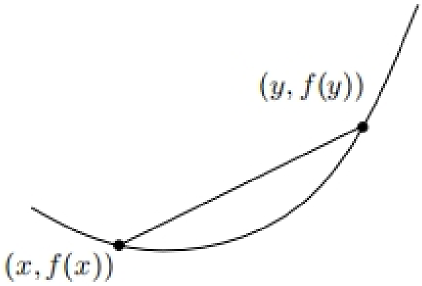

Figure 3.3: For a convex function, the line segment between any two points on the graph lies above the graph

Definition 3.3 (Convex Optimizaiton Problem)

an optimization problem with

- a convex feasible set $\Omega$ and
- a convex objective function $f : \Omega \to \mathbb{R}$

is called a ”convex Optimization problem”

Theorem 3.1 (Local Implies Global Optimality for Convex Problems)

for a convex Optimization problem, every local minimum is also a global one.

*Proof.* Consider a local minimum $x^*$ of the convex optimization problem

$$\min_{x \in \mathbb{R}^n} f(x)$$
$$s.t. x \in \Omega$$

We will show that for each y $\in \Omega$ it holds $f(y) \geq f(x^*)$.

Suppose that $x^*$ is not the global minmium, that is $\exists\, \widetilde{x} \in \Omega \; s.t. \; f(\widetilde{x}) < f(x^*)$.

Consider the line segement $x(t) = tx^* + (1-t)\widetilde{x}, t \in [0,1]$, noting that $x(t) \in \Omega$ by the convexity of $\Omega$. By the convexity of $f$,

$$f(x(t)) \leq tf(x^*) + (1-t)f(\widetilde{x}) < tf(x^*) + (1-t)f(x^*) = f(x^*), \forall t \in [0,1]$$

As $x^*$ is a local minmium, we know that $\exists N$($N$ is a neighbourhood of $x^*$), $\forall x \in N$,

$f(x) \geq f(x^*)$. We can pick $t$ sufficiently to 1 such that $x(t) \in N$. Then $f(x(t)) \geq f(x^*)$. This is a contradiction as $f(x(t)) < f(x^*)$ by the above inequality.

Hence, $x^*$ is the global minimum. □

## 3.2 How to Check Convexity of Functions?

**Theorem 3.2** (Convexity for $C^1$ Functions)

Assume that $f : \Omega \to \mathbb{R}$ is continuously differentiable and $\Omega$ is convex. Then it holds that $f$ is convex if and only if

$$\forall x, y \in \Omega : f(y) \geq f(x) + (\nabla f(x))^T (y - x) \tag{3.3}$$

$i.e.$ tangents lie below the graph.

**Remark.** The gradient (or gradient vector field) of a scalar function $f(x_1, x_2, x_3, \ldots, x_n)$ is denoted $\nabla f$ and is a row vector.

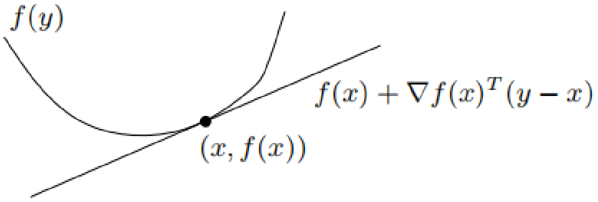

Figure 3.4

*Proof.* "⇒": If $f$ is convex, by definition, $\forall x \in (0, 1]$,

$$f(x + t(y - x)) \leq f(x) + t(f(y) - f(x))$$
$$\implies f(y) - f(x) \geq \frac{f(x + t(y - x)) - f(x)}{t(y - x)} \cdot (y - x)$$

As $t \to 0$, we get

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

"⇐": Take any $x, y \in \Omega$ , $t \in [0, 1]$ and let

$$z = tx + (1 - t)y.$$

As $\Omega$ is convex, we have $z \in \Omega$. Then, by (3.3)

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

Then,

$$\begin{aligned} tf(x) + (1-t)f(y) &\geq tf(z) + t(\nabla f(z))^T(x-z) + (1-t)(f(z) + (1-t)(\nabla f(z))^T(y-z)) \\ &= f(z) + (\nabla f(z))^T(tx + (1-t)y - z) \\ &= f(z) + (\nabla f(z))^T(z-z) \\ &= f(tx + (1-t)y). \end{aligned}$$

By definition3.2, $f$ is a convex function. □

**Definition 3.4** (Generalized inequality for Symmetric Matrices)

$B$ is a symmetric matrix in $\mathbb{R}^{n\times n}$. Define "$B \succcurlyeq 0$" if and only if $B$ is positive semi-definite. i.e.

$$\begin{aligned} B \succcurlyeq 0 &\iff \forall z \in \mathbb{R}^n : z^T B z \geq 0 \\ &\iff \min\{eig(B)\} \geq 0 \end{aligned}$$

**Remark.** $B \succcurlyeq 0 \iff$ all eigenvalues of $B$ are non-negative real value. $B$ is a symmetric matrix $\Rightarrow$ all eigenvalues of $B$ is real value.

**Theorem 3.3** (Convexity for $C^2$ Functions)

Assume that $f : \Omega \to \mathbb{R}$ is twice continuously differentiable and $\Omega$ is convex and open. Then it holds that $f$ is convex if and only if for all $x \in \Omega$ the Hessian is positive semi-definite, i.e.

$$\forall x \in \Omega : \nabla^2 f(x) \succcurlyeq 0. \tag{3.4}$$

*Proof.* Using Taylor expansion. □

**Theorem 3.4**

Consider an unconstrained optimization problem

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

where $f : \mathbb{R}^n \to \mathbb{R}$ is convex and differentiable. Then,

$$\nabla f(x^*) = 0 \iff f(x) \geq f(x^*), \forall x \in dom(f).$$

*Proof.* □

Theorem 3.5

Consider an optimization problem

$$\begin{aligned} \min f(x) \\ s.t.\ x \in \Omega, \end{aligned}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is convex and differentiable and $\Omega$ is convex. Then for any $x^* \in \Omega$,

$$\nabla f(x^*)(y - x^*) \geq 0, \forall y \in \Omega \iff f(x) \geq f(x^*), \forall x \in \Omega.$$

*Proof.* □

# Chapter 4

# Gradient Desent

## 4.1 Motivation

Let's recall our unconstrained optimization problem:

$$\min_x f(x),$$

where $f : \mathbb{R}^n \to \mathbb{R}$.
**Where we stand so far:**

- Learned about some structural properties of local optimal solutions (first and second order conditions for optimality).
- Learned that for convex problems, local optima are automatically global.

But how to find a local optimum? How to even find a stationary point (i.e., a point where the gradient vanishes)? Recall that this would suffice for global optimality if is convex. We now begin to see some algorithms for this purpose, starting with gradient descent algorithms. These will be iterative algorithms: start at a point, jump to a new point that hopefully has a lower objective value and continue.

## 4.2 Iterative Algorithms Overview

### 4.2.1 Iterative Form

**General form of the iterations:**

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

where,

- $k \in \mathbb{Z}_+$ : iteration number
- $x_k \in \mathbb{R}^n$ : current point
- $d_k \in R^n$ : direction to move along at iteration $k$

- $x_{k+1} \in \mathbb{R}^n$: next point
- $\alpha_k \in \mathbb{R}_+$: step size at iteration $k$.

Goal is to make the sequence $\{f(x_k)\}$ decrease as much as possible. But how to choose $d_k$? How to choose $\alpha_k$? In Gradient Desent method, the direction to move along at step is chosen ”based on information from” $\nabla f(x)$”. Why is $\nabla f(x)$ a natural vector to look at? Lemmas 4.1 and 4.2 below provide two reasons.

Lemma 4.1

$f \in C^1$. Consider yourself sitting at a point $x \in \mathbb{R}^n$ and looking (locally) at the value of the function $f$ in all directions around you. The direction with the maximum rate of decrease is along $-\nabla f(x)$.

**Remark.** When we speak of direction, the magnitude of the vector does not matter; e.g., $\nabla f(X), 5\nabla f(x), \frac{\nabla f(x)}{20}$ all are in the same direction.

*Proof.* Consider a point $x$, a direction $d$, and the univariate function

$$g(\alpha) = f(x + \alpha \frac{d}{||d||}).$$

As the rate of change of $f$ at $x$ in direction $d$(derivative in direction $d$) is

$$\begin{aligned}\nabla_d f(x) &= \lim_{\alpha \to 0} \frac{f(x + \alpha \frac{d}{||d||}) - f(x)}{\alpha} = \lim_{\alpha \to 0} \frac{g(\alpha) - g(0)}{\alpha} = g'(0) \\ &= (\nabla f(x))^T \frac{d}{||d||} = \frac{1}{||d||} < \nabla f(x), d >,\end{aligned}$$

by the Cauchy-Schwarz inequality, we have:

$$-\frac{d}{||d||} \cdot ||\nabla f(x)|| \cdot ||d|| \leq \frac{1}{||d||} < \nabla f(x), d > \leq \frac{d}{||d||} \cdot ||\nabla f(x)|| \cdot ||d||,$$

which after simplifying gives

$$-||\nabla f(x)|| \leq \frac{1}{||d||} < \nabla f(x), d > \leq ||\nabla f(x)||.$$

So $\nabla_d f(x)$ cannot be larger than $||f(x)||$, or smaller than $-||f(x)||$. However, if we take $d = \nabla f(x)$, the right inequality is achieved:

$$\frac{1}{||\nabla f(x)||} < \nabla f(x), \nabla f(x) > = \frac{1}{||\nabla f(x)||} \cdot ||\nabla f(x)||^2 = ||\nabla f(x)||.$$

Similarly, if we take $d = -\nabla f(x)$, then the left inequality is achieved. □

Definition 4.1

For a given point $x \in R^n$, a direction $d \in \mathbb{R}^n$ is called a desent direction, if there exists $\bar{\alpha} > 0$ such that

$$f(x + \alpha d) < f(x), \forall \alpha \in (0, \bar{\alpha})$$

Lemma 4.2

Consider a point $x \in \mathbb{R}^n$. Any direction $d$ satisfying

$$< \nabla f(x), d >< 0$$

is a descent direction.(In particular, $-\nabla f(x)$ is a desent direction.)

**Remark.** The condition $< \nabla f(x), d >< 0$ in Lemma 4.2 geometrically means that the vectors and make an angle of more than 90 degrees (on the plane that contains them).
Why?(Hint: $< \nabla f(x), d >= ||\nabla f(x)|| \cdot ||d|| \cdot cos\theta$)

*Proof.* By Taylor's theorem, we have

$$f(x + \alpha d) = f(x) + \alpha \nabla f(x)^T d + o(\alpha).$$

Since $\lim_{\alpha \to 0} \frac{|o(\alpha)|}{\alpha} = 0$, there exists $\bar{\alpha} > 0$ such that

$$\frac{|o(\alpha)|}{\alpha} < |\nabla f(x)^T d|, \forall \alpha \in (0, \bar{\alpha}).$$

This, together with our assumption that $\nabla f(x)^T d < 0$, implies that $\forall \alpha \in (0, \bar{\alpha})$ we must have:

$$\begin{aligned} f(x + \alpha d) &< f(x) + \alpha \nabla f(x)^T d + \alpha |\nabla f(x)^T d| \\ &< f(x) + \alpha \nabla f(x)^T d - \alpha \nabla f(x)^T d \\ &< f(x) \end{aligned}$$

Hence, by Definition 4.1 , $d$ is a descent direction. □

Lemma 4.3

Consider any positive definite matrix $B$. For any point $x$, with $\nabla f(x) \neq 0$, the direction $-B\nabla f(x)$ is a descent direction.

*Proof.* By the assumption that $B$ is positive define, we have

$$< \nabla f(x), -B\nabla f(x) >= -\nabla f(x)^T B \nabla f(x) < 0.$$

□

Lemma 4.3 suggets a general paradigm for our descent algorithms:

$$x_{k+1} = x_k - \alpha_k B_k \nabla f(x_k), (with B_k \succ 0, \forall k).$$

**Common choices of descent direction:**

- Steepest Descent: $B_k = I, \forall k$.

  Simplest descent direction but not always the fastest.

- Newton Direction: $B_k = (\nabla^2 f(x_k))^{-1}$, assuming $\nabla^2 f(x) \succ 0, \forall k$.

  More expensive, but can have much faster convergence.

- Modified Newton Direction: $B_k = (\nabla^2 f(x_0))^{-1}, \forall k$

  Compute Newton direction only at the beginning, or once every M steps.

**Common choices of the step size $\alpha_k$:**

- Constant step size: $\alpha_k = s, \forall k (s > 0)$
  - * Simplest rule to implement, but may not converge if $s$ too large; may be too slow if $s$ too small.
- Diminishing step size: $\alpha_k \to 0$, $\sum\limits_{k=1}^{\infty} \alpha_k = \infty$. (e.g., $\alpha_k = \frac{1}{k}$)
- Minimization rule (exact line search): $\alpha_k = argmin_{\alpha \geq 0} f(x_k + \alpha_k d_k)$
  - * A minimization problem itself, but an easier one (one dimensional).
  - * If $f$ convex, the one dimensional minimization problem also convex
- Successive step size reduction: well-known examples are Armijo rule and Wolf rule. We will cover Armijo in the next chapter.
  - * Try to ensure enough decrease in line search without spending time to solve $\alpha_k$ to optimality.

### 4.2.2 Stopping Criteria

Once we have a rule for choosing the search direction and the step size, we are good to go for running the algorithm. Typically the initial point $x_0$ is picked randomly, or if we have a guess for the location of local minima, we pick $x_0$ close to them. But when to stop the algorithm? Some common choices ( $\epsilon > 0$ is a small prescribed threshold):

- $||\nabla f(x)|| < \epsilon$

  If we have $\nabla f(x_k) = 0$, our iterates stop moving. We have found a point satisfying the first order necessary condition for optimality. This is what we are aiming for.

- $|f(x_{k+1}) - f(x_k)| < \epsilon$

Improvements in function value are saturating.

- $||x_{k+1} - x_k|| < \epsilon$

  Movement between iterates has become small.

## 4.3 Convergence

The descent algorithms discussed so far typically come with proofs of convergence to a stationary point. We state a couple such theorems here; the proofs are a bit tedious, and I won't require you to know them. But those interested can look some of these proofs up in the information I remark.

Theorem 4.1

Assume that $f : \mathbb{R}^n \to \mathbb{R}$ is convex and differentiable, and additionally

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

i.e. $\nabla f$ is Lipschitz continuous with constant $L > 0$. Then Gradient Descent with fixed step size $\alpha \leq \frac{1}{L}$ satisfies

$$f(x_k) - f(x^*) \leq \frac{||x_0 - x^*||^2}{2\alpha k}.$$

**Remark.** proof reference: notes from cmu

## 4.4 Rates of Convergence

Once we know an iterative algorithm converges, the next question is how fast?

Definition 4.2

Let $\{x_k\}$ converge to $x^*$. We say the convergence is of order $p(\geq 1)$ and with factor $\gamma(> 0)$, if $\exists k_0$ such that $\forall k \geq k_0$,

$$||x_{k+1} - x^*|| \leq \gamma||x_k - x^*||^p.$$

Make sure the following comments make sense:

## 4.5 Reference

- lecture notes from princeton university

# Chapter 5

# Line Search

## 5.1 Reference

- lecture note from washington university
- matlab optimizaiton toolbox tutorial

# Part II

# Duality