

Jinchao Xu

Numerical Optimization (MATH/CSE 555)

Spring 2021

Contributors:

This set of notes are based on contributions from many of graduate students, post-doctoral fellows and other collaborators. Here is a partial list:

Qingguo Hong, Limin Ma, Jonathan Siegel

Contents

1	Preliminaries	7
1.1	Examples of optimization problems	7
1.2	Basic facts of calculus	8
1.2.1	Optimality Condition	8
1.2.2	Convex function	11
2	Gradient descent method	15
2.1	Line search and gradient descent method	15
2.1.1	Gradient descent method	15
2.1.2	Convergence of Gradient Descent method	18
3	Quadratic optimization and linear systems of equations	21
3.1	Basic linear iterative methods	21
3.2	Richardson, Jacobi and Gauss-Seidel methods	21
3.3	Conjugate gradient methods and preconditioning	21
3.3.1	Convergence rate in terms of condition number	21
3.3.2	Convergence rate $O(k^{-2})$	21
3.4	Method of subspace corrections	21
4	Examples: Logistic Regressions and Deep Learning	23
4.1	Definition of linearly separable sets	23
4.1.1	Binary classification	23
4.1.2	Multi-class classification	24
4.1.3	Geometric interpretation for multi-label cases ($k > 2$)	25
4.1.4	Two more definitions of linearly separable sets	25
4.1.5	Comparison of different definitions of linearly separable sets	26
4.2	Introduction to logistic regression	29
4.2.1	Logistic regression	29
4.2.2	Regularized logistic regression	33
4.3	KL divergence and cross-entropy	36
4.3.1	KL(Kullback–Leibler) divergence	36

4.3.2	Cross-entropy	37
4.4	Deep learning	38
4.5	MNIST	38
5	Smooth Convex Optimization	39
5.1	Nesterov acceleration methods	39
6	Stochastic Gradient Descent Methods	41
6.1	Stochastic gradient descent method and convergence theory	42
6.1.1	Convergence of SGD	42
6.1.2	SGD with mini-batch	43
6.1.3	Space decomposition versus function decomposition	45
6.1.4	Connection between space decomposition and function decomposition	47
6.2	Practical Training Method in Large-Scale Machine Learning Problems	48
6.2.1	Basic Gradient Descent Type Algorithms	48
6.2.2	Deterministic Algorithm(a specially case): Incremental Gradient Descent Method	49
6.2.3	Convergence for CIGD	50
6.2.4	Some comments	51
6.3	Adam, Momentum etc	53
6.3.1	Adaptive Learning Rate Mini-batch Based Method	53
6.3.2	For linear problems	57
6.3.3	Choosing the Right Optimization Algorithm	59
6.4	Incremental Methods	59
6.4.1	SPD Quadratic problems	59
6.5	General problems	59
6.6	Stochastic gradient descent method	59
6.7	Stochastic gradient descent method and convergence theory	60
6.7.1	Convergence of SGD	60
6.7.2	SGD with mini-batch	61
6.7.3	Space decomposition versus function decomposition	63
6.7.4	Connection between space decomposition and function decomposition	65
6.8	Practical Training Method in Large-Scale Machine Learning Problems	66
6.8.1	Basic Gradient Descent Type Algorithms	66
6.8.2	Deterministic Algorithm(a specially case): Incremental Gradient Descent Method	67
6.8.3	Convergence for CIGD	68
6.8.4	Some comments	69
6.9	Adam, Momentum etc	71
6.9.1	Adaptive Learning Rate Mini-batch Based Method	71
6.9.2	For linear problems	75

6.9.3	Choosing the Right Optimization Algorithm.....	77
6.10	Incremental Methods	77
7	Method of subspace corrections, ADMM.....	79
7.1	Coordinate descent methods.....	79
7.2	Alternating direction method of multipliers (ADMM)	79
8	Unconstraint optimization for non-smooth problems	81
8.1	Subgradient	81
8.2	Soft-thresholding	81
8.3	Forward-Backward Splitting Algorithm	83
8.4	Analysis of Unconstrained Stochastic Gradient Descent.....	84
8.5	Lower Bounds	85
8.6	Douglas-Rachford Splitting	85
9	Unconstraint optimization: second order method	87
9.1	Newton's method and variants	87
9.2	Quasi-Newton Methods: BFGS Method	87
10	Constraint optimization	89
10.1	Necessary and sufficient conditions.....	89
10.2	Conic Constraints	89
10.3	Conic Duality	89
10.4	Linear Programming	89
10.5	Semi-definite Programming	89
10.6	Interior Point Methods	89
11	Homework	91
	References	95

Notation

$$(0.1) \quad \|x\| = \|x\|_2 = \|x\|_{\ell^2} = \left(\sum_{i=1}^n x_i^2 \right)^{\frac{1}{2}}$$

$$(0.2) \quad \|x\|_{\ell^1} = \sum_{i=1}^n |x_i|$$

$$(0.3) \quad \|x\|_p = \left(\sum_{i=1}^n |x_i|^p \right)^{\frac{1}{p}}$$

$$(0.4) \quad \|x\|_{\infty} = \max_{1 \leq j \leq n} |x_j|$$

$$(0.5) \quad o(r) : \quad \lim_{r \rightarrow 0} \frac{o(r)}{r} = 0.$$

Preliminaries

1.1 Examples of optimization problems

Let us start by fixing the mathematical form of our main problem and the standard terminology. Let x be an n -dimensional real vector:

$$x = (x_1, \dots, x_n)^T \in \mathbb{R}^n$$

and $f_1(\cdot), \dots, f_m(\cdot)$ be some real-valued functions defined on a set $\Omega \subseteq \mathbb{R}^n$. In this book, we consider different variants of the following general minimization problem:

$$(1.1) \quad \begin{aligned} & \min f(x) \\ & \text{s.t. } f_j(x) \geq 0, \quad j = 1, \dots, m, \\ & x \in \Omega. \end{aligned}$$

We call $f(\cdot)$ the objective function of our problem, the vector function

$$\mathbf{f}(x) = (f_1(x), \dots, f_m(x))^T$$

is called the vector of functional constraints, the set Ω is called the basic feasible set, and the set

$$\mathcal{F} = \{x \in \Omega \mid f_j(x) \geq 0, j = 1 \dots m\}$$

is called the (entire) feasible set of problem (1.1). We will mainly consider minimization problems in this notes. Instead, we could consider maximization problems with the objective function $-f(\cdot)$.

There exists a natural classification of the types of minimization problems.

- Unconstrained problems: there is no constraint functions $f_j(x)$ in (1.1). Thus,

$$\mathcal{F} = \mathbb{R}^n.$$

- Constrained problems: $\mathcal{F} \subsetneq \mathbb{R}^n$.
- Smooth problems: all $f_j(\cdot)$ and $f(\cdot)$ are differentiable.

- Nonsmooth problems: some components $f_k(\cdot)$ or $f(\cdot)$ are nondifferentiable, say

$$f(x) = \|x\|_{l^1} := \sum_{j=1}^n |x_j|.$$

- Linearly constrained problems: the functional constraints are affine:

$$f_j(x) = \sum_{i=1}^n a_{ij}x_i + b_j \equiv (\mathbf{a}_j, x) + b_j, j = 1 \dots m.$$

Here $\mathbf{a}_j = (a_{1j}, a_{2j}, \dots, a_{nj})$ and (\cdot, \cdot) stands for the inner (or scalar) product in \mathbb{R}^n : $(a, x) = a^T x$, and Ω is a polyhedron. If $f(\cdot)$ is also affine, then (1.1) is a linear optimization problem. If $f(\cdot)$ is quadratic, then (1.1) is a quadratic optimization problem. If all the functions $f(\cdot), \dots, f_m(\cdot)$ are quadratic, then this is a quadratically constrained quadratic problem.

1.2 Basic facts of calculus

To begin with the gradient based optimization, it is necessary to review some multi-variable calculus aspects and definition of convex functions.

1.2.1 Optimality Condition

At the very beginning, let us recall the definition of gradient and Hessian matrix for function $f : \mathbb{R}^n \rightarrow \mathbb{R}$.

Definition 1. Given the objective function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ and $x = (x_1, x_2, \dots, x_n)^T \in \mathbb{R}^n$, the gradient and the Hessian of $f(x)$ are defined by

$$(1.2) \quad g(x) := \nabla f(x) = \begin{pmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{pmatrix} \quad H(x) := \nabla^2 f(x) = \begin{pmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \dots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \frac{\partial^2 f(x)}{\partial x_2^2} & \dots & \frac{\partial^2 f(x)}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x)}{\partial x_n \partial x_1} & \frac{\partial^2 f(x)}{\partial x_n \partial x_2} & \dots & \frac{\partial^2 f(x)}{\partial x_n^2} \end{pmatrix}.$$

Note that the Hessian of $f(x)$ is a symmetric function in $\mathbb{R}^{n \times n}$.

Let us introduce one often-used inequality.

Lemma 1 (Cauchy-Schwarz inequality). For any $\mathbf{p} = (p_1, \dots, p_n)^T$ and $\mathbf{q} = (q_1, \dots, q_n)^T$, we have

$$(1.3) \quad \left(\sum_{i=1}^n p_i q_i \right)^2 \leq \left(\sum_{i=1}^n p_i^2 \right) \left(\sum_{i=1}^n q_i^2 \right),$$

where equality holds if and only if for some $k \in \mathbb{R}$, $\frac{p_i}{q_i} = k$, or in inner form:

$$(1.4) \quad |(\mathbf{p}, \mathbf{q})| \leq \|\mathbf{p}\| \cdot \|\mathbf{q}\|.$$

Proof. In the case of $\mathbf{q} = \mathbf{0}$, the inequality holds. Now assume $\mathbf{q} \neq \mathbf{0}$, then

$$(1.5) \quad \begin{aligned} \|\mathbf{p} - \lambda \mathbf{q}\|^2 &= (\mathbf{p}, \mathbf{p}) - 2\lambda(\mathbf{p}, \mathbf{q}) + (\lambda \mathbf{q}, \lambda \mathbf{q}) \\ &= \|\mathbf{p}\|^2 - 2\lambda(\mathbf{p}, \mathbf{q}) + \lambda^2 \|\mathbf{q}\|^2, \end{aligned}$$

Due to that $\|\mathbf{p} - \lambda \mathbf{q}\|^2 \geq 0$ for any $\lambda \in \mathbb{R}$ and the discriminant of root of quadratic equation,

$$(1.6) \quad \Delta(\lambda) = 4(\mathbf{p}, \mathbf{q})^2 - 4\|\mathbf{p}\|^2\|\mathbf{q}\|^2 \leq 0$$

If the inequality holds as an equality, which means there exists one $k \in \mathbb{R}$ such that $\|\mathbf{p} - k\mathbf{q}\|^2 = 0$, or so-called \mathbf{p} and \mathbf{q} are linearly dependent. Conversely, if \mathbf{p} and \mathbf{q} are linearly dependent, then there is only one solution $\lambda = k$ to equation $\|\mathbf{p} - \lambda \mathbf{q}\|^2 = 0$, therefore, $\Delta(\lambda) = 0$. \square

The next statement is probably the most fundamental fact in optimization theory.

Definition 2. For any real-valued function f defined on a domain Ω ,

1. if $f(x^*) \leq f(x)$ for all x in Ω , $f(x)$ has a global minimum point at x^* ;
2. if $f(x^*) \leq f(x)$ for x near x^* in Ω , $f(x)$ has a local minimum point at x^* ;
3. if $f(x^*) < f(x)$ for all x in Ω , $f(x)$ has a strict global minimum point at x^* ;
4. if $f(x^*) < f(x)$ for x near x^* in Ω , $f(x)$ has a strict local minimum point at x^* .

Theorem 1 (First-Order Optimality Condition). Let x^* be a local minimum of a differentiable function $f(\cdot)$. Then

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

Proof. Since x^* is a local minimum of $f(\cdot)$, there exists an $r > 0$ such that for all $y \in \mathbb{R}^n$, $\|y - x^*\| \leq r$, we have $f(y) \geq f(x^*)$. Since f is differentiable, this implies that

$$f(y) = f(x^*) + (\nabla f(x^*), y - x^*) + o(\|y - x^*\|) \geq f(x^*).$$

Thus, for all $s \in \mathbb{R}^n$, we have

$$(\nabla f(x^*), s) \geq 0.$$

By taking $s = -\nabla f(x^*)$, we get

$$-\|\nabla f(x^*)\|^2 \geq 0.$$

Hence, $\nabla f(x^*) = 0$. \square

In what follows the notation $B \geq 0$, where $B = (b_{ij})$ is a symmetric $(n \times n)$ -matrix, means that B is positive semidefinite:

$$(Bx, x) = \sum_{i,j=1}^n b_{ij}x_i x_j \geq 0 \quad \forall x \in \mathbb{R}^n.$$

The notation $B > 0$ means that B is symmetric positive definite (SPD for short hereinafter), namely there exists some $\lambda > 0$ such that

$$(Bx, x) = \sum_{i,j=1}^n b_{ij}x_i x_j > \lambda \sum_{i,j=1}^n x_i x_j = \lambda \|x\|_2^2, \quad \forall x \in \mathbb{R}^n.$$

Using the second-order approximation, there exist the following second-order optimality conditions.

Theorem 2 (Second-Order Optimality Condition). *Let x^* be a local minimum of a twice differentiable function $f(\cdot)$. Then*

$$\nabla f(x^*) = 0, \quad \nabla^2 f(x^*) \geq 0.$$

Proof. Since x^* is a local minimum of the function $f(\cdot)$, there exists an $r > 0$ such that for all y , $\|y - x^*\| \leq r$, we have

$$f(y) \geq f(x^*)$$

In view of Theorem 1.2.1, $\nabla f(x^*) = 0$. Therefore, for any such y ,

$$f(y) = f(x^*) + \left(\nabla^2 f(x^*)(y - x^*), y - x^* \right) + o(\|y - x^*\|^2) \geq f(x^*).$$

Thus, $\left(\nabla^2 f(x^*)s, s \right) \geq 0$, for all $\|s\| = 1$. \square

Again, the above theorems are necessary (second-order) characteristic of a local minimum. Let us prove now a sufficient condition.

Theorem 3. *Let a function $f(\cdot)$ be twice differentiable on \mathbb{R}^n and let $x^* \in \mathbb{R}^n$ satisfy the following conditions.*

$$\nabla f(x^*) = 0, \quad \nabla^2 f(x^*) > 0.$$

Then x^ is a strict local minimum of $f(\cdot)$.*

Proof. Note that in a small neighborhood of a point x^* the function $f(\cdot)$ can be represented as

$$f(y) = f(x^*) + \frac{1}{2} \left(\nabla^2 f(x^*)(y - x^*), y - x^* \right) + o(\|y - x^*\|^2).$$

Since $\frac{o(r^2)}{r^2} \rightarrow 0$ as $r \downarrow 0$, there exists a value $\bar{r} > 0$ such that for all $r \in [0, \bar{r}]$ we have

$$\left| o(r^2) \right| \leq \frac{r^2}{4} \lambda_{\min}(\nabla^2 f(x^*)).$$

In view of our assumption, this eigenvalue is positive. Therefore, for any $y \in \mathbb{R}^n$, $0 < \|y - x^*\| \leq \bar{r}$, we have

$$\begin{aligned} f(y) &\geq f(x^*) + \frac{1}{2} \lambda_{\min}(\nabla^2 f(x^*)) \|y - x^*\|^2 + o(\|y - x^*\|^2) \\ &\geq f(x^*) + \frac{1}{4} \lambda_{\min}(\nabla^2 f(x^*)) \|y - x^*\|^2 > f(x^*). \end{aligned}$$

□

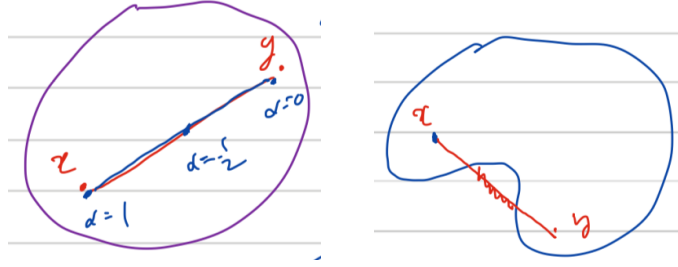
1.2.2 Convex function

Let us first give the definition of convex sets.

Definition 3 (Convex set). A set C is convex, if the line segment between any two points in C lies in C , i.e., if any $x, y \in C$ and any α with $0 \leq \alpha \leq 1$, there holds

$$(1.7) \quad \alpha x + (1 - \alpha)y \in C.$$

Here are two diagrams for this definition about convex and non-convex sets.



Following the definition of convex set, we define convex function as following.

Definition 4 (Convex function). Let $C \subset \mathbb{R}^n$ be a convex set and $f : C \rightarrow \mathbb{R}$:

1. f is called **convex** if for any $x, y \in C$ and $\alpha \in [0, 1]$

$$(1.8) \quad f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha)f(y).$$

2. f is called **strictly convex** if for any $x \neq y \in C$ and $\alpha \in (0, 1)$:

$$(1.9) \quad f(\alpha x + (1 - \alpha)y) < \alpha f(x) + (1 - \alpha)f(y).$$

3. A function f is said to be (strictly) **concave** if $-f$ is (strictly) convex.