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

Notation

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

$$\|x\|_{l^1} = \sum_{i=1}^{n} |x_i| \tag{0.2}$$

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

$$\|x\|_\infty = \max_{1\le j\le n} |x_j| \tag{0.4}$$

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

1

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:

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

and $f_1(\cdot), \ldots, f_m(\cdot)$ 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:

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

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

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

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

$$\mathscr{F} = \left\{ x \in \Omega \mid f_j(x) \geq 0, j = 1 \ldots m \right\}$$

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,

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

- Constrained problems: $\mathscr{F} \subsetneq \mathbb{R}^n$.
- 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:

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

  Here $\mathbf{a}_j = (a_{1j}, a_{2j}, \cdots, a_{nj})$ 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 $f(\cdot), \cdots, f_m(\cdot)$ 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 multivariable 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 $f : \mathbb{R}^n \to \mathbb{R}$.

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

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

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 $\mathbf{p} = (p_1, \ldots, p_n)^T$ and $\mathbf{q} = (q_1, \ldots, q_n)^T$, we have*

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

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

$$(1.4)\qquad |(\mathbf{p}, \mathbf{q})| \le \|\mathbf{p}\| \cdot \|\mathbf{q}\|.$$

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

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

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,

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

If the inequality holds as an equality, which means there exists one $k \in \mathbb{R}$ such that $\|\mathbf{p} - \lambda\mathbf{q}\|^2 = 0$, 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, $\Delta(\lambda) = 0$. □

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, $\nabla f(x^*) = 0$. □

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

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

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,

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

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

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

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

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

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

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

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

□

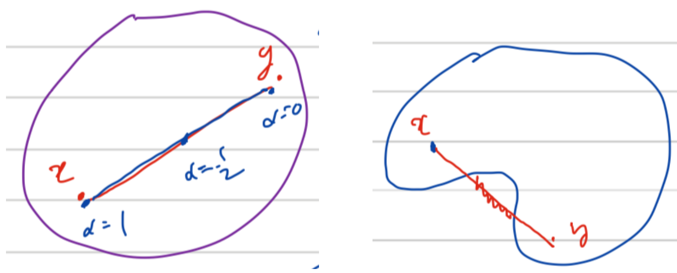
1.2.2 Convex function

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*

$$\alpha x + (1 - \alpha)y \in C. \tag{1.7}$$

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 $f : C \to \mathbb{R}$:*

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

$$f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha)f(y). \tag{1.8}$$

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

$$f(\alpha x + (1 - \alpha)y) < \alpha f(x) + (1 - \alpha)f(y). \tag{1.9}$$

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