Notes on Optimization

Ruichen Luo

November 25, 2022

Contents

1 Convexity

1.1 Inner product and norm

Definition 1.1
We define the **(standard) inner product** $\langle \cdot, \cdot \rangle : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ as

$$\langle u, v \rangle = u^{\mathrm{T}} v,$$

and then refer to the space $\mathbb{R}^d$ as an **Euclidean space**.

Proposition 1.1 (Cauchy-Schwarz inequality)
Let $\mathbb{R}^d$ be an Euclidean space.

$$|\langle u, v \rangle| \leq \sqrt{\langle u, u \rangle \cdot \langle v, v \rangle}, \quad \forall u, v \in \mathbb{R}^d.$$

Definition 1.2
Within vector space $\mathbb{V}$, $\| \cdot \| : \mathbb{V} \to \mathbb{R}$ is a **norm** if, for any $u, v \in \mathbb{V}$ and $\lambda \in \mathbb{R}$,

1. $\|u\| \geq 0$, and it equals 0 iff $u = 0$;
2. $\|\lambda u\| = |\lambda| \cdot \|u\|$;
3. $\|u + v\| \leq \|u\| + \|v\|$.

The vector space with a norm is called **norm space**. The norm space naturally induces a **metric** (or **distance**) by

$$d(u, v) = \|u - v\|.$$

Remark 1.1. *The inner product within Euclidean space can also induce an **Euclidean norm**:*

$$\|u\| = \sqrt{u^{\mathrm{T}} u}, \quad \forall u \in \mathbb{R}^d.$$

*The metric induced by the Euclidean norm is called an **Euclidean distance**, and we also have an **(Euclidean) angle** between $u, v \in \mathbb{R}^d$ given by:*

$$\arccos\left(\frac{\langle u, v \rangle}{\|u\| \cdot \|v\|} \right).$$

Proposition 1.2
For any $u, v \in \mathbb{V}$, $0 \leq \lambda \leq 1$,

$$\lambda \|u\|^2 + (1 - \lambda)\|v\|^2 \geq \lambda(1 - \lambda)(\|u\| + \|v\|)^2 \geq \lambda(1 - \lambda)\|u \pm v\|^2$$

Theorem 1.3 (all norms are equivalent)
Let $\mathbb{V}$ be a finite dimensional vector space.[a] Then for any two norms $\|\cdot\|_A$ and $\|\cdot\|_B$ defined on $\mathbb{V}$, there exists $c, C \in \mathbb{R}_{>0}$ *s.t.*

$$c\|u\|_B \le \|u\|_A \le C\|u\|_B, \quad \forall u \in \mathbb{V}.$$

[a] In infinite dimensional space, this result does not hold.

Proof. See `https://math.mit.edu/~stevenj/18.335/norm-equivalence.pdf`. □

1.2 Convex set and projection

Definition 1.3
A set C is a **convex set** if, for any $x, y \in C$ and $0 \le \lambda \le 1$,

$$\lambda x + (1-\lambda)y \in C.$$

Proposition 1.4
Suppose S is a set of convex sets, *i.e.*, for any $C \in S$, C is a convex set. Then

$$\bigcap_{C \in S} C$$

is a convex set.

Proposition 1.5
$C \subseteq \mathbb{V}$ is a convex set, then $kC = \{kx : x \in C\}$ is also a convex set, where $k \in \mathbb{R}$.

Proposition 1.6
$C, D \subseteq \mathbb{V}$ are convex sets, then $C + D = \{x + y : x \in C, y \in D\}$ is also a convex set.

Definition 1.4
Let the set $Q \subseteq \mathbb{R}^d$ be closed and convex. We define the projection operator $\Pi_Q : \mathbb{R}^d \to Q$ by

$$\Pi_Q(x) = \arg\min_{y \in Q} \|y - x\|,$$

where $\|\cdot\|$ is the Euclidean norm.

Proposition 1.7
For any $x \in Q$ and $y \in \mathbb{R}^d$,

$$\langle \Pi_Q(y) - x, \Pi_Q(y) - y \rangle \leq 0$$

Remark 1.2. *For any $x \neq \Pi_Q(y)$, the Euclidean angle between $\Pi_Q(y) - x$ and $\Pi_Q(y) - y$ is no less than $90°$.*

Proof. When $x = \Pi_Q(y)$, it becomes trivial. Let's consider $x \neq \Pi_Q(y)$.

For any $0 < \lambda \leq 1$, we have $\lambda x + (1-\lambda)\Pi_Q(y) \in Q$, and then

$$\begin{aligned}
\|\Pi_Q(y) - y\|^2 &\leq \|[\lambda x + (1-\lambda)\Pi_Q(y)] - y\|^2 \\
&= \lambda^2 \|x - \Pi_Q(y)\|^2 - 2\lambda \langle x - \Pi_Q(y), y - \Pi_Q(y) \rangle + \|\Pi_Q(y) - y\|^2 \\
&= \lambda \left[\lambda \|x - \Pi_Q(y)\|^2 - 2 \langle x - \Pi_Q(y), y - \Pi_Q(y) \rangle \right] + \|\Pi_Q(y) - y\|^2,
\end{aligned}$$

or

$$\lambda \|x - \Pi_Q(y)\|^2 - 2 \langle x - \Pi_Q(y), y - \Pi_Q(y) \rangle \geq 0.$$

Let $\lambda \downarrow 0$. We must have

$$2 \langle x - \Pi_Q(y), y - \Pi_Q(y) \rangle \leq 0.$$

□

Corollary 1.8
For any $x, y \in \mathbb{R}^d$, we have $\|\Pi_Q(x) - \Pi_Q(y)\| \leq \|x - y\|$.

Corollary 1.9
For any $x \in Q$ and $y \in \mathbb{R}^d$, we have

$$\|x - \Pi_Q(y)\|^2 + \|\Pi_Q(y) - y\|^2 \leq \|x - y\|^2.$$

Theorem 1.10 (separating hyperplane theorem)
Suppose $C, D \subseteq \mathbb{R}^d$ are nonempty disjoint convex sets, *i.e.*, $C \cap D = \emptyset$. Then there exist $v \neq 0$ and b such that $v^{\mathrm{T}} x \leq b$ for all $x \in C$ and $v^{\mathrm{T}} x \geq b$ for all $x \in D$.

Proof. [Wikipedia contributors, 2022] It's equivalent to show that there exists a vector $v \neq 0$, *s.t.* $v^{\mathrm{T}} n \geq 0$ for any $n \in \operatorname{int} K$, where the convex set $K = \{x - y : x \in C, y \in D\}$.

If $\operatorname{int} K$ is empty, then the statement becomes trivial.

Otherwise, we construct v as follows. First, let $K_j = \{x : x \in \operatorname{int} K, d(x, \operatorname{bd} K) \geq \frac{1}{j}\}$. (Feel free to skip those empty ones at the beginning.) Then, for each nonempty $\operatorname{cl} K_j$ (which is convex, closed, and doesn't pass the origin point), let v_j be the projection of the origin point

onto $\operatorname{cl} K_j$. Then $v_j^{\mathrm{T}}(v_j - n) \le 0$ for any $n \in \operatorname{cl} K_j$, so we always have $v_j^{\mathrm{T}} n \ge 0$. Finally, the sequence $\{\frac{v_j}{\|v_j\|}\}$ must contain a infinite subsequence (with index set $\mathcal{I}$) that converges to, say, v. Since $\operatorname{int} K = \cup_{j\in\mathcal{I}} K_j$, we have $v^{\mathrm{T}} n \ge 0$ for any $n \in \operatorname{int} K$. □

Theorem 1.11 (supporting hyperplane theorem)
For any nonempty convex set $C \subseteq \mathbb{R}^d$, and any $x_0 \in \operatorname{bd} C$ (or equivalently, $x_0 \in \operatorname{cl} C \backslash \operatorname{int} C$), there exists $v \neq 0$, *s.t.* $v^{\mathrm{T}} x \le v^{\mathrm{T}} x_0$ for all $x \in C$.

Proof. Apply the separating hyperplane theorem to $\operatorname{int} C$ and $\{x_0\}$. □

1.3 Convex function

1.3.1 Definition and basic properties

Definition 1.5
Let $f : \operatorname{dom}(f) \to \mathbb{R}$, and $\operatorname{dom}(f) \subseteq \mathbb{R}^d$. The **graph** of f is

$$\{(x, f(x)) : x \in \operatorname{dom}(f)\}.$$

Definition 1.6
A function $f : \operatorname{dom}(f) \to \mathbb{R}$ is a **convex function**, if (i) $\operatorname{dom}(f)$ is a convex set, and (ii)

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

for any $x, y \in \operatorname{dom}(f)$, $0 \le \lambda \le 1$.

A function g is a **concave function** if $-g$ is a convex function.

Remark 1.3. *Geometrically, let f be a convex function, and $(x, f(x)), (y, f(y))$ be two points on its graph, then the line segment connecting $(x, f(x))$ and $(y, f(y))$ lies above the graph of f.*

Example 1.1
The following functions are convex:

1. $f(x) = a^{\mathrm{T}} x + b$, $x \in \mathbb{R}^d$, $a \in \mathbb{R}^d$, $b \in \mathbb{R}$
2. $f(x) = \|x\|$, $x \in \mathbb{R}^d$
3. $f(x) = \|x\|^2$, $x \in \mathbb{R}^d$

Remark 1.4. *Please refer to Example 1.2 for some other common convex functions.*

Proposition 1.12
Suppose that $f_1, \cdots, f_m$ are convex functions, and that $\lambda_1, \cdots, \lambda_m \geq 0$. Then

$$f = \sum_{i=1}^{m} \lambda_i f_i$$

is convex with $\mathrm{dom}(f) = \bigcap_{i=1}^{m} \mathrm{dom}(f_i)$.

Proposition 1.13
Suppose that f is a convex function with $\mathrm{dom}(f) \subseteq \mathbb{R}^d$, and that $g(x) = Ax + b$, where $A \in \mathbb{R}^{d \times m}$ and $b \in \mathbb{R}^d$. Then

$$f \circ g$$

is convex on $\mathrm{dom}(f \circ g) = \{x \in \mathbb{R}^m : g(x) \in \mathrm{dom}(f)\}$.

Definition 1.7
The **epigraph** of $f : \mathrm{dom}(f) \to \mathbb{R}$ is

$$\mathrm{epi}(f) = \{(x, \alpha) : x \in \mathrm{dom}(f), \alpha \geq f(x)\}.$$

Theorem 1.14
f is a convex function iff $\mathrm{epi}(f)$ is a convex set.

Proof. (if) For any $x, y \in \mathrm{dom}(f)$, $0 \leq \lambda \leq 1$, we have

$$(x, f(x)), (y, f(y)) \in \mathrm{epi}(f) \Longrightarrow (\lambda x + (1-\lambda)y, \lambda f(x) + (1-\lambda) f(y)) \in \mathrm{epi}(f).$$

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

(only if) For any $(x, \alpha), (y, \beta) \in \mathrm{epi}(f)$, $0 \leq \lambda \leq 1$, we have

$$\alpha \geq f(x), \ \beta \geq f(y);$$

$$\lambda f(x) + (1-\lambda) f(y) \geq f(\lambda x + (1-\lambda)y).$$

Putting them together, we have

$$\lambda\alpha + (1-\lambda)\beta \geq f(\lambda x + (1-\lambda)y) \Longrightarrow (\lambda x + (1-\lambda)y, \lambda\alpha + (1-\lambda)\beta) \in \mathrm{epi}(f).$$

□

Theorem 1.15 (Jensen's inequality)
Let f be convex, $x_1, \cdots, x_m \in \text{dom}(f)$, $\lambda_1, \cdots, \lambda_m \geq 0$, and $\sum_{i=1}^m \lambda_i = 1$. Then

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

Theorem 1.16 (Jensen's inequality)
Let f be convex, and X be a random variable *s.t.* $X \in \text{dom}(f)$ with probability one and $\mathbb{E}[X] \in \text{int}\,\text{dom}(f)$. Then

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)],$$

provided both expectations exist.

Proof. [Shreve, 2005] First, by supporintg hyperplane theorem, for any $x \in \text{int}\,\text{dom}(f)$,

$$f(x) = \sup\left\{l(x) : l \text{ is affine and } l(y) \leq f(y) \text{ for all } y \in \text{dom}(f)\right\}.$$

Then, for compactness, let l be always affine, all expectations taken over X, and $l(y) \leq f(y)$ for all $y \in \text{dom}(f)$ written in brief as $l \preceq f$,

$$\mathbb{E}[f(X)] \geq \mathbb{E}[\sup_{l \preceq f}\{l(X)\}] \geq \sup_{l \preceq f}\{\mathbb{E}[l(X)]\} = \sup_{l \preceq f}\{l(\mathbb{E}[X])\} = f(\mathbb{E}[X]).$$

□

Theorem 1.17 (convex function is continuous)
Suppose that f is convex, and that $\text{dom}(f) \subseteq \mathbb{R}^d$ is open.[a] Then f is continuous.

[a] It's worth mentioning that a convex function on infinite dimensional space is not necessarily continuous, though we mainly deal with functions on finite dimensional space throughout this note.

Proof. Consider first a cube $C \subseteq \text{dom}(f)$ centered at $x \in \text{dom}(f)$, with edge length l. f is bounded above in C (with its maximum, namely $f(y^*) = f(x) + g$, achieved on some vertex). Hence $f(y) - f(x) \leq g$ for all $y \in C$. Let C' be another cube centered at x with edge length kl, where $k = \min(1, \frac{\varepsilon}{g})$.

On one hand, for any $y \in C'$, let $\bar{y} \in C$ *s.t.* $\bar{y} - x = \frac{1}{k}(y - x)$. It follows from the convexity that

$$f(y) \leq (1-k)f(x) + kf(\bar{y}) = f(x) + k[f(\bar{y}) - f(x)] \leq f(x) + kg \leq f(x) + \varepsilon;$$

and on the other hand, $f(y)$ is bounded below in C' at least as tight as $f(x) - \varepsilon$ (since for $y \in C'$, $f(x) - f(y) \leq f(2x - y) - f(x) \leq \varepsilon$). Hence, $|f(x) - f(y)| \leq \varepsilon$ for all $y \in C'$. □

1.3.2 First-order and second-order condition

Theorem 1.18
Suppose that f is differentiable, and that $\mathrm{dom}(f)$ is open. Then the following conditions are equivalent:

1. f is convex.
2. $f(y) \geq f(x) + \nabla f(x)^{\mathrm{T}}(y - x), \quad \forall x, y \in \mathrm{dom}(f)$.
3. $\langle \nabla f(x) - \nabla f(y), x - y \rangle \leq 0, \quad \forall x, y \in \mathrm{dom}(f)$.

Proof. Let's break it into Theorem 1.19 and Theorem 1.20. □

Theorem 1.19 (first-order condition)
Suppose that f is differentiable, and that $\mathrm{dom}(f)$ is open. Then f is convex iff

$$f(y) \geq f(x) + \nabla f(x)^{\mathrm{T}}(y - x), \quad \forall x, y \in \mathrm{dom}(f).$$

Remark 1.5. *Geometrically, it shows that the graph $\{(y, f(y)\}$ is above its every tangent hyperplane $\{(y, f(x) + \nabla f(x)^{\mathrm{T}}(y - x))\}$.*

Proof. (if) For any $x, y \in \mathrm{dom}(f)$, $0 \leq \lambda \leq 1$, denote $z = \lambda x + (1 - \lambda)y$, then

$$f(x) \geq f(z) + \nabla f(z)^{\mathrm{T}}(x - z), \tag{1}$$
$$f(y) \geq f(z) + \nabla f(z)^{\mathrm{T}}(y - z). \tag{2}$$

Thus, we have $\lambda(1) + (1 - \lambda)(2)$ as follows:

$$\lambda f(x) + (1 - \lambda) f(y) \geq f(z) + 0 = f(z).$$

(only if) For any $x, y \in \mathrm{dom}(f)$, $0 \leq \lambda \leq 1$, we have

$$(1 - \lambda) f(x) + \lambda f(y) \geq f((1 - \lambda)x + \lambda y),$$

i.e.,

$$f(x) + \lambda(f(y) - f(x)) \geq f(x + \lambda(y - x)),$$

or

$$f(y) - f(x) \geq \frac{f(x + \lambda(y - x)) - f(x)}{\lambda}.$$

Hence,

$$f(y) - f(x) \geq \lim_{\lambda \downarrow 0} \frac{f(x + \lambda(y - x)) - f(x)}{\lambda} = \nabla f(x)^{\mathrm{T}}(y - x).$$

□

Theorem 1.20 (monotone mapping of ∇f)
Suppose that f is differentiable, and that $\mathrm{dom}(f)$ is open. Then

$$f(y) \le f(x) + \nabla f(x)^{\mathrm{T}}(y-x), \quad \forall x, y \in \mathrm{dom}(f),$$

iff

$$0 \le \langle \nabla f(x) - \nabla f(y), x - y \rangle, \quad \forall x, y \in \mathrm{dom}(f).$$

Proof. (if) For any $x, y \in \mathrm{dom}(f)$,

$$\begin{aligned} f(y) &= f(x) + \int_0^1 \langle \nabla f(x + t(y-x)), y - x \rangle \, dt \\ &= f(x) + \nabla f(x)^{\mathrm{T}}(y-x) + \int_0^1 \langle \nabla f(x + t(y-x)) - \nabla f(x), y - x \rangle \, dt \\ &\ge f(x) + \nabla f(x)^{\mathrm{T}}(y-x). \end{aligned}$$

(only if) For any $x, y \in \mathrm{dom}(f)$,

$$\begin{aligned} f(y) &\ge f(x) + \nabla f(x)^{\mathrm{T}}(y-x); \\ f(x) &\ge f(y) + \nabla f(y)^{\mathrm{T}}(x-y). \end{aligned}$$

Adding the above two inequalities yields

$$\langle \nabla f(x) - \nabla f(y), y - x \rangle \le 0.$$

□

Theorem 1.21 (second-order condition)
Suppose that f is twice continuously differentiable, and that $\mathrm{dom}(f)$ is open. Then f is convex iff for any $x \in \mathrm{dom}(f)$,

$$\nabla^2 f(x) \succeq 0.$$

Proof. (if) For any $x, y \in \mathrm{dom}(f)$,

$$\langle \nabla f(y) - \nabla f(x), y - x \rangle = \int_0^1 \left\langle \nabla^2 f(x + t(y-x))(y-x), y - x \right\rangle dt \ge 0.$$

(only if) For any $x, y \in \mathrm{dom}(f)$, let $y - x = u$. For any $0 < \lambda \le 1$,

$$0 \le \langle \nabla f(x + \lambda u) - \nabla f(x), \lambda u \rangle \le \lambda^2 \left\langle \frac{\nabla f(x + \lambda u) - \nabla f(x)}{\lambda}, u \right\rangle.$$

Let $\lambda \downarrow 0$. Then

$$0 \le \lim_{\lambda \downarrow 0} \left\langle \frac{\nabla f(x + \lambda u) - \nabla f(x)}{\lambda}, u \right\rangle = \left\langle \nabla^2 f(x) u, u \right\rangle.$$

□

Example 1.2
The following functions are convex:

- $f(x) = a^x$, $x \in \mathbb{R}$, $a > 0$ and $a \neq 1$
- $f(x) = -\log_b x$, $x \in (0, +\infty)$, $b > 1$
- $f(x) = |x|^p$, $x \in \mathbb{R}$, $p \geq 1$

1.3.3 Subgradient

Definition 1.8
Let $\mathrm{dom}(f) \subseteq \mathbb{R}^d$, $f : \mathrm{dom}(f) \to \mathbb{R}$, and $x \in \mathbb{R}^d$. Then g is a **subgradient** of f at x if for any $y \in \mathrm{dom}(f)$,

$$f(y) \geq f(x) + \langle g, y - x \rangle.$$

The set of subgradients of f at x is a subdifferential, denoted as $\partial f(x)$.

Remark 1.6. *Suppose that f is convex and differentiable at x. Then $\partial f(x) = \{\nabla f(x)\}$.*

Theorem 1.22
Let $\mathrm{dom}(f) \subseteq \mathbb{R}^d$ be convex. If for any $x \in \mathrm{dom}(f)$, $\partial f(x) \neq \emptyset$, then f is convex. Conversely, if f is convex, then for any $x \in \operatorname{int} \mathrm{dom}(f)$, $\partial f(x) \neq \emptyset$.

Proof. The first claim can be proven in similar way as in Theorem 1.19. The second claim follows from applying the supporting hyperplane thoerem on $\mathrm{epi}(f)$ at x. □

Theorem 1.23
Let $f : \mathrm{dom}(f) \to \mathbb{R}$, $x, y \in \mathrm{dom}(f)$, $x' \in \partial f(x)$, and $y' \in \partial f(y)$. Then $\langle x' - y', x - y \rangle \geq 0$.

Proof. The proof is very similar to the (only if) part of the proof of Theorem 1.20. □

1.3.4 Minima

Theorem 1.24
Suppose that f is convex, and that $\mathrm{dom}(f)$ is open. If $x^* \in \mathrm{dom}(f)$ is a local minimum, then it's a global minimum.

Proof. If there exists $x \in \mathrm{dom}(f)$ *s.t.* $f(x) < f(x^*)$, then, for any $0 < \lambda \leq 1$, $f(x^* + \lambda(x - x^*)) \leq \lambda f(x) + (1 - \lambda) f(x^*) < f(x^*)$. So, for any D *s.t.* $x^* \in \operatorname{int} D$, we always have a sufficiently small λ *s.t.* $x^* + \lambda(x - x^*) \in D$, then x^* will not be a minimum in D. □

Theorem 1.25
Suppose that f is convex. Then x is a global minimum iff $0 \in \partial f(x)$.

Proof. By the definition of subdifferential, $0 \in \partial f(x)$ is equivalent to, for any $y \in \mathrm{dom}(f)$,

$$f(y) \geq f(x).$$

□

Definition 1.9
Let $f : \mathrm{dom}(f) \to \mathbb{R}$ and $\alpha \in \mathbb{R}$. The set $\mathscr{L}_f(\alpha) = \{x \in \mathrm{dom}(f) : f(x) \leq \alpha\}$ is the **α-sublevel set** of f.

Theorem 1.26
Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function, and suppose there is a non-empty and bounded α-sublevel set $\mathscr{L}_f(\alpha)$. Then f has global minima.

Proof. By Theorem 1.17, f is continuous. Hence $\mathscr{L}_f(\alpha)$ is closed and bounded (compact). Hence it has a global minimum within $\mathscr{L}_f(\alpha)$. □

1.4 Dual norm and conjugate function

1.4.1 Dual norm

Definition 1.10
Let $\|\cdot\|$ be a norm (not necessarily the Euclidean norm) defined on $\mathbb{R}^d$. Then its **dual norm** $\|\cdot\|_* : \mathbb{R}^d \to \mathbb{R}$ is defined as

$$\|y\|_* = \sup\{\langle x, y\rangle : \|x\| = 1\}.$$

Theorem 1.27 (Cauchy-Schwarz inequality)
Within $\mathbb{R}^d$, we have norm $\|\cdot\|$ and its dual norm $\|\cdot\|_*$. Then

$$|\langle x, y\rangle| \leq \|x\| \cdot \|y\|_*, \quad \forall x, y \in \mathbb{R}^d.$$

Proof. $\|y\|_* = \sup\{z^{\mathrm{T}} y : \|z\| = 1\} \geq \left|\frac{x^{\mathrm{T}}}{\|x\|} y\right|$. □

Definition 1.11
L_p-norm $\|\cdot\|_p : \mathbb{R}^d \to \mathbb{R}$ (where $1 \le p \le \infty$) is defined as

$$L_p(x) = \|x\|_p = (\sum_{i=1}^{d} |x^{(i)}|^p)^{1/p}, \quad \forall x = [x^{(1)}, \cdots, x^{(d)}]^{\mathrm{T}}.$$

When $p = 2$, it's exactly the Euclidean norm.

Remark 1.7. $\|x\|_\infty = \sup\{|x^{(i)}| : 1 \le i \le d\}, \quad \forall x = [x^{(1)}, \cdots, x^{(d)}]^{\mathrm{T}}.$

Remark 1.8 (Minkowski's Inequality)**.** *The L_p-norm in Definition 1.11 satisfies non-negativity and homogeneity. But it's not entirely obvious for the triangle inequality:*

$$L_p(x + y) \le L_p(x) + L_p(y).$$

Proof. If either x or y is 0, it becomes trivial. So, assume they are both non-zero.
Let $\bar{x} = \frac{x}{L_p(x)}$, $\bar{y} = \frac{y}{L_p(y)}$. Then it's equivalent to prove

$$L_p(L_p(x) \cdot \bar{x} + L_p(y) \cdot \bar{y}) \le L_p(x) + L_p(y).$$

Assume $\frac{L_p(x)}{L_p(x)+L_p(y)} = \lambda$. Then $\frac{L_p(y)}{L_p(x)+L_p(y)} = 1 - \lambda$. And it's equivalent to prove

$$L_p(\lambda\bar{x} + (1 - \lambda)\bar{y}) \le 1.$$

For each coordinate $x^{(i)}$, $y^{(i)}$, we have (by the convexity of $|\cdot|^p$)

$$|\lambda\bar{x}^{(i)} + (1 - \lambda)\bar{y}^{(i)}|^p \le \lambda|\bar{x}^{(i)}|^p + (1 - \lambda)|\bar{y}^{(i)}|^p.$$

Hence

$$L_p(\lambda\bar{x} + (1 - \lambda)\bar{y}) \le \lambda L_p(\bar{x}) + (1 - \lambda)L_p(\bar{y}) = 1.$$

□

Proposition 1.28
L_1-norm and L_∞-norm are a pair of dual norms; L_2-norm is self dual.

Proposition 1.29 (dual of the dual norm)
Within $\mathbb{R}^d$[a], we have norm $\|\cdot\|$ and its dual norm $\|\cdot\|_*$. Then

$$\|x\|_{**} = \|x\|, \quad \forall x \in \mathbb{R}^d.$$

[a] In infinite-dimensional vector space, this result might not hold.

Remark 1.9. *We at this moment treat Proposition 1.29 as a conjecture. See this proof using Lagrangian https://www.stat.cmu.edu/ ryantibs/convexopt-F16/scribes/dual-corres-scribed.pdf.*

Definition 1.12
The norm $\|\cdot\|$ and dual norm $\|\cdot\|_*$ defined on $\mathbb{R}^d$ naturally induce the norm for matrix $\|\cdot\| : \mathbb{R}^{d\times d}$:

$$\|A\| = \sup\{\|Ax\|_* : \|x\| = 1\}.$$

Proposition 1.30
Under the definition in Definition 1.12, we have

$$\|Ax\|_* \le \|A\| \cdot \|x\|.$$

1.4.2 Conjugate function

Definition 1.13
Let $f : \mathbb{R}^d \to \mathbb{R}$. The function $f^* : \mathbb{R}^d \to \mathbb{R}$ defined as

$$f^*(y) = \sup_{x\in\mathrm{dom}(f)} (y^{\mathrm{T}}x - f(x)),$$

is called the **conjugate function** of the function f, with its domain consisting of $y \in \mathbb{R}^n$ for which $f^*(y) < \infty$.

Proposition 1.31
Let f^* be the conjugate function of f. The conjugate function f^* is convex (no matter whether f is convex).

Theorem 1.32 (Fenchel's inequality)
Let f^* be the conjugate function of f. Then, for all $x \in \mathrm{dom}(f)$, $y \in \mathrm{dom}(f^*)$,

$$f(x) + f^*(y) \ge x^{\mathrm{T}}y.$$

Proof. It follows immediately from

$$f^*(y) = \sup_{x\in\mathrm{dom}(f)} (y^{\mathrm{T}}x - f(x)) \ge y^{\mathrm{T}}x - f(x).$$

□

Theorem 1.33 (conjugate of the conjugate)
If f is convex, and f is closed (*i.e.*, $\mathrm{epi}(f)$ is closed), then $f^{**} = f$.

Proof. For any $x \in \mathrm{dom}(f)$,

$$f^{**}(x) = \sup_y x^{\mathrm{T}} y - f^*(y) \leq \sup_y x^{\mathrm{T}} y - (x^{\mathrm{T}} y - f(x)) = f(x).$$

Assume that $f^{**}(x) < f(x)$. We need the following lemma.

Lemma 1.34
There exists a non-vertical hyperplane that passes through $(x, f^{**}(x))$ and is strictly below $\mathrm{epi}(f)$.

Proof. Let $(v, f(v))$ be the projection of $(x, f^{**}(x))$ onto $\mathrm{epi}(f)$. Obviously, $f(v) \neq f^{**}(x)$. The hyperplane passing through $(\frac{v+x}{2}, \frac{f(v)+f^{**}(x)}{2})$ with normal vector $(v-x, f(v)-f^{**}(x))$ is non-vertical, and separates $\mathrm{epi}(f)$ and $(x, f^{**}(x))$. After shifting the aforementioned hyperplane a bit down, we obtain the desired hyperplane, which passes through $(x, f^{**}(x))$ with normal vector $(v-x, f(v)-f^{**}(x))$. □

By Lemma 1.34, we have $y \in \mathbb{R}^d$ and $\varepsilon > 0$, *s.t.*

$$f(z) > y^{\mathrm{T}}(z-x) + f^{**}(x) + \varepsilon, \quad \forall z \in \mathrm{dom}(f),$$

and then

$$y^{\mathrm{T}} x \geq \sup_{z \in \mathrm{dom}(f)} (y^{\mathrm{T}} z - f(z)) + f^{**}(x) + \varepsilon = f^*(y) + f^{**}(x) + \varepsilon \geq y^{\mathrm{T}} x + \varepsilon,$$

where the last step follows from Fenchel's inequality and finally leads to contradiction. □

1.5 Key points and problems

Key Points: inner product (Cauchy-Schwarz inequality, norm, angle); convex set; convex function (definition, graph and geometric meaning, epigraph, Jensen's inequality, continuity of convex functions); first order condition; second order condition; minima (local/global minima, critical point); conjugate function (conjugate function, Fenchel's inequality, conjugate of the conjugate).

Problem 1.1
Suppose that f is continuous, and that $\mathrm{dom}(f)$ is convex. Then f is convex iff $f(\frac{x+y}{2}) \leq \frac{f(x)+f(y)}{2}$ for any $x, y \in \mathrm{dom}(f)$.

Problem 1.2 ([Boyd and Vandenberghe, 2004])
Show that:

- $x \log x$ is convex on $\mathbb{R}_{>0}$;
- $f(x, y) = x^2/y$ is convex on $\mathbb{R} \times \mathbb{R}_{>0}$;
- $\log(e^{x_1} + \cdots + e^{x_n})$ is convex on $\mathbb{R}^n$;
- $(\prod_{i=1}^n x_i)^{1/n}$ is concave on $\mathbb{R}^n_{>0}$;
- $\log \det X$ is concave on $\mathbb{S}^n_{\succ 0}$, where $\mathbb{S}^n_{\succ 0}$ denotes the set of symmetric positive definite matrices.

2 Optimization problems

The Chapter 2 is based on the Section 1.1 in [Nesterov et al., 2018].

2.1 A general formulation of optimization problems

We consider different variants of the following **optimization problem**[1]:

$$\begin{aligned} &\min f_0(x) \\ &s.t.\ f_j(x) \le 0, \quad j = 1, \dots, m \\ &x \in Q, \end{aligned} \tag{3}$$

where $f_0(x)$ is called the **objective function**, Q is called the **basic feasible set**, and the set $\mathscr{F}$ is called the **(entire) feasible set**

$$\mathscr{F} = \{x \in Q : f_j(x) \le 0,\ j = 1, \dots, m\}.$$

A point $x^* \in \mathscr{F}$ is called the **(optimal) global solution** to (3), if

$$f_0(x^*) \le f_0(x) \quad \forall x \in \mathscr{F}.$$

In this case, $f_0(x^*)$ is called the **(global) optimal value** of the problem.

A point $x^* \in \mathscr{F}$ is called a **local solution** to (3), if there exists a set $\hat{\mathscr{F}} \subseteq \mathscr{F}$ *s.t.* $x^* \in \operatorname{int} \hat{\mathscr{F}}$ and

$$f_0(x^*) \le f_0(x) \quad \forall x \in \hat{\mathscr{F}}.$$

A problem class $\mathscr{P}$ contains a **model**, its known part, Σ, an **oracle** $\mathscr{O}$ to answer successive questions, and a **stopping criterion** (often associated with some accuracy $\varepsilon > 0$) $\mathscr{J}_\varepsilon$:

$$\mathscr{P} \equiv (\Sigma, \mathscr{O}, \mathscr{J}_\varepsilon).$$

A **method** $\mathscr{M}$, who knows part of the problem (the model Σ and the stopping criterion $\mathscr{J}_\varepsilon$), will try to solve a problem $P \in \mathscr{P}$ by collecting and handling the answers returned by the oracle $\mathscr{O}$.

In order to solve a problem $P \in \mathscr{P}$ with $\mathscr{M}$, we usually apply to it an iterative process as Algorithm 1.

To measure the performance of a method $\mathscr{M}$, we typically use two different kinds of **computational cost**:

- **Analytical complexity**: The number of calls of the oracle $\mathscr{O}$ in Algorithm 1.
- **Arithmetical complexity**: The total number of arithmetic operations, including the work of oracle $\mathscr{O}$ (line 3 in Algorithm 1) and the work of method $\mathscr{M}$ (line 5 in Algorithm 1).

One standard assumption called **Local Black Box** on the oracle is as follows:

- The **only** information available for the numerical scheme is the answer of the oracle $\mathscr{O}$.

[1] Note that $\le$ sign is also capable of expressing $\ge$ and $=$ constraints.

Algorithm 1 General iterative scheme

Require: Starting point x_0 and accuracy $\varepsilon > 0$.

Initialize $k = 0$, $\mathscr{I}_{-1} = \emptyset$. Here k is the iteration counter and $\mathscr{I}_k$ is the accumulated **information set**.
loop
  Call oracle $\mathscr{O}$ at x_k.
  Update the information set: $\mathscr{I}_k = \mathscr{I}_{k-1} \cup (x_k, \mathscr{O}(x_k))$.
  Apply the rules of the method $\mathscr{M}$ to $\mathscr{I}_k$ and generate a new point x_{k+1}.
  if $\mathscr{I}_\varepsilon$ is yes **then**
    break
  else
    $k := k + 1$.
  end if
end loop

- The oracle is **local**: A small variation of the problem far enough from the test point x, which is compatible with the description of the problem class, does not change the answer at x.

In Black-Box Optimization, our main consideration is analytical complexity. In Structural Optimization, we will also consider arithmetical complexity.

Usually, for solving (3), we adopt some standard assumptions on the level of smoothness of functional components. According to the degree of smoothness, we can apply different types of oracle:

- **Zeroth-order** oracle: returns $f(x)$.
- **First-order** oracle: returns $f(x)$ and $\nabla f(x)$.
- **Second-order** oracle: returns $f(x)$, $\nabla f(x)$, and $\nabla^2 f(x)$.

2.2 Complexity bounds for Lipschitz continuous functions

2.2.1 Lipschitz continuous function

Definition 2.1
The function $f : \mathbb{R}^d \to \mathbb{R}$ is M-**Lipschitz continuous** ($M > 0$) if

$$|f(y) - f(x)| \leq M\|y - x\|.$$

Theorem 2.1
Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable and $\mathrm{dom}(f)$ be open and convex. f is M-Lipschitz continuous iff

$$\|\nabla f(x)\|_* \leq M, \quad \forall x \in \mathrm{dom}(f).$$

Proof. (if) For any $x, y \in \mathrm{dom}(f)$,

$$|f(y) - f(x)| = \left| \int_0^1 \nabla f(x + t(y-x))^{\mathrm{T}} (y - x) dt \right| \leq M \cdot \|y - x\|.$$

(only if) For any $x, y \in \mathrm{dom}(f)$, let $u = y - x$, then for $0 < \lambda \leq 1$,

$$\frac{|f(x + \lambda u) - f(x)|}{\lambda} \leq M \|u\|.$$

Let $\lambda \downarrow 0$. Then

$$M \geq \frac{1}{\|u\|} \cdot \lim_{\lambda \downarrow 0} \frac{|f(x + \lambda u) - f(x)|}{\lambda} = \left| \nabla f(x)^{\mathrm{T}} \frac{u}{\|u\|} \right|.$$

Therefore, $\|\nabla f(x)\|_* \leq M$. □

Now, we try to apply the formal language introduced above in Section 2.1 to a particular problem class (denoted as $\mathscr{P}_\infty$):

- The model Σ.

  $$\min_{x \in B_d} f(x),$$

  where

  $$B_d = \{x \in \mathbb{R}^d : 0 \leq x^{(i)} \leq 1,\ i = 1, \ldots, d\},$$

  and (for some constant $M > 0$)

  $$|f(x) - f(y)| \leq M \|x - y\|_\infty \quad \forall x, y \in B_d.$$

- The oracle $\mathscr{O}$: Zeroth-order Local Black Box.
- The stopping criterion $\mathscr{J}_\varepsilon$: Find $\bar{x} \in B_d$ *s.t.* $f(\bar{x}) - f^* < \varepsilon$.

2.2.2 Uniform grid method

We consider a very simple method (denoted as $\mathscr{G}(p)$) for solving $P \in \mathscr{P}_\infty$ in Algorithm 2.

Algorithm 2 Uniform Grid Method $\mathscr{G}(p)$

Require: $p \in \mathbb{Z}_{\geq 1}$.

1: Form p^d points

$$x_\alpha = \left(\frac{2i_1 - 1}{2p}, \frac{2i_2 - 1}{2p}, \cdots, \frac{2i_d - 1}{2p} \right)^{\mathrm{T}},$$

for all $\alpha \in \{1, \cdots, p\}^d$.

2: Find $\bar{x} \in \{1, \cdots, p\}^d$ that minimizes $f(\bar{x})$.

3: **return** $(\bar{x}, f(\bar{x}))$.

Theorem 2.2
Let f^* be a global optimal value. For the $\bar{x}$ returned by Algorithm 2,

$$f(\bar{x}) - f^* \leq \frac{M}{2p}.$$

Proof. Let x^* be a global solution. Note that there exists an $\alpha^* \in \{1, \cdots, d\}$, *s.t.* $\|x^* - x_{\alpha^*}\|_\infty \leq \frac{1}{2p}$. Therefore,

$$f(\bar{x}) - f(x^*) \leq f(x_{\alpha^*}) - f(x^*) \leq \frac{M}{2p}.$$

□

Remark 2.1. *Following Theorem 2.2, the analytical complexity of $\mathscr{P}_\infty$ for method $\mathscr{G}$ is at most*

$$\left(\left\lfloor \frac{M}{2\epsilon} \right\rfloor + 1\right)^d. \tag{4}$$

*Thus, we've already had an **upper complexity bound** for $\mathscr{P}_\infty$.*

2.2.3 Lower bound for Lipschitz continuous optimization

However, at this point, we neither know whether our analysis in Theorem 2.2 is tight enough, nor whether there is other schemes with much better performance.

Hence, we want to derive a **lower complexity bound** for the problem class $\mathscr{P}_\infty$.

To do this, we will employ a **resisting oracle** to create the worst possible problem $P \in \mathscr{P}_\infty$ for each particular method. It starts from an "empty" function and it tries to answer each call of the method in the worst possible way. However, the answers must be **compatible** with the previous answers and with description of the problem class. Then, after termination of the method it is possible to **reconstruct** a problem which perfectly fits the final informational set accumulated by the algorithm. Moreover, if we run the method on this newborn problem P, it will reproduce the same sequence of test points since it will have the same sequence of answers from the oracle.

Theorem 2.3
For $\varepsilon < \frac{M}{2}$, the analytical complexity of $\mathscr{P}_\infty$ is at least

$$\lfloor \frac{M}{2\varepsilon} \rfloor^d.$$

Proof. Let $p = \lfloor \frac{M}{2\varepsilon} \rfloor$ (≥ 1). Suppose there exists a method $\mathscr{M}$ which needs $N < p^d$ calls of oracle to solve any problem from $\mathscr{P}_\infty$. Then let the resisting oracle return $f(x) = 0$ at any test point x.

Therefore, this method can find only $\bar{x} \in B_d$ with $f(\bar{x}) = 0$. And since $N < p^d$, there exists an $\alpha \in \{1, \ldots, p\}^d$ *s.t.* all points in $\{x \in \mathbb{R}^d : \|x - x_\alpha\}_\infty \leq \frac{1}{2p}$ are not tested.

Define the function

$$\bar{f}(x) = \min\left\{0, M\left\|x - x_\alpha\right\|_\infty - \varepsilon\right\},$$

which satisfies the conditions of the problem class $\mathscr{P}_\infty$, and has global optimal value $-\varepsilon$. Since its returned value is always 0 at test points, the accuracy of $\mathscr{M}$ cannot be better than ε. □

When $\varepsilon \leq O(\frac{M}{d})$, the upper and lower bounds coincide up to an absolute constant multiplicative factor. This means that, for such level of accuracy, $\mathcal{G}(\cdot)$ is **optimal** for the problem calss $\mathscr{P}_\infty$.

Remark 2.2. *Even though this is already the optimal complexity one can achieve for this class of problem, the fact is that it's still very unacceptable in practice. For example, let* $M = 2$, $d = 10$, $\varepsilon = 0.01$. *The size of this problem is actually quite small, but the lower bound shows that it can take as many as* 10^{20} *calls of the oracle.*

3 Smooth optimization

3.1 Local gradient method

3.1.1 Relaxation sequences

The majority of methods in general (possibly non-convex) Nonlinear Optimization are to generate a **relaxation sequence** of function values, *i.e.*,

$$f(x_{k+1}) \leq f(x_k), \quad k = 0, 1, 2, \cdots$$

This type of algorithms has the following desired properties:

1. If f is bounded below, $\{f(x_k)\}$ will converge.
2. We can always improve (or at lease not worsen) the initial value.

Let $\mathrm{dom}(f) = \mathbb{R}^d$. To construct such a sequence, assuming that f is continuously differentiable, we can apply the **first-order approximation** at x_k:

$$f(x) = f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + o(\|x - x_k\|).$$

Consider a direction s *s.t.* $\|s\| = 1$. Then

$$f(x_k + hs) = f(x_k) + h\langle \nabla f(x_k), s \rangle + o(h).$$

Hence

$$\lim_{h \downarrow 0} \frac{f(x_k + hs) - f(x_k)}{h} = \langle \nabla f(x_k), s \rangle \geq -\|\nabla f(x_k)\|_* \cdot \|s\| = -\|\nabla f(x_k)\|_*.$$

When we consider the Euclidean norm $\|\cdot\|_2$, the antigradient direction $-\nabla f(x_k)$ will be the direction of the fastest decrease of function $f(\cdot)$ locally at x_k. So, one basic strategy is to always head towards this antigradient direction, as shown in Algorithm 3.

Algorithm 3 Gradient descent method

1: Initialize $x_0 \in \mathbb{R}^d$.
2: **for** $k = 0, 1, \cdots$ **do**
3: $\quad x_{k+1} = x_k - h_k \nabla f(x_k)$
4: **end for**

3.1.2 Lipschitz continuous gradient and smoothness

To better reason the behavior of the gradient descent method, we often need to characterize the level of continuity of its gradient.

Definition 3.1
Let a function f be differentiable and $\mathrm{dom}(f)$ be open and convex. f has **L-Lipschitz continuous gradient** ($L > 0$) under norm $\|\cdot\|$ if for any $x, y \in \mathrm{dom}(f)$,

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

Definition 3.2
Let $Q \subseteq \mathbb{R}^d$ be open and convex. We denote by $f \in C_L^{k,p}(Q, \|\cdot\|)$ for function f which

1. is k-times continuously differentiable, and
2. has L-Lipschitz continuous p-th derivative (under norm $\|\cdot\|$) on Q.[a]

We also use the notation $f \in C^k(Q, \|\cdot\|)$ for function f which is k-times continuously differentiable on Q.

[a] At this moment, we haven't introduced the definition of Lipschitz continuous p-th derivative for $p > 1$. But don't worry, we'll do that later, and until then, we'll only work on the cases where $p = 1$.

Proposition 3.1
If $f_1 \in C_{L_1}^{k,1}(Q_1, \|\cdot\|)$, and $f_2 \in C_{L_2}^{k,1}(Q_2, \|\cdot\|)$, then, for any $\alpha, \beta \in \mathbb{R}$,

$$\alpha f_1 + \beta f_2 \in C_{|\alpha|L_1+|\beta|L_2}^{k,1}(Q_1 \cap Q_2, \|\cdot\|).$$

Definition 3.3
A function $f \in C^0(Q, \|\cdot\|)$ is **L-smooth** ($L > 0$) if for any $x, y \in Q$ and $0 \le \lambda \le 1$,

$$\lambda f(x) + (1-\lambda) f(y) - f(\lambda x + (1-\lambda) y) \le \lambda(1-\lambda)\frac{L}{2}\|x - y\|^2.$$

Theorem 3.2
Let $f \in C^1(Q, \|\cdot\|)$, $L > 0$. Then the following conditions are equivalent: (for any $x, y \in Q$)

1. f is L-smooth.
2. $f(y) \le f(x) + \nabla f(x)^{\mathrm{T}}(y - x) + \frac{L}{2}\|y - x\|^2$.
3. $\langle \nabla f(x) - \nabla f(y), x - y \rangle \le L\|x - y\|^2$.

Proof. Let's break it into Theorem 3.3 and Theorem 3.4. □

Theorem 3.3 (L-smoothness quadric)
Let $f \in C^1(Q, \|\cdot\|)$, $L > 0$. Then f is L-smooth iff

$$f(y) \le f(x) + \nabla f(x)^{\mathrm{T}}(y - x) + \frac{L}{2}\|y - x\|^2, \quad \forall x, y \in Q.$$

Proof. (if) For any $x, y \in Q$ and $0 \le \lambda \le 1$, let $z = \lambda x + (1-\lambda)y$, then

$$f(x) \le f(z) + \nabla f(z)^{\mathrm{T}}(x - z) + \frac{L}{2}\|x - z\|^2, \tag{5}$$

$$f(y) \le f(z) + \nabla f(z)^{\mathrm{T}}(y - z) + \frac{L}{2}\|y - z\|^2. \tag{6}$$

By $\lambda(5) + (1-\lambda)(6)$, we have

$$\lambda f(x) + (1-\lambda)f(y) - f(z) \le \lambda(1-\lambda)\frac{L}{2}\|x - y\|^2.$$

(only if) For any $x, y \in Q$ and $0 < \lambda \le 1$, since

$$(1-\lambda)f(x) + \lambda f(y) - f((1-\lambda)x + \lambda y) \le (1-\lambda)\lambda\frac{L}{2}\|x - y\|^2,$$

or

$$f(y) \le f(x) + \frac{f(x + \lambda(y-x)) - f(x)}{\lambda} + \frac{L}{2}(1-\lambda)\|y - x\|^2,$$

we have (when $\lambda \downarrow 0$)

$$f(y) \le f(x) + \nabla f(x)^{\mathrm{T}}(y - x) + \frac{L}{2}\|y - x\|^2.$$

□

Theorem 3.4
Let $f \in C^1(Q, \|\cdot\|)$, $L > 0$. Then

$$f(y) \le f(x) + \nabla f(x)^{\mathrm{T}}(y - x) + \frac{L}{2}\|y - x\|^2, \quad \forall x, y \in Q,$$

iff

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \le L\|x - y\|^2, \quad \forall x, y \in Q.$$

Proof. (if) For any $x, y \in Q$,

$$\begin{aligned} f(y) &= f(x) + \nabla f(x)^{\mathrm{T}}(y - x) + \int_0^1 (\nabla f(x + t(y-x)) - \nabla f(x))^{\mathrm{T}}(y - x)dt \\ &\le f(x) + \nabla f(x)^{\mathrm{T}}(y - x) + \int_0^1 Lt\|y - x\|^2 dt \\ &= f(x) + \nabla f(x)^{\mathrm{T}}(y - x) + \frac{L}{2}\|y - x\|^2. \end{aligned}$$

(only if) For any $x, y \in Q$,

$$f(y) \le f(x) + \nabla f(x)^{\mathrm{T}}(y - x) + \frac{L}{2}\|y - x\|^2, \tag{7}$$

$$f(x) \le f(y) + \nabla f(y)^{\mathrm{T}}(x - y) + \frac{L}{2}\|x - y\|^2. \tag{8}$$

By (7) + (8), we have

$$0 \le \langle \nabla f(x) - \nabla f(y), y - x \rangle + L\|x - y\|^2.$$

□

Remark 3.1. *The proof techniques in Theorem 3.3 and Theorem 3.4 are quiet similar to the ones used in the proof of the first-order condition and the monotone gradient mapping of convex function.*

Proposition 3.5
If $f \in C^1(Q, \|\cdot\|)$, f is L-smooth, and $x^* \in Q$ *s.t.* $\nabla f(x^*) = 0$, then

$$f(x) \le f(x^*) + \frac{L}{2}\|x - x^*\|^2, \quad \forall x \in Q.$$

Theorem 3.6
If $f \in C_L^{1,1}(Q, \|\cdot\|)$, then for any $x, y \in Q$,

$$(\nabla f(x) - \nabla f(y))^{\mathrm{T}}(x - y) \le L\|x - y\|^2.$$

Proof. For any $x, y \in Q$,

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \le \|\nabla f(x) - \nabla f(y)\|_* \cdot \|x - y\| \le L\|x - y\|^2.$$

□

Remark 3.2. *Theorem 3.6 shows that $f \in C_L^{1,1}(Q, \|\cdot\|)$ yields the condition in Theorem 3.2. Therefore, if a function has L-Lipschitz continuous gradient, then it is L-smooth.*

Theorem 3.7
Let $f \in C^2(Q, \|\cdot\|)$. Then $f \in C_L^{2,1}(Q, \|\cdot\|)$ iff

$$\|\nabla^2 f(x)\| \le L, \quad \forall x \in Q.$$

Proof. (if) For any $x, y \in Q$,

$$\nabla f(y) - \nabla f(x) = \int_0^1 \nabla^2 f(x + t(y - x))(y - x)dt.$$

Then

$$\begin{aligned}
\|\nabla f(y) - \nabla f(x)\|_* &= \left\| \int_0^1 \nabla^2 f(x + t(y - x))(y - x)dt \right\|_* \\
&\le \int_0^1 \|\nabla^2 f(x + t(y - x))\| \cdot \|y - x\| dt \\
&\le L\|y - x\|.
\end{aligned}$$

(only if) For any $x, y \in Q$, let $u = y - x$. For any $0 < \lambda \leq 1$,

$$\|\nabla f(x + \lambda u) - \nabla f(x)\|_* \leq \lambda L \|u\|.$$

Hence,

$$L\|u\| \geq \lim_{\lambda \downarrow 0} \left\| \frac{\nabla f(x + \lambda u) - \nabla f(x)}{\lambda} \right\|_* = \|\nabla^2 f(x) u\|_*.$$

Therefore,

$$\|\nabla^2 f(x)\| = \sup \left\{ \frac{\|\nabla^2 f(x) u\|_*}{\|u\|} \right\} \leq L.$$

□

3.1.3 Convergence to stationary points

In this paragraph (Paragraph 3.1.3) as well as in the remaining of this section (Section 3.1), we will work on the convergence rate of a generalized version of our Algorithm 3, where we can consider the smoothness under an arbitrary norm, not limited to the Euclidean norm. As shown in Algorithm 4, our Algorithm 3 is actually the special case of $\|\cdot\|$ being $\|\cdot\|_2$.

Algorithm 4 Steepest descent method

1: Initialize $x_0 \in \mathbb{R}^d$.
2: **for** $k = 0, 1, \cdots$ **do**
3: $\quad g_k = \underset{\|v\| = \|\nabla f(x_k)\|_*}{\arg\max} \langle \nabla f(x_k), v \rangle$.
4: $\quad x_{k+1} = x_k - h_k g_k$.
5: **end for**

Theorem 3.8
Suppose that $f \in C^1(\mathbb{R}^d, \|\cdot\|)$, and that f is L-smooth. In Algorithm 4, for $h_k = \frac{2\alpha}{L}$ with $\alpha \in (0, 1)$,

$$f(x_k) - f(x_{k+1}) \geq \frac{2}{L} \alpha (1 - \alpha) \|\nabla f(x_k)\|_*^2.$$

Proof.

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) - \left\langle \nabla f(x_k), \frac{2\alpha}{L} g_k \right\rangle + \frac{L}{2} \left\| \frac{2\alpha}{L} g_k \right\|^2 \\ &= f(x_k) - \frac{2\alpha}{L} \|\nabla f(x_k)\|_* \cdot \|g_k\| + \frac{2\alpha^2}{L} \|g_k\|^2 \\ &= f(x_k) - \frac{2}{L} \alpha (1 - \alpha) \|\nabla f(x_k)\|_*^2. \end{aligned}$$

□

Corollary 3.9 (sufficient decrease)
Suppose that $f \in C^1(\mathbb{R}^d, \|\cdot\|)$, and that f is L-smooth. In Algorithm 4, for $\alpha = \frac{1}{2}$, or $h_k = \frac{1}{L}$, we have

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|_*^2. \tag{9}$$

Corollary 3.10
Suppose that $f \in C^1(\mathbb{R}^d, \|\cdot\|)$, that f is L-smooth, and that f is bounded below by f^*. Then for any $x \in \mathbb{R}^d$,

$$f(x) - f^* \geq \frac{1}{2L}\|\nabla f(x)\|_*^2.$$

Remark 3.3. *Note that* $\mathrm{dom}(f) = \mathbb{R}^d$ *is necessary not only for this theorem but also for the other results in this paragraph (Paragraph 3.1.3). Otherwise, when we start from* $x_0 = x$*, the* x_1 *defined in Algorithm 4 can be outside the domain of* f*.*

Theorem 3.11
Suppose that $f \in C^1(\mathbb{R}^d, \|\cdot\|)$, that f is L-smooth, and that f is bounded below by f^*. In Algorithm 4, for $h_k = \frac{1}{L}$,

$$\frac{1}{T}\sum_{k=0}^{T-1}\|\nabla f(x_k)\|_*^2 \leq \frac{2L(f(x_0) - f(x_T))}{T} \leq \frac{2L(f(x_0) - f^*)}{T}.$$

Proof. Summing over k from 0 to $T-1$ in (9) and multiplying both sides by $2L$, we have

$$\sum_{k=0}^{T-1}\|\nabla f(x_k)\|_*^2 \leq 2L(f(x_0) - f(x_T)) \leq 2L(f(x_0) - f^*).$$

□

Example 3.1 (Example 1.2.2 in [Nesterov et al., 2018])
Suppose $f : \mathbb{R}^2 \to \mathbb{R}$ is defined as

$$f(x) = f(x^{(1)}, x^{(2)}) = \frac{1}{2}(x^{(1)})^2 + \frac{1}{4}(x^{(2)})^4 - \frac{1}{2}(x^{(2)})^2.$$

The local (actually also global) minima are $(0, -1)$ and $(0, 1)$, while $(0, 0)$ is only a stationary point. But if we start from, *e.g.*, $x_0 = (1, 0)$, the gradient descent method (Algorithm 3) will finally converge to $(0, 0)$.

Remark 3.4. *As shown in Theorem 3.11, under the assumption of L-Lipschitz continuous gradient (or the slightly weaker assumption of L-smoothness), we can guarantee the convergence of Algorithm 4 to a stationary point, but we have no guarantee of obtaining an (even local) minimum.*

3.1.4 Smooth convex function

Definition 3.4
Let $Q \subseteq \mathbb{R}^d$ be open and convex. We denote by $f \in \mathscr{F}_L^{k,p}(Q, \|\cdot\|)$ for *convex* function f which

1. is k-times continuously differentiable, and
2. has L-Lipschitz continuous p-th derivative (under norm $\|\cdot\|$) on Q.

We also use the notation $f \in \mathscr{F}^k(Q, \|\cdot\|)$ for *convex* function f which is k-times continuously differentiable on Q.

Theorem 3.12
Let $f \in C^1(\mathbb{R}^d, \|\cdot\|)$, $L > 0$. The following conditions (for any $x, y \in \mathbb{R}^d$ and $0 \le \lambda \le 1$) are equivalent:

$$f \in \mathscr{F}_L^{1,1}(\mathbb{R}^d, \|\cdot\|) \tag{10}$$

$$f \text{ is convex and } L\text{-smooth (under } \|\cdot\|) \tag{11}$$

$$\lambda f(x) + (1-\lambda) f(y) \ge f(\lambda x + (1-\lambda) y) + \frac{\lambda(1-\lambda)}{2L} \|\nabla f(x) - \nabla f(y)\|_*^2 \tag{12}$$

$$f(y) - f(x) - \nabla f(x)^{\mathrm{T}}(y - x) \ge \frac{1}{2L} \|\nabla f(x) - \nabla f(y)\|_*^2 \tag{13}$$

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \ge \frac{1}{L} \|\nabla f(x) - \nabla f(y)\|_*^2. \tag{14}$$

Remark 3.5. *Note that for the equivalence to be held,* $\mathrm{dom}(f) = \mathbb{R}^d$ *is necessary, since we are to use Corollary 3.10 in our proof. Suppose* $f \in C^1(Q, \|\cdot\|)$ *and* $Q \neq \mathbb{R}^d$*, then we can only have the equivalence between* (12) *and* (13) *(leave as an exercise in Problem 3.2). The two equivalent conditions can imply* (14)*, which is often called the* ***co-coercivity of gradient****. The co-coercivity of gradient is (slightly) stronger than* $f \in \mathscr{F}_L^{1,1}(\mathbb{R}^d, \|\cdot\|)$*, which in turn is (slightly) stronger than* (11)*.*

Proof. ((10)→(11)) It follows from Theorem 3.6.

((11)→(12)) Let $z = \lambda x + (1-\lambda) y$, and let

$$\bar{f}(t) = f(t) - \nabla f(z)^{\mathrm{T}} t.$$

Then $\bar{f}$ is convex and L-smooth. Moreover, since $\nabla \bar{f}(z) = 0$, z is a minimum of $\bar{f}$. Therefore,

by Corollary 3.10,

$$\bar{f}(x) - \bar{f}(z) \geq \frac{1}{2L}\|\nabla \bar{f}(x)\|_*^2, \tag{15}$$

$$\bar{f}(y) - \bar{f}(z) \geq \frac{1}{2L}\|\nabla \bar{f}(y)\|_*^2. \tag{16}$$

By $\lambda(15) + (1-\lambda)(16)$, we have

$$\begin{aligned}\lambda f(x) + (1-\lambda)f(y) - f(z) &\geq \frac{1}{2L}\left(\lambda\|\nabla f(x) - \nabla f(z)\|_*^2 + (1-\lambda)\|\nabla f(y) - \nabla f(z)\|_*^2\right) \\ &\geq \frac{1}{2L}\cdot\lambda(1-\lambda)\|\nabla f(x) - \nabla f(y)\|_*^2.\end{aligned}$$

((12)→(13)) Since

$$(1-\lambda)f(x) + \lambda f(y) \geq f((1-\lambda)x + \lambda y) + \frac{(1-\lambda)\lambda}{2L}\|\nabla f(x) - \nabla f(y)\|_*^2,$$

or

$$f(y) - f(x) - \frac{f(x+\lambda(y-x)) - f(x)}{\lambda} \geq \frac{1-\lambda}{2L}\|\nabla f(x) - \nabla f(y)\|_*^2,$$

we have (when $\lambda \downarrow 0$),

$$f(y) - f(x) - \nabla f(x)^{\mathrm{T}}(y-x) \geq \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|_*^2.$$

((13)→(14)) By adding (13) and another inequality obtained by switching x and y of (13), we obtain

$$\langle \nabla f(x) - \nabla f(y), x - y\rangle \geq \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|_*^2.$$

((14)→(10)) f is convex, since

$$\langle \nabla f(x) - \nabla f(y), x - y\rangle \geq \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|_*^2 \geq 0.$$

Moreover,

$$\|\nabla f(x) - \nabla f(y)\|_* \cdot \|x - y\| \geq \langle \nabla f(x) - \nabla f(y), x - y\rangle \geq \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|_*^2,$$

and then

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

□

3.1.5 Convergence to minima

Example 3.2
Let $\{a_n\}$ be a sequence of positive real numbers. If

$$a_k - a_{k+1} \geq a_k^2, \quad \forall k \geq 0,$$

then

1. $\frac{1}{a_{k+1}} - \frac{1}{a_k} \geq 1$.
2. $a_k \leq \frac{1}{k}$.

Proof. 1. Obviously, the sequence $\{a_n\}$ is decreasing. Therefore, $a_k - a_{k+1} \geq a_k^2 \geq a_k a_{k+1}$. Then, the proof is done by dividing both sides of the above inequality by $a_k a_{k+1}$.

2. It follows from $\frac{1}{a_k} \geq \frac{1}{a_0} + k$.

□

Theorem 3.13
Suppose that $f \in \mathscr{F}_L^{1,1}(\mathbb{R}^d, \|\cdot\|)$, and that the $f(x_0)$-sublevel set $\mathscr{L}_f(f(x_0))$ is bounded. Let x^* be a minimum of f, and then, let $R = \sup\{\|x - x^*\| : x \in \mathscr{L}_f(f(x_0))\}$. In Algorithm 4, for $h_k = \frac{1}{L}$,

$$f(x_T) - f(x^*) \leq \frac{2LR^2}{T}.$$

Proof. Since

$$R\|\nabla f(x_k)\|_* \geq \|\nabla f(x_k)\|_* \cdot \|x_k - x^*\| \geq \langle \nabla f(x_k), x_k - x^* \rangle \geq f(x_k) - f(x^*),$$

we have

$$f(x_k) - f(x_{k+1}) \geq \frac{1}{2L}\|\nabla f(x_k)\|_*^2 \geq \frac{1}{2L} \cdot \frac{(f(x_k) - f(x^*))^2}{R^2}.$$

For the sequence $\{\frac{f(x_n) - f(x^*)}{2LR^2}\}$, applying the same techniques as in Example 3.2, we have

$$f(x_T) - f(x^*) \leq \frac{2LR^2}{T}.$$

□

3.1.6 Strongly convex function

Definition 3.5
A function $f \in C^0(Q, \|\cdot\|)$ is **μ-strongly convex** under norm $\|\cdot\|$ ($\mu > 0$), if for any $x, y \in Q$ and $0 \leq \lambda \leq 1$,

$$\lambda f(x) + (1-\lambda) f(y) - f(\lambda x + (1-\lambda) y) \geq \lambda(1-\lambda)\frac{\mu}{2}\|x - y\|^2.$$

Definition 3.6
We use the notation of $f \in \mathscr{S}_\mu(Q, \|\cdot\|)$ to denote the function f that is μ-strongly convex (under norm $\|\cdot\|$) on Q. We also use the notation of $\mathscr{S}^{k,l}_{\mu,L}(Q, \|\cdot\|)$, where k, l and L have the same meaning as for $C^{k,l}_L(Q, \|\cdot\|)$.

Proposition 3.14
If $f_1 \in \mathscr{S}_{\mu_1}(Q_1, \|\cdot\|)$, and $f_2 \in \mathscr{S}_{\mu_2}(Q_2, \|\cdot\|)$, then, for any $\alpha > 0$, $\beta > 0$,

$$\alpha f_1 + \beta f_2 \in \mathscr{S}_{\alpha\mu_1+\beta\mu_2}(Q_1 \cap Q_2, \|\cdot\|).$$

Theorem 3.15
Let $f \in C^1(Q, \|\cdot\|)$, $\mu > 0$. Then the following conditions are equivalent: (for any $x, y \in Q$)

1. $f \in \mathscr{S}^1_\mu(Q, \|\cdot\|)$.
2. $f(y) \geq f(x) + \nabla f(x)^{\mathrm{T}}(y - x) + \frac{\mu}{2}\|y - x\|^2$.
3. $\langle \nabla f(x) - \nabla f(y), x - y\rangle \geq \mu\|x - y\|^2$.

Proof. The proof is quite similar to the one of Theorem 3.2. Leave it as an exercise in Problem 3.5. □

Proposition 3.16
If $f \in \mathscr{S}^1_\mu(Q, \|\cdot\|)$, and $x^* \in Q$ *s.t.* $\nabla f(x^*) = 0$, then

$$f(x) \geq f(x^*) + \frac{\mu}{2}\|x - x^*\|^2, \quad \forall x \in Q.$$

Theorem 3.17
If $f \in \mathscr{S}^1_\mu(Q, \|\cdot\|)$, then for any $x, y \in Q$,

$$\mu\|x - y\| \leq \|\nabla f(x) - \nabla f(y)\|_*,$$

and moreover,

$$\langle \nabla f(x) - \nabla f(y), x - y\rangle \leq \frac{1}{\mu}\|\nabla f(x) - \nabla f(y)\|^2_*.$$

Proof. The proof is quite similar to the one of Theorem 3.6. Leave it as an exercise in Problem 3.6. □

Theorem 3.18
Let $f \in C^2(Q, \|\cdot\|)$. Then $f \in \mathscr{S}^2_\mu(Q, \|\cdot\|)$ iff

$$\|\nabla^2 f(x)\| \geq \mu, \quad \forall x \in Q.$$

Proof. The proof is quite similar to the one of Theorem 3.7. Leave it as an exercise in Problem 3.7. □

Lemma 3.19 (Polyak-Łojasiewicz condition)
Suppose that $f \in \mathscr{S}^1_\mu(\mathbb{R}^d, \|\cdot\|)$, and that f is bounded below by f^*. Then for any $x \in \mathbb{R}^d$,

$$f(x) - f^* \leq \frac{1}{2\mu}\|\nabla f(x)\|_*^2. \tag{17}$$

Proof. The procedure of derivation is quite similar to that of Corollary 3.10. Leave it as an exercise in Problem 3.8. □

Remark 3.6. *The condition of* (17)*, which is often referred to as* μ***-PŁ condition****, is weaker than the condition of* μ*-strong convexity. In fact, a function that satisfies* μ*-PŁ condition may not even be convex.*

Theorem 3.20
Suppose that $f \in \mathscr{S}^1_\mu(\mathbb{R}^d, \|\cdot\|)$. Then for any $x, y \in \mathbb{R}^d$,

$$f(y) - f(x) - \nabla f(x)^{\mathrm{T}}(y - x) \leq \frac{1}{2\mu}\|\nabla f(x) - \nabla f(y)\|_*^2.$$

Proof. The proof is quite similar to the part in the proof of Theorem 3.12. Leave it as an exercise in Problem 3.9. □

3.1.7 Smooth and strongly convex function

Now, we consider $f \in \mathscr{S}^{1,1}_{\mu,L}(Q, \|\cdot\|)$, *i.e.*, with μ-strong convexity and L-Lipschitz gradient.

Definition 3.7
Let $f \in \mathscr{S}^{1,1}_{\mu,L}(Q, \|\cdot\|)$. The **condition number** of the funtion f is defined as

$$\kappa_f = \frac{L}{\mu}.$$

Theorem 3.21
Let $f \in C^2(Q, \|\cdot\|)$. Then $f \in \mathscr{S}^{2,1}_{\mu,L}(Q, \|\cdot\|)$ iff

$$\mu \leq \|\nabla^2 f(x)\| \leq L, \quad \forall x \in Q.$$

Proof. It follows immediately from Theorem 3.7 and Theorem 3.18. □

3.1.8 Faster convergence to minima

Definition 3.8
Let a positive sequence $\{r_n\}$ converges to 0, and satisfies that

$$\limsup_{n\to\infty} \frac{r_{n+1}}{r_n^p} = C(p).$$

If $p^* = 1$, $C(1) = 1$, it is **subliner convergence**.
If $p^* = 1$, $0 < C(1) < 1$, it is **linear convergence**.
If $p^* > 1$ or $C(1) = 0$, it is **superlinear convergence** (among which a frequently used special case is when $p^* = 2$, we also say that it's **quadratic convergence**).

Theorem 3.22
Suppose that $f \in C^1(\mathbb{R}^d, \|\cdot\|)$, that f is L-smooth, and that f satisfies μ-PŁ condition (bounded below by f^*). In Algorithm 4, for $h_k = \frac{1}{L}$,

$$f(x_T) - f^* \leq (1 - \frac{\mu}{L})^T \cdot (f(x_0) - f^*). \tag{18}$$

Proof. We have

$$f(x_k) - f(x_{k+1}) \geq \frac{1}{2L}\|\nabla f(x_k)\|_*^2 \geq \frac{\mu}{L} \cdot (f(x_k) - f^*).$$

Therefore,

$$f(x_{k+1}) - f^* \leq (1 - \frac{\mu}{L})(f(x_k) - f^*).$$

□

Remark 3.7. *The convergence rate in* (18) *under smoothness and PŁ condition is a linear convergence rate. It's faster than the sublinear rate we previously obtained in Theorem 3.13 for smooth convex functions.*

Remark 3.8. *If the conditional number κ_f is smaller (*i.e., *closer to* 1*), then the Algorithm 4 will converge faster.*

Remark 3.9. *Of course, if we replace the "μ-PŁ condition" in Theorem 3.22 with "μ-strong convexity", then the theorem still holds, since "μ-strong convexity" can imply "μ-PŁ condition".*

3.2 Lower complexity bounds

In Section 3.2, we derive the lower complexity bounds for the problem classes mentioned in Section 3.1.

We take the following assumption[2]:

Assumption 3.9
The iterative method $\mathscr{M}$ generates a sequence of test points x_k *s.t.*

$$x_k \in x_0 + \operatorname{Lin}\{\nabla f(x_0), \ldots, \nabla f(x_{k-1})\}, \quad k \geq 1.$$

3.2.1 Lower bound for smooth convex optimization

Lemma 3.23 (Theorem 2.1.7 in [Nesterov et al., 2018])
For any T, $1 \leq T \leq \frac{1}{2}(d-1)$, and any $x_0 \in \mathbb{R}^d$ there exists a function $f \in \mathscr{F}_L^{\infty,1}(\mathbb{R}^d, \|\cdot\|_2)$ *s.t.* for any first-order method $\mathscr{M}$ satisfying Assumption 3.9 we have

$$f(x_T) - f^* \geq \frac{3L\|x_0 - x^*\|_2^2}{32(T+1)^2}.$$

Proof. For $k < d$, let $A_k \in \mathbb{R}^{d\times d}$ be

$$(A_k)_{i,j} = \begin{cases} 2, & i = j, i \leq k \\ -1, & j \in \{i-1, i+1\}, i \leq k, j \neq k+1 \\ 0, & \text{otherwise} \end{cases}$$

Obviously,

$$0 \preceq A \preceq 4I.$$

Without loss of generality, assume $x_0 = 0$. Then, consider

$$f(x) = \frac{L}{8}x^\top A_{2T+1}x - \frac{L}{4}x^\top e_1. \tag{19}$$

By Assumption 3.9, for the function in (19),

$$x_k \in \operatorname{Lin}\{e_1, \cdots, e_{k-1}\}.$$

Hence x_k^* is the unique solution of $A_k x = e_1$ in $\operatorname{Lin}\{e_1, \cdots, e_{k-1}\}$, *i.e.*,

$$x_k^* = \sum_{i=1}^{k}(1 - \frac{i}{k+1}) \cdot e_i.$$

Then

$$f(x_k^*) = \frac{L}{8}(x_k^*)^\top A_k x_k^* - \frac{L}{4}(x_k^*)^\top e_1 = -\frac{L}{8}(x_k^*)^\top e_1 = -\frac{L}{8}\left(1 - \frac{1}{k+1}\right).$$

[2]Under Assumption 3.9, the steepest descent method (Algorithm 4) is excluded.

Since

$$x^* = x^*_{2T+1} = \sum_{i=1}^{2T+1} (1 - \frac{i}{2T+2}) \cdot e_i,$$

we conclude the proof by obtaining

$$\frac{f(x_T) - f^*}{\|x_0 - x^*\|_2^2} \geq \frac{\frac{L}{8}\left(-1 + \frac{1}{T+1} + 1 - \frac{1}{2T+2}\right)}{\frac{1}{3}(2T+2)} = \frac{3}{8}L \cdot \frac{1}{4(T+1)^2}.$$

□

Theorem 3.24
For $\mathscr{F}_L^{1,1}(\mathbb{R}^d, \|\cdot\|_2)$, a first-order method under Assumption 3.9 needs at least

$$\Omega\left(\min(d, \frac{1}{\sqrt{\varepsilon}})\right)$$

iterations to guarantee ε-accuracy.

Proof. It follow immediately from Lemma 3.23. □

3.2.2 Lower bound for smooth and strongly convex optimization

Theorem 3.25 ([Nemirovskij and Yudin, 1983])
For $\mathscr{S}_{\mu,L}^{1,1}(\mathbb{R}^d, \|\cdot\|_2)$, a first-order method under Assumption 3.9 needs at least

$$\Omega\left(\min(d, \sqrt{\kappa}\log(\frac{1}{\varepsilon}))\right)$$

iterations to guarantee ε-accuracy.

3.3 Accelerated gradient method

The Section 3.3 is based on Subsection 2.2.1 in [Nesterov et al., 2018].

3.3.1 Estimating sequences

Definition 3.10
A pair of sequences $\{\phi_k(x)\}_{k=0}^{\infty}$ and $\{\lambda_k\}_{k=0}^{\infty}$, $\lambda_k \geq 0$, are called the **estimating sequences** of the function $f(\cdot)$ if

$$\lambda_k \to 0,$$

and for any $x \in \mathbb{R}^d$ and all $k \geq 0$ we have

$$\phi_k(x) \leq (1 - \lambda_k) f(x) + \lambda_k \phi_0(x). \tag{20}$$

Lemma 3.26
If for some seuqence of points $\{x_k\}$ we have

$$f(x_k) \leq \phi_k^* \stackrel{\text{def}}{=} \min_{x \in \mathbb{R}^n} \phi_k(x),$$

then

$$f(x_k) - f(x^*) \leq \lambda_k [\phi_0(x^*) - f(x^*)] \to 0, \tag{21}$$

where x^* is a minimum of $f(\cdot)$.

Proof.

$$\begin{aligned} f(x_k) \leq \phi_k^* &= \min_{x \in \mathbb{R}^n} \phi_k(x) \leq \min_{x \in \mathbb{R}^n} \left[(1 - \lambda_k) f(x) + \lambda_k \phi_0(x) \right] \\ &\leq (1 - \lambda_k) f(x^*) + \lambda_k \phi_0(x^*). \end{aligned}$$

□

Thus, for any sequence $\{x_k\}$, satisfying (21), we can derive its rate of convergence directly from the convergence rate of the sequence $\{\lambda_k\}$. Then we try to understand the two questions: (i) how to form the estimating sequences, and (ii) how to satisfy inequalities (21). We first try to answer the first question.

Lemma 3.27
Assume that:

1. a function $f(\cdot)$ belongs to $\mathscr{S}_{\mu,L}^{1,1}(\mathbb{R}^d, \|\cdot\|_2)^a$,
2. $\phi_0(\cdot)$ is an arbitrary convex function on $\mathbb{R}^d$,
3. $\{y_k\}_{k=0}^{\infty}$ is an arbitrary sequence of points in $\mathbb{R}^d$,
4. the coefficients $\{\alpha_k\}_{k=0}^{\infty}$ satisfies conditions $\alpha_k \in (0,1)$ and $\sum_{k=0}^{\infty} \alpha_k = \infty$,
5. we choose $\lambda_0 = 1$.

Then the pair of sequences $\{\phi_k(\cdot)\}_{k=0}\infty$ and $\{\lambda_k\}_{k=0}^{\infty}$ defined recursively by the relations

$$\begin{aligned} \lambda_{k+1} &= (1 - \alpha_k) \lambda_k, \\ \phi_{k+1}(x) &= (1 - \alpha_k) \phi_k(x) + \alpha_k \left[f(y_k) + \langle \nabla f(y_k), x - y_k \rangle + \frac{\mu}{2} \|x - y_k\|_2^2 \right], \end{aligned} \tag{22}$$

are estimating sequences.

[a] We hereby use this notation for simplicity, where we specifically allow $\mu = 0$ throughout Section 3.3.

Proof. Indeed, $\phi_0(x) = (1-\lambda_0) f(x) + \lambda_0 \phi_0(x)$. Let (20) holds for k. Then

$$\begin{aligned}\phi_{k+1}(x) &\leq (1-\alpha_k)\,\phi_k(x) + \alpha_k f(x)\\ &= (1-(1-\alpha_k)\,\lambda_k)\,f(x) + (1-\alpha_k)\,(\phi_k(x) - (1-\lambda_k)\,f(x))\\ &\leq (1-(1-\alpha_k)\,\lambda_k)\,f(x) + (1-\alpha_k)\,\lambda_k\phi_0(x)\\ &\leq (1-\lambda_{k+1})\,f(x) + \lambda_{k+1}\phi_0(x).\end{aligned}$$

It remains to notice that $\lambda_k \to 0$. □

At this moment, we are also free in our choice of initial function $\phi_0(x)$. Let us choose it as a simple quadratic function. Then, we can obtain a closed form recurrence for values ϕ_k^*.

Lemma 3.28
Let $\phi_0(x) = \phi_0^* + \frac{\gamma_0}{2}\|x - v_0\|_2^2$. Then, in view of (22), we have

$$\phi_k(x) = \phi_k^* + \frac{\gamma_k}{2}\|x - v_k\|_2^2,$$

where

$$\begin{aligned}\gamma_{k+1} &= (1-\alpha_k)\,\gamma_k + \alpha_k\mu,\\ v_{k+1} &= \frac{1}{\gamma_{k+1}}\left[(1-\alpha_k)\,\gamma_k v_k + \alpha_k\mu y_k - \alpha_k\nabla f(y_k)\right],\\ \phi_{k+1}^* &= (1-\alpha_k)\,\phi_k^* + \alpha_k f(y_k) - \frac{\alpha_k^2}{2\gamma_{k+1}}\|\nabla f(y_k)\|_2^2\\ &\quad + \frac{\alpha_k(1-\alpha_k)\,\gamma_k}{\gamma_{k+1}}\left(\frac{\mu}{2}\|y_k - v_k\|_2^2 + \langle\nabla f(y_k), v_k - y_k\rangle\right).\end{aligned}$$

Proof. Leave as an exercise in Problem 3.12 (just some basic algebraic transformations of (22)). □

3.3.2 Accelerated gradient method

We try to answer the second question of how to satisfy inequalities (21) now. Indeed, assume that we already have x_k *s.t.*

$$\phi_k^* \geq f(x_k).$$

Then let's try to find x_{k+1} satisfying $\phi_{k+1}^* \geq f(x_{k+1})$. Note that we still have the freedom to choose y_k and α_k. In view of Lemma 3.28,

$$\begin{aligned}
\phi_{k+1}^* &= (1-\alpha_k)\,\phi_k^* + \alpha_k f(y_k) - \frac{\alpha_k^2}{2\gamma_{k+1}}\|\nabla f(y_k)\|_2^2 \\
&\quad + \frac{\alpha_k(1-\alpha_k)\gamma_k}{\gamma_{k+1}}\left(\frac{\mu}{2}\|y_k - v_k\|_2^2 + \langle \nabla f(y_k), v_k - y_k\rangle\right) \\
&\geq (1-\alpha_k)\,f(x_k) + \alpha_k f(y_k) - \frac{\alpha_k^2}{2\gamma_{k+1}}\|\nabla f(y_k)\|_2^2 \\
&\quad + (1-\alpha_k)\left\langle \nabla f(y_k), \frac{\alpha_k\gamma_k}{\gamma_{k+1}}(v_k - y_k)\right\rangle \\
&\geq f(y_k) - \frac{\alpha_k^2}{2\gamma_{k+1}}\|\nabla f(y_k)\|_2^2 + (1-\alpha_k)\left\langle \nabla f(y_k), \frac{\alpha_k\gamma_k}{\gamma_{k+1}}(v_k - y_k) + x_k - y_k\right\rangle.
\end{aligned}$$

Hence, it's sufficient to ensure $\phi_{k+1}^* \geq f(x_{k+1})$ by setting α_k and y_k according to

$$\begin{cases} \frac{\alpha_k^2}{2\gamma_{k+1}} = \frac{1}{2L}, \\ \frac{\alpha_k\gamma_k}{\gamma_{k+1}}(v_k - y_k) + x_k - y_k = 0, \end{cases}$$

and x_{k+1} *s.t.*

$$f(x_{k+1}) \leq y_k - \frac{1}{2L}\|\nabla f(y_k)\|_2^2.$$

Therefore, we come to Algorithm 5, which is often addressed as **Accelerated Gradient Method**.

Algorithm 5 Accelerated gradient method

1: Choose $x_0 \in \mathbb{R}^d$, some $\gamma_0 > 0$, and set $v_0 = x_0$.
2: **for** $k = 0, 1, \cdots$ **do**
3: Compute $\alpha_k \in (0,1)$ *s.t.* $L\alpha_k^2 = (1-\alpha_k)\gamma_k + \alpha_k\mu$.
4: $\gamma_{k+1} := (1-\alpha_k)\gamma_k + \alpha_k\mu$.
5: $y_k := \frac{1}{\gamma_k + \alpha_k\mu}[\alpha_k\gamma_k v_k + \gamma_{k+1}x_k]$.
6: Choose x_{k+1} *s.t.* $f(x_{k+1}) \leq y_k - \frac{1}{2L}\|\nabla f(y_k)\|_2^2$.
7: $v_{k+1} := \frac{1}{\gamma_{k+1}}[(1-\alpha_k)\gamma_k v_k + \alpha_k\mu y_k - \alpha_k\nabla f(y_k)]$.
8: **end for**

3.3.3 Convergence analysis

Theorem 3.29
Algorithm 5 generates a sequence of points $\{x_k\}_{k=0}^{\infty}$ *s.t.*

$$f(x_k) - f^* \leq \lambda_k\left[f(x_0) - f^* + \frac{\gamma_0}{2}\|x_0 - x^*\|_2^2\right],$$

where $\lambda_0 = 1$ and $\lambda_k = \Pi_{i=0}^{k-1}(1-\alpha_i)$.

Proof. It follows immediately from Lemma 3.26. □

Theorem 3.30
In Algorithm 5, for $\gamma_0 \in (\mu, 4L+\mu]$, we have

$$\lambda_k \le \frac{4L}{(\gamma_0-\mu)(k+1)^2}.$$

If $\mu > 0$, then for $\gamma_0 = \mu$, we have

$$\lambda_k = \left(1-\sqrt{\frac{\mu}{L}}\right)^k.$$

Proof. 1. Suppose $\gamma_0 \in (\mu, 4L+\mu]$. Note that

$$\gamma_{k+1}-\mu = (1-\alpha_k)(\gamma_k-\mu) = \frac{\lambda_{k+1}}{\lambda_k}(\gamma_k-\mu) = \cdots = \frac{\lambda_{k+1}}{\lambda_0}(\gamma_0-\mu) = \lambda_{k+1}(\gamma_0-\mu).$$

Then we have

$$1-\frac{\lambda_{k+1}}{\lambda_k} = \alpha_k = \sqrt{\frac{\gamma_{k+1}}{L}} = \sqrt{\frac{\mu}{L}+\frac{\gamma_0-\mu}{L}\lambda_{k+1}}. \tag{23}$$

Assume $\lambda_k \le \frac{C}{(k+1)^2}$ holds for some constant C. Then, in view of (23),

$$\lambda_{k+1} \le \frac{C}{(k+2)^2}$$

is equivalent to

$$1-\frac{C}{(k+2)^2\lambda_k} \le \sqrt{\frac{\mu}{L}+\frac{\gamma_0-\mu}{L}\cdot\frac{C}{(k+2)^2}}.$$

In fact,

$$1-\frac{C}{(k+2)^2\lambda_k} \le 1-\frac{(k+1)^2}{(k+2)^2} \le \sqrt{\frac{\mu}{L}+\frac{\gamma_0-\mu}{L}\cdot\frac{C}{(k+2)^2}},$$

where the last inequality simply follows when $C := \frac{4L}{\gamma_0-\mu}$.
It's also required that $\lambda_0 \le C$, *i.e.*, $\gamma_0 \le 4L+\mu$.

2. Suppose $\mu > 0$ and $\gamma_0 = \mu$. Then $\gamma_k \equiv \mu$ and $\alpha_k \equiv \sqrt{\frac{\mu}{L}}$. Hence

$$\lambda_k = \left(1-\sqrt{\frac{\mu}{L}}\right)^k.$$

□

Remark 3.10. *By Theorem 3.29 and Theorem 3.30, Algorithm 5 requires $O(\frac{1}{\sqrt{\varepsilon}})$ iterations to achieve ε-accuracy for smooth convex optimization, and $O(\sqrt{\kappa}\cdot\log\frac{1}{\varepsilon})$ iterations for smooth and strongly convex optimization, respectively. Thus, in view of the lower bounds shown in Section 3.2, Algorithm 5 is considered optimal for both of the optimization problems.*

In fact, if $x_{k+1} := y_k - \frac{1}{L}\|\nabla f(y_k)\|_2^2$, then Algorithm 5 can be rewritten into a simpler form as Algorithm 6. The proof of the equivalence is left as an exercise in Problem 3.13.

Algorithm 6 A simpler form of accelerated gradient method

1: Choose $x_0 \in \mathbb{R}^d$, some $\alpha_0 \in (0,1)$, and set $y_0 = x_0$.
2: **for** $k = 0, 1, \cdots$ **do**
3: $\quad x_{k+1} := y_k - \frac{1}{L}\nabla f(y_k)$.
4: $\quad$ Compute $\alpha_{k+1} \in (0,1)$ *s.t.*

$$\alpha_{k+1}^2 = (1-\alpha_{k+1})\,\alpha_k^2 + \sqrt{\frac{\mu}{L}}\cdot\alpha_{k+1}.$$

5: $\quad \beta_k := \frac{\alpha_k(1-\alpha_k)}{\alpha_k^2+\alpha_{k+1}}$ and $y_{k+1} := x_{k+1} + \beta_k\,(x_{k+1} - x_k)$.
6: **end for**

Further, if $\mu > 0$, we can set $\alpha_k \equiv \sqrt{\frac{\mu}{L}}$, and $\beta_k \equiv \frac{\sqrt{\kappa_f}-1}{\sqrt{\kappa_f}+1}$.

3.4 Key points and problems

Key Points: gradient descent; smooth convex functions; smooth and strongly convex functions; lower bound; accelerated gradient method.

Problem 3.1
$f \in C_L^{1,1}(Q, \|\cdot\|)$ can imply

$$|f(y) - f(x) - \langle\nabla f(x), y - x\rangle| \le \frac{L}{2}\|y-x\|^2, \quad \forall x, y \in Q. \tag{24}$$

Hints. For any $x, y \in Q$,

$$f(y) = f(x) + \nabla f(x)^{\mathrm{T}}(y-x) + \int_0^1 [\nabla f(x + t(y-x)) - \nabla f(x)]^{\mathrm{T}}(y-x)dt.$$

Then

$$\begin{aligned}|f(y)-f(x)-\langle\nabla f(x),y-x\rangle| &= \left|\int_0^1 [\nabla f(x+t(y-x))-\nabla f(x)]^{\mathrm{T}}(y-x)dt\right| \\ &\le \int_0^1 \|\nabla f(x+t(y-x))-\nabla f(x)\|_*\cdot\|y-x\|dt \\ &\le \int_0^1 tL\|y-x\|^2dt \\ &= \frac{L}{2}\|y-x\|^2.\end{aligned}$$

□

Remark 3.11. *In fact,* (24) *is (slightly) stronger than the L-smoothness of f, and is (slightly) weaker than $f\in C_L^{1,1}(Q,\|\cdot\|)$.*

Problem 3.2
Suppose $f\in C^1(Q,\|\cdot\|)$, $L>0$. Show the equivalence between (12) and (13) (for any $x,y\in Q$ and $0\le\lambda\le 1$).

Problem 3.3
Suppose that $f\in\mathscr{F}_L^{1,1}(\mathbb{R}^d,\|\cdot\|_2)$. Let x^* be a minimum of f, and then, let $R_0=\|x_0-x^*\|_2$. In Algorithm 3, for $h_k=\frac{1}{L}$,

$$f(x_T)-f(x^*)\le\frac{LR_0^2}{2T}.$$

Hints. By the convexity of f, we have

$$\begin{aligned}\|x_{k+1}-x^*\|_2^2 &= \|x_k-h_k\nabla f(x_k)-x^*\|_2^2 \\ &= \|x_k-x^*\|_2^2-2h_k\langle x_k-x^*,\nabla f(x_k)\rangle+h_k^2\|\nabla f(x_k)\|_2^2 \\ &\le \|x_k-x^*\|_2^2-2h_k(f(x_k)-f(x^*))+h_k^2\|\nabla f(x_k)\|_2^2 \\ &= \|x_k-x^*\|_2^2-\frac{2}{L}(f(x_k)-f(x^*))+\frac{1}{L^2}\|\nabla f(x_k)\|_2^2,\end{aligned}$$

or

$$2L(f(x_k)-f(x^*))+L^2(\|x_{k+1}-x^*\|_2^2-\|x_k-x^*\|_2^2)\le\|\nabla f(x_k)\|_2^2.$$

Summing over T from 0 to $T-1$ and devided both sides by $2L$, together with Theorem 3.11, we have

$$\begin{aligned}&\left(\sum_{k=0}^{T-1}f(x_k)-f(x^*)\right)+\frac{L}{2}(\|x_T-x^*\|_2^2-\|x_0-x^*\|_2^2) \\ &\le\frac{1}{2L}\sum_{k=0}^{T-1}\|\nabla f(x_k)\|_2^2 \\ &\le f(x_0)-f(x_T),\end{aligned}$$

or

$$\left(\sum_{k=1}^{T} f(x_k) - f(x^*)\right) + \frac{L}{2}(\|x_T - x^*\|_2^2 - \|x_0 - x^*\|_2^2) \leq 0.$$

Hence,

$$\begin{aligned} f(x_T) - f(x^*) &\leq \frac{1}{T}\left(\sum_{k=1}^{T} f(x_k) - f(x^*)\right) \\ &\leq \frac{L(\|x_0 - x^*\|_2^2 - \|x_T - x^*\|_2^2)}{2T} \\ &\leq \frac{L\|x_0 - x^*\|_2^2}{2T}. \end{aligned}$$

□

Problem 3.4 (no overshooting)
Suppose that $\mathrm{dom}(f)$ is open, and that f is differentiable. Let $x \in \mathrm{dom}(f)$ *s.t.* $\nabla f(x) \neq 0$. Suppose that f is L-smooth over the line segment connecting x and $x' = x - h\nabla f(x)$, where $0 < h \leq 1/L$. Then $\langle x' - x, x' - x^* \rangle \leq 0$.

Problem 3.5
Prove Theorem 3.15.

Problem 3.6
Prove Theorem 3.17.

Problem 3.7
Prove Theorem 3.18.

Problem 3.8
Prove Lemma 3.19.

Problem 3.9
Prove Theorem 3.20.

Problem 3.10
Let $f \in C^0(Q, \|\cdot\|_2)$. Then, show that

- f is L-smooth iff $g(x) = \frac{L}{2}x^{\mathrm{T}}x - f(x)$ is convex.
- f is μ-strongly convex iff $h(x) = f(x) - \frac{\mu}{2}x^{\mathrm{T}}x$ is convex.

Hints. $\lambda\|x\|_2^2 + (1-\lambda)\|y\|_2^2 - \|\lambda x + (1-\lambda)y\|_2^2 = \lambda(1-\lambda)\|x-y\|_2^2$. □

Problem 3.11
If $f \in \mathscr{S}_{\mu,L}^{1,1}(\mathbb{R}^d, \|\cdot\|_2)$, then for any $x, y \in \mathbb{R}^d$,

$$\langle \nabla f(x) - \nabla f(y), x - y\rangle \geq \frac{\mu L}{\mu + L}\|x-y\|_2^2 + \frac{1}{\mu+L}\|\nabla f(x) - \nabla f(y)\|_2^2. \tag{25}$$

Hints. Let $\phi(x) = f(x) - \frac{\mu}{2}x^{\mathrm{T}}x$. □

Problem 3.12
Prove Lemma 3.28.

Hints. Proof by induction. □

Problem 3.13
Show that Algorithm 5 and Algorithm 6 are actually equivalent.

4 Constrained optimization

4.1 Optimization over simple sets

Let's now consider the following problem:

$$\min_{x \in Q} f(x), \quad f \in \mathscr{F}^1(Q, \|\cdot\|), \tag{26}$$

where Q is a closed convex set.

Theorem 4.1
Let $f \in \mathscr{F}^1(Q, \|\cdot\|_2)$ and the set Q be closed and convex. A point x^* is a solution of (26) iff

$$\langle \nabla f(x^*), x - x^* \rangle \geq 0$$

for all $x \in Q$.

Proof. (if) For any $x \in Q$, $f(x) \geq f(x^*) + \langle \nabla f(x^*), x - x^* \rangle \geq f(x^*)$.
(only if) For any $x \in Q$, let $\phi(\lambda) = f(x^* + \lambda(x - x^*))$, $\lambda \in [0, 1]$. Then $\phi'(0) \geq 0$. □

Theorem 4.2 (Theorem 2.2.10 in [Nesterov et al., 2018])
Let $f \in \mathscr{S}^1_\mu(Q, \|\cdot\|)$ with $\mu > 0$ and the set Q be closed and convex. Then there exists a unique solution x^* of (26).

Proof. Leave it as an exercise in Problem 4.1. □

4.1.1 Projected gradient descent over (simple) closed convex sets

Theorem 4.3
Let x^* be a solution of (26). Then, for any $\gamma > 0$, we have

$$\Pi_Q(x^* - \gamma \nabla f(x^*)) = x^*.$$

Proof. Let $y = \Pi_Q(x^* - \gamma \nabla f(x^*))$. Then

$$\langle x^* - \gamma \nabla f(x^*) - y, x^* - y \rangle \leq 0.$$

By Theorem 4.1,

$$\langle -\gamma \nabla f(x^*), y - x^* \rangle \leq 0.$$

Adding the above two inequalities, we get $\|x^* - y\|^2 \leq 0$. Hence $y = x^*$. □

Definition 4.1
Fix some $\gamma > 0$. Define

$$x_Q(\bar{x};\gamma) = \arg\min_{x\in Q}\left[f(\bar{x}) + \langle \nabla f(\bar{x}), x-\bar{x}\rangle + \frac{1}{2\gamma}\|x-\bar{x}\|^2\right], \quad g_Q(\bar{x};\gamma) = \frac{1}{\gamma}(\bar{x} - x_Q(\bar{x};\gamma)).$$

We call $x_Q(\bar{x};\gamma)$ the **gradient mapping**, and $g_Q(\bar{x};\gamma)$ the **reduced gradient** of the function f on Q.

Theorem 4.4
Let $f \in \mathscr{S}^{1,1}_{\mu,L}(Q, \|\cdot\|_2)$, $\gamma \le \frac{1}{L}$, and $\bar{x} \in \mathbb{R}^d$. Then for any $x \in Q$, we have

$$f(x) \ge f(x_Q(\bar{x};\gamma)) + \langle g_Q(\bar{x};\gamma), x-\bar{x}\rangle + \frac{1}{2\gamma}\|g_Q(\bar{x};\gamma)\|_2^2 + \frac{\mu}{2}\|x-\bar{x}\|_2^2.$$

Proof. Let x_Q denote $x_Q(\bar{x};\gamma)$, g_Q denote $g_Q(\bar{x};\gamma)$, and

$$\phi(x) = f(\bar{x}) + \langle \nabla f(\bar{x}), x-\bar{x}\rangle + \frac{1}{2\gamma}\|x-\bar{x}\|_2^2.$$

Then $\nabla\phi(x) = \nabla f(\bar{x}) + \frac{1}{\gamma}(x-\bar{x})$. Hence for any $x \in Q$,

$$\langle \nabla f(\bar{x}) - g_Q, x - x_Q\rangle = \langle \nabla\phi(x_Q), x - x_Q\rangle \ge 0.$$

Hence,

$$\begin{aligned}
f(x) - \frac{\mu}{2}\|x-\bar{x}\|_2^2 &\ge f(\bar{x}) + \langle \nabla f(\bar{x}), x-\bar{x}\rangle \\
&= f(\bar{x}) + \langle \nabla f(\bar{x}), x_Q - \bar{x}\rangle + \langle \nabla f(\bar{x}), x - x_Q\rangle \\
&\ge f(\bar{x}) + \langle \nabla f(\bar{x}), x_Q - \bar{x}\rangle + \langle g_Q, x - x_Q\rangle \\
&= f(\bar{x}) + \langle \nabla f(\bar{x}), x_Q - \bar{x}\rangle + \langle g_Q, x - \bar{x}\rangle + \gamma\|g_Q\|_2^2 \\
&\ge f(x_Q) - \frac{1}{2\gamma}\|\bar{x} - x_Q\|_2^2 + \langle g_Q, x-\bar{x}\rangle + \gamma\|g_Q\|_2^2 \\
&= f(x_Q) + \langle g_Q, x-\bar{x}\rangle + \frac{\gamma}{2}\|g_Q\|_2^2.
\end{aligned}$$

□

Remark 4.1. *It immediately follows that*

$$f(x_Q) \le f(\bar{x}) - \frac{\gamma}{2}\|g_Q\|_2^2,$$

and that

$$\langle g_Q, \bar{x} - x^*\rangle \ge \frac{\gamma}{2}\|g_Q\|_2^2 + \frac{\mu}{2}\|\bar{x}-x^*\|_2^2 + \frac{\mu}{2}\|x_Q - x^*\|_2^2.$$

Let's show that we can use the gradient mapping as in Algorithm 7 to solve (26) where $f \in \mathscr{S}^{1,1}_{\mu,L}(Q, \|\cdot\|_2)$.

Algorithm 7 Projected gradient descent

1: Initialize $x_0 \in Q$ and a parameter $\gamma > 0$.
2: **for** $k = 0, 1, \cdots$ **do**
3: $\quad x_{k+1} = x_k - \gamma g_Q(x_k; \gamma)$ (*i.e.*, $x_{k+1} = x_Q(x_k; \gamma) = \Pi_Q(x_k - \gamma \nabla f(x_k))$)
4: **end for**

Theorem 4.5
Let $f \in \mathscr{S}^{1,1}_{\mu,L}(\mathbb{R}^d, \|\cdot\|_2)$. For $\gamma < \frac{2}{L+\mu}$, we have

$$\|x_T - x^*\|_2 \le (1 - \mu\gamma)^T \|x_0 - x^*\|_2.$$

Proof.

$$\begin{aligned}
&\|x_{k+1} - x^*\|_2^2 \\
&= \|\Pi_Q(x_k - \gamma \nabla f(x_k)) - \Pi_Q(x^* - \nabla f(x^*))\|_2^2 \\
&\le \|x_k - x^* - \gamma(\nabla f(x_k) - \nabla f(x^*))\|_2^2 \\
&= \|x_k - x^*\|_2^2 - 2\gamma \langle \nabla f(x_k) - \nabla f(x^*), x_k - x^* \rangle + \gamma^2 \|\nabla f(x_k) - \nabla f(x^*)\|_2^2 \\
&\overset{(25)}{\le} (1 - \mu\gamma)^2 \|x_k - x^*\|_2^2.
\end{aligned}$$

□

4.1.2 Conditional gradient method over compact convex sets

The conditional gradient method (Algorithm 8), also known as Frank-Wolfe method, is an alternative to projected gradient descent, and it does not involve projections.

Algorithm 8 Conditional gradient method

1: Initialize $x_0 \in Q$.
2: **for** $k = 0, 1, \cdots$ **do**
3: $\quad s_k = \arg\min_{s \in Q} \langle s, \nabla f(x_k) \rangle$
4: $\quad x_{k+1} = (1 - \gamma_k) x_k + \gamma_k s_k$, where $\gamma_k = \frac{2}{k+2}$
5: **end for**

Remark 4.2. *The g_k is often referred to as the **Frank-Wolfe gap**. When f is convex, we have*

$$g_k \ge \langle \nabla f(x_k), x_k - x^* \rangle \ge f(x_k) - f(x^*).$$

Lemma 4.6
Let $f \in C_L^{1,1}(Q, \|\cdot\|)$ and Q be convex and compact. We have

$$f(x_{k+1}) \le f(x_k) - \gamma_k g_k + \gamma_k^2 \frac{L}{2} D^2,$$

where $D = \max\{\|x - x'\| : x, x' \in Q\}$.

Proof.

$$\begin{aligned} f(x_{k+1}) &\le f(x_k) + \gamma_k \langle \nabla f(x_k), s_k - x_k \rangle + \frac{L}{2}\gamma_k^2 \|s_k - x_k\|^2 \\ &= f(x_k) - \gamma_k g_k + \frac{L}{2}\gamma_k^2 \|s_k - x_k\|^2 \\ &\le f(x_k) - \gamma_k g_k + \gamma_k^2 \frac{L}{2} D^2. \end{aligned}$$

□

Theorem 4.7
Let $f \in \mathscr{F}_L^{1,1}(Q, \|\cdot\|)$ and Q be ocnvex and compact. We have

$$f(x_T) - f(x^*) \le \frac{2LD^2}{T+1},$$

where $D = \max\{\|x - x'\| : x, x' \in Q\}$.

Proof.

$$\begin{aligned} &(k+1)(k+2)[f(x_{k+1}) - f(x^*)] - k(k+1)[f(x_k) - f(x^*)] \\ &\le (k+1)(k+2)[f(x_k) - \gamma_k g_k + \gamma_k^2 \frac{L}{2} D^2] - k(k+1)[f(x_k) - f(x^*)] \\ &\le (k+1)(k+2)[f(x_k) - \gamma_k (f(x_k) - f(x^*)) + \gamma_k^2 \frac{L}{2} D^2] - k(k+1)[f(x_k) - f(x^*)] \\ &= [(1-\gamma_k)(k+1)(k+2) - k(k+1)](f(x_k) - f(x^*)) + (k+1)(k+2)\gamma_k^2 \frac{L}{2} D^2 \\ &= 0 + \frac{k+1}{k+2} \cdot 2LD^2 \le 2LD^2. \end{aligned}$$

Hence

$$T(T+1)[f(x_T) - f(x^*)] \le 2TLD^2.$$

□

4.2 Optimization with functional constraints

4.2.1 The minimax problem

4.2.2 Optimization over simple sets and with functional constraints

4.3 Key points and problems

Key Points: .

Problem 4.1
Prove Theorem 4.2.

Problem 4.2
In fact, the projected gradient descent algorithm (Algorithm 7) can be accelerated. Try to give the accelerated version of Algorithm 7, and then show the convergence rate.

5 Composite proximal optimization

In this chapter, let's consider the problem of

$$\min_{x\in\mathbb{R}^d} f(x)+g(x), \tag{27}$$

where $f\in\mathscr{F}_L^{1,1}(\mathbb{R}^d,\|\cdot\|_2)$, and $g:\mathbb{R}^d\to(-\infty,+\infty]$ is a proper, closed, and convex[3] function. We have access to a **proximal operator** of g[4], $\mathrm{prox}_\gamma(\cdot):\mathbb{R}^d\to\mathbb{R}^d$ ($\gamma>0$), defined as

$$\mathrm{prox}_\gamma(y)=\arg\min_{x\in\mathbb{R}^d}\left[\frac{1}{2}\|x-y\|_2^2+\gamma g(x)\right].$$

Theorem 5.1 (firm nonexpansiveness)
For any $x,y\in\mathbb{R}^d$,

$$\langle\mathrm{prox}_\gamma(x)-\mathrm{prox}_\gamma(y),x-y\rangle\geq\|\mathrm{prox}_\gamma(x)-\mathrm{prox}_\gamma(y)\|_2^2.$$

Proof. By the optimal condition of minimization,

$$\frac{1}{\gamma}(x-\mathrm{prox}_\gamma(x))\in\partial g(\mathrm{prox}_\gamma(x)),$$

$$\frac{1}{\gamma}(y-\mathrm{prox}_\gamma(y))\in\partial g(\mathrm{prox}_\gamma(y)).$$

Then, since g is convex, we have

$$\frac{1}{\gamma}\left\langle(x-\mathrm{prox}_\gamma(x))-(y-\mathrm{prox}_\gamma(y)),\mathrm{prox}_\gamma(x)-\mathrm{prox}_\gamma(y)\right\rangle\geq 0.$$

□

Corollary 5.2
For any $x,y\in\mathbb{R}^d$,

$$\|\mathrm{prox}_\gamma(x)-\mathrm{prox}_\gamma(y)\|_2\leq\|x-y\|_2$$

5.1 Proximal gradient descent

[3]The domain of g is $\mathrm{dom}(g)=\{x\in\mathbb{R}^d:g(x)<+\infty\}$. Say g is proper if $\mathrm{dom}(g)\neq\emptyset$. Say g is closed if $\mathrm{epi}(g)$ is closed on $\mathrm{dom}(g)$. Say g is convex if g is a convex function on $\mathrm{dom}(g)$.

[4]Usually, the function g is simple (*e.g.*, a norm, a indicator function, *etc.*).

Theorem 5.3
Let x^* be a solution of (27). Then

$$\text{prox}_\gamma(x^* - \gamma\nabla f(x^*)) = x^*.$$

Proof. Let $y = \text{prox}_\gamma(x^* - \gamma\nabla f(x^*))$. Then, by the optimal condition of y,

$$-\frac{1}{\gamma}\left[y - (x^* - \gamma\nabla f(x^*))\right] \in \partial g(y).$$

Also, by the optimal condition of x^*,

$$-\nabla f(x^*) \in \partial g(x^*).$$

Then, by the convexity of g,

$$\langle \frac{1}{\gamma}(x^* - y), y - x^*\rangle = \langle -\frac{1}{\gamma}\left[y - (x^* - \gamma\nabla f(x^*))\right] + \nabla f(x^*), y - x^*\rangle \geq 0.$$

Hence $y = x^*$. □

The proximal gradient descent method (Algorithm 9), also known as iterative shrinkage-thresholding algorithm (ISTA), is a basic method for solving (27).

Algorithm 9 Proximal gradient descent

1: Initialize $x_0 \in \mathbb{R}^d$.
2: **for** $k = 0, 1, \cdots$ **do**
3: $\quad x_{k+1} = \text{prox}_\gamma(x_k - \gamma\nabla f(x_k))$
4: **end for**

Remark 5.1. *The projected gradient descent (Algorithm 7) can be formulated as a special case of proximal gradient descent, where*

$$g(x) = \begin{cases} 0, & x \in Q, \\ +\infty, & x \notin Q. \end{cases}$$

Theorem 5.4
Let $f \in \mathscr{S}^{1,1}_{\mu,L}(\mathbb{R}^d, \|\cdot\|_2)$. For $\gamma < \frac{2}{L+\mu}$, we have

$$\|x_T - x^*\|_2 \leq (1 - \mu\gamma)^T \|x_0 - x^*\|_2.$$

Proof.

$$
\begin{aligned}
&\|x_{k+1} - x^*\|_2^2 \\
&= \|\operatorname{prox}_\gamma(x_k - \gamma\nabla f(x_k)) - \operatorname{prox}_\gamma(x^* - \nabla f(x^*))\|_2^2 \\
&\leq \|x_k - x^* - \gamma(\nabla f(x_k) - \nabla f(x^*))\|_2^2 \\
&= \|x_k - x^*\|_2^2 - 2\gamma\langle\nabla f(x_k) - \nabla f(x^*), x_k - x^*\rangle + \gamma^2\|\nabla f(x_k) - \nabla f(x^*)\|_2^2 \\
&\overset{(25)}{\leq} (1 - \mu\gamma)^2\|x_k - x^*\|_2^2.
\end{aligned}
$$

□

5.2 FISTA

6 Non-smooth optimization

6.1 Subgradient descent method

6.2 Lower bound for non-smooth optimization

7 Mirror descent

7.1 Bregman divergence and its exhaustiveness

[Banerjee et al., 2005] provides necessary and sufficient conditions (roughly speaking, being Bregman divergence) for the general loss functions under which the conditional expectation is the unique optimal predictor.

7.2 Mirror descent method

8 Stochastic optimization

8.1 Lower bound for stochastic optimization

8.2 Optimal method for stochastic composite optimization

8.3 High probability convergence

9 Topics in distributed optimization

9.1 ADMM

9.2 Finite sum problem

9.2.1 Dual formulation (SDCA)

9.2.2 Variance reduction (SVRG)

9.2.3 Accelerating variance-reduced stochastic gradient methods

9.3 Federated optimization

9.3.1 FedAvg

9.3.2 Lower bound for communication complexity

With the bound for communication complexity, due to the fact that the federated learning can be formulated as a composite proximal optimization, we can conclude that the FISTA algorithm is optimal in terms of the proximal oracle complexity.

9.3.3 Proxskip

References

[Banerjee et al., 2005] Banerjee, A., Guo, X., and Wang, H. (2005). On the optimality of conditional expectation as a bregman predictor. *IEEE Transactions on Information Theory*, 51(7):2664–2669.

[Boyd and Vandenberghe, 2004] Boyd, S. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

[Nemirovskij and Yudin, 1983] Nemirovskij, A. S. and Yudin, D. B. (1983). *Problem complexity and method efficiency in optimization*. Wiley-Interscience.

[Nesterov et al., 2018] Nesterov, Y. et al. (2018). *Lectures on convex optimization*, volume 137. Springer.

[Shreve, 2005] Shreve, S. (2005). *Stochastic calculus for finance I: the binomial asset pricing model*. Springer Science & Business Media.

[Wikipedia contributors, 2022] Wikipedia contributors (2022). Hyperplane separation theorem — Wikipedia, the free encyclopedia.