Lecture 2

Smooth and Convex Functions

Classes of differentiable functions

Let $\mathcal{S}$ be a subset of $\mathbb{R}^d$ and $\mathcal{C}_L^{k,p}(\mathcal{S})$ denote a family of functions satisfying

- Any $f \in \mathcal{C}_L^{k,p}(\mathcal{S})$ is k times continuously differentiable over $\mathcal{S}$.
- Its pth derivative is Lipschitz continuous on $\mathcal{S}$ with constant L, i.e.,

$$\|\nabla^p f(\boldsymbol{x}) - \nabla^p f(\boldsymbol{y})\|_2 \leq L\|\boldsymbol{x} - \boldsymbol{y}\|_2, \text{ for any } \boldsymbol{x}, \boldsymbol{y} \in \mathcal{S}.$$

In this course, the function class $\mathcal{C}_L^{1,1}(\mathcal{S})$ is of particular interest and we call any $f \in \mathcal{C}_L^{1,1}(\mathcal{S})$ *L-smooth* on $\mathcal{S}$.

The Lipschitz continuity controls the rate of changes and the following lemma shows the fact in a special case.

Lemma 2.0.1

A function $f : \mathcal{S} \subset \mathbb{R}^d \to \mathbb{R}$ belongs to $\mathcal{C}_L^{2,1}(\mathcal{S}) \subset \mathcal{C}_L^{1,1}(\mathcal{S})$ iff. (if and only if)

$$\|\nabla^2 f(\boldsymbol{x})\| \leq L, \text{ for any } \boldsymbol{x} \in \mathcal{S}.$$

Proof. ($\Rightarrow$) If $f \in \mathcal{C}_L^{2,1}(\mathcal{S})$, we have

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2 \leq L\|\boldsymbol{x} - \boldsymbol{y}\|_2, \text{ for any } \boldsymbol{x}, \boldsymbol{y} \in \mathcal{S}.$$

Let $\boldsymbol{y} := \boldsymbol{x} + \alpha\boldsymbol{d}$ for $\boldsymbol{d} \in \mathbb{R}^d \backslash \{\boldsymbol{0}\}$ and small $\alpha > 0$.

$$\begin{aligned}\|\nabla f(x + \alpha\boldsymbol{d}) - \nabla f(\boldsymbol{x})\|_2 &= \left\|\int_0^\alpha \nabla^2 f(\boldsymbol{x} + \tau\boldsymbol{d})\boldsymbol{d}\mathrm{d}\tau\right\| \\ &\leq \alpha L\|\boldsymbol{d}\|\end{aligned}$$

For the inequality, dividing both sides by α and $\|\boldsymbol{d}\|$ yields

$$\left\|\frac{\int_0^\alpha \nabla^2 f(\boldsymbol{x} + \tau\boldsymbol{d})\boldsymbol{d}\mathrm{d}\tau}{\alpha\|\boldsymbol{d}\|}\right\| \leq L.$$

Taking $\alpha \to 0^+$

$$\lim_{\alpha \to 0^+} \left\| \frac{\int_0^\alpha \nabla^2 f(\boldsymbol{x} + \tau \boldsymbol{d}) \boldsymbol{d} \mathrm{d}\tau}{\alpha \|\boldsymbol{d}\|} \right\| = \left\| \lim_{\alpha \to 0^+} \frac{\int_0^\alpha \nabla^2 f(\boldsymbol{x} + \tau \boldsymbol{d}) \boldsymbol{d} \mathrm{d}\tau}{\alpha \|\boldsymbol{d}\|} \right\|$$
$$= \left\| \nabla^2 f(\boldsymbol{x}) \frac{\boldsymbol{d}}{\|\boldsymbol{d}\|} \right\| \leq L.$$

By the definition of induced norm of matrices, we immediately have $\|\nabla^2 f(\boldsymbol{x})\| \leq L$.

($\Leftarrow$) It is trivial by using the Newton-Leibniz formula

$$\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x}) = \int_0^1 \nabla^2 f(\boldsymbol{x} + \tau(\boldsymbol{y} - \boldsymbol{x})) \cdot (\boldsymbol{y} - \boldsymbol{x}) \mathrm{d}\tau,$$

for any $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{S}$. □

Smooth convex functions

Definition 2.0.2: Convex sets

Let $\mathcal{S} \subseteq \mathbb{R}^d$. $\mathcal{S}$ is a **convex set** if

$$\lambda \boldsymbol{x} + (1 - \lambda) \boldsymbol{y} \in \mathcal{S},$$

for any $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{S}$ and $\lambda \in [0, 1]$.

Definition 2.0.3: Differentiable convex function

A continuously differentiable function f is called *convex* on a convex set $\mathcal{S}$ if for any $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{S}$ we have

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle.$$

Remark.

- For a differentiable convex function, the graph of the function always lies above (or on) its tangent lines at any point
- The definition of convex functions is equivalent to

$$f(\lambda \boldsymbol{x} + (1 - \lambda) \boldsymbol{y}) \leq \lambda f(\boldsymbol{x}) + (1 - \lambda) f(\boldsymbol{y})$$

  for any $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{S}$ and $\lambda \in [0, 1]$, if f is defined and differentiable on the convex set $\mathcal{S}$.
- If $-f$ is convex, we call f *concave*.

Theorem 2.0.4

A continuously differentiable function f is convex on a convex set $\mathcal{S}$ iff. for any $\boldsymbol{x}, \boldsymbol{y}$ we have

$$\langle \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle \geq 0.$$

Proof. ($\Rightarrow$) It is trivial by the definition of differentiable convex functions.

($\Leftarrow$) By the Newton-Leibniz formula, we have

$$\begin{aligned} f(\boldsymbol{y}) - f(\boldsymbol{x}) &= \int_0^1 \langle \nabla f(\boldsymbol{x} + \tau(\boldsymbol{y} - \boldsymbol{x})), \boldsymbol{y} - \boldsymbol{x} \rangle \mathrm{d}\tau \\ &= \int_0^1 \langle \nabla f(\boldsymbol{x} + \tau(\boldsymbol{y} - \boldsymbol{x})) - \nabla f(\boldsymbol{x}) + \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle \mathrm{d}\tau \\ &= \int_0^1 \langle \nabla f(\boldsymbol{x} + \tau(\boldsymbol{y} - \boldsymbol{x})) - \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle \mathrm{d}\tau + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle \\ &\geq \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle. \end{aligned}$$

□

Theorem 2.0.5

Let $\mathcal{S}$ be an open set. A twice continuously differentiable function f is convex iff. for any $\boldsymbol{x} \in \mathcal{S}$ we have

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

Proof. We skip the proof here. □

Theorem 2.0.6: Sufficient condition for optimality

Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth and convex. If $\nabla f(\boldsymbol{x}^*) = 0$ then $\boldsymbol{x}^*$ is the global minimum of f on $\mathbb{R}^d$.

Proof. We also skip the proof. □

The following theorem shows the optimal condition on a subset of $\mathbb{R}^d$.

Theorem 2.0.7: Optimal condition on a closed and convex set

Let $f : \mathcal{S} \subseteq \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable and convex. Suppose $\mathcal{S}$ is a closed and convex set. Then a point $\boldsymbol{x}^*$ is a minimum point of f on $\mathcal{S}$ iif.

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

for any $\boldsymbol{x} \in \mathcal{S}$.

Proof. ($\Rightarrow$) Suppose there exists some $\boldsymbol{x}^\ddagger \in \mathcal{S}$ such that

$$\langle \nabla f(\boldsymbol{x}^*), \boldsymbol{x}^\ddagger - \boldsymbol{x}^* \rangle < 0.$$

By the definition of directional derivative, we have

$$\lim_{t\to 0^+} \frac{f(\boldsymbol{x}^* + t(\boldsymbol{x}^\ddagger - \boldsymbol{x}^*)) - f(\boldsymbol{x}^*)}{t} = \langle \nabla f(\boldsymbol{x}^*), \boldsymbol{x}^\ddagger - \boldsymbol{x}^* \rangle < 0.$$

Hence, for some small t_0 such that $\boldsymbol{x}^* + t_0(\boldsymbol{x}^\ddagger - \boldsymbol{x}^*) \in \mathcal{S}$,

$$f(\boldsymbol{x}^* + t_0(\boldsymbol{x}^\ddagger - \boldsymbol{x}^*)) < f(\boldsymbol{x}^*)$$

which leads to contradiction with the optimal of $\boldsymbol{x}^*$.

($\Leftarrow$) It is trivial. □

Theorem 2.0.8

Let $f : \mathbb{R}^d \to \mathbb{R}$ be L-smooth and convex. The following inequalities hold for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$.

$$f(\boldsymbol{y}) \le f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2$$

$$\frac{1}{2L}\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2^2 \le f(\boldsymbol{y}) - f(\boldsymbol{x}) - \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle.$$

Proof. To show the first inequality, we have

$$\begin{aligned}
f(\boldsymbol{y}) - f(\boldsymbol{x}) &= \int_0^1 \langle \nabla f(\boldsymbol{x} + \tau(\boldsymbol{y} - \boldsymbol{x})), \boldsymbol{y} - \boldsymbol{x} \rangle \mathrm{d}\tau \\
&= \int_0^1 \langle \nabla f(\boldsymbol{x} + \tau(\boldsymbol{y} - \boldsymbol{x})) - \nabla f(\boldsymbol{x}) + \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle \mathrm{d}\tau \\
&= \int_0^1 \langle \nabla f(\boldsymbol{x} + \tau(\boldsymbol{y} - \boldsymbol{x})) - \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle \mathrm{d}\tau + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle \\
&\le \int_0^1 \|\nabla f(\boldsymbol{x} + \tau(\boldsymbol{y} - \boldsymbol{x})) - \nabla f(\boldsymbol{x})\| \cdot \|\boldsymbol{y} - \boldsymbol{x}\| \mathrm{d}\tau + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle \\
&\le \int_0^1 L\tau \|\boldsymbol{y} - \boldsymbol{x}\|^2 \mathrm{d}\tau + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle \\
&= \frac{L}{2}\|\boldsymbol{y} - \boldsymbol{x}\| + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle,
\end{aligned}$$

which is rearranged as

$$f(\boldsymbol{y}) \le f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2. \qquad \text{(Upper-Bound)}$$

To show the second inequality, we first take $\boldsymbol{y} = \boldsymbol{x} - \frac{1}{L}\nabla f(\boldsymbol{x})$ in (Upper-Bound) and obtain

$$\frac{1}{2L}\|\nabla f(\boldsymbol{x})\|_2^2 \le f(\boldsymbol{x}) - f(\boldsymbol{x} - \frac{1}{L}\nabla f(\boldsymbol{x})) \quad \text{for any } \boldsymbol{x} \in \mathbb{R}^d.$$

Let us fix $\boldsymbol{x}$ and define $\phi(\boldsymbol{y}) = f(\boldsymbol{y}) - \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} \rangle$ for any $\boldsymbol{y} \in \mathbb{R}^d$. Note that $\phi(\boldsymbol{z})$ is L-smooth and convex, and hence

$$\frac{1}{2L}\|\nabla \phi(\boldsymbol{y})\|_2^2 \le \phi(\boldsymbol{y}) - \phi(\boldsymbol{y} - \frac{1}{L}\nabla \phi(\boldsymbol{y})).$$

Since $\nabla\phi(\boldsymbol{x}) = 0$, ϕ attains its minimum at $\boldsymbol{x}$. Therefore

$$\begin{aligned}\frac{1}{2L}\|\nabla\phi(\boldsymbol{y})\|_2^2 &\leq \phi(\boldsymbol{y}) - \phi\left(\boldsymbol{y} - \frac{1}{L}\nabla\phi(\boldsymbol{y})\right). \\ &\leq \phi(\boldsymbol{y}) - \phi(\boldsymbol{x}),\end{aligned}$$

and we get the second inequality since $\nabla\phi(\boldsymbol{y}) = \nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x})$. □

Remark.

> These two inequalities are important for convergence analysis of first-order methods.

We know that the convex functions are not necessarily to be differentiable.

Definition 2.0.9: Not-necessarily differentiable convex functions

A function f is called convex on a convex set $\mathcal{S}$ if for any $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{S}$ we have

$$f(\lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{y}) \leq \lambda f(\boldsymbol{x}) + (1-\lambda)f(\boldsymbol{y})$$

for any $\lambda \in [0, 1]$.

We have mentioned that these two definitions of convex functions are equivalent if f is differentiable on $\mathcal{S}$.

Lemma 2.0.10: Jensen's inequality

Let $f : \mathcal{S} \subseteq \mathbb{R}^d \to \mathbb{R}$ be convex and $\mathcal{S}$ is a convex set. Then for any sequence of $(\boldsymbol{x}_i)_{i=1}^n \subseteq \mathcal{S}$ and any $n \in \mathbb{N}$, we have

$$f(\sum_{i=1}^n \alpha_i\boldsymbol{x}_i) \leq \sum_{i=1}^n \alpha_i f(\boldsymbol{x}_i)$$

if $\sum_{i=1}^n \alpha_i = 1$ and $\alpha_i \geq 0 (i = 1, 2, \ldots, n)$.

We call $\boldsymbol{z}$ is a **convex combination** of $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\}$ if there exists $\boldsymbol{\lambda} = (\lambda_i) \in \mathbb{R}^n$ such that

$$\boldsymbol{z} = \sum_{i=1}^n \lambda_i\boldsymbol{x}_i$$

satisfying $\sum_{i=1}^n \lambda_i = 1$ and $\lambda_i \geq 0$ for all i.

Definition 2.0.11: Convex hulls

Let $\mathcal{S} \subseteq \mathbb{R}^d$. We call a set **convex hull** of $\mathcal{S}$, denoted by $\text{conv}(\mathcal{S})$ if any element of this set, can be expressed as a convex combination of points from $\mathcal{S}$. Mathematically, for any $\boldsymbol{z} \in \text{conv}(\mathcal{S})$, there exists a sequence $\{\boldsymbol{x}_i\}_{i=1}^n \subseteq \mathcal{S}$ for $n \in \mathbb{N}$ such that

$$\boldsymbol{z} = \sum_{i=1}^n \lambda_i \boldsymbol{x}_i$$

satisfying $\sum_{i=1}^n \lambda_i = 1$ and $\lambda_i \geq 0$ for all i.

By Jensen's inequality, we immediately reach the following lemma.

Lemma 2.0.12

Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and $\mathcal{S}$ be a subset of $\mathbb{R}^d$. Then we have

$$\max_{\boldsymbol{x} \in \text{conv}(\mathcal{S})} f(\boldsymbol{x}) = \max_{\mathcal{S}} f(\boldsymbol{x}).$$

Proof. Since $\mathcal{S} \subseteq \text{conv}(S)$, we have

$$\max_{\boldsymbol{x} \in \text{conv}(\mathcal{S})} f(\boldsymbol{x}) \geq \max_{\mathcal{S}} f(\boldsymbol{x}).$$

For any $\boldsymbol{z} \in \text{conv}(\mathcal{S})$, there exists a $n \in \mathbb{N}$ and a sequence $(\boldsymbol{x}_i)_{i=1}^n \subseteq \mathcal{S}$ such that

$$\boldsymbol{z} = \sum_{i=1}^n \alpha_i \boldsymbol{x}_i$$

for some $\alpha_i \geq 0 (i = 1, 2, \ldots, n)$ and $\sum_{i=1}^n \alpha_i = 1$.

Therefore, by Jensen's inequality

$$f(\boldsymbol{z}) = f(\sum_{i=1}^n \alpha_i \boldsymbol{x}_i) \leq \sum_{i=1}^n \alpha_i f(\boldsymbol{x}_i) \leq \max_{\mathcal{S}} f(\boldsymbol{x}) \sum_{i=1}^n \alpha_i \leq \max_{\mathcal{S}} f(\boldsymbol{x}).$$

For the arbitrariness of $\boldsymbol{z} \in \text{conv}(\mathcal{S})$, we have

$$\max_{\boldsymbol{x} \in \text{conv}(\mathcal{S})} f(\boldsymbol{x}) \leq \max_{\mathcal{S}} f(\boldsymbol{x}).$$

□

We present in the following that the non-necessarily differentiable (NND) convex functions are locally bounded and locally Lipschitz continuous.

Theorem 2.0.13

Let $f : \mathcal{S} \subseteq \mathbb{R}^d \to \mathbb{R}$ be convex and $\boldsymbol{x}_0 \in \text{int}(\mathcal{S})$. Then f is locally bounded, i.e., $\exists \epsilon > 0$ and $M(\boldsymbol{x}_0, \epsilon) > 0$ such that

$$f(\boldsymbol{x}) \leq M(\boldsymbol{x}_0, \epsilon) \text{ for any } \boldsymbol{x} \in \mathcal{B}_2(\boldsymbol{x}_0, \epsilon) := \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x} - \boldsymbol{x}_0\|_2 \leq \epsilon\}$$

Proof. Since $\boldsymbol{x}_0 \in \text{int}(\mathcal{S})$, there exists $\epsilon > 0$ such that f is defined on the hypercube $\mathcal{B}_\infty(\boldsymbol{x}_0, \epsilon) := \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x} - \boldsymbol{x}_0\|_\infty \leq \epsilon\}$, which is the convex hull of the set $\{\boldsymbol{x}_0 \pm \epsilon \boldsymbol{e}_i\}_{i=1}^d$. The symbol $\boldsymbol{e}_i$ denotes the unit vector along coordinate i.

Therefore

$$\max_{\boldsymbol{x} \in \mathcal{B}_2(\boldsymbol{x}_0, \epsilon)} f(\boldsymbol{x}) \leq \max_{\boldsymbol{x} \in \mathcal{B}_\infty(\boldsymbol{x}_0, \epsilon)} = \max_{\boldsymbol{x} \in \{\boldsymbol{x}_0 \pm \epsilon \boldsymbol{e}_i\}_{i=1}^d} f(\boldsymbol{x}) := M(\boldsymbol{x}_0, \epsilon),$$

where the first inequality attributes to the fact $\mathcal{B}_2(\boldsymbol{x}_0, \epsilon) \subset \mathcal{B}_\infty(\boldsymbol{x}_0, \epsilon)$ and the second equality comes from Lemma 2.0.11. □

Theorem 2.0.14

Let $f : \mathcal{S} \subseteq \mathbb{R}^d \to \mathbb{R}$ be convex and $\boldsymbol{x}_0 \in \text{int}(\mathcal{S})$. Then f is locally Lipschitz continuous, i.e., $\exists \epsilon > 0$ and $\bar{M}(\boldsymbol{x}_0, \epsilon) > 0$ such that

$$|f(\boldsymbol{y}) - f(\boldsymbol{x_0})| \leq \bar{M}(\boldsymbol{x}_0, \epsilon) \|\boldsymbol{y} - \boldsymbol{x}_0\|_2$$

for any $\boldsymbol{y} \in \mathcal{B}_2(\boldsymbol{x}_0, \epsilon) := \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x} - \boldsymbol{x}_0\|_2 \leq \epsilon\}$

Proof. Since $\boldsymbol{x}_0 \in \text{int}(\mathcal{S})$, there exists $\epsilon > 0$ such that f is defined on $\mathcal{B}_2(\boldsymbol{x}_0, \epsilon) := \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x} - \boldsymbol{x}_0\|_2 \leq \epsilon\}$

For $\boldsymbol{y} = \boldsymbol{x}_0$, the result is trivial.

Suppose $\boldsymbol{y} \neq \boldsymbol{x}_0$ and $\boldsymbol{y} \in \mathcal{B}_2(\boldsymbol{x}_0, \epsilon)$. We extend the line segment connecting $\boldsymbol{x}_0$ and $\boldsymbol{y}$ so that it intersect the boundary of $\mathcal{B}_2(\boldsymbol{x}_0, \epsilon)$. The intersection points are denoted by $\boldsymbol{v}$ and $\boldsymbol{u}$, respectively.

Define $\alpha = \frac{\|\boldsymbol{x}_0 - \boldsymbol{y}\|_2}{\epsilon}$. We have

$$\boldsymbol{y} = (1 - \alpha)\boldsymbol{x}_0 + \alpha \boldsymbol{v},$$
$$\boldsymbol{x}_0 = \frac{\alpha}{1 + \alpha}\boldsymbol{u} + \frac{1}{1 + \alpha}\boldsymbol{y}.$$

By the convexity of f,

$$f(\boldsymbol{y}) \leq (1 - \alpha) f(\boldsymbol{x}_0) + \alpha f(\boldsymbol{v})$$
$$f(\boldsymbol{x}_0) \leq \frac{\alpha}{1 + \alpha} f(\boldsymbol{u}) + \frac{1}{1 + \alpha} f(\boldsymbol{y}).$$

We rearrange as

$$f(\boldsymbol{x}_0) - f(\boldsymbol{y}) \leq \alpha(f(\boldsymbol{v}) - f(\boldsymbol{x}_0)) \leq \frac{\|\boldsymbol{x}_0 - \boldsymbol{y}\|_2}{\epsilon}(M(\boldsymbol{x}_0, \epsilon) - f(\boldsymbol{x}_0))$$
$$f(\boldsymbol{y}) - f(\boldsymbol{x}_0) \leq \alpha(f(\boldsymbol{v}) - f(\boldsymbol{x}_0)) \leq \frac{\|\boldsymbol{x}_0 - \boldsymbol{y}\|_2}{\epsilon}(M(\boldsymbol{x}_0, \epsilon) - f(\boldsymbol{x}_0))$$

by the fact that f is locally bounded by $M(\boldsymbol{x}_0, \epsilon)$.

Let $\bar{M}(\boldsymbol{x}_0, \epsilon) = (M(\boldsymbol{x}_0, \epsilon) - f(\boldsymbol{x}_0))/\epsilon$, we complete the proof. □

Smooth and Strongly Convex Functions

Definition 2.0.15

A continuously differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is called *μ-strongly convex* on $\mathbb{R}^d$ if there exists a constant $\mu > 0$ such that for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ we have

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

For a μ-strongly convex function, we have

$$\langle \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle \geq \mu \|\boldsymbol{x} - \boldsymbol{y}\|_2^2.$$

Theorem 2.0.16

If $f : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex, then for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ we have

$$\begin{aligned} f(\boldsymbol{y}) \leq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{1}{2\mu} \|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2^2, \\ \langle \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle \leq \frac{1}{\mu} \|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2^2, \\ \mu \|\boldsymbol{x} - \boldsymbol{y}\|_2 \leq \|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2^2. \end{aligned}$$

Proof. We only show the first inequality here.

For any $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$, we have

$$f(\boldsymbol{u}) \geq f(\boldsymbol{v}) + \langle \nabla f(\boldsymbol{v}), \boldsymbol{u} - \boldsymbol{v} \rangle + \frac{\mu}{2} \|\boldsymbol{u} - \boldsymbol{v}\|_2^2,$$

which implies

$$\begin{aligned} \min_{\boldsymbol{u} \in \mathbb{R}^d} f(\boldsymbol{u}) &\geq \min_{\boldsymbol{u} \in \mathbb{R}^d} \left\{ f(\boldsymbol{v}) + \langle \nabla f(\boldsymbol{v}), \boldsymbol{u} - \boldsymbol{v} \rangle + \frac{\mu}{2} \|\boldsymbol{u} - \boldsymbol{v}\|_2^2 \right\} \\ &= f(\boldsymbol{v}) - \frac{1}{2\mu} \|\nabla f(\boldsymbol{v})\|_2^2. \end{aligned}$$

Let fix some $\boldsymbol{x} \in \mathbb{R}^d$ and define $\phi(\boldsymbol{z}) = f(\boldsymbol{z}) - \langle \nabla f(\boldsymbol{x}), \boldsymbol{z} \rangle$. Function $\phi(\boldsymbol{z})$ is also μ-strongly convex, and therefore

$$\min_{\boldsymbol{z} \in \mathbb{R}^d} \phi(\boldsymbol{z}) \geq \phi(\boldsymbol{y}) - \frac{1}{2\mu} \|\nabla \phi(\boldsymbol{y})\|_2^2 \quad \text{for any } \boldsymbol{y} \in \mathbb{R}^d. \qquad \text{(Lower-Bounded)}$$

Note that $\nabla \phi(\boldsymbol{x}) = \mathbf{0}$, we have $\min_{\boldsymbol{z} \in \mathbb{R}^d} \phi(\boldsymbol{z}) = \phi(\boldsymbol{x})$.

Since $\nabla \phi(\boldsymbol{y}) = \nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x})$, substituting all the ingredients back to (Lower-Bounded) and arranging yield

$$f(\boldsymbol{y}) \le f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{1}{2\mu} \|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2^2$$

□

Theorem 2.0.17

If $f : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex and L-smooth, then for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ we have

$$\langle \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle \ge \frac{\mu L}{\mu + L} \|\boldsymbol{x} - \boldsymbol{y}\|_2^2 + \frac{1}{\mu + L} \|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2^2$$

Theorem 2.0.16 will be useful for the converge analysis of gradient descent. We skip the proof here.

Theorem 2.0.18

Let $f : \mathcal{S} \subseteq \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable and μ-strongly convex on a closed and convex $\mathcal{S}$. Then the minimum point of f on $\mathcal{S}$ exists and is unique.

Proof. Let $\boldsymbol{x}_0 \in \mathcal{S}$ and define $\bar{\mathcal{S}} = \{\boldsymbol{x} \in \mathcal{S} : f(\boldsymbol{x}) \le f(\boldsymbol{x}_0)\}$. Hence minimizing f on $\mathcal{S}$ is equivalent to minimizing f on $\bar{\mathcal{S}}$.

We are going to show that $\bar{\mathcal{S}}$ is bounded.

For any $\boldsymbol{x} \in \bar{\mathcal{S}}$, we have

$$f(\boldsymbol{x}_0) \ge f(\boldsymbol{x}) \ge f(\boldsymbol{x}_0) + \langle \nabla f(\boldsymbol{x}_0), \boldsymbol{x} - \boldsymbol{x}_0 \rangle + \frac{\mu}{2} \|\boldsymbol{x} - \boldsymbol{x}_0\|_2^2.$$

which is rearranged as

$$\frac{\mu}{2} \|\boldsymbol{x} - \boldsymbol{x}_0\|_2^2 \le \langle \nabla f(\boldsymbol{x}_0), \boldsymbol{x} - \boldsymbol{x}_0 \rangle \le \|\nabla f(\boldsymbol{x}_0)\|_2 \|\boldsymbol{x} - \boldsymbol{x}_0\|_2$$

We use the Cauchy-Schwartz inequality in the last inequality and obtain

$$\|\boldsymbol{x} - \boldsymbol{x}_0\|_2 \le \frac{2}{\mu} \|\nabla f(\boldsymbol{x}_0)\|_2.$$

Hence $\bar{\mathcal{S}}$ is bounded and is also closed by the fact that f is continuous on $\mathcal{S}$.

Therefore, the minimum point of f on $\bar{\mathcal{S}}$ exists and so does it on $\mathcal{S}$.

Let $\boldsymbol{x}^*$ and $\boldsymbol{x}^\ddagger$ be two minimum points of f on $\mathcal{S}$.

We have

$$f(\boldsymbol{x}^\ddagger) \ge f(\boldsymbol{x}^*) + \langle \nabla f(\boldsymbol{x}^*), \boldsymbol{x}^\ddagger - \boldsymbol{x}^* \rangle + \frac{\mu}{2} \|\boldsymbol{x}^\ddagger - \boldsymbol{x}^*\|_2^2.$$

□

By the fact that $f(\boldsymbol{x}^\ddagger) = f(\boldsymbol{x}^*)$ and $\langle \nabla f(\boldsymbol{x}^*), \boldsymbol{x}^\ddagger - \boldsymbol{x}^* \rangle \ge 0$, we have

$$\frac{\mu}{2} \|\boldsymbol{x}^\ddagger - \boldsymbol{x}^*\|_2^2 \le 0,$$

which implies $\boldsymbol{x}^{\ddagger} = \boldsymbol{x}^{*}$

Definition 2.0.19: Not-necessarily differentiable μ-strongly convex functions

A function f is called μ-strongly convex on a convex set $\mathcal{S}$ if for any $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{S}$ we have

$$f(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}) \leq \lambda f(\boldsymbol{x}) + (1-\lambda) f(\boldsymbol{y}) - \frac{\mu}{2}\lambda(1-\lambda)\|\boldsymbol{x} - \boldsymbol{y}\|_2^2$$

for any $\lambda \in [0,1]$.

If f is differentiable, the definition is equivalent to the differentiable version.