
CHAPTER 0. PRELIMINARY

Yutong He

1 Norm

Definition 1.1 (Norm). A real-valued function $\|\cdot\|$ defined on linear space \mathbb{E} is called norm, if it satisfy:

1. (being positive definite) $\forall x \in \mathbb{E}$, we have $\|x\| \geq 0$, and $\|x\| = 0$ if and only if $x = 0$.
2. (being absolutely homogeneous) $\forall x \in \mathbb{E}, \alpha \in \mathbb{R}$, we have $\|\alpha x\| = |\alpha| \cdot \|x\|$.
3. (the triangle inequality) $\forall x, y \in \mathbb{E}$, we have $\|x + y\| \leq \|x\| + \|y\|$.

1.1 Vector Norm

We consider vector norm defined on vector space $\mathbb{E} = \mathbb{R}^n$.

Definition 1.2 (ℓ_p -norm). The ℓ_p -norm ($p \geq 1$) is defined as:

$$\|x\|_p = (|x_1|^p + |x_2|^p + \cdots + |x_n|^p)^{1/p},$$

for all $x = (x_1, x_2, \cdots, x_n)^\top \in \mathbb{R}^n$.

Specifically, we have the following commonly used definition of ℓ_1 -norm and ℓ_2 -norm for vector $x = (x_1, x_2, \cdots, x_n)^\top \in \mathbb{R}^n$. When $p = 1$, the ℓ_1 -norm is given by:

$$\|x\|_1 = |x_1| + |x_2| + \cdots + |x_n|.$$

When $p = 2$, the ℓ_2 -norm is the same as the Euclidean norm:

$$\|x\|_2 = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}.$$

We have the following useful Cauchy's inequality for ℓ_2 norm:

Proposition 1.3 (Cauchy's inequality). $\forall x, y \in \mathbb{R}^n$, we have

$$|\langle x, y \rangle| \leq \|x\|_2 \cdot \|y\|_2,$$

where $\langle x, y \rangle = x^\top y$ denotes the inner product of x and y , and the equality holds if and only if x and y are linearly correlated (i.e., $\exists \alpha, \beta \in \mathbb{R}$, s.t. $\alpha^2 + \beta^2 > 0$ and $\alpha x + \beta y = 0$).

Definition 1.4 (ℓ_∞ -norm). The ℓ_∞ -norm is defined as:

$$\|x\|_\infty = \max\{|x_1|, |x_2|, \dots, |x_n|\},$$

for all $x = (x_1, x_2, \dots, x_n)^\top \in \mathbb{R}^n$.

Definition 1.5 (Norm induced by a positive definite matrix A). Given a positive definite matrix $A \in \mathbb{R}^{n \times n}$, we define:

$$\|x\|_A = \sqrt{x^\top A x},$$

for all $x = (x_1, x_2, \dots, x_n)^\top \in \mathbb{R}^n$.

1.2 Matrix norm

We consider matrix norms defined on matrix space $\mathbb{E} = \mathbb{R}^{m \times n}$.

Definition 1.6 (Frobenius norm / F-norm). The Frobenius norm (or F-norm) of matrix $A \in \mathbb{R}^{m \times n}$ is defined as:

$$\|A\|_F = \sqrt{\text{tr}(A^\top A)} = \sqrt{\sum_{i=1}^m \sum_{j=1}^n A_{i,j}^2},$$

where $A_{i,j}$ represents the (i, j) -th element of matrix A .

The Frobenius norm has similar properties as the ℓ_2 -norm. For example, we have the following Cauchy's inequality for the Frobenius norm.

Proposition 1.7 (Cauchy's inequality). $\forall A, B \in \mathbb{R}^{m \times n}$, we have

$$|\langle A, B \rangle| \leq \|A\|_F \cdot \|B\|_F,$$

where $\langle A, B \rangle = \text{tr}(A^\top B) = \text{tr}(AB^\top)$ denotes the Frobenius inner product of A and B , and the equality holds if and only if A and B are linearly correlated.

Definition 1.8 (Spectral norm). The spectral norm of matrix $A \in \mathbb{R}^{m \times n}$ is defined as:

$$\|A\|_2 = \max_{x \in \mathbb{R}^n, \|x\|_2=1} \|Ax\|_2.$$

For a given matrix $A \in \mathbb{R}^{m \times n}$, it can be shown that $\|A\|_2$ equals the largest singular value of A , and $\|A\|_2^2$ equals the largest eigenvalue of $A^\top A$ (and AA^\top).

By definition, we also have the following inequality:

$$\|Ax\|_2 \leq \|A\|_2 \|x\|_2,$$

for all $A \in \mathbb{R}^{m \times n}$ and $x \in \mathbb{R}^n$.

Definition 1.9 (Nuclear norm). The nuclear norm of matrix $A \in \mathbb{R}^{m \times n}$ is defined as:

$$\|A\|_* = \sigma_1 + \sigma_2 + \cdots + \sigma_r,$$

where $r = \text{rank}(A)$ and $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$ are non-zero singular values of A .

We have the following inequalities for the matrix norms:

$$\|AB\|_F \leq \|A\|_2 \|B\|_F,$$

$$|\langle A, B \rangle| \leq \|A\|_2 \|B\|_*.$$

2 Gradient

2.1 Gradient and Hessian matrix

Definition 2.1 (Gradient). Suppose $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is continuously differentiable, its gradient $\nabla f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is defined at point $x = (x_1, x_2, \dots, x_n)^\top$ as:

$$\nabla f(x) = \left(\frac{\partial f(x)}{\partial x_1}, \frac{\partial f(x)}{\partial x_2}, \dots, \frac{\partial f(x)}{\partial x_n} \right)^\top.$$

Definition 2.2 (Hessian matrix). Suppose $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is twice continuously differentiable, its Hessian matrix $\nabla^2 f : \mathbb{R}^n \rightarrow \mathbb{R}^{n \times n}$ is defined at point $x = (x_1, x_2, \dots, x_n)^\top$ as:

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

As an example, we consider gradients and Hessian matrices of two typical functions.

1. (The least squares problem) Let $f(x) := \frac{1}{2} \|Ax - b\|_2^2$, we have $\nabla f(x) = A^\top (Ax - b)$ and $\nabla^2 f(x) = A^\top A$.
2. (The logistic regression problem) Let $f(x) := \frac{1}{M} \sum_{i=1}^M \ln(1 + \exp(-b_i a_i^\top x))$, we have

$$\nabla f(x) = \frac{1}{M} \sum_{i=1}^M \frac{-b_i \exp(-b_i a_i^\top x)}{1 + \exp(-b_i a_i^\top x)} a_i,$$

and

$$[\nabla^2 f(x)]_{j,k} = \frac{1}{M} \sum_{i=1}^M \frac{b_i^2 \exp(-b_i a_i^\top x) a_{i,j} a_{i,k}}{[1 + \exp(-b_i a_i^\top x)]^2}.$$

Like univariate functions, we have Taylor expansions in the multivariate case:

Proposition 2.3 (Taylor expansion). Suppose $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is continuously differentiable, $\forall x, y \in \mathbb{R}^n$, we have

$$f(x + y) = f(x) + \langle \nabla f(x + \theta_1 y), y \rangle,$$

for some $\theta_1 \in (0, 1)$. If f is further twice continuously differentiable, it holds that

$$\nabla f(x + y) = \nabla f(x) + \int_0^1 \nabla^2 f(x + ty) y dt,$$

and

$$f(x + y) = f(x) + \langle \nabla f(x), y \rangle + \frac{1}{2} y^\top \nabla^2 f(x + \theta_2 y) y,$$

for some $\theta_2 \in (0, 1)$.

2.2 L -smoothness

Definition 2.4 (L -smoothness). Function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is said to be L -smooth if $\forall x, y \in \mathbb{R}^n$,

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L \|x - y\|_2,$$

where $L > 0$ is the Lipschitz constant of ∇f .

Theorem 2.5 (L -smooth property). Suppose $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is L -smooth, it holds for $\forall x, y \in \mathbb{R}^n$ that

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

Proof. Let $g(t) := f(x + t(y - x))$, we have

$$g'(t) = \langle \nabla f(x + t(y - x)), y - x \rangle,$$

thus

$$\begin{aligned} & f(y) - f(x) - \langle \nabla f(x), y - x \rangle \\ &= g(1) - g(0) - g'(0) \\ &= \int_0^1 (g'(t) - g'(0)) dt \\ &= \int_0^1 \langle \nabla f(x + t(y - x)) - \nabla f(x), y - x \rangle dt \\ &\leq \int_0^1 \|\nabla f(x + t(y - x)) - \nabla f(x)\|_2 \cdot \|y - x\|_2 dt \\ &\leq \int_0^1 Lt \|y - x\|_2^2 dt \\ &= \frac{L}{2} \|y - x\|_2^2. \end{aligned}$$

□

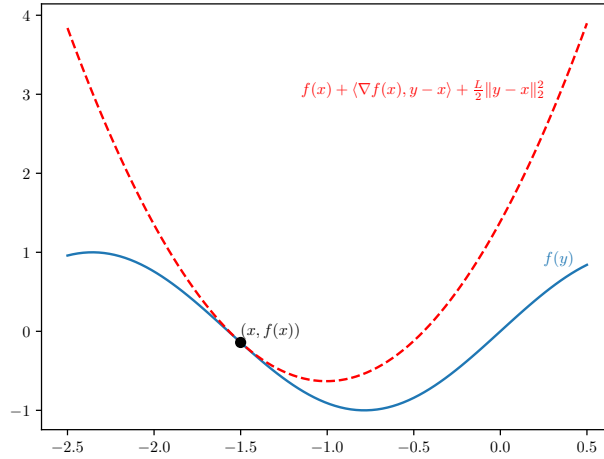


Figure 1: L -smooth property

3 Convexity

3.1 Convex set

Definition 3.1 (Convex set). A set $\mathcal{X} \subseteq \mathbb{R}^n$ is called convex, if for $\forall x, y \in \mathcal{X}$, it holds that

$$\theta x + (1 - \theta)y \in \mathcal{X}, \quad \forall \theta \in [0, 1].$$

Theorem 3.2 (Projection onto closed convex sets). Suppose $\mathcal{X} \subseteq \mathbb{R}^n$ is a closed convex set, then for $\forall y \in \mathbb{R}^n$, there exists a unique $x^* \in \mathcal{X}$ such that $\|x^* - y\|_2 \leq \|x - y\|_2$ for any $x \in \mathcal{X}$. The point x^* is called the projection of y onto \mathcal{X} , denoted by $\mathcal{P}_{\mathcal{X}}(y)$.

Proof. We first prove the existence of point x^* . For a given $y \in \mathbb{R}^n$, let $d := \inf_{x \in \mathcal{X}} \|x - y\|_2$, there exists sequence $\{x_k\}_{k=1}^{\infty} \subseteq \mathcal{X}$ such that $\|x_k - y\|_2^2 \leq d^2 + 1/k$. For any integers $0 < m < n$, by the convexity of \mathcal{X} we have $(x_m + x_n)/2 \in \mathcal{X}$. By the definition of d , we have $\|(x_m + x_n)/2 - y\|_2 \geq d$. Consequently,

$$\begin{aligned} \langle x_m - y, x_n - y \rangle &= 2\|(x_m + x_n)/2 - y\|_2^2 - \frac{\|x_m - y\|_2^2 + \|x_n - y\|_2^2}{2} \geq d^2 - \frac{1}{m}, \\ \Rightarrow \|x_m - x_n\|_2^2 &= \|x_m - y\|_2^2 + \|x_n - y\|_2^2 - 2\langle x_m - y, x_n - y \rangle \leq \frac{4}{m}. \end{aligned}$$

Thus $\{x_k\}_{k=1}^{\infty}$ is a Cauchy sequence, which implies the existence of point $x^* \in \mathbb{R}^n$ such that $x^* = \lim_{k \rightarrow \infty} x_k$. By closedness of \mathcal{X} we know $x^* \in \mathcal{X}$, and by the continuity of ℓ_2 -norm we know $\|x^* - y\|_2 = d$. Next we show the uniqueness of x^* . Otherwise $\exists x_1^*, x_2^* \in \mathcal{X} (x_1^* \neq x_2^*)$ such that $\|x_1^* - y\|_2 = \|x_2^* - y\|_2 = d$, we thus have

$$\|(x_1^* + x_2^*)/2 - y\|_2^2 = \frac{1}{2}\|x_1^* - y\|_2^2 + \frac{1}{2}\|x_2^* - y\|_2^2 - \frac{1}{4}\|x_1^* - x_2^*\|_2^2 < d,$$

a contradiction with the definition of d . □

3.2 Convex function

Definition 3.3 (Convex function). Function $f : \mathcal{X} \rightarrow \mathbb{R}$ is said to be convex if $\mathcal{X} \subseteq \mathbb{R}^n$ is a convex set and $\forall x, y \in \mathcal{X}$, it holds that

$$f(\theta x + (1 - \theta)y) \leq \theta f(x) + (1 - \theta)f(y), \quad \forall \theta \in [0, 1].$$

Here we list some common convex functions:

1. vector norm: $f(x) = \|x\|$,
2. quadratic function: $f(x) = \frac{1}{2}x^\top A x$ where A is positive semi-definite,
3. linear function: $f(x) = \langle a, x \rangle$ for some $a \in \mathbb{R}^n$,
4. combination of convex functions: $f(x) = \alpha_1 f_1(x) + \alpha_2 f_2(x) + \dots + \alpha_m f_m(x)$ where f_1, f_2, \dots, f_m are convex and $\alpha_1, \alpha_2, \dots, \alpha_m$ are non-negative.

Theorem 3.4 (Convex property). Suppose $f : \mathcal{X} \rightarrow \mathbb{R}$ is differentiable, then f is convex if and only if

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle, \quad \forall x, y \in \mathcal{X}. \quad (1)$$

Similarly, f is convex if and only if

$$\langle \nabla f(y) - \nabla f(x), y - x \rangle \geq 0, \quad \forall x, y \in \mathcal{X}. \quad (2)$$

Proof. Step 1: we prove that (1) holds if f is convex. By definition we have for any $\theta \in [0, 1]$ that $f(\theta y + (1 - \theta)x) \leq \theta f(y) + (1 - \theta)f(x)$. Thus,

$$\begin{aligned} f(y) - f(x) &= \lim_{\theta \rightarrow 0+0} \frac{\theta f(y) + (1 - \theta)f(x) - f(x)}{\theta} \\ &\geq \lim_{\theta \rightarrow 0+0} \frac{f(\theta y + (1 - \theta)x) - f(x)}{\theta} \\ &= \lim_{\theta \rightarrow 0+0} \frac{f(x + \theta(y - x)) - f(x)}{\theta} \\ &= \langle \nabla f(x), y - x \rangle. \end{aligned}$$

Step 2: we prove that f is convex if (1) holds. By (1) we have for $\forall x, y \in \mathcal{X}$ and $\theta \in [0, 1]$ that

$$\begin{aligned} &\theta f(x) + (1 - \theta)f(y) - f(\theta x + (1 - \theta)y) \\ &= \theta(f(x) - f(\theta x + (1 - \theta)y)) + (1 - \theta)(f(y) - f(\theta x + (1 - \theta)y)) \\ &\geq \theta \langle \nabla f(\theta x + (1 - \theta)y), (1 - \theta)(x - y) \rangle + (1 - \theta) \langle \nabla f(\theta x + (1 - \theta)y), \theta(y - x) \rangle \\ &= 0. \end{aligned}$$

Step 3: we prove that (1) implies (2). By (1) we have

$$\begin{aligned} f(y) + f(x) &\geq (f(x) + \langle \nabla f(x), y - x \rangle) + (f(y) + \langle \nabla f(y), x - y \rangle) \\ &= f(x) + f(y) - \langle \nabla f(y) - \nabla f(x), y - x \rangle, \end{aligned}$$

which is equivalent to (2).

Step 4: we prove that (2) implies (1). This follows from

$$\begin{aligned} f(y) - f(x) &= \int_0^1 \langle \nabla f(x + t(y - x)), y - x \rangle dt \\ &= \int_0^1 \left(\langle \nabla f(x), y - x \rangle + \frac{1}{t} \langle \nabla f(x + t(y - x)) - \nabla f(x), t(y - x) \rangle \right) dt \\ &\geq \int_0^1 \langle \nabla f(x), y - x \rangle dt \\ &= \langle \nabla f(x), y - x \rangle. \end{aligned}$$

□

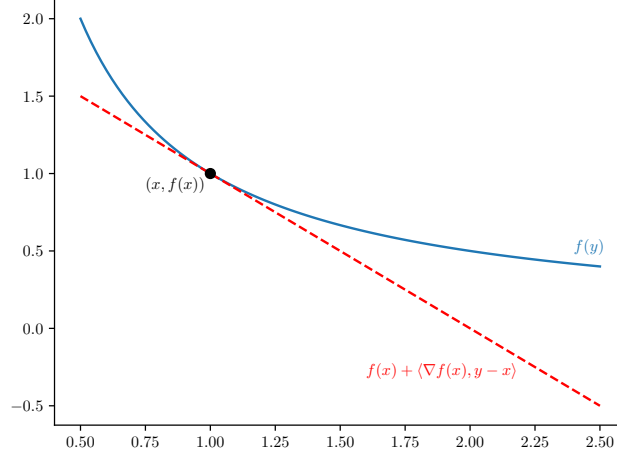


Figure 2: convex property

The following Jensen's inequality is an important property of convex functions.

Proposition 3.5 (Jensen's inequality). Let $f : \mathcal{X} \rightarrow \mathbb{R}$ be convex, then for any $x_1, x_2, \dots, x_m \in \mathcal{X}$ and non-negative $\theta_1, \theta_2, \dots, \theta_m$ satisfying $\theta_1 + \theta_2 + \dots + \theta_m = 1$, it holds that

$$f(\theta_1 x_1 + \theta_2 x_2 + \dots + \theta_m x_m) \leq \theta_1 f(x_1) + \theta_2 f(x_2) + \dots + \theta_m f(x_m).$$

Definition 3.6 (μ -strongly convex function). Function $f : \mathcal{X} \rightarrow \mathbb{R}$ is said to be μ -strongly convex if

$$g(x) := f(x) - \frac{\mu}{2} \|x\|_2^2$$

is a convex function.

It can be proved that $f : \mathcal{X} \rightarrow \mathbb{R}$ is μ -strongly convex if and only if for $\forall x, y \in \mathcal{X}$ and $\theta \in [0, 1]$, it holds that

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

Theorem 3.7 (μ -strongly convex property). Function $f : \mathcal{X} \rightarrow \mathbb{R}$ is μ -strongly convex if and only if

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2} \|y - x\|_2^2, \quad \forall x, y \in \mathcal{X}, \quad (3)$$

if and only if

$$\langle \nabla f(y) - \nabla f(x), y - x \rangle \geq \mu \|y - x\|_2^2, \quad \forall x, y \in \mathcal{X}. \quad (4)$$

Proof. Let $g(x) := f(x) - \frac{\mu}{2}\|x\|_2^2$, then f is μ -strongly convex if and only if g is convex. Note that (3) is equivalent to

$$\begin{aligned} g(y) + \frac{\mu}{2}\|y\|_2^2 &\geq g(x) + \frac{\mu}{2}\|x\|_2^2 + \langle \nabla g(x) + \mu x, y - x \rangle + \frac{\mu}{2}\|y - x\|_2^2, \\ \Leftrightarrow g(y) &\geq g(x) + \langle \nabla g(x), y - x \rangle, \end{aligned}$$

and that (4) is equivalent to

$$\begin{aligned} \langle \nabla g(y) + \mu y - \nabla g(x) - \mu x, y - x \rangle &\geq \mu\|y - x\|_2^2 \\ \Leftrightarrow \langle \nabla g(y) - \nabla g(x), y - x \rangle &\geq 0, \end{aligned}$$

it suffices to apply Theorem 3.4. □

Theorem 3.8 (μ -strongly convex property). Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be μ -strongly convex, then the following inequalities hold:

$$\begin{aligned} f(y) &\leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2\mu}\|\nabla f(y) - \nabla f(x)\|_2^2, \quad \forall x, y \in \mathbb{R}^n, \\ \langle \nabla f(y) - \nabla f(x), y - x \rangle &\leq \frac{1}{\mu}\|\nabla f(y) - \nabla f(x)\|_2^2, \quad \forall x, y \in \mathbb{R}^n. \end{aligned}$$

Proof. Fix $x \in \mathbb{R}^n$, and let $\phi(y) := f(y) - f(x) - \langle \nabla f(x), y - x \rangle$, we know ϕ is μ -strongly convex with minimum $\phi(x) = 0$. Thus,

$$\begin{aligned} 0 = \min_v \phi(v) &\geq \min_v \left(\phi(y) + \langle \nabla \phi(y), v - y \rangle + \frac{\mu}{2}\|v - y\|_2^2 \right) \\ &= \phi(y) - \frac{1}{\mu}\|\nabla \phi(y)\|_2^2 + \frac{\mu}{2}\left\| \frac{1}{\mu}\nabla \phi(y) \right\|_2^2 = \phi(y) - \frac{1}{2\mu}\|\nabla \phi(y)\|_2^2 \\ &= f(y) - f(x) - \langle \nabla f(x), y - x \rangle - \frac{1}{2\mu}\|\nabla f(y) - \nabla f(x)\|_2^2, \end{aligned}$$

which is equivalent to

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

Summing the two copies of the above inequality with x and y interchanged, we obtain

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

□

L -smooth convex functions are widely considered in optimization, which have the fundamental properties below.

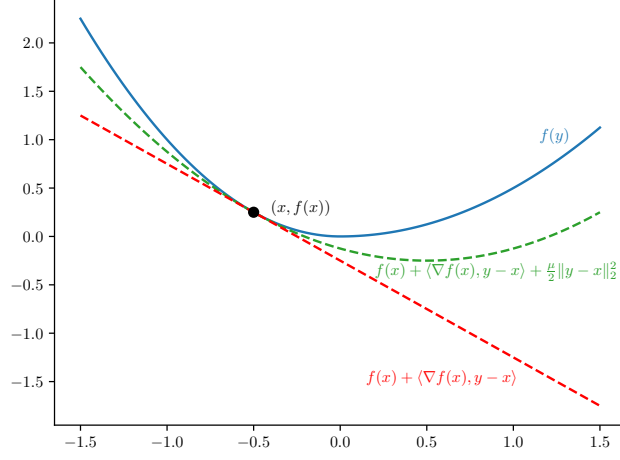


Figure 3: μ -strongly convex property

Theorem 3.9 (L -smooth convex property). Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be differentiable, then f is convex and L -smooth if and only if any one of the following conditions holds:

$$0 \leq f(y) - f(x) - \langle \nabla f(x), y - x \rangle \leq \frac{L}{2} \|y - x\|_2^2, \quad \forall x, y \in \mathbb{R}^n, \quad (5)$$

$$f(y) - f(x) - \langle \nabla f(x), y - x \rangle \geq \frac{1}{2L} \|\nabla f(y) - \nabla f(x)\|_2^2, \quad \forall x, y \in \mathbb{R}^n, \quad (6)$$

$$\langle \nabla f(y) - \nabla f(x), y - x \rangle \geq \frac{1}{L} \|\nabla f(y) - \nabla f(x)\|_2^2, \quad \forall x, y \in \mathbb{R}^n, \quad (7)$$

$$0 \leq \langle \nabla f(y) - \nabla f(x), y - x \rangle \leq L \|y - x\|_2^2, \quad \forall x, y \in \mathbb{R}^n. \quad (8)$$

Proof. Step 1: we show (5) holds when f is L -smooth and convex. In fact, (5) is a direct result of Theorem 3.4 and Theorem 2.5.

Step 2: we show (5) implies (6). Fix $x \in \mathbb{R}^n$ and let $\phi(y) := f(y) - f(x) - \langle \nabla f(x), y - x \rangle$. By (5) we know $\phi(x) = 0$ is the minimum of ϕ . Substituting y by $y - \frac{1}{L} \nabla \phi(y)$ and applying (5), we obtain

$$\begin{aligned} 0 \leq \phi\left(y - \frac{1}{L} \nabla \phi(y)\right) &= f\left(y - \frac{1}{L} \nabla \phi(y)\right) - f(x) - \left\langle \nabla f(x), y - \frac{1}{L} \nabla \phi(y) - x \right\rangle \\ &\leq f(y) - \left\langle \nabla f(y), \frac{1}{L} \nabla \phi(y) \right\rangle + \frac{L}{2} \left\| \frac{1}{L} \nabla \phi(y) \right\|_2^2 - f(x) - \left\langle \nabla f(x), y - \frac{1}{L} \nabla \phi(y) - x \right\rangle \\ &= f(y) - f(x) - \langle \nabla f(x), y - x \rangle - \frac{1}{2L} \|\nabla f(y) - \nabla f(x)\|_2^2. \end{aligned}$$

Step 3: we show (6) implies (7).

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

Step 4: we show (7) implies f is L -smooth and convex. The convexity can be justified by directly applying Theorem 3.4. By Cauchy's inequality, we have

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

which implies

$$\|\nabla f(y) - \nabla f(x)\|_2 \leq L \|y - x\|_2.$$

Step 5: we show (5) implies (8). This can be achieved by directly adding two copies of (5) with x and y interchanged.

Step 6: we show (8) implies (5). By Theorem 3.4, it remains to show the second inequality in (5).

$$\begin{aligned}
f(y) - f(x) - \langle \nabla f(x), y - x \rangle &= \int_0^1 \langle \nabla f(x + t(y - x)) - \nabla f(x), y - x \rangle dt \\
&\leq \int_0^1 \frac{L}{t} \|t(y - x)\|_2^2 dt \\
&= \frac{L}{2} \|y - x\|_2^2.
\end{aligned}$$

□

We also has the following property for L -smooth and μ -strongly convex functions. By Theorem 2.5 and Theorem 3.7 we can easily know the fact that $\mu \leq L$.

Theorem 3.10 (L -smooth μ -strongly convex property). Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be L -smooth and μ -strongly convex, then for $\forall x, y \in \mathbb{R}^n$, the following inequality holds:

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

Proof. If $\mu = L$, we have

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

If $\mu < L$, $g(x) := f(x) - \frac{\mu}{2}\|x\|_2^2$ is $(L - \mu)$ -smooth convex function, implying

$$\langle \nabla g(y) - \nabla g(x), y - x \rangle \geq \frac{1}{L - \mu} \|\nabla g(y) - \nabla g(x)\|_2^2,$$

which is equivalent to

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

□