Chapter *5* Smoothness and convexity

Siheng Zhang

zhangsiheng@cvte.com

2021 年 11 月 1 日

This part corresponds to **Chapter 1,3,4 of PRML, Chapter of UML**. It mainly introduces some important properties regarding with functions: convexity, smoothness, strong convexity and Lipschitzness, which are basis for the next chapters.

目录

1   Convexity

Definition 5.1 (convex set) A set C in a vector space is convex if for any two vectors $\mathbf{u}, \mathbf{v}$ in C, the line segment between them is in C, that is, for any $\alpha \in [0, 1]$, $\alpha\mathbf{u} + (1 - \alpha)\mathbf{v} \in C$.

Definition 5.2 (convex function) Let C be a convex set. A function $f : C \to \mathcal{R}$ is convex if for all $\mathbf{u}, \mathbf{v} \in C$ and $\alpha \in [0, 1]$,

$$f(\alpha\mathbf{u} + (1 - \alpha)\mathbf{v}) \leq \alpha f(\mathbf{u}) + (1 - \alpha)f(\mathbf{v})$$

Claim 5.1 local minimum of convex function is global minimum

Proof *Let $B(\mathbf{u}, r) = \{\mathbf{v} : \|\mathbf{v} - \mathbf{u}\| \leq r\}$, we say that $f(\mathbf{u})$ is a local minimum if there exists some $r > 0$ such that $\forall \mathbf{v} \in B(\mathbf{u}, r), f(\mathbf{v}) \geq f(\mathbf{u})$. Then for any $\mathbf{v}$ (not necessarily in B), there exists a small enough $\alpha > 0$ such that $\mathbf{u} + \alpha(\mathbf{v} - \mathbf{u}) \in B(\mathbf{u}, r)$, and therefore $f(\mathbf{u}) \leq f(\mathbf{u} + \alpha(\mathbf{v} - \mathbf{u}))$. If f is convex, we also have $f(\mathbf{u} + \alpha(\mathbf{v} - \mathbf{u})) \leq (1 - \alpha)f(\mathbf{u}) + \alpha f(\mathbf{v})$, which leads to $f(\mathbf{u}) \leq f(\mathbf{v})$ and hence is a global minimum.*

Claim 5.2 For a convex function f, we can construct a tangent at any point that lies below the function everywhere. If f is differential, then we have

$$\forall \mathbf{u}, f(\mathbf{u}) \geq f(\mathbf{v}) + \nabla f(\mathbf{v})^\top(\mathbf{u} - \mathbf{v})$$

To generalize this inequality to non-differential functions, we should study for sub-gradient.

Definition 5.3 (sub-gradient) Define the sub-gradient of a function f at $\mathbf{x}$ to be a vector $\mathbf{g} \in \mathcal{R}^d$ which satisfies that $f(\mathbf{y}) \geq f(\mathbf{x}) + \mathbf{g}^\top(\mathbf{y} - \mathbf{x}), \forall \mathbf{y} \in \mathcal{R}^d$.

Theorem 5.1 Let f be a differential function, then f is convex *iff.* ∇f is monotonically non-decreasing, and *iff.* $\nabla^2 f$ is non-negative.

Ex1 $f(\mathbf{x}) = \mathbf{x}^\top\mathbf{x}$ is convex. To see this, note that $\nabla f(\mathbf{x}) = 2\mathbf{x}$ and $\nabla^2 f(\mathbf{x}) = 2$.

Ex2 Consider scalar function $f(x) = \log(1 + \exp(x))$. It is convex since $f'(x) = \frac{1}{1+\exp(-x)}$, which is the sigmoid function, and $f''(x) = \frac{\exp(-x)}{(1+\exp(-x))^2} > 0$.

Claim 5.1 (linear transformation preserves convexity) Assume that $g : \mathcal{R}^d \to \mathcal{R}$ can be written as $g(\mathbf{w}) = f(\mathbf{w}^\top\mathbf{x} + b)$, for some $\mathbf{x} \in \mathcal{R}^d, b \in \mathcal{R}$. Then, if f is convex, g is convex too.

Proof *For $\mathbf{w}_1, \mathbf{w}_2 \in \mathcal{R}^d$, and $\alpha \in (0, 1)$*

$$\begin{aligned}
g(\alpha\mathbf{w}_1 + (1 - \alpha)\mathbf{w}_2) &= f(\alpha\mathbf{w}_1^\top\mathbf{x} + (1 - \alpha)\mathbf{w}_2^\top\mathbf{x} + b) \\
&= f(\alpha(\mathbf{w}_1^\top\mathbf{x} + b) + (1 - \alpha)(\mathbf{w}_2^\top\mathbf{x} + b)) \\
&\leq \alpha f(\mathbf{w}_1^\top\mathbf{x} + b) + (1 - \alpha)f(\mathbf{w}_2^\top\mathbf{x} + b) = \alpha g(\mathbf{x}_1) + (1 - \alpha)g(\mathbf{x}_2)
\end{aligned}$$

Following **Claim5.1**,

Ex3 $f(\mathbf{w}) = (\mathbf{w}^\top\mathbf{x} - b)^2$ is convex.

Ex4 $f(\mathbf{w}) = \log(1 + \exp(-y\mathbf{w}^\top\mathbf{x}))$ is convex.

Claim 5.2 (Pointwise supremum of convex functions is convex) For $i = 1, \cdots, r$, let $f_i : \mathcal{R}^d \to \mathcal{R}$ be a convex function, then $g(\mathbf{x}) = \max_i f_i(\mathbf{x})$ is convex.

Proof *For $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{R}^d$, and $\alpha \in (0, 1)$*

$$g(\alpha\mathbf{x}_1 + (1 - \alpha)\mathbf{x}_2) = \max_i f_i(\alpha\mathbf{x}_1 + (1 - \alpha)\mathbf{x}_2) \leq \max_i[\alpha f_i(\mathbf{x}_1) + (1 - \alpha)f(\mathbf{x}_2)]$$

$$\leq \alpha \max_i f_i(\mathbf{x}_1) + (1 - \alpha)\max_i f_i(\mathbf{x}_2) = \alpha g(\mathbf{x}_1) + (1 - \alpha)g(\mathbf{x}_2)$$

Following **Claim5.2**,

Ex5 $g(\mathbf{x}) = |\mathbf{x}|$ is convex. Since $f_1(\mathbf{x}) = \mathbf{x}$, $f_2(\mathbf{x}) = -\mathbf{x}$ are both convex, and $g(\mathbf{x}) = \max\{f_1(\mathbf{x}), f_2(\mathbf{x})\}$.

Claim 5.3 (linear combination with non-negative weights preserves convexity) For $i = 1, \cdots, r$, let $f_i : \mathcal{R}^d \to \mathcal{R}$ be a convex function, then $\forall w_i \geq 0, i \in [1, r]$, $g(\mathbf{x}) = \sum_i w_i f_i(\mathbf{x})$ is convex.

Proof *For $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{R}^d$, and $\alpha \in (0, 1)$*

$$g(\alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2) = \sum_i w_i f_i(\alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2) \leq \sum_i w_i(\alpha f_i(\mathbf{x}_1) + (1-\alpha)f(\mathbf{x}_2))$$

$$= \alpha \sum_i w_i f_i(\mathbf{x}_1) + (1-\alpha) \sum_i w_i f(\mathbf{x}_2) = \alpha g(\mathbf{x}_1) + (1-\alpha)g(\mathbf{x}_2)$$

Theorem 5.2 Jensen's Inequality Let $f : \mathcal{R}^d \to \mathcal{R}$ be a measurable convex function and $x \in \mathcal{R}^d$, $\mathbb{E}[\S]$ exists. Then

$$\mathbb{E}[f(\mathbf{x})] \geq f(\mathbb{E}[\mathbf{x}])$$

Ex6 Consider function $f(x) = |x|$, then the sub-differential set $\partial f(x)$ is

$$\partial f(x) \begin{cases} \{1\}, \ x > 0 \\ [-1, 1], \ x = 0 \\ \{-1\}, \ x < 0 \end{cases}$$

Ex7 *Sub-gradient of Hinge loss.* Consider the loss $l(x) = \max(1 - \mathbf{z}^\top\mathbf{x}, 0)$, then the sub-differential set is

$$\partial f(x) \begin{cases} \{\mathbf{0}\}, \ 1 - \mathbf{z}^\top\mathbf{x} < 0 \\ \{-\alpha\mathbf{z}, \alpha \in [0, 1]\}, \ 1 - \mathbf{z}^\top\mathbf{x} = 0 \\ \{-\mathbf{z}\}, \ 1 - \mathbf{z}^\top\mathbf{x} > 0 \end{cases}$$

2   Strong convexity

Definition 5.5 (strongly convex function) A function f is λ-strongly convex if for all $\mathbf{u}, \mathbf{v} \in C$ and $\alpha \in [0, 1]$,

$$f(\alpha\mathbf{u} + (1-\alpha)\mathbf{v}) \leq \alpha f(\mathbf{u}) + (1-\alpha)f(\mathbf{v}) - \frac{\lambda}{2}\alpha(1-\alpha)\|\mathbf{u} - \mathbf{v}\|_2^2$$

Theorem 5.3 If f is λ-strongly convex then for every $\mathbf{u}, \mathbf{v}$ and $\mathbf{g} \in \partial f(\mathbf{v})$ we have

$$f(\mathbf{v}) \geq f(\mathbf{u}) + \mathbf{g}^\top(\mathbf{u} - \mathbf{v}) + \frac{\lambda}{2}\|\mathbf{u} - \mathbf{v}\|_2^2$$

Proof *According to the definition,*

$$
\begin{aligned}
f(\mathbf{u}) &\geq \frac{f(\alpha\mathbf{u} + (1-\alpha)\mathbf{v})}{\alpha} - \frac{1-\alpha}{\alpha}f(\mathbf{v}) + \frac{\lambda}{2}(1-\alpha)\|\mathbf{u} - \mathbf{v}\|_2^2 \\
&\geq \frac{f(\mathbf{v}) + \mathbf{g}^\top(\alpha\mathbf{u} + (1-\alpha)\mathbf{v}) - \mathbf{v})}{\alpha} - \frac{1-\alpha}{\alpha}f(\mathbf{v}) + \frac{\lambda}{2}(1-\alpha)\|\mathbf{u} - \mathbf{v}\|_2^2 \quad \textit{(Claim 5.2)} \\
&= \frac{f(\mathbf{v}) + \alpha\mathbf{g}^\top(\mathbf{u} - \mathbf{v})}{\alpha} - \frac{1-\alpha}{\alpha}f(\mathbf{v}) + \frac{\lambda}{2}(1-\alpha)\|\mathbf{u} - \mathbf{v}\|_2^2 \\
&= f(\mathbf{v}) + \mathbf{g}^\top(\mathbf{u} - \mathbf{v}) + \frac{\lambda}{2}(1-\alpha)\|\mathbf{u} - \mathbf{v}\|_2^2
\end{aligned}
$$

Note that it holds for all $\alpha \in [0, 1]$, setting $\alpha = 0$ leads to the conclusion.

Corollary 5.1 If f is λ-strongly convex and $\mathbf{v}$ is a minimizer of f, then for any $\mathbf{u}$, $f(\mathbf{v}) \geq f(\mathbf{u}) + \frac{\lambda}{2}\|\mathbf{u} - \mathbf{v}\|_2^2$.

Ex8 $f(\mathbf{x}) = \|\mathbf{x}\|_2^2$ is 2-strongly convex, $f(\mathbf{x}) = \lambda\|\mathbf{x}\|_2^2$ is 2λ-strongly convex. Note that $f(\mathbf{x}') = \lambda\|\mathbf{x}' - \mathbf{x} + \mathbf{x}\|_2^2 = \lambda\|\mathbf{x}\|_2^2 + 2\lambda\mathbf{x}^\top(\mathbf{x}' - \mathbf{x}) + \lambda\|\mathbf{x}' - \mathbf{x}\|_2^2$, in which $2\lambda\mathbf{x}$ is the gradient of $f(\mathbf{x})$.

Claim 5.4 If f is λ-strongly convex and g is convex, then $f + g$ is λ-strongly convex.

Following **Claim 5.4**, if we will to optimize a convex loss function, then with l_2-norm regularization, the loss function becomes strongly convex. This speeds up the convergence of learning procedure, which we will see in the next chapter.

3 Lipschitzness

Definition 5.6 (dual norm)

One way to understand 'dual norm' is that it is a way to measure how "big" are linear functionals. For example, consider the linear function f, we want to try to understand how big it is. So, we can measure 1 that is we measure how big is the output of the linear functional compared to its input x, where x is measured with some norm. Now, you can show that the above is equivalent to the dual norm of 1. The definition of dual norm immediately implies $\mathbf{w}^\top\mathbf{x} \leq \|\mathbf{w}\|_*\|\mathbf{x}\|$.

Ex9 Dual norm of l_2-norm is itself.

Definition 5.7 (Lipschitzness) A function $f : \mathcal{R}^d \to \mathcal{R}^k$ is ρ-Lipschitz over set C if for every $\mathbf{w}_1, \mathbf{w}_2 \in C$, $\|f(\mathbf{w}_1) - f(\mathbf{w}_2)\|_* \leq \rho\|\mathbf{w}_1 - \mathbf{w}_2\|$.

In this chapter, we just focus on l_2-norm. According to the mean value theorem, we have $f(\mathbf{w}_1) - f(\mathbf{w}_2) = g^\top(\mathbf{w}_1 - \mathbf{w}_2)$, where $\mathbf{g} \in \partial f(\mathbf{u})$, $\mathbf{u}$ is some point between $\mathbf{w}_1$ and $\mathbf{w}_2$. This leads to the following theorem.

Theorem 5.4 If f is Lipschitz *w.r.t* l_2-norm, *iff.* for all $\mathbf{g} \in \partial f(\mathbf{x})$, we have $\|\mathbf{g}\|_2 \leq \rho$.

Ex10 The function $f(x) = |x|$ is 1-Lipschitz over $\mathcal{R}$, since $|x_1| - |x_2| = |x_1 - x_2 + x_2| - |x_2| \leq |x_1 - x_2| + |x_2| - |x_2| = |x_1 - x_2|$.

Ex11 The function $f(x) = \log(1 + \exp(x))$ is 1-Lipschitz over $\mathcal{R}$, since $|f'(x)| = |\frac{1}{1+\exp(-x)}| \leq 1$.

Ex12 The function $f(x) = x^2$ is not ρ-Lipschitz over $\mathcal{R}$ for any ρ. However, over the set $C = \{x : |x| \leq \frac{\rho}{2}\}$, this function is ρ-Lipschitz. Indeed, for any $x_1, x_2 \in C$, $|x_1^2 - x_2^2| = |x_1 + x_2||x_1 - x_2| \leq 2 \cdot \frac{\rho}{2} \cdot |x_1 - x_2| = \rho|x_1 - x_2|$.

Ex13 Linear function $f(\mathbf{x}) = \mathbf{w}^\top\mathbf{x} + b$ is $\|\mathbf{w}\|$-Lipschitz following Cauchy-Schwartz inequality $|f(\mathbf{x}_1) - f(\mathbf{x}_2)| = |\mathbf{w}^\top(\mathbf{x}_1 - \mathbf{x}_2)| \leq \|\mathbf{w}\|_2\|\mathbf{x}_1 - \mathbf{x}_2\|_2$.

Claim 5.5 (Lipschitzness preserves on compound functions) Let $f(\mathbf{x}) = g_1(g_2(\mathbf{x}))$, where g_1 is ρ_1-Lipschitz and g_2 is ρ_2-Lipschitz. Then, f is $(\rho_1\rho_2)$-Lipschitz. In particular, if g_2 is the linear function, $g_2(\mathbf{x}) = \mathbf{w}^\top\mathbf{x} + b$, then f is $(\rho_1\|\mathbf{w}\|_2)$-Lipschitz.

Proof

$$|f(\mathbf{x}_1) - f(\mathbf{x}_2)| \leq \rho_1\|g_2(\mathbf{x}_1) - g_2(\mathbf{x}_2)\| \leq \rho_1\rho_2\|\mathbf{w}_1 - \mathbf{w}_2\|_2$$

Ex14 For binary classification $y \in \{-1, 1\}$, $f(\mathbf{w}) = y(\mathbf{w}^\top\mathbf{x} + b)$ is $\|\mathbf{x}\|_2$-Lipschitz following $|f(\mathbf{x}_1) - f(\mathbf{x}_2)| = |y\mathbf{w}_1^\top\mathbf{x} - y\mathbf{w}_2^\top\mathbf{x}_2| \leq \|y\mathbf{x}\|_2\|\mathbf{w}_1 - \mathbf{w}_2\|_2$. Hence, log loss $f(\mathbf{w}) = \log(1 + \exp(-y\mathbf{w}^\top\mathbf{x}))$ is $\|\mathbf{x}\|_2$-Lipschitz.

4 Smoothness

Definition 5.8 (smooth function) A differentiable function f is β-smooth if its gradient is β-Lipschitz

remark1. 1

4.4 According to definition, the dual norm of $\|\cdot\|_p (p \geq 1)$ is $\|\theta\|_* = \max_{\mathbf{x}:\|\mathbf{x}\|_p \leq 1} \langle \theta, \mathbf{x} \rangle$.

Using the Hölder inequality

$$\|\langle \mathbf{z}, \mathbf{x} \rangle\|_1 \leq \|\mathbf{x}\|_q \|\mathbf{z}\|_p$$

where $\frac{1}{p} + \frac{1}{q} = 1$, we have that:

$$\|\theta\|_* \leq \|\mathbf{x}\|_q \|\mathbf{z}\|_p \leq \|\mathbf{x}\|_q$$

So the dual norm is $\|\cdot\|_q$, where $\frac{1}{p} + \frac{1}{q} = 1$.

4.5 Start from strong convexity, we have that

$$\sum_{t=1}^{T} l_t(\mathbf{x}_t) - \sum_{t=1}^{T} l_t(\mathbf{u}) \leq \sum_{t=1}^{T} \left(\frac{1}{2\eta_t} \|\mathbf{x}_t - \mathbf{u}\|_2^2 - \frac{1}{2\eta_t} \|\mathbf{x}_{t+1} - \mathbf{u}\|_2^2 - \frac{\mu_t}{2} \|\mathbf{x}_t - \mathbf{u}\|_2^2 + \frac{\eta_t}{2} \|\mathbf{g}_t\|_2^2 \right)$$

Let all $mu_t \leq \mu$, choose $\eta_t = 1/\mu t$. Given that the loss functions are Lipschitz, we can further bound the regret to be

$$\sum_{t=1}^{T} l_t(\mathbf{x}_t) - \sum_{t=1}^{T} l_t(\mathbf{u}) \leq \sum_{t=1}^{T} \frac{\eta_t}{2} \|\mathbf{g}_t\|_2^2 = \frac{L^2}{2\mu} \sum_{t=1}^{T} \frac{1}{t}$$

Since that the loss functions are smooth, t can be rewrite as $t \geq \sum_{i=1}^{t} \frac{L\|\mathbf{x}-\mathbf{u}\|_2^2}{\|l_i(\mathbf{x})-l_i(\mathbf{u})\|_2^2} \geq \frac{L}{M} \sum_{i=1}^{t} \frac{\|\mathbf{g}_i\|_2^2}{\|l_i(\mathbf{x})-l_i(\mathbf{u})\|_2^2}$, so $\frac{1}{t} \leq \frac{M}{L} \frac{1}{\sum_{i=1}^{t} \frac{\|\mathbf{g}_i\|_2^2}{\|l_i(\mathbf{x})-l_i(\mathbf{u})\|_2^2}}$.

Now using the fact that $\sum_{t=1}^{T} \frac{1}{t} \leq 1 + \ln T$, we can also make integral by $\|l_i(\mathbf{x}) - l_i(\mathbf{u})\|_2^2$ and get that

$$\sum_{t=1}^{T} l_t(\mathbf{x}_t) - \sum_{t=1}^{T} l_t(\mathbf{u}) \leq O(\ln(1 + [\sum_{t=1}^{T} l_t(\mathbf{x}_t) - \sum_{t=1}^{T} l_t(\mathbf{u})] \sum_{t=1}^{T} \sum_{t=1}^{T} l_t(\mathbf{u})))$$

Using the fact that $x \leq a \ln(1 + bx) + c$ leads to $x \leq a \ln(2ab \ln(ab) + 2bc + 2) + c$, it leads to

$$\sum_{t=1}^{T} l_t(\mathbf{x}_t) - \sum_{t=1}^{T} l_t(\mathbf{u}) \leq O(1 + \ln \sum_{t=1}^{T} l_t(\mathbf{u})$$

which is the $O(1 + \ln L*)$ bound.

4.6 With regard to $\|\cdot\|_2$, we need to prove that

$$\|l''(\mathbf{x}_1)\|_2 \leq \frac{1}{4}$$

Note that

$$l'(\mathbf{x}) = \frac{-y\mathbf{z}e^{-y\langle \mathbf{z}, \mathbf{x} \rangle}}{1 + e^{-y\langle \mathbf{z}, \mathbf{x} \rangle}} = -y\mathbf{z}(1 - \exp(-l(\mathbf{x})))$$

$$l''(\mathbf{x}) = y\mathbf{z}\exp(-l(\mathbf{x}))(1 - \exp(-l(\mathbf{x})))$$

Since $y \in \{-1, 1\}$ and $\|\mathbf{z}\|_2 \leq 1$, we can bound it by

$$\|l''(\mathbf{x}_1)\|_2 \leq |\exp(-l(\mathbf{x}))(1 - \exp(-l(\mathbf{x})))|$$

.

Note that $l(\mathbf{x}) > 0$, so $0 < \exp(-l(\mathbf{x})) < 1$, so the upper bound is $1/4$.