Chapter *5* Smoothness and convexity

Siheng Zhang
zhangsiheng@cvte.com

2022 年 7 月 24 日

After discussing Bayesian model in Chapter 3, we made some simplification to get linear models, the easiest but important ones in machine learning. Now, we should take a break for some mathematical foundations, in order to introduce more complex model based on linear models.

This chapter 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. And the next chapter will introduces some optimization techniques, which are the tools for solving machine learning models.

目录

1 Convexity

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

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

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

Claim 5.1 local minimum of convex function is global minimum

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

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

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 x to be a vector $\mathsf{g} \in \mathcal{R}^d$ which satisfies that $f(\mathsf{y}) \geq f(\mathsf{x}) + \mathsf{g}^\top (\mathsf{y} - \mathsf{x}), \forall \mathsf{y} \in \mathcal{R}^d$.

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 $f(\mathsf{x}) = \mathsf{x}^\top \mathsf{x}$ is convex. To see this, note that $\nabla f(\mathsf{x}) = 2\mathsf{x}$ and $\nabla^2 f(\mathsf{x}) = 2$.
- 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 $g(\mathsf{w}) = f(\mathsf{w}^\top \mathsf{x} + b)$, for some $\mathsf{x} \in \mathcal{R}^d$, $b \in \mathcal{R}$. Then, if f is convex, g is convex too.

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

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

Following **Claim5.1**,

- Ex3 $f(\mathsf{w}) = (\mathsf{w}^\top \mathsf{x} - b)^2$ is convex.
- Ex4 $f(\mathsf{w}) = \log(1 + \exp(-y\mathsf{w}^\top \mathsf{x}))$ 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 $g(\mathsf{x}) = \max_i f_i(\mathsf{x})$ is convex.

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

$$\begin{aligned}
g(\alpha\mathsf{x}_1 + (1-\alpha)\mathsf{x}_2) &= \max_i f_i(\alpha\mathsf{x}_1 + (1-\alpha)\mathsf{x}_2) \leq \max_i [\alpha f_i(\mathsf{x}_1) + (1-\alpha) f(\mathsf{x}_2)] \\
&\leq \alpha \max_i f_i(\mathsf{x}_1) + (1-\alpha) \max_i f_i(\mathsf{x}_2) = \alpha g(\mathsf{x}_1) + (1-\alpha) g(\mathsf{x}_2)
\end{aligned}$$

Following **Claim5.2**,

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

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 $\forall w_i \geq 0, i \in [1, r]$, $g(\mathsf{x}) = \sum_i w_i f_i(\mathsf{x})$ is convex.

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

$$\begin{aligned} g(\alpha\mathsf{x}_1 + (1-\alpha)\mathsf{x}_2) &= \sum_i w_i f_i(\alpha\mathsf{x}_1 + (1-\alpha)\mathsf{x}_2) \leq \sum_i w_i(\alpha f_i(\mathsf{x}_1) + (1-\alpha)f(\mathsf{x}_2)) \\ &= \alpha\sum_i w_i f_i(\mathsf{x}_1) + (1-\alpha)\sum_i w_i f(\mathsf{x}_2) = \alpha g(\mathsf{x}_1) + (1-\alpha)g(\mathsf{x}_2) \end{aligned}$$

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

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

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 $l(x) = \max(1 - \mathsf{z}^\top\mathsf{x}, 0)$, then the sub-differential set is

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

2 Strong convexity

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

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

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

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

Proof *According to the definition,*

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

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 v is a minimizer of f, then for any u, $f(\mathsf{v}) \geq f(\mathsf{u}) + \frac{\lambda}{2}\|\mathsf{u}-\mathsf{v}\|_2^2$.

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

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 $\mathsf{w}^\top\mathsf{x} \leq \|\mathsf{w}\|_*\|\mathsf{x}\|$.

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 $\mathsf{w}_1, \mathsf{w}_2 \in C$, $\|f(\mathsf{w}_1) - f(\mathsf{w}_2)\|_* \leq \rho\|\mathsf{w}_1 - \mathsf{w}_2\|$.

In this chapter, we just focus on l_2-norm. According to the mean value theorem, we have $f(\mathsf{w}_1) - f(\mathsf{w}_2) = g^\top(\mathsf{w}_1-\mathsf{w}_2)$, where $\mathsf{g} \in \partial f(\mathsf{u})$, u is some point between w_1 and w_2. This leads to the following theorem.

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

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

Claim 5.5 (Lipschitzness preserves on compound functions) Let $f(\mathsf{x}) = g_1(g_2(\mathsf{x}))$, 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, $g_2(\mathsf{x}) = \mathsf{w}^\top\mathsf{x}+b$, then f is $(\rho_1\|\mathsf{w}\|_2)$-Lipschitz.

Proof

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

Ex14 For binary classification $y \in \{-1,1\}$, $f(\mathsf{w}) = y(\mathsf{w}^\top\mathsf{x}+b)$ is $\|\mathsf{x}\|_2$-Lipschitz following $|f(\mathsf{x}_1)-f(\mathsf{x}_2)| = |y\mathsf{w}_1^\top\mathsf{x} - y\mathsf{w}_2^\top\mathsf{x}_2| \leq \|y\mathsf{x}\|_2\|\mathsf{w}_1-\mathsf{w}_2\|_2$. Hence, log loss $f(\mathsf{w}) = \log(1+\exp(-y\mathsf{w}^\top\mathsf{x}))$ is $\|\mathsf{x}\|_2$-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 $\|\theta\|_* = \max_{\mathsf{x}:\|\mathsf{x}\|_p \leq 1} \langle \theta, \mathsf{x} \rangle$.

Using the Hölder inequality

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

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

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

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

4.5 Start from strong convexity, we have that

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

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

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

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

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

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

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

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

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

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

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

Note that

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

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

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

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

.

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