Contents

1 Probability and training algorithms 3
1.1 Convex functions and convergence of gradient descent 3
1.1.1 Convex function 3
1.1.2 On the Convergence of GD 6

1

Probability and training algorithms

1.1 Convex functions and convergence of gradient descent

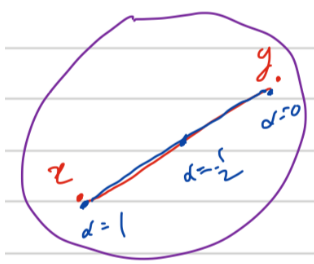
1.1.1 Convex function

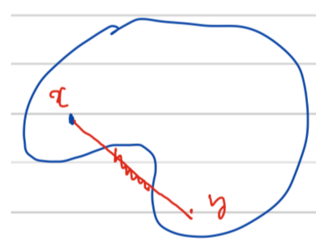
Then, let us first give the definition of convex sets.

Definition 1 (Convex set). *A set C is convex, if the line segment between any two points in C lies in C, i.e., if any $x, y \in C$ and any α with $0 \leq \alpha \leq 1$, there holds*

$$\alpha x + (1 - \alpha) y \in C. \tag{1.1}$$

Here are two diagrams for this definition about convex and non-convex sets.





Following the definition of convex set, we define convex function as following.

Definition 2 (Convex function). *Let $C \subset \mathbb{R}^n$ be a convex set and $f : C \to \mathbb{R}$:*

1. *f is called* **convex** *if for any $x, y \in C$ and $\alpha \in [0, 1]$*

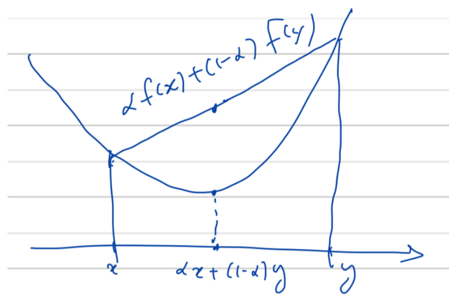
$$f(\alpha x + (1 - \alpha) y) \leq \alpha f(x) + (1 - \alpha) f(y). \tag{1.2}$$

2. *f is called* **strictly convex** *if for any $x \neq y \in C$ and $\alpha \in (0, 1)$:*

$$f(\alpha x + (1 - \alpha) y) < \alpha f(x) + (1 - \alpha) f(y). \tag{1.3}$$

3. *A function f is said to be (strictly)* **concave** *if $-f$ is (strictly) convex.*

We also have the next diagram for convex function definition.



Lemma 1. *If $f(x)$ is differentiable on $\mathbb{R}^n$, then $f(x)$ is convex if and only if*

$$f(x) \geq f(y) + \nabla f(y) \cdot (x - y), \forall x, y \in \mathbb{R}^n. \tag{1.4}$$

Based on the lemma, we can first have the next new diagram for convex functions.

Proof. Let $z = \alpha x + (1 - \alpha)y, 0 \leq \alpha \leq 1, \forall x, y \in \mathbb{R}^n$, we have these next two Taylor expansion:

$$\begin{aligned} f(x) &\geq f(z) + \nabla f(z)(x - z) \\ f(y) &\geq f(z) + \nabla f(z)(y - z). \end{aligned} \tag{1.5}$$

Then we have

$$\begin{aligned} &\alpha f(x) + (1 - \alpha) f(y) \\ \geq &f(z) + \nabla f(z)[\alpha(x - z) + (1 - \alpha)(y - z)] \\ = &f(z) \\ = &f(\alpha x + (1 - \alpha)y). \end{aligned} \tag{1.6}$$

Thus we have

$$\alpha f(x) + (1-\alpha) f(y) \geq f(\alpha x + (1-\alpha) y). \tag{1.7}$$

This finishes the proof.

On the other hand (**homework**): if $f(x)$ is differentiable on $\mathbb{R}^n$, then $f(x) \geq f(y) + \nabla f(y) \cdot (x-y), \forall x, y \in \mathbb{R}^n$ if $f(x)$ is convex. □

Definition 3 (λ-strongly convex). *We say that $f(x)$ is λ-strongly convex if*

$$f(x) \geq f(y) + \nabla f(y) \cdot (x-y) + \frac{\lambda}{2}\|x-y\|^2, \quad \forall x, y \in C, \tag{1.8}$$

for some $\lambda > 0$.

Example 1. Consider $f(x) = \|x\|^2$, then we have

$$\frac{\partial f}{\partial x_i} = 2x_i, \nabla f = 2x \in R^n. \tag{1.9}$$

So, we have

$$\begin{aligned} & f(x) - f(y) - \nabla f(y)(x-y) \\ =& \|x\|^2 - \|y\|^2 - 2y(x-y) \\ =& \|x\|^2 - \|y\|^2 - 2xy + 2\|y\|^2 \\ =& \|x\|^2 - 2xy + \|y\|^2 \\ =& \|x-y\|^2 \\ =& \frac{\lambda}{2}\|x-y\|^2, \quad \lambda = 2. \end{aligned} \tag{1.10}$$

Thus, $f(x) = \|x\|^2$ is 2-strongly convex

Example 2 (Homework). Actually, the loss function of the logistic regression model

$$L(\theta) = -\log P(\theta), \tag{1.11}$$

is convex as a function of θ.

Furthermore, the loss function of the regularized logistic regression model

$$L_\lambda(\theta) = -\log P(\theta) + \lambda\|\theta\|_F^2, \lambda > 0 \tag{1.12}$$

is λ'-strongly convex (λ' is related to λ) as a function of θ.

We also have these following interesting properties of convex function.

Properties 1 (basic properties of convex function) *[Homework]*

1. *If $f(x)$, $g(x)$ are both convex, then $\alpha f(x) + \beta g(x)$ is also convex, if $\alpha, \beta \geq 0$.*

2. *Linear function is both convex and concave. Here, $f(x)$ is concave if and only if $-f(x)$ is convex.*
3. *If $f(x)$ is a convex convex function on $\mathbb{R}^n$, then $g(y) = f(Ay + b)$ is a convex function on $\mathbb{R}^m$. Here $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$.*
4. *If $g(x)$ is a convex function on $\mathbb{R}^n$, and the function $f(u)$ is convex function on $\mathbb{R}$ and non-decreasing, then the composite function $f \circ g(x) = f(g(x))$ is convex.*

Proof. **Homework**: prove them by definition. □

1.1.2 On the Convergence of GD

For the next optimization problem

$$\min_{x\in\mathbb{R}^n} f(x). \tag{1.13}$$

We assume that $f(x)$ is convex. Then we say that x^* is a minimizer if $f(x^*) = \min_{x\in\mathbb{R}^n} f(x)$.

Let recall that, for minimizer x^* we have

$$\nabla f(x^*) = 0. \tag{1.14}$$

Then we have the next tw properties of minimizer for convex functions:

1. If $f(x) \geq c_0$, for some $c_0 \in \mathbb{R}$, then we have

$$\arg\min f \neq \emptyset. \tag{1.15}$$

2. If $f(x)$ is λ-strongly convex, then $f(x)$ has a unique minimizer, namely, there exists a unique $x^* \in \mathbb{R}^n$ such that

$$f(x^*) = \min_{x\in\mathbb{R}^n} f(x). \tag{1.16}$$

To investigate the convergence of gradient descent method, let recall the gradient descent method:

Algorithm 1 FGD

For: $t = 1, 2, \cdots$

$$x_{t+1} = x_t - \eta_t \nabla f(x_t), \tag{1.17}$$

where η_t is the stepsize / learning rate.

Assumption 1.18 *We make the following assumptions*

1. *$f(x)$ is λ-strongly convex for some $\lambda > 0$. Recall the definition, we have*

$$f(x) \geq f(y) + \nabla f(y) \cdot (x - y) + \frac{\lambda}{2}\|x - y\|^2,$$

then note $x^* = \arg\min f(x)$. Then we have

- *Take $y = x^*$, this leads to*

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

- *Take $x = x^*$, this leads to*

$$0 \geq f(x^*) - f(y) \geq \nabla f(y) \cdot (x^* - y) + \frac{\lambda}{2}\|x^* - y\|^2,$$

which means that

$$\nabla f(x) \cdot (x - x^*) \geq \frac{\lambda}{2}\|x - x^*\|^2. \tag{1.19}$$

2. *∇f is Lipschitz for some $L > 0$, i.e.,*

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \forall x, y. \tag{1.20}$$

Thus, we have the next theorem about the convergence of gradient descent method.

Theorem 2. *For Algorithm 1, if $f(x)$ is λ-strongly convex and ∇f is Lipschitz for some $L > 0$, then*

$$\|x_t - x^*\|^2 \leq \alpha^t \|x_0 - x^*\|^2 \tag{1.21}$$

if $0 < \eta_t \leq \eta_0 = \frac{\lambda}{2L^2}$ and $\alpha = 1 - \frac{\lambda^2}{4L^2} < 1$.

Proof. If we minus any $x \in \mathbb{R}^n$, we can only get:

$$x_{t+1} - x = x_t - \eta_t \nabla f(x_t) - x. \tag{1.22}$$

If we take L^2 norm for both side, we get:

$$\|x_{t+1} - x\|^2 = \|x_t - \eta_t \nabla f(x_t) - x\|^2. \tag{1.23}$$

So we have the following inequality and take $x = x^*$:

$$\begin{aligned}
\|x_{t+1} - x^*\|^2 &= \|x_t - \eta_t \nabla f(x_t) - x^*\|^2 \\
&= \|x_t - x^*\|^2 - 2\eta_t \nabla f(x_t)^\top (x_t - x^*) + \eta_t^2 \|\nabla f(x_t) - \nabla f(x^*)\|^2 \\
&\leq \|x_t - x^*\|^2 - \eta_t \lambda \|x_t - x^*\|^2 + \eta_t^2 L^2 \|x_t - x^*\|^2 \quad (\lambda - \text{strongly convex and Lipschitz}) \\
&\leq (1 - \eta_t \lambda + \eta_t^2 L^2) \|x_t - x^*\|.
\end{aligned} \tag{1.24}$$

So, if $\eta_t \leq \frac{\lambda}{2L^2}$, then $\alpha = (1 - \eta_t \lambda + \eta_t^2 L^2) \leq 1 - \frac{\lambda^2}{4L^2} < 1$, which finishes the proof. □

Some issues on GD:

- $\nabla f(x_t)$ is very expensive to compete.
- GD does not yield generalization accuracy.

The stochastic gradient descent (SGD) method which we will discuss in the next section will focus on these two issues.