Jinchao Xu

Deep Learning Algorithms and Analysis

Summer 2020

Contributors:

This set of notes are based on contributions from many of graduate students, post-doctoral fellows and other collaborators. Here is a partial list:

Juncai He, Qingguo Hong, Li Jiang.....

Contents

1

Logistic Regression

1.1 Optimization: gradient descent method

We finish the study of logistic regression with solving the next optimization problem

$$\min_{\boldsymbol{\theta}} L_\lambda(\boldsymbol{\theta}). \tag{1.1}$$

For simplicity, let us just consider the next general optimization problem

$$\min_{x\in\mathbb{R}^n} f(x). \tag{1.2}$$

A general approach: line search method

Here we propose the next descent scheme to produce $\{x_t\}_{t=1}^{\infty}$

$$x_{t+1} = x_t + \eta_t p_t, \tag{1.3}$$

where η_t is called the step size in optimization and also learning rate in machine learning. p_t is called the descent direction, which is the critical component of this algorithm. There is a series of optimization algorithms which follow the above form just using different choices of p_t.

The main method that we will discuss here is the so-called gradient descent method.

1.1.1 Multi-variable calculus

To begin with the gradient based optimization, it is necessary to review some multi-variable calculus aspects and definition of convex functions.

At the very beginning, let us recall the definition of gradient and Hessian matrix for function $f : \mathbb{R}^n \to \mathbb{R}$.

Definition 1. *Given objective function $f : \mathbb{R}^n \to \mathbb{R}$ and $x = (x_1, x_2, \ldots, x_n)^T \in \mathbb{R}^n$, the gradient of $f(x)$ is defined by*

$$g(x) := \nabla f(x) = \begin{pmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{pmatrix}, \tag{1.4}$$

Then, the next natural question is what a good choice of p_t is? We have the next theorem to show why gradient direction is a good choice for p_t.

Before that, let us introduce one often-used inequality.

Lemma 1 (Cauchy-Schwarz inequality). *For any $\mathbf{x} = (x_1, \ldots, x_n)^T$ and $\mathbf{x} = (y_1, \ldots, y_n)^T$, we have*

$$\left(\sum_{i=1}^{n} x_i y_i\right)^2 \leq \left(\sum_{i=1}^{n} x_i^2\right)\left(\sum_{i=1}^{n} y_i^2\right), \tag{1.5}$$

where equality holds if and only if for some $k \in \mathbb{C}$, $\frac{x_i}{y_i} = k$, or in inner form:

$$|\langle \mathbf{x}, \mathbf{y} \rangle| \leq \|\mathbf{x}\| \cdot \|\mathbf{y}\|. \tag{1.6}$$

Proof. In the case of $\mathbf{y} = \mathbf{0}$, the inequality holds. Now assume $\mathbf{y} \neq \mathbf{0}$, then

$$\begin{aligned} \|\mathbf{x} - \lambda\mathbf{y}\|^2 =& \langle \mathbf{x}, \mathbf{x} \rangle - \langle \mathbf{x}, \lambda\mathbf{y} \rangle - \langle \lambda\mathbf{y}, \mathbf{x} \rangle + \langle \lambda\mathbf{y}, \lambda\mathbf{y} \rangle \\ =& \langle \mathbf{x}, \mathbf{x} \rangle - \bar{\lambda}\langle \mathbf{x}, \mathbf{y} \rangle - \lambda\langle \mathbf{y}, \mathbf{x} \rangle + \bar{\lambda}\lambda\langle \mathbf{y}, \mathbf{y} \rangle \\ =& \|\mathbf{x}\|^2 - \bar{\lambda}\langle \mathbf{x}, \mathbf{y} \rangle - \lambda\overline{\langle \mathbf{x}, \mathbf{y} \rangle} + \bar{\lambda}\lambda\|\mathbf{y}\|^2, \end{aligned} \tag{1.7}$$

Let $\lambda = \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{y}\|^2}$, the above formula can be

$$0 \leq \|\mathbf{x} - \lambda\mathbf{y}\|^2 = \|\mathbf{x}\|^2 - \frac{|\langle \mathbf{x}, \mathbf{y} \rangle|^2}{\|\mathbf{y}\|^2}. \tag{1.8}$$

If the inequality holds as an equality, $\mathbf{x}$ and $\mathbf{y}$ are linearly dependent. □

Theorem 1. *To choose descent directions $p \in \mathbb{R}^n$, we have*

$$-\frac{\nabla f(x)}{\|\nabla f(x)\|} = \arg\min_{p\in\mathbb{R}^n, \|p\|=1} \left.\frac{\partial f(x+\eta p)}{\partial \eta}\right|_{\eta=0}. \tag{1.9}$$

Proof. First, we have

$$\left.\frac{\partial f(x+\eta p)}{\partial \eta}\right|_{\eta=0} = \nabla f(x)\cdot p. \tag{1.10}$$

Then notice the Cauchy-Schwarz inequality and the constrain that $\|p\| = 1$,

$$|\nabla f(x)\cdot p| \le \|\nabla f(x)\|\|p\| = \|\nabla f(x)\|, \tag{1.11}$$

which mean that

$$\nabla f(x)\cdot p \ge -\|\nabla f(x)\|, \tag{1.12}$$

and the equality holds when $p = -\frac{\nabla f(x)}{\|\nabla f(x)\|}$. □

Corollary 1. *Locally, $f(x)$ decreases most rapidly along the negative gradient direction.*

Here is a simple diagram for this property.

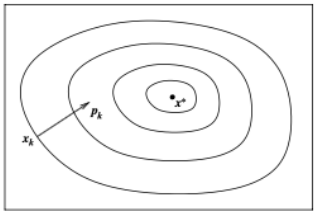

Fig. 1.1. Negative Gradient Direction: x_k is current point, p_k is the negative gradient of x_k, i.e., $-\nabla f(x_k)$

.

Since at each point, $f(x)$ decreases most rapidly along the negative gradient direction, it is then natural to choose the search direction in (1.3) in the negative gradient direction and the resulting algorithm is the so-called gradient descent method.

Algorithm 1 Gradient Descent Method

Given the initial guess x_0, learning rate $\eta_t > 0$
For t=1,2,$\cdots$,

$$x_{t+1} = x_t - \eta_t \nabla f(x_t), \tag{1.13}$$

In practice, we need a "stopping criterion" that determines when the above gradient descent method to stop. One possibility is

While $S(x_t; f) = \|\nabla f(x_t)\| \leq \epsilon$ or $t \geq T$

for some small tolerance $\epsilon > 0$ or maximal number of iterations T. In general, a good stopping criterion is hard to come by and it is subject that has called a lot of research in optimization for machine learning.

1.1.2 Convex function

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

Definition 2 (Convex set). *A set C is convex, if the line segment between any two points in C lies in C, i.e., if any $x_1, x_2 \in C$ and any α with $0 \leq \alpha \leq 1$, there holds*

$$\alpha x_1 + (1 - \alpha) x_2 \in C. \tag{1.14}$$

Based on the definition above, we have the next property of convex sets.

Lemma 2. *Let C be a convex set, with $x_1, \ldots, x_k \in C$, and let $\alpha_1, \ldots, \alpha_k \in \mathbb{R}$ satisfy $\alpha_i \geq 0$ and $\sum_{i=1}^{k} \alpha_i = 1$, then*

$$\sum_{i=1}^{k} \alpha_i x_i \in C, \tag{1.15}$$

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

Definition 3 (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 $x_1, x_2 \in C$ and $\alpha \in [0, 1]$*

$$f(\alpha x_1 + (1 - \alpha) x_2) \leq \alpha f(x_1) + (1 - \alpha) f(x_2). \tag{1.16}$$

2. *f is called* **strictly convex** *if for any $x_1 \neq x_2 \in C$ and $\alpha \in (0, 1)$:*

$$f(\alpha x_1 + (1 - \alpha) x_2) < \alpha f(x_1) + (1 - \alpha) f(x_2). \tag{1.17}$$

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

With mentioned previously property of convex sets, it is easy to expend the above definition of convex function. Here we write as one lemma without proof.

Lemma 3. *Let C be a convex set, with $x_1, \ldots, x_k \in C$, and let $\alpha_1, \ldots, \alpha_k \in \mathbb{R}$ satisfy $\alpha_i \geq 0$ and $\sum_{i=1}^{k} \alpha_i = 1$, then*

$$f(\sum_{i=1}^{k} \alpha_i x_i) \leq \sum_{i=1}^{k} \alpha_i f(x_i). \tag{1.18}$$

We also have these following interesting properties of convex function. You can prove them as exercises just with the above definitions.

Properties 2 (basic properties of convex function)

1. *Linear combination of convex functions with positive coefficients is convex function.*
2. *Linear function is both convex and concave.*
3. *A convex function composited with a linear function is convex.*
4. *If the function $u = g(x)$ is concave, and the function $y = f(u)$ is convex and non-increasing, then the composite function $f \circ g$ is convex.*
5. *If the function $u = g(x)$ is convex, and the function $y = f(u)$ is convex and non-decreasing, then the composite function $f \circ g$ is convex.*

Definition 4. *The Hessian matrix of $f(x)$*

$$H(x) := \nabla^2 f(x) = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial {x_1}^2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x)}{\partial {x_2}^2} & \cdots & \frac{\partial^2 f(x)}{\partial x_2 \partial x_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial^2 f(x)}{\partial x_n \partial x_1} & \frac{\partial^2 f(x)}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial {x_n}^2} \end{bmatrix}, \tag{1.19}$$

From the perspective of multi-variable calculus, we also have the next important theorem to describe differentiable convex function.

Theorem 3. *If $f(x)$ is twice continuously differentiable on the non-empty and open convex set $C \subset \mathbb{R}^n$, then*

- *$f(x)$ is convex if and only if its Hessian matrix $H(x) = \nabla^2 f(x) \geq 0$ for every $x \in C$.*
- *$f(x)$ is strictly convex if $H(x) > 0$ for every $x \in C$.*

Here, for any matrix $A \in \mathbb{R}^{n \times n}$, and

$$A \geq 0 \Leftrightarrow v^T A v \geq 0, \quad \forall v \in \mathbb{R}^n, \tag{1.20}$$

we call A as a positive semidefinite matrix. Correspondingly, $A > 0$ is defined with $v^T A v > 0$ for any $v \neq 0$, which is so-called positive definite matrix.

Theorem 4. *If $f(x)$ is differentiable on the non-empty and open convex set $C \subset \mathbb{R}^n$, then $f(x)$ is convex if and only if*

$$f(x_1) - f(x_2) \geq \nabla f(x_2)(x_1 - x_2), \forall x_1, x_2 \in C. \tag{1.21}$$

$f(x)$ is convex if and only if

$$f(x_1) - f(x_2) > \nabla f(x_2)(x_1 - x_2), \forall x_1, x_2 \in C, x_1 \neq x_2. \tag{1.22}$$

Following the theorem above, we can prove the next important theorem for convex function. Before that, we need to clearly define several solution points of optimization problem. One general optimal setting:

$$\min_{x \in C} f(x). \tag{1.23}$$

For the above objective function,

Definition 5. *Let $x^* \in C$, x^* is a local minima, if exists $\delta > 0$, for any x satisfies $\|x - x^*\| < \delta$, $f(x^*) \leq f(x)$. Especially if adding that $x \neq x^*$, $f(x^*) < f(x)$, then x^* is one strictly local minima of f.*

Theorem 5 (First order requirement). *Let $f : C \subset \mathbb{R}^n \to \mathbb{R}$ is continuous and differentiable on open set C, if x^* is a local minima, then*

$$\nabla f(x^*) = 0. \tag{1.24}$$

Theorem 6. *Suppose that $f : C \to \mathbb{R}$ is differentiable and convex over set C, any local minima of f is also its global minima.*

1.1.3 Convergence of gradient Descent Method

Based on Theorem 6, to optimize $f(x)$, we hope to find a sequence $\{x_t\}_{t=1}^{\infty}$ such that

$$\lim_{t \to \infty} \|x_t - x^*\| \quad \text{or} \quad \lim_{t \to \infty} f(x_t) = \inf_x f(x) \quad \text{or} \quad \lim_{t \to \infty} \|\nabla f(x_t)\| = 0, \tag{1.25}$$

for some $x^* \in \arg\min_x f(x)$.

Theorem 7. *Suppose that $f(x)$ is convex, differentiable, and Lipschitz continuous with constant L for $\nabla f(x)$, i.e., $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$. Then, for any small constant step size*

$$\eta_t = \eta \leq = \frac{2\alpha}{L}, \tag{1.26}$$

while $\alpha \in (01,)$ then we will have

$$\min_{1 \leq t \leq T} \{\|\nabla f(x_t)\|\} \leq \frac{C}{\sqrt{T}}. \tag{1.27}$$