Notes on Optimization

Liu Zhizhou

First Created: August 6, 2022
Last Modified: August 7, 2022

Contents

1 Derivatives

Let $f : \mathbb{R}^n \to \mathbb{R}$. The *gradient* of f at x is defined as the column vector

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix}.$$

If f is a vector-valued function, i.e. $f : \mathbb{R}^n \to \mathbb{R}^m$, with component functions $f_1, \ldots, f_m$, then

$$\nabla f(x) = \begin{bmatrix} \nabla f_1(x) & \cdots & \nabla f_m(x) \end{bmatrix}.$$

The transpose of ∇f is called the *Jacobian* of f. The Jacobian of f is the matrix whose ij-th entry is equal to the partial derivative $\frac{\partial f_i}{\partial x_j}$.

The *Hessian* of $f : \mathbb{R}^n \to \mathbb{R}$ is the matrix whose ij-th entry is equal to $\frac{\partial^2 f}{\partial x_i \partial x_j}$, denoted by $\nabla^2 f$.

Be careful that, for $f : \mathbb{R}^n \to \mathbb{R}$, $\nabla^2 f \neq \nabla(\nabla f)$, but $\nabla^2 f = \nabla(\nabla f^T)$.

Proposition 1.1 (chain rule). Let $f : \mathbb{R}^k \to \mathbb{R}^m$ and $g : \mathbb{R}^m \to \mathbb{R}^n$ be smooth functions, and $h = g(f(x))$. Then

$$\nabla h(x) = \nabla f(x) \nabla(g(f(x)))$$

for all $x \in \mathbb{R}^k$.

Some useful relations:

1. $\nabla(Ax) = A^T$;

2. $\nabla(x^T Ax) = (A+A^T)x$; in particular, if Q is symmetric, then $\nabla(x^T Qx) = 2Qx$ and $\nabla(\|x\|^2) = \nabla(x^T x) = 2x$;
3. $\nabla(f(Ax)) = A^T \nabla f(Ax)$;
4. $\nabla^2(f(Ax)) = A^T \nabla^2 f(Ax) A$;

The shape of the left hand side would be helpful to memorize the right hand side.

Theorem 1.2 (Second Order Taylor Expansions). Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable over an open sphere S centered at a vector x. Then for all d such that $x + d \in S$,

1. we have
$$f(x+d) = f(x) + d^T \nabla f(x) + \frac{1}{2} d^T \left(\int_0^1 \left(\int_0^t \nabla^2 f(x + \tau d) \mathrm{d}\tau \right) \mathrm{d}t \right) d.$$
2. there exists
$$f(x+d) = f(x) + d^T \nabla f(x) + \frac{1}{2} d^T \nabla^2 f(x + \alpha d) d.$$
3. there holds
$$f(x+d) = f(x) + d^T \nabla f(x) + \frac{1}{2} d^T \nabla^2 f(x) d + o(\|d\|^2).$$

2 Convexity

Definition 2.1 (convex set, convex function). A subset C of $\mathbb{R}^n$ is called *convex* if
$$\alpha x + (1-\alpha) y \in C$$
for all $x, y \in C$ and $\alpha \in [0,1]$. A function $f : C \to \mathbb{R}$ is called *convex* if
$$f(\alpha x + (1-\alpha) y) \le \alpha f(x) + (1-\alpha) f(y) \tag{2.1}$$
for $x, y \in C$ and $\alpha \in [0,1]$. The function is called *concave* if $-f$ is convex.

Definition 2.2 (strictly convex). The function f is called *strictly convex* if Eq.(2.1) is strict for all $x \neq y$ and $\alpha \in (0,1)$.

Proposition 2.3 (First Derivative Characterizations). Let C be a convex subset of $\mathbb{R}^n$ and let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable over $\mathbb{R}^n$. Then

1. f is convex over C if and only if
$$f(z) \ge f(x) + (z-x)^T \nabla f(x) \tag{2.2}$$
for all $x, z \in C$.

2. f is strictly convex over C if and only if the above inequality is strict whenever $x \neq z$.

Definition 2.4 (strongly convex). A function $f : \mathbb{R}^n \to \mathbb{R}$ is called *strongly convex* if for some $\sigma > 0$, we have

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{\sigma}{2} \|x - y\|^2 \tag{2.3}$$

for all $x, y \in \mathbb{R}^n$.

It can be shown that an equivalent definition is that

$$(\nabla f(x) - \nabla f(y))^T (x - y) \geq \sigma \|x - y\|^2 \tag{2.4}$$

for all $x, y \in \mathbb{R}^n$.

3 Main Optimality Conditions

Theorem 3.1 (Necessary Optimality Conditions). Let x^* be an unconstrained local minimum of $f : \mathbb{R}^n \to \mathbb{R}$, and assume that f is continuously differentiable in an open set S containing x^*. Then we have the *First Order Necessary Condition*:

$$\nabla f(x^*) = 0. \tag{3.1}$$

If in addition f is twice continuously differentiable within S, then we have the *Second Order Necessary Condition*:

$$\nabla^2 f(x^*) \succeq 0. \tag{3.2}$$

The intuition of this theorem is considering

$$f(x^* + \Delta x) - f(x^*) \approx \nabla f(x^*)^T \Delta x,$$

and similarly for second order,

$$f(x^* + \Delta x) - f(x^*) \approx \nabla f(x^*)^T \Delta x + \frac{1}{2} \Delta x^T \nabla^2 f(x^*) \Delta x.$$

Read rigorous proof to see the reason.

Proof. Fix some $d \in \mathbb{R}^n$. Consider $g(\alpha) \triangleq f(x^* + \alpha d)$. Then

$$0 \leq \lim_{\alpha \to 0} \frac{f(x^* + \alpha d) - f(x^*)}{\alpha} = \frac{\mathrm{d}g}{\mathrm{d}\alpha}(0) = d^T \nabla f(x^*).$$

The "$\leq$" is because x^* is the local minimum. Replace d by $-d$, then it must be $\nabla f(x^*) = 0$.
Assume f is twice differentiable. Then the second order expansion of $g(\alpha)$ in $\alpha = 0$ yields

$$g(\alpha) = g(0) + \frac{\mathrm{d}g}{\mathrm{d}\alpha}(0)\alpha + \frac{1}{2} \frac{\mathrm{d}^2 g}{\mathrm{d}\alpha^2}(0)\alpha^2 + o(\alpha^2).$$

Equivalently,

$$f(x^* + \alpha d) - f(x^*) = d^T \nabla f(x^*)\alpha + \frac{\alpha^2}{2} d^T \nabla^2 f(x^*) d + o(\alpha^2).$$

Since $\nabla f(x^*) = 0$, for α positive and near 0, we have

$$0 \leq \frac{f(x^* + \alpha d) - f(x^*)}{\alpha^2} = \frac{1}{2} d^T \nabla^2 f(x^*) d + \frac{o(\alpha^2)}{\alpha^2}.$$

Then let $\alpha \to 0$, we obtain $d^T \nabla^2 f(x^*) d \geq 0$, which means $\nabla^2 f(x^*) \succeq 0$. □

Proposition 3.2. If X is a convex subset of $\mathbb{R}^n$ and $f : \mathbb{R}^n \to \mathbb{R}$ is convex over X, then a local minimum of f is also a global minimum. If in addition f is strictly convex over X, then f has at most one global minimum over X. Moreover, if f is strongly convex and X is closed, then f has a unique global minimum over X.

Theorem 3.3 (Convex Case - Necessary and Sufficient Conditions). Let X be a convex set and let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function over X. Then

1. If f is continuously differentiable, then

$$\nabla f(x^*)^T (x - x^*) \geq 0$$

   for all $x \in X$ is a necessary and sufficient condition for x^* to be a global minimum of f over X.

2. If X is open and f is continuously differentiable over X, then $\nabla f(x^*) = 0$ is a necessary and sufficient condition for x^* to be a global minimum of f over X.

Note that in the second statement, we require X to be open.
The intuition of this theorem is also

$$f(x^* + \Delta x) - f(x^*) \approx \nabla f(x^*)^T \Delta x.$$

The proof of this need the first order characterization of convexity,

$$f(x) \geq f(x^*) + \nabla f(x^*)^T (x - x^*)$$

for all $x \in X$.

A geometric illustration of $\nabla f(x^*)^T (x - x^*)$ is that: $\nabla f(x^*)$ is the direction that f increase the most, the condition means that the connection of x^* and all feasible points x in X has angle less than $\frac{\pi}{2}$ with the gradient; in other words, all the direction makes f increase.

Theorem 3.4 (Second Order Sufficient Optimality Conditions). Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable over an open set S. Suppose that a vector $x^* \in S$ satisfies the conditions: (i) $\nabla f(x^*) = 0$ and (ii) $\nabla^2 f(x^*) \succ 0$. Then x^* is a strict unconstrained local minimum of f. In particular, there exists scalars $\gamma > 0$ and $\epsilon > 0$ such that

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

for all $\|x - x^*\| < \epsilon$.

Proof. Denote λ the smallest eigenvalue of $\nabla^2 f(x^*)$. Since $\nabla^2 f(x^*) \succ 0$, $\lambda > 0$. We have $d^T \nabla^2 f(x^*) d \geq \lambda \|d\|^2$ for all $d \in \mathbb{R}^n$. By the second order Taylor expansion

$$\begin{aligned} f(x^* + d) - f(x^*) &= \nabla f(x^*)^T d + \frac{1}{2} d^T \nabla^2 f(x^*) d + o(\|d\|^2) \\ &\geq \frac{\lambda}{2} \|d\|^2 + o(\|d\|^2) \\ &= \left(\frac{\lambda}{2} + \frac{o(\|d\|^2)}{\|d\|^2} \right) \|d\|^2 . \end{aligned}$$

Then choose $\epsilon > 0$ and $\gamma > 0$ such that for $\|d\| < \epsilon$,

$$\frac{\lambda}{2} + \frac{o(\|d\|^2)}{\|d\|^2} \geq \frac{\gamma}{2}.$$

Then the proof is complete. $\square$

References