Notes on Optimization

Liu Zhizhou

First Created: August 6, 2022
Last Modified: August 14, 2022

[Ber99] is brilliant textbook for optimization, which is also the main reference of this notes.

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 $\nabla^2 f = \nabla(\nabla f^\top)$.

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. $\nabla(Ax) = A^\top$;

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

3. $\nabla(f(Ax)) = A^\top \nabla f(Ax)$;

4. $\nabla^2(f(Ax)) = A^\top \nabla^2 f(Ax) A$;

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

$$f(x + d) = f(x) + d^\top \nabla f(x) + \frac{1}{2} d^\top \left(\int_0^1 \left(\int_0^\top \nabla^2 f(x + \tau d) \mathrm{d}\tau \right) \mathrm{d}t \right) d.$$

2. there exists

$$f(x + d) = f(x) + d^\top \nabla f(x) + \frac{1}{2} d^\top \nabla^2 f(x + \alpha d) d.$$

3. there holds

$$f(x + d) = f(x) + d^\top \nabla f(x) + \frac{1}{2} d^\top \nabla^2 f(x) d + o(\|d\|^2).$$

2 Convex Sets and Functions

2.1 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

$$f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha)f(y) \tag{2.1}$$

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

$$f(z) \geq f(x) + (z - x)^\top \nabla f(x) \tag{2.2}$$

    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

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

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

It can be shown that an equivalent definition is that

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

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

2.2   Convex Hull and Affine Hull

Definition 2.5 (convex combination, convex hull). A *convex combination* of elements of X is a vector of the form $\sum_{i=1}^m \alpha_i x_i$, where $x_i \in X$ and $\alpha_i \in \mathbb{R}$ such that $\alpha_i \geq 0$ for $i = 1, \ldots, m$ and $\sum_{i=1}^m \alpha_i = 1$.
The *convex hull* of X, denoted $\operatorname{conv} X$, is the set of all convex combinations of elements of X, i.e.

$$\operatorname{conv}(X) = \left\{ \sum_{i \in I} \alpha_i x_i : \alpha_i \geq 0, \sum_{i \in I} \alpha_i = 1, x_i \in X, I \subseteq \mathbb{N} \right\}. \tag{2.5}$$

Recall that a *linear manifold* (or linear affine) is a set of the form $x + S$, where S is a subspace.

Definition 2.6 (affine hull). If $S \subset \mathbb{R}^n$, the *affine hull* of S, denoted $\operatorname{aff}(S)$, is the intersection of all linear manifolds containing S.

Definition 2.7 (cone). A set $C \subset \mathbb{R}^n$ is said to be *cone* if $ax \in C$ for all $a \geq 0$ and $x \in C$. The *cone generated by* X, denoted $\mathrm{cone}(X)$, is the set of all nonnegative combinations of elements of X.

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

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

and similarly for second order,

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

Read rigorous proof to see the reason.

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

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

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,

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

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

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

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

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

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

   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

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

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

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

for all $x \in X$.

A geometric illustration of $\nabla f(x^*)^\top (x - x^*)$ 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 $d^\top \nabla^2 f(x^*) d \geq \lambda \|d\|^2$ for all $d \in \mathbb{R}^n$. By the second order Taylor

expansion

$$f(x^* + d) - f(x^*) = \nabla f(x^*)^\top d + \frac{1}{2} d^\top \nabla^2 f(x^*) d + o(\|d\|^2)$$

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

$$= \left(\frac{\lambda}{2} + \frac{o(\|d\|^2)}{\|d\|^2} \right) \|d\|^2 .$$

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. $\qquad\square$

4   Algorithms: Gradient Methods

Optimality conditions often provide the basis for the development and the analysis of the algorithms. The idea of the algorithms rely on an important idea, called *iterative descent*, i.e. $f(x^{k+1}) < f(x^k)$. Gradient methods, also called gradient descent methods, implement the idea of iterative descent. The iteration is

$$x^{k+1} = x^k + \alpha^k d^k, \tag{4.1}$$

where $k = 0, 1, \ldots$, $\nabla f(x^k)^\top d^k < 0$, $\alpha^k \in \mathbb{R}$ and $d^k \in \mathbb{R}^n$. There is a large variety of possibilities for choosing direction d^k and stepsize α^k.

4.1   Descent Direction

To make sure $\nabla f(x^k)^\top d^k < 0$, most gradient methods take the form $d^k = -D^k \nabla f(x^k)$, where D^k is a positive definite symmetric matrix. The iteration becomes

$$x^{k+1} = x^k - \alpha^k D^k \nabla f(x^k). \tag{4.2}$$

Different choice of D^k result in different methods, see Table 1.

4.2   Stepsize

There are a number of rules for choosing the stepsize α^k in a gradient method. We give some that are used widely in practice in Table 2.

4.3   Mathematical Statements for Convergence Results

There is a common line of proof for the convergence results. The main idea is that the cost function is improved at each iteration and the improvement is "substantial" near a non-stationary point, i.e. it is bounded away from zero. We then argue that

Name of Method	Choice of D^k
Steepest Descent	$D^k = I$
Newton's Methods	$D^k = \left(\nabla^2 f(x^k)\right)^{-1}$
Diagonally Scaled Steepest Descent	$D^k = \mathrm{diag}(d_1^k, \ldots, d_n^k)$
Modified Newton's Method	$D^k = \left(\nabla^2 f(x^0)\right)^{-1}$
Gauss Newton Method	$D^k = (\nabla g(x^k)\nabla g(x^k)^\top)^{-1}$

Table 1: Various choice of the positive definite matrix D^k, where $d^k = -D^k\nabla f(x^k)$. Gauss Newton Method is widely used when the cost function $f(x)$ is of the form $f(x) = \frac{1}{2}\|g(x)\|^2 = \frac{1}{2}\sum_{i=1}^m (g_i(x))^2$, where $g = (g_1, \ldots, g_m)$, which is a problem often encountered in statistical data analysis and in the context of neural network training.

Name of Method	Choice of α^k
Minimization Rule	$\alpha^k = \arg\min_{\alpha\geq 0} f(x^k + \alpha d^k)$
Limited Minimization Rule	$\alpha^k = \arg\min_{\alpha\in[0,s]} f(x^k + \alpha d^k)$
Armijo Rule	$\alpha^k = \beta^{m_k} s$
Constant Stepsize	$\alpha^k = s$
Diminishing Stepsize	$\alpha^k \to 0$, where $\sum_{k=0}^{\infty} \alpha^k = \infty$

Table 2: Various choice of stepsize α^k. In Arimijo rule, first choose fix scalars s, β and σ, with $0 < \beta < 1$ and $0 < \sigma < 1$, let m_k be the first non-negative integer m such that $f(x^k) - f(x^k + \beta^m s d^k) \geq -\sigma\beta^m s\nabla f(x^k)^\top d^k$. Note that here β^m means β to the m-th power. In diminishing method, we require $\sum_{k=0}^{\infty} \alpha^k = \infty$ to guarantees that $\{x^k\}$ does not converge to a non-stationary point. Indeed, if $x^k \to \bar{x}$, then for large m, n, $x^m \approx x^n \approx \bar{x}$, also $x^m \approx x^n - (\sum_{k=n}^{m-1} \alpha^k)\nabla f(\bar{x})$, which shows $\nabla f(\bar{x})$ must be zero.

the algorithm cannot approach a non-stationary point, since in this case the total cost improvement would accumulate to infinity.

I use "Proposition" instead of "Theorem" for the convergence results because in my opinion they are just some theoretic backgrounds (results) we should know in applications of the algorithms.

> **Definition 4.1 (gradient related).** We say that the direction $\{d^k\}$ is *gradient related* to $\{x^k\}$ if for any subsequence $\{x^k\}_{k \in \mathcal{K}}$ that converges to a non-stationary point, the corresponding subsequence $\{d^k\}_{k \in \mathcal{K}}$ is bounded and satisfies
>
> $$\limsup_{k \to \infty, k \in \mathcal{K}} \nabla f(x^k)^\top d^k < 0. \tag{4.3}$$

> **Proposition 4.2 (Stationary of Limit Points).** Let $\{x_k\}$ be a sequence generated by a gradient method $x^{k+1} = x^k + \alpha^k d^k$ and assume that $\{d^k\}$ is gradient related and α^k is chosen by the minimization rule, or the limited minimization rule, or the Armijo rule. Then every limit point of $\{x^k\}$ is a stationary point.

Proof. The key to prove this is noting that if $\{x^k\}$ converges to a non-stationary point, by the assumption of gradient related property of $\{d^k\}$, and by the definition of Armijo rule, we must have for some index $\bar{k} \geq 0$,

$$f(x^k) - f(x^k + (\alpha^k/\beta)d^k) < -\sigma(\alpha^k/\beta)\nabla f(x^k)^\top d^k,$$

for all $k \in \mathcal{K}$ and $k \geq \bar{k}$; in other words, if $\alpha^k \to 0$ there would exist a slight larger α^k/β that does not make "substantial" movement, which will result in contradiction.

For minimization rule, just note that

$$f(x^k) - f(x^{k+1}) \geq f(x^k) - f(\tilde{x}^{k+1}) \geq -\sigma(\tilde{\alpha})\nabla f(x^k)^\top d^k,$$

where $\{\tilde{x}\}$ is the sequence generated via Armijo rule and $\{\tilde{\alpha}\}$ is the corresponding stepsize. Then repeat the method to reach contradiction. $\qquad\square$

> **Lemma 4.3 (Descent Lemma).** Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable, and let x and y be two vectors in $\mathbb{R}^n$. Suppose that
>
> $$\|\nabla f(x + ty) - \nabla f(x)\| \leq Lt \|y\|$$
>
> for all $t \in [0, 1]$, where L is some scalar. Then
>
> $$f(x + ty) \leq f(x) + y^\top \nabla f(x) + \frac{L}{2} \|y\|^2 .$$

We can interpret this lemma as: if f is Lipschitz continuous then the variation of $f(x)$ is controlled by its slope plus $\frac{L}{2} \|y\|^2$. In fact, Lipschitz condition requires roughly that the "curvature" of f is no more than L at all points and in all directions.

> **Proposition 4.4 (Constant Stepsize).** Let $\{x^k\}$ be a sequence generated by a gradient method $x^{k+1} = x^k + \alpha^k d^k$, where $\{d^k\}$ is gradient related. Assume that f

is L-Lipschitz continuous and that for all k we have $d^k \neq 0$ and

$$\epsilon \leq \alpha^k \leq (2 - \epsilon)\bar{\alpha}^k, \tag{4.4}$$

where

$$\bar{\alpha}^k = \frac{\left|\nabla f(x^k)^\top d^k\right|}{L \left\|d^k\right\|^2},$$

and $\epsilon \in (0, 1]$ is a fixed scalar. Then every limit point of $\{x^k\}$ is a stationary point of f.

The intuition of the proof is that by descent lemma, $f(x^k) - f(x^k + \alpha^k d^k)$ is bounded by a quadratic overestimation.

Proof. By using descent lemma combined with the right hand side of Eq. (4.4), we have

$$f(x^k) - f(x^k + \alpha^k d^k) \geq \frac{1}{2}\epsilon^2 \left|\nabla f(x^k)^\top d^k\right|.$$

$\square$

In the case of steepest descent, the condition on stepsize becomes

$$\epsilon \leq \alpha^k \leq \frac{2 - \epsilon}{L}.$$

Proposition 4.5 (Diminishing Stepsize). Let $\{x^k\}$ be a sequence generated by a gradient method $x^{k+1} = x^k + \alpha^k d^k$. Assume that f is L-Lipschitz continuous and there exist positive scalars c_1, c_2 such that for all k we have

$$c_1 \left\|\nabla f(x^k)\right\|^2 \leq -\nabla f(x^k)^\top d^k, \quad \left\|d^k\right\|^2 \leq c_2 \left\|\nabla f(x^k)\right\|^2. \tag{4.5}$$

Suppose also that

$$\alpha^k \to 0, \quad \sum_{k=0}^{\infty} \alpha^k = \infty.$$

Then either $f(x^k) \to -\infty$ or else $\{f(x^k)\}$ converges to a finite value and $\nabla f(x^k) \to 0$. Furthermore, every limit point of $\{x^k\}$ is a stationary point of f.

Proof. Tedious. First use descent lemma to show $\liminf_{k \to \infty} \left\|\nabla f(x^k)\right\| = 0$. Then separate the sequence into $\left\|\nabla f(x^k)\right\| > \epsilon/3$ and $\left\|\nabla f(x^k)\right\| \leq \epsilon/3$ to reach contradiction. $\square$

Theorem 4.6 (Capture Theorem). Let f be continuously differentiable and let $\{x^k\}$ be a sequence satisfying $f(x^{k+1}) \leq f(x^k)$ for all k and generated by a gradient method $x^{k+1} = x^k + \alpha^k d^k$, which is convergent in the sense that every limit point of sequences that it generates is a stationary point of f. Assume that there exist scalars $s > 0$ and $c > 0$ such that for all k there holds

$$\alpha^k \leq s, \quad \left\|d^k\right\| \leq c \left\|\nabla f(x^k)\right\|.$$

Let x^* be a local minimum of f, which is the only stationary point of f within

some open set. Then there exists an open set S containing x^* such that if $x^{\bar{k}} \in S$ for some $\bar{k} \geq 0$, then $x^k \in S$ for all $k \geq \bar{k}$ and $\lim_{k \to \infty} x^k = x^*$. Furthermore, given any scalar $\bar{\epsilon} > 0$, the set S can be chosen so that $\|x - x^*\| < \bar{\epsilon}$ for all $x \in S$.

4.4 Rate of Convergence

Rate of convergence is evaluated using an *error function* $e : \mathbb{R}^n \to \mathbb{R}$ satisfying $e(x) \geq 0$ for all $x \in \mathbb{R}^n$ and $e(x^*) = 0$. Typical choices are $e(x) = \|x - x^*\|$ and $e(x) = |f(x) - f(x^*)|$.

> **Definition 4.7 (linear convergence).** We say that $\{e(x^k)\}$ converges *linearly* if there exist $q > 0$ and $\beta \in (0, 1)$ such that for all k, $e(x^k) \leq q\beta^k$.

It is possible to show that linear convergence is obtained if for some $\beta \in (0, 1)$ we have

$$\limsup_{k \to \infty} \frac{e(x^{k+1})}{e^k} \leq \beta.$$

> **Definition 4.8 (superlinear convergence).** If for every $\beta \in (0, 1)$, there exists q such that the condition $e(x^k) \leq q\beta^k$ holds for all k, we say that $\{e(x^k)\}$ converges *superlinearly*.

This is true in particular, if

$$\lim_{k \to \infty} \frac{e(x^{k+1})}{e(x^k)} = 0.$$

Consider

$$f(x) = \frac{1}{2} x^\top Q x,$$

where Q is positive definite and symmetric. Let m, M be the smallest and biggest eigenvalue of Q, respectively.

> **Proposition 4.9.** Suppose we use the method of the steepest descent
>
> $$x^{k+1} = x^k - \alpha^k \nabla f(x^k),$$
>
> where the stepsize α^k is chosen to be constant, i.e. $\alpha^k \triangleq \alpha$. Then
>
> $$\frac{\|x^{k+1}\|}{\|x^k\|} \leq \frac{M - m}{M + m}.$$
>
> If the stepsize α^k is chosen according to the minimization rule
>
> $$\alpha^k = \arg\min_{\alpha \geq 0} f(x^k - \alpha \nabla f(x^k)).$$
>
> Then, for all k,
>
> $$f(x^{k+1}) \leq \left(\frac{M - m}{M + m} \right)^2 f(x^k).$$

References

[Ber99] D.P. Bertsekas. *Nonlinear Programming.* Athena Scientific, 1999.