Notes on Optimization

Liu Zhizhou

First Created: August 6, 2022
Last Modified: September 5, 2022

[Ber99] is brilliant textbook for optimization, which is also the main reference of this notes. The structure of the this notes is 100% from the lectures by Zhang Jin, in SUSTech.

Contents

Chapter 1

Preliminaries

1.1 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 1.1.1 (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. $\nabla(x^\top Ax) = (A + A^\top)x$; in particular, if Q is symmetric, then $\nabla(x^\top Qx) = 2Qx$ 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 1.1.2 (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

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

Chapter 2

Fundamental Concepts

2.1 Convexity

2.1.1 normal convexity

Definition 2.1.1 (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

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y) \tag{2.1.1}$$

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

An optimization problem is convex if both the objective function and feasible set are convex.

Definition 2.1.2 (strictly convex). The function f is called *strictly convex* if Eq.(2.1.1) is strict for all $x \neq y$ and $\alpha \in (0,1)$.

Proposition 2.1.3 (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

$$f(z) \geq f(x) + (z-x)^\top \nabla f(x) \tag{2.1.2}$$

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 2.1.4 (epigraph). Assume C is a convex subset of $\mathbb{R}^n$. The *epigraph* of $f : C \to [-\infty, \infty]$ is the subset of $\mathbb{R}^{n+1}$ given by

$$\text{epi}(f) = \{(x, w) : x \in C, w \in \mathbb{R}, f(x) \leq w\}.$$

By definition, it is easy to see that if f is a convex function, then epi(f) is a

convex set.

2.1.2 strong convexity

Definition 2.1.5 (σ-strongly convex). A function $f : \mathbb{R}^n \to \mathbb{R}$ is called *σ-strongly convex* if for some $\sigma > 0$, we have

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{\sigma}{2} \|x - y\|^2 \tag{2.1.3}$$

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

It can be shown that an equivalent definition is that

$$(\nabla f(x) - \nabla f(y))^\top (x - y) \geq \sigma \|x - y\|^2 \tag{2.1.4}$$

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

2.2 Convex Hull and Affine Hull

Definition 2.2.1 (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.

$$\operatorname{conv}(X) = \left\{ \sum_{i \in I} \alpha_i x_i : \alpha_i \geq 0, \sum_{i \in I} \alpha_i = 1, x_i \in X, I \subseteq \mathbb{N} \right\}. \tag{2.2.1}$$

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

Definition 2.2.2 (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 2.2.3 (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 $\operatorname{cone}(X)$, is the set of all nonnegative combinations of elements of X.

2.3 Fundamental Terminologies

Consider the problem

$$\begin{aligned} &\min f(x) \\ &\text{subject to } x \in \Omega. \end{aligned} \tag{2.3.1}$$

Then Ω is called the *feasible set*. If $\Omega = \mathbb{R}^n$, then the problem is called *unconstrained*; otherwise, the problem is called *constrained*.

Definition 2.3.1 (minimizer). x^* is a *local minimizer* if there is $\delta > 0$ such that $f(x) \geq f(x^*)$ for all $x \in \Omega$ and $\|x - x^*\| < \delta$. x^* is a *global minimizer* if

$f(x) \geq f(x^*)$ for all $x \in \Omega$. If $f(x) > f(x^*)$ for all $x \in \Omega - \{x^*\}$, then it is called the corresponding strict minimizer (strict local minimizer and strict global minimizer).

Theorem 2.3.2. Any local minimizer of a convex optimization problem is a global minimizer.

Proof. Assume that x^* is a local minimizer but not a global one. Use the convexity to arrive at a contradiction. □

Definition 2.3.3 (feasible direction). A vector $d \in \mathbb{R}^n$ is a feasible direction at $x \in \Omega$ if $d \neq 0$ and $x + \alpha d \in \Omega$ for some small $\alpha > 0$.

Note that $x + d$ is not necessarily in Ω.

2.4 General Optimality Condition

Theorem 2.4.1 (First Order Necessary Condition). Let Ω be a subset of $\mathbb{R}^n$ and $f \in C^1$ a real-valued function on Ω. If x^* is a local minimizer of f over Ω, then for any feasible direction d at x^*, we have

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

Proof. Let d be any feasible direction. Consider first order Taylor expansion at $\alpha = 0$:

$$f(x^* + \alpha d) = f(x^*) + \alpha d^\top \nabla f(x^*) + o(\alpha).$$

Since x^* is a local minimizer, then $\alpha d^\top \nabla f(x^*) + o(\alpha) \geq 0$, then $d^\top \nabla f(x^*) \geq 0$ if divided by α and let $\alpha \to 0$. Note that we can only consider $\alpha > 0$. □

The geometric interpretation of this result is that: if x^* is a local minimizer, then every feasible direction has a less than right angle with the direction that increase the function for the most, i.e. every feasible direction makes it increase.

Corollary 2.4.2 (Interior Case). Let Ω be a subset of $\mathbb{R}^n$ and $f \in C^1$ a real-valued function on Ω. If x^* is a local minimizer of f over Ω and a interior point, then

$$\nabla f(x^*) = 0.$$

Proof. Set $d = -\nabla f(x^*)$, i.e. the direction that decrease the most. □

There are two cases in first order necessary condition: $d^\top \nabla f(x^*) > 0$ for all feasible direction d; or $d^\top \nabla f(x^*) = 0$ for some feasible direction d. In the first case, we must have x^* a local minimizer; while in the second case, the following theorem provides a higher order derivatives check.

Theorem 2.4.3 (Second Order Necessary Condition (SONC)). Let $\Omega \subset \mathbb{R}^n$, $f \in C^2$ a function on Ω. Let x^* be a local minimizer of f over Ω, d a feasible

direction at x^*. If $d^\top \nabla f(x^*) = 0$, then

$$d^\top \nabla^2 f(x^*) d \geq 0$$

Proof. The proof is similar to the first order necessary condition. □

Corollary 2.4.4 (Interior Case). Let x^* be a interior point of $\Omega \subset \mathbb{R}^n$. If x^* is a local minimizer of $f : \Omega \to \mathbb{R}^n$, $f \in C^2$, then $\nabla f(x^*) = 0$ and $\nabla^2 f(x)$ is positive semidefinite, i.e. $\nabla^2 f(x) \succeq 0$.

Theorem 2.4.5 (Second Order Sufficient Condition (SOSC), Interior Case). Let $\Omega \subset \mathbb{R}^n$, $f \in C^2$ a function on Ω. Suppose that $\nabla f(x^*) = 0$, $\nabla^2 f(x^*) \succ 0$ and x^* is an interior point. Then x^* is a strict local minimizer of f.

Note that the condition is not necessary for strict local minimizer.

Proof. We have

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

Show that $\alpha^2/2 \cdot d^\top \nabla^2 f(x^*) d + o(\alpha^2) > 0$. This is true otherwise divided by α^2 we would have $d^\top \nabla^2 f(x^*) d \leq 0$, contradicts with the assumption that $\nabla^2 f(x^*) \succ 0$. □

Chapter 3

Unconstrained Optimization Algorithms

3.1 Gradient Descent Method

The *level set* of f at c, denoted $\mathcal{L}_f(c)$, is $\{x : f(x) = c\}$.

Lemma 3.1.1. The vector $\nabla f(x_0)$ is orthogonal to the tangent vector to an arbitrary smooth curve passing through x_0 on the level set $\mathcal{L}_f(f(x_0))$.

Proof. Suppose the curve γ is parameterized y $g : \mathbb{R} \to \mathbb{R}^n$ and $g(t_0) = x_0$. Let $\nabla g(t_0)^\top = v \neq 0$, so v is the tangent vector to γ at x_0. We should show $v^\top \nabla f(x_0) = 0$. Let $h(t) = f(g(t))$. Then $h(t)$ is constant since $g(t)$ lies on the level set of f. Then

$$0 = \frac{\mathrm{d}h(t)}{\mathrm{d}t} = \nabla g(t) \nabla f(g(t)).$$

So that $v^\top \nabla f(x_0) = 0$. □

The gradient descent algorithm is the following: given initial x^0. Generate x^{k+1} from x^k by

$$x^{k+1} = x^k - \alpha^k \nabla f(x^k),$$

where α_k is called the step size.

The algorithm works: first order Taylor expansion of $f(x + \alpha d)$ at $\alpha = 0$ is

$$f(x + \alpha d) = f(x) + \alpha d^\top \nabla f(x) + o(\alpha).$$

Let $d = -\nabla f(x)$, then

$$f(x - \alpha \nabla f(x)) - f(x) = -\alpha \left\| \nabla f(x) \right\|^2 + o(\alpha).$$

Then for sufficiently small α, $f(x - \alpha \nabla f(x)) < f(x)$.

Another interpretation of the algorithm is considered it as a consequence of optimization problem. Note that

$$x^{k+1} = \arg\min_x \frac{1}{2\alpha^k} \left\| x - x^k + \alpha^k \nabla f(x^k) \right\|^2,$$

and

$$\frac{1}{2\alpha^k}\left\|x - x^k + \alpha^k \nabla f(x^k)\right\|^2 = \frac{\alpha^k}{2}\left\|\nabla f(x^k)\right\| + \left\langle \nabla f(x^k), x - x^k \right\rangle + \frac{1}{2\alpha^k}\left\|x - x^k\right\|^2.$$

So that x^{k+1} is obtained by minimizing the linearization of f at x^k and a proximal term that keeps x^{k+1} close to x^k.

3.2 The Method of Steepest Descent

The method of steepest descent is choosing $\alpha^k = \arg\min_{\alpha \geq 0} f(x^k - \alpha \nabla f(x^k))$. It is used mostly for quadratic programs:

$$\min f(x) = x^\top Q x - b^\top x, \tag{3.2.1}$$

where Q is assumed to be symmetric. Otherwise it is not worth the effort to solve the sub-problem exactly.

Proposition 3.2.1. If (x^k) is a steepest descent sequence for a given function $f : \mathbb{R}^n \to \mathbb{R}$, then for each k, $x^{k+1} - x^k$ is orthogonal to the vector $x^{k+2} - x^{k+1}$.

Proof. It suffices to show $\left\langle \nabla f(x^k), \nabla f(x^{k+1}) \right\rangle = 0$. Let $\phi_k(\alpha) = f(x^k - \alpha \nabla f(x^k))$. Then $\phi_k'(\alpha^k) = 0$. Note that

$$\frac{\mathrm{d}\phi_k}{\mathrm{d}\alpha}(\alpha^k) = -\nabla f(x^k)^\top \nabla f(x^k - \alpha^k \nabla f(x^k)) = \left\langle \nabla f(x^k), \nabla f(x^{k+1}) \right\rangle.$$

□

For $f(x) = x^\top Q x - b^\top x$, $\nabla f(x) = Qx - b$. Then $\alpha^k = \arg\min_{\alpha \geq 0} f(x^k - \alpha \nabla f(x^k))$. The solution to this using first order necessary condition is

$$\alpha^k = \frac{\nabla f(x^k)^\top \nabla f(x^k)}{\nabla f(x^k)^\top Q \nabla f(x^k)}.$$

3.3 Rate of Convergence

Definition 3.3.1 (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 3.3.2 (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

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

In summary, suppose

$$\lim_{k\to\infty} \frac{e^{k+1}}{e^k} = \mu.$$

1. if $\mu = 1$, then (x^k) converges sublinearly;
2. if $\mu \in (0, 1)$, then (x^k) converges linearly;
3. if $\mu = 0$, then (x^k) converges superlinearly.

Also, to distinguish superlinear rates of convergence, we check that

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

1. if $q = 2$, it is quadratic convergence;
2. if $q = 3$, it is cubic convergence;
3. q can be non-integer.

3.4 Convergence Rate of Steepest Descent

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 $e(x) = |f(x) - f(x^*)|$.

First that's see a special case: if $e^k = x^k - x^*$, where x^* is the local minimizer. Note that $Qe^k = Qx^k - Qx^* = Qx^k - b = \nabla f(x^k)$. The special case is that e^k is the eigenvalue of Q, then $Qe^k = \lambda e^k$. Then

$$e^{k+1} = e^k - \frac{\nabla f(x^k)^\top \nabla f(x^k)}{\nabla f(x^k)^\top Q \nabla f(x^k)} Qe^k = 0.$$

So $x^{k+1} = x^*$.

For the general case, we would work with

$$V(x) = f(x) + \frac{1}{2} x^{*\top} Q x^* = \frac{1}{2}(x - x^*)^\top Q (x - x^*). \tag{3.4.1}$$

The solution point x^* is obtained by solving $Qx = b$, $x^* = Q^{-1}b$. We now investigate the convergence result of minimizing $V(x)$.

Investigate

$$\frac{V(x^k) - V(x^{k+1})}{V(x^k)}.$$

Plug in $V(x)$. Note that formally $x^\top Q y = \langle x, y \rangle_Q$ which I mean is bilinear. Then

$$\begin{aligned}
&\left\langle x^{k+1} - x^*, x^{k+1} - x^* \right\rangle_Q = \left\langle x^k - \alpha^k \nabla f(x^k) - x^*, x^k - \alpha^k \nabla f(x^k) - x^* \right\rangle_Q \\
&= \left\langle x^{k+1} - x^*, x^{k+1} - x^* \right\rangle_Q - 2\alpha^k \left\langle x^k - x^*, \nabla f(x^k) \right\rangle_Q + (\alpha^k)^2 \left\langle \nabla f(x^k), \nabla f(x^k) \right\rangle_Q .
\end{aligned}$$

So

$$\begin{aligned}
&V(x^k) - V(x^{k+1}) = 2\alpha^k \left\langle x^k - x^*, \nabla f(x^k) \right\rangle_Q - (\alpha^k)^2 \left\langle \nabla f(x^k), \nabla f(x^k) \right\rangle_Q \\
&= 2\alpha^k (x^k - x^*)^\top Q \nabla f(x^k) - (\alpha^k)^2 \nabla f(x^k)^\top Q \nabla f(x^k) \\
&= 2\alpha^k (e^k)^\top Q \nabla f(x^k) - (\alpha^k)^2 \nabla f(x^k)^\top Q \nabla f(x^k).
\end{aligned}$$

Plug in

$$\alpha^k = \frac{\nabla f(x^k)^\top \nabla f(x^k)}{\nabla f(x^k)^\top Q \nabla f(x^k)},$$

and $e^k = Q^{-1}\nabla f(x^k)$, we would have

$$\frac{V(x^k) - V(x^{k+1})}{V(x^k)} = \frac{(\nabla f(x^k)^\top \nabla f(x^k))^2}{[\nabla f(x^k)^\top Q \nabla f(x^k)][\nabla f(x^k)^\top Q^{-1} \nabla f(x^k)]}$$

Therefore,

$$V(x^{k+1}) = \left\{ 1 - \frac{(\nabla f(x^k)^\top \nabla f(x^k))^2}{[\nabla f(x^k)^\top Q \nabla f(x^k)][\nabla f(x^k)^\top Q^{-1} \nabla f(x^k)]} \right\} V(x^k).$$

Lemma 3.4.1 (Kantorovich Inequality). Assume $Q \succ 0$. For any $x \in \mathbb{R}^n$,

$$\frac{(x^\top x)^2}{(x^\top Q x)(x^\top Q^{-1} x)} \geq \frac{4\lambda_{\min}\lambda_{\max}}{(\lambda_{\min} + \lambda_{\max})^2},$$

where $\lambda_{\min}, \lambda_{\max}$ is the smallest and biggest eigenvalue respectively.

Using this lemma,

$$\begin{aligned}
V(x^{k+1}) &\leq \left\{ 1 - \frac{4\lambda_{\min}\lambda_{\max}}{(\lambda_{\min} + \lambda_{\max})^2} \right\} V(x^k) \\
&= \left(\frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}} \right)^2 V(x^k).
\end{aligned}$$

Let $\|e\|_Q = \sqrt{e^\top Q e}$ and $\kappa = \lambda_{\max}/\lambda_{\min}$, then

$$\left\|e^k\right\|_Q \leq \left(\frac{\kappa - 1}{\kappa + 1} \right)^k \left\|e^0\right\|_Q. \tag{3.4.2}$$

Therefore, the convergence rate is linear.

3.5 Convergence Rate of Fixed Stepsize Gradient Descent

Observe that

$$\begin{aligned}
&\left\|x^{k+1} - x^*\right\|^2 = \left\|x^k - x^* - \alpha \nabla f(x^k)\right\| \\
&= \left\|x^k - x^*\right\|^2 - 2\alpha \left\langle \nabla f(x^k), x^k - x^* \right\rangle + \alpha^2 \left\|\nabla f(x^k)\right\|^2.
\end{aligned}$$

In order to have $\left\|x^{k+1}-x^*\right\| \leq \left\|x^k-x^*\right\|$, we must have

$$\frac{\alpha}{2}\left\|\nabla f(x^k)\right\|^2 \leq \left\langle \nabla f(x^k), x^k-x^*\right\rangle,$$

which is equivalent to

$$\frac{\alpha}{2}\left\|\nabla f(x^k)-\nabla f(x^*)\right\|^2 \leq \left\langle \nabla f(x^k)-\nabla f(x^*), x^k-x^*\right\rangle. \tag{3.5.1}$$

3.5.1 when f is convex

Theorem 3.5.1 (Baillon-Haddad Theorem). If $f \in C^1$ is a convex function, then L-Lipschitz differentiable if and only if

$$\|\nabla f(x)-\nabla f(y)\|^2 \leq L\left\langle \nabla f(x)-\nabla f(y), x-y\right\rangle,$$

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

Proposition 3.5.2. Let $f \in C^1$ be a convex function and L-Lipschitz differentiable. If $0 < \alpha \leq 2/L$, then

$$\frac{\alpha}{2}\left\|\nabla f(x^k)-\nabla f(x^*)\right\|^2 \leq \left\langle \nabla f(x^k)-\nabla f(x^*), x^k-x^*\right\rangle.$$

Consequently, $\left\|x^{k+1}-x^*\right\| \leq \left\|x^k-x^*\right\|$ for $k \in \mathbb{N}$.

Proof. Directly from Theorem 3.5.1 and observations above. □

This proposition tells us the appropriate stepsize, which is $\alpha \in [0, 2/L]$ when f is convex and L-Lipschitz differentiable.

Lemma 3.5.3 (Fundamental Gradient Inequality). Assume f is convex. If there exists $L > 0$ for any $y \in \mathbb{R}^n$,

$$f(T_L(y)) \leq f(y) + \langle \nabla f(y), T_L(y)-y\rangle + \frac{L}{2}\|T_L(y)-y\|^2,$$

it holds that

$$f(x)-f(T_L(y)) \geq \frac{L}{2}\|x-T_L(y)\|^2 - \frac{L}{2}\|x-y\|^2 + L_f(x,y), \tag{3.5.2}$$

where $T_L(y) = y - \nabla f(y)/L$ and $L_f(x,y) = f(x)-f(y)-\langle \nabla f(y), x-y\rangle$.

Proof. Consider

$$\phi(u) = f(y) + \langle \nabla f(y), u-y\rangle + \frac{L}{2}\|u-y\|^2.$$

Set $\nabla\phi(u) = 0$, then we find that $T_L(y) = \arg\min_u \phi(u)$. Then

$$\begin{aligned}
&\phi(x) - \phi(T_L(y)) \\
&= \langle \nabla f(y), x - y\rangle + \frac{L}{2}\|x-y\|^2 - \left\langle \nabla f(y), -\frac{\nabla f(y)}{L}\right\rangle - \frac{L}{2}\left\|-\frac{\nabla f(y)}{L}\right\|^2 \\
&= \langle \nabla f(y), x - y\rangle + \frac{L}{2}\|x-y\|^2 + \frac{\|\nabla f(y)\|^2}{2L} \\
&= \frac{L}{2}\|x - T_L(y)\|^2 .
\end{aligned}$$

Note that the condition implies $\phi(T_L(y)) \geq f(T_L(y))$. Thus,

$$\phi(x) - f(T_L(y)) \geq \phi(x) - \phi(T_L(y)) = \frac{L}{2}\|x - T_L(y)\|^2 .$$

The result follows as plugging in $\phi(x)$. □

Corollary 3.5.4. Assume f is convex and L-Lipschitz differentiable, then

$$f(x) - f(T_L(y)) \geq \frac{L}{2}\|x - T_L(y)\|^2 - \frac{L}{2}\|x-y\|^2 + L_f(x,y),$$

where $T_L(y) = y - \nabla f(y)/L$ and $L_f(x,y) = f(x) - f(y) - \langle \nabla f(y), x - y\rangle$.

Proof. Use Descent Lemma 3.5.5. □

Let $x = x^*$, $y = x^k$ in Eq. (3.5.2), then

$$\begin{aligned}
f(x^*) - f(x^{k+1}) &\geq \frac{L}{2}\left\|x^* - x^{k+1}\right\|^2 - \frac{L}{2}\left\|x^* - x^k\right\|^2 + L_f(x^*, x^k) \\
&\geq \frac{L}{2}(\left\|x^* - x^{k+1}\right\|^2 - \left\|x^* - x^k\right\|^2),
\end{aligned} \tag{3.5.3}$$

where we use $L_f(x^*, x^k) \geq 0$ since f is convex. Now summing together,

$$\sum_{i=0}^{k}(f(x^{i+1} - f(x^*))) \leq \frac{L}{2}(\left\|x^* - x^0\right\|^2 - \left\|x^* - x^k\right\|^2) \leq \frac{L}{2}\left\|x^* - x^0\right\|^2 .$$

Since actually we choose $\alpha = 1/L \leq 2/L$, $f(x^n)$ is decreasing, then

$$f(x^k) - f(x^*) \leq \frac{L}{2k}\left\|x^k - x^0\right\|^2 .$$

Therefore, the convergence rate of constant stepsize gradient method is $O(1/k)$ when f is convex and Lipschitz differentiable.

3.5.2 when f is not necessarily convex

Lemma 3.5.5 (Descent Lemma). Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable, and let x and d be two vectors in $\mathbb{R}^n$. Suppose that

$$\|\nabla f(x + \alpha d) - \nabla f(x)\| \leq L\alpha\|d\|$$

for all $\alpha \in [0,1]$, where L is some scalar. Then

$$f(x+d) \leq f(x) + d^\top \nabla f(x) + \frac{L}{2} \|d\|^2 .$$

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

Lemma 3.5.6 (sufficient decrease). Suppose f is L_0-Lipschitz differentiable. Let $L > L_0/2$ and $\alpha = 1/L$ be the stepsize. Then

$$f(x) - f(T_L(x)) \geq \frac{L - L_0/2}{L^2} \|\nabla f(x)\|^2 ,$$

where $T_L(x) = x - \nabla f(x)/L$.

Proof. By descent lemma,

$$\begin{aligned} f(T_L(x)) &\leq f(x) - \alpha \nabla f(x)^\top \nabla f(x) + \frac{L_0 \alpha^2}{2} \|\nabla f(x)\|^2 \\ &= f(x) - \frac{1}{L} \|\nabla f(x)\|^2 + \frac{L_0/2}{L^2} \|\nabla f(x)\|^2 . \end{aligned}$$

The result follows. □

Therefore, if we take $\alpha = 1/L$, where $L > L_0/2$, then $f(x^{k+1}) < f(x^k)$ if and only if x^k is not a stationary point, i.e. $\nabla f(x^k) \neq 0$. Since we assume the minimizer exists f is bounded below, so that $\lim_{n\to\infty} f(x^n)$ exists. Thus $(f(x_n))$ is Cauchy so that $f(x^k) - f(x^{k+1}) \to 0$, which implies $\nabla f(x^k) \to 0$.

Let

$$M = \frac{L - L_0/2}{L^2}.$$

Using $f(x^k) - f(x^{k+1}) \geq M \left\|\nabla f(x^k)\right\|^2$, we could obtain

$$f(x^0) - f(x^{k+1}) \geq M \sum_{i=0}^{k} \left\|\nabla f(x^k)\right\|^2 \geq M(k+1) \min_{i=0,\ldots,k} \left\|\nabla f(x^i)\right\|^2 .$$

Therefore,

$$\min_{i=0,\ldots,k} \left\|\nabla f(x^i)\right\| \leq \frac{\sqrt{f(x^0) - f(x^*)}}{\sqrt{M(k+1)}},$$

where we used $f(x^{k+1}) \geq f(x^*)$.

Proposition 3.5.7. Assume the objective function f is Lipschitz differentiable. Suppose (x^k) is the sequence generated by constant stepsize gradient descent method. All limiting points of (x^k) are stationary points.

Proof. Observe that

$$\|\nabla f(\bar{x})\| \leq \left\|\nabla f(\bar{x}) - \nabla f(x^{k_j})\right\| + \left\|\nabla f(x^{k_j})\right\| .$$

The first term converges to zero by the Lipschitz differentiable condition; the

second term is illustrated in above paragraphs. □

3.6 Convergence of (x^k) Generated by Gradient Descent

Definition 3.6.1 (Fejér monotone). A sequence $(x^k) \in \mathbb{R}^n$ is called *Fejér monotone* with respect to a set $S \subseteq \mathbb{R}^n$ if

$$\|x^{k+1} - y\| \leq \|x^k - y\|$$

for every $k \geq 0$ and $y \in S$.

Lemma 3.6.2. Let $(x^k) \in \mathbb{R}^n$. Let S be a set satisfying $D \subset S$, where D is the set consists of all the limit points of (x^k). If (x^k) is Fejér monotone with respect to S, then it converges to a point in D.

This lemma tells us that if (x^k) is Fejér monotone, then it can only has one limit, i.e. it converges.

Proof. Since (x^k) is Fejér monotone with respect to S, then x^k is bounded. Then every limit point of (x^k) belongs to $D \subseteq S$. For $\bar{x} \in D$, we have

$$\|x^{k+1} - \bar{x}\| \leq \|x^k - \bar{x}\|$$

for all $k \in \mathbb{N}$. The sequence $(\|x^k - \bar{x}\|)$ is bounded below and decreasing thus convergent. Note that $\bar{x}$ is a limit point of (x^k), we must have $x^k \to \bar{x}$. □

Proposition 3.6.3. Suppose $f : \mathbb{R}^n \to R$ is L-Lipschitz differentiable and convex. (x^k) is a sequence generated by constant stepsize, $x^{k+1} = T_L(x^k)$, gradient descent. Then for any $x^* \in X$, where X is the solution set,

1. $\|x^{k+1} - x^*\| \leq \|x^k - x^*\|$;
2. (x^k) converges to an optimal solution.

Proof. Substitute $x = x^*$, $y = x^k$ into the fundamental inequality, we get Eq. (3.5.3). Since the left hand side is less than zero, we obtain item 1.
Then (x^k) is Fejér monotone with respect to X, thus converges to a point in X. □

3.7 The Method of Backtracking Line Search

3.7.1 when f is not necessarily convex

Assume f is L-Lipschitz differentiable.

Backtracking line search procedure (also known as Armijo Rule) is a method of choosing the stepsize $\alpha_k = 1/L_k$. There are three parameters in this method, (s, γ, η), where $s, \gamma \in (0, 1)$ and $\eta > 1$. The choice of L_k is:

1. $L_k = s$ initially, so s is the initial guess;

2. while
$$f(x^k) - f(T_{L_k}(x^k)) < \frac{\gamma}{L_k} \left\|\nabla f(x^k)\right\|^2,$$
we set multiply L_k by η, where $\eta > 1$. In other words, L_k is finally chosen as $s\eta^{i_k}$, where i_k is the smallest non-negative integer for which
$$f(x^k) - f(T_{s\eta^{i_k}}(x^k)) \geq \frac{\gamma}{s\eta^{i_k}} \left\|\nabla f(x^k)\right\|^2 \tag{3.7.1}$$
is satisfied.

Note that such i_k must exist, because we have Lemma 3.5.6 that
$$f(x^k) - f(T_{L_k}(x^k)) \geq \frac{L_k - L}{L_k^2} \left\|\nabla f(x^k)\right\|^2.$$
If $L_k \geq L/[2(1-\gamma)]$, then $(L_k - L/2)/L \geq \gamma$ and then Eq. (3.7.1) is satisfied. Therefore, either $L_k = s$ or the backtracking is invoked so that $L_k/\eta < L/[2(1-\gamma)]$. We can summarize this observation as
$$L_k \leq \max\{s, \frac{\eta L}{2(1-\gamma)}\}.$$
Then
$$f(x^k) - f(x^{k+1}) \geq \frac{\gamma}{\max\{s, \frac{\eta L}{2(1-\gamma)}\}} \left\|\nabla f(x^k)\right\|^2.$$

3.7.2 when f is convex

In convex case, we have two parameters (s, η), where $s > 0, \eta > 1$. Then the stepsize L_k is defined as

1. L_k is set to be L_{k-1}, where $L_{-1} = s$. (Compared with the case that f is not necessary convex, we can inherited the L_k when iteration.)
2. Choose L_k to be $L_{k-1}\eta^{i_k}$, where i_k is the smallest non-negative integer for which
$$f(T_L(x^k)) \leq f(x^k) + \left\langle \nabla f(x^k), T_{L_{k-1}\eta^{i_k}}(x^k) - x^k \right\rangle + \frac{L_k}{2} \left\| T_{L_{k-1}\eta^{i_k}}(x^k) - x^k \right\|^2.$$

Recall that when the stepsize is $1/L$, where L is the Lipschitz constant of f, then the inequality is satisfied. And L_k/η makes the inequality not satisfied. So we must have $L_k/\eta < L$, $L_k < \eta L$. Then
$$s \leq L_k \leq \max\{s, \eta L\}.$$

Proposition 3.7.1. Assume (x^k) is the sequence generated by either constant stepsize or backtracking procedure. Then

1. $(f(x^k))$ is non-increasing.
2. $(\left\|x^k - x^*\right\|)$ is non-increasing for any $x^* \in X$.

3. $f(x^k) - f(x^*) \leq \frac{\alpha L \left\|x^0 - x^*\right\|^2}{2k}$ for any $k \in \mathbb{N}$ and $x^* \in X$.

4. $\{x^k\}$ converges to some optimal solution as $k \to \infty$.

5. for any $k \in \mathbb{N}$ and $x^* \in X$, we have

$$\min_{i=0,\ldots,k} \left\|\nabla f(x^i)\right\| \leq \frac{2\alpha^{1.5} L \left\|x^0 - x^*\right\|}{\sqrt{\beta} k}.$$

3.7.3 when f is strongly convex

Proposition 3.7.2. Suppose f is L-Lipschitz differentiable and σ-strongly convex. (x^k) is generated by either constant stepsize or backtracking procedure. Let

$$\alpha = \begin{cases} 1, & \text{constant stepsize} \\ \max\{\eta, s/L\}, & \text{backtracking procedure.} \end{cases}$$

Then

1. $\left\|x^{k+1} - x^*\right\|^2 \leq (1 - \frac{\sigma}{\alpha L}) \left\|x^k - x^*\right\|^2$;

2. $\left\|x^{k+1} - x^*\right\|^2 \leq (1 - \frac{\sigma}{\alpha L})^k \left\|x^0 - x^*\right\|^2$;

3. $f(x^k) - f(x^*) \leq \frac{\alpha L}{2}(1 - \frac{\sigma}{\alpha L})^k \cdot \left\|x^0 - x^*\right\|^2$.

Proof. Use Eq. (3.5.3) and f is σ-strongly convex. □

3.8 Newton's Method

Assume $f \in C^2$.

The idea of Newton's Method is to use local quadratic function to get $f(x^{k+1}) < f(x^k)$. Using Taylor's expansion, the approximation quadratic function of $f(x)$ at x^k is

$$q(x) = f(x^k) + \nabla f(x^k)^\top (x - x^k) + \frac{1}{2}(x - x^k)^\top \nabla^2 f(x^k)(x - x^k).$$

The minimizer of $q(x)$ is

$$x^{k+1} = x^k - (\nabla^2 f(x^k))^{-1} \nabla f(x^k)$$

by setting $\nabla q(x) = 0$.

However, note that even if $\nabla^2 f(x^k) \succ 0$, objective descent is not guaranteed. Instead, we have the following result.

Lemma 3.8.1. If the Hessian $\nabla^2 f(x^k) \succ 0$ and $\nabla f(x^k) \neq 0$, then the search direction

$$d^k = -(\nabla^2 f(x^k))^{-1} \nabla f(x^k)$$

is a descent direction, that is, there exists $\overline{\alpha} > 0$ such that

$$f(x^k_\alpha d^k) < f(x^k)$$

for all $\alpha \in (0, \overline{\alpha})$.

The appearance of searching direction would result in more general descent algorithms.

Example 3.8.2 (quadratic function minimization). The objective function is

$$f(x) = \frac{1}{2}x^\top Qx - b^\top x.$$

Assume that Q is symmetric and invertible. Then $\nabla f(x) = Qx - b$, $\nabla^2 f(x) = Q$. Then solve $\nabla f(x) = 0$, $x^* = Q^{-1}b$. By Newton's Method,

$$x^1 = x^0 - (\nabla^2 f(x^0))^{-1}\nabla f(x^0) = x^0 - Q^{-1}(Qx^0 - b) = Q^{-1}b = x^*.$$

The solution is obtained in one step.

There are some issues in Newton's method:

1. when the dimension n is large, obtain $\nabla^2 f(x^k)$ can be computationally expensive;
2. when Hessian is not positive definite, the direction is not necessarily descending.

3.9 Convergence Rate of Newton's Method

Let $e^k = x^k - x^*$.

Proposition 3.9.1. Suppose $f \in C^2$, F is Lipschitz continuous near x^* and $\nabla f(x^*) = 0$. If x^k is sufficiently close to x^* and $\nabla^2 f(x) \succ 0$, then there exists $C > 0$ such that

$$\left\|e^{j+1}\right\| \leq C \left\|e^j\right\|^2,$$

for $j = k, k+1, \ldots$.

3.10 Modified Newton's Method

In this section, we introduce several modified Newton's methods. There are Greenstadt Method, Levenberg-Marquardt Method and modified Choleky/ Gill-Murray Method.

3.11 Gauss-Newton Method

The Gauss-Newton Method solves a particular class of problems: the non-linear least squares: given functions $r_i : \mathbb{R}^n \to \mathbb{R}$, $i = 1, \ldots, m$,

$$\min_x \frac{1}{2}\sum_{i=1}^m (r_i(x))^2. \tag{3.11.1}$$

If we define $r = (r_1, \ldots, r_m)^\top$. Then the problem becomes

$$\min_x f(x) \triangleq \frac{1}{2} r(x)^\top r(x). \tag{3.11.2}$$

The gradient $\nabla f(x)$ is formed by components

$$(\nabla f(x))_j = \frac{\partial f}{\partial x_j}(x) = \sum_{i=1}^m r_i(x) \frac{\partial r_i}{\partial x_j}(x).$$

Define the Jocobian of r,

$$J(x) = \left(\frac{\partial r_i}{x_j}(x) \right)_{1 \le i \le m, 1 \le j \le n}.$$

Then we have

$$\nabla f(x) = J(x)^\top r(x). \tag{3.11.3}$$

Now calculate $\nabla^2 f(x)$. We have

$$\begin{aligned} \frac{\partial^2 f}{\partial x_k \partial x_j} &= \frac{\partial}{\partial x_k} \left(\sum_{i=1}^m r_i(x) \frac{\partial r_i}{\partial x_j}(x) \right) \\ &= \sum_{i=1}^m \left(\frac{\partial r_i}{\partial x_k}(x) \frac{\partial r_i}{\partial x_j}(x) + r_i(x) \frac{\partial^2 r_i}{\partial x_k \partial x_j}(x) \right). \end{aligned} \tag{3.11.4}$$

Compare this with $J(x)$, let

$$(S(x))_{k,j} = r_i(x) r_i(x) \frac{\partial^2 r_i}{\partial x_k \partial x_j}(x).$$

Then

$$\nabla^2 f(x) = J(x)^\top J(x) + S(x). \tag{3.11.5}$$

Therefore, if we use the Newton's Method, then

$$x^{k+1} = x^k - (J(x)^\top J(x) + S(x))^{-1} J(x)^\top r(x). \tag{3.11.6}$$

Guass-Newton just ignored $S(x)$ in Eq. (3.11.6) to save computation, so it is

$$x^{k+1} = x^k - (J(x)^\top J(x))^{-1} J(x)^\top r(x). \tag{3.11.7}$$

The class of problems formed like non-linear least square appears when data-fitting.

Example 3.11.1 (non-linear data-fitting). Given a model

$$y = A \sin(\omega t + \phi).$$

Determine A, ω, ϕ by observations $(t_i, y_i)_{i=1}^n$. Then the optimization problem

should be

$$\min_{A,\omega,\phi} \sum_{i=1}^{n} (y_i - A\sin(\omega t_i + \phi))^2.$$

The $r_i(x)$ is $y_i - A\sin(\omega t_i + \phi)$ here.

Chapter 4

Linear Programming

4.1 Geometric Concepts

Consider in $\mathbb{R}^n$, $\{x : a^\top x = b\}$ is called a *hyperplane*. $\{x : a\top x \geq b\}$ is called a *halfspace*. The intersection of finitely many halfspaces is called a *polyhedron*.

Definition 4.1.1 (extreme points). Consider the polyhedron $P = \{x : Ax \geq b\} \subseteq \mathbb{R}^n$. $x \in P$ is an *extreme point* of P if there is no $y, z \in P$ that not equals to x such that $x = \lambda y + (1 - \lambda)z$ for $0 < \lambda < 1$.

In other words, an extreme point is not strictly within the line segment connecting two other points in P.

Definition 4.1.2 (vertex). Still in P. $x \in P$ is a *vertex* of P if there exists $c \in \mathbb{R}^n$ such that $c^\top x < c^\top z$ for all $z \in P - \{x\}$.

In other words, a vertex is the unique minimizer of some linear function over P.

Lemma 4.1.3. A vertex or extreme point has n linearly independent active constraints.

Proof. If not, then the point will lie in a line, which contradicts the fact that there is no line connecting it for extreme points and the minimizer is unique for vertex. □

4.2 Standard Form of Linear Programming

A standard linear programming problem consists of the following:

1. a variable $x \in \mathbb{R}^n$;
2. a cost vector $c \in \mathbb{R}^n$;
3. a right hand side vector $b \in \mathbb{R}^m$;
4. coefficient matrix $A \in \mathbb{M}_{m\times n}(\mathbb{R})$.

The standard form is

$$\begin{aligned} &\min c^\top x \\ &\text{subject to } Ax = b, x \geq 0 \end{aligned}. \tag{4.2.1}$$

Any non-standard form can be reformulated to the standard form. We have the following methods:

1. "maximum" objective function: turn to minimize its negative;
2. $\leq$ constraint: add non-negative slack variable;
3. $\geq$ constraint: subtract non-negative slack variable;
4. $x_i \leq 0$: substitute x_i by $-x_i$ throughout;
5. free x_i: introduce $u_i, v_i \geq 0$ and substitute $u_i - v_i$ throughout;
6. constraint $|x_i| \leq b_i$: replace by $x_i \leq b_i$ and $-x_i \leq b_i$ then substitute x_i by $u_i - v_i$;
7. objective function contains $|x_i|$: introduce $u_i, v_i \geq 0$ and substitute x_i by $u_i - v_i$, $|x_i|$ by $u_i + v_i$.

4.3 Assumptions in Linear Programming

Now assume we have a linear programming problem with standard form (4.2.1), with $\operatorname{rank}(A) = m$, i.e. it is full rank on rows, the constraints are linearly independent. Further assume that $m < n$ since if $m \geq n$, then $Ax = b$ would have only one or no solution. We can also assume $b \geq 0$ (means every component greater than zero), since we can force the corresponding rows of A to change sign.

For convenience, we often reorder the columns of A so that the columns we are interested in appear first. Specifically, let B be a matrix whose columns are m linearly independent of A. We would often reorder the columns of A so that the columns in B appears first: $A = [B, D]$, where $B \in \mathbb{M}_{m\times m}(\mathbb{R})$ is full rank and $D \in \mathbb{M}_{m\times(n-m)}$.

Therefore, $Bx_B = b$ is solvable with $x_B = B^{-1}b$. Note that $x = [x_B^\top, 0^\top]^\top$ is a solution to $Ax = b$ since

$$\begin{bmatrix} B & D \end{bmatrix} \begin{bmatrix} x_B \\ 0 \end{bmatrix} = Bx_B = b.$$

Definition 4.3.1.

1. We call $[x_B^\top, 0^\top]^\top$ a *basic solution* to $Ax = b$ with respect to B. We refer to the components of the vector x_B as *basic variables* and the columns of B as basic columns.
2. If some of the basic variables in x_B are zero, then the basic solution is said to be a *degenerate basic solution*.
3. A vector x satisfying $Ax = b$, $x \geq 0$ is said to be a *feasible solution*.
4. A feasible solution that is also basic is called a *basic feasible solution*.
5. If the basic feasible solution is a degenerate basic solution, then it is called a *degenerate basic feasible solution*.

In the proceeding sections, we would use the above assumptions and terminologies.

4.4 Properties of Basic Solution

Recall Lemma 4.1.3 that any extreme point of P is given by n equality constraints. We have already have m in $Ax = b$, so we should select $(n - m)$ constraints from $x_1 \geq 0, \ldots, x_n \geq 0$.

Note that if $A = [B, D]$ and B is of full rank m, then

$$\operatorname{rank} \begin{bmatrix} B & D \\ 0 & I \end{bmatrix} = \operatorname{rank} \begin{bmatrix} B & 0 \\ 0 & I \end{bmatrix} = n.$$

Therefore, setting $(n - m)$ components to zero would uniquely determine a feasible point in $\mathbb{R}^n$.

Conversely, in order to select an extreme point x of P, we should find B with rank m and compute $x = [x_B^\top, 0^\top]^\top$ and check if $x_B \geq 0$. If yes, then it should be an extreme point of P (we haven't prove this, see Theorem 4.5.1).

Definition 4.4.1 (optimal feasible solution). Any vector x that yields the minimum value of the function $c^\top x$ over the set of all feasible vectors is said to be an *optimal feasible solution*.

An optimal feasible solution that is basic is said to be an *optimal basic feasible solution*.

Theorem 4.4.2 (The Fundamental Theorem of Linear Programming). Consider a linear program in standard form satisfied our assumptions.

1. If there exists a feasible solution, then there exists a basic feasible solution.
2. If there exists an optimal feasible solution, then thre exists a optimal basic feasible solution.

Proof. Consider two cases. Let $A = [a_1, \ldots, a_p, \ldots, a_n]$, where p is the number of positive components of the given solution. Note that the order of the columns is reordered correspond to $x = [x_1, \ldots, x_p, \ldots, x_n]^\top$. Then

$$x_1 a_1 + x_2 a_2 + \ldots x_p a_p = b.$$

Case 1 for $a_1, \ldots, a_p$ linearly independent, which is an easy case. Case 2 for linearly dependent. Try eliminate p to $p-1$ repeatly. Note that linear depending means that there exists a non-zero solution $y = [y_1, \ldots, y_p, 0, \ldots, 0]$ such that

$$y_1 a_1 + y_2 a_2 + \cdots + y_p a_p = 0.$$

Then the solution $x - \epsilon y$ would still satisfy $A[x - \epsilon y] = b$, which makes x not basic since it is not an extreme point [a]. □

[a] Actually the proof does not need to x is not basic. Instead, it needs eliminating components to make it to be basic. The illustrating here just help understand the process of the proof. For the reason, see preceding sections.

Due to item 2, in order to find optimal feasible solution, we only need to find optimal basic solution.

4.5 Connecting Basic Feasible Solutions and Extreme Points

Theorem 4.5.1. Let Ω be the convex set consisting of all feasible solutions. Then x is an extreme point of Ω if and only if x is a basic feasible solution to $Ax = b$.

Proof. Suppose x is an extreme point, then

$$z_1 = [x_1 + \epsilon y_1, x_2 + \epsilon y_2, \ldots, x_p + \epsilon y_2, 0, \ldots, 0],$$

and

$$z_2 = [x_1 - \epsilon y_1, x_2 - \epsilon y_2, \ldots, x_p - \epsilon y_2, 0, \ldots, 0],$$

where x_i, y_i is the same in the proof of Theorem 4.4.2. Note that for small ϵ, z_1, z_2 are both feasible. Then $x = 1/2(z_1 + z_2)$ is also feasible. Since x is an extreme point, $z_1 = z_2$ so that $y_i = 0$. Then $a_1, \ldots, a_p$ are linearly independent.
On the other side, assume $x \in \Omega$ a basic feasible solution. Let $y, z \in \Omega$ such that $x = \alpha y + (1 - \alpha)z$ for some $\alpha \in (0, 1)$. We need to show $y = z$. Since the last $n - m$ components of x are zero, the last $n - m$ components of y, z are zero as well since $y, z \geq 0$. And note that

$$(y_1 - z_1)a_1 + \ldots (y_m - z_m)a_m = 0.$$

We have $y_i = z_i$, $i = 1, \ldots, m$. □

Combined with Theorem 4.4.2 and Theorem 4.5.1, we can see that in solving linear programming problems, we need only examine the extreme points of the constraint set.

4.6 Simplex Method

In the remainder of this section, we assume that every basic feasible solution of the linear programming problem is a non-degenerated basic feasible solution. We make this assumption primarily for convenience – all arguments can be extended to include degeneracy.

The essence of the simplex method is to move from one basic feasible solution to another until an optimal basic feasible solution is found.

An edge in $\mathbb{R}^n$ is obtained from an basic feasible solution by removing one equation from the n linearly independent equations that define it. More precisely, consider the basic feasible solution $x = [x_B^\top, 0^\top]^\top$ with $A = [B, D]$. Denotes the index sets $I_B = \{1, \ldots, m\}$ and $I_D = \{m + 1, \ldots, n\}$. Then x is in fact controlled by

$$\{x\} = \{x : Ax = b, x_i = 0, \forall i \in I_D\}.$$

Then pick $j \in I_D$, the edge connected to x is then

$$\{x \geq 0 : Ax = b, x_i = 0, \forall i \in I_D - \{j\}\}.$$

The edge direction for the basic feasible point x corresponding to x_j is

$$\delta^j = \begin{bmatrix} -B^{-1}a_j \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix},$$

where a_j is the j-th column of A and $a_j = B\mathbb{M}_B(a_j)$. So that $B^{-1}a_j$ can be understood as the representation of a_j under basis B. Note that $A\delta^j = 0$. Therefore,

$$\{x \geq 0 : Ax = b, x_i = 0, \forall i \in I_D - \{j\}\} = \{y \geq 0 : y = x + \epsilon\delta^j, \epsilon \geq 0\}. \quad (4.6.1)$$

The unit change in $c^\top x$ along δ^j is

$$\bar{c}_j = c^\top \delta^j = c_j - c_B^\top B^{-1} a_j.$$

If $\bar{c}_j < 0$, then the direction will decrease the objective function.

The *reduced cost vector* is defined as

$$\bar{c}^\top \triangleq c^T - c_B^\top B^{-1} A,$$

which is also $\bar{c} = [\bar{c}_1, \ldots, \bar{c}_n]^\top$. Note that in basic part $\bar{c}_B^\top = c_B^T - c_B^\top B^{-1} B = 0$.

Theorem 4.6.1. Let $A = [B, D]$, where $\text{rank}(B) = m$. Suppose that $x = [x_B^\top, 0^\top]^\top$ is a basic feasible solution. Then x is optimal if and only if

$$\bar{c}^\top = c^\top - c_B^\top B^{-1} A \geq 0^\top.$$

The theorem states that if all the direction of reduced cost cannot decrease the objective function, then the basic feasible solution is the optimal.

Proof. A Rigorous proof need more tedious matrix algebra, see Chapter 16 in [CZ04]. □

At this point we have all the necessary steps for the simplex algorithm.

1. Select linearly independent columns in A to form B;
2. Calculate the corresponding basic feasible solution x;
3. check if $\bar{c}^\top \geq 0^\top$. If yes, then return x as the optimal basic feasible solution. If not:
   (a) choose j such that $\bar{c}_j < 0$ and then move along δ^j to get a new basic feasible solution x';

(b) update basis B' corresponding to x';

(c) Return to step 3.

Note that in Step 3(c) there might be a case that $\delta_B^j = -B^{-1}a_j \geq 0$, then for any $\alpha \geq 0$, $x+\alpha\delta^j \geq 0$ and it is feasible, the problem is thus unbounded since $c^\top(x+\alpha\delta^j)$ is unbounded.

For step 3(a), recall that the feasible direction is (4.6.1). To arrive at another basic feasible solution, let

$$\epsilon_{\min} = \min\left\{\frac{x_i}{-\delta_i^j} : i \in I_B, \delta_i^j < 0\right\}.$$

In other words, $\epsilon_{\min}$ is the smallest scalar such that for some index i in I_B, $x_i+\alpha\epsilon = 0$ and $x_j + \alpha\epsilon \geq 0$ for all $j \in I_B$.

In step 1, a question is how to select linearly independent columns. An easy case is when the original constraint looks look $Ax \leq b$ and $x \geq 0$. Then adding slack variable, we would have

$$[A, I]\begin{bmatrix} x \\ s \end{bmatrix} = b.$$

Then an obvious basis is I. And the corresponding basic feasible solution is $x_B = b$. For general case, we do not have any quick methods other than checking the rank of trials.

A final question is: why we assume in the beginning of the section that x is non-degenerate? What would happen if the case occurs? The issue happens in step 3(a), that $\epsilon_{\min}$ might be zero so that the basic feasible solution remains unchanged.

Chapter 5

Optimization Problem with Simple Constraints

5.1 Optimality Conditions

We now consider function $f : C \to \mathbb{R}$, where C is a convex closed set of $\mathbb{R}^n$ and f is a differentiable function (not necessarily continuous differentiable as compared with the conditions in Section 2.4).

Theorem 5.1.1 (First Order Necessary Condition). Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function and C is a convex closed set of $\mathbb{R}^n$. Then

1. if x^* is a local minimizer of f over $C \subset \mathbb{R}^n$, then
$$\nabla f(x^*)^\top (x - x^*) \geq 0$$
for all $x \in C$.
2. Moreover, suppose that f is a convex function on C. Then x^* is a global minimum of f over C if and only if
$$\nabla f(x^*)^\top (x - x^*) \geq 0$$
for all $x \in C$.

Rather than the Taylor's expansion treatment in the proofs of Section 2.4, we would make full use of the convex property of C.

Proof. To prove item 1, let $x \in C$. Since C is a convex set and x, x^* are both in C, for all $\alpha \in [0, 1]$, $\alpha x + (1 - \alpha)x^* = x^* + \alpha(x - x^*) \in C$. Therefore,

$$f(x^* + \alpha(x - x^*)) - f(x^*) \geq 0,$$

for all $\alpha \in [0, 1]$. Dividing the above inequality by α and let $t \to 0$, the result follows as the definition of derivative. More precisely, let $\phi(\alpha) = f(x^* + \alpha(x - x^*))$. Then $\phi'(\alpha) = \nabla f(x^* + \alpha(x - x^*))^\top (x - x^*)$, so that $\phi'(0) = \nabla f(x^*)^\top (x - x^*)$.

To prove item 2, the necessity follows from item 1. Assume that $\nabla f(x^*)^\top (x - x^*) \geq 0$ for all $x \in C$. Then by the characterization of convex function, Proposition 2.1.3,

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

for all $x \in C$. □

Definition 5.1.2 (normal cone). Let $C \subset \mathbb{R}^n$ be a convex set and $x^* \in C$. We say η is a *normal vector* to C at x^* if

$$\eta^\top (x - x^*) \leq 0$$

for all $x \in C$. The set of normal vectors to C at x^* is called the *normal cone* of C at x^* and is denoted by

$$N_C(x^*) = \{\eta \in \mathbb{R}^n : \eta^\top (x - x^*) \leq 0, \forall x \in C\}.$$

The geometric meaning of normal cone at x^* is the direction that is "outside" the tangent of x. Therefore if x^* is an interior point, then $N_C(x^*) = \{0\}$.

Corollary 5.1.3. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function and C is a convex closed set of $\mathbb{R}^n$. Then

1. if x^* is local minimizer of f over C, then

$$0 \in \nabla f(x^*) + N_C(x^*).$$

2. Moreover, suppose that f is a convex function on C. Then x^* is a global minimum of f over C if and only if

$$0 \in \nabla f(x^*) + N_C(x^*).$$

Proof. Just note that $0 \in \nabla f(x^*) + N_C(x^*)$ if and only if $-\nabla f(x^*) \in N_C(x^*)$ if and only if $\nabla f(x^*)^\top (x - x^*) \geq 0$. □

5.2 Functions on Convex Set

Theorem 5.2.1. Suppose that f is a convex function on a bounded closed convex set C. Further assume that f has a global maximum on C. Then one can find a global maximum which lies at the boundary point of C.

Proof. If not, then a line through x intersects the boundary will leads to a contradiction. □

Recall that the propositions and theorems in Hilbert space of closest points, see my notes on Analysis in the part of Hilbert space, functional analysis. It is easy to state the following theorems and propositions.

Theorem 5.2.2 (The Closest Point Theorem). Let C be a closed convex set in $\mathbb{R}^n$ and $y \notin C$. Then there exists a unique $x^* \in C$ that is the closest point in C to y. And for every $x \in C$,

$$\langle y - x^*, x - x^* \rangle \leq 0.$$

Corollary 5.2.3 (Basic Separation). Suppose that C is a closed convex set in $\mathbb{R}^n$ and $y \notin C$. Then there exists $0 \neq a \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$ such that

$$\langle a, x \rangle \leq \alpha < \langle a, y \rangle .$$

This theorem means that there exists a linear function such that $f(x) < f(y)$ for all $x \in C$, i.e. a separation function.

Lemma 5.2.4 (Support Lemma). Suppose that C is a non-empty convex set in $\mathbb{R}^n$ and z is a boundary point of C. Then there exists $0 \neq a \in \mathbb{R}^n$ such that

$$\langle a, x \rangle \leq \langle a, z \rangle ,$$

for all $x \in C$.

Proof. Use the basic separation to construct a sequence and consider $\overline{C}$. □

This lemma shows that there exists a linear function that "supports" the convex set, i.e. $f(x) \leq f(z)$ for all $x \in C$ and $z \in \partial C$.

5.3 Subgradient

Lemma 5.3.1 (Existence of Subgradient). Suppose that C is a non-empty convex set and $f : C \to \mathbb{R}$ is convex. If $x^* \in C^\circ$, then there is a vector $d \in \mathbb{R}^n$ such that

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

for all $x \in C$.

Proof. Since f is convex, the epigraph of g is a convex set. It is obvious that $(x^*, f(x^*)) \in \text{epi}(f)$ and hence $\text{epi}(f)$ is non-empty convex set. Note that $(x^*, f(x^*))$ lies on the boundary of $\text{epi}(f)$. By the Support Lemma, there exists $(b, c) \in \mathbb{R}^n \times \mathbb{R}$ such that

$$b^\top x + cr \leq b^\top x^* + cf(x^*),$$

for any $(x, r) \in \text{epi}(f)$, which implies that

$$b^\top (x - x^*) \leq c(f(x^*) - r) \tag{5.3.1}$$

for all $r \geq f(x)$. Since r can go to ∞, $c \leq 0$.
We now show that $c < 0$. If $c = 0$, then $b^\top (x - x^*) \leq 0$ for all $x \in C$. However, this is not possible since $x^* \in C^\circ$. Hence $c < 0$. Then let $r = f(x)$ and set $d^\top = -b^\top / c$ in Eq. (5.3.1), the result follows. □

Definition 5.3.2 (subgradient). Let $C \subseteq \mathbb{R}^n$ be a convex set and f be a convex function defined on C. A vector $d \in \mathbb{R}^n$ satisfying

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

for all $x \in C$ is called a *subgradient* of f at x^*. We denote the set of all subgradients at x by

$$\partial f(x^*) \triangleq \{d \in \mathbb{R}^n : f(x) \geq f(x^*) + d^\top (x - x^*), \forall x \in C\}.$$

Lemma 5.3.1 guarantees that if $x^* \in C^\circ$ and f is a convex function, then the subgradient exists.

In the case when f is convex and differentiable at x^*,

$$\partial f(x^*) = \{\nabla f(x^*)\} = \nabla f(x^*).$$

The following is a trivial but important fact.

Theorem 5.3.3. Let $C \subseteq \mathbb{R}^n$ be an open convex set and $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function. Then x^* is a global minimizer of f on C if and only if $0 \in \partial f(x^*)$.

Proof. Note that $f(x) \geq f(x^*)$ if and only if $f(x) \geq f(x^*) + 0^\top (x - x^*)$. □

5.4 Value Functions and Envelop Theorems

Suppose the objective function f is of the form

$$f(x, \alpha) = f(x_1, \ldots, x_n, \alpha_1, \ldots, \alpha_k),$$

where $x \in S \subseteq \mathbb{R}^n$ and $\alpha \in \mathbb{R}^k$. Assume that we have found the minimum of $f(x, \alpha)$ for each fixed α. We denote this value by $V(\alpha)$ and call it the *value function*, i.e.

$$V(\alpha) = \min_{x \in S} f(x, \alpha).$$

If we denote $x^*(\alpha)$ the minimizer of $f(x, \alpha)$ for fixed α, then $V(\alpha) = f(x^*(\alpha), \alpha)$. Note that the minimizer might not be unique and $x^*(\alpha)$ just denotes one of them.

Envelope theorem illustrates the "sensitivity" of the minimal, i.e. the change of $V(\alpha)$. For example,

$$\begin{aligned}\frac{\partial V}{\partial \alpha_i}(\alpha) &= \frac{\partial}{\partial \alpha_i} f(x^*(\alpha), \alpha) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(x^*(\alpha), \alpha) \frac{\mathrm{d}x_i^*}{\mathrm{d}\alpha_i}(\alpha) + \frac{\partial f}{\partial \alpha_i}(x^*(\alpha), \alpha) \\ &= \frac{\partial f}{\partial \alpha_i}(x^*(\alpha), \alpha),\end{aligned}$$

where we use $\partial f(x^*(\alpha), \alpha)/\partial x_i = 0$ since $x = x^*(\alpha)$ minimizes $f(x, \alpha)$. The result of $\nabla V(\alpha) = \nabla_\alpha f(x^*(\alpha), \alpha)$ is also deserved. We summarize it in the following theorems with different conditions.

Theorem 5.4.1 (Envelop Theorem A). Let $f(x, \alpha) : \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R}$. Let $S \subseteq \mathbb{R}^n$. Consider the problem $\min_{x \in S} f(x, \alpha)$. Suppose that $x^*(\alpha)$ is a solution of this problem for every α in some open ball $B(\bar{\alpha}, \delta)$ with $\delta > 0$. Furthermore, assume the map $\alpha \mapsto f(x^*(\bar{\alpha}), \alpha)$ (with $\bar{\alpha}$ fixed!) and the value function $V(\alpha)$ are both differentiable at $\bar{\alpha}$. Then

$$\nabla V(\bar{\alpha}) = \nabla_\alpha f(x^*(\bar{\alpha}), \bar{\alpha}).$$

Proof. Consider the function $\phi(\alpha) = f(x^*(\bar{\alpha}), \alpha) - V(\alpha)$. Show that $\bar{\alpha}$ is the minimizer of ϕ. □

Corollary 5.4.2 (Envelop Theorem B). Let $f(x, \alpha)$ be a C^2 function for all x in an open convex set $S \subseteq \mathbb{R}^n$ and for each α in an open ball $B(\bar{\alpha}, \delta) \subset \mathbb{R}^k$, the function $x \mapsto f(x, \alpha)$ is convex, and when $\alpha = \bar{\alpha}$, the Hessian matrix of f respect to x is positive definite. Moreover, assume that x^* is a minimum for $x \mapsto f(x, \bar{x})$ in S. Then $V(\alpha) = \min_{x \in S} f(x, \alpha)$ is defined for all $a \in B(\bar{\alpha}, \delta)$. Moreover, the value function is continuously differentiable at $\bar{\alpha}$ and

$$\nabla V(\bar{\alpha}) = \nabla_\alpha f(x^*, \bar{\alpha}).$$

Proof. Consider $\nabla_x f(x, \alpha) = 0$. By the implicit function theorem, $x(\alpha)$ is a C^1 function for $\alpha \in B(\bar{\alpha}, \epsilon)$. And $x(\alpha)$ is the minimum of $f(x, \alpha)$ for fixed α. Since $x(\alpha)$ is differentiable at $\bar{\alpha}$ so is $V(\alpha) = f(x(\alpha), \alpha)$ since $f \in C^2$. Then apply previous Envelop Theorem. □

Corollary 5.4.3 (Envelop Theorem C). Suppose that $V(\alpha) = \inf_{x \in S} f(x, \alpha)$ is finite and convex in $\alpha \in A$, where A is an open convex set in $\mathbb{R}^k$, and $S \subseteq \mathbb{R}^n$. Assume that the point $(x^*, \bar{\alpha}) \in S \times A$ satisfies $f(x^*, \bar{\alpha}) = V(\bar{\alpha})$ and the gradient vector $\nabla_\alpha f$ exists at $(x^*, \bar{\alpha})$. Then the value function $V(\alpha)$ is differentiable at $\bar{\alpha}$ and

$$\nabla V(\bar{\alpha}) = \nabla_\alpha f(x^*, \bar{\alpha}).$$

Proof. Since A is convex and $V(\alpha)$ is a convex function, there exists $\xi \in \partial V(\bar{\alpha})$ such that

$$f(x^*, \alpha) - f(x^*, \bar{\alpha}) \geq V(\alpha) - V(\bar{\alpha}) \geq \xi^\top (\alpha - \bar{\alpha}).$$

This means that $\alpha \mapsto f(x^*, \alpha)$ has a subgradient ξ at $\bar{\alpha}$. But by assumption, it is differentiable at $(x^*, \bar{\alpha})$. Therefore, $\xi = \nabla_\alpha f(x^*, \bar{\alpha})$, i.e.

$$\nabla V(\bar{\alpha}) = \nabla_\alpha f(x^*, \bar{\alpha}).$$

□

Bibliography

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

[CZ04] Edwin KP Chong and Stanislaw H Zak. *An introduction to optimization.* John Wiley & Sons, 2004.