Optimization Algorithm Notes

Depu Meng

Feb. 2019

1 Introduction to Optimization Algorithms

1.1 Goal of the Course

- Understand foundations of optimization
- Learn to analyze widely used optimization algorithms
- Be familiar with implementation of optimization algorithms

1.2 Topics Involved

- Unconstrained optimization
- Constrained optimization
- Convex optimization
- Sparse optimization
- Stochastic optimization
- Combinational optimization
- Global optimization

1.3 Basic Concepts

Problem Definition Find the value of the decision variable s.t. objective function is maximized/minimized under certain conditions.

$$\min f(x) \tag{1}$$
$$s.t. x \in \mathcal{S} \subset \mathbb{R}^n \tag{2}$$

Here, we call $\mathcal{S}$ *feasible region.*

We often denote constrained optimization Problem as

$$\min f(x) \tag{3}$$
$$s.t. \quad g_i(x) \geq 0, i = 1, ..., n \tag{4}$$
$$b_i(x) = 0, i \in 1, ..., m \tag{5}$$

Definition 1. *Global Optimality. For global optimal value $x^* \in \mathcal{S}$,*

$$f(x^*) \leq f(x), \forall x \in \mathcal{S} \tag{6}$$

Definition 2. *Local Optimality. For local optimal value $x^* \in \mathcal{S}$, $\exists U(x^*)$, such that*

$$f(x^*) \leq f(x), \forall x \in \mathcal{S} \cap U(x^*) \tag{7}$$

Definition 3. *Feasible direction. Let $x \in \mathcal{S}$, $d \in \mathbb{R}^n$ is a non-zero vector. if $\exists \delta > 0$, such that*

$$x + \lambda d \in \mathcal{S}, \forall \lambda \in (0, \delta) \tag{8}$$

Then d is a ***feasible direction*** *at x. We denote $F(x, \mathcal{S})$ as the set of feasible directions at x.*

Definition 4. *Descent direction. $f(x) : \mathbb{R}^n \to \mathbb{R}$, $x \in \mathbb{R}^n$, d is a non-zero vector. If $\exists \delta > 0$, such that*

$$f(x + \lambda d) < f(x), \forall \lambda \in (0, \delta) \tag{9}$$

Then d is a ***descent direction*** *at x. We denote $D(x, f) = \{d | \bigtriangledown f(x)^T d < 0\}$ as the set of descent direction at x.*

1.4 Optimal Conditions

Unconstrained Optimization

First-order necessery condition: $f(x)$ is differentiable at x,

$$\bigtriangledown f(x) = 0 \tag{10}$$

Second-order necessery condition: $f(x)$ is second-order differentiable at x,

$$\bigtriangledown f(x) = 0 \tag{11}$$

$$\bigtriangledown^2 f(x) \geq 0 \tag{12}$$

Constrained Optimization

Theorem 1. *Fritz-John Condition*

For constrained optimization problem

$$\min f(x) \tag{13}$$

$$s.t. \quad g_i(x) \geq 0, i = 1, ..., n \tag{14}$$

$$h_i(x) = 0, i \in 1, ..., m \tag{15}$$

Denote $I(x) = \{i \in \{1, ..., n\} | g_i(x) = 0\}$. For $x \in \mathcal{S}$, f and $g_i, i \in I(x)$ is differentiable at x, $h_j(x)$ is continuously differentiable at x. If x is local optimal, then there exists non-trivial $\lambda_0, \lambda_i \geq 0, i \in I(x)$ and μ_j, such that

$$\lambda_0 \bigtriangledown f(x) - \sum_{i \in I(x)} \lambda_i \bigtriangledown g_i(x) - \sum_{j=1}^{m} \mu_j \bigtriangledown h_j(x) = 0 \tag{16}$$

Proof. (i) If $\{\bigtriangledown h_j(x)\}$ is linearly dependent, then there exists non-trivial μ_j, such that

$$\sum_{j=1}^{m} \bigtriangledown h_j(x) = 0 \tag{17}$$

Let $\lambda_0, \lambda_i, i \in I(x) = 0$, then (13) holds.

(ii) If $\{\bigtriangledown h_j(x)\}$ is linearly independent, Denote

$$F_g = F(x,g) = \{d | \bigtriangledown g_i(x)^T d > 0, i \in I(x)\} \tag{18}$$
$$F_h = F(x,h) = \{d | \bigtriangledown h_j(x)^T d = 0, j = 1, ..., m\} \tag{19}$$

If x is a optimal value, then appearently $F(x, \mathcal{S}) \cap D(x,f) = \varnothing$. Due to the independence of $\{\bigtriangledown h_j(x)\}$, we have $F_g \cap F_h \subset F(x, \mathcal{S})$, then

$$F_g \cap F_h \cap D(x,f) = \varnothing \tag{20}$$

that is

$$\begin{cases} \bigtriangledown f(x)^T d < 0 \\ \bigtriangledown g_i(x)^T d > 0, i \in I(x) \\ \bigtriangledown h_j(x)^T d = 0, j = 1, ..., m \end{cases} \tag{21}$$

has no solution. Let

$$A = \{\bigtriangledown f(x)^T, \bigtriangledown g_i(x)\}^T, i \in I(x) \tag{22}$$
$$B = \{- \bigtriangledown h_j(x)\}, j = 1, ..., m \tag{23}$$

Then (21) is equivalent to

$$\begin{cases} A^T d < 0 \\ B^T d = 0 \end{cases} \tag{24}$$

has no solution.
Denote

$$S_1 = \left\{ \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} | y_1 = A^T d, y_2 = B^T d, d \in \mathbb{R}^n \right\} \tag{25}$$
$$S_2 = \left\{ \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} | y_1 < 0, y_2 = 0 \right\} \tag{26}$$

S_1, S_2 are non-trivial convex sets, and $S_1 \cap S_2 = \varnothing$. From *Hyperplane Separation Theorem*: $\exists \begin{pmatrix} p_1 \\ p_2 \end{pmatrix}$, so that

$$p_1^T A^T d + p_2^T B^T d \geq p_1^T y_1 + p_2^T y_2, \forall d \in \mathbb{R}^n, \forall \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \in CL(S_2) \tag{27}$$

Let $y_2 = 0, d = 0, y_1 < 0$, we have

$$p_1 \geq 0 \tag{28}$$

Let $\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \in CL(S_2)$ So that

$$(p_1^T A^T + p_2^T B^T)d \geq 0 \tag{29}$$

$$(Ap_1 + Bp_2)^T d \geq 0 \tag{30}$$

Let $d = -(Ap_1 + Bp_2)$, we have

$$Ap_1 + Bp_2 = 0 \tag{31}$$

From above, we have

$$\begin{cases} Ap_1 + Bp_2 = 0 \\ p_1 \geq 0 \end{cases} \tag{32}$$

Let $p_1 = \{\lambda_0, ..., \lambda_{I(x}\}$, $p_2 = \{\mu_1, ..., \mu_m\}$, i.e.,

$$\begin{cases} \lambda_0 \bigtriangledown f(x) - \sum_{i \in I(x)} \lambda_i \bigtriangledown g_i(x) - \sum_{j=1}^{m} \mu_j \bigtriangledown h_j(x) = 0 \\ \lambda_i \geq 0 \end{cases} \tag{33}$$

Theorem 2. *Kuhn-Tucker Condition*

For constrained optimization problem

$$\min f(x) \tag{34}$$

$$s.t. \quad g_i(x) \geq 0, i = 1, ..., n \tag{35}$$

$$h_i(x) = 0, i \in 1, ..., m \tag{36}$$

Denote $I(x) = \{i \in \{1, ..., n\} | g_i(x) = 0\}$*. For* $x \in \mathcal{S}$*,* f *and* $g_i, i \in I(x)$ *is differentiable at* x*,* $h_j(x)$ *is continuously differentiable at* x*.* $\{\bigtriangledown g_i(x), i \in I(x); \bigtriangledown h_j(x), j = 1, ..., m\}$ *is linearly independent. If* x *is local optimal, then* $\exists \lambda_i \geq 0$ *and* μ_j*, such that*

$$\bigtriangledown f(x) - \sum_{i \in I(x)} \lambda_i \bigtriangledown g_i(x) - \sum_{j=1}^{m} \mu_j \bigtriangledown h_j(x) = 0 \tag{37}$$

1.5 Descent function

Definition 5. *Descent function. Denote solution set* $\Omega \in X$*,* $\mathcal{A}$ *is an algorithm on* X*,* $\psi : X \to \mathbb{R}$*. If*

$$\psi(y) < \psi(x), \quad \forall x \notin \Omega, y \in \mathcal{A}(x) \tag{38}$$

$$\psi(y) \leq \psi(x), \quad \forall x \in \Omega, y \in \mathcal{A}(x) \tag{39}$$

Then ψ *is a* ***descent function*** *of* $(\Omega, \mathcal{A})$*.*

1.6 Convergence of Algorithm

Theorem 3. *$\mathcal{A}$ is an algorithm on X, Ω is the solution set, $x^{(0)} \in X$. If $x^{(k)} \in \Omega$, then the iteration stops. Otherwise set $x^{(k+1)} = \mathcal{A}(x^{(k)}), k := k+1$. If*

- *$\{x^{(k)}\}$ in a compact subset of X*
- *There exists a continuous function ψ, ψ is a descent function of $(\Omega, \mathcal{A})$*
- *$\mathcal{A}$ is closed on Ω^C*

Then, any convergent subsequence of $\{x^{(k)}\}$ converges to $x, x \in \Omega$.

Proof.

1.7 Search Methods

Line Search

Generate $d^{(k)}$ from $x^{(k)}$,

$$x^{(k+1)} = x^{(k)} + \alpha_k d^{(k)} \tag{40}$$

. search α_k in 1-D space.

Trust Region

Generate local model $Q_k(s)$ of $x^{(k)}$,

$$s^{(k)} = \arg\min Q_k(s) \tag{41}$$

$$x^{(k+1)} = x^{(k)} + s^{(k)} \tag{42}$$

2 Unconstrained Optimization

2.1 Gradient Based Methods

Algorithm 1: Example of gradient based algorithm

Data: Solution set Ω, cost function f
$x^{(0)} \in \mathbb{R}^n, k := 0;$
while $x^{(k)} \notin \Omega$ **do**
  $d^{(k)} = -H_k \bigtriangledown f(x^{(k)})$, ($H_k$ is a positive definite symmetrical matrix);
  solve $\min_{\alpha_k \geq 0} f(x^{(k)} + \alpha_k d^{(k)})$;
  $x^{(k+1)} = x^{(k)} + \alpha_k d^{(k)}$, $k := k + 1$
end

2.2 Determine Search Direction

First-order gradient method

For unconstrained optimization problem

$$\min_{x \in \mathbb{R}^n} f(x) \tag{43}$$

We have

$$f(x) = f(x^{(k)}) + \bigtriangledown f(x^{(k)})^T (x - x^{(k)}) + O(\| x - x^{(k)} \|^2) \tag{44}$$

Set $d^{(k)} = - \bigtriangledown f(x^{(k)})$, when α_k is sufficiently small,

$$f(x^{(k)} + \alpha_k d^{(k)}) < f(x^{(k)}) \tag{45}$$

Second-order gradient method – Newton Direction

$$f(x) = f(x^{(k)}) + \bigtriangledown f(x^{(k)})^T (x - x^{(k)}) \tag{46}$$

$$+ \frac{1}{2}(x - x^{(k)})^T \bigtriangledown^2 f(x^{(k)})(x - x^{(k)}) + O(\| x - x^{(k)} \|^3) \tag{47}$$

Set $d^{(k)} = -G_k^{-1} \bigtriangledown f(x^{(k)})$, where $G_k = \bigtriangledown^2 f(x^{(k)})$, i.e., Hesse matrix of f at $x^{(k)}$.

2.3 Determine Step Factor