5 Algorithms for Minimizing Unconstrained Functions

5.1 General Minimization Problem and Terminologies

Definition 5.1 We define the *general minimization problem* as follows

$$\begin{cases} \text{minimize} & f(\boldsymbol{x}) \\ \text{subject to} & f_j(\boldsymbol{x}) \ \& \ 0, \quad j = 1, 2, \ldots, m \\ & \boldsymbol{x} \in S, \end{cases} \tag{3}$$

where $f : \mathbb{R}^n \to \mathbb{R}$, $f_j : \mathbb{R}^n \to \mathbb{R}$ $(j = 1, 2, \ldots, m)$, the symbol & could be $=$, $\geq$, or $\leq$, and $S \subseteq \mathbb{R}^n$.

Definition 5.2 The *feasible set* Q of (3) is

$$Q = \{\boldsymbol{x} \in S \mid f_j(\boldsymbol{x}) \ \& \ 0, \ (j = 1, 2, \ldots, m)\}.$$

In the following items we assume $S \equiv \mathbb{R}^n$.

- If $Q \equiv \mathbb{R}^n$, (3) is a *unconstrained optimization problem.*
- If $Q \subsetneq \mathbb{R}^n$, (3) is a *constrained optimization problem.*
- If all functionals $f(\boldsymbol{x}), f_j(\boldsymbol{x})$ are differentiable, (3) is a *smooth optimization problem.*
- If one of functionals $f(\boldsymbol{x})$, $f_j(\boldsymbol{x})$ is non-differentiable, (3) is a *non-smooth optimization problem.*
- If all constraints are linear $f_j(\boldsymbol{x}) = \langle \boldsymbol{a}_j, \boldsymbol{x} \rangle + b_j$ $(j = 1, 2, \ldots, m)$, (3) is a *linear constrained optimization problem.*
  - In addition, if $f(\boldsymbol{x})$ is linear, (3) is a *linear programming problem.*
  - In addition, if $f(\boldsymbol{x})$ is quadratic, (3) is a *quadratic programming problem.*
- If $f(\boldsymbol{x})$, $f_j(\boldsymbol{x})$ $(j = 1, 2, \ldots, m)$ are quadratic, (3) is a *quadratically constrained quadratic programming problem.*

Definition 5.3 $\boldsymbol{x}^*$ is called a *global optimal solution* of (3) if $f(\boldsymbol{x}^*) \leq f(\boldsymbol{x}), \quad \forall \boldsymbol{x} \in Q$. Moreover, $f(\boldsymbol{x}^*)$ is called the *global optimal value.* $\boldsymbol{x}^*$ is called a *local optimal solution* of (3) if there exists an open ball $B(\boldsymbol{x}^*, \varepsilon) := \{\boldsymbol{x} \in \mathbb{R}^n \mid \|\boldsymbol{x} - \boldsymbol{x}^*\|_2 < \varepsilon\}$ such that $f(\boldsymbol{x}^*) \leq f(\boldsymbol{x}), \quad \forall \boldsymbol{x} \in B(\boldsymbol{x}^*, \varepsilon) \cap Q$. Moreover, $f(\boldsymbol{x}^*)$ is called a *local optimal value.*

5.2 Complexity Bound for a Global Optimization Problem on the Unit Box

Consider one of the simplest problems in optimization, that is, minimizing a function in the n-dimensional box.

$$\begin{cases} \text{minimize} & f(\boldsymbol{x}) \\ \text{subject to} & \boldsymbol{x} \in B_n := \{\boldsymbol{x} \in \mathbb{R}^n \mid 0 \leq [\boldsymbol{x}]_i \leq 1, \ i = 1, 2, \ldots, n\}. \end{cases} \tag{4}$$

To be coherent, we use the ℓ_∞-norm:

$$\|\boldsymbol{x}\|_\infty = \max_{1 \leq i \leq n} |[\boldsymbol{x}]_i|.$$

Let us also assume that $f(\boldsymbol{x})$ is *Lipschitz continuous on* B_n:

$$|f(\boldsymbol{x}) - f(\boldsymbol{y})| \leq L \|\boldsymbol{x} - \boldsymbol{y}\|_\infty, \quad \forall \boldsymbol{x}, \boldsymbol{y} \in B_n.$$

Let us define a very simple method to solve (4), the **uniform grid method**.

Given a positive integer $p > 0$,

1. Form $(p+1)^n$ points
$$\boldsymbol{x}_{i_1,i_2,\ldots,i_n} = \left(\frac{i_1}{p}, \frac{i_2}{p}, \ldots, \frac{i_n}{p}\right)^T$$
where $(i_1, i_2, \ldots, i_n) \in \{0, 1, \ldots, p\}^n$.

2. Among all points $\boldsymbol{x}_{i_1,i_2,\ldots,i_n}$, find a point $\bar{\boldsymbol{x}}$ which has the minimal value for the objective function.

3. Return the pair $(\bar{\boldsymbol{x}}, f(\bar{\boldsymbol{x}}))$ as the result.

Theorem 5.4 Let $f^* := f(\boldsymbol{x}^*)$ be the global optimal value for (4). Then the uniform grid method yields
$$f(\bar{\boldsymbol{x}}) - f(\boldsymbol{x}^*) \leq \frac{L}{2p}.$$

Proof:
Let $\boldsymbol{x}^*$ be a global optimal solution. Then there are coordinates $(i_1, i_2, \ldots, i_n)$ such that $\boldsymbol{x} := \boldsymbol{x}_{i_1,i_2,\ldots,i_n} \leq \boldsymbol{x}^* \leq \boldsymbol{x}_{i_1+1,i_2+1,\ldots,i_n+1} =: \boldsymbol{y}$. Observe that $[\boldsymbol{y}]_i - [\boldsymbol{x}]_i = 1/p$ for $i = 1, 2, \ldots, n$ and $[\boldsymbol{x}^*]_i \in [[\boldsymbol{x}]_i, [\boldsymbol{y}]_i]$ $(i = 1, 2, \ldots, n)$.

Consider $\hat{\boldsymbol{x}} = (\boldsymbol{x} + \boldsymbol{y})/2$ and form a new point $\tilde{\boldsymbol{x}}$ as:
$$[\tilde{\boldsymbol{x}}]_i := \begin{cases} [\boldsymbol{y}]_i, & \text{if } [\boldsymbol{x}^*]_i \geq [\hat{\boldsymbol{x}}]_i \\ [\boldsymbol{x}]_i, & \text{otherwise.} \end{cases}$$

It is clear that $|[\tilde{\boldsymbol{x}}]_i - [\boldsymbol{x}^*]_i| \leq 1/(2p)$ for $i = 1, 2, \ldots, n$. Then $\|\tilde{\boldsymbol{x}} - \boldsymbol{x}^*\|_\infty = \max_{1 \leq i \leq n} |[\tilde{\boldsymbol{x}}]_i - [\boldsymbol{x}^*]_i| \leq 1/(2p)$. Since $\tilde{\boldsymbol{x}}$ belongs to the grid,
$$f(\bar{\boldsymbol{x}}) - f(\boldsymbol{x}^*) \leq f(\tilde{\boldsymbol{x}}) - f(\boldsymbol{x}^*) \leq L\|\tilde{\boldsymbol{x}} - \boldsymbol{x}^*\|_\infty \leq L/(2p).$$

∎

Let us define our goal

$$\boxed{\text{Find } \boldsymbol{x} \in B_n \text{ such that } f(\boldsymbol{x}) - f(\boldsymbol{x}^*) < \varepsilon.}$$

Corollary 5.5 The number of iterations necessary for the problem (4) for the uniform grid method is at most
$$\left(\left\lfloor \frac{L}{2\varepsilon} \right\rfloor + 2\right)^n.$$

Proof:
Take $p = \lfloor L/(2\varepsilon) \rfloor + 1$. Then, $p > L/(2\varepsilon)$ and from the previous theorem, $f(\bar{\boldsymbol{x}}) - f(\boldsymbol{x}^*) \leq L/(2p) < \varepsilon$. Observe that we constructed $(p+1)^n$ points. ∎

Consider the class of problems $\mathcal{P}$ defined as follows:

Model:	$\min_{\boldsymbol{x} \in B_n} f(\boldsymbol{x})$, $f(\boldsymbol{x})$ is ℓ_∞-Lipschitz continuous on B_n.
Oracle:	Only function values are available
Approximate solution:	Find $\bar{\boldsymbol{x}} \in B_n$ such that $f(\bar{\boldsymbol{x}}) - f(\boldsymbol{x}^*) < \varepsilon$

Theorem 5.6 For $\varepsilon < \frac{L}{2}$, the number of iterations necessary for the class of problems $\mathcal{P}$ using any method which uses only function evaluations is always at least $(\lfloor \frac{L}{2\varepsilon} \rfloor)^n$.

Proof:
Let $p = \lfloor \frac{L}{2\varepsilon} \rfloor$ (which is ≥ 1 from the hypothesis).

Suppose that there is a method which requires $N < p^n$ calls of the oracle to solve the problem in $\mathcal{P}$.

Then, there is a point $\hat{\boldsymbol{x}} \in B_n = \{\boldsymbol{x} \in \mathbb{R}^n \mid 0 \leq [\boldsymbol{x}]_i \leq 1,\ i = 1, 2, \ldots, n\}$ where there is no test points in the interior of $B := \{\boldsymbol{x} \mid \hat{\boldsymbol{x}} \leq \boldsymbol{x} \leq \hat{\boldsymbol{x}} + \boldsymbol{e}/p\}$ where $\boldsymbol{e} = (1, 1, \ldots, 1)^T \in \mathbb{R}^n$.

Let $\boldsymbol{x}^* := \hat{\boldsymbol{x}} + \boldsymbol{e}/(2p)$ and consider the function $\bar{f}(\boldsymbol{x}) := \min\{0, L\|\boldsymbol{x} - \boldsymbol{x}^*\|_\infty - \varepsilon\}$. Clearly, $\bar{f}$ is ℓ_∞-Lipschitz continuous with constant L and its global minimum is $-\varepsilon$. Moreover, $\bar{f}(\boldsymbol{x})$ is non-zero valued only inside the box $B' := \{\boldsymbol{x} \mid \|\boldsymbol{x} - \boldsymbol{x}^*\|_\infty \leq \varepsilon/L\}$.

Since $2p \leq L/\varepsilon$, $B' \subseteq \{\boldsymbol{x} \mid \|\boldsymbol{x} - \boldsymbol{x}^*\|_\infty \leq 1/(2p)\} \subseteq B$.

Therefore, $\bar{f}(\boldsymbol{x})$ is equal to zero to all test points of our method and the accuracy of the method is ε.

If the number of calls of the oracle is less than p^n, the accuracy can not be better than ε. ∎

Theorem 5.6 supports the claim that the *general optimization problem are unsolvable.*

Example 5.7 Consider a problem defined by the following parameters. $L = 2$, $n = 10$, and $\varepsilon = 0.01$ (1%).

lower bound $(L/(2\varepsilon))^n$	:	10^{20} calls of the oracle
computational complexity of the oracle	:	at least n arithmetic operations
total complexity	:	10^{21} arithmetic operations
CPU	:	1GHz or 10^9 arithmetic operations per second
total time	:	10^{12} seconds
one year	:	$\leq 3.2 \times 10^7$ seconds
we need	:	≥ 10000 years

- If we change n by $n + 1$, the # of calls of the oracle is multiplied by 100.
- If we multiply ε by 2, the arithmetic complexity is reduced by 1000.

We know from Corollary 5.5 that the number of iterations of the uniform grid method is at least $(\lfloor L/(2\varepsilon) \rfloor + 2)^n$. Theorem 5.6 showed that any method which uses only function evaluations requires at least $(\lfloor L/(2\varepsilon) \rfloor)^n$ calls to have a better performance than ε. If for instance we take $\varepsilon = \mathcal{O}(L/n)$, these two bounds coincide up to a constant factor. In this sense, the uniform grid method is an *optimal method for the class of problems $\mathcal{P}$.*

5.3 Steepest Descent Method

Consider $f : \mathbb{R}^n \to \mathbb{R}$ a differentiable function in its domain.

Steepest Descent Method	
Choose:	$\boldsymbol{x}_0 \in \mathbb{R}^n$
Iterate:	$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - h_k f'(\boldsymbol{x}_k),\ \ k = 0, 1, \ldots$

We consider four strategies for the step-size h_k:

1. **Constant Step**

   The sequence $\{h_k\}_{k=0}^\infty$ is chosen in *advance*. For example

$$h_k := h > 0,$$

$$h_k := \frac{h}{\sqrt{k+1}}.$$

This is the simplest strategy.

2. **Exact Line Search (Cauchy Step-Size)**

   The sequence $\{h_k\}_{k=0}^{\infty}$ is chosen such that

$$h_k := \arg\min_{h \geq 0} f(\boldsymbol{x}_k - hf'(\boldsymbol{x}_k)).$$

   This choice is only theoretical since even for the one dimensional case, it is very difficult and expensive.

3. **Goldstein-Armijo Rule**

   Find a sequence $\{h_k\}_{k=0}^{\infty}$ such that

$$\begin{array}{rcl} \alpha\langle f'(\boldsymbol{x}_k), \boldsymbol{x}_k - \boldsymbol{x}_{k+1}\rangle & \leq & f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k+1}), \\ \beta\langle f'(\boldsymbol{x}_k), \boldsymbol{x}_k - \boldsymbol{x}_{k+1}\rangle & \geq & f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k+1}), \end{array}$$

   where $0 < \alpha < \beta < 1$ are fixed parameters.

   Since $f(\boldsymbol{x}_{k+1}) = f(\boldsymbol{x}_k - h_k f'(\boldsymbol{x}_k))$,

$$f(\boldsymbol{x}_k) - \beta h_k \|f'(\boldsymbol{x}_k)\|_2^2 \leq f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{x}_k) - \alpha h_k \|f'(\boldsymbol{x}_k)\|_2^2.$$

   The acceptable steps exist unless $f(\boldsymbol{x}_{k+1}) = f(\boldsymbol{x}_k - hf'(\boldsymbol{x}_k))$ is not bounded from below.

4. **Barzilai-Borwein Step-Size**[1]

   Let us define $\boldsymbol{s}_{k-1} := \boldsymbol{x}_k - \boldsymbol{x}_{k-1}$ and $\boldsymbol{y}_{k-1} := f'(\boldsymbol{x}_k) - f'(\boldsymbol{x}_{k-1})$. Then, we can define the Barzilai-Borwein (BB) step sizes $\{h_k^1\}_{k=1}^{\infty}$ and $\{h_k^2\}_{k=1}^{\infty}$:

$$h_k^1 := \frac{\|\boldsymbol{s}_{k-1}\|_2^2}{\langle \boldsymbol{s}_{k-1}, \boldsymbol{y}_{k-1}\rangle},$$

$$h_k^2 := \frac{\langle \boldsymbol{s}_{k-1}, \boldsymbol{y}_{k-1}\rangle}{\|\boldsymbol{y}_{k-1}\|_2^2}.$$

   The first step-size is the one which minimizes the following secant condition $\|\frac{1}{h}\boldsymbol{s}_{k-1} - \boldsymbol{y}_{k-1}\|_2^2$ while the second one minimizes $\|\boldsymbol{s}_{k-1} - h\boldsymbol{y}_{k-1}\|_2^2$.

Now, consider the problem

$$\boxed{\begin{array}{c} \min_{\boldsymbol{x} \in \mathbb{R}^n} f(\boldsymbol{x}) \\[1em] \text{where } f \in \mathcal{C}_L^{1,1}(\mathbb{R}^n), \text{ and } f(\boldsymbol{x}) \text{ is bounded from below.} \end{array}}$$

[1] J. Barzilai and J. M. Borwein, "Two-point step size gradient methods," *IMA Journal of Numerical Analysis*, **8** (1988), pp. 141–148.