Foundation of Computing and Mathematical Sciences — Optimization —

Tokyo Institute of Technology
Dept. Mathematical and Computing Sciences
Mituhiro Fukuda

Fall/Winter Semester of 2012

"In our opinion, the main fact, which should be known to any person dealing with optimization models, is that in general *optimization problems are unsolvable*."

— Yurii Nesterov

Bibliography

[DASPREMONT2008] A. d'Aspremont, "Smooth optimization with approximate gradient", *SIAM Journal on Optimization* **19** (2008), pp. 1171–1183.

[GK2008] C. C. Gonzaga and E. W. Karas, "Fine tuning Nesterov's steepest descent algorithm for differentiable convex programming", *Mathematical Programming*, to appear.

[LLM2006] G. Lan, Z. Lu, and R. D. C. Monteiro, "Primal-dual first-order methods with $\mathcal{O}(1/\varepsilon)$ iteration-complexity for cone programming", *Mathematical Programming*, **126** (2011), pp.1–29.

[NESTEROV2004] Yu. Nesterov, *Introductory Lecture on Convex Optimization: A Basic Course*, (Kluwer Academic Publishers, Boston, 2004).

[NESTEROV2005] Yu. Nesterov, "Smooth minimization of non-smooth functions", *Mathematical Programming* **103** (2005), pp. 127–152.

[NESTEROV2005-2] Yu. Nesterov, "Excessive gap technique in nonsmooth convex minimization", *SIAM Journal on Optimization* **16** (2005), pp. 669–700.

[NESTEROV2007] Yu. Nesterov, "Smoothing technique and its applications in semidefinite optimization", *Mathematical Programming* **110** (2007), pp. 245–259.

[NESTERVO2009] Yu. Nesterov, "Primal-dual subgradient methods for convex problems", *Mathematical Programming* **120** (2009), pp. 221-259.

[NOCEDAL2006] J. Nocedal and S. J. Wright, *Numerical Optimization*, 2nd edition, (Springer, New York, 2006).

[TSENG2010] P. Tseng, "Approximation accuracy, gradient methods, and error bound for structured convex optimization", *Mathematical Programming* **12** (2010), pp. 263–295.

[YUAN2010] Y.-X. Yuan, "A short note on the Q-linear convergence of the steepest descent method", *Mathematical Programming* **123** (2010), pp. 339–343.

Chapter 1

Nonlinear Optimization

1.1 General minimization problem and terminologies

Definition 1.1.1 We define the *general minimization problem* as follows

$$\begin{cases} \text{minimize} & f(\boldsymbol{x}) \\ \text{subject to} & f_j(\boldsymbol{x}) \ \& \ 0, \quad j = 1, 2, \ldots, m \\ & \boldsymbol{x} \in S, \end{cases} \tag{1.1}$$

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 1.1.2 The *feasible set* Q of (1.1) 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 $Q \equiv \mathbb{R}^n$, (1.1) is a *unconstrained optimization problem.*
- If $Q \subsetneq \mathbb{R}^n$, (1.1) is a *constrained optimization problem.*
- If all functionals $f(\boldsymbol{x}), f_j(\boldsymbol{x})$ are differentiable, (1.1) is a *smooth optimization problem.*
- If one of functionals $f(\boldsymbol{x}),\ f_j(\boldsymbol{x})$ is non-differentiable, (1.1) is a *non-smooth optimization problem.*
- If all constraints are linear $f_j(\boldsymbol{x}) = \sum_{i=1}^n [\boldsymbol{a}]_{ji}[\boldsymbol{x}]_i + [\boldsymbol{b}]_j$ $(j = 1, 2, \ldots, m)$, (1.1) is a *linear constrained optimization problem.*
  - In addition, if $f(\boldsymbol{x})$ is linear, (1.1) is a *linear programming problem.*
  - In addition, if $f(\boldsymbol{x})$ is quadratic, (1.1) is a *quadratic programming problem.*
- If $f(\boldsymbol{x}),\ f_j(\boldsymbol{x})$ $(j = 1, 2, \ldots, m)$ are quadratic, (1.1) is a *quadratically constrained quadratic programming problem.*

Definition 1.1.3 $\boldsymbol{x}^*$ is called a *global optimal solution* of (1.1) if $f(\boldsymbol{x}^*) \le f(\boldsymbol{x}), \quad \forall \boldsymbol{x} \in Q$. Moreover, $f(\boldsymbol{x}^*)$ is called the *global optimal value*. $\boldsymbol{x}^*$ is called a *local optimal solution* of (1.1) if there exists an open ball $B(\varepsilon) = \{\boldsymbol{x} \in \mathbb{R}^n \mid \|\boldsymbol{x} - \boldsymbol{x}^*\| < \varepsilon\} \subseteq Q$ such that $f(\boldsymbol{x}^*) \le f(\boldsymbol{x}), \quad \forall \boldsymbol{x} \in B(\varepsilon)$. Moreover, $f(\boldsymbol{x}^*)$ is called a *local optimal value*.

General Iterative Scheme
Input: A starting point $\boldsymbol{x}_0$ and an accuracy $\varepsilon > 0$. **Initialization:** Set the *iteration counter* $k := 0$, and the *information set* $I_{-1} := \emptyset$.
MAIN LOOP **1.** Call oracle $\mathcal{O}$ at $\boldsymbol{x}_k$. **2.** Update the information set: $I_k := I_{k-1} \cup (\boldsymbol{x}_k, \mathcal{O}(\boldsymbol{x}_k))$. **3.** Apply the rules of the *method* $\mathcal{M}$ to I_k and compute $\boldsymbol{x}_{k+1}$. **4.** Check *stopping criterion* $\mathcal{T}_\varepsilon$. If **Yes**, **output** $\bar{\boldsymbol{x}}$. **Otherwise** set $k := k + 1$ and go to Step **1.**

Definition 1.1.4 The *analytical complexity* of a method is the number of calls of an oracle which is required to solve a problem $\mathcal{P}$ up to the given accuracy ε.

Definition 1.1.5 The *arithmetical complexity* of a method is the total number of arithmetic operations (including the work of the oracle and the method) which is required to solve a problem $\mathcal{P}$ up to the given accuracy ε.

Assumption 1.1.6 (Local black box)

1. The only information available for the numerical scheme is the answer of the oracle.
2. The oracle is local, that is, a small variation of the problem far enough from the test point $\boldsymbol{x}$ does not change the answer at $\boldsymbol{x}$.

Definition 1.1.7

1. The *zero-order oracle* returns the value $f(\boldsymbol{x})$.
2. The *first-order oracle* returns the value $f(\boldsymbol{x})$, and the gradient $f'(\boldsymbol{x})$.
3. The *second-order oracle* returns the value $f(\boldsymbol{x})$, $f'(\boldsymbol{x})$ and the Hessian $f''(\boldsymbol{x})$.

1.2 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.

$$\begin{cases} \text{minimize} & f(\boldsymbol{x}) \\ \text{subject to} & \boldsymbol{x} \in B_n = \{\boldsymbol{x} \in \mathbb{R}^n \mid 0 \le [\boldsymbol{x}]_i \le 1,\ i = 1, 2, \ldots, n\}. \end{cases} \tag{1.2}$$

To be coherent, we use the ℓ_∞-norm:

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

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

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

Let us define a very simple method to solve (1.2), 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 1.2.1 Let f^* be the global optimal value for (1.2). Then the uniform grid method yields

$$f(\bar{\boldsymbol{x}}) - f^* \le \frac{L}{2p}.$$

Proof: Let $\boldsymbol{x}^*$ be a global optimal solution. Then there are coordinates $(i_1, i_2, \ldots, i_n)$ such that $\boldsymbol{x} \equiv \boldsymbol{x}_{i_1,i_2,\ldots,i_n} \le \boldsymbol{x}^* \le \boldsymbol{x}_{i_1+1,i_2+1,\ldots,i_n+1} \equiv \boldsymbol{y}$. 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:

$$[\tilde{\boldsymbol{x}}]_i = \begin{cases} [\boldsymbol{y}]_i, & \text{if } [\boldsymbol{x}^*]_i \ge [\hat{\boldsymbol{x}}]_i \\ [\boldsymbol{x}]_i, & \text{otherwise.} \end{cases}$$

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

$$f(\bar{\boldsymbol{x}}) - f(\boldsymbol{x}^*) \le f(\tilde{\boldsymbol{x}}) - f(\boldsymbol{x}^*) \le L\|\tilde{\boldsymbol{x}} - \boldsymbol{x}^*\|_\infty \le L/(2p).$$

∎

Let us define our goal

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

Corollary 1.2.2 The analytical complexity of the problem (1.2) 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 $\mathcal{C}$ defined as follows:

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

Theorem 1.2.3 For $\varepsilon < \frac{L}{2}$, the analytical complexity of class of problems $\mathcal{C}$ using zero-order methods is 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 $\mathcal{P}$.

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

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 1.2.3 supports our initial claim that the *general optimization problem are unsolvable.*

Example 1.2.4 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
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 analytical complexity estimate is multiplied by 100.
- If we multiply ε by 2, the arithmetic complexity is reduced by 1000.

We know from Corollary 1.2.2 that the analytical complexity for the uniform grid method is $(\lfloor L/(2\varepsilon)\rfloor + 2)^n$. Theorem 1.2.3 showed that any method with zero-order oracle requires at least $(\lfloor L/(2\varepsilon)\rfloor)^n$ calls to have a better performance that 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* $\mathcal{C}$.

1.3 Optimality conditions for unconstrained minimization problems

Let $f(\boldsymbol{x})$ be differentiable at $\bar{\boldsymbol{x}}$. Then for $\boldsymbol{y} \in \mathbb{R}^n$, we have

$$f(\boldsymbol{y}) = f(\bar{\boldsymbol{x}}) + \langle f'(\bar{\boldsymbol{x}}), \boldsymbol{y} - \bar{\boldsymbol{x}} \rangle + o(\|\boldsymbol{y} - \bar{\boldsymbol{x}}\|),$$

where $o(r)$ is some function of $r > 0$ such that

$$\lim_{r \to 0} \frac{1}{r} o(r) = 0, \;\; o(0) = 0.$$

Let $\boldsymbol{s}$ be a direction in $\mathbb{R}^n$ such that $\|\boldsymbol{s}\| = 1$. Consider the local decrease of $f(\boldsymbol{x})$ along $\boldsymbol{s}$:

$$\Delta(\boldsymbol{s}) = \lim_{\alpha \to 0} \frac{1}{\alpha} \left[f(\bar{\boldsymbol{x}} + \alpha \boldsymbol{s}) - f(\bar{\boldsymbol{x}}) \right].$$

Since $f(\bar{\boldsymbol{x}} + \alpha \boldsymbol{s}) - f(\bar{\boldsymbol{x}}) = \alpha \langle f'(\bar{\boldsymbol{x}}), \boldsymbol{s} \rangle + o(\|\alpha \boldsymbol{s}\|)$, we have $\Delta(\boldsymbol{s}) = \langle f'(\bar{\boldsymbol{x}}), \boldsymbol{s} \rangle$.
Using the Cauchy-Schwartz inequality $-\|\boldsymbol{x}\|\|\boldsymbol{y}\| \leq \langle \boldsymbol{x}, \boldsymbol{y} \rangle \leq \|\boldsymbol{x}\|\|\boldsymbol{y}\|$,

$$\Delta(\boldsymbol{s}) = \langle f'(\bar{\boldsymbol{x}}), \boldsymbol{s} \rangle \geq -\|f'(\bar{\boldsymbol{x}})\|.$$

Choosing the direction $\bar{\boldsymbol{s}} = -f'(\bar{\boldsymbol{x}})/\|f'(\bar{\boldsymbol{x}})\|$,

$$\Delta(\bar{\boldsymbol{s}}) = -\left\langle f'(\bar{\boldsymbol{x}}), \frac{f'(\bar{\boldsymbol{x}})}{\|f'(\bar{\boldsymbol{x}})\|} \right\rangle = -\|f'(\bar{\boldsymbol{x}})\|.$$

Thus, the direction $-f'(\bar{\boldsymbol{x}})$ is the direction of the *fastest local decrease of* $f(\boldsymbol{x})$ *at point* $\bar{\boldsymbol{x}}$.

Theorem 1.3.1 (First-order necessary optimality condition) Let $\boldsymbol{x}^*$ be a local minimum of the differentiable function $f(\boldsymbol{x})$. Then

$$f'(\boldsymbol{x}^*) = \boldsymbol{0}.$$

Proof: Let $\boldsymbol{x}^*$ be the local minimum of $f(\boldsymbol{x})$. Then, there is $r > 0$ such that for all $\boldsymbol{y}$ with $\|\boldsymbol{y} - \boldsymbol{x}^*\| \leq r$, $f(\boldsymbol{y}) \geq f(\boldsymbol{x}^*)$.
Since f is differentiable,

$$f(\boldsymbol{y}) = f(\boldsymbol{x}^*) + \langle f'(\boldsymbol{x}^*), \boldsymbol{y} - \boldsymbol{x}^* \rangle + o(\|\boldsymbol{y} - \boldsymbol{x}^*\|) \geq f(\boldsymbol{x}^*).$$

Dividing by $\|\boldsymbol{y} - \boldsymbol{x}^*\|$, and taking the limit $\boldsymbol{y} \to \boldsymbol{x}^*$,

$$\langle f'(\boldsymbol{x}^*), \boldsymbol{s} \rangle \geq 0, \quad \forall \boldsymbol{s}, \quad \|\boldsymbol{s}\| = 1.$$

Consider the opposite direction $-\boldsymbol{s}$, and then we conclude that

$$\langle f'(\boldsymbol{x}^*), \boldsymbol{s}\rangle = 0, \quad \forall \boldsymbol{s}, \quad \|\boldsymbol{s}\| = 1.$$

Choosing $\boldsymbol{s} = \boldsymbol{e}_i \quad (i = 1, 2, \ldots, n)$, we conclude that $f'(\boldsymbol{x}^*) = 0$. ∎

Corollary 1.3.2 Let $\boldsymbol{x}^*$ be a local minimum of a differentiable function $f(\boldsymbol{x})$ subject to linear equality constraints

$$\boldsymbol{x} \in \mathcal{L} \equiv \{\boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}\} \neq \emptyset,$$

where $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, $\boldsymbol{b} \in \mathbb{R}^m$, $m < n$.

Then, there exists a vector of multipliers $\boldsymbol{\lambda}^*$ such that

$$f'(\boldsymbol{x}^*) = \boldsymbol{A}^T\boldsymbol{\lambda}^*.$$

Proof: Consider the vectors $\boldsymbol{u}_i$ $(i = 1, 2, \ldots, k)$ with $k \geq n - m$ which form an orthonormal basis of the null space of $\boldsymbol{A}$. Then, $\boldsymbol{x} \in \mathcal{L}$ can be represented as

$$\boldsymbol{x} = \boldsymbol{x}(\boldsymbol{t}) \equiv \boldsymbol{x}^* + \sum_{i=1}^{k} t_i\boldsymbol{u}_i, \quad \boldsymbol{t} \in \mathbb{R}^k.$$

Moreover, the point $\boldsymbol{t} = \boldsymbol{0}$ is the local minimal solution of the function $\phi(\boldsymbol{t}) = f(\boldsymbol{x}(\boldsymbol{t}))$.

From Theorem 1.3.1, $\phi'(\boldsymbol{0}) = \boldsymbol{0}$. That is,

$$\frac{d\phi}{dt_i}(\boldsymbol{0}) = \langle f'(\boldsymbol{x}^*), \boldsymbol{u}_i\rangle = 0, \quad i = 1, 2, \ldots, k.$$

Now there is $\boldsymbol{t}^*$ and $\boldsymbol{\lambda}^*$ such that

$$f'(\boldsymbol{x}^*) = \sum_{i=1}^{k} t_i^*\boldsymbol{u}_i + \boldsymbol{A}^T\boldsymbol{\lambda}^*.$$

For each $i = 1, 2, \ldots, k$,

$$\langle f'(\boldsymbol{x}^*), \boldsymbol{u}_i\rangle = t_i^* = 0.$$

Therefore, we have the result. ∎

If $f(\boldsymbol{x})$ is twice differentiable at $\bar{\boldsymbol{x}}$, then for $\boldsymbol{y} \in \mathbb{R}^n$, we have

$$f'(\boldsymbol{y}) = f'(\bar{\boldsymbol{x}}) + f''(\bar{\boldsymbol{x}})(\boldsymbol{y} - \bar{\boldsymbol{x}}) + \boldsymbol{o}(\|\boldsymbol{y} - \bar{\boldsymbol{x}}\|),$$

where $\boldsymbol{o}(r)$ is such that $\lim_{r\to 0} \|\boldsymbol{o}(r)\|/r = 0$ and $\boldsymbol{o}(0) = 0$.

Theorem 1.3.3 (Second-order necessary optimality condition) Let $\boldsymbol{x}^*$ be a local minimum of a twice continuously differentiable function $f(\boldsymbol{x})$. Then

$$f'(\boldsymbol{x}^*) = 0, \qquad f''(\boldsymbol{x}^*) \succeq \boldsymbol{O}.$$