Contents

I Introduction

1. General Formulation of the Optimization Problem

Let x be an n-dimensional real vector

$$\boldsymbol{x} = \left(x^{(1)}, \ldots, x^{(n)}\right)^T \in \mathbb{R}^n$$

and $f_0(x), \ldots, f_m(x)$ be some **real-valued** function defined on a set $S \subseteq \mathbb{R}^n$. We consider different variants of the following general minimization problem:

$$\begin{aligned} &\min f_0(\boldsymbol{x}) \\ &\text{s.t. } f_j(\boldsymbol{x}) \& 0, j = 1 \ldots m \\ &\boldsymbol{x} \in S \end{aligned} \tag{I.1}$$

where the sign & can be $\leq, \geq$ or $=$.

& 表示受约束. 可用 $\leq$ 表示 $\geq$, 然后用 $\leq, \geq$ 表示 $=$.

We call $f_0(x)$ the **objective function** (目标函数) of our problem, the vector function

$$f(\boldsymbol{x}) = (f_1(x), \ldots, f_m(x))^T$$

is called the vector of **functional constraints** (泛函约束), the set S is the **basic feasible set** (基本可行集), and the set

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

is the **(entire) feasible set** (完整可行集) of problem I.1. It's just a convention to consider minimization problems.

1.1 Natural Classification

There exists a natural classification of the types of minimization problems:

- Constrained problems (带约束问题): $Q \subset \mathbb{R}^n$
- Unconstrained problems (无约束问题): $Q \equiv \mathbb{R}^n$
- smooth problems (光滑问题): all $f_j(x)$ are differentiable
- nonsmooth problems: some $f_k(x)$ are non-differentiable
- linearly constrained problems: the functional constraints are affine (仿射) like

$$f_j(\boldsymbol{x}) = \sum_{i=1}^{n} a_j^{(i)} \boldsymbol{x}^{(i)} + b_j \equiv \langle a_j, \boldsymbol{x} \rangle + b_j, \ j = 1, \ldots, m$$

  - linear optimization problem: $f_0(\cdot)$ is also affine
  - quadratic optimization problem: $f_0(\cdot)$ is quadratic
  - quadratically constrained problem: all the function $f_0(\cdot), \ldots, f_m(\cdot)$ are quadratic

1.2 Classification based on the properties of feasible set

There is also a classification based on properties of the feasible set.

- Problem I.1 is called **feasible**, if $Q \neq \emptyset$
- Problem I.1 is called **strictly feasible**, if there exists an $\boldsymbol{x} \in \text{int } Q$ such taht $f_j(\boldsymbol{x}) < 0$ (or > 0) for all inequality constraints and $f_j(\boldsymbol{x}) = 0$ for all equality constraints (Slater condition)

int Q 表示不包含边界点的 Q.

1.3 Different Types of Solution

Finally, we distinguish different types of solution to I.1:

- $\boldsymbol{x}^*$ is called the **global optimal solution** to I.1 if $f_0(\boldsymbol{x}^*) \leq f_0(x)$ for all $x \in Q$. In this case, $f_0(\boldsymbol{x}^*)$ is called the (global) **optimal value** of the problem. ($\boldsymbol{x}^* = \arg\min \cdots$, 是个集合)
- $\boldsymbol{x}^*$ is called a **local optimal solution** to I.1 if $\exists \delta$, for all $\boldsymbol{x} \in NBR(\boldsymbol{x}^*, \delta) \cap Q$ (NBR means Neighborhood), we have

$$f_0(\boldsymbol{x}^*) \leq f_0(\boldsymbol{x})$$

1.4 Example

Example 1 Let $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}$ be our **design variables**. Then we can fix some functional characteristics of our decision vector $\boldsymbol{x}$: $f_0(\boldsymbol{x}), \ldots, f_m(\boldsymbol{x})$. For example, we can consider a price of the project, amount of required resources, reliability of the system, etc. We fix the most important characteristics, $f_0(\boldsymbol{x})$, as our objective. For all others, we impose some bounds: $a_j \leq f_j(\boldsymbol{x}) \leq b_j$. Thus, we come to the problem:

$$\begin{aligned} &\min f_0(\boldsymbol{x}) \\ \text{s.t. } &a_j \leq f_j(\boldsymbol{x}) \leq b_j, j = 1 \ldots m \\ &\boldsymbol{x} \in S \end{aligned}$$

where S stands for the **structural** constraints like non-negativity, boundedness of some variables, etc.

Example 2 Let our initial problem be as follow:

$$\text{Find } \boldsymbol{x} \in \mathbb{R}^n \text{ such that } f_j(\boldsymbol{x}) = a_j, j = 1, \ldots, m \qquad \text{(I.2)}$$

Then we consider the problem

$$\min_{\boldsymbol{x}} \sum_{j=1}^{m} (f_j(\boldsymbol{x}) - a_j)^2$$

perhaps evne with some additional constraints on $\boldsymbol{x}$. If the optimal value of the latter problem is zero, we conclude that our initial problem I.2 has a solution.

Example 3 Sometimes our decision variables $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}$ must be integers. This can be described by the following constraint:

$$\sin(\pi \boldsymbol{x}^{(i)}) = 0, i = 1, \ldots, n$$

Thus, we can also treat integer optimization problems:

$$\begin{aligned} &\min f_0(\boldsymbol{x}) \\ \text{s.t. } &a_j \leq f_j(\boldsymbol{x}) \leq b_j, j = 1, \ldots, m \\ &\boldsymbol{x} \in S \\ &\sin(\pi \boldsymbol{x}^{(i)}) = 0, i = 1, \ldots, n \end{aligned}$$

2. Interplay of Optimization and Machine Learning

2.1 The Machine Learning (Supervised) Paradigm

Consider the following **spaces**: a space of *example* $\mathcal{X}$, a space of **labels** $\mathcal{Y}$, and a space of *hypothesis* $\mathcal{H}$ that contains functions mapping $\mathcal{X}$ to $\mathcal{Y}$. Also consider a **loss** function $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ and a **probability measure** P over the space $\mathcal{X} \times \mathcal{Y}$.

Definition I.1. *For any hypothesis $h \in \mathcal{H}$, we define the risk of h with respect to L and P to be:*

$$\mathcal{L}_P(h) = E_P(L(h(\boldsymbol{X}), Y)), \text{ with } (\boldsymbol{X}, Y) \sim P$$

The general problem of machine learning can be then cast as finding the hypothesis $h \in \mathcal{H}$ that solves the following optimization problems:

$$\min_{h \in \mathcal{H}} \mathcal{L}_P(h)$$

2.2 Empirical Risk Minimization

The practical way is to solve an *empirical risk minimization* problem. That is,

$$\min_{h \in \mathcal{H}} \sum_{i=1}^{m} L(h(\boldsymbol{x}_i), y_i) = \min_{\theta} \sum_{i=1}^{m} L(h_\theta(\boldsymbol{x}_i), y_i)$$

since we do not know the P exactly. That is

$$\theta_* = \operatorname*{argmin}_{\theta} \sum_{i=1}^{m} L(h_\theta(\boldsymbol{x}_i), y_i)$$

- Classification: the output variable takes **class labels**. Especially, for **binary classification**, $y_i \in \{1, -1\}$.
- Regression: the output variable takes **continuous values**. That is $y_i \in \mathbb{R}$.

2.3 Classification: The $0 \sim 1$ Loss

By using the $0 \sim 1$ *loss*, we can rewrite the objective function as:

$$\sum_{i=1}^{m} \mathbf{1}_{\text{sign}(h(\boldsymbol{x}_i)) \neq y_i} = \sum_{i=1}^{m} \mathbf{1}_{y_i h(\boldsymbol{x}_i) < 0}$$

where $\mathbf{1}_z$ is just the indicator function, that is

$$\mathbf{1}_z = \begin{cases} 1 & z \text{ is true} \\ 0 & \text{otherwise} \end{cases}$$

and $\text{sign}(x)$ is sign function,

$$\text{sign}(x) = \begin{cases} 1 & x > 0 \\ 0 & x = 0 \\ -1 & x < 0 \end{cases}$$

In fact, this problem is **NP-hard** so it's pointless to try to minimize this function.

The Surrogate of The $0 \sim 1$ Loss To deal with this problem, we can introduce some *surrogate*.

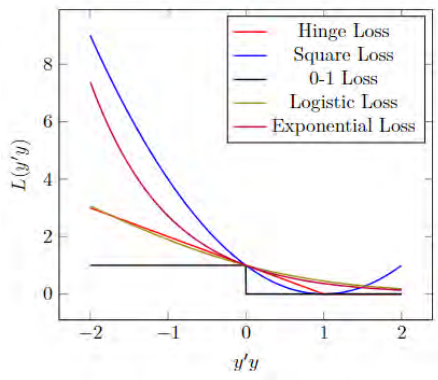

Figure I.1: The Surrogate of The $0 \sim 1$ Loss

Some common loss function

- hinge loss or margin loss

$$\phi_h(y, y') = \max(0, 1 - y'y)$$

- Square loss:

$$\phi_s(y, y') = (1 - yy')^2$$

- Logistic loss:

$$\phi_l(y, y') = \frac{1}{\ln 2} \ln(1 + e^{-yy'})$$

- Exponential loss:

$$\phi_e(y, y') = e^{-yy'}$$

- Cross entropy loss:

$$\phi_c(y, y') = -t \ln(y') - (1 - t) \ln(1 - y')$$

where $t = \frac{(1+y)}{2}$

2.4 Maximum Likelihood Estimation View

Let's say, we have a likelihood function $P(\boldsymbol{X}, Y|\theta)$. Then the MLE for θ, the parameter we want to infer, is

$$\begin{aligned} \theta_{MLE} &= \underset{\theta}{\text{argmax}}\, P(\boldsymbol{X}, Y|\theta) \\ &= \underset{\theta}{\text{argmax}} \log P(\boldsymbol{X}, Y|\theta) \\ &= \underset{\theta}{\text{argmax}} \log \prod_i P(x_i, y_i|\theta) \\ &= \underset{\theta}{\text{argmax}} \sum_{i=1}^{n} \log P(x_i, y_i|\theta) \end{aligned}$$

For case that $P(x_i, y_i|\theta) \sim \exp(-L(h_\theta(\boldsymbol{x}_i), y_i))$ ($\sim$ 表示成比例), we have

$$\theta_{MLE} = \underset{\theta}{\text{argmin}} \sum_{i=1}^{n} L(h_\theta(\boldsymbol{x}_i), y_i)$$

2.5 Maximum A posteriori Estimation View

If we replace the likelihood in the MLE formula above with the posterior, we get:

$$\begin{aligned} \theta_{MAP} &= \underset{\theta}{\text{argmax}}\, P(\boldsymbol{X}, Y|\theta) P(\theta) \\ &= \underset{\theta}{\text{argmax}} \log\{P(\theta) P(\boldsymbol{X}, Y|\theta)\} \\ &= \underset{\theta}{\text{argmax}} \log \left\{ P(\theta) \prod_i P(\boldsymbol{x}_i, y_i|\theta) \right\} \\ &= \underset{\theta}{\text{argmax}} \left\{ \log P(\theta) + \sum_{i=1}^{n} \log P(\boldsymbol{x}_i, y_i|\theta) \right\} \end{aligned}$$

For case that $P(x_i, y_i|\theta) \sim \exp(-L(h_\theta(\boldsymbol{x}_i), y_i))$ and $P(\theta) \sim \exp(-\lambda\|\theta\|_2^2)$, we have

$$\theta_{MAP} = \underset{\theta}{\text{argmax}} \left\{ \sum_{i=1}^{n} L(h_\theta(\boldsymbol{x}_i), y_i) + \lambda\|\theta\|_2^2 \right\}$$

$P(\theta)$ 作为先验, 推导出正则项. 正则化可以让函数凸性更好, 更容易找到最小值.

3. Performance of Numerical Methods

3.1 Method and Class of Problem

考虑一种解决问题 I.1 的方法：除了 $x^* = 0$ 之外, 什么都不做. 对于那些在原点恰好有最优解的问题, 这种方法的 "性

能" 是无法超越的. 然而, 这种方法对于其他任何问题都无法正常工作. 因此, 我们不能说某个特定问题 P 的最佳方法是什么, 但我们可以针对一类问题 $\mathcal{F} \ni P$ 来讨论.

一个方法 $\mathcal{M}$ 在整个类别 $\mathcal{P}$ 上的性能可以自然地衡量其效率. 考虑 $\mathcal{M}$ 在类别 $\mathcal{P}$ 上的性能. 我们应该假设方法 $\mathcal{M}$ 并没有关于特定问题 $\mathcal{P}$ 的完整信息.

- **Model:** 已知 (对数值方案)"部分" 问题 $\mathcal{P}$ 被称为问题的模型. 我们用 Σ 表示模型. 通常, 模型包括问题的表述、功能组件类别的描述等.
- **Oracle:** 我们通过预言的概念描述收集关于 $\mathcal{P}$ 的特定信息的过程. 预言 $\mathcal{O}$ 只是一个回答方法连续问题的单位.

一般来说, 每个问题都可以用不同的模型来描述. 此外, 对于每个问题, 我们都可以开发出不同类型的预言.

Let us fix Σ and $\mathcal{O}$. In this case, it's natural to define the performance of $\mathcal{M}$ on $(\Sigma, \mathcal{O})$, as its performance on the worst $\mathcal{P}_w$ from $(\Sigma, \mathcal{O})$. Note that this $\mathcal{P}_w$ can be bad only fo $\mathcal{M}$.

Definition I.2. *The performance of $\mathcal{M}$ on $\mathcal{P}$ is the total amount of computational effort required by method $\mathcal{M}$ to solve the problem $\mathcal{P}$.*

Solving the problem means finding an approximate solution on $\mathcal{P}$ with some accuracy $\epsilon > 0$ (找逼近解就是解问题了).

3.2 General Iterative Scheme

- Input: Starting point x_0 and accuracy $\epsilon > 0$
- Initialization: Set $k = 0, I_{-1} = \emptyset$

Main Loop:

1) Call oracle $\mathcal{O}$ at point x_k
2) Update the information set: $I_k = I_{k-1} \cup (x_k, \mathcal{O}(x_k))$
3) Apply the rules of method $\mathcal{M}$ to I_k and generate a new point x_{k+1}
4) Check criterion $\mathcal{T}_\epsilon$. If **yes** then form an output $\bar{\boldsymbol{x}}$. Otherwise set $k := k + 1$ and go to Step 1.

3.3 Complexity

In above scheme, we can see two potentially expensive steps. The first one is Step 1, where we call the oracle. The second one is Step 3, where we form the new test point. Thus, we can introduce two measures of complexity of problem $\mathcal{P}$ for method $\mathcal{M}$:

- Analytical Complexity: The number of calls of the oracle which is necessary to solve problem $\mathcal{P}$ up to accuracy ϵ.
- Arithmetical Complexity: The total number of arithmetic operations (including the work of both oracle and method), which is necessary to solve problem $\mathcal{P}$ up to accuracy ϵ.

For a particular method $\mathcal{M}$ as applied to problem $\mathcal{P}$, arithmetical complexity can be easily obtained from the analytical complexity and complexity of the oracle.

3.4 Oracle

There is one standard assumption on the oracle which allows us to obtain the majority of results on analytical complexity for optimization schemes. This assumption, called the **Local Black Box Concept**, is listed in the following lines:

1) The only information available for the numerical scheme is the answer of the oracle.
2) The oracle is local: A small variation of the problem far enough from the test point x, which is compatible with the description of the problem class, does not change the answer at x

The standard formulation I.1 is called a functional model of optimization problems. Usually, for such models the standard assumptions are related to the level of smoothness of functional components. According to the degree of smoothness we can apply different types of oracle:

- zero-order oracle: return $f(x)$
- first-order oracle: return $f(x)$ and gradient $\nabla f(x)$
- second-order oracle: return $f(x)$, gradient $\nabla f(x)$ and Hessian $\nabla^2 f(x)$

4. Complexity Bounds for Global Optimization

4.1 N-dimensional Box Constraint Problem

Consider the following problem:

$$\min_{x \in \mathbb{B}_n} f(x) \tag{I.3}$$

In our terminology, this is a constrained minimization problem with no functional constraints. The basic feasible set of this problem is $\mathbb{B}_n$, an n-dimensional box in $\mathbb{R}^n$:

$$\mathbb{B}_n = \{\boldsymbol{x} \in \mathbb{R}^n \mid 0 \le \boldsymbol{x}^{(i)} \le 1,\ i = 1, \dots, n\};$$

Let us measure distances in $\mathbb{R}^n$ by the ℓ_∞:

$$\|\boldsymbol{x}\|_\infty = \max_{1 \le i \le n} |\boldsymbol{x}^{(i)}|$$

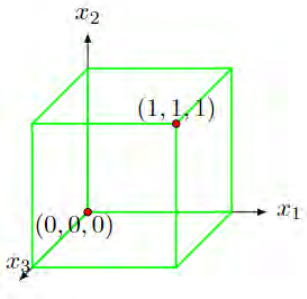

Figure I.2: N-dimensional Box Constraint Problem

Assume that, with respect to this norm, the objective function $f(\boldsymbol{x}) : \mathbb{R}^n \to \mathbb{R}$ is **Lipschitz continuous** on $\mathbb{B}_n$:

$$|f(\boldsymbol{x}) - f(\boldsymbol{y})| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|_\infty, \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{B}_n \tag{I.4}$$

with some constant L (**Lipschitz constant**).

4.2 Uniform Grid Method

Method $\mathcal{G}(p)$:

1) Form $(p+1)^n$ points

$$\boldsymbol{x}_{(i_1,\ldots,i_n)} = \left(\frac{i_1}{p}, \frac{i_2}{p}, \ldots, \frac{i_n}{p}\right)^T$$

where $(i_1, \ldots, i_n) \in \{0, \ldots, p\}^n$.

2) Among all points $\boldsymbol{x}_{(i_1,\ldots,i_n)}$, find the point $\bar{\boldsymbol{x}}$ with the minimal value of the objective function.

3) The pair $(\bar{\boldsymbol{x}}, f(\bar{\boldsymbol{x}}))$ is the output of the method.

This method forms a **uniform grid** of the test points inside the box $\mathbb{B}_n$, computes the best value of the objective over this grid, and returns this value as an approximate solution to problem I.3.

- Zeroorder iterative method,
- Without any influence from the accumulated information on the sequence of test points.

Theorem I.3. *Let f^* be a global optimal value of problem I.3. Then*

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

Proof. For a multiindex $(i_1, i_2, \ldots, i_n)$, let $\boldsymbol{x}_*$ be a global minimum of our problem, Then there exist coordinates $(i_1, i_2, \ldots, i_n)$ such that

$$\boldsymbol{x} \equiv \boldsymbol{x}_{(i_1,i_2,\ldots,i_n)} \leq \boldsymbol{x}_* \leq \boldsymbol{x}_{(i_1+1,i_2+1,\ldots,i_n+1)} \equiv \boldsymbol{y}$$

Note that

$$\boldsymbol{y}^{(i)} - \boldsymbol{x}^{(i)} = \frac{1}{p}$$

for $i = 1, \ldots, n$ and

$$\boldsymbol{x}_*^{(i)} \in [\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}], i = 1, \ldots, n$$

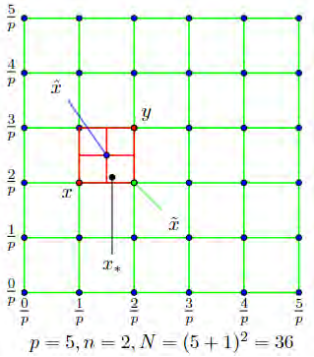

Figure I.3: Uniform Grid Method

Denote $\hat{\boldsymbol{x}} = (x + y)/2$. Let's form a point $\tilde{\boldsymbol{x}}$ as follows:

$$\tilde{\boldsymbol{x}} = \begin{cases} \boldsymbol{y}^{(i)} & \text{if } \boldsymbol{x}_*^{(i)} \geq \hat{\boldsymbol{x}}^{(i)} \\ \boldsymbol{x}^{(i)} & \text{otherwise} \end{cases}$$

It's clear that $|\tilde{\boldsymbol{x}}^{(i)} - \boldsymbol{x}_*^{(i)}| \leq \frac{1}{2p}, i = 1, \ldots, n$. Therefore

$$\|\tilde{\boldsymbol{x}} - \boldsymbol{x}_*\|_\infty = \max_{1 \leq i \leq n} |\tilde{\boldsymbol{x}}^{(i)} - \boldsymbol{x}_*^{(i)}| \leq \frac{1}{2p}$$

Since $\tilde{\boldsymbol{x}}$ belongs to our grid, we conclude that

$$f(\bar{\boldsymbol{x}}) - f(\boldsymbol{x}_*) \leq f(\tilde{\boldsymbol{x}}) - f(\boldsymbol{x}_*) \leq L\|\tilde{\boldsymbol{x}} - \boldsymbol{x}_*\|_\infty \leq \frac{L}{2p}$$

Q.E.D.

4.3 Upper Complexity Bound

Let us conclude with the definition of our problem class. We fix our goal as follows:

$$\text{Find } \bar{\boldsymbol{x}} \in \mathbb{B}_n : f(\bar{\boldsymbol{x}}) - f^* \leq \epsilon \tag{I.5}$$

Then we immediately get the following result.

Corollary I.4. *The analytical complexity of problem class I.3, I.4, I.5 for method $\mathcal{G}$ is at most*

$$\mathcal{A}(\mathcal{G}) = \left(\left\lfloor \frac{L}{2\epsilon} \right\rfloor + 2\right)^n$$

Proof. Take $p = \left\lfloor \frac{L}{2\epsilon} \right\rfloor + 1$. Then $p \geq \frac{L}{2\epsilon}$. and , in view of Theorem I.3, we have

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

Note that we construct $(p+1)^n$ points. Q.E.D.

Remark:

- Each point has an access to oracle, so the number of accessing is equal to the number of points.
- Thus, $\mathcal{A}(\mathcal{G})$ justifies an upper complexity bound for our problem class.

4.4 Lower Complexity Bound

我们需要进一步探讨两种情况：

- 首先, 可能我们的证明过于粗糙, 方法 $\mathcal{G}(p)$ 的实际性能要好得多.
- 其次, 我们仍然不能确定 $\mathcal{G}(p)$ 是否是解决 I.3 的合理方法. 可能存在其他性能更高的方案.

为了回答这些问题, 我们需要为问题类别 I.3, I.4, I.5 推导出更低的复杂性界限. 这些界限的主要特点如下：

- 它们基于黑箱概念.
- 这些界限对所有合理的迭代方案都有效. 因此, 它们为我们提供了问题类别的分析复杂性的下限估计.
- 这样的界限往往采用抵抗预言 (resisting oracle) 的思想.

一个抵抗预言试图为每一种特定方法 (例如, $\mathcal{G}(p)$) 创建最糟糕的问题.

- 它从一个 “空” 函数开始, 并试图以最糟糕的方式回答方法的每个调用.
- 然而, 答案必须与先前的答案和问题类别的描述相符.

在方法终止后, 可以重构一个完全符合算法累积的最终信息集的问题. 此外, 如果我们在这个新生问题上运行该方法, 由于它将得到预言相同的答案序列, 因此它将重现相同的测试点序列.

Let us show how this works for problem *I.3*. Consider the class of problems $\mathcal{C}$ defined as follows:

- Model: $\min_{\boldsymbol{x} \in \mathbb{B}_n} f(x)$, where $f(x)$ is ℓ_∞-Lipschitz continuous on $\mathbb{B}_n$.
- Oracle: Zero-order Local Black Box.
- Approximate solution: Find $\bar{\boldsymbol{x}} \in \mathbb{B}_n : f(\bar{\boldsymbol{x}} - f^*) \leq \epsilon$

Theorem I.5. *For $\epsilon < \frac{1}{2}L$, the analytical complexity of problem class $\mathcal{C}$ is at least $\left(\frac{L}{2\epsilon}\right)^n$.*

Proof. Denote $q = \left\lfloor \frac{L}{2\epsilon} \right\rfloor (\geq 1)$. Assume that there exists a method, which need $N < q^n$ calls of oracle to solve any problem from $\mathcal{C}$. Let us apply this method to the following resisting strategy:

$$\text{Oracle returns } f(\boldsymbol{x}) = 0 \text{ at any test point } \boldsymbol{x}$$

Therefore this method can find only $\bar{\boldsymbol{x}} \in \mathbb{R}^n$ with $f(\bar{\boldsymbol{x}}) = 0$. However, note that there exists $\hat{\boldsymbol{x}} \in \mathbb{B}_n$ such that

$$\hat{\boldsymbol{x}} + \frac{1}{q} e \in \mathbb{B}_n, e = (1, \ldots, 1)^T \in \mathbb{R}^n$$

and there were no test points inside the box $\mathbb{B} = \{\boldsymbol{x} | \hat{\boldsymbol{x}} \leq \boldsymbol{x} \leq \hat{\boldsymbol{x}} + \frac{1}{q} e\}$.
(“每个由 $\hat{\boldsymbol{x}}$ 构造的 $\mathbb{B}$ 至少包含一个点” 这个表述是错误的, 因为和假设不符合, 因此没有测试点的 $\mathbb{B}$ 一定存在)

Denotes $\boldsymbol{x}_* = \hat{\boldsymbol{x}} + \frac{1}{2q} e$. Consider the function

$$\bar{f}(\boldsymbol{x}) = \min\{0, L\|\boldsymbol{x} - \boldsymbol{x}_*\|_\infty - \epsilon\}$$

Clearly,

1) this function is ℓ_∞-Lipschitz continuous with the constant L and
2) its global optimal value is $-\epsilon$.

Moreover, $\bar{f}(x)$ differs from zero only inside the box $\mathbb{B}' = \{\boldsymbol{x} | \|\boldsymbol{x} - \boldsymbol{x}_*\|_\infty \leq \frac{\epsilon}{L}\}$. Since $2q \leq L/\epsilon$, we conclude that

$$\mathbb{B}' \subseteq \mathbb{B} \equiv \{\|\boldsymbol{x} - \boldsymbol{x}_*\|_\infty \leq \frac{1}{2q}\}$$

Thus, $\bar{f}(x)$ is equal to zero at all test points of our methods. Since the accuracy of the rest of our method is ϵ, we come to the following conclusion: If the number of calls of the oracle is less that q^n, then the accuracy of the result cannot be better that ϵ. Q.E.D.

Now we can say much more about the performance of the Uniform Grid Method.Let us compare its efficiency estimate with the lower bound:

$$\mathcal{G} : \left(\left\lfloor \frac{L}{2\epsilon} \right\rfloor + 2\right)^n, \text{ Lower Bound: } \left(\left\lfloor \frac{L}{2\epsilon} \right\rfloor\right)^n$$

If $\epsilon \leq O(\frac{L}{n})$, then the lower and upper bounds coincide up to an absolute constant multiplicative factor. This means that, for such level of accuracy, $\mathcal{G}(p)$ is optimal for the problem class $\mathcal{C}$.

Meanwhile, the Theorem I.5 supports our original statement that general optimization problems are unsolvable.

Remark: If $\epsilon \leq O(\frac{L}{n})$, we have

$$\frac{\left(\lfloor \frac{L}{2\epsilon} \rfloor + 2\right)^n}{\left(\lfloor \frac{L}{2\epsilon} \rfloor\right)^n} = \left(1 + \frac{1}{\frac{1}{2}\lfloor \frac{L}{2\epsilon} \rfloor}\right)^n \leq \left(1 + \frac{1}{cn}\right)^n = \left(\left(1 + \frac{1}{cn}\right)^{cn}\right)^{\frac{1}{c}} \underset{n\to\infty}{=} e^{\frac{1}{c}}$$

where c is a constant.

II Foundations of Smooth Optimization

1. Relaxation and Approximation

1.1 Concepts of Relaxation and Approximation

We call the sequence $\{a_k\}_{k=0}^{\infty}$ a **relaxation sequence** if $\forall k \geq 0$,

$$a_{k+1} \leq a_k$$

Consider several methods for solving the following unconstrained minimization problem

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x}) \tag{II.1}$$

where $f(\boldsymbol{x})$ is a smooth function. In order to do so, we generate a relaxation sequence

$$\{f(\boldsymbol{x})\}_{k=0}^{\infty},\ f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{x}_k),\ k = 0, 1, \ldots$$

This strategy has the following important advantages:

1) If $f(\boldsymbol{x})$ is bounded below on $\mathbb{R}^n$, then the sequence $\{f(\boldsymbol{x}_k)\}_{k=0}^{\infty}$ converges.

2) In any case we improve the initial value of the objective function.

In general, to **approximate** means to replace an initial complex object by a simplified one, which is close by its properties to the original.

In nonlinear optimization we usually apply local approximations based on derivatives(导数) of non-linear functions. These are the first order and the second order approximations (or, the linear and quadratic approximations).

1.2 First Order Approximation

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

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

where $o(r)$(peano, 皮亚诺余项) is some function of $r \geq 0$, such that

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

In the sequel we fix the notation $\|\cdot\|$ for the standard Euclidean norm on $\mathbb{R}^n$:

$$\|\boldsymbol{x}\| = \left[\sum_{i=1}^{n} \left(\boldsymbol{x}^{(i)}\right)^2\right]^{1/2}$$

The linear function $f(\bar{\boldsymbol{x}}) + \langle \nabla f(\bar{\boldsymbol{x}}, \boldsymbol{y} - \bar{\boldsymbol{x}}) \rangle$ is called the linear approximation of f at $\bar{\boldsymbol{x}}$.

- The vector $\nabla f(\boldsymbol{x})$ is called the **gradient** of function f at $\boldsymbol{x}$.

  Considering the points $\boldsymbol{y}_i = \bar{\boldsymbol{x}} + \epsilon e_i$, where e_i is the i-th coordinate vector in $\mathbb{R}^n$, and taking the limit in $\epsilon \to 0$, we obtain the following coordinate representation of the gradient

$$\nabla f(\boldsymbol{x}) = \left(\frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}^{(1)}}, \dots, \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}^{(n)}} \right)^\top$$

- Denote by $\mathcal{L}_f(\alpha)$ the **level set**(等高线) of $f(\boldsymbol{x})$:

$$\mathcal{L}_f(\alpha) = \{\boldsymbol{x} \in \mathbb{R}^n | f(\boldsymbol{x}) \le \alpha\}$$

- Consider the *set* of directions that are **tangent** to $\mathcal{L}_f(f(\bar{\boldsymbol{x}}))$ at $\bar{\boldsymbol{x}}$

$$S_f(\bar{\boldsymbol{x}}) = \left\{ \boldsymbol{s} \in \mathbb{R}^n \,\middle|\, \lim_{\boldsymbol{y}_k \to \bar{\boldsymbol{x}},\ f(\boldsymbol{y}_k) = f(\bar{\boldsymbol{x}})} = \frac{\boldsymbol{y}_k - \bar{\boldsymbol{x}}}{\|\boldsymbol{y}_k - \bar{\boldsymbol{x}}\|} \right\}$$

Lemma II.1. *If $\boldsymbol{s} \in S_f(\bar{\boldsymbol{x}})$, then $\langle \nabla f(\bar{\boldsymbol{x}}), \boldsymbol{s} \rangle = 0$*

Proof. For $f(\boldsymbol{y}_k) = f(\bar{\boldsymbol{x}})$, we have,

$$\begin{aligned} f(\boldsymbol{y}_k) &= f(\bar{\boldsymbol{x}}) + \langle \nabla f(\bar{\boldsymbol{x}}), \boldsymbol{y}_k - \bar{\boldsymbol{x}} \rangle + o(\|\boldsymbol{y}_k - \bar{\boldsymbol{x}}\|) \\ &= f(\bar{\boldsymbol{x}}) \end{aligned}$$

Therefore, $\langle \nabla f(\bar{\boldsymbol{x}}), \boldsymbol{y}_k - \bar{\boldsymbol{x}} \rangle + o(\|\boldsymbol{y}_k - \bar{\boldsymbol{x}}\|) = 0$. Dividing this equation by $\|\boldsymbol{y}_k - \bar{\boldsymbol{x}}\|$, and taking the limit $\boldsymbol{y}_k \to \bar{\boldsymbol{x}}$, we obtain

$$\langle \nabla f(\bar{\boldsymbol{x}}), \lim_{\boldsymbol{y}_k \to \bar{\boldsymbol{x}}} \frac{\boldsymbol{y}_k - \bar{\boldsymbol{x}}}{\|\boldsymbol{y}_k - \bar{\boldsymbol{x}}\|} \rangle = \langle \nabla f(\bar{\boldsymbol{x}}), \boldsymbol{s} \rangle = 0$$

Q.E.D.

The directions $-\nabla f(\bar{\boldsymbol{x}})$ (the **antigradient**) is the directions of the **fastest local decrease** of $f(\boldsymbol{x})$ at point $\bar{\boldsymbol{x}}$. (证明: 梯度负方向下降最快)

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

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

Note that $f(\bar{\boldsymbol{x}} + \alpha \boldsymbol{s}) - f(\bar{\boldsymbol{x}}) = \alpha \langle \nabla f(\bar{\boldsymbol{x}}), \boldsymbol{s} \rangle + o(\alpha \|\boldsymbol{s}\|)$. Therefore, we have

$$\Delta(\boldsymbol{s}) = \langle \nabla f(\bar{\boldsymbol{x}}), \boldsymbol{s} \rangle$$

By leveraging the Cauchy-Schwartz inequality, that is $-\|x\| \cdot \|y\| \le \langle x, y \rangle \le \|x\| \cdot \|y\|$, we obtain,

$$\Delta(\boldsymbol{s}) = \langle \nabla f(\bar{\boldsymbol{x}}), \boldsymbol{s} \rangle \ge -\|\nabla f(\bar{\boldsymbol{x}})\|$$

For the lower bound $\bar{\boldsymbol{s}} = -\frac{\nabla f(\bar{\boldsymbol{x}})}{\|\nabla f(\bar{\boldsymbol{x}})\|}$ (取到下界), we have

$$\Delta(\bar{\boldsymbol{s}}) = -\frac{\langle \nabla f(\bar{\boldsymbol{x}}), \nabla f(\bar{\boldsymbol{x}}) \rangle}{\|\nabla f(\bar{\boldsymbol{x}})\|} = -\|\nabla f(\bar{\boldsymbol{x}})\|$$

Q.E.D.

Theorem II.2 (First-order Optimality Condition)**.** *Let $\boldsymbol{x}^*$ be a local minimum of differentiable function $f(\boldsymbol{x})$. Then $\nabla f(\boldsymbol{x}^*) = 0$.*

Proof. Since $\boldsymbol{x}^*$ is a local minimum of $f(\boldsymbol{x})$, then there exists $r > 0$ such that $\forall \boldsymbol{y}$, $\|\boldsymbol{y} - \boldsymbol{x}^*\| \le r$, we have $f(\boldsymbol{y}) \ge f(\boldsymbol{x}^*)$. Since f is differentiable, this implies that

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

Thus, $\forall \boldsymbol{s}, \|\boldsymbol{s}\| = 1$, we have $\langle \nabla f(\boldsymbol{x}^*), \boldsymbol{s} \rangle \ge 0$. Consider the directions $\boldsymbol{s}$ and $-\boldsymbol{s}$, we get

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

Finally, choosing $\boldsymbol{s} = e_i, i = 1 \dots n$, where e_i is the i-th coordinate vector in $\mathbb{R}^n$, we obtain $\nabla f(\boldsymbol{x}^*) = 0$. Q.E.D.

仅必要条件, 反过来是 stationary points (静态点).
e.g $f(x) = x^3 . x \in \mathbb{R}^1$ at $x = 0$.

Corollary II.3. *Let $\boldsymbol{x}^*$ be a local minimum of a differentiable function $f(\boldsymbol{x})$ s.t. linear equality constraints*

$$\boldsymbol{x} \in \mathcal{L} \equiv \{\boldsymbol{x} \in \mathbb{R}^n | A\boldsymbol{x} = \boldsymbol{b}\} \ne \emptyset$$

where A is an $m \times n$ matrix and $\boldsymbol{b} \in \mathbb{R}^m, m < n$. Then there exists a vector of multipliers λ^* such that

$$\nabla f(\boldsymbol{x}^*) = A^\top \lambda^* \tag{II.2}$$

$\mathcal{L}$ is the Null space of matrix A.

Proof. Consider some vector $\boldsymbol{u}_i, i = 1 \dots k$, that form a basis of the Null space (零空间) of matrix A. Then any $\boldsymbol{x} \in \mathcal{L}$ can be represented as follows:

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

Moreover, the point $\boldsymbol{y} = 0$ is a local minimum of the function $\phi(\boldsymbol{y}) = f(\boldsymbol{x}(\boldsymbol{y}))$. In view of **Theorem** II.2, $\nabla \phi(0) = 0$. This means that

$$\frac{\partial \phi(0)}{\partial \boldsymbol{y}^{(i)}} = \frac{\partial \phi(0)}{\partial \boldsymbol{x}(\boldsymbol{y})} \cdot \frac{\partial \boldsymbol{x}(\boldsymbol{y})}{\partial \boldsymbol{y}^{(i)}} = \langle \nabla f(\boldsymbol{x}^*), \boldsymbol{u}_i \rangle = 0, i = 1 \dots k$$

and II.2 follows. (因为零空间与行空间正交) Q.E.D.

1.3 Second Order Approximation

Let function $f(\boldsymbol{x})$ be twice differentiable at $\bar{\boldsymbol{x}}$. Then

$$f(\boldsymbol{y}) = f(\bar{\boldsymbol{x}}) + \langle \nabla f(\bar{\boldsymbol{x}}), \boldsymbol{y} - \bar{\boldsymbol{x}} \rangle + \frac{1}{2} \langle \nabla^2 f(\bar{\boldsymbol{x}})(\boldsymbol{y} - \bar{\boldsymbol{x}}), \boldsymbol{y} - \bar{\boldsymbol{x}} \rangle + o(\|\boldsymbol{y} - \bar{\boldsymbol{x}}\|^2)$$

The quadratic function

$$f(\bar{\boldsymbol{x}}) + \langle \nabla f(\bar{\boldsymbol{x}}), \boldsymbol{y} - \bar{\boldsymbol{x}} \rangle + \frac{1}{2} \langle \nabla^2 f(\bar{\boldsymbol{x}})(\boldsymbol{y} - \bar{\boldsymbol{x}}), \boldsymbol{y} - \bar{\boldsymbol{x}} \rangle$$

is called the **quadratic** (or **second order**) approximation of function f at $\bar{\boldsymbol{x}}$.

Recall that the $(n \times n)$ matrix $\nabla^2 f(\boldsymbol{x})$ has the following entries:

$$(\nabla^2 f(\boldsymbol{x}))^{(i,j)} = \frac{\partial^2 f(\boldsymbol{x})}{\partial \boldsymbol{x}^{(i)} \partial \boldsymbol{x}^{(j)}}$$

This matrix is called **Hessian** of function f at $\boldsymbol{x}$.

Note that the Hessian is a **symmetric** matrix:

$$\nabla^2 f(\boldsymbol{x}) = [\nabla^2 f(\boldsymbol{x})]^\top$$

Remark:

- Notation $A \succeq 0$ means that A is **positive semidefinite**(半正定):

$$\boldsymbol{x}^\top A \boldsymbol{x} \geq 0 \ \forall \boldsymbol{x} \in \mathbb{R}^n$$

- Notation $A \succ 0$ means that A is **positive definite**(正定):

$$\boldsymbol{x}^\top A \boldsymbol{x} > 0 \ \forall \boldsymbol{x} \in \mathbb{R}^n$$

Theorem II.4. *Let $\boldsymbol{x}^*$ be a local minimum of a twice differentiable function $f(\boldsymbol{x})$. Then*

$$\nabla f(\boldsymbol{x}^*) = 0, \ \nabla^2 f(\boldsymbol{x}^*) \succeq 0$$

Proof. Since $\boldsymbol{x}^*$ is a local minimum of function $f(\boldsymbol{x})$, there exists $r > 0$ such that $\forall \boldsymbol{y}, \|\boldsymbol{y} - \boldsymbol{x}^*\| \leq r$, we have

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}^*)$$

In view of **Theorem** II.2, $\nabla f(\boldsymbol{x}^*) = 0$.

Therefore, for any such $\boldsymbol{y}$,

$$f(\boldsymbol{y}) = f(\boldsymbol{x}^*) + \frac{1}{2} \langle \nabla^2 f(\boldsymbol{x}^*)(\boldsymbol{y} - \boldsymbol{x}^*), \boldsymbol{y} - \boldsymbol{x}^* \rangle + o(\|\boldsymbol{y} - \boldsymbol{x}^*\|^2) \geq f(\boldsymbol{x}^*)$$

两项都除以 $\|\boldsymbol{y} - \boldsymbol{x}^*\|^2$, let $\boldsymbol{y} \to \boldsymbol{x}^*$, and let $\boldsymbol{s} = \frac{\boldsymbol{y} - \boldsymbol{x}^*}{\|\boldsymbol{y} - \boldsymbol{x}^*\|}$

$$\lim_{\boldsymbol{y} \to \boldsymbol{x}^*} \left(\frac{1}{2} \frac{\langle \nabla^2 f(\boldsymbol{x}^*)(\boldsymbol{y} - \boldsymbol{x}^*), \boldsymbol{y} - \boldsymbol{x}^* \rangle}{\|\boldsymbol{y} - \boldsymbol{x}^*\|^2} + \frac{o(\|\boldsymbol{y} - \boldsymbol{x}^*\|^2)}{\|\boldsymbol{y} - \boldsymbol{x}^*\|^2} \right) \geq 0$$

$$\langle \nabla^2 f(\boldsymbol{x}^*) \boldsymbol{s}, \boldsymbol{s} \rangle \geq 0$$

$\forall \boldsymbol{s}, \|\boldsymbol{s}\| = 1$ Q.E.D.

Theorem II.5. *Let function $f(\boldsymbol{x})$ be twice differentiable on $\mathbb{R}^n$ and let $\boldsymbol{x}^*$ satisfy the following conditions:*

$$\nabla f(\boldsymbol{x}^*) = 0, \ \nabla^2 f(\boldsymbol{x}^*) \succ 0$$

Then $\boldsymbol{x}^*$ is a strict local minimum of $f(\boldsymbol{x})$.

Remark: A point $\bar{\boldsymbol{x}} \in \mathbb{R}^n$ is an unconstrained strict local minimum of a function $f : \mathbb{R}^n \to \mathbb{R}$ if $\exists \epsilon > 0$ such that $f(\bar{\boldsymbol{x}}) < f(\boldsymbol{x}) \ \forall \boldsymbol{x} \in B(\bar{\boldsymbol{x}}, \epsilon), \boldsymbol{x} \neq \bar{\boldsymbol{x}}$, where $B(\bar{\boldsymbol{x}}, \epsilon) := \{\boldsymbol{x} | \|\boldsymbol{x} - \bar{\boldsymbol{x}}\| \leq \epsilon\}$

$\lambda_1(A)$ 表示 A 的最小特征值, $\lambda_{max}(A)$ 表示最大的特征值.

2. Classes of differentiable function

2.1 Class $C_L^{k,p}(\mathbb{R}^n)$

Consider a classes of **differentiable** functions which meet a **Lipschitz conditon** for a derivative of certain order.

Let $Q \subseteq \mathbb{R}^n$. We denote by $C_L^{k,p}(Q)$ the class of functions with the following properties:

- Any $f \in C_L^{k,p}(Q)$ is k times continuously **differentiable** on Q.
- Its p-th derivative is **Lipschitz continous** on Q with the constant L:

$$\left\| f^{(p)}(\boldsymbol{x}) - f^{(p)}(\boldsymbol{y}) \right\| \leq L \|\boldsymbol{x} - \boldsymbol{y}\|$$

$\forall x, y, \in Q$.

Clearly, we always have

1) $p \leq k$ 显然成立
2) If $q \geq k$, then $C_L^{q,p}(Q) \subseteq C_L^{k,p}$.
3) Note also that these classes possess the following property:

If $f_1 \in C_{L_1}^{k,p}(Q), f_2 \in C_{L_1}^{k,p}(Q)$ and $\alpha, \beta \in \mathbb{R}^1$, then for

$$L_3 = |\alpha| L_1 + |\beta| L_2$$

we have $\alpha f_1 + \beta f_2 \in C_{L_3}^{k,p}(Q)$.

Remark. We use notation $f \in C^k(Q)$ for a function f which is k times continuously differ entiable on Q.

2.2 Class $C_L^{1,1}(\mathbb{R}^n)$

Consider $C_L^{1,1}(\mathbb{R}^n)$, the class of functions with **Lipschitz continous gradient**. By definition, the inclusion $f \in C_L^{1,1}(\mathbb{R}^n)$ implies that $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$,

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\| \leq L \|\boldsymbol{x} - \boldsymbol{y}\|$$

Let us give a sufficient condition for that inclusion.

Lemma II.6. *A function $f(\boldsymbol{x})$ belongs to $C_L^{2,1}(\mathbb{R}^n) \subset C_L^{1,1}(\mathbb{R}^n)$, if and only if*

$$\|\nabla^2 f(\boldsymbol{x})\| \leq L, \ \forall \boldsymbol{x} \in \mathbb{R}^n$$

矩阵二范数

$g(a) = \nabla f(x + a(y - x))$, $g(0) = \nabla f(x), g(1) = \nabla f(y)$

$g(1) = g(0) + \int_0^1 g'(\tau) d\tau$

$g'(\tau) = \frac{\partial \nabla f(x+\tau(y-x))}{\partial (x+\tau(y-x))} \frac{\partial (x+\tau(y-x))}{\partial \tau} = \nabla^2 f(x + \tau(y - x))(y - x)$

e.g.

geometric interpretation The next statement is important for the geometric interpretation of function from $C_L^{1,1}(\mathbb{R}^n)$.

Lemma II.7. *Let $f \in C_L^{1,1}(\mathbb{R}^n)$. Then $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, we have*

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

f is called smooth.

Remark.

$f(\boldsymbol{y})$ 和其一阶逼近 $g(\boldsymbol{y}) = f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle$ 的距离的上界.

强凸

2.3 Class $C_M^{2,1}(\mathbb{R}^n)$

Consider class $C_M^{2,2}(\mathbb{R}^n)$. That is $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, we have

$$\|\nabla^2 f(\boldsymbol{x}) - \nabla^2 f(\boldsymbol{y})\| \leq M \|\boldsymbol{x} - \boldsymbol{y}\|$$

Lemma II.8. *Let $f \in C_M^{2,2}(\mathbb{R}^n)$. Then for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, we have*

$$\|\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x}) - \nabla^2 f(\boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x})\| \leq \frac{M}{2} \|\boldsymbol{y} - \boldsymbol{x}\|^2$$

Corollary II.9. *Let $f \in C_M^{2,2}(\mathbb{R}^n)$ adn $\|\boldsymbol{y} - \boldsymbol{x}\| = r$. Then*

$$\nabla^2 f(\boldsymbol{x}) - MrI_n \succeq \nabla^2 f(\boldsymbol{y}) \succeq \nabla^2 f(\boldsymbol{x}) + MrI_n$$

where I_n is the unit matrix $\mathbb{R}^n$.

对于矩阵 A 和 B, $A \succeq B \iff A - B \succeq 0$.

3. Convex Function

3.1 Definition of The Convex Function and The Class $\mathcal{F}^k$

Consider the unconstrained minimization problem

$$\min_{\boldsymbol{x} \in \mathbb{R}^n} f(\boldsymbol{x}) \tag{II.3}$$

where the objective function $f(\boldsymbol{x})$ is smooth enough

定义 smooth 足够的问题, 但这个约束很弱, 在局部最小不能保证收敛, 不能保证上下界.

增加几个假设

Assumption II.10. *For any $f \in \mathcal{F}$, first-order optimality condition is **sufficient** for a point to be a **global solution** to II.3*

The possibility to verify the inclusion $f \in \mathcal{F}$ in a simple way.

Assumption II.11. *If $f_1, f_2 \in \mathcal{F}$ and $\alpha, \beta \geq 0$, then $\alpha f_1 + \beta f_2 \in \mathcal{F}$.*

Assumption II.12. *Any linear function $f(\boldsymbol{x}) = \alpha + \langle a, \boldsymbol{x} \rangle$ belongs to $\mathcal{F}$.*

Definition of The Convex Function Consider

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}_0) + \langle \nabla f(\boldsymbol{x}_0), \boldsymbol{y} - \boldsymbol{x}_0 \rangle$$

函数大于线性逼近.

Definition II.13 (Convex Set)**.** *A set $\mathcal{Q} \subseteq \mathbb{R}^n$ is called **convex** if $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{Q}$, and $\alpha \in [0, 1]$, we have*

$$\alpha \boldsymbol{x} + (1 - \alpha \boldsymbol{y}) \in \mathcal{Q}$$

(集合中任取两点, 连线都是属于集合的, $\mathcal{F}$ 表示是凸的)

We denote by $\mathcal{F}^k(\mathcal{Q})$ the class we discussed above, and call it **class of the convex function**:

- Any $f \in \mathcal{F}^k(\mathcal{Q})$ is a **convex function**
- Any $f \in \mathcal{F}^k(\mathcal{Q})$ is k times continuously **differentiable** on $\mathcal{Q}$
- We assume $\mathcal{Q} = \mathbb{R}^n$ in this chapter

Definition II.14 (Convex Function)**.** *A continuously differentiable function $f(\cdot)$ is called **convex** on a convex set $\mathcal{Q}$ (notation $f \in \mathcal{F}^1(\mathcal{Q}^n)$) if $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{Q}$ we have*

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle$$

*If $-f(\boldsymbol{x})$ is convex, we call $f(\boldsymbol{x})$ **concave**.*

3.2 Properties of The Convex Function

Theorem II.15 (Global Property)**.** *If $f \in \mathcal{F}^1(\mathbb{R}^n)$ and $\nabla(f\boldsymbol{x}^*) = 0$, then $\boldsymbol{x}^*$ is the global minimum of $f(\boldsymbol{x})$ on $\mathbb{R}^n$*

Lemma II.16 (Conic Combination)**.** *If f_1 and f_2 belong to $\mathcal{F}^1(\mathbb{R}^n)$, and $\alpha, \beta \geq 0$, then the function $f = \alpha f_1 + \beta f_2$ also belongs to $\mathcal{F}^1(\mathbb{R}^n)$.*

Lemma II.17 (Affine Composition)**.** *If $f \in \mathcal{F}^1(\mathbb{R}^n), \boldsymbol{b} \in \mathbb{R}^m$, and $A : \mathbb{R}^n \to \mathbb{R}^m$, then*

$$\phi(\boldsymbol{x}) = f(A\boldsymbol{x} + \boldsymbol{b}) \in \mathcal{F}^1(\mathbb{R}^m)$$

Lemma II.18 (Pointwise maximum and supremum)**.**

If $f_i(\boldsymbol{x}), i \in I$, are convex, then

$$g(\boldsymbol{x}) = \max_{i \in I} f_i(\boldsymbol{x})$$

is also convex.

Lemma II.19 (Convex monotone composition)**.**

1) If f is a convex function on $\mathbb{R}^n$ and $F(\cdot)$ is a convex and non-decreasing function on $\mathbb{R}$, then $g(\boldsymbol{x}) = F(f(\boldsymbol{x}))$ is convex.

2) If $f_i, i = 1, \ldots, m$ are convex functions on $\mathbb{R}^n$ and $F(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m)$ is convex and non-decreasing (componentwise) in each argument, then

$$g(\boldsymbol{x}) = F(f_1(\boldsymbol{x}), \ldots, f_m(\boldsymbol{x}))$$

is convex.

Lemma II.20 (Partial minimization)**.** *If $f(\boldsymbol{x}, \boldsymbol{y})$ is convex in $(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}^n$ and Y is a convex set, then*

$$g(\boldsymbol{x}) = \inf_{\boldsymbol{y} \in Y} f(\boldsymbol{x}, \boldsymbol{y})$$

is convex.

3.3 Equivalent Definitions

Theorem II.21. *A continuously **differentiable** function f belongs to the class $\mathcal{F}^1(\mathbb{R}^n)$ if and only if $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ and $\alpha \in [0, 1]$ we have*

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

(凸组合的函数 $\leq$ 函数的凸组合)

Theorem II.22. *A continuously differentiable function f belongs to the class $\mathcal{F}^1(\mathbb{R}^n)$ if and only if $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, we have*

$$\langle \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle \geq 0$$

Theorem II.23. *A twice continuously differentiable function f belongs to class $\mathcal{F}^2(\mathbb{R}^n)$ if and only if $\forall x \in \mathbb{R}^n$, we have*

$$\nabla^2 f(\boldsymbol{x}) \succeq 0$$

3.4 Examples

1)

4. Smooth and Convex Function

4.1 The Class $\mathcal{F}_L^{k,l}(\mathbb{R}^n)$

1) Any function $f \in \mathcal{F}_L^{k,l}(\mathbb{R}^n)$ is convex

2) the meaning of the index is the same as $C_L^{k,l}(\mathbb{R}^n)$

3) The most important class of that type is $\mathcal{F}_L^{1,1}(\mathbb{R}^n)$, the class of convex functions with **Lipschitz continous gradient**.

4.2 Necessary and Sufficient Conditions for The Class $\mathcal{F}_L^{1,1}(\mathbb{R}^n)$

Theorem II.24. *All the conditions below, holding $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ and $\alpha \in [0, 1]$, are equivalent to the inclusion $f \in \mathcal{F}_L^{1,1}(\mathbb{R}^n)$:*

$$0 \leq f(\boldsymbol{y}) - f(\boldsymbol{x}) - \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle \leq \frac{L}{2} \|\boldsymbol{x} - \boldsymbol{y}\|^2 \tag{II.4}$$

$$f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{1}{2L} \|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|^2 \leq f(\boldsymbol{y}) \tag{II.5}$$

$$\frac{1}{L} \|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|^2 \leq \langle \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle \tag{II.6}$$

$$0 \leq \langle \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle \leq L \|\boldsymbol{x} - \boldsymbol{y}\|^2 \tag{II.7}$$

$$\begin{aligned} \alpha f(\boldsymbol{x}) + (1-\alpha) f(\boldsymbol{y}) \geq & f(\alpha \boldsymbol{x} + (1-\alpha)\boldsymbol{y}) \\ & + \frac{\alpha(1-\alpha)}{2L} \|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|^2 \end{aligned} \tag{II.8}$$

$$\begin{aligned} & 0 \leq \alpha f(\boldsymbol{x}) + (1-\alpha) f(\boldsymbol{y}) - f(\alpha \boldsymbol{x} + (1-\alpha)\boldsymbol{y}) \\ & \leq \alpha(1-\alpha) \frac{L}{2} \|\boldsymbol{x} - \boldsymbol{y}\|^2 \end{aligned} \tag{II.9}$$

4.3 Necessary and Sufficient Conditions for The Class $\mathcal{F}_L^{2,1}(\mathbb{R}^n)$

Theorem II.25. *Twice continuously differentiable function $f \in \mathcal{F}_L^{2,1}(\mathbb{R}^n)$ if and only if $\forall \boldsymbol{x} \in \mathbb{R}^n$, we have*

$$0 \preceq \nabla^2 f(\boldsymbol{x}) \preceq L I_n$$

5. Strongly Convex Function

5.1 Definition of The Strongly Convex and The Class $\mathcal{S}_\mu^1(\mathbb{R}^n)$

Definition II.26. *A continuously differentiable function $f(\boldsymbol{x})$ is called **strongly convex** on $\mathbb{R}^n$ (notation $f \in \mathcal{S}_\mu^1(\mathbb{R}^n)$) if there exists a constant $\mu > 0$ such that $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, we have*

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{1}{2}\mu \|\boldsymbol{y} - \boldsymbol{x}\|^2$$

*μ is called **convexity parameter**(凸参数) of f.*

5.2 Property of Strongly Convex Function

Theorem II.27. *If $f \in \mathcal{S}_\mu^1(\mathbb{R}^n)$ and $\nabla f(\boldsymbol{x}^*) = 0$, then*

$$f(\boldsymbol{x}) \geq f(\boldsymbol{x}^*) + \frac{1}{2}\mu \|\boldsymbol{x} - \boldsymbol{x}^*\|^2, \ \forall \boldsymbol{x} \in \mathbb{R}^n$$

Lemma II.28. *If $f_1 \in \mathcal{S}_{\mu_1}^1(\mathbb{R}^n), f_2 \in \mathcal{S}_{\mu_2}^1(\mathbb{R}^n)$ and $\alpha, \beta \geq 0$ then,*

$$f = \alpha f_1 + \beta f_2 \in \mathcal{S}_{\alpha\mu_1 + \beta\mu_2}^1(\mathbb{R}^n)$$

加线性函数增加凸性, 让多解变为单解.

Theorem II.29. *If $f \in \mathcal{S}_{\mu}^{1}(\mathbb{R}^n)$, then $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, we have*

$$f(\boldsymbol{y}) \leq f(\boldsymbol{x}) + \langle f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{1}{2\mu} \|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|^2 \quad \text{(II.10)}$$

$$\langle \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle \leq \frac{1}{\mu} \|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|^2 \quad \text{(II.11)}$$

$\phi(\nu)$

5.3 Equivalent Definitions

Theorem II.30. *Let f be continuously differentiable. Both cnoditions below, holding $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ and $\alpha \in [0, 1]$, are equivalent to inclusion $\mathcal{S}_{\mu}^{1}(\mathbb{R}^n)$*

$$\langle \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle \geq \mu \|\boldsymbol{x} - \boldsymbol{y}\|^2 \quad \text{(II.12)}$$

$$\alpha f(\boldsymbol{x}) + (1-\alpha) f(\boldsymbol{y}) \geq f(\alpha \boldsymbol{x} + (1-\alpha)\boldsymbol{y}) + \alpha(1-\alpha)\frac{\mu}{2} \|\boldsymbol{x} - \boldsymbol{y}\|^2 \quad \text{(II.13)}$$

$\mathcal{S}_{\mu}^{2}(\mathbb{R}^n) \subseteq \mathcal{S}_{\mu}^{1}(\mathbb{R}^n)$

Theorem II.31. *Two times continuously differentiable functions f belongs to the class $\mathcal{S}_{\mu}^{2}(\mathbb{R}^n)$ if and only if $\boldsymbol{x} \in \mathbb{R}^n$*

$$\nabla^2 f(\boldsymbol{x}) \succeq \mu I_n$$

$\boldsymbol{s} \in \mathbb{R}^n$

$\langle I_n \boldsymbol{s}, \boldsymbol{s} \rangle$

5.4 Examples

1)

6. Smooth and Strongly Convex Function

6.1 The Class $\mathcal{S}_{\mu,L}^{1,1}(\mathbb{R}^n)$

The value $Q_f = \frac{L}{\mu} \geq 1$ is called the **condition number**(条件数) of function f.

6.2 Property of Smooth and Strongly Convex Function

Theorem II.32. *If $f \in \mathcal{S}_{\mu,L}^{1,1}(\mathbb{R}^n)$, then $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, we have*

$$\langle \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle \geq \frac{\mu L}{\mu + L} \|\boldsymbol{x} - \boldsymbol{y}\|^2 + \frac{1}{\mu + L} \|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|^2$$

退化后和... 与... 相关

7. Conclusion

7.1 Upper Bounds on Functional Components

Lipschitz Continuity: $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$,

- Zerothorder Condition:

$$|f(\boldsymbol{x}) - f(\boldsymbol{y})| \leq L \|\boldsymbol{x} - \boldsymbol{y}\|$$

- First-order Condition:

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\| \leq L \|\boldsymbol{x} - \boldsymbol{y}\|$$

- p-order Condition

$$\|\nabla^p f(\boldsymbol{x}) - \nabla^p f(\boldsymbol{y})\|_* \leq L \|\boldsymbol{x} - \boldsymbol{y}\|, \ p \geq 2$$

7.2 Lower Bounds on Functional Components

Convexity: $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, $\alpha \in [0, 1]$,

- Zerothorder Condition:

$$\alpha f(\boldsymbol{x}) + (1-\alpha) f(\boldsymbol{y}) - f(\alpha \boldsymbol{x} + (1-\alpha)\boldsymbol{y}) \geq 0$$

- First-order Condition:

$$D_f(\boldsymbol{x}, \boldsymbol{x}) \triangleq f(\boldsymbol{y}) - \{f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle\} \geq 0$$
$$\langle \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle \geq 0$$

- Second-order Condition:

$$\nabla^2 f(\boldsymbol{x}) \succeq 0$$

Strong Convexity: $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$,

- Zerothorder Condition:

$$\alpha f(\boldsymbol{x}) + (1-\alpha) f(\boldsymbol{y}) - f(\alpha \boldsymbol{x} + (1-\alpha)\boldsymbol{y}) \geq \alpha(1-\alpha)\frac{\mu}{2} \|\boldsymbol{x} - \boldsymbol{y}\|^2$$

- First-order Condition:

$$D_f(\boldsymbol{x}, \boldsymbol{y}) \geq \frac{\mu}{2} \|\boldsymbol{x} - \boldsymbol{y}\|^2$$
$$\langle \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle \geq \mu \|\boldsymbol{x} - \boldsymbol{y}\|^2$$

- Second-order Condition

$$\nabla^2 f(\boldsymbol{x}) \succeq \mu I_n$$

7.3 Other Lower Bounds on Functional Components

- Weak Strong Convexity(WSC)

$$D_f(\boldsymbol{x}^*, \boldsymbol{y}) \geq \frac{\mu}{2} \|\boldsymbol{x}^* - \boldsymbol{x}\|^2, \ \forall \boldsymbol{x} \in \mathbb{R}^n$$

- Restricted Secant Inequality(RSI)

$$\langle \nabla f(\boldsymbol{x}), \boldsymbol{x} - \boldsymbol{x}^* \rangle \geq \mu \|\boldsymbol{x}^* - \boldsymbol{x}\|^2, \ \forall \boldsymbol{x} \in \mathbb{R}^n$$

- Polyak-Łojaciewicz(PL)

$$\frac{1}{2} \|\nabla f(\boldsymbol{x})\|^2 \geq \mu(f(\boldsymbol{x}) - f(\boldsymbol{x}^*)), \ \forall \boldsymbol{x} \in \mathbb{R}^n$$

- Error Bounds(EB)

$$\|\nabla f(\boldsymbol{x})\| \geq \mu \|\boldsymbol{x}^* - \boldsymbol{x}\|, \ \forall \boldsymbol{x} \in \mathbb{R}^n$$

- Quadratic Growth(QG)

$$f(\boldsymbol{x}) - f(\boldsymbol{x}^*) \geq \frac{\mu}{2} \|\boldsymbol{x}^* - \boldsymbol{x}\|, \ \forall \boldsymbol{x} \in \mathbb{R}^n$$

(SC)→(WSC)→(RSI)→(EB)≡(PL)→(QG)
If f is convex, (RSI)≡(EB)≡(PL)≡(QG)

Theorem II.33. *μ-strongly convex function → μ-PL*

III Descent Meethod

1. Gradient Descent

1.1 Basic Scheme

Gradient Descent Formulation We will
Gradient Method

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

Step Size

1.2 Performance for $C_L^{1,1}(\mathbb{R}^n)$

1) in advance
2) Full relaxation (贪心)
3) G-A (折中)

$$f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{x}) + ..$$
$$f(\boldsymbol{x}_{k+1}) \geq f(\boldsymbol{x}) + ..$$

Geometric interpretation

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

where $f \in C_L^{1,1}(\mathbb{R}^n)$
ref (1.2.5)

$$f(\boldsymbol{y}) \leq f(\boldsymbol{x}) - h\left(1 - \frac{h}{2}L\right) \|\nabla f(\boldsymbol{x})\|^2$$

Remark: $\forall h \in \left(0, \frac{2}{L}\right)$, $h\left(1 - \frac{h}{2}L\right) \|\nabla f(\boldsymbol{x})\|^2 \geq 0$.
step-size strategies:

1) $h_k = \frac{1}{L}$
2) $h_k = \frac{1}{L}$
3) $h_k \geq \frac{2}{L}(1 - \beta)$

步长选取影响 ω

$$f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}_k) \leq -\frac{\omega}{L} \|\nabla f(\boldsymbol{x}_k)\|^2$$

where ω is some positive constant.

$$\|\nabla f(\boldsymbol{x}_k)\| \to 0 \text{ as } k \to \infty$$

方法的收敛和梯度与步长相关. 每一步下降但最后是不一定收敛的 (需要一个下界).

$$g_T^* = \min g_k$$

$$g_T^* \leq \frac{1}{\sqrt{T+1}} []$$

速度是 $\frac{1}{\sqrt{T+1}}$

可得到一个复杂度上界 $g_T^* \leq \epsilon$

步长收敛相当于主动构造序列, 保证了递减与有下界, 但收敛点的意义不能保证.

e.g.

1.3 Performance for $\mathcal{F}_L^{1,1}(\mathbb{R}^n)$

Theorem III.1.

$$f(\boldsymbol{x}_k) - f^* \leq \frac{2(f(\boldsymbol{x}_0) - f^*)\left\|\boldsymbol{x}_0 - \boldsymbol{x}^*\right\|^2}{2\left\|\boldsymbol{x}_0 - \boldsymbol{x}^*\right\|^2 + kh(2-Lh)(f(\boldsymbol{x}_0) - f^*)}$$

Corollary III.2. *If*

$$f(\boldsymbol{x}_k) - f^* \leq \frac{2L\left\|\boldsymbol{x}_0 - \boldsymbol{x}^*\right\|^2}{k+4}$$

1.4 Performance for $\mathcal{S}_{\mu,L}^{1,1}(\mathbb{R}^n)$

Theorem III.3.

$$\left\|\boldsymbol{x}_k - \boldsymbol{x}^*\right\|^2 \leq \left(1 - \frac{2h\mu L}{\mu + L}\right)^k \left\|\boldsymbol{x}_0 - \boldsymbol{x}^*\right\|^2$$

2. General Descent Directions

2.1 Choosing the Direction

3. Newton Method

3.1 Basic Scheme

Historical Origins 寻找单变量函数的根.

可扩展到解非线性系统. non degenerate

Basic Scheme 由非约束最小化问题到寻找非线性系统的根.

Remark:

- ss
- 是可能 diverge(发散) 的

Example:

Damped Newton Method 加阻尼防止发散

3.2 Local Convergence of The Newton Method

1)

考虑:

$$111 \tag{III.1}$$

第一步 $\nabla f(\boldsymbol{x}_k)$ 进行 Taylor 零阶展开.

contracting mapping(收缩映射)

Bound for $[\nabla^2 f(\boldsymbol{x}_k)]^{-1}$

The rate of convergence of this type is called quadratic

$$\lim_{k\to\infty} \frac{a_{k+1}}{a_k^2} = r$$

Theorem III.4. *f(x)*

3.3 Convergence Analysis

- GD
- Lightweight Newton G_k
- Variable Metric(可变度量方法) or Quasi-Newton (拟牛顿)
- New Inner Product (W.R.t)

Example: (α 应该是 $\boldsymbol{a}$)

Variable metric method

1) Choose
2) iter

   a.

Quasi-Newton rule: 符合公式, 能迭代, 能收敛

Example:

- Rank-one correction
- (DFP)
- (BFGS)

Remark:

1) n iter
2) superlinear

$$\lim_{k\to\infty} \frac{a_{k+1}}{a_k} = 0$$

3) 对全局最优并不必 GD 好.

4) 拟牛顿最大的问题是有过多附加运算, 用以存储与计算 $n \times n$ 矩阵. 以此提出新的方法.

4. Conjugate Gradient

4.1 Historical Origins

迭代就是一个线性组合, 直接在子空间中寻找解.

Krylov Subspace To solve (这是一类方法, 例如)

Remark: 这些方法可以被看做为投影.

CG CG

$$\boldsymbol{x}_k = \text{argmin}\{f(\boldsymbol{x}) | \boldsymbol{x} \in \boldsymbol{x}_0 + \mathcal{L}_k\},\ k \geq 1$$

4.2 Fundamental Theory

Lemma III.5. $\forall k \geq 1$

Lemma III.6. $\forall k, i \geq 0$

Corollary III.7. *finite*

Corollary III.8. $\forall p$

Lemma III.9. *Let* δ_i, $\forall k \neq i$

4.3 CG Algorithm

Conjugate Gradient Method

1) 00
2) 11

a.

The specification of the coefficient β_k:

-

Recall:

- 11
- 做很多次 n 次最后可以收敛于 global

local 达到了 quadratic convergence

但全局并没有理论支持其更好.

IV Acceleration Methods

1. Lower Complexity Bounds for $\mathcal{F}_L^{\infty,1}(\mathbb{R}^n)$

1.1 Problem Class

Consider an optimization probelm where objectives function comes from $\mathcal{F}_L^{\infty,1}(\mathbb{R}^n)$(and $\mathcal{F}_L^{1,1}(\mathbb{R}^n)$). The probelm class is as following.

Model	$\min\limits_{\boldsymbol{x}\in\mathbb{B}^n} f(\boldsymbol{x}),\ f \in \mathcal{F}_L^{1,1}(\mathbb{R}^n)$
Oracle	First-order Black Box
Approximate solution	$\bar{\boldsymbol{x}} \in \mathbb{R}^n : f(\bar{\boldsymbol{x}}) - f^* \leq \epsilon$

Assumption IV.1.

$$\boldsymbol{x}_k = \boldsymbol{x}_0 + Lin\{\nabla f(\boldsymbol{x}_0), \ldots, \nabla f(\boldsymbol{x}_{k-1})\},\ k \geq 1$$

不把 $\boldsymbol{x}_0$ 写进 *Lin*, 表明对起点不关注.

1.2 Worst Function in $\mathcal{F}_L^{\infty,1}(\mathbb{R}^n)$

$$f_k(\boldsymbol{x}) = \frac{L}{4}\left\{\frac{1}{2}\left[\left(\boldsymbol{x}^{(1)}\right)^2 + \sum_{i=1}^{k-1}\left(\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(i+1)}\right)^2 + \left(\boldsymbol{x}^{(k)}\right)^2\right] - \boldsymbol{x}^{(1)}\right\}$$

for $k = 1, \ldots, n$

Remark: We have $f_k(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^\top\left(\frac{L}{4}A_k\right)\boldsymbol{x} - \frac{L}{4}e_1\boldsymbol{x}$.

$\nabla^2 f(\boldsymbol{x}) = \frac{L}{4}A_k$

$$A_k =$$

($A_k \preceq 4I$, for $k \in [1, n]$)

As assumption $0 \preceq \nabla^2 f(\boldsymbol{x}) \preceq LI_n$

About the optimal value:

$$\bar{\boldsymbol{x}}(i) = \begin{cases} 1 - \frac{i}{k+1} & i \in [1, k] \\ 0 & i \in [k+1, n] \end{cases}$$

Therefore, f_k

$$f_k =$$

Bounds of $\bar{\boldsymbol{x}}$

$\mathbb{R}^{k,n}$ sub-space

1.3 Theoretical Analysis and Main Results

Lemma IV.2.

$$\boldsymbol{x}_k \in \mathcal{L} =$$

Corollary IV.3.

$$f_p(\boldsymbol{x}_k) \geq f_k^*$$

Theorem IV.4. $1 \le k \le \frac{1}{2}(n-1)$

$$f(\boldsymbol{x}_k) - f^* \ge \frac{3L\left\|\boldsymbol{x}_0 - \boldsymbol{x}^*\right\|^2}{32(k+1)^2}$$

$$\left\|\boldsymbol{x}_k - \boldsymbol{x}^*\right\|^2 \ge \frac{1}{8}\left\|\boldsymbol{x}_0 - \boldsymbol{x}^*\right\|^2$$

where $f^* = f(\boldsymbol{x}^*)$

- 1
- 有限维可能可以有更好的复杂度.

2. Lower Complexity Bounds for $\mathcal{S}_L^{\infty,1}(\mathbb{R}^n)$

2.1 Worst Function in $\mathcal{S}_L^{\infty,1}(\mathbb{R}^n)$

$$\|\boldsymbol{x}\|^2 = \sum_{i=1}^{\infty}\left(\boldsymbol{x}^{(i)}\right)^2 < \infty$$

收敛

$$f_{\mu,Q_f}(\boldsymbol{x}) = \frac{\mu(Q_f-1)}{8}\left\{\left(\boldsymbol{x}^{(1)}\right)^2 + \sum_{i=1}^{\infty}\left(\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(i+1)}\right)^2 + -2\boldsymbol{x}^{(1)}\right\} + \frac{\mu}{2}\|\boldsymbol{x}\|^2$$

$$Q_{f_{\mu,Q_f}} = \frac{\mu Q_f}{\mu} = Q_f$$

Theorem IV.5.

$$\left\|\boldsymbol{x}_k - \boldsymbol{x}^*\right\|^2 \ge ()^{2k}\left\|\boldsymbol{x}_0 - \boldsymbol{x}^*\right\|^2$$

$$f(\boldsymbol{x}_k) - f^* \ge \frac{\mu}{2}()^{2k}\left\|\boldsymbol{x}_0 - \boldsymbol{x}^*\right\|^2$$

3. Basic Scheme

3.1 Difference bewteen Lower Bounds and Real Efficiency

Efficiency Estimation upper vs. lower

Estimate Sequences

Definition IV.6. *of function of* $f(\boldsymbol{x})$

Lemma IV.7.

$$f(\boldsymbol{x}_k) \le \phi_k^* \equiv \min_{\boldsymbol{x}\in\mathbb{R}^n}\phi_k(\boldsymbol{x})$$

Then $f(\boldsymbol{x}_k) - f^* \le \lambda_k[\phi^*(\boldsymbol{x}^*) - f^*] \to 0$

However,

1) 不知道怎么估计
2) 不知道如何满足条件

Lemma IV.8. *Assume that*

1)

Lemma IV.9. *Let*

Constructing $\boldsymbol{x}_k$

3.2 Optimal Scheme

4. Theoretical Analysis and Variants

4.1 Analysis of Optimal Scheme

Lemma IV.10. *www*

Lemma IV.11.

$$\lambda_k \le \min\{\}$$

Lemma IV.12.

$$f(\boldsymbol{x}_k) - f^* \le L\min\{\}$$

4.2 Variants of Optimal Scheme

...

Theorem IV.13.

...

Other methods

1) 1
2) 2
3) 3

Lemma IV.14.

V General Convex Problem

1. Motivation and Definitions

1.1 Motivation

$$\begin{aligned} &\min f_0(\boldsymbol{x}) \\ \text{s.t. } &f_i(\boldsymbol{x}) \le 0,\ i = 1, \dots, m \\ &\boldsymbol{x} \in Q \subseteq \mathbb{R}^n \end{aligned}$$

non-differentiable

Definition V.1 (Interior Point). *An element $\boldsymbol{x} \in C \subset \mathbb{R}^n$ is called an interior point of C if there exists an $\epsilon > 0$ for which*

$$\{\boldsymbol{y} \,|\, \|\boldsymbol{y} - \boldsymbol{x}\|_2 \le \epsilon\} \subset C$$

以及另一个 related interior point

Definition V.2 (Open and Closed Set). *A set C is open if int $C = C$, i.e. every point in C is an interior point. A set $C \subset \mathbb{R}^n$ is closed if its complement $\mathbb{R}^n / C = \{\boldsymbol{x} \in \mathbb{R}^n | \boldsymbol{x} \notin C\}$ is open.*

经常出现不光滑的优化问题.

- max-type functions
- 隐式给出问题

Denoted by

$$\operatorname{dom} f = \{\boldsymbol{x} \in \mathbb{R}^n : |f(\boldsymbol{x})| < \infty\}$$

the domain of function f. We always assume that $\operatorname{dom} f \neq \emptyset$

1.2 Definition of Convex Function

Definition V.3. *A function $f(\boldsymbol{x})$ is called convex, if its domain is convex and $\forall \boldsymbol{x}, \boldsymbol{y} \in \operatorname{dom} f$ and $\alpha \in [0, 1]$ the following inequality holds*

$$f(\alpha \boldsymbol{x} + (1 - \alpha)\boldsymbol{y}) \le \alpha f(\boldsymbol{x}) + (1 - \alpha) f(\boldsymbol{y})$$

If this inequality is strict(严格小于), the function is called strictly convex. We call f concave if $-f$ is Convex.

之前学的优化都是 gradients, 对于不光滑的, 需要更多的技巧.

Lemma V.4 (Jensen's Inequality). *For ang $\boldsymbol{x}_1, \dots, \boldsymbol{x}_m \in \operatorname{dom} f$ and positive coefficients $\alpha_1, \dots, \alpha_m$ such that*

$$\sum_{i=1}^m \alpha_i = 1,\ \alpha \ge 0,\ i = 1, \dots, m \tag{V.1}$$

we have

$$f(\sum_{i=1}^m \alpha_i \boldsymbol{x}_i) \le \alpha_i f(\boldsymbol{x}_i)$$

A point $\boldsymbol{x} = \sum_{i=1}^m \alpha_i \boldsymbol{x}_i$ with positive coefficients α_i satisfying the normalizing condition V.1 is called a convex combination(凸组合) of points $\{\boldsymbol{x}_i\}_{i=1}^m$.

Corollary V.5. *Let $\boldsymbol{x}$ be a convex combination of points $\boldsymbol{x}_1, \dots, \boldsymbol{x}_m$, then*

$$f(\boldsymbol{x}) \le \max_{1 \le i \le m} f(\boldsymbol{x}_i)$$

Corollary V.6. *Let*

$$\begin{aligned} \Delta &= Conv\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_m\} \\ &= \left\{ \boldsymbol{x} = \sum_{i=1}^m \alpha_i \boldsymbol{x}_i \,\middle|\, \alpha_i \ge 0,\ \sum_{i=1}^m \alpha_i = 1 \right\} \end{aligned}$$

Conv means convex combination.

Then $\max_{\boldsymbol{x} \in \Delta} f(\boldsymbol{x}) = \max_{1 \le i \le m} f(\boldsymbol{x}_i)$

Theorem V.7. *A function f is convex iff $\forall \boldsymbol{x}, \boldsymbol{y} \in \operatorname{dom} f$ and $\beta \ge 0$ such that $\boldsymbol{y} + \beta(\boldsymbol{y} - \boldsymbol{x}) \in \operatorname{dom} f$, we have*

$$f(\boldsymbol{y} + \beta(\boldsymbol{y} - \boldsymbol{x})) \ge f(\boldsymbol{y}) + \beta(f(\boldsymbol{y}) - f(\boldsymbol{x}))$$

Theorem V.8. *A function f is convex iff its epigraph(函数线以上的区域)*

$$epi(f) = \{(\boldsymbol{x}, t) \in \operatorname{dom} f \times \mathbb{R} | t \ge f(\boldsymbol{x})\}$$

is a convex set.

Theorem V.9. *If a function f is convex, then all level sets(层集)*

$$\mathcal{L}_f(\beta) = \{\boldsymbol{x} \in \operatorname{dom} f | f(\boldsymbol{x}) \le \beta\}$$

are either convex or empty.

1.3 Other Properties

Definition V.10. *A function f is called closed and convex if its epigraph is a closed set.*

Theorem V.11. *If convex function f is closed, then all its level sets are either empty or closed.*

pathological(病态)

2. Operation with Convex Function

2.1 Invariant Operations

Definition V.12 (Lower Semicontinuity). *A function f is lower-semicontinuity at a given vector $\bar{\boldsymbol{x}}$ if for every sequence $\{\boldsymbol{x}_k\}$ converging to $\bar{\boldsymbol{x}}$, we have*

$$\lim \inf_{k\to\infty} f(\boldsymbol{x}_k) \geq f(\bar{\boldsymbol{x}})$$

We say that f is lower-semicontinuity over a set X if f is lower-semicontinuity at every $\boldsymbol{x} \in X$.

Note:

$$\lim \inf_{k\to\infty} \boldsymbol{x}_k \iff \lim_{k\to\infty} \left(\inf_{m\geq k} \boldsymbol{x}_m \right)$$

Theorem V.13. *For a function $f : \mathbb{R}^n \to \mathbb{R} \cup \{-\infty, +\infty\}$ the following statements are equivalent*

1) *f is closed*
2) *Every level set of f is closed*
3) *f is lower-semicontinuity oevr $\mathbb{R}^n$*

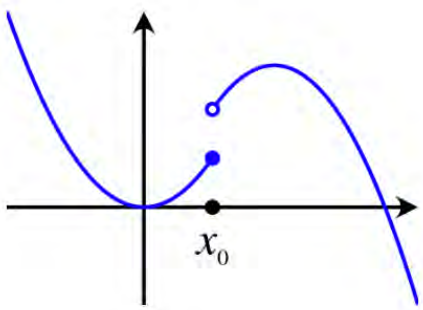

Figure V.1: f

Let us describe a set of invariant operations to create more complicated objects.

Theorem V.14. *Let function f_1 and f_2 be closed and convex and let $\beta \geq 0$. Then all functions below are closed and convex:*

1) $f(\boldsymbol{x}) = \beta f_1(\boldsymbol{x}),\ \operatorname{dom} f = \operatorname{dom} f_1$
2) $f(\boldsymbol{x}) = f_1(\boldsymbol{x}) + f_2(\boldsymbol{x}),\ \operatorname{dom} f = (\operatorname{dom} f_1) \cap (\operatorname{dom} f_2)$
3) $f(\boldsymbol{x}) = \max\{f_1(\boldsymbol{x}), f_2(\boldsymbol{x})\},\ \operatorname{dom} f = (\operatorname{dom} f_1) \cap (\operatorname{dom} f_2)$

The following theorem demonstrates that convexity is an affine-invariant property.

Theorem V.15. *Let function $\phi(\boldsymbol{y}),\ \boldsymbol{y} \in \mathbb{R}^n$ be convex and closed. Consider a linear operator*

$$\mathcal{A}(\boldsymbol{x}) = A\boldsymbol{x} + b : \mathbb{R}^n \to \mathbb{R}^m$$

Then $f(\boldsymbol{x}) = \phi(\mathcal{A}(\boldsymbol{x}))$ is a closed and convex function with domain

$$\operatorname{dom} f = \{\boldsymbol{x} \in \mathbb{R}^n | \mathcal{A}(\boldsymbol{x}) \in \operatorname{dom} \phi\}$$

The next theorem is one of the main suppliers of convex functions with implicit structure.

Theorem V.16. *Let Δ be some set and*

$$f(\boldsymbol{x}) = \sup_{\boldsymbol{y}} \{\phi(\boldsymbol{y},\ \boldsymbol{x}) | \boldsymbol{y} \in \Delta\}$$

Suppose that for any fixed $\boldsymbol{y} \in \Delta$, the function $\phi(\boldsymbol{y},\ \boldsymbol{x})$ is closed and convex in $\boldsymbol{x}$. Then $f(\boldsymbol{x})$ is a closed convex function with domain

$$\operatorname{dom} f = \left\{ \boldsymbol{x} \in \bigcap_{\boldsymbol{y}\in\Delta} \operatorname{dom} \phi(\boldsymbol{y},\ \cdot) \,\middle|\, \exists \gamma : \phi(\boldsymbol{y}, \boldsymbol{x}) \leq \gamma,\ \forall \boldsymbol{y} \in \Delta \right\}$$

3. Continuity and Differentiability

3.1 Continuity of Convex Function

Lemma V.17. *Let f be convex and $\boldsymbol{x}_0 \in int(\operatorname{dom} f)$. Then f is locally upper bounded at $\boldsymbol{x}_0$.*

Theorem V.18. *Let f be convex and $\boldsymbol{x}_0 \in int(\operatorname{dom} f)$. Then f is locally Lipschitz continuous at $\boldsymbol{x}_0$.*

Remark: 构造 $\boldsymbol{z}$ 没有超出领域, $\boldsymbol{y}$ 是 $\boldsymbol{x}_0$ 与 $\boldsymbol{z}$ 的凸组合.

3.2 Differentiability of Convex Function

Definition V.19. *Let $\boldsymbol{x} \in \operatorname{dom} f$. We call f differentiable in a direction p at point $\boldsymbol{x}$ if the following limit exists*

$$f'(\boldsymbol{x}; \boldsymbol{p}) = \lim_{\alpha \downarrow 0} \frac{1}{\alpha} [f(\boldsymbol{x} + \alpha \boldsymbol{p}) - f(\boldsymbol{x})]$$

The value $f'(\boldsymbol{x}; \boldsymbol{p})$ is called the directional derivative of f at $\boldsymbol{x}$.

Theorem V.20. *Convex f is differentiable in any direction at any interior point of its domain.*

Lemma V.21. *Let f be a convex function and $\boldsymbol{x} \in int(\operatorname{dom} f)$. Then $f'(\boldsymbol{x}; \boldsymbol{p})$ is a convex function of $\boldsymbol{p}$, which is homogeneous of degree 1. For any $\boldsymbol{y} \in \operatorname{dom} f$ we have*

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + f'(\boldsymbol{x}; \boldsymbol{y} - \boldsymbol{x})$$

4. Separation Theorem

4.1 Projection

Definition V.22 (Hyperplane). *Let Q be a convex set. We say that hyperplane*

$$\mathcal{H}(\boldsymbol{g}, \gamma) = \{\boldsymbol{x} \in \mathbb{R}^n | \langle \boldsymbol{g}, \boldsymbol{x} \rangle = \gamma\},\ \boldsymbol{g} \neq 0$$

is supporting to Q if any $\boldsymbol{x} \in Q$ satisfies inequality $\langle \boldsymbol{g}, \boldsymbol{x} \rangle \leq \gamma$.

We say that the hyperplane $\mathcal{H}(\boldsymbol{g}, \gamma)$ separates a point $\boldsymbol{x}_0$ from Q if

$$\langle \boldsymbol{g}, \boldsymbol{x} \rangle \leq \gamma \leq \langle \boldsymbol{g}, \boldsymbol{x}_0 \rangle \tag{V.2}$$

$\forall \boldsymbol{x} \in Q$. If the second inequality in V.2 is strict, we call the separation strict.

Definition V.23 (Projection)**.** *Let Q be a closed set and $\boldsymbol{x}_0 \in \mathbb{R}^n$. Denote*

$$\pi_Q(\boldsymbol{x}_0) = \operatorname{argmin} \{\|\boldsymbol{x} - \boldsymbol{x}_0\| : \boldsymbol{x} \in Q\}$$

We call $\pi_Q(\boldsymbol{x}_0)$ the projection of point $\boldsymbol{x}_0$ onto the set Q.

Theorem V.24. *If Q is a convex set, then there exists a unique projection $\pi_Q(\boldsymbol{x}_0)$*

Remark: It's clear that $\pi_Q(\boldsymbol{x}_0) = \boldsymbol{x}_0 \iff \boldsymbol{x}_0 \in Q$.

Lemma V.25. *Let Q be a closed and convex set and $\boldsymbol{x}_0 \notin Q$. Then $\forall \boldsymbol{x} \in Q$, we have*

$$\langle \pi_Q(\boldsymbol{x}_0) - \boldsymbol{x}_0, \boldsymbol{x} - \pi_Q(\boldsymbol{x}_0) \rangle \geq 0$$

从 $\boldsymbol{x}$ 到 $\pi_Q(\boldsymbol{x}_0)$ 的向量与从 $\pi_Q(\boldsymbol{x}_0)$ 到 $\boldsymbol{x}_0$ 的向量的夹角是锐角.

Lemma V.26. *$\forall \boldsymbol{x} \in Q$, we have*

$$\|\boldsymbol{x} - \pi_Q(\boldsymbol{x}_0)\|^2 + \|\pi_Q(\boldsymbol{x}_0) - \boldsymbol{x}_0\|^2 \leq \|\boldsymbol{x} - \boldsymbol{x}_0\|^2$$

4.2 Main Theorems

Theorem V.27. *Let Q be a closed set and $\boldsymbol{x}_0 \notin Q$. Then there exists a hyperplane $\mathcal{H}(\boldsymbol{g}, \gamma)$, which strictly separates $\boldsymbol{x}_0$ from Q. Namely, we can take*

$$\boldsymbol{g} = \boldsymbol{x}_0 - \pi_Q(\boldsymbol{x}_0) \neq 0$$
$$\gamma = \langle \boldsymbol{x}_0 - \pi_Q(\boldsymbol{x}_0), \pi_Q(\boldsymbol{x}_0) \rangle$$

$\psi(\boldsymbol{g}) = \sup \{\langle \boldsymbol{g}, \boldsymbol{x} \rangle \, | \boldsymbol{x} \in Q\}$

Corollary V.28. *Let Q_1 and Q_2 be two closed convex sets.*

1) If $\forall \boldsymbol{g} \in \operatorname{dom} \psi_{Q_2}$ we have $\psi_{Q_1}(\boldsymbol{g}) \leq \psi_{Q_2}(\boldsymbol{g})$, then $Q_1 \subseteq Q_2$.

2) Let $\operatorname{dom} \psi_{Q_1} = \operatorname{dom} \psi_{Q_2}$ and $\forall \boldsymbol{g} \in \operatorname{dom} \psi_{Q_1}$, we have $\psi_{Q_1}(\boldsymbol{g}) = \psi_{Q_2}(\boldsymbol{g})$. Then $Q_1 \equiv Q_2$.

Theorem V.29. *Let Q be a closed and convex set, and $\boldsymbol{x}_0$ belong to boundary of set Q. Then there exists a hyperplane $\mathcal{H}(\boldsymbol{g}, \gamma)$, supporting to Q and passing through $\boldsymbol{x}_0$.*

Such a vector $\boldsymbol{g}$ is called supporting to Q at $\boldsymbol{x}_0$.

5. Subgradient

5.1 Definition of Subgradient

Definition V.30. *Let f be a convex function. A vector $\boldsymbol{g}$ is called a subgradient of function f at point $\boldsymbol{x}_0 \in \operatorname{dom} f$ if for any $\boldsymbol{x} \in \operatorname{dom} f$, we have*

$$f(\boldsymbol{x}) \geq f(\boldsymbol{x}_0) + \langle \boldsymbol{g}, \boldsymbol{x} - \boldsymbol{x}_0 \rangle$$

The set of all subgradient of f at $\boldsymbol{x}_0$, is called the subdifferential of function f at point $\boldsymbol{x}_0$.

Subgradient and Convexity

Lemma V.31. *Let for any $\boldsymbol{x} \in \operatorname{dom} f$ subdifferential $\partial f(\boldsymbol{x})$ be nonempty. Then f is a convex function.*

Theorem V.32. *Let f be closed and convex and $\boldsymbol{x}_0 \in int(\operatorname{dom} f)$. Then $\partial f(\boldsymbol{x}_0)$ is a nonempty bounded set.*

Theorem V.33. *Let f be a closed convex function. For any $\boldsymbol{x}_0 \in int(\operatorname{dom} f)$ and $\boldsymbol{p} \in \mathbb{R}^n$, we have*

$$f'(\boldsymbol{x}_0; \boldsymbol{p}) = \max\{\langle \boldsymbol{g}, \boldsymbol{p} \rangle \, | \boldsymbol{g} \in \partial f(\boldsymbol{x}_0)\}$$

5.2 Properties of Subgradient

Theorem V.34. *We have*

$$f(\boldsymbol{x}^*) = \min_{\boldsymbol{x} \in \operatorname{dom} f} f(\boldsymbol{x}) \iff 0 \in \partial f(\boldsymbol{x}^*)$$

Theorem V.35. *For any $\boldsymbol{x}_0 \in \operatorname{dom} f$, all vector $\boldsymbol{g} \in \partial f(\boldsymbol{x}_0)$ are supporting to the level set $\mathcal{L}_f(f(\boldsymbol{x}_0))$:*

$$\langle \boldsymbol{g}, \boldsymbol{x}_0 - \boldsymbol{x} \rangle \geq 0$$
$$\forall \boldsymbol{x} \in \mathcal{L}_f(f(\boldsymbol{x}_0)) \equiv \{\boldsymbol{x} \in \operatorname{dom} f : f(\boldsymbol{x}) \leq f(\boldsymbol{x}_0)\}$$

Corollary V.36. *Let $Q \subseteq \operatorname{dom} f$ be a closed convex set, $\boldsymbol{x}_0 \in Q$ and*

$$\boldsymbol{x}x^* = \operatorname{argmin}\{f(\boldsymbol{x}) | \boldsymbol{x} \in Q\}$$

Then $\forall \boldsymbol{g} \in \partial f(\boldsymbol{x}_0)$, we have $\langle \boldsymbol{g}, \boldsymbol{x}_0 - \boldsymbol{x}^* \rangle \geq 0$.

5.3 Rules for Computing

Lemma V.37. *Let f be closed and convex. Assume that it's differentiable on its domain. Then*

$$\partial f(\boldsymbol{x}) = \{\nabla f(\boldsymbol{x})\}$$

$\forall \boldsymbol{x} \in int(\operatorname{dom} f)$

Lemma V.38. *Let $f(\boldsymbol{y})$ be closed and convex with $\operatorname{dom} f \subseteq \mathbb{R}^m$. Consider a linear operator*

$$\mathcal{A}(\boldsymbol{x}) = A\boldsymbol{x} + \boldsymbol{b} : \mathbb{R}^n \to \mathbb{R}^m$$

Then $\phi(\boldsymbol{x}) = f(\mathcal{A}(\boldsymbol{x}))$ is closed convex function with domain $\operatorname{dom}\phi = \{\boldsymbol{x}|\mathcal{A}(\boldsymbol{x}) \in \operatorname{dom} f\}$. $\forall \boldsymbol{x} \in int(\operatorname{dom}\phi)$, we have

$$\partial\phi(\boldsymbol{x}) = A^\top \partial f(\mathcal{A}(\boldsymbol{x}))$$

Lemma V.39. *Let $f_1(\boldsymbol{x})$ and $f_2(\boldsymbol{x})$ be closed convex function and $\alpha_1, \alpha_2 \geq 0$. Then function $f(\boldsymbol{x}) = \alpha_1 f_1(\boldsymbol{x}) + \alpha_2 f_2(\boldsymbol{x})$ is closed and convex and*

$$\partial f(\boldsymbol{x}) = \alpha_1 \partial f_1(\boldsymbol{x}) + \alpha_2 \partial f_2(\boldsymbol{x})$$

$\forall \boldsymbol{x} \in int(\operatorname{dom} f) = int(\operatorname{dom} f_1) \cap int(\operatorname{dom} f_2)$

Lemma V.40. *Let functions $f_i(\boldsymbol{x})$, $i = 1, \ldots, m$ be closed and convex. Then function*

$$f(\boldsymbol{x}) = \max_{1\leq i\leq m} f_i(\boldsymbol{x})$$

is also closed and convex. $\forall \boldsymbol{x} \in int(\operatorname{dom} f) = \bigcap_{i=1}^{m} int(\operatorname{dom} f_i)$, we have

$$\partial f(\boldsymbol{x}) = Conv\{\partial f_i(\boldsymbol{x}) | i \in I(\boldsymbol{x})\}$$

where $I(\boldsymbol{x}) = \{i : f_i(\boldsymbol{x}) = f(\boldsymbol{x})\}$

Lemma V.41. *Let Δ be a set and $f(\boldsymbol{x}) = \sup\{\phi(\boldsymbol{y}, \boldsymbol{x}) | \boldsymbol{y} \in \Delta\}$. Suppose that for any fixed $\boldsymbol{y} \in \Delta$, the function $\phi(\boldsymbol{y}, \boldsymbol{x})$ is closed and convex in $\boldsymbol{x}$. Then $f(\boldsymbol{x})$ is closed convex.*

Moreover, for any $\boldsymbol{x}$ from

$$\operatorname{dom} f = \{\boldsymbol{x} \in \mathbb{R}^n | \exists \gamma : \phi(\boldsymbol{y}, \boldsymbol{x}) \leq \gamma,\ \forall \boldsymbol{y} \in \Delta\}$$

we have

$$\partial f(\boldsymbol{x}) \supseteq Conv\{\partial \phi_{\boldsymbol{x}}(\boldsymbol{y}, \boldsymbol{x}) | \boldsymbol{y} \in I(\boldsymbol{x})\}$$

where $I(\boldsymbol{x}) = \{\boldsymbol{y} | \phi(\boldsymbol{y}, \boldsymbol{x}) = f(\boldsymbol{x})\}$

Theorem V.42. *Let $\|\cdot\|$ be a vector norm in $\mathbb{R}^n$, then*

$$\partial \|\cdot\| = \left\{ V(\boldsymbol{x}) \triangleq \{\boldsymbol{v} \in \mathbb{R}^n | \langle \boldsymbol{v}, \boldsymbol{x} \rangle = \|\boldsymbol{x}\|,\ \|\boldsymbol{v}\| \leq 1\} \right\}$$

where $\|\boldsymbol{v}\|_*$ is the dual norm of $\|\cdot\|$, defined as

$$\|\boldsymbol{v}\|_* \triangleq \sup_{\|\boldsymbol{u}\|\leq 1} \langle \boldsymbol{x}, \boldsymbol{u} \rangle$$

6. General Lower Complexity Bounds

6.1 Problem Class

We have

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

Table V.1: Problem Class

Model
Oracle
Approximate solution
Method

6.2 Resisting Oracle

Special Function family

Let us fix some constant $\mu > 0$ and $\gamma > 0$,

$$f_k(\boldsymbol{x}) = \gamma \max_{1\leq i\leq k} \boldsymbol{x}^{(i)} + \frac{\mu}{2} \|\boldsymbol{x}\|^2,\ k = 1, \ldots, n$$

$$\partial f_k(\boldsymbol{x}) = \mu\boldsymbol{x} + \gamma Conv\{e_i | i \in I(\boldsymbol{x})\}$$
$$I(\boldsymbol{x}) = \left\{ 1 \leq j \leq k, \boldsymbol{x}^{(j)} = \max_{1\leq i\leq k} \boldsymbol{x}^{(i)} \right\}$$

1) Loacl Lipschitz Continuity of f_k

Therefore $\forall \boldsymbol{x}, \boldsymbol{y} \in B_2(0, \rho), \rho > 0$, and $g_k(\boldsymbol{y}) \in \partial f_k(\boldsymbol{y})$, we have

$$\begin{aligned} f_k(\boldsymbol{y}) - f_k(\boldsymbol{x}) &\leq \langle g_k(\boldsymbol{y}), \boldsymbol{y} - \boldsymbol{x} \rangle \\ &\leq \|g_k(\boldsymbol{y})\| \cdot \|\boldsymbol{y} - \boldsymbol{x}\| \\ &\leq (\mu\rho + \gamma) \|\boldsymbol{y} - \boldsymbol{x}\| \end{aligned}$$

Thus f_k is Lipschitz continuous on $B_2(0, \rho)$ with Lipschitz constant

$$M = \mu\rho + \gamma$$

2) minimum of f_k

Further, consider the point $\boldsymbol{x}_k^*$ with the coordinates

$$(\boldsymbol{x}_k^*)^{(i)} = \begin{cases} -\frac{\gamma}{\mu k} & 1 \leq i \leq k \\ 0 & K = 1 \leq i \leq n \end{cases}$$

It's easy to check that $0 \in \partial f_k(\boldsymbol{x}_k^*)$, and therefore $\boldsymbol{x}_k^*$ is the minimum of $f_k(\boldsymbol{x})$. Note that

$$R_k = \|\boldsymbol{x}_k^*\| = \frac{\gamma}{\mu\sqrt{k}}$$
$$f_k^* = -\frac{\gamma^2}{\mu k} + \frac{\mu}{2} R_k^2 = -\frac{\gamma^2}{2\mu k}$$

3) The subgradient of f_k

4) Characteristic of $\boldsymbol{x}_{i+1}$ generated by resisting oracle
Let us choose starting point $\boldsymbol{x}_0 = 0$. Denote

$$\mathbb{R}^{p,n} = \left\{\boldsymbol{x} \in \mathbb{R}^n | \boldsymbol{x}^{(i)} = 0, p+1 \leq i \leq n\right\}$$

6.3 Lower Bound

Theorem V.43. *For nay class $\mathcal{P}(\boldsymbol{x}_0, R, M)$ and any k, $0 \leq k \leq n-1$, there exists a function $f \in \mathcal{P}(\boldsymbol{x}_0, R, M)$ such that*

$$\boldsymbol{x}_k \in \boldsymbol{x}_0 + Lin\left\{g(\boldsymbol{x}_0), \ldots, g(\boldsymbol{x}_{k-1})\right\}$$

for any optimization scheme, which generates a sequence $\{\boldsymbol{x}_k\}$ satisfying the condition

$$f(\boldsymbol{x}_k) - f^* \geq \frac{MR}{2(1+\sqrt{k+1})}$$

7. Subgradient Method

7.1 property of Subgradient

$$\min\left\{f(\boldsymbol{x}) | \boldsymbol{x} \in Q\right\}$$

$$\langle g(\boldsymbol{x}), \boldsymbol{x} - \boldsymbol{x}^* \rangle \geq 0$$

- The distance between $\boldsymbol{x}$ and $\boldsymbol{x}^*$ is decreasing in the direction $-g(\boldsymbol{x})$
- In equality (4) cuts $\mathbb{R}^n$ into two galf-spaces. Only one of them contains $\boldsymbol{x}^*$

7.2 Main Lemma

Let us fix some $\bar{\boldsymbol{x}} \in \mathbb{R}^n$. For $\boldsymbol{x} \in \mathbb{R}^n$ with $g(\boldsymbol{x}) \neq 0$ define

$$v_f(\bar{\boldsymbol{x}}, \boldsymbol{x}) = \frac{1}{\|g(\boldsymbol{x})\|} \langle g(\boldsymbol{x}), \boldsymbol{x} - \bar{\boldsymbol{x}} \rangle$$

If $g(\boldsymbol{x}) = 0$, then define $v_f(\bar{\boldsymbol{x}}, \boldsymbol{x}) = 0$. Clearly, $v_f(\bar{\boldsymbol{x}}, \boldsymbol{x}) \leq \|\boldsymbol{x} - \bar{\boldsymbol{x}}\|$

Let us introduce a function that measures the variation of function f with respect to the point $\bar{\boldsymbol{x}}$. For $t \geq 0$, define

$$\omega_f(\bar{\boldsymbol{x}}; t) = \max\left\{f(\boldsymbol{x}) - f(\bar{\boldsymbol{x}}) | \, \|\boldsymbol{x} - \bar{\boldsymbol{x}}\| \leq t\right\}$$

If $t < 0$, we set $\omega_f(\bar{\boldsymbol{x}}; t) = 0$. Clearly, the function ω_f possesses the following Properties:

1) For all $t < 0$, $\omega_f(\bar{\boldsymbol{x}}; 0) = 0$
2) $\omega_f(\bar{\boldsymbol{x}}; t)$ is a nondecreasing function of t, $t \in \mathbb{R}^1$
3) $f(\boldsymbol{x}) - f(\bar{\boldsymbol{x}}) \leq \omega_f(\bar{\boldsymbol{x}}; \|\boldsymbol{x} - \bar{\boldsymbol{x}}\|)$

Lemma V.44. *For any $\boldsymbol{x} \in \mathbb{R}^n$, we have*

$$f(\boldsymbol{x}) - f(\bar{\boldsymbol{x}}) \leq \omega_f(\bar{\boldsymbol{x}}; v_f(\bar{\boldsymbol{x}}, \boldsymbol{x}))$$

If $f(\boldsymbol{x})$ is Lipschitz continuous on $B_2(\bar{\boldsymbol{x}}, R)$ with some M, then

$$f(\boldsymbol{x}) - f(\bar{\boldsymbol{x}}) \leq M(v_f(\bar{\boldsymbol{x}}, \boldsymbol{x}))_+$$

$\forall \boldsymbol{x} \in \mathbb{R}^n$ *with* $\omega_f(\bar{\boldsymbol{x}}; \boldsymbol{x}) \leq \mathbb{R}$

Definition V.45. *Let $\{\boldsymbol{x}_i\}_{i=0}^{\infty}$ be a sequence in Q. Define*

$$S_k = \{\boldsymbol{x} \in Q | \, \langle g(\boldsymbol{x}_i), \boldsymbol{x}_i - \boldsymbol{x} \rangle \geq 0, i = 0, \ldots, k\}$$

We call this set the location set of problem (3) generated by sequence $\{\boldsymbol{x}_i\}_{i=0}^{\infty}$.

Note that in view of inequality (4) $\forall k \geq 0$, we have $\boldsymbol{x}^* \in S_k$. Denote

$$v_i = v_f(\boldsymbol{x}^*; \boldsymbol{x}_i) \geq 0$$
$$v_k^* = \min_{0 \leq i \leq k} v_i$$

Thus

$$v_k^* = \max\left\{r | \, \langle g(\boldsymbol{x}_i), \boldsymbol{x}_i - \boldsymbol{x} \rangle \geq 0, i = 0, \ldots, k, \forall \boldsymbol{x} \in B_2(\boldsymbol{x}^*, r)\right\}$$

Lemma V.46. *Let $f_k^* = \min\limits_{0 \leq i \leq k} f(\boldsymbol{x}_i)$. Then $f_k^* - f^* \leq \omega_f(\boldsymbol{x}^*; v_k^*)$*

7.3 Scheme for Non-smooth Problem

7.4 Main Theorem

Theorem V.47. *Let f be Lipschitz continuous on $B_2(\boldsymbol{x}^*, R)$ with constant M and $\boldsymbol{x}_0 \in B(\boldsymbol{x}^*, R)$. Then*

$$f_k^* - f^* \leq M \frac{R^2 + \sum_{i=0}^k h_i^2}{2\sum_{i=0}^k h_i}$$

8. Frank-Wolfe Algorithm

8.1 Problems

Algorithm 1 Frank-Wolfe Algorithm

8.2 Examples

8.3 Convergence Theory

Definition V.48. *We will say that $\boldsymbol{x}^* \in \mathcal{D}$ is a stationary point if*

$$\langle \nabla f(\boldsymbol{x}^*), \boldsymbol{x} - \boldsymbol{x}^* \rangle \geq 0$$

$\forall \boldsymbol{x} \in \mathcal{D}$

Definition V.49. *We denote by g_k the Frank-Wolfe gap, defined as*

$$g_k \equiv \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{x} - \boldsymbol{s}_k \rangle$$

Lemma V.50.

VI Beyond The Black-box Model

1. Proximal Gradient Method

1.1 Proximal Operator

1.2 Properties of Proximal Operator

1.3 Analysis for Proximal Gradient Method

1.4 Accelerated Proximal Gradient Method

1.5 Special case: Proximal Point Method

2. Douglas-Rachford Splitting

2.1 Different Setting for Convex Problem

2.2 Fixed Point for Nonsmooth Composition

2.3 Splitting Algorithm

3. Duality Principle

3.1 Duality Principle

Example: Duality in Linear Programs

Primal LP:

$$\min_{\boldsymbol{x}} c^\top \boldsymbol{x}$$

Duality Problem

$$\min_{\boldsymbol{x}} f(\boldsymbol{x}),\ \text{s.t. } g_i(\boldsymbol{x}) \le 0,\ i = 1, \dots, m$$

4. Lagrangian Duality and Algorithms

4.1 Lagrangian Duality

$$\begin{aligned} &\min_{\boldsymbol{x}} f(\boldsymbol{x}) \\ \text{s.t. } &h(\boldsymbol{x}) \le 0 \\ &l_i(\boldsymbol{x}) = 0 \end{aligned}$$

因为 min max 的互换, 所以有个 gap

$$\min_{\boldsymbol{x}\in\mathcal{X}} f(\boldsymbol{x})$$

where f is convex and nonsmooth. Approximate $f(\boldsymbol{x})$. by a smooth and convex $f_u(\boldsymbol{x})$.

$$\min_{\boldsymbol{x}\in\mathcal{X}} f_u(\boldsymbol{x})$$

where f_u is a L_u-Lipschitz continuous, smooth and convex.

e.g. (Motivation)

Huber function

$$f_u(\boldsymbol{x}) = \begin{cases} \frac{x^2}{2u} & |\boldsymbol{x}| \le u \\ |\boldsymbol{x}| - \frac{u}{2} & |\boldsymbol{x}| > u \end{cases}$$

1) $f_u(\boldsymbol{x})$ is clearly continuous and differentiable everywhere.

2) $f(\boldsymbol{x}) - \frac{u}{2} \le f_u(\boldsymbol{x}) \le f(\boldsymbol{x})$

3) If $u \to 0$, then $f_u(\boldsymbol{x}) \to f(\boldsymbol{x})$

4) $|f_u''(\boldsymbol{x})| \le \frac{1}{u}$. This implies that $f_u(\boldsymbol{x})$ is $\frac{1}{u}$-Lipschitz continuous.

Major Techniques:

1) Nesterov's smoothing Techniques: proximity function(近邻函数)

2) Moreau-Yosida smoothing regularization: envelope(包络)

3) Ben-Tal-Teboulle smoothing based on recession function

4) Randomized smoothing

6.2 Nesterov's Smoothing

$$\min_{\boldsymbol{x}\in\mathcal{X}} f(\boldsymbol{x}) \iff f_u(\boldsymbol{x}) = \max_{\boldsymbol{y}\in\operatorname{dom} f^*} \left\{\boldsymbol{x}^\top \boldsymbol{y} - f^*(\boldsymbol{y}) - ud(\boldsymbol{y})\right\}$$

Assume that function f can be represented by

$$f(\boldsymbol{x}) = g(A\boldsymbol{x} + b) \triangleq \max_{y\in\mathcal{Y}} \{\langle A\boldsymbol{x} + b, \boldsymbol{y}\rangle - \phi(\boldsymbol{y})\}$$

where $\phi(\boldsymbol{y})$ is a convex and continuous function and $\mathcal{Y}$ is a convex and compact set.

Proximity Function The function $d(\boldsymbol{y})$

Proposition

Theorem VI.1. *For any $u > 0$, let $D_{\mathcal{Y}}^2 = \max_{\boldsymbol{y}\in\mathcal{Y}} d(\boldsymbol{y})$, we have*

$$f(\boldsymbol{x}) - uD_{\mathcal{Y}}^2 \le f_u(\boldsymbol{x}) \le f(\boldsymbol{x})$$

Analysis of Nesterov's smoothing

1)

e.g.

6.3 Moreau-Yosida Regularization

7. Generalized Distance: Mirror Descent

两大动机: general, bound 界更好

7.1 Motivation

7.2 Bregman Divergence

7.3 Mirror Descent

VII Stochastic Optimization