优化算法复杂度分析简介

王奇超[1], 文再文[*1], 蓝光辉[†2], and 袁亚湘[‡3]

[1] 北京大学北京国际数学研究中心

[2] 乔治亚理工学院 H. Milton Stewart 工业和系统工程学院

[3] 中国科学院数学与系统科学研究院，计算数学与科学工程计算研究所

摘要

优化算法的收敛性分析是优化中很重要的一个领域。然而收敛性并不足以作为比较不同算法效率的标准，因此我们需要另外一套衡量优化问题难易程度以及优化算法效率高低的理论。这套理论被称为优化算法的复杂度分析理论。本论文分共为五个部分。第一章介绍复杂度分析的背景和理论框架，给出复杂度分析的定义，方法和例子，并总结本论文中的复杂度结论。第二章介绍光滑优化问题的复杂度分析，给出了不同优化问题的复杂度上界和下界，并给出加速梯度法收敛性分析的框架。第三章介绍非光滑优化问题的复杂度上界，介绍了次梯度法，重心法，椭球法和近似点梯度法的复杂度分析。第四章介绍了条件梯度法的复杂度分析，介绍了条件梯度法的复杂度上界和下界，以及加速条件梯度法的框架。第五章介绍随机优化算法的复杂度分析，比较了随机优化算法在凸和非凸问题下收敛的置信水平以及复杂度。

目录

[*] wenzw@pku.edu.cn

[†] george.lan@isye.gatech.edu

[‡] yyx@lsec.cc.ac.cn

1 复杂度分析的理论基础

1.1 引言

优化算法的复杂度分析一直以来是我们理解优化算法和设计新的优化算法的工具。考虑通常形式下的优化问题，记做 $\mathcal{P}$：

$$f^* = \min_{x \in X} \ f(x). \tag{1.1}$$

根据约束集合 X 类型和目标函数 $f(x)$ 类型对上述优化问题进行分类：

- 约束/无约束优化问题：$X \subset \mathbb{R}^n / X \equiv \mathbb{R}^n$；

- 光滑/非光滑优化问题：$f(x)$ 在 X 上可/不可导；
- 凸/强凸/非凸优化问题：$f(x)$ 是凸/强凸/非凸函数；
- 随机优化问题：$f(x)=\mathbb{E}_{\xi}[F(x,\xi)]$。

为了分析优化算法在求解问题 (1.1) 时的效率，我们将引入优化算法的复杂度分析的概念。

将 (1.1) 的最优解，最优值分别记为 x^*，f^*。我们知道一旦给定 $f(x)$ 和约束集合 X 的具体形式，我们可以找到不同的数值算法，记做 $\mathcal{S}$，来求得 x^* 和 f^* 给定误差 ϵ 的近似值。为了寻找更好的求解算法，我们需要衡量不同算法 $\mathcal{S}$ 的效率。而复杂度分析理论就提供了一套衡量算法效率的框架。

当 (1.1) 中 $f(x)$ 和 X 具体形式给定时，我们将 (1.1) 称为问题 $\mathcal{P}$。把具备某种共同性质的问题 $\mathcal{P}$ 集合称为问题集合 $\mathcal{F}$。我们定义问题 $\mathcal{P}$ 的 **解算法集合 $\mathcal{M}:=\mathcal{M}(\mathcal{P},\epsilon)$ (Solution methods set)**，其中包含所有可以求解 $\mathcal{P}$ 到某一给定误差精度 ϵ 的**解算法 (Solution method)**，记做 $\mathcal{S}:=\mathcal{S}(\mathcal{P},\epsilon)\in\mathcal{M}$。一个自然的问题是，在 $\mathcal{P}$ 的解算法集合 $\mathcal{M}$ 中是否存在效率最好的解算法 $\mathcal{S}$?

当然，我们还没有正式给出衡量算法效率的度量。由于无法完全穷尽 $\mathcal{M}$ 的集合，即使我们设定了算法效率的度量，也很难找到效率最好的算法。事实上，考虑这样一个求解问题 $\mathcal{P}$ 的解算法 $\mathcal{S}_0$。当调用 $\mathcal{S}_0$ 的时候，它只执行一个操作：返回 $x^*=0$。对于最优解不是 $x^*=0$ 的某些问题 $\mathcal{P}$，$\mathcal{S}_0$ 不是它的解算法，但是当 $\mathcal{P}$ 的最优解刚好是 $x^*=0$ 时，对于该问题 $\mathcal{P}$ 来说，$\mathcal{S}_0$ 在其解算法集合 $\mathcal{M}$ 中的效率是无法超越的。然而 $\mathcal{S}_0$ 这样的解算法并不是我们寻找的效率最好的解算法。

因此，我们无法寻找求解某一特定问题 $\mathcal{P}$ 时效率最高的解算法，但是我们可以寻找求解某一类问题集合 $\mathcal{F}\ni\mathcal{P}$ 时候效率最好的解算法。事实上，设计数值优化算法的目的都是想用它解决某一类具有相同特征的问题。因此，我们需要定义解算法 $\mathcal{S}$ 在一类问题集合 $\mathcal{F}$ 上的表现来衡量一个解算法的效率。

本章我们介绍优化算法复杂度分析的理论基础。复杂度分析的理论分为三个部分，第一部分是对所要求解的优化问题进行分类，建立问题模型；第二部分是对求解优化问题的数值算法进行抽象，建立算法模型；第三部分是定义算法模型在问题模型上的度量，即复杂度。这一章我们先分别介绍复杂度分析的三个部分，最后用两个具体的例子介绍复杂度分析的分析方法。

1.2 复杂度分析的问题模型

复杂度分析的问题模型分为三个部分：全局信息，局部信息，以及解的精度。问题模型是对问题集合 $\mathcal{F}$ 更精确的描述，也是对所要求解的优化任务更具体的要求。

当我们考察数值解算法 $\mathcal{S}$ 在优化问题集合 $\mathcal{F}$ 上的表现时，假设算法 $\mathcal{S}$ 无法获得所要求解的具体问题 $\mathcal{P}$（$\mathcal{P}\in\mathcal{F}$）的全部信息，只能获取 $\mathcal{P}$ 在 $\mathcal{F}$ 中的共有特征信息，比如问题集合 $\mathcal{F}$ 中的目标函数是否光滑，可微，其约束集合的类型，是否有界等。我们将 $\mathcal{F}$ 中所具有的共有特征信息记为 Σ，表示问题集合 $\mathcal{F}$ 的**全局信息**。

为了认识和求解 $\mathcal{F}$ 中的某一具体问题 $\mathcal{P}$，$\mathcal{S}$ 需要逐步去收集有关 $\mathcal{P}$ 的局部信息，获得 $\mathcal{P}$ 的局部几何结构，然后根据收集的历史局部信息给出寻找最优解的策略。我们将解算法 $\mathcal{S}$ 收集**局部信息**数据的过程称记做**子程序 $\mathcal{O}$ (Oracle)**。子程序 $\mathcal{O}$ 是一个单元，算法 $\mathcal{S}$ 可以连续调用它来获得一系列有关 $\mathcal{P}$ 的局部信息。一个具体的子程序例子是梯度法中求解函数导数的过程，导数反应了目标函数的局部几何信息，梯度法的执行过程需要连续的调用有关求解目标函数导数的子程序。

为了衡量算法的效率，我们需要事先给出数值算法**解的精度**的信息，记做 $\mathcal{T}_\epsilon$。不同的问题集合 $\mathcal{F}$ 接受不同类型解的精度的度量方式。我们也将解的精度信息固定到问题集合 $\mathcal{F}$ 里面，作为算法求解 $\mathcal{F}$ 时可调用的信息。

因此，我们扩充问题集合 $\mathcal{F}$ 得到复杂度分析理论中的问题模型 $\mathcal{F}$：

$$\mathcal{F} \equiv (\Sigma, \mathcal{O}, \mathcal{T}_\epsilon). \tag{1.2}$$

在解算法 $\mathcal{S}$ 求解问题集合 $\mathcal{F}$ 中的优化问题时，它只能连续调用有关 $\mathcal{F}$ 的三部分信息：全局信息 Σ，局部信息 $\mathcal{O}$，解的精度 $\mathcal{T}_\epsilon$，并利用这些信息来获得最优解的近似值。下面我们对问题模型的三部分信息做更细致的描述。

1.2.1 全局信息 Σ

全局信息由 $\mathcal{F}$ 中问题的目标函数信息和约束集合信息组成。对于目标函数信息，我们主要研究求解下列函数空间中的问题时候的算法复杂度，

- $f(x)$ 是凸函数、强凸函数、非凸函数的情况；$f(x)$ 是光滑函数、非光滑函数的情况。
- $f(x)$ 是合成函数，
$$f(x) = g(x) + h(x). \tag{1.3}$$
在机器学习问题中 $g(x)$ 光滑，正则项 $h(x)$ 可能光滑也可能不光滑。
- $f(x)$ 是随机优化问题，
$$\min_{x\in X}\{f(x) := \mathbb{E}[F(x,\xi)]\}, \tag{1.4}$$
其中 ξ 是随机变量。

更具体的，我们可以将凸函数空间做划分。

定义 1.1 (凸函数子集) 设 $X \subset \mathbb{R}^n$ 是闭凸集合，$f : X \to \mathbb{R}$ 是凸函数，X^* 是(1.1)最优解的集合，x_* 是 x 在最优解集合 X^* 上的投影。定义如下 f 的凸函数子集：

1. $C(X) = C^0(X)$ 是所有连续函数的集合。
2. $C_L(X) = C_L^{0,0}(X)$：若 $f(x) \in C_L(X)$，则 $f(x)$ 具有 *Lipschitz* 连续性
$$|f(x) - f(y)| \le L\|x - y\|, \quad \forall x, y \in X. \tag{1.5}$$
3. $C_L^{1,1}(X)$：若 $f(x) \in C_L^{1,1}(X)$，则 $f(x)$ 一阶可导且导数具有 *Lipschitz* 连续性
$$|\nabla f(x) - \nabla f(y)| \le L\|x - y\|, \quad \forall x, y \in X. \tag{1.6}$$
4. $C_L^{1,\alpha}(X)$：若 $f(x) \in C_L^{1,\alpha}(X)$，则 $f(x)$ 一阶可导且导数具有 *Hölder* 连续性，其中 $\alpha \in (0, 1]$
$$|\nabla f(x) - \nabla f(y)| \le L\|x - y\|^\alpha, \quad \forall x, y \in X. \tag{1.7}$$
5. $\mathcal{F}_{L,\mu}^{0,1}(X)$ 是 $C_L(X)$ 中所有强凸函数组成的集合，$f(x) \in \mathcal{F}_{L,\mu}(X)$，则
$$f(y) \ge f(x) + \langle \partial f(x), y - x\rangle + \frac{\mu}{2}\|y - x\|^2, \quad \forall x, y \in X. \tag{1.8}$$
6. $\mathcal{F}_{L,\mu}^{1,1}(X)$ 是 $C_L^{1,1}(X)$ 中所有强凸函数组成的集合，$f(x) \in \mathcal{F}_{L,\mu}^{1,1}(X)$，则
$$f(y) \ge f(x) + \langle \partial f(x), y - x\rangle + \frac{\mu}{2}\|y - x\|^2, \quad \forall x, y \in X. \tag{1.9}$$

7. $\mathcal{W}_{L,\mu}^{1,1}(X)$ 是 $C_L^{1,1}(X)$ 中所有弱强凸函数组成的集合，$f(x)\in\mathcal{W}_{L,\mu}^{1,1}(X)$，则

$$f^*\geq f(x)+\langle\nabla f(x),x_*-x\rangle+\frac{\mu}{2}\|x-x_*\|^2. \tag{1.10}$$

8. $\mathcal{S}_{L,\mu}^{1,1}(X)$：若 $f(x)\in\mathcal{S}_{L,\mu}^{1,1}(X)$，则 $f(x)\in C_L^{1,1}(X)$ 且具有二阶增长性

$$f(x)-f^*\geq\frac{\mu}{2}\|x-x_*\|^2. \tag{1.11}$$

对于弱强凸函数集合 $\mathcal{W}_{L,\mu}^{1,1}(X)$ 和具有二阶增长性的函数集合 $\mathcal{S}_{L,\mu}^{1,1}(X)$，我们会在后面的章节介绍其具体定义。若记 $\mathcal{F}_L^{1,1}(X)$ 表示 $C_L^{1,1}(X)$ 和凸函数集合的交集，文献 [32] 中给出了凸函数集合之间的包含关系。

定理 1.1 $\mathcal{F}_{L,\mu}^{1,1}(X)\subset\mathcal{W}_{L,\mu}^{1,1}(X)\subset\mathcal{S}_{L,\mu}^{1,1}(X)\subset\mathcal{F}_L^{1,1}(X)$.

对于约束集合信息，我们可以假定 X 是凸集且是闭的，可以在 X 上做投影，用投影梯度法（Projected gradient method）求解 (1.1)。另一种情况下，当 $X=\{x\in\mathbb{R}^n:\ \sum_{i=1}^n x_n=1,\ x_i\geq 0,\ i=1,\cdots,n\}$ 是单纯形状或凸多面体时，我们可以使用条件梯度法（Conditional gradient method）去求解 (1.1)。

1.2.2 局部信息 $\mathcal{O}$

在复杂度分析中，算法通过子程序来获得所求问题的局部信息。不同的算法需要获得不同的局部信息。比如，对于梯度法来说，对任意给定输入 $x_0\in X$，子程序返回函数值信息 $f(x_0)$ 和梯度信息 $\nabla f(x_0)$；对于次梯度法，则需要返回次梯度信息 $\partial f(x_0)$；对于牛顿法，需要返回二阶导数信息 $\nabla^2 f(x_0)$。

这些子程序（oracle）都是事先给定的，复杂度分析理论并不衡量这些子程序的求解复杂度。我们假设这些子程序 $\mathcal{O}$ 具有局部性和黑盒性：

1. 子程序 $\mathcal{O}$ 是唯一局部信息来源：数值算法唯一可获得的有关目标优化问题的局部信息来源于子程序；
2. 子程序 $\mathcal{O}$ 具有局部稳定性：调用子程序时，对测试点 x 做微小扰动，返回的局部信息 $\mathcal{O}(x)$ 变化不大。

其中第二条是对算法进行收敛性分析的关键假设。例如，在分析梯度法的收敛性时，由于 $\mathcal{O}(x)=f(x)$ 或 $\nabla f(x)$，我们需要假设存在某一常数 L 使得

$$\|\mathcal{O}(x)-\mathcal{O}(y)\|\leq L\|x-y\|, \tag{1.12}$$

也就是假设目标函数具有 Lipschitz 连续性，或者目标函数的导数具有 Lipschitz 连续性。对于没有这样性质的优化问题的目标函数，数值算法有可能很难收敛，或者收敛性质很差。

对于不同的算法，我们总能抽象出该算法所需要的子程序，大致分为如下类别：

- $\mathcal{ZO}$ (zero order oracle)：适用于无导数优化算法，对于任意给定的 x_0 返回函数值 $f(x_0)$.
- $\mathcal{FO}$ (first order oracle): 适用于梯度法，返回函数在 x_0 处的函数值和一阶导数信息，$f(x_0)$，$\nabla f(x_0)$ 或 $\partial f(x_0)$.
- 2nd$\mathcal{O}$ (second order oracle): 适用于牛顿法，返回函数在 x_0 处的函数值和一，二阶导数信息，$f(x_0)$，$\nabla f(x_0)$ 以及 $\nabla^2 f(x_0)$.

- $\mathcal{SFO}$ (stochastic first order oracle): 适用于随机梯度法，返回 x_0 在 $F(x,\xi)$ 的函数值和一阶随机梯度信息 $F(x_0,\xi_0)$, $G(x_i,\xi_0)$.
- $\mathcal{PO}$ (projection/prox-operator oracle): 适用于投影梯度法，返回 x_0 在 X 上的投影

$$y \in \operatorname{argmin}_{x\in X} \|x_0 - x\|^2. \tag{1.13}$$

  对于求解合成函数 $f(x) = g(x) + h(x)$ 的邻近点梯度法，返回 x_0 的邻近点投影

$$y \in \operatorname{argmin}_x \{h(x) + g(x_0) + \langle \nabla g(x_0), x - x_0\rangle + \frac{L}{2}\|x_0 - x\|^2\}. \tag{1.14}$$

- $\mathcal{LO}$ (linear optimization oracle): 适用于条件梯度法，当 X 是多面体的时，对给定 x_0 返回线性规划的解 $y \in \operatorname{argmin}_{x\in X}\langle x_0, x\rangle$.
- $\mathcal{SO}$ (separation oracle): 适用于椭球法，对于有界闭约束集合 $X \subset \mathbb{R}^n$，给定点 $c_0 \in \mathbb{R}^n$，如果 $c_0 \in X$ 则返回真，否则返回一个向量 w 在 c_0 处形成一个分割超平面：

$$w^T(x - c_0) \le 0, \quad \forall x \in X. \tag{1.15}$$

上述子程序（Oracle）的任意组合可以形成不同算法所需要的子程序。比如：$\mathcal{FO} + \mathcal{SO}$ 组成椭球法和重心法每一步所需要调用的子程序；$\mathcal{FO} + \mathcal{PO}$ 组成投影梯度法所需的子程序；$\mathcal{FO} + \mathcal{LO}$、$\mathcal{SFO} + \mathcal{PO}$、$\mathcal{SFO} + \mathcal{LO}$ 分别组成条件梯度法、随机梯度法、随机条件梯度法所需要调用的子程序。

1.2.3 解的精度 $\mathcal{T}_\epsilon$

对于每个求解 (1.1) 的算法，我们都会要求其求解精度。对于不同的问题，我们会使用不同解的精度来衡量算法的复杂度。大致分为以下两类：

- **确定性的优化问题**，当目标函数 $f(x)$ 是凸函数的时候，所求局部最优解也是全局最优解，所以算法第 k 步输出 x_k 满足如下条件之一时停止算法：

$$f(x_k) - f^* \le \epsilon, \quad \frac{f(x_k) - f^*}{f(x_k)} \le \delta, \quad \|\nabla f(x_k)\| \le \epsilon, \quad \|x_k - x^*\| \le \epsilon. \tag{1.16}$$

  当目标函数 $f(x)$ 是非凸函数的时候，局部最优解不一定是全局最优解，因此对于某一局部极小值 f^*，会出现 $f(x_k) - f^* \le 0$ 的情形，因此不采用目标函数的误差来衡量解的精度，而是采用 (1.16) 后两个条件，解梯度的范数或解与最优解误差来作为停止条件。

- **随机优化问题**，由于随机优化算法的解是一个随机变量，我们不光需要衡量解的精度，还需要衡量解的置信度。采取两种衡量解的条件，其中一个是 ϵ 解：

$$\mathbb{E}[f(x_k) - f^*] \le \epsilon \quad 或 \quad \mathbb{E}[\|\nabla f(x_R)\|^2] \le \epsilon. \tag{1.17}$$

  ϵ 解衡量的是算法多次运行求得解的期望收敛效率。另一个 (ϵ,δ) 衡量的是算法一次运行求得的解精度达到 ϵ 的置信率：

$$\mathbf{Prob}\{f(x_k) - f^* \ge \epsilon\} \le \delta \quad 或 \quad \mathbf{Prob}\{\|\nabla f(x_R)\|^2 \ge \epsilon\} \le \delta. \tag{1.18}$$

1.3 复杂度分析的算法模型

为了求解问题 $\mathcal{P} \in \mathcal{F}$，我们可以采用迭代过程，即，迭代调用子程序的策略。我们对梯度法，随机梯度法等迭代的优化算法的步骤进行抽象，得到迭代算法框架，以此来衡量优化算法在执行过程中的复杂度。首先，我们给出确定优化问题的抽象迭代算法的框架：

算法 1.1 抽象迭代算法 $\mathcal{S}$ 的框架（确定优化问题）

输入：给定精度 $\epsilon > 0$ 和初始点 $x_0 \in X$，以及初始信息集合 $I_{-1} = \emptyset$，对于 $k = 0, 1, \cdots$，执行迭代

1. 在 x_k 处调用子程序 $\mathcal{O}$，获得目标函数 $f(x)$ 和约束集合 X 在 x_k 处的局部信息 $\mathcal{O}(x_k)$，
2. 更新收集的信息集合 $I_k = I_{k-1} \cup (x_k, \mathcal{O}(x_k))$，
3. 应用算法 $\mathcal{S}$ 规则处理信息集合 I_k 得到新的迭代点 x_{k+1}，
4. 验证是否满足停止条件 $\mathcal{T}_\epsilon$。如果满足，输出 x_k，否则 $k := k + 1$ 转步 1。

输出：$\bar{x} = \mathcal{S}(x_0)$

对于随机优化问题，需要对上述算法框架 1.1 做一些更改。在执行上述算法框架的第 1 步时，需首先根据 ξ 的分布随机抽样出 ξ_k（如果分布未知，则采用蒙特卡罗方法抽样），然后调用子程序 $\mathcal{SFO}$ 得到局部信息 $\mathcal{SFO}(x_k, \xi_k)$。然后执行第 2 步，更新信息集合 $I_k = I_{k-1} \cup (x_k, \xi_k, \mathcal{SFO}(x_k, \xi_k))$。后面的算法步和确定优化问题相同。

有了抽象迭代算法的框架，我们可以定义解算法 $\mathcal{S}$ 和解算法集合 $\mathcal{M}$。对于无约束的梯度法来说，迭代过程

$$x_{k+1} = x_k - \gamma_k \nabla f(x_k). \tag{1.19}$$

可以看到 x_{k+1} 是 x_k 和 $\nabla f(x_k)$ 的函数，即

$$x_{k+1} = F_k(x_k, \nabla f(x_k)). \tag{1.20}$$

我们将 F_k 称为迭代规则函数。实际上在抽象迭代算法框架 1.1 中，

$$x_{k+1} = F_k(x_0, ..., x_k, \nabla f(x_0), ..., \nabla f(x_k), f(x_0), ..., f(x_k)). \tag{1.21}$$

每一个具体的解算法 $\mathcal{S}$ 都对应着一组迭代规则函数 $F = (F_1, F_2, ...)$。我们将不同 F 的集合对应的解算法 $\mathcal{S}$ 的集合记做解算法集合 $\mathcal{M}$。例如

$$x_{k+1} = x_0 + \text{span}\{\nabla f(x_0), \nabla f(x_1), ..., \nabla f(x_k)\}. \tag{1.22}$$

就对应一个解算法集合 $\mathcal{M}$。

1.3.1 复杂度分析的度量

有了问题模型和算法模型，我们就可以定义算法 $\mathcal{S}$ 在问题 $\mathcal{P}$ 上的效率。对于算法 1.1，我们注意到其计算主要集中在两个迭代步骤。一个是步 1，调用子程序，另一个是步 3，更新新的迭代点。因此，我们可以定义两种测度来衡量算法 $\mathcal{S}$ 在问题 $\mathcal{P}$ 上的计算复杂度：

- **分析复杂度 (Analytical complexity)**：将问题 $\mathcal{P}$ 求解到精度 ϵ 总共所需要调用子程序 $\mathcal{O}$ 的次数。

- **算术复杂度 (Arithmetical complexity)**：将问题 $\mathcal{P}$ 求解到精度 ϵ 总共所需要执行的算术操作（包括子程序 $\mathcal{O}$ 内部的操作和算法本身的操作）。

比较上面两种衡量算法效率的复杂度概念，我们发现算术复杂度更能实际体现算法的效率。但当我们用特定算法 $\mathcal{S}$ 求解某一具体问题 $\mathcal{P} \in \mathcal{F}$ 时候，如果可以知道子程序 $\mathcal{O}$ 的复杂度，我们可以很容易的从分析复杂度推出算术复杂度。因此在本文中，我们主要研究算法 $\mathcal{S}$ 在问题集合 $\mathcal{F}$ 上的分析复杂度。

至此，我们给了问题集合 $\mathcal{F}$，具体问题 $\mathcal{P}$ ($\mathcal{P} \in \mathcal{F}$)，解算法集 $\mathcal{M} := \mathcal{M}(\mathcal{P}, \epsilon)$，解算法 $\mathcal{S} := \mathcal{S}(\mathcal{P}, \epsilon)$ ($\mathcal{S} \in \mathcal{M}$) 四个概念的定义。记解算法 $\mathcal{S}$ 求解 $\mathcal{P}$ 的分析复杂度为 $N_{\mathcal{S}}(\mathcal{P}, \epsilon)$。我们定义问题集合 $\mathcal{F}$ 的复杂度上界和下界：

- **$\mathcal{F}$ 的复杂度上界**：对 $\mathcal{F}$ 某一给定的解算法 $\mathcal{S}$，$\mathcal{F}$ 的复杂度上界定义为：

$$\mathrm{Compl}_{\mathcal{S}}(\epsilon) = \sup_{\mathcal{P} \in \mathcal{F}} N_{\mathcal{S}}(\mathcal{P}, \epsilon). \tag{1.23}$$

- **$\mathcal{F}$ 的复杂度下界**：对 $\mathcal{F}$ 某一给定的解算法集 $\mathcal{M}$，$\mathcal{F}$ 的复杂度上界定义为：

$$\mathrm{Compl}(\epsilon) = \inf_{\mathcal{S} \in \mathcal{M}} \mathrm{Compl}_{\mathcal{S}}(\epsilon) := \inf_{\mathcal{S} \in \mathcal{M}} \sup_{\mathcal{P} \in \mathcal{F}} N_{\mathcal{S}}(\mathcal{P}, \epsilon). \tag{1.24}$$

为了求 $\mathcal{F}$ 的复杂度上界，我们需要找到可以求解 $\mathcal{F}$ 中所有问题 $\mathcal{P}$ 的解算法 $\mathcal{S}$。而求解 $\mathcal{F}$ 的复杂度下界，我们需要找到 $\mathcal{F}$ 中的一类病态问题，使得解算法集合 $\mathcal{M}$ 中的算法的效率都很低。

分析复杂度与优化收敛性分析理论中的收敛率有一定的关系。

- 次线性收敛率 (Sublinear rate)：$f(x_k) - f^* \leq \frac{c}{\sqrt{k}}$ 其中 c 为常数，令 $\frac{c}{\sqrt{k}} \leq \epsilon$ 得到 $k \geq \frac{c^2}{\epsilon^2}$，对应的分析复杂度为 $\mathcal{O}(\frac{1}{\epsilon^2})$。
- 线性收敛率 (Linear rate)：$\|x_k - x^*\| \leq c(1-q)^k$ 其中 c 为常数，令 $\leq c(1-q)^k \leq \epsilon$ 得到 $k \geq \frac{1}{q}(\ln c + \ln \frac{1}{\epsilon})$，对应的分析复杂度为 $\mathcal{O}(\ln \frac{1}{\epsilon})$。
- 二阶收敛率 (Quadratic rate)：$\|x_{k+1} - x^*\| \leq c\|x_k - x^*\|^2$，对应的分析复杂度为 $\mathcal{O}(\ln \ln \frac{1}{\epsilon})$。

1.4 复杂度分析的简单例子

复杂度分析的研究可以参考文献 [6]，[40]，[33]，[8]，[36]，[24] 等。这一节我们利用文献 [40] 和 [33] 中的两个例子来应用前面的复杂度分析理论。

例 1.1（盒约束全局优化问题的复杂度上界和下界）

考虑如下全局优化问题：

$$\min_{x \in B_n} \ f(x), \tag{1.25}$$

其中约束集合 B_n 是一个 n 维的盒子 $B_n = \{x \in \mathbb{R}^n \mid 0 \leq x^{(i)} \leq 1,\ i = 1, ..., n\}$，其中 $x^{(i)}$ 表示 x 第 i 个元素的值。在 $\mathbb{R}^n$ 中我们使用 ℓ_∞ 范数来作为距离的度量。假设函数 $f(x)$ 在 B_n 上相对于 ℓ_∞ 范数是 *Lipschitz* 连续的，

$$|f(x) - f(y)| \leq L\|x - y\|_\infty \quad \forall\ x, y \in B_n. \tag{1.26}$$

问题模型 1.1 盒约束全局优化问题模型 $\mathcal{F} = (\Sigma, \mathcal{O}, \mathcal{T}_\epsilon)$ 的三个部分：

- 全局信息 Σ：$f(x)$ 在 B_n 上 ℓ_∞-*Lipschitz* 连续，
- 局部信息 $\mathcal{O}$：$\mathcal{ZO}$ 子程序，对于任意给定的 x_0 返回函数值 $f(x_0)$，

- 解的精度 $\mathcal{T}_\epsilon$：求近似解 $\bar{x} \in B_n$，使得 $f(\bar{x}) - f^* \leq \epsilon$。

对于问题集合 1.1 $\mathcal{F}$ 中的任何一个具体优化问题 $\mathcal{P} \in \mathcal{F}$，我们考虑求解它的一个简单解算法 $\mathcal{S}(p)$:

算法 1.2 无导数全局优化算法 $\mathcal{S}(p)$

输入：盒约束集合每一维划分数 p，问题维数 n.

1. 构造包含 $(p+1)^n$ 个点的**测试点列**：

$$x_{(i_1,\ldots,i_n)} = \left(\frac{i_1}{p}, \frac{i_2}{p}, \cdots, \frac{i_n}{p}\right). \tag{1.27}$$

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

2. 遍历点列 $x_{(i_1,\ldots,i_n)}$，寻找使目标函数值 $f(x_{(i_1,\ldots,i_n)})$ 最小的点，记为 $\bar{x}$.

输出：$(\bar{x}, f(\bar{x}))$

无导数全局优化算法 $\mathcal{S}(p)$ 将 B_n 均匀分成 $(p+1)^n$ 个网格点，然后计算每个网格点上的函数值，最后返回函数值最小的网格点作为(1.25)的近似解。容易看到，$\mathcal{S}(p)$ 也属于我们的抽象迭代算法框架，它需要迭代 $(p+1)^n$ 次，全部遍历测试点列才能找到函数值最小的点。只是它是按顺序更新新的点，并未对收集的历史点列信息做任何实际操作。于是，我们可以衡量算法 $\mathcal{S}(p)$ 的效率。

定理 1.2 *记 f^* 为问题模型 1.25 的全局最优解，利用无导数全局优化算法 1.2 求解 (1.25) 有，*

$$f(\bar{x}) - f^* \leq \frac{L}{2p}. \tag{1.28}$$

对于问题模型 1.25 中的每一个具体问题 $\mathcal{P} \in \mathcal{F}$，算法 $\mathcal{S}(p)$ 的分析复杂度满足：

$$N_{\mathcal{S}(p)}(\mathcal{P}, \epsilon) \leq \left(\left\lfloor \frac{L}{2\epsilon} \right\rfloor + 2\right)^n. \tag{1.29}$$

其中 $\lfloor a \rfloor$ 表示 a 的整数部分。

证明： 由算法 1.2 不难看出问题模型 (1.25) 的全局最优解 x_* 要么落在测试点列上，要么被测试点列组成的网格包含。即存在 $(i_1, \ldots, i_n) \in \{0, \ldots, p\}^n$ 使得，

$$x \equiv x_{(i_1,\ldots,i_n)} \leq x_* \leq x_{(i_1+1,\ldots,i_n+1)} \equiv y. \tag{1.30}$$

这里 $a \equiv (a^{(1)}, \ldots, a^{(n)}) \leq b \equiv (b^{(1)}, \ldots, b^{(n)})$，$a, b \in \mathbb{R}^n$ 当且仅当 $a^{(i)} \leq b^{(i)}$ 对任意 $i = 1, \ldots, n$ 成立。注意到 $y^{(i)} - x^{(i)} = \frac{1}{p}$ 对任意 $i = 1, \ldots, n$ 成立，且

$$x_*^{(i)} \in [x^{(i)}, y^{(i)}], \qquad i = 1, \ldots, n. \tag{1.31}$$

记 $\hat{x} = (x + y)/2$，构造点 $\tilde{x}$ 如下，

$$\tilde{x}^{(i)} = \begin{cases} y^{(i)}, & \text{如果 } x_*^{(i)} \geq \hat{x}^{(i)}, \\ x^{(i)}, & \text{其它.} \end{cases} \tag{1.32}$$

可以看到 $|\tilde{x}^{(i)} - x_*^{(i)}| \leq \frac{1}{2p}$，$i = 1, \ldots, n$。因此，

$$\|\tilde{x} - x_*\|_\infty = \max_{1 \leq i \leq n} |\tilde{x}^{(i)} - x_*^{(i)}| \leq \frac{1}{2p}. \tag{1.33}$$

注意到 $\tilde{x}$ 也属于测试点列，因此

$$f(\bar{x}) - f(x_*) \leq f(\tilde{x}) - f(x_*) \leq L\|\tilde{x} - x_*\|_\infty \leq \frac{L}{2p}. \tag{1.34}$$

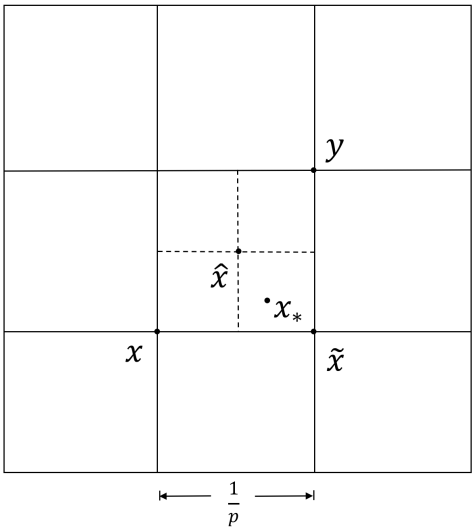

图 1: 当 $n = 2$ 时 $x, y, \tilde{x}, \hat{x}, x_*$ 示意图

当 $n = 2$ 时 $x, y, \tilde{x}, \hat{x}, x_*$ 的关系如图 1。取 $p = \lfloor \frac{L}{2\epsilon} \rfloor + 1$，则 $p \geq \frac{L}{2\epsilon}$。由定理 1.2 我们有，

$$f(\bar{x}) - f^* \leq \frac{L}{2p} \leq \epsilon. \tag{1.35}$$

注意到我们需要调用 $(p+1)^n$ 次的函数值，因此算法 $\mathcal{S}(p)$ 的分析复杂度满足 (1.29)。 □

可以看出算法 $\mathcal{S}(p)$ 的收敛速度与盒约束集合每一维划分点数 p 有关。也就是与网格的划分密度有关。复杂度分析中，我们更关心算法的分析复杂度，也就是 $\mathcal{S}(p)$ 在得到 ϵ 精度的近似解时，需要调用多少次 $\mathcal{ZO}$ 子程序。于是我们有如下关于 $\mathcal{F}$ 复杂度上界的推论，

在 (1.29) 两边关于问题集合 $\mathcal{F}$ 中取极大，我们可以得到问题集合 $\mathcal{F}$ 的复杂度上界的一个估值：

$$\mathrm{Compl}_{\mathcal{S}(p)}(\epsilon) \;=\; \max_{\mathcal{P}\in\mathcal{F}} N_{\mathcal{S}(p)}(\mathcal{P}, \epsilon) \;\leq\; \left(\left\lfloor \frac{L}{2\epsilon} \right\rfloor + 2 \right)^n. \tag{1.36}$$

也就是说，存在一个算法（比如 $\mathcal{S}(p)$）在解决 $\mathcal{F}$ 中每一个具体问题 $\mathcal{P}$ 时（近似解满足 $f(\bar{x}) - f^* \leq \epsilon$, $\bar{x} \in B_n$），所需要调用的 $\mathcal{ZO}$ 子程序次数之多不超过 $\left(\lfloor \frac{L}{2\epsilon} \rfloor + 2 \right)^n$。

关于结论 (1.36)，我们会提出一些问题。第一，我们在估计算法 $\mathcal{S}(p)$ 时太过粗略，是否存在比 (1.36) 更好的界？第二，是否存在其他算法，其效率比 $\mathcal{S}(p)$ 要好？要回答这两个问题，我们就需要推导问题集合 $\mathcal{F}$ 的复杂度下界。

我们首先需要定义问题集合 $\mathcal{F}$ 的解算法集合 $\mathcal{M}$。

定理 1.3 令 $\epsilon < \frac{1}{2}L$，对于 (1.25) 对应的问题集合 $\mathcal{F}$ 来说，对于其中每一个具体问题 $\mathcal{P} \in \mathcal{F}$，对于其解算法集合 $\mathcal{M} := \mathcal{M}(\mathcal{F}, \mathcal{ZO})$ 中的任意一个算法 $\mathcal{S} \in \mathcal{M}$，其分析复杂度满足：

$$\left(\left\lfloor \frac{L}{2\epsilon} \right\rfloor \right)^n \;\leq\; N_{\mathcal{S}}(\mathcal{P}, \epsilon). \tag{1.37}$$

证明: 考虑 $\mathcal{ZO}$ 子程序定义的解算法集合 $\mathcal{M} := \mathcal{M}(\mathcal{F}, \mathcal{ZO})$ 中的任意一个解算法 $\mathcal{S}$。$\mathcal{S}$ 通过对约束集合 B_n 进行采样得到测试点列 $\{x_k\}$，然后在每个测试点列上调用 $\mathcal{ZO}$ 子程序，通过处理（排序）子程序返回的信息（函数值信息）得到最优近似解。$\mathcal{M}$ 中解算法之间的差异仅在于测试点列的选取方式，比如算法 1.2 是在 B_n 上采取均匀采样的方式。我们利用反证法证明定理 1.3。

令采样的测试点列个数 $p = \lfloor \frac{L}{2\epsilon} \rfloor \geq 1$，若存在 $\mathcal{M}$ 中的一个解算法 $\mathcal{S}_0$ 在求解问题模型 1.1 得到 ϵ 精度时调用 $\mathcal{ZO}$ 子程序的次数 $N < p^n$。我们只需要证明存在某一目标函数，使得 $\mathcal{S}_0$ 在求解该目标函数时的精度大于等于 ϵ。

由于 $\mathcal{S}_0$ 在 B_n 上采样的测试点列个数 $N < p^n$，因此存在某一非测试点 $\hat{x}$ 以及集合 $B = \{x | \hat{x} \leq x \leq \hat{x} + \frac{1}{p}e\} \subseteq B_n$ 使得 B 中不包含任何测试点。令 $x_* = \hat{x} + \frac{1}{2p}e$，在集合 B 上构造函数

$$f_p(x) = \min\{0, L\|x - x_*\|_\infty - \epsilon\}, \quad \text{当 } x \in B. \tag{1.38}$$

令 $f_p(x) = 0$ 当 $x \in B_n \setminus B$。注意到 $\|x_* - \hat{x}\|_\infty \geq \frac{\epsilon}{L}$，因此 $f_p(\hat{x}) = 0$ 其图像如图 2。

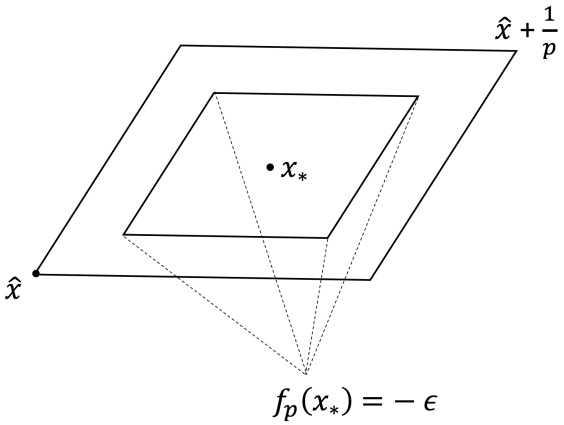

图 2: $f_p(x)$ 在集合 B 上的示意图

注意到 $f_p(x)$ 是 ℓ_∞-Lipschitz 连续（Lipschitz 常数为 L），全局最优解为 $f_p(x_*) = -\epsilon$。利用 $\mathcal{S}_0$ 求解 $f_p(x)$ 时，由于在所要测试点列处调用 $\mathcal{ZO}$ 子程序都返回 $f_p(x_k) = 0$，因此求得近似最优解为 $f_p(\bar{x}) = 0$，于是有

$$f_p(\bar{x}) - f_p(x_*) \geq \epsilon. \tag{1.39}$$

因此我们得到结论：若测试点列的个数 $N < \left(\lfloor \frac{L}{2\epsilon} \rfloor\right)^n$ （子程序 $\mathcal{ZO}$ 的调用次数）则对应的解算法求得的精度不可能比 ϵ 更好。即问题模型 1.1 的复杂度下界为 $\left(\lfloor \frac{L}{2\epsilon} \rfloor\right)^n$。 □

例 1.2（一维凸优化问题的复杂度上界和下界）

考虑一维凸优化问题，

$$\min_{x \in [a,b]} \ f(x). \tag{1.40}$$

问题模型 1.2 一维凸优化问题模型 $\mathcal{F} = (\Sigma, \mathcal{O}, \mathcal{T}_\epsilon)$，

- 全局信息 Σ：求优化问题 $\min\{f(x) \mid x \in [a,b]\}$，$f(x)$ 是凸函数、连续，且满足 $0 \leq f(x) \leq V$，$a < b$ 是任意常数，

- 局部信息 $\mathcal{O}$：$\mathcal{FO}$ 子程序，对于任意给定的 x_0 返回函数值 $f(x_0)$ 和 x_0 处次梯度 $\partial f(x_0)$，
- 解的精度 $\mathcal{T}_\epsilon$：求近似解 $\bar{x} \in [a,b]$，使得 $f(\bar{x}) - f^* \leq \epsilon$。

在文献 [33] 中给出了问题模型 1.2 的复杂度上下界。

定理 1.4 对于问题模型 *1.2* 来说，其复杂度满足：

$$\left\lceil \frac{1}{5}\log_2\left(\frac{V}{\epsilon}\right)\right\rceil \leq N(\epsilon) \leq \left\lfloor \log_2\left(\frac{V}{\epsilon}\right)\right\rfloor. \tag{1.41}$$

证明：a). 复杂度上界：为了求得问题模型 1.2 的复杂度上界，我们只需要提出 1.2 的一个解算法 $\mathcal{S}(\epsilon)$，它求解 1.2 时候的复杂度就是该问题模型的复杂度上界。我们利用最简单的二分法来作为解算法，其求解步骤如下：记目标集合为 $X_k = [l_k, u_k]$，迭代点为 x_k，取 $X_0 = [a,b]$ 即 $l_0 = a$，$u_0 = b$，对 $k = 1, ..., N$ 执行迭代，

1. 更新 $x_k = (l_{k-1} + u_{k-1})/2$，
2. 更新 X_k：在 x_k 处调用子程序得到次梯度 $f'(x_k)$，按图 3 所示三种情况：若 $f'(x_k) < 0$，则 $l_k = l_{k-1}$，$u_k = x_k$；若 $f'(x_k) = 0$，则输出 $\bar{x} = x_k$，停止迭代；若 $f'(x_k) > 0$，则 $l_k = x_k$，$u_k = u_{k-1}$。

最后输出解 $\bar{x} \in \{x_1, ..., x_N\}$ 满足

$$f(\bar{x}) = \min_{1\leq i\leq N} f(x_i). \tag{1.42}$$

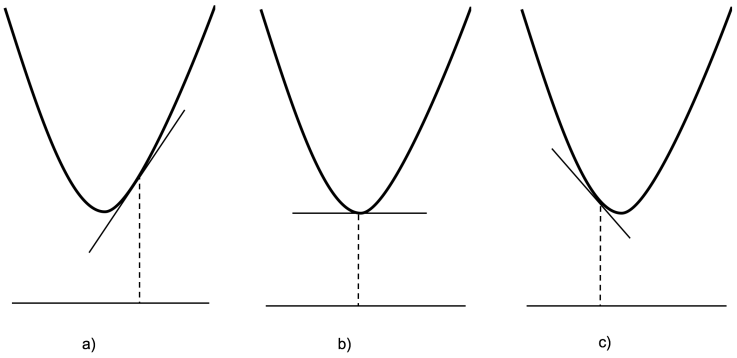

图 3: 二分法示意图

首先我们可以看到迭代集合 X_k 的长度为 $(b-a)/2^k$。同时对任意 $x \in X \setminus X_k$ 我们有 $f(x) \geq f(\bar{x}) = \min_{1\leq i\leq N} f(x_i)$ 成立。取 α 满足 $2^{-N} < \alpha < 1$，构造约束集合 $X = [a,b]$ 的 α 倍收缩集合，

$$X_\alpha = (1-\alpha)x^* + \alpha X \equiv \{(1-\alpha)x^* + \alpha z | z \in X\}. \tag{1.43}$$

不难证明 X_α 是 x^* 处长度为 $\alpha(b-a)$ 的线段。考虑到 α 的取值，可以知道 X_α 无法包含在 X_N 中。即存在 $y \in X_\alpha$ 使得 $y \notin X_N$。设 $y = (1-\alpha)x^* + \alpha z$ 对某一 $z \in X$。由 f 的凸性，

$$f(y) \leq (1-\alpha)f(x^*) + \alpha f(z). \tag{1.44}$$

可以推出

$$f(y) - f(x^*) \leq \alpha(f(z) - f^*) \leq \alpha V. \tag{1.45}$$

因此 y 是原问题的一个 αV 精度的近似解。注意到 $y \notin X_N$，因此有 $f(\bar{x}) \leq f(y)$，即

$$f(\bar{x}) - f^* \leq f(y) - f^* \leq \alpha V. \tag{1.46}$$

由于 α 可以无限接近 2^{-N}，因此取 $N = \left\lfloor \log_2 \left(\frac{V}{\epsilon}\right) \right\rfloor$ 有，

$$f(\bar{x}) - f^* \leq 2^{-N} V \leq 2^{-\left\lfloor \log_2 \left(\frac{V}{\epsilon}\right) \right\rfloor} V \leq \epsilon. \tag{1.47}$$

我们证明了问题模型 1.2 的复杂度上界为 $\left\lfloor \log_2 \left(\frac{V}{\epsilon}\right) \right\rfloor$。

b). 复杂度下界：与证明复杂度上界不同，复杂度下界的证明需要构造 1.2 中一个最坏情况的函数。不失一般性，我们令 $X = [-1, 1]$，$V = 1$，则我们需要证明对任意给定的精度 $\epsilon \in (0, 1)$，问题集合 1.2 的复杂度都不小于

$$\left\lceil \frac{1}{5} \log_2 \left(\frac{1}{\epsilon}\right) \right\rceil. \tag{1.48}$$

也就是说，我们需要证明对于问题模型 1.2 任意的解算法 $\mathcal{S}(\epsilon)$，我们总能构造 1.2 中的一个函数，使得 $\mathcal{S}(\epsilon)$ 在求解该函数时的分析复杂度都不小于 (1.48)。

令 $N = \left\lceil \frac{1}{5} \log_2 \left(\frac{1}{\epsilon}\right) \right\rceil$，对任意的 $0 \leq i \leq N$ 构造函数 $f_i(x)$ 满足如下性质：

1. $f_i(x)$ 凸、连续且存在长度为 $\delta_i = 2^{1-2i}$ 的子集合 $\Delta_i \subseteq X = [-1, 1]$ 使得

$$f_i(x) = a_i + 2^{-3i}|x - c_i|, \quad x \in \Delta_i. \tag{1.49}$$

   其中 c_i 是 Δ_i 的中点。

2. 解算法 $\mathcal{S}(\epsilon)$ 求解 $f_i(x)$ 时产生的前 i 个点列 $x_1, ..., x_i$ 落在 $X \setminus \Delta_i$ 上。

我们用递推法证明存在这样的函数序列 $f_i(x)$。当 $i = 0$ 时，取

$$f_0(x) = |x|, \quad \Delta_0 = X = [-1, 1]. \tag{1.50}$$

满足第一、二条性质。假设存在 f_i 和 Δ_i 满足第一、二条性质，我们利用 f_i 和 Δ_i 来构造 f_{i+1} 和 Δ_{i+1}。令 $x_1, ..., x_{i+1}$ 表示解算法 $\mathcal{S}(\epsilon)$ 求解 $f_i(x)$ 时产生的前 $i+1$ 个点列。由假设我们知道搜索点列 $x_1, ..., x_i$ 落在 $X \setminus \Delta_i$ 上，我们根据 x_{i+1} 的位置构造 f_{i+1}，满足如下性质：

1. 当 $x \in X \setminus \Delta_i$ 时 $f_{i+1}(x) = f_i(x)$。

2. 若 x_{i+1} 落在 c_i 右边，取 c_{i+1} 为 Δ_i 左半边中点，构造如图 4 的函数 f_{i+1} 和 Δ_{i+1}。其中 $\overline{ab}$，$\overline{bc_{i+1}}$，$\overline{c_{i+1}d}$，$\overline{c_ie}$ 的斜率依次为 -2^{-3i}，$-2^{-3(i+1)}$，$2^{-3(i+1)}$，2^{-3i}。

3. 若 x_{i+1} 落在 c_i 左边，构造方式与落在右边对称。

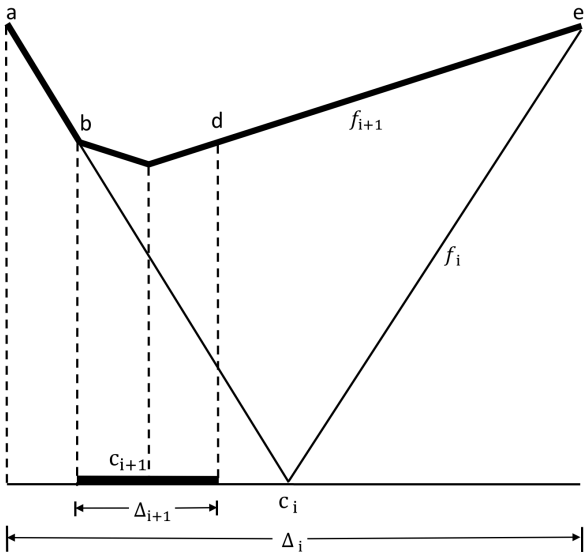

图 4: 由 f_i 构造 f_{i+1} 示意图

我们可以得到满足性质的 f_{i+1} 和 Δ_{i+1}。注意到 f_i 是凸函数，可以知道 c_i 是 f_i 的全局最优解。考虑 f_N 和 Δ_N，由于 $x_1, ..., x_N$ 都落在 $X \setminus \Delta_N$ 上，因此解算法 $\mathcal{S}(\epsilon)$ 求解 $f_N(x)$ 时产生的解 $\bar{x}(\mathcal{S}, f_N)$（由 $x_1, ..., x_N$ 得到）与最优解 c_N 之间的函数值差至少为 $2^{-3N} \times 2^{-2N} = 2^{-5N}$（线段 $\overline{ac_N}$ 的垂直高度）。因此我们有

$$f_N(\bar{x}(\mathcal{S}, f_N)) - f_K^* > 2^{-5N}. \tag{1.51}$$

令 $2^{-5N} < \epsilon$，得到 $N > \frac{1}{5}\log(\frac{1}{\epsilon})$。我们得到问题模型 1.2 的算法复杂度下界至少为 (1.48)。 □

1.5 算法复杂度表

我们将论文中研究的优化算法求解确定性优化问题的收敛性和复杂度列在表 1 中。

其中 gravity 代表重心法，ellipsoid 代表椭球法，PGD （Projected Gradient Method ）代表投影梯度法，AGD (Accelerated Gradient Method) 代表加速梯度法，CndG （Conditional Gradient Method ）代表条件梯度法，CGS （Conditional Gradient Sliding Method ）代表加速条件梯度法。表中求解的函数都是凸函数，且 $X \subseteq \mathbb{R}^n$ 是闭且凸的满足 $\mathcal{B}(r) \subseteq X \subseteq \mathcal{B}(R)$。其中 $Q = L/\mu$，L 表示梯度的 Lipschitz 常数 μ 表示强凸函数对应的常数。

2 光滑优化问题的复杂度分析

2.1 引言

这一章我们给出光滑优化问题的复杂度分析。考虑优化问题

$$f^* = \min_{x \in X} f(x). \tag{2.1}$$

本章中我们考虑优化问题 (2.1) 中目标函数 $f(x)$ 是光滑的情况。首先我们定义光滑函数的一些子集合，并列出它们的一系列性质，用于收敛性分析。这些性质的证明可以在文献 [38] 中找到。

问题集合 $\mathcal{F}$	算法 $\mathcal{S}$	子程序 $\mathcal{O}$	收敛速率	分析复杂度
$C^0(X)$	gravity	$\mathcal{FO}+\mathcal{SO}$	$\exp(-\frac{k}{n})$	$n\log(\frac{B}{\epsilon})$
$C^0(X)$	ellipsoid	$\mathcal{FO}+\mathcal{SO}$	$\frac{R}{r}\exp(-\frac{k}{n^2})$	$n^2\log(\frac{BR}{r\epsilon})$
$C_L^{0,1}(X)$	PGD	$\mathcal{FO}+\mathcal{PO}$	$\frac{LR}{\sqrt{k}}$	$\frac{L^2R^2}{\epsilon^2}$
$\mathcal{F}_{L,\mu}^{0,1}(X)$	PGD	$\mathcal{FO}+\mathcal{PO}$	$\frac{L^2}{\mu k}$	$\frac{L^2}{\mu\epsilon}$
$C_L^{1,1}(X)$	PGD	$\mathcal{FO}+\mathcal{PO}$	$\frac{LR^2}{k}$	$\frac{LR^2}{\epsilon}$
$C_L^{1,1}(X)$	AGD	$\mathcal{FO}+\mathcal{PO}$	$\frac{LR^2}{k^2}$	$\frac{\sqrt{L}R}{\sqrt{\epsilon}}$
$C_L^{1,1}(X)$	CndG	$\mathcal{FO}+\mathcal{LO}$	$\frac{LR^2}{k}$	$\frac{LR^2}{\epsilon}$
$C_L^{1,1}(X)$	CGS	$\mathcal{FO}+\mathcal{LO}$	$\frac{LR^2}{k^2}$	$\mathcal{FO}:\sqrt{LR^2/\epsilon},\ \mathcal{LO}:\frac{LR^2}{\epsilon}$
$\mathcal{S}_{L,\mu}^{1,1}(X)$	PGD	$\mathcal{FO}+\mathcal{PO}$	$LR^2(\frac{Q}{Q+1})^k$	$\log(\frac{LR^2}{\epsilon})/\log(\frac{Q+1}{Q})$
$\mathcal{W}_{L,\mu}^{1,1}(X)$	PGD	$\mathcal{FO}+\mathcal{PO}$	$LR^2(\frac{Q-1}{Q+1})^k$	$\log(\frac{LR^2}{\epsilon})/\log(\frac{Q+1}{Q-1})$
$\mathcal{F}_{L,\mu}^{1,1}(X)$	PGD	$\mathcal{FO}+\mathcal{PO}$	$LR^2(\frac{Q-1}{Q+1})^{2k}$	$\log(\frac{LR^2}{\epsilon})/\log(\frac{Q+1}{Q-1})^2$
$\mathcal{F}_{L,\mu}^{1,1}(X)$	AGD	$\mathcal{FO}+\mathcal{PO}$	$LR^2(\frac{\sqrt{Q}-1}{\sqrt{Q}})^k$	$\log(\frac{LR^2}{\epsilon})/\log(\frac{\sqrt{Q}}{\sqrt{Q}-1})$
$\mathcal{F}_{L,\mu}^{1,1}(X)$	CndG	$\mathcal{FO}+\mathcal{LO}$	$\mu R/2^t$	$Q\log\left(\frac{\mu R}{\epsilon}\right)$
$\mathcal{F}_{L,\mu}^{1,1}(X)$	CGS	$\mathcal{FO}+\mathcal{LO}$	$\delta_0/2^t$	$\mathcal{FO}:\sqrt{Q}\log\frac{\delta_0}{\epsilon},\ \mathcal{LO}:\frac{LR^2}{\epsilon}$
$C_L^{1,\alpha}(\mathbb{R}^n)$	AGD	$\mathcal{FO}$	$\frac{2LR^{1+\alpha}}{(1+3\alpha)\log k}$	$\left(\frac{LR^{1+\alpha}}{\epsilon}\right)^{(2/1+3\alpha)}$

表 1: 确定性优化问题的复杂度分析表

定义 2.1（光滑函数子集合）设 X 为 $\mathbb{R}^n$ 的一个子集，记 $C_L^{k,p}(X)$ 为具有如下性质的函数集合：

1. 任何函数 $f \in C_L^{k,p}(X)$ 在 X 上 k 次连续可微，
2. 任何函数 $f \in C_L^{k,p}(X)$ 的 p 阶导数在 X 上 *Lipschitz* 连续（对于某常数 L），

$$\|f^{(p)}(x) - f^{(p)}(y)\| \leq L\|x-y\|, \quad \text{对任意 } x, y \in X. \tag{2.2}$$

记 $\mathcal{F}_L^{k,p}(X)$ 表示函数集合 $C_L^{k,p}(X)$ 与凸函数集合的交集。

性质 2.1 若 $f_1 \in C_{L_1}^{k,p}(X)$, $f_2 \in C_{L_2}^{k,p}$, 且 $\alpha, \beta \in \mathbb{R}^1$, 则对 $L_3 = |\alpha|L_1 + |\beta|L_2$ 我们有 $\alpha f_1 + \beta f_2 \in C_{L_3}^{k,p}(X)$。

性质 2.2 若 $f \in C_L^{2,1}(\mathbb{R}^n) \subset C_L^{1,1}(\mathbb{R}^n)$ 当且仅当对任意 $x \in \mathbb{R}^n$, $\|\nabla^2 f(x)\| \leq L$.

性质 2.3 令 $f \in C_L^{1,1}(\mathbb{R}^n)$ 则对任意 $x, y \in \mathbb{R}^n$，我们有

$$|f(y) - f(x) - \langle \nabla f(x), y - x\rangle| \leq \frac{L}{2}\|y-x\|^2. \tag{2.3}$$

性质 2.4 令 $f \in C_L^{2,2}(\mathbb{R}^n)$ 则对任意 $x, y \in \mathbb{R}^n$，我们有

$$\|\nabla f(y) - \nabla f(x) - \langle \nabla^2 f(x), y - x\rangle\| \leq \frac{M}{2}\|y-x\|^2, \tag{2.4}$$

$$|f(y) - f(x) - \langle \nabla f(x), y - x\rangle - \frac{1}{2}\langle \nabla^2 f(x)(y-x), y-x\rangle| \leq \frac{M}{6}\|y-x\|^3. \tag{2.5}$$

性质 2.5 令 $f \in C_L^{2,2}(\mathbb{R}^n)$ 且 $\|y - x\| = r$ 则有，

$$\nabla^2 f(x) - MrI_n \preceq \nabla^2 f(y) \preceq \nabla^2 f(x) + MrI_n. \tag{2.6}$$

性质 2.6 令 $f \in \mathcal{F}_L^{1,1}(\mathbb{R}^n)$，则对任意 $x, y \in \mathbb{R}^n$、$\alpha \in [0,1]$ 下列不等式成立，

$$0 \leq f(y) - f(x) - \langle \nabla f(x), y - x\rangle \leq \frac{L}{2}\|y-x\|^2, \tag{2.7}$$

$$f(x) + \langle \nabla f(x), y - x\rangle + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|^2 \leq f(y), \tag{2.8}$$

$$\frac{1}{L}\|\nabla f(y) - \nabla f(x)\|^2 \leq \langle \nabla f(x) - \nabla f(y), x - y\rangle, \tag{2.9}$$

$$\langle \nabla f(x) - \nabla f(y), x - y\rangle \leq L\|x-y\|^2, \tag{2.10}$$

$$\alpha f(x) + (1-\alpha) f(y) \leq f(\alpha x + (1-\alpha)y) + \frac{\alpha(1-\alpha)}{2L}\|\nabla f(x) - \nabla f(y)\|^2, \tag{2.11}$$

$$\alpha f(x) + (1-\alpha) f(y) \leq f(\alpha x + (1-\alpha)y) + \alpha(1-\alpha)\frac{L}{2}\|x-y\|^2. \tag{2.12}$$

2.2 复杂度上界

这一节，我们分析不同算法在不同光滑函数子集合中的收敛性和复杂度，并给出不同光滑函数子集合的复杂度上界。考虑 $X = \mathbb{R}^n$ 情形的优化问题

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

我们分析梯度算法在求解上述问题时的复杂度。

算法 2.1 梯度法

输入：$x_0 \in \mathbb{R}^n$，步长序列 $\{h_k\}$，对 $k = 0, 1, \cdots$ 执行迭代，

$$x_{k+1} = x_k - h_k \nabla f(x_k). \tag{2.13}$$

梯度法的步长 h_k 有如下三种选择方式：

1. 固定步长和变步长策略：取固定步长 $h_k = h > 0$，取变步长 $h_k = \frac{h}{\sqrt{k+1}}$；
2. 精确先搜索步长法：$h_k = \operatorname{argmin}_{h \geq 0} \ f(x_k - h\nabla f(x_k))$；
3. Goldenstein-Armijo 准则：令 $x_{k+1} = x_k - h_k \nabla f(x_k)$，寻找 h_k 满足不等式，

$$\alpha \langle \nabla f(x_k), x_k - x_{k+1} \rangle \ \leq \ f(x_k) - f(x_{k+1}), \tag{2.14}$$

$$\beta \langle \nabla f(x_k), x_k - x_{k+1} \rangle \ \geq \ f(x_k) - f(x_{k+1}). \tag{2.15}$$

其中 $0 < \alpha < \beta < 1$ 为固定常数。

下面，我们列出算法 2.1 在求解无约束，约束优化问题时候的复杂度结论。

问题模型 2.1 *考虑无约束非凸优化问题集合 $\mathcal{F} = (\Sigma, \mathcal{O}, \mathcal{T}_\epsilon)$：*

- *全局信息 Σ：目标函数 $f \in C_L^{1,1}(\mathbb{R}^n)$，且 $f(x)$ 不一定是凸函数；$f(x)$ 下有界（存在常数 M 使得 $f(x) \geq M, \ \forall x \in \mathbb{R}^n$），约束集合 $X \equiv \mathbb{R}^n$，*
- *局部信息 $\mathcal{O}$：$\mathcal{FO}$ 子程序，对于任意给定的 x_0 返回函数值 $f(x_0)$ 和一阶导数 $\nabla f(x_0)$，*
- *解的精度 $\mathcal{T}_\epsilon$：求局部极小值的近似解 $\bar{x} \in \mathbb{R}^n$，使得 $\|\nabla f(\bar{x})\| \leq \epsilon$。*

定理 2.1 *算法 2.1 在求解问题模型 2.1 时，取 $\bar{x}$ 满足 $\|\nabla f(\bar{x})\| = \min_{0 \leq k \leq N} \|\nabla f(x_k)\|$，则算法的收敛速度为*

$$\|\nabla f(\bar{x})\| \leq \frac{1}{\sqrt{N+1}} \left[\frac{L}{\omega} (f(x_0) - f^*) \right]^{1/2}. \tag{2.16}$$

问题模型 2.1 的分析复杂度上界为

$$N(\epsilon) \leq \frac{L(f(x_0) - f^*)}{\omega \epsilon^2}, \tag{2.17}$$

其中 ω 为常数。

证明： 首先，我们可以估计梯度法迭代一步的下降量。对于固定步长 $h_k \equiv h$，在性质 (2.3) 中取 $y = x_{k+1} = x_k - h\nabla f(x_k)$ 以及 $x = x_k$，我们有

$$f(x_{k+1}) \ \leq \ f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2} \|x_{k+1} - x_k\|^2, \tag{2.18}$$

$$= \ f(x_k) - h\|\nabla f(x_k)\|^2 + \frac{h^2}{2} L \|\nabla f(x_k)\|^2, \tag{2.19}$$

$$= \ f(x_k) - h(1 - \frac{h}{2} L) \|\nabla f(x_k)\|^2. \tag{2.20}$$

将 (2.20) 右边关于 h 求极小，我们可以得到梯度法迭代一步时，目标函数的最优下降量，

$$\min_{h>0} \left\{ f(x_k) - h(1 - \frac{h}{2} L) \|\nabla f(x_k)\|^2 \right\} \ = \ f(x_k) - \frac{1}{2L} \|\nabla f(x_k)\|^2. \tag{2.21}$$

当 $h^* = \frac{1}{L}$ 时取到最小值。因此，取恒定步长 $h \equiv \frac{1}{L}$ 则梯度法迭代一步的估计为，

$$f(x_k) - f(x_{k+1}) \geq \frac{1}{2L}\|\nabla f(x_k)\|^2. \tag{2.22}$$

同样的，我们可以取固定步长 $h_k \equiv \frac{2\alpha}{L}$，$\alpha \in (0,1)$，则

$$f(x_k) - f(x_{k+1}) \geq \frac{2}{L}\alpha(1-\alpha)\|\nabla f(x_k)\|^2. \tag{2.23}$$

对于精确线搜索步长，下降量不会比固定步长 $h_k \equiv \frac{1}{L}$ 差，因此也有下降量不等式 (2.22)。对于 Goldenstein-Armijo 准则，结合不等式 (2.15)和 (2.20) 有

$$\beta\langle\nabla f(x_k), x_k - x_{k+1}\rangle \geq f(x_k) - f(x_{k+1}) \geq h_k\left(1 - \frac{h_k}{2}L\right)\|\nabla f(x_k)\|^2. \tag{2.24}$$

由于 $\beta\langle\nabla f(x_k), x_k - x_{k+1}\rangle = \beta h_k\|\nabla f(x_k)\|^2 \geq h_k\left(1 - \frac{h_k}{2}L\right)\|\nabla f(x_k)\|^2$ ，我们得到 $h_k \geq \frac{2}{L}(1-\beta)$，结合不等式 (2.14) 得到

$$f(x_k) - f(x_{k+1}) \geq \alpha h_k\|(x_k)\|^2 \geq \frac{2\alpha(1-\beta)}{L}\|\nabla f(x_k)\|^2. \tag{2.25}$$

综上，梯度法的三种步长选择都有如下下降量估计

$$f(x_k) - f(x_{k+1}) \geq \frac{\omega}{L}\|\nabla f(x_k)\|^2. \tag{2.26}$$

现在，我们可以估计梯度法的复杂度了，将上述不等式对 $k = 0, \cdots, N$ 求和，得到

$$\frac{\omega}{L}\sum_{k=0}^{N}\|\nabla f(x_k)\|^2 \leq f(x_0) - f(x_{N+1}) \leq f(x_0) - f^*. \tag{2.27}$$

由于上式右边是常数，因此当 $N \to \infty$，我们有 $\|\nabla f(x_N)\| \to 0$。注意到 $\|\nabla f(\bar{x})\| = \min_{0\leq k\leq N}\|\nabla f(x_k)\|$，利用不等式 (2.27) 可以得到不等式 (2.16)。 □

注意到梯度法求解该问题模型时候，表现为全局次线性收敛，分析复杂度上界为 $\mathcal{O}(\frac{L}{\epsilon^2})$。

问题模型 2.2 *考虑无约束非凸优化问题集合 $\mathcal{F} = (\Sigma, \mathcal{O}, \mathcal{T}_\epsilon)$:*

- *全局信息 Σ: 目标函数 $f \in C_L^{2,2}(\mathbb{R}^n)$，且 $f(x)$ 不一定是凸函数；$f(x)$ 存在局部极小点 x^*，且 $\nabla^2 f(x^*)$ 正定；存在常数 $0 < m \leq M < \infty$ 使得 $mI_n \preceq \nabla^2 f(x^*) \preceq MI_n$；初始点 x_0 距离 x^* 足够近；约束集合 $X \equiv \mathbb{R}^n$，*
- *局部信息 $\mathcal{O}$: $\mathcal{FO}$ 子程序，对于任意给定的 x_0 返回函数值 $f(x_0)$ 和一阶导数 $\nabla f(x_0)$，*
- *解的精度 $\mathcal{T}_\epsilon$: 求局部极小值的近似解 $\bar{x} \in \mathbb{R}^n$，使得 $\|\bar{x} - x^*\| \leq \epsilon$。*

定理 2.2 *假设梯度法的初始点 x_0 距离局部极小点 x^* 足够近，满足 $r_0 = \|x_0 - x^*\| < \bar{r} = \frac{2m}{L}$，且步长取 $h_k \equiv \frac{2}{m+M}$，则梯度法求解问题模型 2.2 的收敛速率为，*

$$\|x_k - x^*\| \leq \frac{\bar{r}r_0}{\bar{r} - r_0}\left(1 - \frac{2m}{M+3m}\right)^k. \tag{2.28}$$

复杂度上界为

$$\frac{M+3m}{2m}\left[\ln\left(\frac{\bar{r}r_0}{\bar{r} - r_0}\right) + \ln\frac{1}{\epsilon}\right]. \tag{2.29}$$

该定理的证明可以在文献 [38] 中找到。注意到梯度法求解该问题模型时候，表现为局部线性收敛，且复杂度上界为 $\mathcal{O}(\ln \frac{1}{\epsilon})$。

定义 2.2 强凸函数集合 $\mathcal{F}_{L,\mu}^{1,1}(\mathbb{R}^n)$*：若* $f \in \mathcal{F}_{L,\mu}^{1,1}(\mathbb{R}^n)$，*则* $f \in C_L^{1,1}(\mathbb{R}^n)$ *且* f *是强凸函数，即满足*

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|x - y\|^2. \tag{2.30}$$

我们可以将强凸函数的条件放松，得到弱强凸函数和二阶增长性函数。记 X^* 表示凸函数 $f(x)$ 在 $\mathbb{R}^n$ 上最小值对应的点组成的集合。文献 [32] 给出了弱强凸函数集合 $\mathcal{W}_{L,\mu}^{1,1}(\mathbb{R}^n)$ 和二阶增长函数 $\mathcal{S}_{L,\mu}^{1,1}(\mathbb{R}^n)$ 的定义，并分析梯度法在该函数集合上的复杂度上界。

定义 2.3 弱强凸函数 $\mathcal{W}_{L,\mu}^{1,1}(\mathbb{R}^n)$*：若* $f \in \mathcal{F}_{L,\mu}^{1,1}(\mathbb{R}^n)$，*则* $f \in C_L^{1,1}(\mathbb{R}^n)$ *且对任意* $x \in \mathbb{R}^n$ *满足*

$$f^* \geq f(x) + \langle \nabla f(x), \bar{x} - x \rangle + \frac{\mu}{2}\|x - \bar{x}\|^2, \tag{2.31}$$

其中 $\bar{x} = \operatorname{argmin}_{y \in X^*} \|y - x\|$，*即* x *在* X^* *的投影。*

可以看到弱强凸函数只是在最优点处满足不等式 (2.30)，因此 $\mathcal{F}_{L,\mu}^{1,1}(\mathbb{R}^n) \subset \mathcal{W}_{L,\mu}^{1,1}(\mathbb{R}^n)$。

定义 2.4 二阶增长函数 $\mathcal{S}_{L,\mu}^{1,1}(\mathbb{R}^n)$*：若* $f \in \mathcal{S}_{L,\mu}^{1,1}(\mathbb{R}^n)$，*则* $f \in C_L^{1,1}(\mathbb{R}^n)$ *且对任意* $x \in \mathbb{R}^n$ *满足*

$$f(x) - f^* \geq \frac{\mu}{2}\|x - \bar{x}\|^2. \tag{2.32}$$

文献 [32] 证明了包含关系，

$$\mathcal{F}_{L,\mu}^{1,1}(\mathbb{R}^n) \subset \mathcal{W}_{L,\mu}^{1,1}(\mathbb{R}^n) \subset \mathcal{S}_{L,\mu}^{1,1}(\mathbb{R}^n) \subset \mathcal{F}_L^{1,1}(\mathbb{R}^n). \tag{2.33}$$

问题模型 2.3 *考虑优化问题集合模型* $\mathcal{F} = (\Sigma, \mathcal{O}, \mathcal{T}_\epsilon)$*：*

- *全局信息* Σ*：* $f \in \mathcal{F}_L^{1,1}(\mathbb{R}^n)$，*（或* $f \in \mathcal{F}_{L,\mu}^{1,1}(\mathbb{R}^n)$，*或* $f \in \mathcal{W}_{L,\mu}^{1,1}(\mathbb{R}^n)$，*或* $f \in \mathcal{S}_{L,\mu}^{1,1}(\mathbb{R}^n)$*），*
- *局部信息* $\mathcal{O}$*：* $\mathcal{FO}$ *子程序（或* $\mathcal{PO}$ *子程序）*
- *解的精度* $\mathcal{T}_\epsilon$*：求全局极小点* $\bar{x} \in \mathbb{R}^n$，*使得* $f(\bar{x}) - f^* \leq \epsilon$ *（或* $\|\bar{x} - x^*\| \leq \epsilon$*）。*

由文献 [38] 和 [32] 我们可以得到强凸函数集合，弱强凸函数集合以及二阶增长类函数集合的复杂度上界。定理 2.3，2.4，2.5 给出了这三类集合的复杂度分析结论，其证明可以在文献 [38] 和 [32] 中找到。

定理 2.3 *若* $f \in \mathcal{F}_L^{1,1}(\mathbb{R}^n)$，*令步长取* $h_k \equiv h = \frac{2}{L}$，*则梯度法求解问题模型 2.3 的收敛速率为*

$$f(x_k) - f^* \leq \frac{2L\|x_0 - x^*\|^2}{k + 4}. \tag{2.34}$$

记 $D_0 = \|x_0 - x^*\|$，*则复杂度上界为*

$$\mathcal{O}\left(\frac{LD_0}{\epsilon}\right). \tag{2.35}$$

注意到梯度法求解该问题模型时候，表现为全局次线性收敛，且复杂度上界为 $\mathcal{O}(\frac{1}{\epsilon})$。令 $Q = L/\mu$。

定理 2.4 *若* $f \in \mathcal{F}_{L,\mu}^{1,1}(\mathbb{R}^n)$，*令步长取* $h_k \equiv h = \frac{1}{L}$，*则梯度法求解问题模型 2.3 的收敛速率为*

$$\|x_k - x^*\|^2 \leq \left(\frac{Q-1}{Q+1}\right)^{2k} \|x_0 - x^*\|^2, \tag{2.36}$$

$$f(x_k) - f^* \leq \frac{L}{2}\left(\frac{Q-1}{Q+1}\right)^{2k} \|x_0 - x^*\|^2. \tag{2.37}$$

目标函数值对应的复杂度上界为

$$\log\left(\frac{LD_0^2}{\epsilon}\right) / \log\left(\frac{Q-1}{Q+1}\right)^2. \tag{2.38}$$

注意到梯度法求解该问题模型时候，表现为全局线性收敛，且复杂度上界为 $\mathcal{O}(\ln\frac{1}{\epsilon})$。

定理 2.5 若 $f\in\mathcal{W}^{1,1}_{L,\mu}(\mathbb{R}^n)$，令步长取 $h_k\equiv h=\frac{1}{L}$，则梯度法求解问题模型 2.3 的收敛速率为，

$$\|x_k-\bar{x}^k\|^2 \leq \left(\frac{Q-1}{Q+1}\right)^k\|x_0-\bar{x}^k\|^2, \tag{2.39}$$

$$f(x_k)-f^* \leq \frac{L}{2}\left(\frac{Q-1}{Q+1}\right)^{k-1}\|x_0-\bar{x}^k\|^2. \tag{2.40}$$

对应的复杂度上界分别为

$$\log\left(\frac{LD_0^2}{\epsilon}\right)/\log\left(\frac{Q-1}{Q+1}\right). \tag{2.41}$$

注意到梯度法求解该问题模型时候，表现为全局线性收敛，且复杂度上界为 $\mathcal{O}(\ln\frac{1}{\epsilon})$。

定理 2.6 若 $f\in\mathcal{S}^{1,1}_{L,\mu}(\mathbb{R}^n)$，令步长取 $h_k\equiv h=\frac{1}{L}$，则梯度法求解问题模型 2.3 的收敛速率为，

$$\|x_k-\bar{x}^k\|^2 \leq \left(\frac{Q}{Q+1}\right)^k\|x_0-\bar{x}^k\|^2, \tag{2.42}$$

$$f(x_k)-f^* \leq \frac{L}{2}\left(\frac{Q}{Q+1}\right)^{k-1}\|x_0-\bar{x}^k\|^2. \tag{2.43}$$

对应的复杂度上界分别为

$$\log\left(\frac{LD_0^2}{\epsilon}\right)/\log\left(\frac{Q}{Q+1}\right). \tag{2.44}$$

注意到梯度法求解该问题模型时候，表现为全局线性收敛，且复杂度上界为 $\mathcal{O}(\ln\frac{1}{\epsilon})$。

比较定理2.4，2.5 以及2.6，注意到

$$\log\left(\frac{Q-1}{Q+1}\right)^2 \leq \log\left(\frac{Q-1}{Q+1}\right) \leq \log\left(\frac{Q}{Q+1}\right). \tag{2.45}$$

因此 $\mathcal{F}^{1,1}_{L,\mu}(\mathbb{R}^n)$，$\mathcal{W}^{1,1}_{L,\mu}(\mathbb{R}^n)$，$\mathcal{S}^{1,1}_{L,\mu}(\mathbb{R}^n)$ 的复杂度上界依次递减。

在文献 [40] 中，给出了牛顿法的复杂度分析。首先我们给出牛顿法的具体形式。

算法 2.2 牛顿法

输入：$x_0\in\mathbb{R}^n$，对 $k=0,1,\cdots$ 执行迭代，

$$x_{k+1}=x_k-[\nabla^2 f(x_k)]^{-1}\nabla f(x_k). \tag{2.46}$$

问题模型 2.4 考虑无约束非凸优化问题集合 $\mathcal{F}=(\Sigma,\mathcal{O},\mathcal{T}_\epsilon)$:

- 全局信息 Σ: 目标函数 $f\in C^{2,2}_L(\mathbb{R}^n)$，且 $f(x)$ 不一定是凸函数；$f(x)$ 存在局部极小点 x^*，且 $\nabla f(x^*)$ 正定；存在常数 $m>0$ 使得 $\nabla^2 f(x^*)\succeq mI_n$；初始点 x_0 距离 x^* 足够近；约束集合 $X\equiv\mathbb{R}^n$，
- 局部信息 $\mathcal{O}$: 2nd$\mathcal{O}$ 子程序，返回 $f(x)$ 在 x_0 处的函数值，一阶梯度信息和二阶梯度信息，$f(x_0)$，$\nabla f(x_0)$，$\nabla^2 f(x_0)$，
- 解的精度 $\mathcal{T}_\epsilon$: 求局部极小值的近似解 $\bar{x}\in\mathbb{R}^n$，使得 $\|\bar{x}-x^*\|\leq\epsilon$。

定理 2.7 假设牛顿法的初始点 x_0 距离局部极小点 x^* 足够近，满足 $\|x_0 - x^*\| < \bar{r} = \frac{3m}{L}$，则对任意 k 满足 $\|x_k - x^*\| \leq \bar{r}$。牛顿法求解问题模型 2.4 的收敛速率为，

$$\|x_{k+1} - x^*\| \leq \frac{L\|x_k - x^*\|^2}{2(m - L\|x_k - x^*\|)}. \tag{2.47}$$

复杂度上界为 $c \ln\ln \frac{\gamma}{\epsilon}$，其中 c, γ 为常数。

定理 2.7 的证明可以在文献 [40] 中找到。注意到 Newton 法求解该问题模型时候，表现为局部二阶收敛，且复杂度上界为 $\mathcal{O}(\ln\ln \frac{1}{\epsilon})$。

2.3 复杂度下界

在文献 [40] 中，Nesterov 给出了光滑凸优化问题和光滑强凸优化问题的复杂度下界。

定理 2.8（光滑凸优化问题的复杂度下界） 对任意 $x_0 \in \mathbb{R}^n$ 和 $1 \leq k \leq \frac{1}{2}(n-1)$，存在 $f \in C_L^{\infty,1}(\mathbb{R}^n)$，使得对任意 $x_k \in x_0 + \text{span}\{\nabla f(x_0), ..., \nabla f(x_{k-1})\}$ 都有

$$f(x_k) - f^* \geq \frac{3L\|x_0 - x^*\|^2}{32(k+1)^2}. \tag{2.48}$$

令 $D_0 = \|x_0 - x^*\|$，将定理 2.8 不等式 (2.48) 右端等于 ϵ，可以得到一阶算法（只利用梯度信息的算法）求解凸优化问题 $C_L^{\infty,1}(\mathbb{R}^n)$ 的复杂度下界为：

$$\mathcal{O}\left(\sqrt{\frac{L}{\epsilon}} D_0\right). \tag{2.49}$$

与梯度法的复杂度上界 (2.35) $\mathcal{O}(LD_0/\epsilon)$ 比较，发现梯度法并不是最优算法。

定理 2.9（光滑强凸优化问题的复杂度下界） 对任意 $x_0 \in \mathbb{R}^\infty$, $\mu, L > 0$ 和 $Q = L/\mu > 1$，存在 $f \in C_{L,\mu}^{\infty,1}(\mathbb{R}^\infty)$，使得对任意 $x_k \in x_0 + \text{span}\{\nabla f(x_0), ..., \nabla f(x_{k-1})\}$ 都有

$$f(x_k) - f^* \geq \frac{\mu}{2}\left(\frac{\sqrt{Q}-1}{\sqrt{Q}+1}\right)^{2k} \|x_0 - x^*\|^2. \tag{2.50}$$

将不等式 (2.50) 右边等于 ϵ，利用不等式 $(1+a)^k \leq e^{ka}$ 我们可以得到一阶算法（只利用梯度信息的算法）求解强凸优化问题 $\mathcal{F}_{L,\mu}^{\infty,1}(\mathbb{R}^n)$ 的复杂度下界为：

$$\mathcal{O}\left[\sqrt{\frac{L}{\mu}} \max\left(\log \frac{\mu D_0}{\epsilon}, 1\right)\right]. \tag{2.51}$$

2.4 Nesterov 加速梯度法框架

由于现实问题中数据量规模的增加，优化算法越来越对迭代速度有了要求。因此加速一阶优化算法在近十年内受到了非常广泛的关注。文献 [48] [38]，[40]，[39]，[4]，[18] 介绍了加速一阶优化算法的发展历史和分析技巧。

2.4.1 凸优化问题

上一节我们分析了光滑凸优化问题的复杂度下界。由定理 2.8 ，我们知道函数集合 $C_L^{\infty,1}(\mathbb{R}^n)$ 相对于解算法集合 $x_k \in x_0 + \text{span}\{\nabla f(x_0), ..., \nabla f(x_{k-1})\}$ 的复杂度下界是 $\mathcal{O}(1/\sqrt{\epsilon})$。又由于 (2.35) 我们知道梯度法相对于 $C_L^{\infty,1}(\mathbb{R}^n)$ 的复杂度上界为 $\mathcal{O}(1/\epsilon)$。因此梯度法相对于该解算法集合不是最优的，我们需要寻找达到函数集合 $C_L^{\infty,1}(\mathbb{R}^n)$ 复杂度下界的最优解算法。Nesterov 在文献 [38] 中提出了可以达到 $\mathcal{O}(1/\sqrt{\epsilon})$ 复杂度的加速梯度法。为了将 Nesterov 加速梯度法移植到其它新的算法中，我们需要总结文献 [38] 中加速梯度法的结构，构造 Nesterov 加速梯度法的一般框架，并给出加速类优化算法收敛性统一的分析工具和技巧。首先，我们给出求解问题(2.1)的 Nesterov 加速梯度法框架。

算法 2.3 Nesterov 加速梯度算法框架

输入：令 $x_0 = y_0$, 选取序列 $\{\gamma_k\}$ 和 $\{\beta_k\}$ 使其满足 $L\gamma_k \le \beta_k$, $\gamma_1 = 1$。

对于$k = 1, ..., N$，执行迭代

1. $z_k = (1-\gamma_k)y_{k-1} + \gamma_k x_{k-1}$,
2. $x_k = \text{argmin}_{x\in X}\left\{\langle \nabla f(z_k), x\rangle + \frac{\beta_k}{2}\|x - x_{k-1}\|_2^2\right\}$,
3. $y_k = (1-\gamma_k)y_{k-1} + \gamma_k x_k$。

输出: y_N

算法 2.3 共有三个迭代序列：$\{x_k\}$，$\{y_k\}$ 和 $\{z_k\}$，以及两个待定参数序列 $\{\gamma_k\}$ 和 $\{\beta_k\}$。其中收敛点列为 $\{y_k\}$。如果约束集合 $X \equiv \mathbb{R}^n$，则迭代的第二步等价于 $x_k = x_{k-1} - \frac{1}{\beta_k}\nabla f(z_k)$，此时我们可以画出迭代序列的关系图如下：

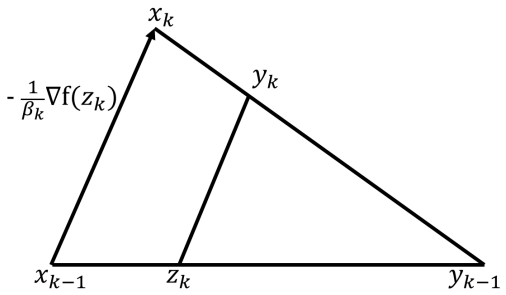

图 5: Nesterov 加速梯度法图

注意到 y_k, z_k, y_{k-1} 组成的三角形与 x_k, x_{k-1}, y_{k-1} 组成的三角形相似，且线段 z_k, y_{k-1} 与线段 x_{k-1}, y_{k-1} 的比为 γ_k。

通过选取不同的参数 γ_k 和 β_k，我们可以得到加速梯度算法框架不同的收敛速度。下面我们给出一般性的分析工具，通过构造序列 $\{\Gamma_k\}$ 来分析收敛性。

引理 2.1 令 $\gamma_t \in (0, 1]$，$t = 1, 2, ...$，构造序列

$$\Gamma_t = \begin{cases} 1 & t = 1 \\ (1-\gamma_t)\Gamma_{t-1} & t \ge 2 \end{cases}. \tag{2.52}$$

如果序列 $\{\Delta_t\}_{t\ge 0}$ 满足

$$\Delta_t \le (1-\gamma_t)\Delta_{t-1} + B_t \quad t = 1, 2, ... \tag{2.53}$$

则对任意的 k 我们对 Δ_k 有估计

$$\Delta_k \leq \Gamma_k(1-\gamma_1)\Delta_0 + \Gamma_k \sum_{t=1}^{k} \frac{B_t}{\Gamma_t}. \tag{2.54}$$

在分析算法的收敛速度时，我们一般根据所要估计的收敛序列，令 $\Delta_k = f(x_k) - f(x^*)$ 或 $\Delta_k = \|x_k - x^*\|_2^2$，然后构造引理 2.1 中的不等式 (2.53)

$$f(x_k) - f(x^*) \leq (1-\gamma_k)(f(x_{k-1}) - f(x^*)) + B_k. \tag{2.55}$$

或者

$$\|x_k - x^*\|_2^2 \leq (1-\gamma_k)\|x_{k-1} - x^*\|_2^2 + B_k. \tag{2.56}$$

通常，我们构造序列 $\{\gamma_t\}_{t\geq 1}$ 使其满足 $\gamma_1 = 1$。根据引理 2.1 不等式 (2.54) 可以得到收敛速度的估计不等式:

$$f(x_k) - f(x^*) \leq \Gamma_k \sum_{i=1}^{k} \frac{B_i}{\Gamma_i}. \tag{2.57}$$

或者

$$\|x_k - x^*\|_2^2 \leq \Gamma_k \sum_{i=1}^{k} \frac{B_i}{\Gamma_i}. \tag{2.58}$$

通过估计上式右端的收敛阶数，我们可以得到算法的收敛速度。可以看到收敛速度与 B_k 和 Γ_k 有关。对于 B_k 我们需要通过放缩估计其上界，而对于 Γ_k 我们可以采取不同的构造方式。由式 (2.52) 可得

$$\Gamma_k = (1-\gamma_k)(1-\gamma_{k-1})...(1-\gamma_2), \quad k = 2, \cdots. \tag{2.59}$$

因此在保证 $\gamma_1 = 1$ 的情况下，我们可以构造序列 $\{\gamma_t\}_{t\geq 1}$ 和 $\{\Gamma_t\}_{t\geq 1}$，例如

$$\gamma_k = \frac{1}{k} \quad \Rightarrow \quad \Gamma_k = \frac{1}{k}, \tag{2.60}$$

$$\gamma_k = \frac{2}{k+1} \quad \Rightarrow \quad \Gamma_k = \frac{2}{k(k+1)}, \tag{2.61}$$

$$\gamma_k = \frac{3}{k+2} \quad \Rightarrow \quad \Gamma_k = \frac{6}{k(k+1)(k+2)}. \tag{2.62}$$

令 $D_X = \sup_{x,y\in X} \|x-y\|$，下面我们证明算法 2.3 在取不同 β_k，γ_k 序列时的收敛速度。

定理 2.10 *若 $f \in C_L^{1,1}(\mathbb{R}^n)$ 且是凸函数，则 Nesterov 加速梯度算法求解问题模型 2.3 的收敛速率为:*

1. *取 $\beta_k = L$，$\gamma_k = \frac{1}{k}$ 则 $\Gamma_k = \frac{1}{k}$，$\frac{\beta_k\gamma_k}{\Gamma_k} = L$，我们有*

$$f(y_k) - f(x^*) \leq \frac{L}{2k} D_X^2, \quad f(y_k) - f(x^*) \leq \frac{L}{2k}\|x_0 - x^*\|^2. \tag{2.63}$$

2. *取 $\beta_k = \frac{2L}{k}$，$\gamma_k = \frac{2}{k+1}$ 则 $\Gamma_k = \frac{2}{k(k+1)}$，$\frac{\beta_k\gamma_k}{\Gamma_k} = 2L$，我们有*

$$f(y_k) - f(x^*) \leq \frac{2L}{k(k+1)} D_X^2, \quad f(y_k) - f(x^*) \leq \frac{4L}{k(k+1)}\|x_0 - x^*\|^2. \tag{2.64}$$

3. *取 $\beta_k = \frac{3L}{k+1}$，$\gamma_k = \frac{3}{k+2}$ 则 $\Gamma_k = \frac{6}{k(k+1)(k+2)}$，$\frac{\beta_k\gamma_k}{\Gamma_k} = \frac{3Lk}{2} \geq \frac{\beta_{k-1}\gamma_{k-1}}{\Gamma_{k-1}}$，我们有*

$$f(y_k) - f(x^*) \leq \frac{9L}{2(k+1)(k+2)} D_X^2. \tag{2.65}$$

证明: 对于算法 2.3，其收敛点列为 $\{y_k\}$。如果我们可以证明不等式

$$f(y_k) - f(x^*) \leq (1-\gamma_k)[f(y_{k-1}) - f(x^*)] + B_k. \tag{2.66}$$

成立。令 $\Delta_k = f(y_k) - f(x^*)$ 利用引理 2.1 并取 γ_k 满足 $\gamma_0 = 1$，由 (2.54) 我们可以得到收敛速度的估计不等式，

$$f(y_k) - f(x^*) \leq \Gamma_k \sum_{i=1}^{k} \frac{B_i}{\Gamma_i}. \tag{2.67}$$

选择合适的 γ_k 使得上式右端收敛，我们就可以得到 Nesterov 加速梯度算法的收敛性。

至此，我们可以用的条件为 $f(x) \in C_L^{1,1}(X)$，$f(x)$ 的凸性，算法第 1，3 步的迭代关系，算法第 2 步的最优性条件。首先，利用 $f(x) \in C_L^{1,1}(X)$ 我们有

$$f(y_k) \leq f(z_k) + \langle \nabla f(z_k), y_k - z_k \rangle + \frac{L}{2}\|y_k - z_k\|^2. \tag{2.68}$$

用迭代步 3 减去步 1 有 $y_k - z_k = \gamma_k(x_k - x_{k-1})$，将此等式与 $y_k = (1-\gamma_k)y_{k-1} + \gamma_k x_k$ 带入上式右端得

$$f(y_k) \leq f(z_k) + \langle \nabla f(z_k), (1-\gamma_k)y_{k-1} + \gamma_k x_k - z_k \rangle + \frac{L\gamma_k^2}{2}\|x_k - x_{k-1}\|^2. \tag{2.69}$$

注意到

$$f(z_k) + \langle \nabla f(z_k), (1-\gamma_k)y_{k-1} + \gamma_k x_k - z_k \rangle \tag{2.70}$$

$$= (1-\gamma_k)[f(z_k) + \langle \nabla f(z_k), y_{k-1} - z_k \rangle] + \gamma_k[f(z_k) + \langle \nabla f(z_k), x_k - z_k \rangle]. \tag{2.71}$$

由 $f(x)$ 的凸性我们有

$$f(y_k) \leq (1-\gamma_k)f(y_{k-1}) + \gamma_k[f(z_k) + \langle \nabla f(z_k), x_k - z_k \rangle] + \frac{L\gamma_k^2}{2}\|x_k - x_{k-1}\|^2. \tag{2.72}$$

由步 2，$x_k = \operatorname{argmin}_{x \in X} \left\{ \langle \nabla f(z_k), x \rangle + \frac{\beta_k}{2}\|x - x_{k-1}\|_2^2 \right\}$，其最优性条件为，

$$\langle \nabla f(z_k) + \beta_k(x_k - x_{k-1}), x_k - x \rangle \ \leq \ 0, \quad \forall\, x \in X. \tag{2.73}$$

即

$$\langle x_{k-1} - x_k, x_k - x \rangle \ \leq \ \frac{1}{\beta_k}\langle \nabla f(x_k), x - x_k \rangle. \tag{2.74}$$

由

$$\frac{1}{2}\|x_k - x_{k-1}\|^2 \ = \ \frac{1}{2}\|x_{k-1} - x\|^2 - \langle x_{k-1} - x_k, x_k - x \rangle - \frac{1}{2}\|x_k - x\|^2. \tag{2.75}$$

$$\leq \ \frac{1}{2}\|x_{k-1} - x\|^2 + \frac{1}{\beta_k}\langle \nabla f(z_k), x - x_k \rangle - \frac{1}{2}\|x_k - x\|^2. \tag{2.76}$$

上式左右两边同乘 $L\gamma_k^2$，考虑到 $L\gamma_k \leq \beta_k$ 我们有

$$\frac{L\gamma_k^2}{2}\|x_k - x_{k-1}\|^2 \ \leq \ \frac{L\gamma_k^2}{2}\|x_{k-1} - x\|^2 + \gamma_k\langle \nabla f(z_k), x - x_k \rangle - \frac{L\gamma_k^2}{2}\|x_k - x\|^2. \tag{2.77}$$

将式 (2.72) 与式 (2.77) 相加，利用 $f(x)$ 的凸性 $f(z_k) + \langle \nabla f(z_k), x - z_k \rangle \leq f(x)$ 得，

$$f(y_k) - f(x) \leq (1-\gamma_k)[f(y_{k-1}) - f(x)] + \frac{L\gamma_k^2}{2}\left(\|x_{k-1} - x\|^2 - \|x_k - x\|^2\right). \tag{2.78}$$

令 $\Delta_k = f(y_k) - f(x)$，$B_k = \frac{L\gamma_k^2}{2}\left(\|x_{k-1} - x\|^2 - \|x_k - x\|^2\right)$ 利用引理 2.1 有，

$$f(y_k) - f(x) \leq \frac{\Gamma_k(1-\gamma_1)}{\Gamma_1}\left(f(y_0) - f(x)\right) + \frac{\Gamma_k}{2}\sum_{i=1}^{k}\frac{\beta_i\gamma_i}{\Gamma_i}\left(\|x_{i-1} - x\|^2 - \|x_i - x\|^2\right). \tag{2.79}$$

分析上面不等式，注意到，

$$\sum_{i=1}^{k}\frac{\beta_i\gamma_i}{\Gamma_i}\left(\|x_{i-1} - x\|^2 - \|x_i - x\|^2\right) \tag{2.80}$$

$$= \frac{\beta_1\gamma_1}{\Gamma_1}\|x_0 - x\|^2 + \sum_{i=2}^{k}\left(\frac{\beta_i\gamma_i}{\Gamma_i} - \frac{\beta_{i-1}\gamma_{i-1}}{\Gamma_{i-1}}\right)\|x_{i-1} - x\|^2 - \beta_k\gamma_k\Gamma_k\|x_k - x\|^2 \tag{2.81}$$

$$\leq \frac{\beta_1\gamma_1}{\Gamma_1}\|x_0 - x\|^2 + \sum_{i=2}^{k}\left(\frac{\beta_i\gamma_i}{\Gamma_i} - \frac{\beta_{i-1}\gamma_{i-1}}{\Gamma_{i-1}}\right)\cdot D_X^2 \tag{2.82}$$

$$\leq \frac{\beta_k\gamma_k}{\Gamma_k}D_X^2. \tag{2.83}$$

因此，对任意序列 $\{\beta_k\}$，$\{\gamma_k\}$，$\{\Gamma_k\}$ 满足 $L\gamma_k \leq \beta_k$ 时有，

$$f(y_k) - f(x) \leq \frac{\beta_k\gamma_k}{2}D_X^2. \tag{2.84}$$

若序列 $\{\beta_k\}$，$\{\gamma_k\}$，$\{\Gamma_k\}$ 满足 $L\gamma_k \leq \beta_k$ 且

$$\frac{\beta_k\gamma_k}{\Gamma_k} \geq \frac{\beta_{k-1}\gamma_{k-1}}{\Gamma_{k-1}}, \qquad \text{对任意} \quad k \geq 2. \tag{2.85}$$

我们有

$$f(y_k) - f(x) \leq \Gamma_k\frac{\beta_1\gamma_1}{2}\|x_0 - x\|^2. \tag{2.86}$$

综上，我们可以通过构造序列 $\{\beta_k\}$，$\{\gamma_k\}$ 来得到算法的收敛速度，将 β_k，γ_k 以及 Γ_k 分别代入不等式 (2.84) 和 (2.86) 即得收敛性。 □

2.4.2 非凸优化问题

上一节我们介绍了 Nesterov 加速梯度算法框架在凸函数上的收敛性质。对于非凸函数，如果直接利用 Nesterov 加速梯度算法框架，不一定有收敛性结果。对于非凸函数我们衡量 $\min_{0\leq k\leq N}\|\nabla f(x_k)\|$ 的收敛速度。由 (2.1) 我们知道梯度算法对于非凸和凸情况下的函数集合 $C_L^{1,1}(\mathbb{R}^n)$，近似解 x_k 满足

$$\min_{0\leq k\leq N}\|\nabla f(x_k)\|^2 \leq \frac{2L(f(x_0) - f^*)}{N+1}. \tag{2.87}$$

文献 [16] 给出了 Nesterov 加速梯度算法框架的变形，从而得到求解无约束非凸优化问题的收敛性。首先，我们给出非凸优化问题模型。

问题模型 2.5 *考虑带无约束的非凸优化问题集合* $\mathcal{F} = (\Sigma, \mathcal{O}, \mathcal{T}_\epsilon)$:

- *全局信息* Σ: *目标函数* $f \in C_L^{1,1}(\mathbb{R}^n)$，*可能非凸函数，约束集合* $X \equiv \mathbb{R}^n$，
- *局部信息* $\mathcal{O}$: $\mathcal{FO}$ *子程序*，
- *解的精度* $\mathcal{T}_\epsilon$: *求局部极小点* $\bar{x} \in \mathbb{R}^n$，*使得* $\|\nabla f(\bar{x})\| \leq \epsilon$。

我们对 Nesterov 加速梯度算法框架做一些变形，加入新的参数序列 λ_k 并修改 y_k 的更新方式，得到非凸函数的加速梯度法。

算法 2.4 非凸函数的加速梯度法

输入： 令 $x_0 = y_0$，$\gamma_k \in (0, 1]$。

对于$k = 1, ..., N$，执行迭代

1. $z_k = (1 - \gamma_k)y_{k-1} + \gamma_k x_{k-1}$，
2. $x_k = x_{k-1} - \frac{1}{\beta_k}\nabla f(z_k)$，
3. $y_k = z_k - \frac{1}{\lambda_k}\nabla f(z_k)$.

输出： z_N

定理 2.11 取 $\gamma_k = \frac{2}{k+1}$ 和 $\lambda_k = 2L$，利用算法 2.4 求解问题模型 2.5 有如下收敛性结果：

1. 若取 $\beta_k \in \left[\frac{4k+4}{2k+3}L, 2L\right]$，则

$$\min_{0\le k\le N} \|\nabla f(z_k)\|^2 \le \frac{6L\left(f(x_0) - f^*\right)}{N}. \tag{2.88}$$

2. 如果问题模型 2.5 中 $f(x)$ 是凸函数，取 $\beta_k = \frac{4L}{k}$，则

$$\min_{0\le k\le N} \|\nabla f(z_k)\|^2 \le \frac{96L^2\|x_0 - x^*\|^2}{N(N+1)(N+2)}. \tag{2.89}$$

该定理的证明可以在文献 [16] 中找到。注意到，在凸函数情况下，算法 2.4 迭代点梯度 $\|\nabla f(z_k)\|$ 的收敛速度为 $\mathcal{O}(1/N^{3/2})$，而梯度法的收敛速度为 $\mathcal{O}(1/N^{1/2})$。而在非凸函数情况下，算法 2.4 和梯度法的收敛速度相同。

3 非光滑优化问题的复杂度分析

这一章，我们研究非光滑优化问题的复杂度分析。我们介绍三部分：次梯度法和镜像次梯度法的收敛性和复杂度，求解非光滑优化问题的重心法和椭球法，复合优化问题的加速算法。次梯度法和镜像梯度法可以参考文献 [40]，[39]，[6]；非光滑优化问题的复杂度下界可以参考文献 [40] 和 [39]；重心法和椭球法可以参考文献 [8]，[29]，[42] 。复合优化问题可参考文献 [40, 48, 4, 5, 37, 39, 16, 17, 15]。

3.1 次梯度法与镜像梯度法

考虑优化问题

$$f^* = \min_{x\in X} f(x). \tag{3.1}$$

当目标函数光滑且无约束的时候，对给定初始点 $x_0 \in \mathbb{R}^n$，梯度法根据如下迭代公式更新迭代点 x_k，

$$x_{k+1} = x_k - \gamma_k \nabla f(x_k). \tag{3.2}$$

其中 $\gamma_k > 0$ 是第 k 步迭代的步长。为了求解非光滑有约束的优化问题 3.1，我们需要对梯度法 (3.2) 做两个改变。第一，由于目标函数 f 不一定可导，我们需要将导数 $\nabla f(x_k)$ 用一个次梯度替换 $g_k \in \partial f(x_k)$。

第二，由于迭代公式 (3.2) 只适用于无约束情形，对于有约束问题 $X \neq \mathbb{R}^n$，(3.2) 确定的新迭代点 x_{k+1} 有可能落在约束集合 X 外面，因此需要将 x_{k+1} 投影回约束集合 X。我们得到非光滑约束优化问题的投影次梯度算法，

算法 3.1 投影次梯度法

输入：$x_0 \in \mathbb{R}^n$，对 $k = 0, 1, \cdots$ 执行迭代，

$$x_{k+1} = \operatorname*{argmin}_{x \in X} \|x - (x_k - \gamma_k g(x_k))\|_2. \tag{3.3}$$

其中 $\gamma_k > 0$ 且 $g(x_k) \in \partial f(x_k)$.

我们可以用对目标函数做近似逼近的角度来理解投影次梯度法的迭代 (3.3)。事实上，(3.3) 可以被写成，

$$\begin{aligned} x_{k+1} &= \operatorname*{argmin}_{x \in X} \frac{1}{2}\|x - (x_k - \gamma_k g(x_k))\|_2^2 && (3.4)\\ &= \operatorname*{argmin}_{x \in X} \gamma_k \langle g(x_k), x - x_k\rangle + \frac{1}{2}\|x - x_k\|_2^2 && (3.5)\\ &= \operatorname*{argmin}_{x \in X} \gamma_k \langle g(x_k), x\rangle + \frac{1}{2}\|x - x_k\|_2^2 && (3.6)\\ &= \operatorname*{argmin}_{x \in X} f(x_k) + \langle g(x_k), x - x_k\rangle + \frac{1}{2\gamma_k}\|x - x_k\|_2^2. && (3.7) \end{aligned}$$

这意味着我们需要最小化 $f(x)$ 在 X 上的线性近似 $f(x_k) + \langle g(x_k), x - x_k\rangle$ 同时求得的解 x_{k+1} 离上一步迭代点 x_k 不能太远（被 $\frac{1}{2\gamma_k}\|x - x_k\|_2^2$ 控制）。为了得到投影梯度法的收敛速率，我们需要对 x_{k+1} 的性质做一定的刻画，有下述引理。

引理 3.1 *投影次梯度法 (3.3) 确定的 x_{k+1}，对任意 $x \in X$，我们有*

$$\gamma_k \langle g(x_k), x_{k+1} - x\rangle + \frac{1}{2}\|x_{k+1} - x_k\|_2^2 \leq \frac{1}{2}\|x - x_k\|_2^2 + \frac{1}{2}\|x - x_{k+1}\|_2^2. \tag{3.8}$$

证明： 记 $\phi(x) = \gamma_k \langle g(x_k), x\rangle + \frac{1}{2}\|x - x_k\|_2^2$，由 $\phi(x)$ 的强凸性，我们有，

$$\phi(x) \geq \phi(x_{k+1}) + \langle \phi'(x_{k+1}), x - x_{k+1}\rangle + \frac{1}{2}\|x - x_{k+1}\|_2^2. \tag{3.9}$$

注意到 $x_{k+1} = \operatorname{argmin}_{x \in X} \phi(x)$，由其一阶最优性条件，我们有，

$$\langle \phi'(x_{k+1}), x - x_{k+1}\rangle \geq 0, \quad \text{对任意 } x \in X. \tag{3.10}$$

因此我们有，

$$\phi(x) - \phi(x_{k+1}) \geq \frac{1}{2}\|x - x_{k+1}\|_2^2. \tag{3.11}$$

整理上式子即可得 (3.8)。 □

我们考虑 $X \subset \mathbb{R}^n$ 闭且凸，$f: X \to \mathbb{R}^n$ 连续且凸，其次梯度满足

$$\|g(x)\| \leq M. \tag{3.12}$$

对任意 $x \in X$ 成立。

定理 3.1 令 x_k，$k=1,\cdots,N$，由 (3.3) 产生，且定义

$$\bar{x}_s^N = \frac{\gamma_s x_s + \cdots + \gamma_N x_N}{\gamma_s + \cdots + \gamma_N} = \left(\sum_{k=s}^{N}\gamma_k\right)^{-1} \sum_{k=s}^{k}\gamma_k x_k. \tag{3.13}$$

则有

$$f(\bar{x}_s^N) - f^* \le \left(2\sum_{k=s}^{N}\gamma_k\right)^{-1}\left[\|x_s - x^*\|_2^2 + M^2\sum_{k=s}^{N}\gamma_k^2\right]. \tag{3.14}$$

证明: 由 f 的凸性，以及引理 (3.1) 有，

$$\begin{aligned}
& \gamma_k[f(x_k) - f(x)] \le \gamma_k\langle g(x_k), x_k - x\rangle && (3.15)\\
= \ & \gamma_k\langle g(x_k), x_k - x_{k+1}\rangle + \gamma_k\langle g(x_k), x_{k+1} - x\rangle && (3.16)\\
\le \ & \gamma_k\langle g(x_k), x_k - x_{k+1}\rangle - \frac{1}{2}\|x_{k+1} - x_k\|_2^2 + \frac{1}{2}\|x - x_k\|_2^2 - \frac{1}{2}\|x - x_{k+1}\|_2^2 && (3.17)\\
\le \ & \frac{\gamma_k^2}{2}\|g(x_k)\|_2^2 + \frac{1}{2}\|x - x_k\|_2^2 - \frac{1}{2}\|x - x_{k+1}\|_2^2 && (3.18)\\
\le \ & \frac{M^2}{2}\|g(x_k)\|_2^2 + \frac{1}{2}\|x - x_k\|_2^2 - \frac{1}{2}\|x - x_{k+1}\|_2^2. && (3.19)
\end{aligned}$$

其中不等式 (3.18) 成立时因为 $bc^2 + ac^2/2 \le b^2/(2a)$。将上式对 $k=s$ 到 N 求和，并利用 f 的凸性，我们可以得到不等式 (3.14)。 □

推论 3.1 对于投影次梯度法 *3.1*，选取不同的步长策略，有如下收敛性结果:

1. 固定步长策略: 假设投影次梯度法 *3.1* 的总迭代步数 N 固定，令 $D_X = \max_{x_1,x_2\in X}\|x_1 - x_2\|$，取固定步长，

$$\gamma_k = \sqrt{\frac{D_X^2}{NM^2}}, \quad k=1,\cdots,N, \tag{3.20}$$

则，

$$f(\bar{x}_1^N) - f^* \le \frac{MD_X}{2\sqrt{N}}, \quad \text{对任意 } N \ge 1. \tag{3.21}$$

2. 变步长策略: 取变步长，

$$\gamma_k = \sqrt{\frac{D_X^2}{kM^2}}, \quad k=1,\cdots, \tag{3.22}$$

则，

$$f(\bar{x}_{\lceil k/2\rceil}^N) - f^* \le \mathcal{O}(1)\frac{MD_X}{\sqrt{k}}, \tag{3.23}$$

其中 $\mathcal{O}(1)$ 是固定常数。

证明: 当去固定步长时候，令 $\gamma_k = \gamma$，$k=1,\cdots,N$，代入定理 (3.1) 得不等式，

$$f(\bar{x}_1^N) - f^* \le \frac{1}{2}\left[\frac{D_X^2}{N\gamma} + M^2\gamma\right].$$

对上式右端求极小，可得所求的固定步长 γ 取值。对变步长策略，只需代如验证即可。 □

考虑到投影次梯度法 3.1 与 $\mathbb{R}^n$ 的结构有关。更确切的说，其中的投影步的距离度量选择的是欧几里得范数（即 ℓ_2 范数），同时 D_X 与 M 的值也是在 ℓ_2 范数下计算的。这就提醒我们可以对投影次梯度法做一定的推广，使得其在非欧几里得结构下执行，这就是镜像梯度法的由来。

令 $\|\cdot\|$ 是 $\mathbb{R}^n$ 上的一个范数，$\|\xi\|_* = \max\{\langle \xi, x\rangle : \|x\| \le 1\}$ 是它的对偶范数。定义相对于该范数 $\|\cdot\|$ 的距离生成函数 $\omega : X \to \mathbb{R}$，$\omega(x)$ 连续可微且强凸（相对于范数 $\|\cdot\|$ 和参数 μ），即

$$\langle \nabla\omega(x) - \nabla\omega(y), x - y\rangle \ge \mu\|x - y\|^2, \quad \text{对任意 } x, y \in X. \tag{3.24}$$

最简单的距离生成函数是 $\omega(x) = \|x\|_2^2/2$（取 ℓ_2 范数且 $\mu = 1$）。有了距离生成函数，我们可以定义 Bregman 距离，

$$V(x, y) = \omega(y) - [\omega(x) + \langle \nabla\omega(x), y - x\rangle]. \tag{3.25}$$

注意到 $V(x,\cdot)$ 是强凸函数（相对于范数 $\|\cdot\|$ 以及参数 μ）。当取 $\omega(x) = \|x\|_2^2/2$ 时，$V(x, y) = \|y - x\|_2^2/2$。

注意到投影次梯度法 (3.1) 等价于 (3.6)，利用 Bregman 距离，我们可以构造在新的范数 $\|\cdot\|$ 下的投影次梯度法（即镜像梯度法）。

算法 3.2 镜像梯度法

输入：$x_0 \in \mathbb{R}^n$，对 $k = 0, 1, \cdots$ 执行迭代，

$$x_{k+1} = \operatorname*{argmin}_{x\in X} \ \gamma_k\langle g(x_k), x\rangle + V(x_k, x). \tag{3.26}$$

其中 $\gamma_k > 0$ 且 $g(x_k) \in \partial f(x_k)$

注意到当镜像梯度法 3.2 中 $V(x, y) = \|y - x\|_2^2/2$ 时，镜像梯度法 3.2 就是投影次梯度法 3.1。与引理 (3.1) 类似，我们也有对镜像梯度法 x_{k+1} 性质的刻画，有下述引理。

引理 3.2 *镜像梯度法 (3.26) 确定的 x_{k+1}，对任意 $x \in X$，我们有*

$$\gamma_k\langle g(x_k), x_{k+1} - x\rangle + V(x_k, x_{k+1}) \le V(x_k, x) - V(x_{k+1}, x). \tag{3.27}$$

证明： 由 (3.26) 的最优性条件，对任意 $x \in X$ 不等式

$$\langle \gamma_k g(x_k) + \nabla V(x_k, x_{k+1}), x - x_k\rangle \ge 0 \tag{3.28}$$

成立。其中 $\nabla V(x_k, x_{k+1})$ 表示 $V(x_k,\cdot)$ 在 x_{k+1} 处的导数。由 $V(x, y)$ 的定义 (3.25)，对任意 $x \in X$ 等式

$$V(x_k, x) = V(x_k, x_{k+1}) + \langle \nabla V(x_k, x_{k+1}), x - x_{k+1}\rangle + V(x_{k+1}, x). \tag{3.29}$$

成立。将 (3.26) 带入上面等式即得结果。 □

定理 3.2 *令 x_k，$k = 1, \cdots, N$，由 (3.26) 产生，且定义*

$$\bar{x}_s^N = \frac{\gamma_s x_s + \cdots + \gamma_N x_N}{\gamma_s + \cdots + \gamma_N} = \left(\sum_{k=s}^N \gamma_k\right)^{-1} \sum_{k=s}^k \gamma_k x_k. \tag{3.30}$$

则有

$$f(\bar{x}_s^N) - f^* \le \left(\sum_{k=s}^N \gamma_k\right)^{-1} \left[V(x_s, x^*) + \frac{M^2}{2\mu}\sum_{k=s}^N \gamma_k^2\right]. \tag{3.31}$$

仿照定理 3.1 的证明方式，利用引理 3.2 我们可以证明定理 3.2。

推论 3.2 对于镜像梯度法 *3.2*，选取不同的步长策略，令 $D_{\omega,X}^2 = \max_{x_1,x_2\in X} V(x_1,x_2)$，取步长

$$\gamma_k = \sqrt{\frac{2\mu D_{\omega,X}^2}{kM^2}}, \quad k = 1,\cdots \tag{3.32}$$

则，

$$f(\bar{x}_1^k) - f^* \le \frac{\sqrt{2}MD_{\omega,X}}{\sqrt{\mu k}}, \quad \text{对任意 } k \ge 1. \tag{3.33}$$

证明：将 (3.32) 代入 (3.31) 即得(3.33) 结果。 □

3.2 重心法与椭球法

这一节我们列出重心法和椭球法的收敛性和复杂度。其收敛性证明可以在文献 [8] 中找到。

问题模型 3.1 考虑约束凸优化问题集合 $\mathcal{F} = (\Sigma, \mathcal{O}, \mathcal{T}_\epsilon)$:

- 全局信息 Σ: 目标函数 $f(x) \in C(X)$ 且是凸函数，存在常数 $B > 0$ 使得对任意 $x \in X$ 有 $-B \le f(x) \le B$；约束集合 $X \subset \mathbb{R}^n$ 闭且凸，
- 局部信息 $\mathcal{O}$: $\mathcal{FO}$ 子程序，对于任意给定的 x_0 返回函数值 $f(x_0)$ 和次梯度 $\partial f(x_0)$，
- 解的精度 $\mathcal{T}_\epsilon$: 求全局极小点 $\bar{x} \in \mathbb{R}^n$，使得 $f(\bar{x}) - f^* \le \epsilon$。

算法 3.3 重心法

令 $\mathcal{S}_1 = X$，对 $k = 1, ..., N$ 执行迭代，

1. 计算集合 $\mathcal{S}_k$ 的重心
$$c_k = \frac{1}{\mathrm{vol}(\mathcal{S}_k)} \int_{x\in\mathcal{S}_k} x dx. \tag{3.34}$$
2. 在 c_k 处调用 $\mathcal{FO}$ 子程序得 $w_k \in \partial f(c_k)$，更新集合 $\mathcal{S}_{k+1}$
$$\mathcal{S}_{k+1} = \mathcal{S}_k \cap \{x \in \mathbb{R}^n : \ (x - c_k)^T w_k \le 0\}. \tag{3.35}$$

输出: $x_N \in \mathrm{argmin}_{1\le r\le N} f(c_r)$

引理 3.3 设 $\mathcal{S}$ 是重心在原点的凸集合，即 $\int_{x\in\mathcal{S}} x dx = 0$，$\mathrm{vol}(\mathcal{S})$ 表示 $\mathcal{S}$ 的体积，则对任意 $w \in \mathbb{R}^n$，$w \ne 0$ 有

$$\mathrm{vol}\left(\mathcal{S} \cap \{x \in \mathbb{R}^n : \ x^T w \ge 0\}\right) \ \ge \ \frac{1}{e}\mathrm{vol}(\mathcal{S}). \tag{3.36}$$

定理 3.3 重心法 *3.3* 在求解问题模型 (3.1) 时的收敛速率为，

$$f(x_N) - f^* \ \le \ 2B\left(1 - \frac{1}{e}\right)^{N/n}. \tag{3.37}$$

复杂度上界为

$$\mathcal{O}\left(n \log \frac{2B}{\epsilon}\right). \tag{3.38}$$

从计算角度来说，由于每一步都需要计算集合的体积，即进行积分计算，因此重心法计算量过大，往往很难执行。重心法只是提供一个理论上的可行策略。而椭球法借用了重心法的思想，改进了其计算可行性。

椭球是具有如下形式的凸集合，

$$\mathcal{E} = \{x \in \mathbb{R}^n : \ (x-c)^T H^{-1}(x-c) \leq 1\}. \tag{3.39}$$

其中 $c \in \mathbb{R}^n$，H 是一个对称正定矩阵。几何上 c 是椭球 $\mathcal{E}$ 的重心，而 H 的特征向量是椭球 $\mathcal{E}$ 的半轴，半轴的长度是对应特征值的正平方根。

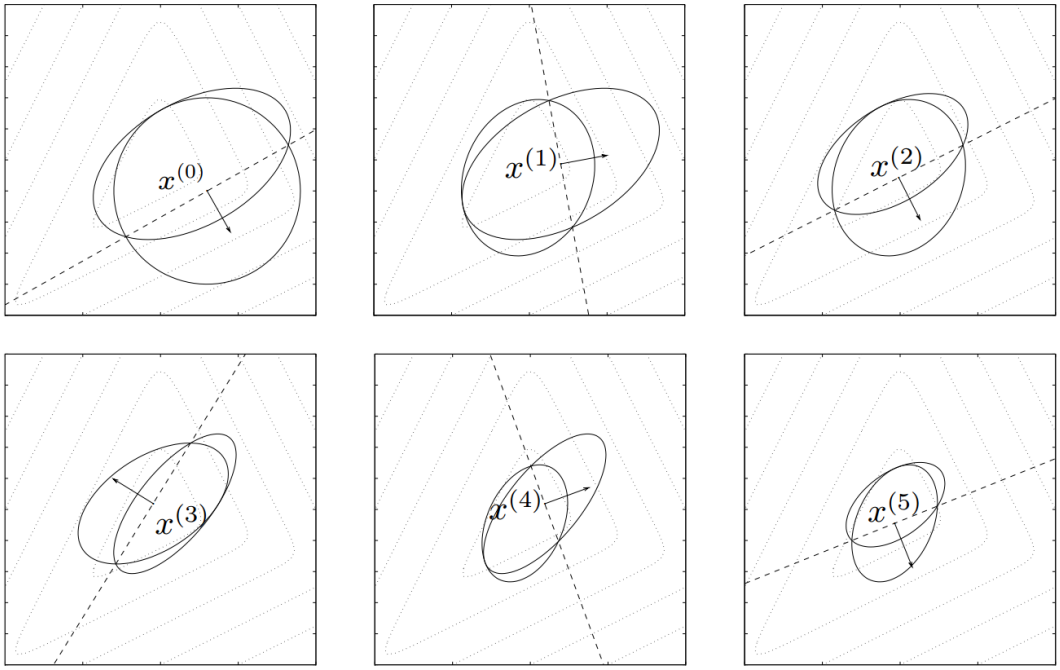

图 6: 椭球法示意图

引理 3.4 令 $\mathcal{E}_0 = \{x \in \mathbb{R}^n : \ (x-c_0)^T H_0^{-1}(x-c_0) \leq 1\}$，对任意 $w \in \mathbb{R}^n$，$\omega \neq 0$，可以构造椭球 $\mathcal{E}$ 使得

$$\mathcal{E} \supset \{x \in \mathcal{E}_0 : \ w^T(x-c_0) \leq 0\}. \tag{3.40}$$

且原椭球 $\mathcal{E}_0$ 和新椭球 $\mathcal{E}$ 之间体积有如下关系，

$$\mathrm{vol}(\mathcal{E}) \ \leq \ \exp\left(-\frac{1}{2n}\right)\mathrm{vol}(\mathcal{E}_0). \tag{3.41}$$

当 $n \geq 2$ 时，新椭球 $\mathcal{E} = \{x \in \mathbb{R}^n : \ (x-c)^T H^{-1}(x-c) \leq 1\}$ 其中，

$$c \ = \ c_0 - \frac{1}{n+1}\frac{H_0 w}{\sqrt{w^T H_0 w}}, \tag{3.42}$$

$$H \ = \ \frac{n^2}{n^2-1}\left(H_0 - \frac{2}{n+1}\frac{H_0 w w^T H_0)}{w^T H_0 w}\right). \tag{3.43}$$

算法 3.4 椭球法（The ellipsoid method）

输入：令 $\mathcal{E}_0$ 为包含约束集合 X 且半径为 R 中心为 c_0 的球，令 $H_0 = R^2 I_n$，

对 $k = 1, ..., N$ 执行迭代

1. 若 $c_k \notin X$，则调用子程序返回向量 $w_k \in \mathbb{R}^n$，过 c_k 且垂直 w_k 的平面将椭球 $\mathcal{E}_k$ 分割，使得约束集 $X \subset \{x : (x - c_k)^T w_k \leq 0\}$；

   若 $c_k \in X$，则调用子程序返回向量 $w_k \in \partial f(c_k)$.

2. 构造新的椭球 $\mathcal{E}_{k+1} = \{x : (x - c_{k+1})^T H_{k+1}^{-1}(x - c_{k+1}) \leq 1\}$ 使得

$$\{x \in \mathcal{E}_k : (x - c_k)^T w_k \leq 0\} \subset \mathcal{E}_{k+1}. \tag{3.44}$$

   其中 c_{k+1}，H_{k+1} 满足，

$$c_{k+1} = c_k - \frac{1}{n+1} \frac{H_k w}{\sqrt{w^T H_k w}}, \tag{3.45}$$

$$H_{k+1} = \frac{n^2}{n^2 - 1} \left(H_k - \frac{2}{n+1} \frac{H_k w w^T H_k)}{w^T H_k w} \right). \tag{3.46}$$

输出：若 $\{c_1, ..., c_N\} \cap X \neq \varnothing$，则输出

$$x_N \in \mathrm{a\,rgmin}_{c \in \{c_1, ..., c_N\} \cap X} f(c). \tag{3.47}$$

问题模型 3.2 考虑约束凸优化问题集合 $\mathcal{F} = (\Sigma, \mathcal{O}, \mathcal{T}_\epsilon)$:

- 全局信息 Σ：目标函数 $f(x) \in C(X)$ 且是凸函数，存在常数 $B > 0$ 使得对任意 $x \in X$ 有 $-B \leq f(x) \leq B$；约束集合 $X \subset \mathbb{R}^n$ 闭且凸，且 $\mathcal{B}(r) \subset X \subset \mathcal{B}(R)$，
- 局部信息 $\mathcal{O}$：$\mathcal{FO}$ 子程序，对于任意给定的 x_0 返回函数值 $f(x_0)$ 和次梯度 $\partial f(x_0)$；$\mathcal{SO}$ 子程序，返回分割平面垂直向量 ω 使得，

$$w^T(x - c_0) \leq 0, \quad \forall x \in X. \tag{3.48}$$

- 解的精度 $\mathcal{T}_\epsilon$：求全局极小点 $\bar{x} \in \mathbb{R}^n$，使得 $f(\bar{x}) - f^* \leq \epsilon$。

我们可以估计椭球法的如下收敛速率和复杂度。

定理 3.4 椭球法 *3.4* 在求解问题模型 *3.2* 时的收敛速率为

$$f(x_N) - f^* \leq \frac{2BR}{r} \exp\left(-\frac{N}{2n^2} \right). \tag{3.49}$$

复杂度上界为

$$\mathcal{O}\left(n^2 \log \frac{BR}{r\epsilon} \right). \tag{3.50}$$

从上面的定理我们可以看出，

1. 从子程序调用的分析复杂度来看，椭球法要比重心法差，前者需要调用 $\mathcal{FO}$ 子程序 $\mathcal{O}(n^2 \log(\frac{2BR}{r\epsilon}))$ 次，而后者仅需要 $\mathcal{O}(n \log(\frac{2B}{\epsilon}))$。
2. 从计算角度来看，椭球法要比重心法更容易执行。

3.3 复合优化的 Nesterov 加速算法

这一节我们介绍复合优化问题的复杂度分析。考虑复合优化问题

$$\min_{x\in\mathbb{R}^n} \Phi(x) := \phi(x) + h(x), \quad \phi(x) = f(x) + g(x). \tag{3.51}$$

其中 $f(x) \in C^{1,1}_{L_f}(\mathbb{R}^n)$ 是非凸函数，$g(x) \in C^{1,1}_{L_g}(\mathbb{R}^n)$ 是凸函数，于是我们有 $\phi(x) \in C^{1,1}_{L_\phi}$，其中 $L_\phi = L_f + L_g$。$h(x)$ 是定义域有界的简单凸函数且有可能不光滑（例如，$h(x) = \mathcal{I}_X(\cdot)$，其中 $\mathcal{I}_X(\cdot)$ 是凸紧集合 X 的示性函数；$h(x) = \|x\|_1$，是机器学习损失函数中常用的 ℓ_1 正则项）。可以看到优化问题 (3.51) 由光滑部分 $\phi(x)$ 和非光滑部分 $h(x)$ 组成，其中光滑部分又分为非凸函数 $f(x)$ 和凸函数 $g(x)$。这一节，我们介绍复合优化问题的复杂度分析。当 $f(x) = 0$ 时，目标函数 (3.51) 是凸函数，我们使用近似点梯度法（Proximal gradient method）和加速梯度法（FISTA 和 Nesterov 加速算法）求解。具体可参考文献 [40, 48, 4, 5, 37, 39]。当 $f(x) \neq 0$ 时，目标函数 (3.51) 是非凸函数，文献 [16, 17, 15] 给出了非凸函数情形下的复杂度分析。

3.3.1 凸复合优化问题

当 $f(x) = 0$ 时，复合优化问题 (3.51) 是一个凸优化问题

$$\min_{x\in\mathbb{R}^n} \Phi(x) := g(x) + h(x). \tag{3.52}$$

令 $\mathbf{prox}_h(\cdot)$ 是函数 $h(x)$ 的近似点算子。对任意 $x \in \mathbb{R}^n$，函数 $h(x)$ 的近似点 $\mathbf{prox}_h(x)$ 定义为：

$$\mathbf{prox}_h(x) = \operatorname*{argmin}_u \left(h(u) + \frac{1}{2}\|u - x\|_2^2 \right). \tag{3.53}$$

当 $h(x)$ 是闭凸函数时，$\mathbf{prox}_h(x)$ 存在且唯一。我们可以用如下近似点梯度法求解问题 (3.52)。

算法 3.5 近似点梯度法

输入：$x_0 \in \mathbb{R}^n$，步长序列 $\{t_k\}$，对 $k = 1, \cdots$ 执行迭代，

$$x_k = \mathbf{prox}_{t_k h}\left(x_{k-1} - t_k \nabla g(x_{k-1})\right). \tag{3.54}$$

当 $h(x) = 0$ 时，$\mathbf{prox}_h(x) = x$，算法 3.5 等价于光滑函数的梯度法。当 $h(x) = \mathcal{I}_C(x)$ 时，$\mathbf{prox}_h(x) = \operatorname{argmin}_{u\in C} \|u - x\|^2$，算法 3.5 就是光滑函数的投影梯度法。当 $h(x) = \|x\|_1$ 时，$\mathbf{prox}_h(x)$ 被称为软阈算子（soft-threshold operator），算法 3.5 可以用来求解机器学习中带 ℓ_1 正则的目标优化函数。

定理 3.5 *算法 3.5 求解凸复合优化问题 (3.52) 时若取固定步长 $t_k = \frac{1}{L_g}$，则*

$$\Phi(x_k) - \Phi(x_*) \leq \frac{L_g}{2k}\|x_0 - x_*\|^2. \tag{3.55}$$

可以看到，算法 3.5 的复杂度为 $\mathcal{O}(\frac{1}{\epsilon})$。文献 [4] 给出了如下加速版本的近似点梯度法 FISTA。

算法 3.6 FISTA

输入：$x_0 = x_{-1} \in \mathbb{R}^n$，对 $k = 1, \cdots$ 执行迭代，

$$y = x_{k-1} + \frac{k-2}{k+1}(x_{k-1} - x_{k-2}), \tag{3.56}$$

$$x_k = \mathbf{prox}_{t_k h}(y - t_k \nabla g(y)). \tag{3.57}$$

下面定理给出了 FISTA 算法的复杂度结果。

定理 3.6 算法 *3.6* 求解凸复合优化问题 (3.52) 时若取固定步长 $t_k = \frac{1}{L_g}$，则

$$\Phi(x_k) - \Phi(x_*) \leq \frac{2L_g}{(k+1)^2}\|x_0 - x_*\|^2. \tag{3.58}$$

复合优化问题 (3.52)还有另一种加速算法形式：

算法 3.7 复合函数的 Nesterov 加速算法

输入： 令 $x_0 = y_0$, 选取序列 $\{\gamma_k\}$ 和 $\{\beta_k\}$ 使其满足 $\gamma_k \leq L_g\beta_k$, $\gamma_1 = 1$。

对于$k = 1, ..., N$，执行迭代

1. $z_k = (1 - \gamma_k)y_{k-1} + \gamma_k x_{k-1}$,
2. $x_k = \mathbf{prox}_{(\beta_k/\gamma_k)h}\left(x_{k-1} - \frac{\beta_k}{\gamma_k}\nabla g(z_k)\right)$,
3. $y_k = (1 - \gamma_k)y_{k-1} + \gamma_k x_k$。

输出： y_N

同样，该算法也有 $\mathcal{O}(\frac{1}{\sqrt{\epsilon}})$ 复杂度结果。

定理 3.7 算法 *3.7* 求解凸复合优化问题 (3.52) 时，取 $\gamma_k = \frac{2}{k+1}$，$\beta_k = \frac{1}{L_g}$，则

$$\Phi(x_k) - \Phi(x_*) \leq \frac{2L_g}{(k+1)^2}\|x_0 - x_*\|^2. \tag{3.59}$$

3.3.2 非凸复合优化问题

当 $f(x) \neq 0$ 时，复合优化问题 (3.51) 是一个非凸优化问题。在非凸函数设定下，我们无法用函数值与最优值的差来衡量优化算法解的精度，因为 $f(x_k) - f(x_*)$ 并非恒大于零。类似于非凸光滑函数利用梯度作为停止准则，对于非凸复合函数 (3.51)，利用梯度映射（gradient mapping）作为停止准则。其定义如下，

$$G_\alpha(x, y) = \frac{1}{\alpha}[x - \mathbf{prox}_{\alpha h}(x - \alpha y)]. \tag{3.60}$$

由近似点梯度法 3.5 可以看到 $G_{t_k}(x_{k-1})$ 是近似点迭代步的负方向，即

$$\begin{aligned} x_k - x_{k-1} &= \mathbf{prox}_{t_k h}\left(x_{k-1} - t_k\nabla g(x_{k-1})\right) - x_{k-1} &(3.61)\\ &= -t_k G_{t_k}(x_{k-1}, \nabla g(x_{k-1})). &(3.62)\end{aligned}$$

同时，由近似点算子次梯度的定义我们有，

$$G_\alpha(x, \nabla\phi(x)) \in \nabla\phi(x) + \partial h(x - \alpha G_\alpha(x, \nabla\phi(x))). \tag{3.63}$$

由此我们可以得到结论：$G_\alpha(x, \nabla\phi(x)) = 0$ 当且仅当 x 是 $\phi(x) + h(x)$ 的局部极小点。将算法 2.4 做一些修改，文献 [16] 给出了非凸复合函数的加速梯度法框架，并用梯度映射作为停止条件分析了算法的复杂度。

算法 3.8 复合函数加速算法框架

输入：令 $x_0 = y_0 \in R^n$，取 $\{\gamma_k\}$ 使得 $\gamma_1 = 1$ 且当 $k \geq 2$ 时 $\gamma_k \in (0,1)$。

对$k = 1, ..., N$，执行迭代

1. $z_k = (1-\gamma_k)y_{k-1} + \gamma_k x_{k-1}$，
2. $x_k = \mathbf{prox}_{\lambda_k h}\left(x_{k-1} - \lambda_k \nabla\phi(z_k)\right)$，
3. $y_k = \mathbf{prox}_{\beta_k h}\left(z_k - \beta_k \nabla\phi(z_k)\right)$.

输出: 输出 z_N。

我们利用 $\|G_{\beta_k}(z_k, \nabla\phi(z_k))\|$ 作为算法 3.8 的停止准则。注意到

$$G_{\beta_k}(z_k, \nabla\phi(z_k)) = \frac{1}{\beta_k}(z_k - y_k), \tag{3.64}$$

该停止条件实际上等价于解序列 $\{z_k\}$ 和 $\{x_k\}$ 之间的距离。

定理 3.8 *假设对任意 $\alpha \in (0, +\infty)$ 以及任意 $x, y \in \mathbb{R}^n$ 存在常数 $M > 0$ 使得 $\|\mathbf{prox}_{\alpha h}(x - \alpha y)\| \leq M$，算法 3.8 求解凸复合优化问题 (3.51) 时，取 $\gamma_k = \frac{2}{k+1}$，$\beta_k = \frac{1}{2L_\phi}$，$\lambda_k = \frac{k\beta_k}{2}$，则对任意的 $N \geq 1$ 有，*

$$\min_{k=1,...,N} \|G_{\beta_k}(z_k, \nabla\phi(z_k))\|^2 \;\leq\; 24L_\phi \left[\frac{4L_\phi\|x_0 - x_*\|^2}{N^2(N+1)} + \frac{L_f}{N}(\|x_*\|^2 + 2M^2)\right]. \tag{3.65}$$

另外，当 $f(x) = 0$ 时，有

$$\Phi(y_N) - \Phi(x_*) \;\leq\; \frac{4L_g\|x_0 - x_*\|^2}{N(N+1)}. \tag{3.66}$$

由定理 3.8 可以看出，当目标函数 (3.51) 凸时（$f(x) = 0$），算法 3.8 的复杂度与 FISTA 相同，两者都为 $\mathcal{O}(L_g^2/\sqrt{\epsilon})$。当目标函数 (3.51) 非凸时（$f(x) \neq 0$），算法 3.8 也收敛，且复杂度为 $\mathcal{O}(L_\phi^{\frac{2}{3}}/\epsilon^{\frac{1}{3}} + L_\phi L_f/\epsilon)$。

4 条件梯度算法的复杂度分析

4.1 引言

考虑带线性约束的优化问题:

$$\min_{x \in X} f(x). \tag{4.1}$$

其中 X 是线性约束集合。我们可以用投影梯度法求解上述问题。设上述问题的线性约束集合为 X，在第 k 步迭代点 x_k 处，通过极小化目标函数的局部二次逼近来更新 x_{k+1}，

$$x_{k+1} = \operatorname*{argmin}_{x \in X} \left\{ f(x_k) + \langle \nabla f(x_k), x - x_k\rangle + \frac{1}{2\alpha_k}\|y - x_k\|^2 \right\}. \tag{4.2}$$

即投影梯度迭代，

$$x_{k+1} = \mathcal{P}_X(x_k - \alpha_k \nabla f(x_k)). \tag{4.3}$$

注意到约束集合 X 都是线性约束，X 的几何形状是一个多面体。而子迭代 (4.2) 是一个二次规划问题。如果线性约束集合 X 很复杂，则 (4.3) 投影步的计算量很大，投影梯度法效率会很低。但是注意到子迭代步 (4.2) 是对优化问题 (4.1) 的一个局部二次逼近，如果去掉其二次项只做一次逼近，则子问题 (4.3) 会变成一个线性规划问题。

$$x_{k+1} = \operatorname*{argmin}_{x \in X} \left\{ f(x_k) + \langle \nabla f(x_k), x - x_k\rangle \right\}. \tag{4.4}$$

这样子迭代的复杂度会降低很多。这就是条件梯度法的核心想法，通过局部线性逼近来更新 x_{k+1}。这样不必做约束集合的投影计算，降低了子问题的计算复杂度。

本章我们主要用复杂度分析的框架来研究条件梯度法的复杂度。注意到子迭代 (4.4) 实际上等价于，

$$x_{k+1} = \operatorname*{argmin}_{x \in X} \langle \nabla f(x_k), x \rangle. \tag{4.5}$$

在做复杂度分析时，我们假设存在一个关于线性约束集合 X 的子程序 $\mathcal{LO}$ ，对任意给定向量 $p \in \mathbb{R}^n$ 返回线性规划的解 y ，满足，

$$y \in \operatorname{argmin}_{x \in X} \langle p, x \rangle. \tag{4.6}$$

对任意给定的线性约束集合 X 和向量 $p \in \mathbb{R}^n$，我们能够很容易确定子程序 $\mathcal{LO}$ 的计算复杂度（例如内点法求解线性规划的计算复杂度）。因此我们可以研究条件梯度法关于子程序 $\mathcal{LO}$ 分析复杂度（即条件梯度法为了达到解精度 ϵ 子程序 $\mathcal{LO}$ 所需调用的次数），以及条件梯度法的复杂度下界，并且研究加速条件梯度法的可能性。

条件梯度法（The Conditional Gradient(CndG) Method ）又被称为 Frank-Wolfe 算法，其算法原型由 Frank and Wolfe (1956) [11] 提出，其发展历史和近几年的研究可以参考文献 [1]，[2]，[3]，[9]，[12]，[21]，[19]，[20]，[26]，[28]。近几年来，条件梯度法受到了机器学习和优化领域的广泛关注，主要由于如下原因：

- 子迭代复杂度低：条件梯度法的子问题是求解一个线性逼近问题，在很多情况下会比梯度法的非线性逼近子问题要简单。在下一节里面，我们会给出带范数不等式约束优化问题的例子，比较投影梯度法和条件梯度法在求解它们时的效率。
- 简单性：条件梯度法在实现时不需要精心挑选步长，且步长不依赖于导数的 Lipschitz 常数。而梯度法的实现却非常依赖步长的选择，需要反复的调整以达到收敛的目的。文献 [21]，[19]，[20] 给出更详细的解释。
- 解具有结构性：下一节我们可以看到，条件梯度法的解是约束集合 X 极点的凸组合，因此具有稀疏性和低秩性。

本章我们主要介绍条件梯度法在凸函数和强凸函数下的复杂度分析，以及具有 $\mathcal{LO}$ 子程序的算法在求解约束光滑优化问题时的复杂度下界，以及利用 Nesterov 加速梯度法框架导出加速条件梯度法，并分析其复杂度。

4.2 条件梯度法

4.2.1 凸优化问题

这一节，我们研究条件梯度法求解凸函数和强凸函数问题时的收敛性和复杂度，详细可参考文献 [26]，[28]。首先我们给出条件梯度法求解凸函数优化问题的具体算法：

算法 4.1 条件梯度法 (CndG，凸函数版本)

输入：给定 $x_0 \in X$，令 $y_0 = x_0$，取步长 $\alpha_k \in [0,1]$，

对 $k = 1, ..., N$ 执行迭代

1. 在 y_{k-1} 处调用子程序 $\mathcal{FO}$，返回 $\nabla f(y_{k-1})$;
2. 在 $\nabla f(y_{k-1})$ 处调用子程序 $\mathcal{LO}$，返回 $x_k \in \operatorname{argmin}_{x\in X}\langle \nabla f(y_{k-1}), x\rangle$;
3. 令 $y_k = (1-\alpha_k)y_{k-1} + \alpha_k x_k$.

输入：y_N。

条件梯度法有两个迭代点列 $\{x_k\}$ 和 $\{y_k\}$，其中 $\{y_k\}$ 是收敛点列。经过简单变化可以得到 $y_k = \operatorname{conv}\{x_0, x_1, ..., x_k\}$，即收敛点 y_k 是点集合 $\{x_0, x_1, ..., x_k\}$ 的凸组合。另外由 $y_k = y_{k-1} + \alpha_k(x_k - y_{k-1})$，条件梯度法可以看做 y_k 按照方向 $x_k - y_{k-1}$ 的方向按照步长 α_k 进行更新。为了保证条件梯度法的收敛性，我们有两种步长的选择方式，一种是选取严格递减到零的变步长序列：

$$\alpha_k = \frac{2}{k+1}, \quad k = 1, 2, ... \tag{4.7}$$

令一种是采取精确搜索的步长，求解一个一维的优化问题：

$$\alpha_k = \operatorname*{argmin}_{\alpha\in[0,1]} f((1-\alpha)y_{k-1} + \alpha x_k). \tag{4.8}$$

在研究条件梯度法的收敛性和复杂度之前，我们给出几个条件梯度法比投影梯度法效率高的例子。

例 4.1 范数约束问题

考虑带某一范数 $\|\cdot\|$ 约束的凸优化问题，

$$\min_x \ f(x) \quad \text{s.t.} \quad \|x\| \le t. \tag{4.9}$$

用条件梯度法求解该问题时，需要计算子问题，

$$\begin{aligned} x_k &\in \operatorname*{argmin}_{\|x\|\le t} \langle \nabla f(y_{k-1}), x\rangle && (4.10)\\ &= -t\cdot\left(\operatorname*{argmax}_{\|x\|\le 1} \langle \nabla f(y_{k-1}), x\rangle\right) && (4.11)\\ &= -t\cdot\partial\|\nabla f(y_{k-1})\|_*. && (4.12) \end{aligned}$$

其中 $\|z\|_* = \sup\{z^T x, \|x\| \le 1\}$ 是 $\|\cdot\|$ 的对偶范数。注意到 (4.12) 条件梯度法的子问题相当于计算一个对偶范数的次梯度。如果计算 $\|\cdot\|$ 范数的次梯度比计算在约束集合 $X = \{x \in \mathbb{R}^n : \|x\| \le t\}$ 上的投影要简单，条件梯度法比投影梯度法效率更高。

例 4.2 ℓ_1 范数约束问题

考虑带 ℓ_1 范数约束的凸优化问题，

$$\min_x \ f(x) \quad \text{s.t.} \quad \|x\|_1 \le t. \tag{4.13}$$

由于 ℓ_1 范数的对偶范数是 ℓ_∞ 范数，因此用条件梯度法求解该问题时子问题为，

$$x_k \in -t\cdot\partial\|\nabla f(y_{k-1})\|_\infty. \tag{4.14}$$

考虑到 ℓ_∞ 范数的次梯度为 $\partial\|x\|_\infty = \{v : \langle v, x\rangle = \|x\|_\infty, \|v\|_1 \le 1\}$，子问题等价于，

$$i_k \in \underset{i=1,\dots,n}{\operatorname{argmax}} |\nabla_i f(y_{k-1})| \tag{4.15}$$

$$x_k = -t \cdot \mathbf{sign}\left[\nabla_{i_k} f(y_{k-1})\right] \cdot e_{i_k}. \tag{4.16}$$

其中 $\nabla_i f(y_{k-1})$ 表示向量 $\nabla f(y_{k-1})$ 的第 i 个元素，e_i 表示第 i 个元素为 1 的单位向量。可以看到计算 $\|\cdot\|_\infty$ 的次梯度和计算集合 $X := \{x \in \mathbb{R}^n : \|x\|_1 \le t\}$ 上的投影都需要 $\mathcal{O}(n)$ 的计算复杂度，但是条件梯度法子问题计算 (4.15) 和 (4.16) 明显要更简单直接。

例 4.3 ℓ_p **范数约束问题**

考虑带 ℓ_p 范数约束的凸优化问题，其中 $1 \le p \le \infty$

$$\min_x \ f(x) \quad \text{s.t.} \quad \|x\|_p \le t. \tag{4.17}$$

由于 ℓ_p 范数的对偶范数是 ℓ_q 范数，其中 $1/p + 1/q = 1$，因此用条件梯度法求解该问题时子问题为，

$$x_k \in -t \cdot \partial\|\nabla f(y_{k-1})\|_q. \tag{4.18}$$

注意到 ℓ_q 范数的次梯度为 $\partial\|x\|_q = \{v : \langle v, x\rangle = \|x\|_q, \|v\|_p \le 1\}$，子问题等价于，

$$x_k^{(i)} = -\beta \cdot \mathbf{sign}\left[\nabla_i f(y_{k-1})\right] \cdot |\nabla_i f(y_{k-1})|^{p/q}. \tag{4.19}$$

其中 β 是使得 $\|x_k\|_q = t$ 的归一化常数。可以看到，除过 $p = 1, 2, \infty$ 这些特殊情形，条件梯度法的子问题计算复杂度比直接计算点在集合 $X = \{x \in \mathbb{R}^n : \|x\|_p \le t\}$ 上的投影要简单，后者投影计算需要单独解一个优化问题。

例 4.4 核范数约束问题

考虑带核范数约束的矩阵优化问题，

$$\min_X \ f(X) \quad \text{s.t.} \quad \|X\|_* \le t. \tag{4.20}$$

其中 $X \in \mathbb{R}^{m\times n}$，矩阵核范数 $\|\cdot\|_*$ 的对偶范数是其谱范数 $\|\cdot\|_2$，它们的定义如下，

$$\|X\|_* = \sum_{i=1}^{\min\{m,n\}} \sigma_i(X), \qquad \|X\|_2 = \max_{i=1,\dots,\min\{m,n\}} \sigma_i(X). \tag{4.21}$$

因此条件梯度法的子问题为，

$$X_k \in -t \cdot \partial\|\nabla f(Y_{k-1})\|_2. \tag{4.22}$$

对矩阵范数的次梯度: $\partial\|X\| = \{Y : \langle Y, X\rangle = \|X\|, \|Y\|_* \le 1\}$，设 u, v 分别是矩阵 $\nabla f(Y_{k-1})$ 最大奇异值对应的左、右奇异向量，注意到，

$$\langle uv^T, \nabla f(Y_{k-1})\rangle = u^T \nabla f(Y_{k-1}) v = \sigma_{\max}(\nabla f(Y_{k-1})) = \|\nabla f(Y_{k-1})\|_2. \tag{4.23}$$

且 $\|uv^T\|_* = 1$，因此矩阵 $uv^T \in \partial\|\nabla f(Y_{k-1})\|_2$。则条件梯度法子问题等价于，

$$X_k \in -t \cdot uv^T. \tag{4.24}$$

可以看到，条件梯度法计算子问题时只需要计算矩阵最大的奇异值对应的左、右奇异向量。如果采用投影梯度法，其子问题是计算 X 到集合 $\{X \in \mathbb{R}^{m\times n} : \|X\|_* \le t\}$ 的投影，需要对矩阵做全奇异值分解，计算量比条件梯度法复杂很多。

现在我们给出条件梯度法的收敛性和复杂度分析，首先我们按照复杂度分析的框架给出我们的问题模型 $\mathcal{F}$。我们考虑凸函数和强凸函数两种情形，对于强凸函数，条件梯度法无法在理论上建立线性收敛的性质，因此我们需要对条件梯度法进行改进，引入改进的 $\mathcal{LO}$ 子程序。首先，我们给出凸函数情形下的问题模型。

问题模型 4.1 *考虑带约束的凸优化问题集合 $\mathcal{F} = (\Sigma, \mathcal{O}, \mathcal{T}_\epsilon)$:*

- *全局信息 Σ: 目标函数 $f \in C_L^{1,1}(X)$，且 $f(x)$ 是凸函数（或强凸函数），约束集合 X 凸、闭有界，*
- *局部信息 $\mathcal{O}$: $\mathcal{FO}$ 子程序和 $\mathcal{LO}$ 子程序（或改进的 $\mathcal{LO}$ 子程序），*
- *解的精度 $\mathcal{T}_\epsilon$: 求全局极小点 $\bar{x} \in \mathbb{R}^n$，使得 $f(\bar{x}) - f^* \le \epsilon$。*

衡量条件梯度法的分析复杂度时，我们需要分别衡量 $\mathcal{FO}$ 和 $\mathcal{LO}$ 子程序的调用次数。对于凸优化问题，我们有如下收敛性质和分析复杂度。

定理 4.1 *设 $f(x)$ 是凸函数，条件梯度法 4.1 在求解问题模型 4.1 时取步长 (4.7) 或 (4.8)，则产生的序列 $\{x_k\}$ 对任意 $k = 1, 2, \ldots$ 满足，*

$$f(y_k) - f^* \ \le\ \frac{2L}{k(k+1)} \sum_{i=1}^{k} \|x_i - y_{i-1}\|^2 \ \le\ \frac{2L}{k+1} D_X^2. \tag{4.25}$$

且为达到 ϵ 精度需要调用 $\mathcal{FO}$ 和 $\mathcal{LO}$ 子程序的次数至多为，

$$\left\lceil \frac{2LD_X^2}{\epsilon} \right\rceil - 1. \tag{4.26}$$

证明: 令 $\gamma_k = \frac{2}{k+1}$，记 $\bar{y}_k = (1-\gamma_k)y_{k-1} + \gamma_k x_k$，则不管 α_k 按下列哪种方式选取，

$$\alpha_k = \frac{2}{k+1} \quad 或 \quad \alpha_k = \operatorname*{argmin}_{\alpha \in [0,1]} f((1-\alpha)y_{k-1} + \alpha x_k).$$

对 $y_k = (1-\alpha_k)y_{k-1} + \alpha_k x_k$，我们都有 $f(y_k) \le f(\bar{y}_k)$。注意到 $\bar{y}_k - y_{k-1} = \gamma_k(x_k - y_{k-1})$，由 $f(x) \in C_L^{1,1}(X)$，利用性质 2.6 有，

$$f(y_k) \le f(\bar{y}_k) \ \le\ f(y_{k-1}) + \langle \nabla f(y_{k-1}), \bar{y}_k - y_{k-1} \rangle + \frac{L}{2}\|\bar{y}_k - y_{k-1}\|^2 \tag{4.27}$$

$$\le\ (1-\gamma_k)[f(y_{k-1}) + \gamma_k[f(y_{k-1}) + \langle \nabla f(y_{k-1}), x - y_{k-1} \rangle] + \frac{L\gamma_k^2}{2}\|x_k - y_{k-1}\|^2 \tag{4.28}$$

$$\le\ (1-\gamma_k) f(y_{k-1}) + \gamma_k f(x) + \frac{L\gamma_k^2}{2}\|x_k - y_{k-1}\|^2, \qquad 对任意\ x \in X. \tag{4.29}$$

其中不等式 (4.28) 是因为 $x_k \in \min_{x \in X} \langle \nabla f(y_{k-1}), x \rangle$，由最优性条件我们可以得到对任意 $x \in X$ 有 $\langle x - x_k, \nabla f(y_{k-1}) \rangle \ge 0$。将不等式 (4.29) 稍做变换，对任意 $x \in X$，

$$f(y_k) - f(x) \le (1-\gamma_k)[f(y_{k-1}) - f(x)] + \frac{L}{2}\gamma_k^2 \|x_k - y_{k-1}\|^2. \tag{4.30}$$

由引理 2.1 我们有，

$$f(y_k) - f(x) \le \Gamma_k(1-\gamma_1)[f(y_0) - f(x)] + \frac{\Gamma_k L}{2} \sum_{i=1}^{k} \frac{\gamma_i^2}{\Gamma_i} \|x_i - y_{i-1}\|^2. \tag{4.31}$$

由 $\gamma_k = \frac{2}{k+1}$，$\gamma_1 = 1$ 得到 $\Gamma_k = \frac{2}{k(k+1)}$，我们可以得到收敛性不等式，

$$f(y_k) - f^* \ \leq \ \frac{2L}{k(k+1)} \sum_{i=1}^{k} \|x_i - y_{i-1}\|^2 \ \leq \ \frac{2L}{k+1} D_X^2. \tag{4.32}$$

令 $\frac{2L}{k+1} D_X^2 \leq \epsilon$，可以得到分析复杂度结论。 □

通过定理 4.1 我们可以看到：首先，条件梯度法寻找 ϵ 近似解所需要调用子程序 $\mathcal{LO}$ 和 $\mathcal{FO}$ 的次数相同，都不超过

$$\mathcal{O}\left(\frac{LD_X^2}{\epsilon}\right). \tag{4.33}$$

在下一节我们可以看到，在函数集合 $C_L^{1,1}(X)$ 条件下，条件梯度法关于子程序 $\mathcal{LO}$ 的调用次数 $\mathcal{O}(1/\epsilon)$ 达到了算法复杂度下界，因此是最优的。但 $\mathcal{FO}$ 的调用次数 $\mathcal{O}(1/\epsilon)$ 并没有达到 $C_L^{1,1}(X)$ 函数集合的复杂度下界 $\mathcal{O}(1/\sqrt{\epsilon})$。其次，注意到 Lipschitz 常数 L 和 D_X 都与设定的范数 $\|\cdot\|$ 有关，即，$L \equiv L_{\|\cdot\|}$，$D_X \equiv D_{X,\|\cdot\|}$。尽管条件梯度法的分析复杂度结论 (4.33) 与范数无关，但实际上，我们可以利用这一性质选择合适的范数，使得分析复杂度关于范数达到极小

$$\mathcal{O}(1) \inf_{\|\cdot\|} \left\{ \frac{L_{\|\cdot\|} D_{X,\|\cdot\|}}{\epsilon} \right\}. \tag{4.34}$$

例如当约束集合 X 是单纯形时，我们可以选取 $\|\cdot\| = \|\cdot\|_1$。最后，注意到 (4.25) 中收敛速度依赖于 $\|x_i - y_{i-1}\|^2$，但实际上其值并不一定随着 k 增加而趋于零。例如，设 X 是多面体，则由子问题是线性规划，知道其解 x_i 是 X 的顶点。考虑到 $\{y_i\} \to y^*$，其中 x^* 是最优解，除非 x^* 也是 X 多面体的定点，则当 k 足够大时 $\|x_i - y_{i-1}\|$ 并不会趋于零。

4.2.2 强凸优化问题

这一节我们考虑条件梯度法求解强凸优化问题的收敛性。为此，我们需要对 $\mathcal{LO}$ 子程序做一些改进，从而提高条件梯度法在求解强凸问题时的 $\mathcal{LO}$ 分析复杂度。更具体的，对改进的 $\mathcal{LO}$ 子程序输入某一给定的 $p \in \mathbb{R}^n$ 和 $x_0 \in X$ ，能够返回下列优化问题的最优解：

$$\min\{\langle p, x\rangle : \ x \in X, \|x - x_0\| \leq R\}. \tag{4.35}$$

改进的 $\mathcal{LO}$ 子程序的复杂度与范数的选择有关。实际上，我们可以选择特定的范数来降低(4.35) 优化问题的复杂度。例如，当 X 是多面体的时候，我们可以取 $\|\cdot\| = \|\cdot\|_\infty$ 或 $\|\cdot\| = \|\cdot\|_1$，则改进的 $\mathcal{LO}$ 子程序与原 $\mathcal{LO}$ 的复杂度相同，都是求解一个线性规划问题。但是 $\|\cdot\|$ 的选取会改变 L 和 μ 的值。在改进的 $\mathcal{LO}$ 基础上文献 [26] 给出了求解强凸优化问题的收缩条件梯度法（The Shrinking Conditional Gradient (CndG) Method）。

算法 4.2 收缩条件梯度法 (S-CndG)

输入：$p_0 \in X$，令 $R_0 = D_X$.

对 $t = 1, \ldots$ 执行迭代：

令 $y_0 = p_{t-1}$.

对 $k = 1, \ldots, 8L/\mu$ 执行迭代：

1. 调用改进的 $\mathcal{LO}$ 子程序计算 $x_k \in \operatorname{argmin}_{x \in X_{t-1}} \langle \nabla f(y_{k-1}), x \rangle$，
   其中 $X_{t-1} := \{x \in X : \|x - p_{t-1}\| \le R_{t-1}\}$.
2. 令 $y_k = (1 - \alpha_k) y_{k-1} + \alpha_k x_k$ 其中 $\alpha_k \in [0, 1]$.

令 $p_t = y_k$ 和 $R_t = R_{t-1}/\sqrt{2}$.

定理 4.2 设 $f(x) \in \mathcal{F}^{1,1}_{L,\mu}$，收缩条件梯度法 *4.2* 在求解问题模型 *4.1* 时取步长 (4.7) 或 (4.8)，产生的序列 $\{p_t\}$ 对任意 $t = 1, 2, \ldots$ 满足

$$f(p_t) - f^* \le \frac{\mu R_0}{2^t}, \tag{4.36}$$

且为达到 ϵ 精度需要调用 $\mathcal{FO}$ 和改进的 $\mathcal{LO}$ 子程序的次数至多是

$$\frac{8L}{\mu} \left\lceil \max\left(\log_2 \frac{\mu R_0}{\epsilon}, 1 \right) \right\rceil. \tag{4.37}$$

证明：令 $K \equiv 8L/\mu$，我们首先证明对任意的 $t \ge 0$ 有 $x^* \in X_t$。我们使用归纳法，当 $t = 0$ 时 $\|x^* - p_0\| \le R_0 = D_X$，因此 $x^* \in X_0$。假设对任意 $t \ge 1$，$x^* \in X_{t-1}$ 成立，在定理 4.1 的证明过程中令 $X = X_t$，可知，第 t 次外迭代，对 $k = 1, ..., K$ 有

$$f(y_k) - f^* \le \frac{2L}{k(k+1)} \sum_{i=1}^{k} \|x_i - y_{i-1}\|^2 \le \frac{2L}{k+1} R_{t-1}^2, \quad k = 1, ..., K. \tag{4.38}$$

当 $k = K$ 时，$p_t = y_K$，由强凸函数的性质 $f(y_K) - f^* \ge \frac{\mu}{2} \|y_K - x^*\|^2$，我们有

$$\|p_t - x^*\|^2 \le \frac{2}{\mu}[f(p_t) - f^*] \le \frac{4L}{\mu(K+1)} R_{t-1}^2 \le \frac{1}{2} R_{t-1}^2 = R_t^2. \tag{4.39}$$

因此我们证明了 $x^* \in X_t$ 对任意 $t \ge 0$ 成立。我们现在可以给出 ϵ 近似解所需要的 $\mathcal{LO}$ 和 $\mathcal{FO}$ 上界。注意到 (4.39) 以及 R_t 的定义，我们有

$$f(p_t) - f^* \le \frac{\mu}{2} R_t^2 = \frac{\mu R_0}{2^t}. \tag{4.40}$$

因此为得到 ϵ 近似解所需要的外迭代次数不超过 $\lceil \max\{\log_2(\mu R_0/\epsilon), 1\} \rceil$，考虑到内迭代次数恒定为 K 可以证明总改进的 $\mathcal{LO}$ 子程序调用次数不超过 (4.37) 。 □

4.2.3 加速框架下的条件梯度法

这一节，我们在加速梯度法框架中引入条件梯度法的 $\mathcal{LO}$ 子程序，具体的，在加速梯度法中，将非线性逼近子问题修改为线性逼近的子问题，从而可以得到不同的收敛性结果。加速框架下的条件梯度法（The Primal Averaging Conditional Gradient (PA-CndG) Method ）由 [26] 提出，具体算法如下。

算法 4.3 加速框架下的条件梯度法 (PA-CndG)

输入： $x_0 \in X$，步长序列 $\{\alpha_k\}$，$\alpha_k \in [0,1]$，令 $y_0 = x_0$。

for $k = 1, \ldots$ **do**

1. $z_k = \frac{k-1}{k+1} y_{k-1} + \frac{2}{k+1} x_{k-1}$，
2. 令 $p_k = \nabla f(z_k)$ 调用 $\mathcal{LO}$ 子程序计算 $x_k \in \operatorname{argmin}_{x \in X} \langle p_k, x \rangle$，
3. $y_k = (1-\alpha_k) y_{k-1} + \alpha_k x_k$。

end for

定理 4.3 *条件梯度法 (PA-CndG) 4.3 在求解问题模型 4.1 时取步长 (4.7) 或 (4.8)，则产生的序列 $\{x_k\}$ 和 $\{y_k\}$ 对任意 $k = 1, 2, \ldots$ 满足，*

$$f(y_k) - f^* \ \le\ \frac{2L}{k(k+1)} \sum_{i=1}^{k} \|x_i - x_{i-1}\|^2 \ \le\ \frac{2L}{k+1} D_X^2. \tag{4.41}$$

证明： 令 $\gamma_k = \frac{2}{k+1}$，记 $\bar{y}_k = (1-\gamma_k) y_{k-1} + \gamma_k x_k$，则不管 α_k 按 (4.7) 或 (4.8) 哪种方式选取，对 $y_k = (1-\alpha_k) y_{k-1} + \alpha_k x_k$，我们都有 $f(y_k) \le f(\bar{y}_k)$。由算法 4.3 的定义，$z_{k-1} = (1-\gamma_k) y_{k-1} + \gamma_k x_{k-1}$，因此 $\bar{y}_k - z_{k-1} = \gamma_k (x_k - x_{k-1})$ 由性质 2.6 我们有，

$$\begin{aligned}
& f(y_k) \le f(\bar{y}_k) \ \le\ f(z_{k-1}) + \langle \nabla f(z_{k-1}), \bar{y}_k - z_{k-1} \rangle + \frac{L}{2} \|\bar{y}_k - z_{k-1}\|^2 && (4.42)\\
=\ & (1-\gamma_k)[f(z_{k-1}) + \langle \nabla f(z_{k-1}), y_{k-1} - z_{k-1} \rangle] \\
& + \gamma_k [f(z_{k-1}) + \langle \nabla f(z_{k-1}), x_k - z_{k-1} \rangle] + \frac{L\gamma_k^2}{2} \|x_k - x_{k-1}\|^2 && (4.43)\\
\le\ & (1-\gamma_k) f(y_{k-1}) + \gamma_k [f(z_{k-1}) + \langle \nabla f(z_{k-1}), x - z_{k-1} \rangle] + \frac{L\gamma_k^2}{2} \|x_k - x_{k-1}\|^2 && (4.44)\\
\le\ & (1-\gamma_k) f(y_{k-1}) + \gamma_k f(x) + \frac{L\gamma_k^2}{2} \|x_k - x_{k-1}\|^2, \qquad \text{对任意 } x \in X. && (4.45)
\end{aligned}$$

其中不等式 (4.44) 是利用了 f 的凸性以及 $x_k \in \min_{x \in X} \langle \nabla f(z_{k-1}), x \rangle$ 的最优性条件。由最优性条件，对任意 $x \in X$ 有不等式 $\langle x - x_k, \nabla f(z_{k-1}) \rangle \ge 0$ 成立。将不等式 (4.45) 两边同时减去 $f(x)$，对任意 $x \in X$ 我们有

$$f(y_k) - f(x) \le (1-\gamma_k)[f(y_{k-1}) - f(x)] + \frac{L}{2} \gamma_k^2 \|x_k - x_{k-1}\|^2. \tag{4.46}$$

由引理 2.1 我们有，

$$f(y_k) - f(x) \le \Gamma_k (1-\gamma_1)[f(y_0) - f(x)] + \frac{\Gamma_k L}{2} \sum_{i=1}^{k} \frac{\gamma_i^2}{\Gamma_i} \|x_i - x_{i-1}\|^2. \tag{4.47}$$

由 $\gamma_k = \frac{2}{k+1}$，$\gamma_1 = 1$ 得到 $\Gamma_k = \frac{2}{k(k+1)}$，于是对任意 $x \in X$

$$f(y_k) - f^* \ \le\ \frac{2L}{k(k+1)} \sum_{i=1}^{k} \|x_i - x_{i-1}\|^2 \ \le\ \frac{2L}{k+1} D_X^2. \tag{4.48}$$

上述收敛性不等式成立。 □

比较条件梯度法的收敛性结果 (4.25) 和加速框架下的条件梯度法收敛性结果 (4.41)，我们可以发现：对于条件梯度法来说，收敛性速度与 $\|x_k - y_{k-1}\|$ 的变化有关，而 $\|x_k - y_{k-1}\|$ 不一定随着 k 增加收敛到零；对于加速框架下的条件梯度法来说，收敛速度与 $\|x_k - x_{k-1}\|$ 有关，而 $\|x_k - x_{k-1}\|$ 是否随 k 增加

收敛到零依赖于约束集合 X 的结构以及 $\|p_{k+1}-p_k\|$。令 $\gamma_k=\frac{2}{k+1}$，注意到 $z_k=(1-\gamma_k)y_{k-1}+\gamma_k x_{k-1}$，$y_k=(1-\alpha_k)y_{k-1}+\alpha_k x_k$，因此

$$z_{k+1}-z_k = (y_k-y_{k-1})+\gamma_{k+1}(x_k-y_{k-1})+\gamma_k(x_{k-1}-y_{k-1}) \tag{4.49}$$

$$= \alpha_k(x_k-y_{k-1})+\gamma_{k+1}(x_k-y_{k-1})+\gamma_k(x_{k-1}-y_{k-1}). \tag{4.50}$$

令 $\alpha_k=\gamma_k$，我们有 $\|z_{k+1}-z_k\|\le 3\gamma_k D_X$，注意到 $f(x)\in C_L^{1,1}(X)$，

$$\|p_{k+1}-p_k\|_* = \|\nabla f(z_{k+1})-\nabla f(z_k)\|_* \le 3\gamma_k L D_X. \tag{4.51}$$

因此 p_{k+1} 和 p_k 的差随着 k 增加趋向零（因为 γ_k 随着 k 增加单调递减到零）。我们可以加入一些关于 $\mathcal{LO}$ 更强的假设（在某些局部下满足）来提高 PA-CndG 的收敛性。例如，

$$\|x_k-x_{k-1}\|\le Q\|p_k-p_{k-1}\|_*^\rho, \qquad \text{对某一 } k\ge N. \tag{4.52}$$

在文献 [26] 给出了类似 (4.52) 条件下加速框架下条件梯度法的收敛速度。可以看到，在此类条件下，收敛速度有了明显的提高。

定理 4.4 *加速框架下的条件梯度法 PA-CndG 4.3 取步长 $\alpha_k=\frac{2}{k+1}$，假设子程序 $\mathcal{LO}$ 满足*

$$\|x_k-x_{k-1}\|\le Q\|p_k-p_{k-1}\|_*^\rho, \qquad k\ge 2. \tag{4.53}$$

对某些给定的 $\rho\in(0,1]$ 和 $Q>0$，对任意的 $k\ge 1$ 我们有如下收敛性结果

$$f(y_k)-f^* \le \mathcal{O}(1)\begin{cases}\frac{Q^2L^{2\rho+1}D_X^{2\rho}}{(1-2\rho)k^{2\rho+1}} & \rho\in(0,0.5),\\ \frac{Q^2L^2D_X\log(k+1)}{k^2} & \rho=0.5,\\ \frac{Q^2L^{2\rho+1}D_X^{2\rho}}{(2\rho-1)k^2} & \rho\in(0.5,1].\end{cases}$$

.

4.3 复杂度下界

这一节我们研究问题模型 4.1 关于 $\mathcal{LO}$ 子程序的分析复杂度下界。首先我们定义带 $\mathcal{LO}$ 子程序的解算法 $\mathcal{S}$，考虑其最一般的框架形式。

算法 4.4 条件梯度法的一般框架

输入：$x_0\in X$

对 $k=1,\dots,N$ 执行迭代：

1. 确定线性逼近函数 $\langle p_k,\cdot\rangle$，
2. 调用 $\mathcal{LO}$ 子程序计算 $x_k\in\operatorname{argmin}_{x\in X}\langle p_k,x\rangle$，
3. 输出 $y_k\in\operatorname{conv}\{x_0,\dots,x_k\}$。

注意到上述算法框架的变化形式很多。首先，条件梯度法 4.1 和加速框架下的条件梯度法 4.3 都是其特殊情形。其次，对于线性逼近函数 $\langle p_k,\cdot\rangle$ 没有任何限制。例如，当 f 光滑时，p_k 可以是某个可行点处的梯度，或者是历史梯度的一些线性组合。当 f 非光滑时，p_k 可以是 f 光滑逼近函数的梯度。我们也

可以考虑 p_k 包含噪音，或者一些二阶梯度信息。最后，输出的解 y_k 是点集合 $\{x_0, ..., x_k\}$ 的凸组合，因此输出的解可以有多种多样的组合方式，与 $\{x_k\}$ 中任何一个点都不同。

将上述算法框架与一阶算法进行比较。可以发现，一方面，条件梯度法的一般框架解决线性逼近的子问题，而一阶算法需要求解非线性子问题（投影或近似点算子的计算）。另一方面，条件梯度法的一般框架对于算法的迭代方向 p_k 和输出的点 y_k 比一阶算法更为灵活。

为了研究优化问题关于 $\mathcal{LO}$ 子程序分析复杂度下界，按照复杂度分析理论，我们需要定义问题模型 $\mathcal{F}$、解算法 $\mathcal{S}$、以及解算法集合 $\mathcal{M}$。对于问题模型 $\mathcal{F}$ 我们考虑其中全局信息为 $f \in C^{1,1}_{L,\|\cdot\|}(X)$ 且是凸函数的情形。具体的解算法 $\mathcal{S}$ 定义为算法 4.4 框架里可以将问题模型 $\mathcal{F}$ 求解到 ϵ 精度的算法。解算法集合 $\mathcal{M}$ 定义为所有这样形式的解算法 $\mathcal{S}$ 组成的集合。因此我们需要估计

$$\mathrm{Compl}(\epsilon) = \inf_{\mathcal{S}\in\mathcal{M}} \mathrm{Compl}_{\mathcal{S}}(\epsilon). \tag{4.54}$$

其中 $\mathrm{Compl}_{\mathcal{S}}(\epsilon)$ 是解算法 $\mathcal{S}$ 求解问题集合 $\mathcal{F}$ 时 $\mathcal{LO}$ 分析复杂度的上界。我们没法直接计算出 $\mathrm{Compl}(\epsilon)$，只能给出 $\mathrm{Compl}(\epsilon)$ 的一个下界估计。但按照复杂度分析理论 [[36] Nemirovski 、& Nesterov (1983)] 和 [[40] Nesterov(2004)] 我们需要构造 $\mathcal{F}$ 中的一个函数，对该函数调用子程序 $\mathcal{LO}$ 所获得的局部信息很少，导致解算法集合 $\mathcal{M}$ 中的所有算法在求解它时的效率都不高。[26] 给出了如下复杂度下界结果。

定理 4.5（光滑优化问题的 $\mathcal{LO}$ 复杂度下界） *为了达到 $\epsilon > 0$ 近似解，解算法集合 $\mathcal{M}$ 中的任意解算法 $\mathcal{S}$ 在求解全局信息为凸函数和 $C^{1,1}_{L,\|\cdot\|}(X)$ 的问题模型 $\mathcal{F}$ 时所调用的 $\mathcal{LO}$ 子程序次数 $\mathrm{Compl}_{\mathcal{S}}(\epsilon)$ 满足*

$$\mathrm{Compl}_{\mathcal{S}}(\epsilon) \geq \left\lfloor \min\left\{\frac{n}{4}, \frac{LD_X^2}{8\epsilon}\right\}\right\rfloor - 1. \tag{4.55}$$

对任意 $\mathcal{S} \in \mathcal{M}$ 成立。

证明: 考虑凸优化问题

$$f^* = \min_{x\in X}\left\{f(x) := \frac{L}{2}\sum_{i=1}^{n}(x^{(i)})^2\right\}. \tag{4.56}$$

其中 $x^{(i)}$ 是 x 的第 i 个元素，$X := \{x \in \mathbb{R}^n : \sum_{i=1}^n x^{(i)} = D,\ x^{(i)} > 0\}$，$D > 0$，容易看到 (4.56) 的最优解 x^* 和最优值 f^* 分别为

$$x^* = \{\frac{D}{n}, ..., \frac{D}{n}\}, \qquad f^* = \frac{LD^2}{n}, \tag{4.57}$$

且函数 $f(x) \in C^{1,1}_{L,\|\cdot\|}(X)$ 其中 $\|\cdot\|$ 取 $\|\cdot\|_2$。不失一般性，我们可以假设算法的初始点是 $x_0 = De_1$ 其中 $e_1 = (1, 0, ..., 0)$ 是单位向量。如果选择随机初始点 $x_0 \in X$，我们可以构造相似问题

$$\min_x \quad (x^{(1)})^2 + \sum_{i=2}^{n}(x^{(i)} - x_0^{(i)})^2. \tag{4.58}$$

$$\text{s.t.} \quad x^{(1)} + \sum_{i=2}^{n}(x^{(i)} - x_0^{(i)}), \tag{4.59}$$

$$x^{(1)} \geq 0, \tag{4.60}$$

$$x^{(i)} - x_0^{(i)} \geq 0, \quad i = 2, ..., n. \tag{4.61}$$

然后对该问题进行分析。

现在计算法集合 $\mathcal{M}$ 中的解算法 $\mathcal{S}$ 求解 (4.56)。在第 k 次迭代，算法调用 $\mathcal{LO}$ 子程序，对任意的输入 p_k 输出新的迭代点 $x_k \in \mathrm{Argmin}_{x\in X}\langle p_k, x\rangle$。由于 X 是单纯形，则 $x_k \in \{De_1, De_2, ..., De_n\}$，$e_i$，

$i=1,...,n$ 是 $\mathbb{R}^n$ 第 i 个单位向量。注意到 x_k 会有重复相同的情形，我们记 $x_k = De_{p_k}$，其中 $1 \le p_k \le n$。由 4.4 算法定义我们有 $y_k \in D\ \mathrm{Conv}\{x_0, x_1, ..., x_k\}$。因此

$$y_k \in D\ \mathrm{Conv}\{e_1, e_{p_1}, ..., e_{p_k}\}. \tag{4.62}$$

假设 $\{e_1, e_{p_1}, ..., e_{p_k}\}$ 中 q 个向量是线性独立的，其中 $1 \le q \le k+1 \le n$。不失一般性我们记 $\{e_1, e_{p_1}, ..., e_{p_{q-1}}\}$ 线性独立。因此

$$\begin{aligned} f(y_k) &\ge \min_x\{f(x) : x \in D \cdot \mathrm{conv}\{e_1, e_{p_1}, ..., e_{p_k}\}\} && (4.63)\\ &= \min_x\{f(x) : x \in D \cdot \mathrm{conv}\{e_1, e_{p_1}, ..., e_{p_{q-1}}\}\} && (4.64)\\ &= \frac{LD^2}{2q} \ge \frac{LD^2}{2(k+1)}. && (4.65) \end{aligned}$$

上述不等式与 (4.57) 结合，对任意 $k = 1, ..., n-1$ 我们有

$$f(y_k) - f^* \ge \frac{LD^2}{2(k+1)} - \frac{LD^2}{n} \tag{4.66}$$

成立。注意到 $D_X = \sqrt{2D}$，定义

$$K := \left\lfloor \min\left\{\frac{n}{4}, \frac{LD_X^2}{8\epsilon}\right\}\right\rfloor - 1. \tag{4.67}$$

由 (4.66) 对任意的 $1 \le k \le K$ 我们有

$$\begin{aligned} f(y_k) - f^* &\ge \frac{LD^2}{2(K+1)} - \frac{LD^2}{n} && (4.68)\\ &\ge \frac{LD^2}{\min\left\{\frac{n}{2}, \frac{LD^2}{2\epsilon}\right\}} - \frac{LD^2}{n} && (4.69)\\ &= \frac{LD^2}{\min\left\{n, \frac{LD^2}{\epsilon}\right\}} + \left(\frac{LD^2}{\min\left\{n, \frac{LD^2}{\epsilon}\right\}} - \frac{LD^2}{n}\right) && (4.70)\\ &\ge \frac{LD^2}{\frac{LD^2}{\epsilon}} + \left(\frac{LD^2}{n} - \frac{LD^2}{n}\right). && (4.71) \end{aligned}$$

□

由上述定理可以看出，当 n 足够大时，光滑优化问题的 $\mathcal{LO}$ 复杂度下界为

$$\mathcal{O}\left(\frac{LD_X^2}{\epsilon}\right). \tag{4.72}$$

4.4 加速条件梯度法

利用加速梯度法框架，我们同样可以得到加速的条件梯度法。这里的加速，指的是 $\mathcal{FO}$ 的调用次数减少，算法效率的提高。从上一节我们知道条件梯度法已经在 $\mathcal{LO}$ 的调用次数上达到了最优，但对于 $\mathcal{FO}$ 的调用次数并没有达到最优。因此，我们可以利用加速梯度法的框架来减少对 $\mathcal{FO}$ 的调用次数。文献 [28] 给出了加速条件梯度法的框架，又称为条件梯度滑动算法（Conditional gradient sliding method ），简称为 CGS。

算法 4.5 加速条件梯度法

输入: $x_0 \in X$ 和总迭代次数 N，取 $\beta_k > 0,\ \eta_k \geq 0,\ \gamma_k \in [0,1]$，令 $y_0 = x_0$

对 $k = 1, \ldots, N$ **执行**

$$z_k = (1-\gamma_k)y_{k-1} + \gamma_k x_{k-1}, \tag{4.73}$$

$$x_k = \mathrm{CndG}(f'(z_k), x_{k-1}, \beta_k, \eta_k), \tag{4.74}$$

$$y_k = (1-\gamma_k)y_{k-1} + \gamma_k x_k. \tag{4.75}$$

输出: y_N.

加速条件梯度法也有三个迭代序列 $\{x_k\}$、$\{y_k\}$ 和 $\{z_k\}$，以及三个待定的参数序列 $\{\gamma_k\}$、$\{\beta_k\}$ 和 $\{\eta_k\}$。与 Nesterov 加速梯度法框架 2.3 不同之处在于算法 4.5 第二步中 x_k 的更新方式。加速梯度法通过直接精确求解非线性优化问题得到 x_k，而加速条件梯度法实际上是通过非精确求解非线性优化问题来更新 x_k，具体来说，加速条件梯度法是用条件梯度法来求解非线性优化的子问题。其中更新 x_k 的 **子程序 CndG** 如下。

子程序 CndG: $u^+ = \mathrm{CndG}(g, u, \beta, \eta)$

1: 令 $u_1 = u$ 和 $t = 1$

2: 调用 $\mathcal{LO}$ 求解下面子问题，令 v_t 表示其最优解

$$V_{g,u,\beta}(u_t) = \max_{x \in X} \langle g + \beta(u_t - u_1), u_t - x \rangle. \tag{4.76}$$

3: 若 $V_{g,u,\beta}(u_t) \leq \eta$，令 $u^+ = u_t$ 并且停机.

4: 令 $u_{t+1} = (1-\alpha_t)u_t + \alpha_t v_t$, 其中

$$\alpha_t = \min\left\{1, \frac{\langle \beta(u-u_t) - g, v_t - u_t \rangle}{\beta \|v_t - u_t\|^2}\right\}. \tag{4.77}$$

5: 取 $t \leftarrow t+1$ 转步 2.

可以看到 **子程序 CndG** 的收敛点列是 $\{u_t\}$。记 $\phi(x) := \langle g, x \rangle + \frac{\beta}{2}\|x - u\|^2$，**子程序 CndG** 可以被看做是利用条件梯度法求解 $\min_{x \in X} \phi(x)$。注意到 $\nabla\phi(x) = g + \beta\langle x - u \rangle$，因此 4.76 等价于

$$v_t \in \operatorname*{argmax}_{x \in X} \langle \nabla\phi(u_t), u_t - x \rangle = \operatorname*{argmin}_{x \in X} \langle \nabla\phi(u_t), x \rangle. \tag{4.78}$$

因此我们可以写出**子程序 CndG** 更新 x_k 的等价形式。求解优化问题

$$\min_{x \in X}\left\{\phi(x) := \langle g, x \rangle + \frac{\beta_k}{2}\|x - u_1\|^2\right\}. \tag{4.79}$$

其中 $g = \nabla f(z_k)$，$u_1 = x_{k-1}$，对 $t = 1, 2, \ldots$ 执行迭代

1. $v_t = \operatorname{argmin}_{x \in X} \langle \nabla\phi(u_t), x \rangle$,
2. $u_{t+1} = (1-\alpha_t)u_t + \alpha_t v_t$,
3. 当 $V_{g,u_1,\beta}(u_t) := \max_{x \in X} \langle \nabla\phi(u_t), u_t - x \rangle \leq \eta$ 时停止迭代。

其中 $V_{g,u,\beta}$ 又被称为 Wolfe 间隔，表示约束优化问题最优性条件的满足程度。步长 α_t 的选取实际上等价于下面的优化问题。

$$\alpha_t = \operatorname*{argmin}_{\alpha \in [0,1]} \phi((1-\alpha)u_t + \alpha v_t). \tag{4.80}$$

可以看出，加速条件梯度法的迭代点 x_k 是非线性投影子问题

$$\min_{x\in X}\left\{\phi_k(x) := \langle \nabla f(z_k), x\rangle + \frac{\beta_k}{2}\|x - x_{k-1}\|^2\right\}. \tag{4.81}$$

满足约束条件

$$\langle \nabla \phi_k(x_k), x_k - x\rangle := \langle \nabla f(z_k) + \beta_k\langle x_k - x_{k-1}, x_k - x\rangle \le \eta_k, \quad \forall x \in X. \tag{4.82}$$

的近似解。

为了证明加速条件梯度法的收敛性和复杂度，我们采取分步的策略。第一步，确定**子程序 CndG** 的收敛性和复杂度；第二步，确定外迭代的收敛性和复杂度。记 ϕ_k^* 表示第 k 步所近似求解的非线性投影问题 (4.81) 的精确解。我们首先考虑内迭代**子程序 CndG** 的收敛性和复杂度。

1. 子程序 CndG 的复杂度:

首先，我们考虑子程序求解的优化问题 (4.81) $\min_{x\in X}\phi_k(x)$ 的收敛性，即 $\phi_k(u_t) - \phi_k^*$ 的收敛性和复杂度。为了简化记号，我们在分析中省略外迭代下标 k，即用 $\phi(x)$ 代替 $\phi_k(x)$，用 β、η 分别代替 β_k、η_k。利用引理 2.1 的分析方法，构造序列 $\{\lambda_t\}$ 和 $\{\Gamma_t\}$，满足

$$\Lambda_{t+1} = \Lambda_t(1 - \lambda_{t+1}),\ \forall t \ge 2. \tag{4.83}$$

我们可以取 $\lambda_t := 2/t$，于是 $\Lambda_t = 2/t(t-1)$。令 $\bar{u}_{t+1} = (1-\lambda_{t+1})u_t + \lambda_{t+1}v_t$ 可以推出 $\bar{u}_{t+1} - u_t = \lambda_{t+1}(v_t - u_t)$。注意到 $u_{t+1} = (1-\alpha_t)u_t + \alpha_t v_t$ 和 $\alpha_t = \operatorname{argmin}_{\alpha\in[0,1]}\phi((1-\alpha)u_t + \alpha v_t)$ 可以推出 $\phi(u_{t+1}) \le \phi(\bar{u}_{t+1})$。由 $\phi(x) \in C_L^{1,1}(X)$ 利用引理 2.1 我们有，

$$\phi(u_{t+1}) \le \phi(\bar{u}_{t+1}) \le \phi(u_t) + \langle \phi'(u_t), \bar{u}_{t+1} - u_t\rangle + \frac{\beta}{2}\|\bar{u}_{t+1} - u_t\|^2 \tag{4.84}$$

$$= \phi(u_t) + \lambda_{t+1}\langle \phi'(u_t), v_t - u_t\rangle + \frac{\beta\lambda_{t+1}^2}{2}\|v_t - u_t\|^2 \tag{4.85}$$

$$\le (1-\lambda_{t+1})\phi(u_t) + \lambda_{t+1}[\phi(u_t) + \langle \phi'(u_t), v_t - u_t\rangle] + \frac{\beta\lambda_{t+1}^2}{2}\|v_t - u_t\|^2 \tag{4.86}$$

$$\le (1-\lambda_{t+1})\phi(u_t) + \lambda_{t+1}\phi(x) + \frac{\beta\lambda_{t+1}^2}{2}\|v_t - u_t\|^2. \tag{4.87}$$

上面不等式两边同时减去 $\phi(x)$ 可以得到 $\phi_k(u_{t+1}) - \phi_k^*$ 误差的估计不等式:

$$\phi(u_{t+1}) - \phi(x) \le (1-\lambda_{t+1})[\phi(u_t) - \phi(x)] + + \frac{\beta\lambda_{t+1}^2}{2}\|v_t - u_t\|^2, \quad \forall x \in X. \tag{4.88}$$

将上式从 1 到 t 求和得到，

$$\phi(u_{t+1}) - \phi(x) \le \Lambda_{t+1}(1-\lambda_2)[\phi(u_1) - \phi(x)] + \Lambda_{t+1}\beta\sum_{j=1}^{t}\frac{\lambda_{j+1}^2}{2\Lambda_{j+1}}\|v_j - u_j\|^2 \le \frac{2\beta D_X^2}{t+1}. \tag{4.89}$$

取 $x = x^*$ 我们得到 $\phi_k(u_t) - \phi_k^*$ 的收敛速度估计:

$$\phi(u_{t+1}) - \phi^* \le \frac{2\beta D_X^2}{t+1}. \tag{4.90}$$

有了收敛速度,我们可以估计子程序的复杂度。注意到**子程序 CndG** 的复杂度的停止条件 $V_{g,u_1,\beta}(u_t) \le \eta$，因此我们需要建立 $V_{g,u_1,\beta}(u_t)$ 与 $\phi(u_t) - \phi^*$ 之间的关系。记 $\Delta_t := \phi(u_t) - \phi^*$ 以及 $V(u_t) := V_{g,u_1,\beta}(u_t) = \langle \nabla\phi(u_t), u_t - v_t\rangle$ 并代入不等式 (4.85) 得到

$$\lambda_{t+1}V(u_t) \le \phi(u_t) - \phi(u_{t+1}) + \frac{\beta\lambda_{t+1}^2}{2}\|v_t - u_t\|^2 \tag{4.91}$$

$$= \Delta_t - \Delta_{t+1} + \frac{\beta\lambda_{t+1}^2}{2}\|v_t - u_t\|^2. \tag{4.92}$$

将上式两边同除以 Λ_{t+1} 并从 $t=1$ 加到某一给定的 t，

$$\sum_{j=1}^{t}\frac{\lambda_{j+1}}{\Lambda_{j+1}}V(u_j) \leq \sum_{j=1}^{t}\left(\frac{\Delta_j}{\Lambda_{j+1}}-\frac{\Delta_{j+1}}{\Lambda_{j+1}}\right)+\sum_{j=1}^{t}\frac{\beta\lambda_{j+1}^2}{2\Lambda_{j+1}}\|v_j-u_j\|^2 \tag{4.93}$$

$$= -\frac{\Delta_{t+1}}{\Lambda_{t+1}}+\sum_{j=2}^{t}\left(\frac{1}{\Lambda_{j+1}}-\frac{1}{\Lambda_j}\right)\Delta_j+\frac{\Delta_1}{\Lambda_2}+\sum_{j=1}^{t}\frac{\beta\lambda_{j+1}^2}{2\Lambda_{j+1}}\|v_j-u_j\|^2 \tag{4.94}$$

$$\leq \sum_{j=1}^{t} j\Delta_j+t\beta D_X^2. \tag{4.95}$$

不等式 (4.95) 成立是因为 $\lambda_t:=2/t$ 和 $\Lambda_t=2/t(t-1)$，因此 $\Lambda_2=1$，$1/\Lambda_{j+1}-1/\Lambda_j=j$。由 (4.90) 我们知道 $\Delta_j=\phi(u_j)-\phi^*\leq\frac{2\beta D_X^2}{t+1}$，因此有

$$\sum_{j=1}^{t} j\Delta_j+t\beta D_X^2 \leq \frac{2\beta D_X^2}{t+1}\sum_{j=1}^{t} j+t\beta D_X^2 = 3t\beta D_X^2. \tag{4.96}$$

将不等式 (4.96) 代入 (4.95) 可以得到，

$$\min_{j=1,\ldots,t}V(u_j)\sum_{j=1}^{t}\frac{\lambda_{j+1}}{\Lambda_{j+1}} \leq \sum_{j=1}^{t}\frac{\lambda_{j+1}}{\Lambda_{j+1}}V(u_j)\leq 3t\beta D_X^2. \tag{4.97}$$

注意到

$$\sum_{j=1}^{t}\frac{\lambda_{j+1}}{\Lambda_{j+1}}=\frac{t(t+1)}{2}. \tag{4.98}$$

代入 (4.97) 得

$$\min_{j=1,\ldots,t}V(u_j) \leq \frac{6\beta D_X^2}{t+1}. \tag{4.99}$$

考虑在外迭代第 k 步时，令上面不等式右端小于给定 η_k，我们可以得到**子程序 CndG** 的复杂度 T_k 为

$$T_k = \left\lceil\frac{6\beta_k D_X^2}{\eta_k}\right\rceil. \tag{4.100}$$

2. 外迭代的复杂度:

对于外迭代，其框架和加速梯度法相同。由 (2.72) 可以得到，

$$f(y_k)\leq(1-\gamma_k)f(y_{k-1})+\gamma_k[f(z_k)+\langle\nabla f(z_k),x_k-z_k\rangle]+\frac{L\gamma_k^2}{2}\|x_k-x_{k-1}\|^2. \tag{4.101}$$

由于 x_k 是子问题 $\min_{x\in X}\left\{\langle\nabla f(z_k),x\rangle+\frac{\beta_k}{2}\|x-x_{k-1}\|_2^2\right\}$ 的 η_k 精度的近似解，即最优性条件满足，

$$\langle\nabla f(z_k)+\beta_k(x_k-x_{k-1}),x-x_k\rangle \geq -\eta_k, \quad \forall\, x\in X. \tag{4.102}$$

等价于

$$\langle x_{k-1}-x_k,x_k-x\rangle \leq \frac{1}{\beta_k}\langle\nabla f(x_k),x-x_k\rangle+\frac{\eta_k}{\beta_k}. \tag{4.103}$$

利用上述不等式，我们可以估计 x_k 与 x_{k-1} 的逼近程度，

$$\frac{1}{2}\|x_k-x_{k-1}\|^2 = \frac{1}{2}\|x_{k-1}-x\|^2-\langle x_{k-1}-x_k,x_k-x\rangle-\frac{1}{2}\|x_k-x\|^2 \tag{4.104}$$

$$\leq \frac{1}{2}\|x_{k-1}-x\|^2+\frac{1}{\beta_k}\langle\nabla f(z_k),x-x_k\rangle-\frac{1}{2}\|x_k-x\|^2+\frac{\eta_k}{\beta_k}. \tag{4.105}$$

上述不等式两边同乘以 $L\gamma_k^2$，并注意到 $L\gamma_k \le \beta_k$ 我们可以推出，

$$\begin{aligned}\frac{L\gamma_k^2}{2}\|x_k - x_{k-1}\|^2 \quad \le \quad & \frac{L\gamma_k^2}{2}\|x_{k-1} - x\|^2 + \gamma_k\langle \nabla f(z_k), x - x_k\rangle \\ & -\frac{L\gamma_k^2}{2}\|x_k - x\|^2 + \gamma_k\eta_k. \end{aligned} \tag{4.106}$$

将 (4.106) 代入 (4.101) 并利用 $f(x)$ 的凸性，我们可以得到

$$f(y_k) - f(x) \le (1-\gamma_k)[f(y_{k-1}) - f(x)] + \frac{\beta_k\gamma_k}{2}(\|x_{k-1} - x\|^2 - \|x_k - x\|^2) + \gamma_k\eta_k. \tag{4.107}$$

利用引理 2.1 我们得到收敛性估计的不等式，

$$\begin{aligned} f(y_k) - f(x) \quad \le \quad & \frac{\Gamma_k(1-\gamma_1)}{\Gamma_1}\left(f(y_0) - f(x)\right) \\ & + \frac{\Gamma_k}{2}\sum_{i=1}^{k}\frac{\beta_i\gamma_i}{\Gamma_i}\left(\|x_{i-1} - x\|^2 - \|x_i - x\|^2\right) + \Gamma_k\sum_{i=1}^{k}\frac{\gamma_i\eta_i}{\Gamma_i}. \end{aligned} \tag{4.108}$$

令 $D_X = \sup_{x,y\in X}\|x - y\|$，$D_0 = \|x_0 - x^*\|^2$，分析上面不等式，注意到，

$$\sum_{i=1}^{k}\frac{\beta_i\gamma_i}{\Gamma_i}\left(\|x_{i-1} - x\|^2 - \|x_i - x\|^2\right) \tag{4.109}$$

$$= \frac{\beta_1\gamma_1}{\Gamma_1}\|x_0 - x\|^2 + \sum_{i=2}^{k}\left(\frac{\beta_i\gamma_i}{\Gamma_i} - \frac{\beta_{i-1}\gamma_{i-1}}{\Gamma_{i-1}}\right)\|x_{i-1} - x\|^2 - \beta_k\gamma_k\Gamma_k\|x_k - x\|^2 \tag{4.110}$$

$$\le \frac{\beta_1\gamma_1}{\Gamma_1}\|x_0 - x\|^2 + \sum_{i=2}^{k}\left(\frac{\beta_i\gamma_i}{\Gamma_i} - \frac{\beta_{i-1}\gamma_{i-1}}{\Gamma_{i-1}}\right)\cdot D_X^2 \tag{4.111}$$

$$\le \frac{\beta_k\gamma_k}{\Gamma_k}D_X^2. \tag{4.112}$$

因此，对任意序列 $\{\beta_k\}$，$\{\gamma_k\}$，$\{\Gamma_k\}$ 满足 $L\gamma_k \le \beta_k$ 时有，

$$f(y_k) - f(x) \le \frac{\beta_k\gamma_k}{2}D_X^2 + \Gamma_k\sum_{i=1}^{k}\frac{\gamma_i\eta_i}{\Gamma_i}. \tag{4.113}$$

若序列 $\{\beta_k\}$，$\{\gamma_k\}$，$\{\Gamma_k\}$ 满足 $L\gamma_k \le \beta_k$ 且

$$\frac{\beta_k\gamma_k}{\Gamma_k} \ge \frac{\beta_{k-1}\gamma_{k-1}}{\Gamma_{k-1}}, \qquad \text{对任意} \quad k \ge 2. \tag{4.114}$$

我们有

$$f(y_k) - f(x) \le \Gamma_k\frac{\beta_1\gamma_1}{2}\|x_0 - x\|^2 + \Gamma_k\sum_{i=1}^{k}\frac{\gamma_i\eta_i}{\Gamma_i}. \tag{4.115}$$

综上，我们可以通过构造序列 $\{\beta_k\}$，$\{\gamma_k\}$，$\{\eta_k\}$ 来得到算法的收敛速率。

定理 4.6 *若 $f \in C_L^{1,1}(\mathbb{R}^n)$ 且是凸函数，则加速条件梯度法求解问题模型 2.4 的收敛速率为：*

1. *取 $\beta_k = \frac{3L}{k+1}$，$\gamma_k = \frac{3}{k+2}$，以及 $\eta_k = \frac{LD_X^2}{k(k+1)}$ 时，外迭代收敛速度为，*

$$f(y_k) - f(x^*) \le \frac{15LD_X^2}{2(k+1)(k+2)}. \tag{4.116}$$

若外迭代总次数（$\mathcal{FO}$ 的分析复杂度）为 N，则内迭代总次数（$\mathcal{LO}$ 的分析复杂度）上界为，

$$\sum_{k=1}^{N} T_k \leq 9N^2 + 10N. \tag{4.117}$$

加速条件梯度法为了得到 ϵ 近似解所需要的 $\mathcal{FO}$ 和 $\mathcal{LO}$ 复杂度上界分别为，

$$N(\mathcal{FO}) = \sqrt{\frac{15LD_X^2}{2\epsilon}}, \qquad N(\mathcal{LO}) = \frac{135LD_X^2}{2\epsilon} + 10\sqrt{\frac{15LD_X^2}{2\epsilon}}. \tag{4.118}$$

2. 令外迭代次总次数 N 固定，取 $\beta_k = \frac{2L}{k}$，$\gamma_k = \frac{2}{k+1}$，以及 $\eta_k = \frac{2LD_0^2}{Nk}$，则外迭代收敛速度为

$$f(y_N) - f^* \leq \frac{6LD_0^2}{N(N+1)}. \tag{4.119}$$

内迭代总次数（$\mathcal{LO}$ 的分析复杂度）为

$$\sum_{k=1}^{N} T_k \leq \frac{6N^2 D_X^2}{D_0^2} + N. \tag{4.120}$$

加速条件梯度法为了得到 ϵ 近似解所需要的 $\mathcal{FO}$ 和 $\mathcal{LO}$ 复杂度上界分别为，

$$N(\mathcal{FO}) = \mathcal{O}\left(D_0\sqrt{\frac{L}{\epsilon}}\right), \qquad N(\mathcal{LO}) = \mathcal{O}\left(\frac{LD_X^2}{\epsilon} + D_0\sqrt{\frac{L}{\epsilon}}\right). \tag{4.121}$$

证明: 对于定理的第一部分，当取 $\beta_k = \frac{3L}{k+1}$，$\gamma_k = \frac{3}{k+2}$ 和 $\eta_k = \frac{LD_X^2}{k(k+1)}$ 时满足 $L\gamma_k \leq \beta_k$ 根据，Γ_k 的定义我们有

$$\Gamma_k = \frac{6}{k(k+1)(k+2)}, \qquad \frac{\eta_i\gamma_i}{\Gamma_i} = 3LD_X^2. \tag{4.122}$$

因此由 (4.113) 有，

$$f(y_k) - f(x^*) \leq \frac{9LD_X^2}{2(k+1)(k+2)} + \frac{6}{k(k+1)(k+2)}\sum_{i=1}^{k}\frac{\eta_i\gamma_i}{\Gamma_i} \leq \frac{15LD_X^2}{2(k+1)(k+2)}. \tag{4.123}$$

令不等式 (4.123) 右端小于 ϵ，我们可以得到子程序 $\mathcal{FO}$ 的分析复杂度为，

$$N(\mathcal{FO}) = \sqrt{\frac{15LD_X^2}{2\epsilon}}. \tag{4.124}$$

若外迭代次数为 N，则总内迭代次数的上界为，

$$\sum_{k=1}^{N} T_k \leq \sum_{k=1}^{N}\left(\frac{6\beta_k D_X^2}{\eta_k} + 1\right) = 18\sum_{k=1}^{N} k + N = 9N^2 + 10N. \tag{4.125}$$

而子程序 $\mathcal{LO}$ 的分析复杂度为，

$$N(\mathcal{LO}) = 9N^2(\mathcal{FO}) + 10N(\mathcal{FO}) = \frac{135LD_X^2}{2\epsilon} + 10\sqrt{\frac{15LD_X^2}{2\epsilon}}. \tag{4.126}$$

对于定理的第二部分，当外迭代总次数 N 提前固定，取 $\beta_k = \frac{2L}{k}$，$\gamma_k = \frac{2}{k+1}$，以及 $\eta_k = \frac{2LD_0^2}{Nk}$，则 $\Gamma_k = \frac{2}{k(k+1)}$，$\frac{\gamma_k\eta_k}{\Gamma_k} = \frac{2LD_0^2}{N}$。于是由 (4.115) 外迭代收敛速度为

$$f(y_N) - f^* \leq \frac{\beta_1\Gamma_N}{2}\|x_0 - x\|^2 + \Gamma_N\sum_{i=1}^{N}\frac{\gamma_i\eta_i}{\Gamma_i} = \frac{6LD_0^2}{N(N+1)}. \tag{4.127}$$

令不等式 4.127 右端小于 ϵ，我们可以得到子程序 $\mathcal{FO}$ 的分析复杂度为，

$$N(\mathcal{FO}) = \mathcal{O}\left(D_0\sqrt{\frac{L}{\epsilon}}\right). \tag{4.128}$$

由 (4.100) 得到内迭代总次数（$\mathcal{LO}$ 的分析复杂度）为

$$\sum_{k=1}^{N} T_k \leq \sum_{k=1}^{N}\left(\frac{6\beta_k D_X^2}{\eta_k}+1\right) = \frac{6N^2D_X^2}{D_0^2}+N. \tag{4.129}$$

因此子程序 $\mathcal{LO}$ 的分析复杂度为，

$$N(\mathcal{LO}) = \frac{6D_X^2}{D_0^2}N^2(\mathcal{FO}) + N(\mathcal{FO}) = \mathcal{O}\left(\frac{LD_X^2}{\epsilon}+D_0\sqrt{\frac{L}{\epsilon}}\right). \tag{4.130}$$

□

由上述定理我们可以看出，加速条件梯度法的 $\mathcal{FO}$ 和 $\mathcal{LO}$ 分析复杂度分别达到了函数集合 $C_L^{1,1}(X)$ 对应的复杂度下界，即 $\mathcal{O}(1/\sqrt{\epsilon})$ 和 $\mathcal{O}(1/\epsilon)$。

5 随机优化算法的复杂度分析

5.1 引言

这一章我们研究随机优化问题的复杂度分析。首先，我们定义随机优化问题的具体形式。考虑随机优化问题，

$$\min_{x\in X}\{f(x) := \mathbb{E}_\xi[F(x,\xi)]\}, \tag{5.1}$$

其中 $X \subset \mathbb{R}^n$ 是一个非空有界闭凸集合。ξ 是一个随机向量，其分布 P 的支撑集满足 $\Xi \subset \mathbb{R}^d$。定义函数 $F: X\times\Xi \to \mathbb{R}$，对任意 $\xi\in\Xi$，$F(x,\xi)$ 都是凸函数，且关于 ξ 的期望

$$\mathbb{E}[F(x,\xi)] = \int_\Xi F(x,\xi)dP(\xi) \tag{5.2}$$

是有意义的，且对任意 $x\in X$ 都为有限值。因此，我们可以推出 $f(\cdot)$ 在 X 上是凸的并且具有有限值。同样，由凸函数在有界闭集上连续，我们推出 $f(\cdot)$ 在 X 上连续。

有了这些假设 (5.1) 变成了一个确定的凸优化问题。但在求解优化问题 (5.1) 前，我们需要首先计算出期望 (5.2) 的具体形式。从计算复杂度角度讲，即使 ξ 的分布 P 已知，随着维数的增加，(5.2) 的计算也会变得很困难。事实上，对任意 $x\in X$，我们需要计算 $f(x)$ 的一个近似逼近值 $\hat{f}(x)$，使得 $|\hat{f}(x)-f(x)|\leq\epsilon$，而这需要在不同的 $\xi\in\Xi$ 处计算 $\mathcal{O}(\frac{1}{\epsilon^m})$ 次 $f(x,\xi)$。因此在高维情况下，直接求解优化问题 (5.1)很困难，我们需要采用随机算法，比如蒙特卡洛抽样技巧来降低求解 (5.1)的复杂度。为此，我们需要对问题 (5.1) 做如下假设，

A.1 对于随机向量 ξ，存在已知的蒙特卡洛策略，能够产生一系列独立同分布的样本：$\{\xi_i\}_{i=1}^N$。

对于随机优化问题 (5.1)，我们给出随机向量 ξ 分布已知和未知情形下的两个常见例子，随机效用模型 (stochastic utility model) 和机器学习模型。

例 5.1 随机效用模型

$$\min_{x\in X}\left\{f(x) = \mathbb{E}_\xi\left[\phi\left(\sum_{i=1}^{n}(\frac{i}{n}+\xi_i)x_i\right)\right]\right\}, \tag{5.3}$$

其中 $X=\{x\in\mathbb{R}^n: x\geq 0, \sum_{i=1}^n x_i=1\}$，$\xi_i$ 独立且 $\xi_i\sim N(0,1)$，$\phi(\cdot)$ 是逐段线性的凸函数，

$$\phi(t)=\max\{a_1t+b_1,...,a_mt+b_m\}, \tag{5.4}$$

其中 a_i，b_i 是已知常数。

例 5.2 机器学习模型

对于机器学习中的分类或回归问题，我们有 n 个训练样本 $\{(x_1,y_1),...,(x_n,y_n)\}$，其中样本特征 $x_i\in\mathcal{X}$，样本标签 $y_i\in\mathcal{Y}$。假设其中的样本 (x_i,y_i) 都是从某一未知分布 D 中独立抽样出来。我们需要求解分类或回归模型 $h(\,\cdot\,;\omega):\mathcal{X}\to\mathcal{Y}$，其中 ω 是预测函数的参数。为此我们构造误差函数 $\ell(h(\,\cdot\,;\omega),\,\cdot\,):\mathcal{X}\times\mathcal{Y}\to\mathbb{R}$，并定义其在分布 D 上的期望风险 *(expected risk)* $R(w)$，

$$R(w)=\mathbb{E}_{(x,y)\sim D}[\ell(h(x;\omega),y)]. \tag{5.5}$$

通过求解优化问题

$$\min_w\ R(w)+r(w) \tag{5.6}$$

来求得分类或回归模型 $h(x;\omega)$ 的最优参数 w^*，其中 $r(\cdot)$ 是正则项，可以取 $r(w)=\|w\|_2^2$ 或 $r(w)=\|w\|_1$，用来控制模型 $h(\,\cdot\,;\omega)$ 的复杂度，防止过拟合。

对例 5.2 来说，由于数据分布 D 未知，期望风险 (5.5) 无法计算，但考虑到数据集 $(x_i,y_i)\sim D$ 独立同分布，因此我们用经验风险（empirical risk）

$$R_n(w):=\frac{1}{n}\sum_{i=1}^n\ell(h(x_i;\omega),y_i) \tag{5.7}$$

来近似 $R(w)$，通过求解优化问题

$$\min_w\ R_n(w)+r(w) \tag{5.8}$$

得到 $h(x;\omega^*)$ 的一个近似。这样会出现一个问题，我们如何衡量优化问题 (5.8) 的解对优化问题 (5.6) 解之间的关系?

这时，我们需要首先定义随机优化问题 (5.1) 解的概念。如果将随机优化问题 (5.1) 看成是一个确定性问题，先计算出 (5.1) 期望的具体形式，再用确定性优化算法（比如梯度法，牛顿法）求解其 ϵ 近似解，即，

$$\text{寻找}\ \ \bar{x}\in X:\quad f(\bar{x})-f^*\leq\epsilon. \tag{5.9}$$

然而计算 (5.1) 的具体形式非常困难。因此，为了绕过积分计算，我们需要采用一些随机算法求解 (5.1)。由假设 A1，对随机变量 ξ 采样得到样本数据 $\{\xi_i\}$，通过处理 $\{f(x,\xi_i)\}$ 我们可以得到 (5.1) 的一个近似解 $\tilde{x}$。注意到近似解 $\tilde{x}$ 是一个关于 ξ 的函数，因此它是一个随机变量。传统确定性优化问题的解概念 (5.9) 不再适应于 $\tilde{x}$。我们可以引入如下解概念。

定义 5.1 由随机优化算法产生的随机变量 $\tilde{x}\in X$，如果其满足，

$$\mathbb{E}_{\tilde{x}}[f(\tilde{x})-f^*]\leq\epsilon. \tag{5.10}$$

则称其为随机优化问题 (5.1) 的 ϵ 近似解。

由于同样无法精确计算期望，解概念 (5.10) 与 (5.9) 并没有很大的区别。因此我们可以弱化用期望来衡量随机变量解 $\tilde{x}$ 的精确性，使用关于解不等式成立的概率置信水平来衡量其解的好坏，即引入 (ϵ,δ) 近似解。

定义 5.2 由随机优化算法产生的随机变量 $\tilde{x} \in X$，如果其满足，

$$\mathbf{Prob}\left(f(\tilde{x}) - f^* \geq \epsilon\right) \leq \delta. \tag{5.11}$$

则称其为随机优化问题 (5.1) 的 (ϵ, δ) 近似解

在随机优化问题的 (ϵ, δ) 近似解概念中，δ 代表解收敛到 ϵ 精度的置信水平。当 $\delta = 0$ 时，近似解 (5.11) 等价于解 (5.9)。当 δ 越趋近于 0，表示随机算法得到的近似解 $\tilde{x}$ 收敛到 ϵ 精度的可能性越高（置信水平越高）。

有了上述随机优化问题的解概念，我们可以研究随机优化算法的收敛性和复杂度。本章的结构如下：我们先介绍求解随机优化问题的两个策略：采样平均逼近和随机逼近算法，并比较它们收敛性的置信水平；然后介绍随机梯度法在非光滑问题中的复杂度；最后介绍随机梯度法在光滑凸问题和光滑非凸问题中的复杂度分析。

5.2 随机优化算法

这一节我们介绍三种不同的随机优化算法：采样平均逼近算法（Sample Average Approximation Method）、随机逼近算法（Stochastic Approximation Method）、和集成随机逼近算法（Ensemble Stochastic Approximation Method）。其中采样平均逼近算法首先通过采样（例如蒙特卡罗算法）对目标函数进行近似，然后求解近似函数的最优值，更详细的可以参考文献 [30]，[31]，[46]，[47]，[49]。随机逼近算法的一个典型例子是随机梯度法，它直接利用目标函数的随机梯度信息来更新迭代点求解最优近似解。随机逼近算法的历史可以追溯到 1951 年 Robbins 和 Monro 的开创性工作 [45]，从那开始随机逼近算法（随机梯度法）被广泛应用于求解随机优化问题，可参考文献 [10]，[43]，[23]。随着随机梯度法在机器学习问题中的广泛应用，产生了一些新形式的随机梯度法，具体可以参考文献 [43]，[44]，[18]，[13]，[14]，[25]，[15]，[16]，[34]，[22]，[7]。集成随机逼近算法是利用机器学习中集成算法的思想，将多次随机逼近算法的解综合起来得到置信水平最高的近似解，具体的可以参考文献 [41]。

5.2.1 采样平均逼近算法的复杂度分析

在求解随机优化问题 5.1 时，采样平均逼近算法利用假设 A.1 产生样本序列 $\{\xi_i\}_{i=1}^N$，构造 (5.1) 目标函数的逼近，

$$\hat{f}_N(x) = \frac{1}{N}\sum_{i=1}^{N} F(x, \xi_i). \tag{5.12}$$

然后再用确定的优化算法（比如梯度法，牛顿法）求 $\hat{f}_N(x)$ 在 X 上的极小值。例如，在例 5.2 中我们采用经验风险 $R_n(w)$ 来逼近期望风险 $R(w)$。采样平均逼近算法具体如下：

算法 5.1 采样平均逼近算法

1. 对随机变量 ξ 进行抽样，产生样本序列 ξ_i，$i = 1, ..., N$
2. 构造 $f(x)$ 的逼近函数：

$$\hat{f}_N(x) = \frac{1}{N}\sum_{i=1}^{N} F(x, \xi_i). \tag{5.13}$$

3. 求解确定优化问题：

$$\min_{x \in X}\left\{\hat{f}_N(x) = \frac{1}{N}\sum_{i=1}^{N} F(x, \xi_i)\right\}. \tag{5.14}$$

当采样的样本总量 N 变大，对任意的 $x \in X$ 以及 $\epsilon > 0$，我们有结论，

$$\mathbf{Prob}\left(|\frac{1}{N}\sum_{i=1}^{N} F(x,\xi_i) - f(x)| \geq \epsilon\right) \to 0. \tag{5.15}$$

从计算角度来讲，上述结论非常弱，因为当 ξ 的维数很高时在某一确定的点 x 处对 (5.1) 中 $f(x)$ 做逼近会很困难。但是，当 $f(x,\xi)$ 关于 ξ 一致有界时，我们可以得到比 (5.14) 更强的结论：可以设计在多项式时间内逼近 $f(x)$ 的方法。我们可以衡量 $\hat{f}(x)$ 逼近 $f(x)$ 的复杂度。

定理 5.1 *假定 (5.1) 中函数 $F(x,\xi)$ 任意变差有界，即存在 $V < \infty$，其中*

$$V = \max\{F(x_1,\xi_1) - F(x_2,\xi_2) : \ x_1, x_2 \in X, \ \xi_1, \xi_1 \in \Xi\}. \tag{5.16}$$

则对任意 $\epsilon > 0$ 和 $\delta \in (0,1)$，当抽样样本数量取 $N = \lceil \frac{V^2}{2\epsilon^2}\log(\frac{2}{\delta})\rceil$ 时我们有，

$$\mathbf{Prob}\{|\hat{f}_N(x) - f(x)| > \epsilon\} \leq \delta. \tag{5.17}$$

为了证明定理 5.1，我们需要介绍随机优化算法收敛性证明的常用概率分析工具：Hoeffding 不等式。

引理 5.1（Hoeffding 不等式） 设 $Z_1, ..., Z_N$ 是具有相同期望 $\mu > 0$ 相互独立的随机变量，且 $\mathbf{Prob}(0 \leq Z_i \leq V) = 1$, $i = 1, ..., N$，则对于 $\bar{Z}_N = \frac{Z_1+\cdots+Z_N}{N}$，我们有

$$\mathbf{Prob}(\bar{Z}_N - \mu \geq \epsilon) \leq \exp\left(-\frac{2N\epsilon^2}{V^2}\right). \tag{5.18}$$

有了引理 5.1 的不等式工具，我们可以证明定理 5.1 的收敛性结果。

证明： 令 $\underline{f} = \inf_{x,\xi} F(x,\xi)$，定义随机变量 $Z(\xi) = F(x,\xi) - \underline{f}$ 和 $W(\xi) = V - Z(\xi)$，我没有，$0 \leq Z(\xi) \leq V$ 和 $0 \leq W(\xi) \leq V$，同时，

$$\mu = \mathbb{E}(Z) = f(x) - \underline{f} \quad 和 \quad \mu' = \mathbb{E}(W) = V - \mu. \tag{5.19}$$

注意到，

$$\begin{aligned} \mathbf{Prob}(|\bar{Z}_N - \mu|) &= \mathbf{Prob}(\bar{Z}_N - \mu \geq \epsilon) + \mathbf{Prob}(\bar{Z}_N - \mu \leq -\epsilon) &(5.20)\\ &= \mathbf{Prob}(\bar{Z}_N - \mu \geq \epsilon) + \mathbf{Prob}(\bar{W}_N - \mu' \geq \epsilon). &(5.21)\end{aligned}$$

对 Z 和 W 利用引理 5.1，并取 $N = \lceil \frac{V^2}{2\epsilon^2}\log(\frac{2}{\delta})\rceil$，

$$\mathbf{Prob}\{|\hat{f}_N(x) - f(x)| > \epsilon\} \leq 2\exp(-2N\epsilon^2/V^2) \leq \delta. \tag{5.22}$$

□

从定理 5.1 我们知道，对任意 $x \in X$，为了对 5.1 中 $f(x)$ 做 ϵ 精度的逼近，且逼近的置信水平为 δ，我们需要至少对随机变量抽样 $\lceil \frac{V^2}{2\epsilon^2}\log(\frac{2}{\delta})\rceil$ 次。设 x^* 表示优化问题 5.1 的最优解，$\hat{x}^*$ 表示优化问题 5.14 的最优解。定理 5.1 告诉我们，

$$\mathbf{Prob}\{|\hat{f}_N(x^*) - f(x^*)| > \epsilon\} \leq \delta. \tag{5.23}$$

而我们更关注是否有，

$$\mathbf{Prob}\{|\hat{f}_N(\hat{x}^*) - f(x^*)| > \epsilon\} \leq \delta. \tag{5.24}$$

成立。因为采样平均逼近算法实际中求解的是优化问题 5.14。另外，考虑到数值优化算法只能对 $\hat{x}^*$ 求解到某一 ϵ 精度，我们需要知道 $\hat{x}^*$ 的 ϵ 近似解是否也是 x^* 的 (ϵ,δ) 近似解？Shapiro 和 Nemirovski 在文献 [35] 中给出了如下定理。

定理 5.2 若 $X \subset \mathbb{R}^n$，令 $D := \sup_{x,y\in X} \|x-y\|$，假设存在 $L>0$，使得对任意 $x,y \in X$ 和 $\xi \in \Xi$ 有 $|F(x,\xi)-F(y,\xi)| \le L\|x-y\|$。如果 5.14 中 N 满足

$$N \ge \mathcal{O}(1)\left(\frac{DL}{\epsilon}\right)^2\left[n\ln\left(\frac{DL}{\epsilon}\right)+\ln\left(\frac{1}{\delta}\right)\right]. \tag{5.25}$$

则优化问题 5.14 每个精度为 $\epsilon/2$ 的近似解都是随机优化问题 5.1 的 (ϵ,δ) 近似解。

有了上述定理 5.2 我们就可以分析抽样平均逼近算法的复杂度。考虑为求得随机优化问题 5.1 (ϵ,δ) 近似解所需要对 $F(x,\xi)$ 求导多少次。对 $\hat{f}_N(x)$ 求一次导数需要求 N 次 $F(x,\xi)$ 的导数。于是我么有如下复杂度结论：

1. 若 $\hat{f}_N(x) \in C_L^{0,1}(X)$ 且是凸函数，利用投影次梯度法求 5.14 精度为 $\epsilon/2$ 的近似解需要对 $\hat{f}_N(x)$ 求 $\mathcal{O}(L^2D^2/\epsilon^2)$ 次导数，若使该解为 5.1 的 (ϵ,δ) 近似解，则总共需要对 $F(x,\xi)$ 求导的次数不少于：

$$\mathcal{O}\left(\frac{D^4L^4}{\epsilon^4}\left[n\ln\left(\frac{DL}{\epsilon}\right)+\ln\left(\frac{1}{\delta}\right)\right]\right). \tag{5.26}$$

忽略常数项后，抽样平均逼近算法求解随机优化问题 5.1 的复杂度为

$$\mathcal{O}\left(\frac{n}{\epsilon^4}\ln\frac{1}{\epsilon}+\frac{1}{\epsilon^4}\ln\frac{1}{\delta}\right). \tag{5.27}$$

2. 若 $\hat{f}_N(x) \in C_L^{0,1}(X)$，利用投影梯度法求 5.1 的 (ϵ,δ) 近似解至少需要求 $F(x,\xi)$ 导数的次数不少于：

5.2.2 随机逼近算法的复杂度分析

另一个策略随机逼近算法是随机梯度算法的原型，它与采样平均逼近算法的区别在于每次迭代不用计算 N 次 $F(x,\xi)$ 的导数，而是利用产生一个随机梯度来近似 $f(x)$ 的梯度。例如，当 5.1 的目标函数 $f(x)$ 非光滑时，随机逼近算法利用如下假设。

A.2 存在子程序 $\mathcal{SFO}$ 对任意 $x \in X$ 和随机样本 $\xi \in \Xi$，返回函数值 $F(x,\xi)$ 和一阶随机次梯度 $G(x,\xi)$，满足

$$\mathbb{E}_\xi[G(x,\xi)] = g(x) \in \partial f(x). \tag{5.28}$$

在假设 A.2 下，随机逼近算法通过调用子程序 $\mathcal{SFO}$ 产生随机次梯度 $G(x,\xi)$ 信息，每次迭代只有一次次梯度计算过程，而采样平均逼近算法需要计算 N 次。

算法 5.2 随机逼近算法框架 $\mathcal{S}$

输入：$x_0 \in X$，初始信息集合 $I_{-1}=\emptyset$，对 $k=0,1,\cdots,N$ 执行迭代

1. 根据 ξ 的分布生成样本 ξ_k，
2. 在 ξ_k 处调用子程序 $\mathcal{SFO}$，更新信息集合 $I_{k+1} = I_k \cup (x_k,\xi_k,\mathcal{SFO}(x_k,\xi_k))$，
3. 根据信息集合 I_{k+1} 更新迭代点 x_{k+1}，

输出：$\bar{x} = \mathcal{S}(x_0)$.

利用机器学习中集成学习的概念，我们可以对随机逼近算法进行集成，得到集成随机逼近算法。该算法最早被 Nesterov 和 Vial [41] 提出。

算法 5.3 集成随机逼近算法 $\mathcal{M}$

输入：$x_0 \in \mathbb{R}^n$，随机逼近基算法 $\mathcal{S}$，每个基算法迭代次数 N, 集成次数 K。

对 $j = 1, ..., K$ 执行迭代，

1. 调用随机逼近基算法 $\mathcal{S}$，得到近似解 $\bar{x}_j = \mathcal{S}(x_0)$.

输出：$\hat{x} = \mathcal{M}(x_0) = \frac{1}{K}\sum_{j=1}^{K} \bar{x}_j$.

我们可以研究算法 5.3 的解 $\hat{x} = \mathcal{M}(x_0)$ 和算法 5.2 的解 $\bar{x} = \mathcal{S}(x_0)$ 之间的关系，并比较它们求 (ϵ, δ) 近似解的复杂度。为了分析复杂度，我们对 (5.1) 的目标函数做如下假设，

1. $f(x)$ 在 X 上的总变差有界，即 $V_f = \sup\{f(x) - f^* \ : \ x \in X\} < \infty$。
2. $F(x, \xi)$ 在 $X \times \Xi$ 上的任意变差有界。

对算法 5.2 的解 $\bar{x}$，若我们可以证明存在关于迭代次数 N 的单调递减函数 $C_{\mathcal{S}}(N)$ 使得，

$$\mathbb{E}\ [\ f(\bar{x}) - f^*\ ] \ \leq \ C_{\mathcal{S}}(N). \tag{5.29}$$

则当 N 满足 $C_{\mathcal{S}}(N) \leq \epsilon\delta$ 时，$\bar{x}$ 是 5.1 的 (ϵ, δ) 近似解。因为，由 Markov 不等式，我们有，

$$\mathbf{Prob}\{f(\bar{x}) - f^* \geq \epsilon\} \leq \frac{\mathbb{E}\ [\ f(\bar{x}) - f^*\ ]}{\epsilon} \leq \frac{C_{\mathcal{M}}(N)}{\epsilon} \leq \delta. \tag{5.30}$$

而对于算法 5.3 的解 $\hat{x}$ 我们有如下引理，

引理 5.2 *设 $\hat{x}$ 为算法 (5.3) 的解，则对任意 $\beta > 0$ 我们有，*

$$\mathbf{Prob}\{f(\hat{x}) - f^* \geq C_{\mathcal{M}}(N) + \beta\} \leq \exp\{-\frac{2K\beta^2}{V_f^2}\}. \tag{5.31}$$

证明: 利用 $f(x)$ 的凸性得到 $f(\hat{x}) \leq \frac{1}{K}\sum_{j=1}^{K} f(\bar{x}_j)$，于是

$$\mathbf{Prob}\{f(\hat{x}) - f^* \geq C_{\mathcal{M}}(N) + \beta\} \leq \mathbf{Prob}\{\frac{1}{K}\sum_{j=1}^{K} f(\bar{x}_j) - f^* \geq C_{\mathcal{M}}(N) + \beta\}. \tag{5.32}$$

令 $Z = f(\bar{x}) - f^*$，$\bar{Z}_K = \frac{1}{K}\sum_{j=1}^{K} f(\bar{x}_j) - f^*$, $\mu = \mathbb{E}[f(\bar{x}) - f^*] = \mathbb{E}(Z)$，由假设我们有 $0 \leq Z \leq V_f$。对 Z 利用 Hoeffding 不等式，

$$\begin{aligned}
& \mathbf{Prob}\{\frac{1}{K}\sum_{j=1}^{K} f(\bar{x}_j) - f^* \geq C_{\mathcal{M}}(N) + \beta\} && (5.33)\\
= \ & \mathbf{Prob}\{\bar{Z}_K \geq \mathbb{E}[f(\bar{x}) - f^*] + \beta\} && (5.34)\\
\leq \ & \mathbf{Prob}\{\bar{Z}_K \geq \mu + \beta\} && (5.35)\\
\leq \ & \exp\{\frac{-2K\beta^2}{V_f^2}\}. && (5.36)
\end{aligned}$$

于是我们得到不等式 5.31。 □

定理 5.3 *令 N 和 K 按如下方式选取，*

$$N = \left\lceil C_{\mathcal{S}}^{-1}\left(\frac{\epsilon V_f}{2}\right)\right\rceil, \qquad K = \left\lceil \frac{2}{\epsilon^2}\ln\frac{1}{\delta}\right\rceil. \tag{5.37}$$

则算法 5.3 的解 $\hat{x}$ 是随机优化问题 5.1 的 $(\epsilon V_f, \delta)$ 近似解，即，

$$\mathbf{Prob}\{f(\hat{x}) - f^* \geq \epsilon V_f\} \leq \delta. \tag{5.38}$$

且算法 5.3 总共需要调用子程序 $\mathcal{SFO}$ 的次数，即计算复杂度 T 满足，

$$T = K \cdot N \leq \left(1 + C_{\mathcal{S}}^{-1}\left(\frac{\epsilon V_f}{2}\right)\right) \cdot \left(1 + \frac{2}{\epsilon^2}\ln\frac{1}{\delta}\right). \tag{5.39}$$

证明：令 $\beta = \frac{1}{2}\epsilon V_f$，根据引理 5.2 将 N 和 K 的取值 (5.37) 代入 (5.31) 即得结论。 □

由随机梯度法的收敛性结论:

$$\mathbb{E}[f(\bar{x}_1^k) - f^*] \leq \frac{DM}{\sqrt{k}}, \tag{5.40}$$

因此有 $C_{\mathcal{S}}(N) = DM/\sqrt{N}$。由 (5.30) 令 $C_{\mathcal{S}}(N) \leq \epsilon V_f \delta$，则为了得到 $(\epsilon V_f, \delta)$ 近似解，算法 5.2 迭代次数 N 需要满足，

$$N = \mathcal{O}\left[\frac{1}{\epsilon^2\delta^2}\left(\frac{DM}{V_f}\right)^2\right]. \tag{5.41}$$

若基算法 $\mathcal{S}$ 5.2 是随机梯度法，我们可以考虑算法 $\mathcal{M}$ 5.3 $(\epsilon V_f, \delta)$ 近似解的复杂度，由定理 5.37 得，

$$N = \mathcal{O}\left[\frac{1}{\epsilon^2}\left(\frac{MR}{V_f}\right)^2\right], \qquad T = K \cdot N = \mathcal{O}\left[\frac{1}{\epsilon^4}\ln\frac{1}{\delta}\left(\frac{MR}{V_f}\right)^2\right]. \tag{5.42}$$

我们将算法 $\mathcal{S}$ 与算法 $\mathcal{M}$ 的复杂度结论总结到下表。

	算法 $\mathcal{S}$	算法 $\mathcal{M}$
基算法个数 K	1	$\frac{1}{\epsilon^2}\ln\frac{1}{\delta}$
基算法迭代次数 N	$\frac{1}{\epsilon^2\delta^2}\left(\frac{MR}{V_f}\right)^2$	$\frac{1}{\epsilon^2}\left(\frac{MR}{V_f}\right)^2$
总计算复杂度	$\frac{1}{\epsilon^2\delta^2}\left(\frac{MR}{V_f}\right)^2$	$\frac{1}{\epsilon^4}\ln\frac{1}{\delta}\left(\frac{MR}{V_f}\right)^2$

表 2: 随机次梯度法与集成随机次梯度法计算复杂度的比较

5.3 非光滑随机优化问题

与其梯度法和镜像下降法类似，我们可以推广得到随机逼近算法框架下的随机梯度法和随机镜像下降法。

算法 5.4 随机梯度法（Stochastic Gradient Method）

输入：$x_0 \in \mathbb{R}^n$ 和步长序列 $\{\gamma_k\}$，

对 $k = 0, 1, \cdots$ 执行迭代，

1. 生成样本 ξ_k，
2. 在 (x_k, ξ_k) 处调用子程序 $\mathcal{SFO}$ 得到随机梯度 $G(x_k, \xi_k)$，并更新

$$x_{k+1} = \operatorname*{argmin}_{x \in X} \|x_k - \gamma_k G(x_k, \xi_k)\|_2. \tag{5.43}$$

算法 5.5 随机镜像下降法（Stochastic Mirror Descent Method）

输入：$x_0 \in \mathbb{R}^n$，步长序列 $\{\gamma_k\}$ 以及 Bregman 距离 $V(x,y)$。

对 $k = 0, 1, \cdots$ 执行迭代，

1. 根据 ξ 分布，生成样本 ξ_k，
2. 在 (x_k, ξ_k) 处调用子程序 $\mathcal{SFO}$ 得到随机梯度 $G(x_k, \xi_k)$，并更新

$$x_{k+1} = \operatorname*{argmin}_{x \in X} \ \gamma_k \langle G(x_k, \xi_k), x \rangle + V(x_k, x). \tag{5.44}$$

当取 Bregman 距离为 $V(x,y) = \frac{1}{2}\|x-y\|^2$ 时，随机梯度法 5.4 与随机镜像下降法 5.7 等价。因此我们考虑随机镜像下降法得复杂度分析。我们考虑随机优化问题 5.1 在下列随机梯度假设时的 ϵ 近似解和 (ϵ, δ) 近似解的复杂度。假设 $\{\xi_t\}$ 是关于 ξ 的独立同分布随机变量序列，考虑对任意的 $x \in X$，存在常数 $M_* > 0$，条件 A.2 产生的随机次梯度 $G(x, \xi_t)$ 满足下列条件之一，

A.3 存在 $M_* > 0$，使得 $\|G(x, \xi_t)\|_* \le M_*$。

A.4 存在 $M_* > 0$，使得 $\mathbb{E}_{\xi_t}\left[\exp\left(\|G(x, \xi_t)\|_*^2 / M_*^2\right)\right] \le \exp(1)$。

A.5 存在 $M_* > 0$，使得 $\mathbb{E}_{\xi_t}[\|G(x, \xi_t)\|_*^2] \le M_*^2$。

不难看出上述假设的包含关系如下，

$$\textbf{A.3} \ \Rightarrow \ \textbf{A.4} \ \Rightarrow \ \textbf{A.5}. \tag{5.45}$$

A.3 假设条件最强，**A.5** 最弱。因此，我们先考虑在 **A.5** 假设条件下 ϵ 近似解的复杂度分析。

定理 5.4 *若假设条件 A.1，A.2，A.5 成立，令 $\bar{x}_s^k = (\sum_{t=s}^k \gamma_t x_t)/(\sum_{t=s}^k \gamma_t)$，其中 x_t 表示随机镜像下降法 5.7 的迭代点，则有如下收敛性结果，*

$$\mathbb{E}[f(\bar{x}_s^k)] - f^* \le \left(\sum_{t=s}^k \gamma_t\right)^{-1} \left[\mathbb{E}[V(x_s, x^*)] + (2\nu)^{-1} M_*^2 \sum_{t=s}^k \gamma_t^2\right]. \tag{5.46}$$

证明： 记 $g_t := g(x_t) \in \partial f(x_t)$，$G_t := G(x_t, \xi_t) \in \partial_x F(x_t, \xi_t)$，令 $\Delta_t = g_t - G_t$，$t = 1, ..., k$。由 f 的凸性和引理 3.2 有，

$$\begin{aligned}
\gamma_t[f(x_t) - f(x)] &\le \gamma_t \langle G_t, x_t - x\rangle + \gamma_t \langle \delta_t, x_t - x\rangle && (5.47)\\
&\le V(x_t, x) - V(x_{t+1}) + \gamma_t \langle G_t, x_t - x_{t+1}\rangle - V(x_t, x_{t+1}) \\
&\quad + \gamma_t \langle \delta_t, x_t - x\rangle && (5.48)\\
&\le V(x_t, x) - V(x_{t+1}) + \nu^{-1}\gamma_t^2 \|G_t\|^2 + \gamma_t \langle \delta_t, x_t - x\rangle. && (5.49)
\end{aligned}$$

其中 (5.49) 成立利用了 Cauchy-Swartz 不等式以及对任意 $a > 0$ 不等式 $bt - at^2/2 \le b^2/(2a)$ 都成立的事实。对上述不等式左右两边取期望，考虑到假设 A.2 和假设 A.5，我们有，

$$\gamma_t \mathbb{E}[f(x_t) - f(x)] \le \mathbb{E}[V(x_t, x) - V(x_{t+1}, x)] + (2\nu)^{-1}\gamma_t^2 M^2. \tag{5.50}$$

将上述不等式从 $t = s$ 到 $t = k$ 求和即得结论 (5.46)。 □

推论 5.1 对随机镜像下降法 5.7 取固定步长 $\gamma_t = \sqrt{2\nu}D_{w,X}/(M\sqrt{k})$，则有收敛性结果，

$$\mathbb{E}[f(\bar{x}_1^k) - f^*] \leq D_{w,X}M\sqrt{\frac{2}{\nu k}}. \tag{5.51}$$

其 ϵ 近似解得复杂度为，

$$\mathcal{O}\left(\frac{D_{w,X}^2M^2}{\epsilon^2}\right). \tag{5.52}$$

定理 5.5 若假设条件 A.1，A.2，A.5 成立，则随机镜像下降法 5.7 求得 (ϵ,δ) 近似解的复杂度为，

$$\mathcal{O}\left(\frac{1}{\epsilon^2\delta^2}\right). \tag{5.53}$$

上述定理的证明可以参考文献 [36, 34]。

定理 5.6 若假设条件 A.1，A.2，A.4 成立，则随机镜像下降法 5.7 求得 (ϵ,δ) 近似解的复杂度为，

$$\mathcal{O}\left(\frac{1}{\epsilon^2}\ln^2\frac{1}{\delta}\right). \tag{5.54}$$

证明：记 $g(x_t) := \mathbb{E}_\xi[F(x_t,\xi)] \in \partial f(x_t)$，注意到

$$\|g(x_t)\|_* = \|\mathbb{E}(G(x_t,\xi_t)|\xi_{[t-1]})\|_* \leq \sqrt{\mathbb{E}(\|G(x_t,\xi_t)\|_*^2|\xi_{[t-1]})} \leq M_*. \tag{5.55}$$

记 $\Delta_t = g(x_t) - G(x_t,\xi_t)$，由于

$$\begin{aligned}\|\Delta_t\|_*^2 = \|g(x_t) - G(x_t,\xi_t)\|_*^2 &\leq (\|G(x_t,\xi_t)\|_* + \|g(x_t)\|_*)^2\\ &\leq 2\|G(x_t,\xi_t)\|_*^2 + 2\|g(x_t)\|_*^2.\end{aligned} \tag{5.56}$$

我们有

$$\mathbb{E}\left[\exp\left(\frac{\|G(x,\xi)\|^2}{M_*^2}\right)\right] \leq \exp\{1\} \quad \Rightarrow \quad \mathbb{E}\left[\exp\left(\frac{\|\Delta(x,\xi)\|^2}{(2M_*)^2}\right)\right] \leq \exp\{1\}. \tag{5.57}$$

利用 Markov 不等式的变形，对任意 $t > 0$

$$\mathbf{Prob}\{X \geq \epsilon\} \leq \frac{\mathbb{E}[e^{tX}]}{e^{t\epsilon}}. \tag{5.58}$$

若 $\mathbb{E}[e^X] \leq \exp(1)$，对任意 $\lambda > 0$ 我们有

$$\mathbf{Prob}\{X \geq 1+\lambda\} \leq \frac{E[e^X]}{\exp(1+\lambda)} = \exp(-\lambda). \tag{5.59}$$

因此

$$\begin{aligned}\mathbf{Prob}\{\|G(x,\xi)\|^2 \geq (1+\lambda)M_*^2\} &\leq \exp(-\lambda),\\ \mathbf{Prob}\{\|\Delta(x,\xi)\|^2 \geq 4(1+\lambda)M_*^2\} &\leq \exp(-\lambda).\end{aligned} \tag{5.60}$$

由收敛性估计不等式

$$\gamma_t[f(x_t) - f(x)] \leq V(x_t,x) - V(x_{t+1},x) + (2\nu)^{-1}\gamma_t^2\|G(x_t,\xi_t)\|^2 + \gamma_t\langle\Delta_t, x_t - x\rangle. \tag{5.61}$$

将上式从 1 加到 k 我们得到

$$\begin{aligned}f(\bar{x}_1^k) - f^* &\leq \left(\sum_{t=1}^k\gamma_t\right)^{-1}\left[V(x_1,x^*) + \sum_{t=1}^k\frac{\gamma_t^2}{\nu}\|G(x_t,\xi_t)\|^2 - \sum_{t=1}^k\gamma_t\langle\Delta_t, x_t - x^*\rangle\right]\\ &\leq \left(\sum_{t=1}^k\gamma_t\right)^{-1}\left[V(x_1,x^*) + \sum_{t=1}^k\frac{\gamma_t^2}{\nu}\|G(x_t,\xi_t)\|^2 + \sum_{t=1}^k\gamma_t\|\Delta_t\|\|x_t - x^*\|\right].\end{aligned} \tag{5.62}$$

取 $V(x,y)$, $D_{\omega,X}$ 和 $\{\gamma_t\}$ 如下

$$V(x,y) \geq \frac{\nu}{2}\|x-y\|^2, \quad D_{\omega,X} = \sup_{x,y\in X} V(x,y), \quad \gamma_t = \frac{\sqrt{2\nu}D_{\omega,X}}{M_* k}. \tag{5.63}$$

在假设 $\mathbb{E}\left[\exp\left(\|G(x,\xi)\|_*^2/M_*^2\right)\right] \leq \exp(1)$ 下，我们有

$$\mathbf{Prob}\left\{f(\bar{x}_1^k) - f(x_*) > \frac{\sqrt{2}M_* D_{\omega,X}(12+2\lambda)}{\sqrt{\nu N}}\right\} \leq 2\exp(-\lambda). \tag{5.64}$$

将上式右端项等于 ϵ 解得 $\lambda(\epsilon)$ 并带入左边的估计值可以得到结论。 □

引理 5.3（Azuma-Hoeffding 不等式） 设 $Z_1, Z_2, \ldots$ 是一个鞅差序列，且对 $i = 1, 2, \ldots$ 满足 $|Z_i| \leq V$，则下列不等式成立

$$\mathbf{Prob}\left(\sum_{i=1}^N Z_i \geq c\right) \leq \exp\left(-\frac{2c^2}{NV^2}\right), \tag{5.65}$$

$$\mathbf{Prob}\left(\sum_{i=1}^N Z_i \leq c\right) \leq \exp\left(-\frac{2c^2}{NV^2}\right). \tag{5.66}$$

定理 5.7 *若假设条件 A.1，A.2，A.3 成立，则随机镜像下降法 5.7 求得 (ϵ,δ) 近似解的复杂度为，*

$$\mathcal{O}\left(\frac{1}{\epsilon^2}\ln\frac{1}{\delta}\right). \tag{5.67}$$

证明： 将不等式 5.49 左右两边从 $t=1$ 到 k 求和得，

$$\gamma_t[f(x_t) - f(x)] \leq V(x_t,x) - V(x_{t+1}) + \nu^{-1}\gamma_t^2\|G_t\|^2 + \gamma_t\langle\delta_t, x_t - x\rangle. \tag{5.68}$$

$$\mathbf{Prob}\left\{f(\bar{x}_N) - f^* \geq \frac{3DM_*}{2\sqrt{N}} + \frac{DM_*\sqrt{2\ln\frac{1}{\delta}}}{\sqrt{N}}\right\} \leq \delta. \tag{5.69}$$

□

我们将不同条件下随机次梯度算法的 (ϵ,δ) 计算复杂度总结到下表。

假设条件	(ϵ,δ) 复杂度
$\|G(x,\xi_t)\|_* \leq M_*$	$\mathcal{O}\left(\frac{1}{\epsilon^2\delta^2}\right)$
$\mathbb{E}_{\xi_t}\left[\exp\left(\|G(x,\xi_t)\|_*^2/M_*^2\right)\right] \leq \exp(1)$	$\mathcal{O}\left(\frac{1}{\epsilon^2}\ln^2\frac{1}{\delta}\right)$
$\mathbb{E}_{\xi_t}[\|G(x,\xi_t)\|_*^2] \leq M_*^2$	$\mathcal{O}\left(\frac{1}{\epsilon^2}\ln\frac{1}{\delta}\right)$

表 3: 随机次梯度法 (ϵ,δ) 计算复杂度的比较

5.4 光滑随机优化问题

5.4.1 凸随机优化问题

这一节我们仍旧考虑随机优化问题 (5.1)，但对目标函数 f 做如下假设：

F.1 $f(x)$ 光滑且是凸函数，

F.2 $f(x)$ 的导数满足，$\|\nabla f(x) - \nabla f(y)\|_* \le L\|x-y\|$, $\forall\, x, y \in X$。

并考虑 $f(x)$ 的随机梯度在如下假设条件下的收敛性。

A.6 存在子程序 $\mathcal{SFO}$ 对任意 $x \in X$ 和随机样本 $\xi \in \Xi$，返回函数值 $F(x,\xi)$ 和随机梯度 $G(x,\xi)$。且存在 $\sigma > 0$，对任意 $x \in X$ 有，

$$a) \qquad \mathbb{E}[G(x,\xi)] \equiv \nabla f(x). \tag{5.70}$$

$$b) \qquad \mathbb{E}[\|G(x,\xi) - \nabla f(x)\|_*^2] \le \sigma^2. \tag{5.71}$$

A.7 对任意 $x \in X$ 我们有，

$$\mathbb{E}\left[\exp\left\{\|G(x,\xi_t) - \nabla f(x)\|_*^2/\sigma^2\right\}\right] \le \exp(1). \tag{5.72}$$

其中条件 A.6 说明随机梯度 $G(x,\xi_t)$ 是无偏的，且方差小于 σ^2。由 Jensen 不等式，我们有如下包含关系，

$$\textbf{A.7} \;\Rightarrow\; \textbf{A.6.b)} \tag{5.73}$$

当 $f(x)$ 非光滑时，我们利用条件 A.3，即 $\|G(x,\xi_t)\|_* \le M_*$，考虑到，

$$\begin{aligned}
\mathbb{E}[\|G(x_t,\xi_t)\|_*^2] &= \|\mathbb{E}[G(x_t,\xi_t)]\|_*^2 + \mathbb{E}[\|G(x_t,\xi_t) - \mathbb{E}[G(x_t,\xi_t)]\|_*^2] \\
&= \|\nabla f(x_t)\|_*^2 + \mathbb{E}[\|G(x_t,\xi_t) - \nabla f(x_t)\|_*^2] \\
&\le \|\nabla f(x_t)\|_*^2 + \sigma^2 \\
&= \|\nabla f(x_1) + \nabla f(x_t) - \nabla f(x_1)\|_*^2 + \sigma^2 \\
&\le 2\|\nabla f(x_1)\|_*^2 + 2\|\nabla f(x_t) - \nabla f(x_1)\|_*^2 + \sigma^2 \\
&\le 2\|\nabla f(x_1)\|_*^2 + 2L^2\|x_t - x_1\|^2 + \sigma^2 \\
&\le 2\|\nabla f(x_1)\|_*^2 + 4\nu^{-1}L^2 D_{\omega,X}^2 + \sigma^2.
\end{aligned}$$

利用收敛性结论，

$$\gamma_t[f(x_t) - f(x)] \le V(x_t,x) - V(x_{t+1},x) + (2\nu)^{-1}\gamma_t^2\|G(x_t,\xi_t)\|^2 + \gamma_t\langle \Delta_t, x_t - x\rangle. \tag{5.74}$$

对上述不等式两边取关于 ξ_t 的期望并从 $t=1$ 到 $t=k$ 求和，代如 [ref] 我们有

$$f(\bar{x}_1^k) - f^* \le \frac{D_{\omega,X}(\|\nabla f(x_1)\|_* + L\nu^{-1/2}D_{\omega,X} + \sigma)}{\sqrt{\nu k}}. \tag{5.75}$$

上述结论并没有利用 f 的光滑性。事实上，当 $\sigma = 0$ 时，随机优化问题变成确定性优化问题，则上述收敛性结论和梯度法求解确定性光滑优化问题的收敛性不一致（即满足 $\mathcal{O}(1/k)$ 的收敛速度）。通过利用 f 的光滑性，新的收敛性分析技巧和步长选择，我们可以得到更好的收敛性结果。考虑随机镜像梯度法输出如下解，

$$\tilde{x}_{k+1} = \frac{\sum_{t=1}^k \gamma_t x_{t+1}}{\sum_{t=1}^k \gamma_t} = \frac{\gamma_1 x_2 + \gamma_2 x_3 + \cdots + \gamma_k x_{k+1}}{\sum_{t=1}^k \gamma_t}. \tag{5.76}$$

我们可以证明随机梯度法在求解光滑随机凸优化问题时的复杂度。首先我们引入一个辅助引理，其证明可以在文献 [27] 中找到。

引理 5.4 设 $\xi_1,...,\xi_2$ 是独立同分布的随机变量序列，记 $\xi_{[t-1]}=(\xi_1,...,\xi_t)$，设 $Z_t=Z_t(\xi_{[t]})$ 是关于 $\xi_{[t]}$ 的 *Borel* 函数，且 $\mathbb{E}[Z_t|\xi_{[t]}]=0$ 和 $\mathbb{E}[\exp\{Z_t^2/\sigma_t^2\}|\xi_{[t]}]\le\exp(1)$ 几乎处处成立，其中 $\sigma_t>0$ 为常数，则对任意 $\lambda>0$ 有

$$\mathbf{Prob}\left(\frac{\sum_{t=1}^N Z_t}{\sum_{t=1}^N \sigma_t^2}\ge\lambda\right)\le\exp\{-\frac{\lambda^2}{3}\}. \tag{5.77}$$

由上述引理，我们可以证明如下复杂度结论 [25]。

定理 5.8 取步长 γ_t 满足 $0<\gamma_t<\nu/(2L)$，令 $\tilde{x}$ 是 (5.76) 随机镜像梯度法产生的解，则，

1) 在条件 *A.6* 下，

$$f(\tilde{x}_{k+1})-f^*\le K_0(k). \tag{5.78}$$

其中

$$K_0(k):=\frac{D_{\omega,X}^2+2\nu^{-1}\sigma^2\sum_{t=1}^k\gamma_t^2}{\sum_{t=1}^k\gamma_t}. \tag{5.79}$$

2) 在条件 *A.6*，*A.7* 下，

$$\mathbf{Prob}\{f(\tilde{x}_{k+1})-f^*\ge K_0(k)+\lambda K_1(k)\}\le e^{-\lambda^2/3}+e^{-\lambda}. \tag{5.80}$$

其中

$$K_1(k):=\frac{2D_{\omega,X}\sqrt{\frac{2\sigma^2}{\nu}\sum_{t=1}^k\gamma_t^2}+\frac{2\sigma^2}{\nu}\sum_{t=1}^k\gamma_t^2}{\sum_{t=1}^k\gamma_t}. \tag{5.81}$$

证明：令 $d_t=x_{t+1}-x_t$，$\Delta_t=G(x_t,\xi_t)-g(x_t)$，注意到 $\nu\|d_t\|^2/2\le V(x_t,x_{t+1})$，

$$\begin{aligned}
&\gamma_t f(x_{t+1})\le\gamma_t[f(x_t)+\langle g(x_t),d_t\rangle+\frac{L}{2}\|d_t\|^2] && (5.82)\\
=\;&\gamma_t[f(x_t)+\langle g(x_t),d_t\rangle]+\frac{\nu}{2}\|d_t\|^2+\frac{L-\nu}{2}\|d_t\|^2 && (5.83)\\
\le\;&\gamma_t[f(x_t)+\langle g(x_t),d_t\rangle]+V(x_{t+1},x_t)+\frac{L-\nu}{2}\|d_t\|^2 && (5.84)\\
=\;&\gamma_t[f(x_t)+\langle G(x_t,\xi_t),d_t\rangle]-\gamma_t\langle\Delta_t,d_t\rangle+V(x_{t+1},x_t)+\frac{L-\nu}{2}\|d_t\|^2 && (5.85)\\
\le\;&\gamma_t[f(x_t)+\langle G(x_t,\xi_t),d_t\rangle]+V(x_{t+1},x_t)+\frac{L-\nu}{2}\|d_t\|^2+\gamma_t\|\Delta_t\|_*\|d_t\| && (5.86)\\
\le\;&\gamma_t[f(x_t)+\langle G(x_t,\xi_t),d_t\rangle]+V(x_{t+1},x_t)+\frac{L-\nu}{2}\|d_t\|^2+\frac{\gamma_t^2\|\Delta_t\|_*^2}{2(\nu-L\gamma_t)}. && (5.87)
\end{aligned}$$

由引理 5.4 我们有

$$\begin{aligned}
&\gamma_t[f(x_t)+\langle G(x_t,\xi_t),x_{t+1}-x_t\rangle]+V(x_{t+1},x_t) && (5.88)\\
\le\;&\gamma_t f(x_t)+[\gamma_t\langle G(x_t,\xi_t),x-x_t\rangle+V(x_t,x)-V(x_{t+1},x)] && (5.89)\\
=\;&\gamma_t[f(x_t)+\langle g(x_t),x-x_t\rangle]+\gamma_t\langle\Delta_t,x-x_t\rangle+V(x_t,x)-V(x_{t+1},x) && (5.90)\\
\le\;&\gamma_t f(x)+\gamma_t\langle\Delta_t,x-x_t\rangle+V(x_t,x)-V(x_{t+1},x). && (5.91)
\end{aligned}$$

于是我们有，

$$\gamma_t[f(x_{t+1})-f(x)]+V(x_{t+1},x)\le\gamma_t\langle\Delta_t,x-x_t\rangle+V(x_t,x)+\frac{\gamma_t^2\|\Delta_t\|_*^2}{2(\nu-L\gamma_t)}. \tag{5.92}$$

将上式从 $t=1$ 到 k 求和并取 $x=x^*$

$$\sum_{t=1}^{k}\gamma_t(f(x_{t+1})-f^*)\le V(x_1,x^*)-V(x_{k+1},x^*)+\sum_{t=1}^{k}\left(\gamma_t\langle\Delta_t,x^*-x_t\rangle+\frac{\gamma_t^2\|\Delta_t\|_*^2}{2(\nu-L\gamma_t)}\right). \tag{5.93}$$

考虑到 f 的凸性我们有，

$$f(\tilde{x}_{k+1})=f\left(\frac{\sum_{t=1}^{k}\gamma_t x_{t+1}}{\sum_{t=1}^{k}\gamma_t}\right)\le\frac{\sum_{t=1}^{k}\gamma_t f(x_{t+1})}{\sum_{t=1}^{k}\gamma_t}. \tag{5.94}$$

利用 $\gamma_t\le\nu/(2L)$ 可以得到 $\nu/2\le 2(\nu-L\gamma_t)$，将 (5.93) 两边同除 $\sum_{t=1}^{k}\gamma_t$ 并带入 (5.94) 得到，

$$f(\tilde{x}_{k+1})-f^*\le\frac{V(x_1,x^*)}{\sum_{t=1}^{k}\gamma_t}+\frac{1}{\sum_{t=1}^{k}\gamma_t}\sum_{t=1}^{k}\left(\gamma_t\langle\Delta_t,x^*-x_t\rangle+\frac{2\gamma_t^2\|\Delta_t\|_*^2}{\nu}\right). \tag{5.95}$$

注意到 $(\tilde{x}_{k+1},x_t)$ 是随机变量 $\xi_{[t-1]}=(\xi_1,...,\xi_{t-1})$ 的函数，因此有

$$\mathbb{E}[\langle\Delta_t,x^*-x_t\rangle|\xi_{[t-1]}]=0. \tag{5.96}$$

由假设条件 A.7 得 $\mathbb{E}[\Delta_t|\xi_{[t-1]}]\le\sigma^2$，对 (5.95) 两边关于随机变量 $\xi_{[t-1]}$ 求期望得，

$$f(\tilde{x}_{k+1})-f^*\le\frac{D_{\omega,X}^2+2\nu^{-1}\sigma^2\sum_{t=1}^{k}\gamma_t^2}{\sum_{t=1}^{k}\gamma_t}. \tag{5.97}$$

□

有了上述定理，我们可以通过选取特定的步长序列 $\{\gamma_t\}$ 得到收敛结果。假设随机梯度算法的总迭代步数事先固定，比如 N 步。我们去固定步长 $\gamma_t=\gamma$，带入不等式 (5.78) 得

$$\mathbb{E}[f(x_{k+1}^{av})-f^*]\le\frac{D_{\omega,X}^2}{k\gamma}+\frac{2\gamma}{\nu}\sigma^2. \tag{5.98}$$

上式右端在约束 $\gamma\in(0,\nu/(2L)]$ 下关于 γ 求极小得

$$\gamma^*=\min\left\{\frac{\nu}{2L},\sqrt{\frac{\nu D_{\omega,X}^2}{2k\sigma^2}}\right\}. \tag{5.99}$$

令 $\Omega_{\omega,X}=\sqrt{2/\nu}D_{\omega,X}$，将 (5.99) 带入 (5.98) 得

$$\mathbb{E}[f(x_{k+1}^{av})-f^*]\le\frac{L\Omega_{\omega,X}^2}{k}+\frac{2\Omega_{\omega,X}\sigma}{\sqrt{k}}. \tag{5.100}$$

取 $\gamma_t=\gamma^*$ 代入不等式 (5.79) 和 (5.81) 得

$$\begin{aligned}K_0(k)&=\frac{L\Omega_{\omega,X}^2}{k}+\frac{2\Omega_{\omega,X}\sigma}{\sqrt{k}},\\ K_1(k)&=\frac{2\Omega_{\omega,X}\sigma}{\sqrt{k}}+\frac{2\gamma\sigma^2}{\nu}\le\frac{2\Omega_{\omega,X}\sigma}{\sqrt{k}}+\sqrt{\frac{2}{\nu}}D_{\omega,X}\frac{\sigma^2}{\sqrt{k\sigma^2}}=\frac{3\Omega_{\omega,X}\sigma}{\sqrt{k}}.\end{aligned} \tag{5.101}$$

将上式代入 (5.80) 得

$$\mathbf{Prob}\left\{f(x_{k+1}^{av}) - f^* \geq \frac{L\Omega^2}{k} + \frac{\Omega\sigma}{\sqrt{k}}(2+3\lambda)\right\} \leq e^{-\lambda^2/3} + e^{-\lambda}. \tag{5.102}$$

取 λ 使得上式右端小于 $\delta < 0$，我们可以得到光滑随机凸优化问题的 (ϵ, δ) 复杂度上界为

$$\mathcal{O}\left[\frac{L\Omega^2}{\epsilon} + \frac{\Omega^2\sigma^2}{\epsilon^2}\ln^2\frac{1}{\delta}\right]. \tag{5.103}$$

5.4.2 非凸随机优化问题

这一节我们介绍随机梯度法在求解非凸随机优化问题时的复杂度分析，具体可参考文献 [15]，[16]。为了方便分析，我们将随机优化算法做一些变形。算法如下：

算法 5.6 随机迭代的随机梯度法（Randomized stochastic gradient(RSG) method）

输入：初始点 $x_1 \in \mathbb{R}^n$，最大迭代次数 N，步长序列 $\{\gamma_k\}$，随机迭代次数 R 的概率分布为 $P_R(\cdot)$。

1. 设 R 是概率分布为 $P_R(\cdot)$ 的随机变量，
2. 对 $k = 1, ..., R$，在 (x_k, ξ_k) 处调用子程序 $\mathcal{SFO}$ 得到随机梯度 $G(x_k, \xi_k)$，并更新

$$x_{k+1} = x_k - \gamma_k G(x_k, \xi_k). \tag{5.104}$$

可以看到随机迭代的随机梯度法与随机梯度法的区别在于其总迭代步数事先确定，且服从某一概率分布。随机迭代的随机梯度法收敛性结论如下：

定理 5.9 *选择步长 $\{\gamma_k\}$ 使得 $\gamma_k \leq 2/L$，且概率密度函数 $P_R(\cdot)$ 取，*

$$P_R(k) := \mathbf{Prob}\{R = k\} = \frac{2\gamma_k - L\gamma_k^2}{\sum_{k=1}^N (2\gamma_k - L\gamma_k^2)}. \tag{5.105}$$

在条件 A.6 下，

1. *对任意的 $N \geq 1$ 我们有*

$$\frac{1}{L}\mathbb{E}_{R,\xi_{[N]}}[\|\nabla f(x_R)\|^2] \leq \frac{D_f^2 + \sigma^2\sum_{k=1}^N \gamma_k^2}{\sum_{k=1}^N (2\gamma_k - L\gamma_k^2)}. \tag{5.106}$$

2. *若 f 是凸函数，则对任意 $N \geq 1$ 我们有，*

$$\mathbb{E}_{R,\xi_{[N]}}[f(x_R) - f^*] \leq \frac{D_X^2 + \sigma^2\sum_{k=1}^N \gamma_k^2}{\sum_{k=1}^N (2\gamma_k - L\gamma_k^2)}. \tag{5.107}$$

有上面的收敛性结论，我们可以得到随机迭代的随机梯度法在求解非凸光滑的随机优化问题时的复杂度。

推论 5.2 *对任意的常数 γ，取步长 $\{\gamma_k\}$，*

$$\gamma_k = \min\left\{\frac{1}{L}, \frac{\gamma}{\sigma\sqrt{N}}\right\}. \tag{5.108}$$

在条件 A.6 下，

1. 对任意的 $N \geq 1$ 我们有，

$$\frac{1}{L}\mathbb{E}[\|\nabla f(x_R)\|^2] \;\leq\; C(N) := \frac{LD_f^2}{N} + \left(\gamma + \frac{D_f^2}{\gamma}\right)\frac{\sigma}{\sqrt{N}}. \tag{5.109}$$

2. 若 f 是凸函数，则对任意 $N \geq 1$ 我们有，

$$\mathbb{E}[f(x_R) - f^*] \;\leq\; \frac{LD_X^2}{N} + \left(\gamma + \frac{D_X^2}{\gamma}\right)\frac{\sigma}{\sqrt{N}}. \tag{5.110}$$

利用 Markov 不等式 $\mathbb{E}[\|\nabla f(x_R)\|^2] \leq C(N)$，随机迭代的随机梯度法跑一次的 (ϵ, δ) 复杂度为

$$\mathbf{Prob}\left\{\|\nabla f(x_R)\|^2 \geq \frac{L^2D_f^2}{N}\lambda + \frac{2LD_f\sigma}{\sqrt{N}}\lambda\right\} \leq \frac{1}{\lambda}. \tag{5.111}$$

令 $\delta = 1/\lambda$ 并取

$$\frac{L^2D_f^2}{N}\lambda + \frac{2LD_f\sigma}{\sqrt{N}}\lambda \leq \epsilon. \tag{5.112}$$

则随机迭代的随机梯度法 (ϵ, δ) 复杂度至多为，

$$\mathcal{O}\left\{\frac{1}{\delta\epsilon} + \frac{\sigma^2}{\delta^2\epsilon^2}\right\}. \tag{5.113}$$

可以看到上述复杂度关于置信水平 δ 的结果不是很好。但是我们可以提高它。受到集成随机梯度法的启发，我们也可以对随机迭代的随机梯度法进行集成。

算法 5.7 集成随机迭代的随机梯度法（Two-phase RGS (2-RSG) method ）

输入：初始点 $x_1 \in \mathbb{R}^n$，集成次数 S，最大迭代次数 N，采样次数 T。

对 $d = 1, ..., S$ **执行迭代**

1. 调用 RSG 算法，其中输入为 x_1，迭代最大次数为 N，步长 $\{\gamma_k\}$，RSG 迭代次数 R 的概率分布为 $P_R(\cdot)$。令 $\bar{x}_s$ 表示 RSG 的输出。

输出：从 $\{\bar{x}_1, ..., \bar{x}_S\}$ 中选择 $\bar{x}^*$ 输出，满足，

$$\|g(\bar{x}^*)\| = \min_{s=1,...,S}\|g(\bar{x}_s)\|, \qquad g(\bar{x}_s) := \frac{1}{T}\sum_{k=1}^{T} G(\bar{x}_s, \xi_k). \tag{5.114}$$

定理 5.10 在条件 *A.6* 下，集成随机迭代的随机梯度法满足，

1. 令 $C(N)$ 按照推论 *5.2* 中定义，对任意的 $\lambda > 0$ 我们有，

$$\mathbf{Prob}\left\{\|\nabla f(\bar{x}^*)\|^2 \geq 2\left(4LC(N) + \frac{3\lambda\sigma^2}{T}\right)\right\} \leq \frac{S+1}{\lambda} + 2^{-S}. \tag{5.115}$$

2. 令 $\epsilon > 0$，$\delta \in (0, 1)$ 给定，若参数 (S, N, T) 取

$$S = S(\delta) := \left\lceil \log(\frac{2}{\delta}) \right\rceil.$$

$$N = N(\epsilon) := \left\lceil \max\left\{\frac{32L^2D_f^2}{\epsilon}, \left\lceil 32L\left(\gamma + \frac{D_f^2}{\gamma}\right)\frac{\sigma}{\epsilon}\right\rceil\right\}\right\rceil.$$

$$T = T(\epsilon, \delta) := \left\lceil \frac{24(S+1)\sigma^2}{\delta\epsilon} \right\rceil.$$

则集成随机迭代的随机梯度法为得到 (ϵ, δ) 近似解至多需要调用 $S(\delta)[N(\epsilon) + T(\epsilon, \delta)]$ 次 $\mathcal{SFO}$ 子程序。

综上，随机梯度法在求解光滑函数时有如下 (ϵ,δ) 复杂度结论：

- 随机迭代的随机梯度法在条件 A.6 下 (ϵ,δ) 复杂度为，

$$\mathcal{O}\left\{\frac{1}{\delta\epsilon}+\frac{\sigma^2}{\delta^2\epsilon^2}\right\}. \tag{5.116}$$

- 集成随机迭代的随机梯度法在条件 A.6 下 (ϵ,δ) 复杂度为，

$$\mathcal{O}\left\{\frac{\log(1/\delta)}{\epsilon}+\frac{\sigma^2}{\epsilon^2}\log\frac{1}{\delta}+\frac{\log^2(1/\delta)\sigma^2}{\delta\epsilon}\right\}. \tag{5.117}$$

- 集成随机迭代的随机梯度法在条件 A.6 和 A.7 下 (ϵ,δ) 复杂度为，

$$\mathcal{O}\left\{\frac{\log(1/\delta)}{\epsilon}+\frac{\sigma^2}{\epsilon^2}\log\frac{1}{\delta}+\frac{\log^2(1/\delta)\sigma^2}{\epsilon}\right\}. \tag{5.118}$$

6 结论

本论文通过对优化算法复杂度分析的研究，给出了不同优化算法在求解不同优化问题时效率的衡量准则，并比较了不同算法的收敛性和复杂度。

第二章介绍复杂度分析的理论基础，给出了复杂度分析里面的概念，度量以及具体例子，构建了复杂度分析的理论框架。后面的章节利用第二章的复杂度分析框架，给出了优化问题的复杂度上界以及复杂度下界。第三章介绍光滑优化问题的复杂度分析，给出了优化算法求解不同光滑函数集合时候的收敛性和复杂度分析。第四章介绍非光滑优化问题的复杂度分析。第五章介绍条件梯度法的复杂度分析，给出了优化问题在 $\mathcal{LO}$ 子程序下的复杂度下界，以及加速的条件梯度法。第六章介绍随机优化算法的复杂度分析，在第二章定义的随机优化问题解的度量下，衡量了随机优化算法置信水平的复杂度。

本论文对优化算法的收敛性和复杂度有了一个全面的论述和研究。并将优化算法的收敛性分析纳入一个统一的框架里面，即复杂度分析理论。但论文还有一些领域没有考虑进来，比如复合函数（$f(x)=g(x)+h(x)$ 其中 $g(x)$ 凸且光滑 $h(x)$ 凸但不一定光滑）的收敛性，鞍点问题的收敛性分析，二阶优化算法的收敛性分析。我们将在以后的工作中继续考虑这些问题。

参考文献

[1] S Damla Ahipaşaoğlu and Michael J Todd. A modified frank–wolfe algorithm for computing minimum-area enclosing ellipsoidal cylinders: Theory and algorithms. *Computational Geometry*, 46(5):494–519, 2013.

[2] Francis Bach, Simon Lacoste-Julien, and Guillaume Obozinski. On the equivalence between herding and conditional gradient algorithms. *arXiv preprint arXiv:1203.4523*, 2012.

[3] Amir Beck and Marc Teboulle. A conditional gradient method with linear rate of convergence for solving convex linear systems. *Mathematical Methods of Operations Research*, 59(2):235–247, 2004.

[4] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

[5] Amir Beck and Marc Teboulle. Gradient-based algorithms with applications to signal recovery. *Convex optimization in signal processing and communications*, pages 42–88, 2009.

[6] Ahron Ben-Tal and Arkadi Nemirovski. *Lectures on modern convex optimization: analysis, algorithms, and engineering applications*, volume 2. Siam, 2001.

[7] Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

[8] Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

[9] Kenneth L Clarkson. Coresets, sparse greedy approximation, and the frank-wolfe algorithm. *ACM Transactions on Algorithms (TALG)*, 6(4):63, 2010.

[10] Yuri Ermoliev. Stochastic quasigradient methods and their application to system optimization. *Stochastics: An International Journal of Probability and Stochastic Processes*, 9(1-2):1–36, 1983.

[11] Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval Research Logistics (NRL)*, 3(1-2):95–110, 1956.

[12] Robert M Freund and Paul Grigas. New analysis and results for the frank–wolfe method. *Mathematical Programming*, 155(1-2):199–230, 2016.

[13] Saeed Ghadimi and Guanghui Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization i: A generic algorithmic framework. *SIAM Journal on Optimization*, 22(4):1469–1492, 2012.

[14] Saeed Ghadimi and Guanghui Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, ii: shrinking procedures and optimal algorithms. *SIAM Journal on Optimization*, 23(4):2061–2089, 2013.

[15] Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

[16] Saeed Ghadimi and Guanghui Lan. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Mathematical Programming*, 156(1-2):59–99, 2016.

[17] Saeed Ghadimi, Guanghui Lan, and Hongchao Zhang. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, 155(1-2):267–305, 2016.

[18] Chonghai Hu, Weike Pan, and James T Kwok. Accelerated gradient methods for stochastic optimization and online learning. In *Advances in Neural Information Processing Systems*, pages 781–789, 2009.

[19] Martin Jaggi. *Sparse convex optimization methods for machine learning*. PhD thesis, ETH Zurich, 2011.

[20] Martin Jaggi. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *ICML (1)*, pages 427–435, 2013.

[21] Martin Jaggi, Marek Sulovsk, et al. A simple algorithm for nuclear norm regularized problems. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 471–478, 2010.

[22] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in neural information processing systems*, pages 315–323, 2013.

[23] Harold Kushner and G George Yin. *Stochastic approximation and recursive algorithms and applications*, volume 35. Springer Science & Business Media, 2003.

[24] Guanghui Lan. *Convex optimization under inexact first-order information*. PhD thesis, Georgia Institute of Technology, 2009.

[25] Guanghui Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1-2):365–397, 2012.

[26] Guanghui Lan. The complexity of large-scale convex programming under a linear optimization oracle. *arXiv preprint arXiv:1309.5550*, 2013.

[27] Guanghui Lan, Arkadi Nemirovski, and Alexander Shapiro. Validation analysis of mirror descent stochastic approximation method. *Math. Program.*, 134(2, Ser. A):425–458, 2012.

[28] Guanghui Lan and Yi Zhou. Conditional gradient sliding for convex optimization. *SIAM Journal on Optimization*, 26(2):1379–1409, 2016.

[29] Anatoly Yur'evich Levin. An algorithm for minimizing convex functions. In *Doklady Akademii Nauk*, volume 160, pages 1244–1247. Russian Academy of Sciences, 1965.

[30] Jeff Linderoth, Alexander Shapiro, and Stephen Wright. The empirical behavior of sampling methods for stochastic programming. *Annals of Operations Research*, 142(1):215–241, 2006.

[31] Wai-Kei Mak, David P Morton, and R Kevin Wood. Monte carlo bounding techniques for determining solution quality in stochastic programs. *Operations research letters*, 24(1-2):47–56, 1999.

[32] Ion Necoara, Yurii Nesterov, and Francois Glineur. Linear convergence of first order methods for non-strongly convex optimization. *Mathematical Programming*, pages 1–39, 2016.

[33] Arkadi Nemirovski. Information-based complexity of convex programming. *Lecture Notes*, 1995.

[34] Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

[35] Arkadi Nemirovski and Alexander Shapiro. Convex approximations of chance constrained programs. *SIAM Journal on Optimization*, 17(4):969–996, 2006.

[36] Arkadi Nemirovski and David Borisovich Yudin. *Problem complexity and method efficiency in optimization*. John Wiley & Sons, Inc., New York, 1983.

[37] Yu. Nesterov. Gradient methods for minimizing composite functions. *Math. Program.*, 140(1, Ser. B):125–161, 2013.

[38] Yurii Nesterov. A method for unconstrained convex minimization problem with the rate of convergence $\mathcal{O}(1/k^2)$. In *Doklady AN USSR*, volume 269, pages 543–547, 1983.

[39] Yurii Nesterov. Smooth minimization of non-smooth functions. *Mathematical programming*, 103(1):127–152, 2005.

[40] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

[41] Yurii Nesterov and J-Ph Vial. Confidence level solutions for stochastic programming. *Automatica*, 44(6):1559–1568, 2008.

[42] Donald J Newman. Location of the maximum on unimodal surfaces. *Journal of the ACM (JACM)*, 12(3):395–398, 1965.

[43] Boris T Polyak. New stochastic approximation type procedures. *Automat. i Telemekh*, 7(98-107):2, 1990.

[44] Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

[45] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

[46] Alexander Shapiro. Monte carlo sampling approach to stochastic programming. In *ESAIM: Proceedings*, volume 13, pages 65–73. EDP Sciences, 2003.

[47] Alexander Shapiro and Arkadi Nemirovski. On complexity of stochastic programming problems. In *Continuous optimization*, pages 111–146. Springer, 2005.

[48] Paul Tseng. On accelerated proximal gradient methods for convex-concave optimization. *Manuscript, University of Washington, Seattle*, May 2008.

[49] Bram Verweij, Shabbir Ahmed, Anton J Kleywegt, George Nemhauser, and Alexander Shapiro. The sample average approximation method applied to stochastic routing problems: a computational study. *Computational Optimization and Applications*, 24(2-3):289–333, 2003.