arXiv:2301.03139v1 [math.OC] 9 Jan 2023

一种基于牛顿共轭梯度的增广拉格朗日方法：以复杂度保证求解非凸等式约束优化的二阶驻点

Chuan He* Zhaosong Lu* Ting Kei Pong†

2022年4月10日（修订日期：2022年9月22日；2022年12月31日）

摘要

本文研究在已知近似可行点的情况下，如何求解非凸等式约束优化问题的二阶驻点（SOSP）。具体而言，我们首先提出一种新的牛顿共轭梯度方法，用于求解无约束优化问题的近似二阶驻点，并证明其复杂度显著优于牛顿共轭梯度方法 [56]。随后，我们提出一种基于牛顿共轭梯度的增广拉格朗日（AL）方法，用于求解非凸等式约束优化问题的近似二阶驻点，并采用所提出的牛顿共轭梯度方法作为子问题求解器。我们证明，在广义线性无关约束品性条件（GLICQ）下，所提出的增广拉格朗日方法以高概率求得非凸等式约束优化问题的（$\epsilon, \sqrt{\epsilon}$）-SOSP 时，总内层迭代复杂度为 $\widetilde{\mathcal{O}}(\epsilon^{-7/2})$，运算复杂度为 $\widetilde{\mathcal{O}}(\epsilon^{-7/2}\min\{n, \epsilon^{-3/4}\})$；这两项复杂度均显著优于近端增广拉格朗日方法 [60]所达到的结果。此外，我们证明，即使 GLICQ 不成立，其总内层迭代复杂度仍为 $\widetilde{\mathcal{O}}(\epsilon^{-11/2})$，运算复杂度为 $\widetilde{\mathcal{O}}(\epsilon^{-11/2}\min\{n, \epsilon^{-5/4}\})$。据我们所知，本文获得的所有复杂度结果，都是关于以高概率求解非凸等式约束优化问题近似二阶驻点的新结果。初步数值结果也表明，所提出的方法优于文献 [56, 60]中的方法。

关键词：非凸等式约束优化、二阶驻点、增广拉格朗日方法、牛顿共轭梯度法、迭代复杂度、运算复杂度

数学主题分类：49M15, 68Q25, 90C06, 90C26, 90C30, 90C60

1 引言

本文考虑非凸等式约束优化问题

$$\min_{x\in\mathbb{R}^n} \ f(x) \quad \text{s.t. } c(x) = 0, \tag{1}$$

其中 $f:\ \mathbb{R}^n \to \mathbb{R}$ 和 $c:\ \mathbb{R}^n \to \mathbb{R}^m$ 均为二阶连续可微函数，并且我们假设问题 (1) 至少有一个最优解。由于 (1) 是一个非凸优化问题，它可能存在许多局部但非全局的极小点，而找到其全局极小点通常是 NP难的。实践中通常转而寻找其一阶驻点（FOSP）。然而，仅有一阶驻点有时可能无法满足我们的需求，因此需要寻找二阶驻点（SOSP）。例如，在线性半定规划（SDP）中，一种强大的求解方法是求解一个与之等价的非凸

*工业与系统工程系，明尼苏达大学，美国（电子邮箱：he000233@umn.edu,zhaosong@umn.edu）。第二作者的研究部分得到了美国国家科学基金会资助项目 IIS-2211491 的支持。

†应用数学系，香港理工大学，中国香港特别行政区（电子邮箱： tk.pong@polyu.edu.hk）。该作者的研究部分得到了中国香港特别行政区研究资助局研究计划（项目编号：T22-504/21R）的支持。

等式约束优化问题 [17, 18]。 [18, 15] 中的研究表明，在一些温和条件下，后一个问题的二阶驻点可以给出线性 SDP 的最优解，而仅有一阶驻点通常无法做到这一点。因此，找到问题 (1) 的二阶驻点非常重要。

近年来，人们提出了许多具有复杂度保证的方法，用于寻找多种类型非凸优化问题的近似二阶驻点。例如，三次正则化牛顿法[52, 25, 1, 22], 加速梯度法 [23, 24], 信赖域法 [34, 35, 50], 二次正则化法 [12], 二阶线搜索法 [57], 以及牛顿共轭梯度（Newton-CG）法[56] 均是为非凸无约束优化问题开发的。此外，人们还提出了内点法 [8] 和对数障碍法 [54] ，用于求解带符号约束的非凸优化问题。文献 [8] 还将内点法推广至 [38] ，以求解带符号约束及额外线性等式约束的非凸优化问题。此外，文献 [47] 提出了一种带随机扰动的投影梯度下降法，用于求解带线性不等式约束的非凸优化问题。这些方法寻找近似二阶驻点的迭代复杂度均已得到确立。此外，文献 [1, 23, 34, 41, 24, 57, 22, 56]还研究了运算复杂度，该复杂度以梯度计算和矩阵-向量乘积等基本运算的数量来衡量。

已有多种方法被提出用于求解问题 (1) 的二阶驻点（SOSP），包括信赖域法 [21, 33], 序列二次规划法 [14],两阶段法 [9, 30, 32] 以及增广拉格朗日（AL）类方法 [4, 10, 58, 60] 。然而，其中只有少数方法对求解问题 (1) 的近似二阶驻点提供了复杂度保证。具体而言，非精确增广拉格朗日法 [58] 在二阶预言机调用次数方面具有最坏情况复杂度保证，但其运算复杂度（以梯度计算和海森-向量积等基本运算的数量衡量）尚不明确。据我们所知， [60] 中的近端增广拉格朗日方法似乎是目前唯一一种在基本运算数量方面，对求解问题 (1) 的近似二阶驻点具有最坏情况复杂度保证的方法。在该方法中，给定第 k 次迭代的迭代点 x^k 和乘子估计 λ^k ，通过求解近端增广拉格朗日子问题的一个近似随机二阶驻点，得到下一个迭代点 x^{k+1} ：

$$\min_{x\in\mathbb{R}^n} \ \mathcal{L}(x,\lambda^k;\rho)+\beta\|x-x^k\|^2/2$$

对某些适当的正数 ρ 和 β ，采用 [56], 中提出的牛顿共轭梯度法，其中 $\mathcal{L}$ 是问题 (1) 的增广拉格朗日函数，定义如下

$$\mathcal{L}(x,\lambda;\rho) := f(x)+\lambda^T c(x)+\rho\|c(x)\|^2/2.$$

随后，采用经典方案更新乘子估计，即 $\lambda^{k+1}=\lambda^k+\rho c(x^{k+1})$（例如，参见 [39, 55]）。 [60] 的作者研究了其近端增广拉格朗日方法的最坏情况复杂度，包括：(i) 总内层迭代复杂度，即衡量该方法中牛顿共轭梯度法 [56] 所执行的迭代总次数；(ii) 运算复杂度，即衡量该方法中计算增广拉格朗日函数的海森矩阵所涉及的梯度计算和矩阵向量积的总次数。在一些适当的假设下，包括所有迭代点处均满足广义线性无关约束品性条件（GLICQ）， [60] 证明了其近端增广拉格朗日方法以高概率找到问题 (1) 的 $(\epsilon,\sqrt{\epsilon})$-二阶驻点，其总内层迭代复杂度为 $\widetilde{\mathcal{O}}(\epsilon^{-11/2})$，运算复杂度为 $\widetilde{\mathcal{O}}(\epsilon^{-11/2}\min\{n,\epsilon^{-3/4}\})$。然而，这些复杂度与 [1, 24, 57, 56] 中的方法在高概率下找到非凸无约束优化问题的 $(\epsilon,\sqrt{\epsilon})$-二阶驻点所达到的迭代复杂度 $\widetilde{\mathcal{O}}(\epsilon^{-3/2})$ 和运算复杂度 $\widetilde{\mathcal{O}}(\epsilon^{-3/2}\min\{n,\epsilon^{-1/4}\})$ 之间存在很大差距；该无约束问题是问题 (1) 在 $c\equiv 0$ 时的一个特例。此外，当 GLICQ 不成立时，该近端增广拉格朗日方法缺乏复杂度保证。还需指出的是，研究者也开发了基于牛顿共轭梯度法的增广拉格朗日方法，以高效求解各种凸优化问题（例如，参见 [61, 62]），但其复杂度仍是未知的。

本文提出一种基于牛顿共轭梯度法的增广拉格朗日方法，以高概率求解问题 (1) 的近似二阶驻点，并研究在有无广义线性无关约束资格条件假设下的最坏情况复杂度。在

[1]事实上，文献 [60] 已证明，以高概率找到问题 (1) 的 (ϵ,ϵ)-二阶驻点的总内层迭代复杂度为 $\widetilde{\mathcal{O}}(\epsilon^{-7})$，运算复杂度为 $\widetilde{\mathcal{O}}(\epsilon^{-7}\min\{n,\epsilon^{-1}\})$；参见 [60, 定理 4(ii)、推论 3(ii)、定理 5]。尽管如此，对它们进行修改后，仍可获得前述复杂度，以高概率找到 (1) 的 $(\epsilon,\sqrt{\epsilon})$- 二阶驻点。

具体而言，我们证明，在广义线性无关约束资格条件成立时，该方法求得问题 (1) 的随机 $(\epsilon, \sqrt{\epsilon})$-二阶驻点所需的总内层迭代复杂度为 $\widetilde{\mathcal{O}}(\epsilon^{-7/2})$，运算复杂度为 $\widetilde{\mathcal{O}}(\epsilon^{-7/2}\min\{n, \epsilon^{-3/4}\})$；这显著优于前文提到的近端增广拉格朗日方法 [60] 所达到的复杂度。此外，当广义线性无关约束资格条件不成立时，我们证明该方法求得问题 (1) 的随机 $(\epsilon, \sqrt{\epsilon})$-二阶驻点所需的总内层迭代复杂度为 $\widetilde{\mathcal{O}}(\epsilon^{-11/2})$，运算复杂度为 $\widetilde{\mathcal{O}}(\epsilon^{-11/2}\min\{n, \epsilon^{-5/4}\})$，填补了该主题的研究空白。具体而言，我们的增广拉格朗日方法（算法 2）按以下方式进行。它不直接求解问题 (1)，而是求解问题 (1) 的一个扰动问题，其中 c 被其扰动对应项 $\tilde{c}$ 替代；该对应项利用问题 (1) 的近似可行点构造（详见 (25)）。在第 k 次迭代中，我们采用新提出的牛顿共轭梯度法（算法 1），针对罚参数 ρ_k 和截断拉格朗日乘子 λ^k，求得该扰动问题的增广拉格朗日子问题的一个近似随机二阶驻点 x^{k+1}。其中，截断拉格朗日乘子由经典方案 $\tilde{\lambda}^k = \lambda^{k-1} + \rho_k \tilde{c}(x^k)$ 得到的标准乘子估计 $\tilde{\lambda}^k$ 投影到欧几里得球上而得。[2] 随后，按照以下实用方案更新罚参数 ρ_{k+1}（例如，参见 [7, 第 4 节2]）：

$$\rho_{k+1} = \begin{cases} r\rho_k & \text{if } \|\tilde{c}(x^{k+1})\| > \alpha\|\tilde{c}(x^k)\|, \\ \rho_k & \text{otherwise} \end{cases}$$

对于某个 $r > 1$ 和 $\alpha \in (0, 1)$。需要指出的是，与经典增广拉格朗日方法不同，我们的方法有两个显著特点：（i）增广拉格朗日函数在迭代点处的取值有上界；（ii）与增广拉格朗日子问题相关的乘子估计有界。此外，为了以更好的复杂度保证求解增广拉格朗日子问题，我们提出了一种牛顿共轭梯度法的变体，用于求解无约束优化问题中的近似随机二阶驻点，其复杂度对目标函数海森矩阵的利普希茨常数的依赖显著低于 [56], 中的牛顿共轭梯度法，同时对容差参数的依赖阶数相同或更优。鉴于增广拉格朗日子问题中的这一利普希茨常数通常较大，从理论复杂度的角度看，我们的牛顿共轭梯度法（算法 1）比 [56] 中用于 [60] 近端增广拉格朗日方法的牛顿共轭梯度法更适合作为子问题求解器。

本文的主要贡献总结如下。

- 我们提出了一种新的牛顿共轭梯度法，用于求解无约束优化问题的近似二阶驻点，并证明该方法的迭代复杂度和运算复杂度对目标函数海森矩阵的利普希茨常数呈二次依赖，优于 [56], 中牛顿共轭梯度法所达到的三次依赖，同时对容差参数的依赖阶数不变或更优。此外，我们的复杂度结果是在以下假设下建立的：目标函数的海森矩阵在其某个水平集的凸邻域内利普希茨连续。此假设弱于 [56] 中的牛顿共轭梯度法及其他一些方法（例如 [12, 35]）通常采用的假设；后者要求目标函数的海森矩阵在包含该邻域以及这些方法在线搜索或信赖域步中产生的所有试探点的凸集内利普希茨连续（更详细的讨论见第 3 节）。

- 我们提出了一种基于牛顿共轭梯度法的增广拉格朗日方法，以高概率求解非凸等式约束优化问题 (1) 的近似二阶驻点，并研究了在有无广义线性无关约束资格条件假设下该方法的最坏情况复杂度。在此之前，尚无研究在不施加广义线性无关约束资格条件的情况下，求解问题 (1) 近似二阶驻点的复杂度。此外，在广义线性无关约束资格条件及其他一些适当假设下，我们证明，该方法以高概率求解问题 (1) 的 $(\epsilon, \sqrt{\epsilon})$-二阶驻点时，总内层迭代复杂度为 $\widetilde{\mathcal{O}}(\epsilon^{-7/2})$，运算复杂度为 $\widetilde{\mathcal{O}}(\epsilon^{-7/2}\min\{n, \epsilon^{-3/4}\})$，显著优于文献 [60] 中近端增广拉格朗日方法分别达到的复杂度 $\widetilde{\mathcal{O}}(\epsilon^{-11/2})$ 和 $\widetilde{\mathcal{O}}(\epsilon^{-11/2}\min\{n, \epsilon^{-3/4}\})$。据我们所知，本文获得的所有复杂度结果，都是关于以高概率求解非凸等式约束优化问题近似二阶驻点的新结果。

[2]相关文献中，也将通过投影 $\tilde{\lambda}^k$ 到一个紧集上而得到的 λ^k 称为受保护的拉格朗日乘子 [11, 42, 13], ，并已证明其具有许多实践和理论上的优势（讨论见 [11] ）。

为便于比较，我们在表1中总结了我们的方法与文献 [60] 中的近端增广拉格朗日方法，在假设或不假设广义线性无关约束资格条件时，求解问题 (1) 的随机 $(\epsilon,\sqrt{\epsilon})$-二阶驻点所需的总内层迭代复杂度和运算复杂度。

表1：T　寻找（$\epsilon,\sqrt{\epsilon}$）-随机二阶驻点的总内层迭代次数和运算复杂度，用于解决 (　　1).

方法	广义线性无关约束资格条件	总内层迭代复杂度	运算复杂度
近端增广拉格朗日方法 [60]	✓	$\widetilde{\mathcal{O}}(\epsilon^{-11/2})$	$\widetilde{\mathcal{O}}(\epsilon^{-11/2}$ 最小值$\{n,\epsilon^{-3/4}\})$
近端增广拉格朗日方法 [60]	✗	未知	未知
我们的增广拉格朗日方法	✓	$\widetilde{\mathcal{O}}(\epsilon^{-7/2})$	$\widetilde{\mathcal{O}}(\epsilon^{-7/2}$ 最小值$\{n,\epsilon^{-3/4}\})$
我们的增广拉格朗日方法	✗	$\widetilde{\mathcal{O}}(\epsilon^{-11/2})$	$\widetilde{\mathcal{O}}(\epsilon^{-11/2}$ 最小$\{n,\epsilon^{-5/4}\})$

需要指出的是，除 [60] 之外，还有许多研究工作考察了非凸约束优化中增广拉格朗日方法的复杂度。然而，这些工作旨在寻找问题的一阶驻点近似解，而不是二阶驻点近似解（例如，参见 [40, 37, 13, 51, 45]）。由于我们主要关注利用增广拉格朗日方法寻找近似二阶驻点的复杂度，因此未在上表中列出这些工作进行比较。

本文其余部分的组织结构如下。第 2 节介绍一些记号和最优性条件。第 3 节提出一种用于无约束优化的牛顿共轭梯度法，并研究其最坏情况复杂度。第 4 节针对问题（1）提出一种基于牛顿共轭梯度法的增广拉格朗日方法，并研究其最坏情况复杂度。我们分别在第 5 节和第 6 节给出数值结果以及主要结果的证明。第 7 节讨论一些未来的研究方向。

2 记号与预备知识

在本文中，我们用 $\mathbb{R}^n$ 表示 n 维欧几里得空间。我们用 $\|\cdot\|$ 表示向量的欧氏范数或矩阵的谱范数。对于实对称矩阵 H，我们用 $\lambda_{\min}(H)$ 表示其最小特征值。以原点为中心、半径为 $R\geq 0$ 的欧几里得球记作 $\mathcal{B}_R:=\{x:\|x\|\leq R\}$，并用 $\Pi_{\mathcal{B}_R}(v)$ 表示向量 v 在 $\mathcal{B}_R$上的欧氏投影。对于给定的有限集 $\mathcal{A}$，我们用 $|\mathcal{A}|$ 表示其基数。对于任意 $s\in\mathbb{R}$，若 $s\geq 0$，则令 sgn(s) 为 1；否则令其为 -1。此外，$\mathcal{O}(\cdot)$ 表示省略对数项后的 $\mathcal{O}(\cdot)$。

假设 x^* 是问题 (1) 的局部极小值点，且线性无关约束规格在 x^*处成立，即 $\nabla c(x^*):=[\nabla c_1(x^*)\ \ \nabla c_2(x^*)\ \ \cdots\ \ \nabla c_m(x^*)]$ 列满秩。则存在一个拉格朗日乘子 $\lambda^*\in\mathbb{R}^m$，使得

$$\nabla f(x^*)+\nabla c(x^*)\lambda^*=0, \tag{2}$$

$$d^T\left(\nabla^2 f(x^*)+\sum_{i=1}^m\lambda_i^*\nabla^2 c_i(x^*)\right)d\geq 0,\quad \forall d\in\mathcal{C}(x^*), \tag{3}$$

其中 $\mathcal{C}(\cdot)$定义为

$$\mathcal{C}(x):=\{d\in\mathbb{R}^n:\nabla c(x)^T d=0\}. \tag{4}$$

关系式 (2) 和 (3) 在文献中分别称为问题 (1) 的一阶和二阶最优性条件（例如，参见 [53]）。请注意，一般而言，不可能找到一个能够精确满足 (2) 和 (3) 的点。因此，我们转而关注满足其近似形式的点。具体而言，我们给出近似一阶驻点（FOSP）和二阶驻点（SOSP）的以下定义，这些定义与 [4, 10, 60] 中所考虑的定义类似。它们的合理性可通过对约束优化的序列最优性条件的研究来论证[3, 4]。

定义 2.1（ϵ_1-一阶驻点）。设 $\epsilon_1>0$。如果 $x\in\mathbb{R}^n$ 与某个 $\lambda\in\mathbb{R}^m$ 一起满足以下条件，我们称其为问题 (1) 的 ϵ_1-一阶驻点（ϵ_1-FOSP）：

$$\|\nabla f(x)+\nabla c(x)\lambda\|\leq\epsilon_1,\quad \|c(x)\|\leq\epsilon_1. \tag{5}$$

定义 2.2（(ϵ_1,ϵ_2)-二阶驻点）。设 $\epsilon_1,\epsilon_2>0$。如果 $x\in\mathbb{R}^n$ 与某个 $\lambda\in\mathbb{R}^m$ 一起满足 (5) 以及以下附加条件，我们称其为问题 (1) 的 (ϵ_1,ϵ_2)-二阶驻点（(ϵ_1,ϵ_2)-SOSP）：

$$d^T\left(\nabla^2 f(x)+\sum_{i=1}^m \lambda_i\nabla^2 c_i(x)\right)d\geq-\epsilon_2\|d\|^2,\quad \forall d\in\mathcal{C}(x), \tag{6}$$

其中 $\mathcal{C}(\cdot)$ 的定义见式 (4)。

3 一种用于无约束优化的牛顿共轭梯度法

本节提出一种牛顿共轭梯度法的变体 [56, Algorithm 3]，用于求解一类无约束优化问题的近似二阶驻点；该变体将在下一节提出的增广拉格朗日方法中用作子问题求解器。具体而言，我们考虑如下无约束优化问题

$$\min_{x\in\mathbb{R}^n} F(x), \tag{7}$$

其中函数 F 满足以下假设。

假设 3.1. (a) 水平集 $\mathscr{L}_F(u^0):=\{x:\ F(x)\leq F(u^0)\}$ 对某个 $u^0\in\mathbb{R}^n$ 是紧集。

(b) 函数 F 在 $\mathscr{L}_F(u^0)$ 的一个凸开邻域（记为 Ω）内二阶 Lipschitz 连续可微，也就是说，存在 $L_H^F>0$ 使得

$$\|\nabla^2F(x)-\nabla^2F(y)\|\leq L_H^F\|x-y\|,\quad \forall x,y\in\Omega. \tag{8}$$

根据假设 3.1，存在 $F_{\rm low}\in\mathbb{R}$、$U_g^F>0$ 和 $U_H^F>0$，使得

$$F(x)\geq F_{\rm low},\quad \|\nabla F(x)\|\leq U_g^F,\quad \|\nabla^2F(x)\|\leq U_H^F,\quad \forall x\in\mathscr{L}_F(u^0). \tag{9}$$

近期，有人提出了一种牛顿共轭梯度法 [56, 算法 3]，用于求解问题 (7) 的近似随机二阶驻点。该方法不仅易于实现，而且具有一个优点：其主要计算仅包括与函数 F 相关的梯度计算和海森-向量积。假设 ∇^2F 在一个包含 $\mathscr{L}_F(u^0)$ 以及该方法线搜索步骤中产生的所有试探点的凸开集上利普希茨连续（见 [56, 假设 2]），则已在 [56,定理 4、推论 2] 中证明，该方法求解 (7) 的随机 (ϵ_g,ϵ_H)-二阶驻点的迭代复杂度和运算复杂度（即，找到一个点 x，使得 $\|\nabla F(x)\|\leq\epsilon_g$ 确定性成立，且 $\lambda_{\min}(\nabla^2F(x))\geq-\epsilon_H$ 以高概率成立）为

$$\mathcal{O}((L_H^F)^3\max\{\epsilon_g^{-3}\epsilon_H^3,\epsilon_H^{-3}\})\quad\text{and}\quad\widetilde{\mathcal{O}}((L_H^F)^3\max\{\epsilon_g^{-3}\epsilon_H^3,\epsilon_H^{-3}\}\min\{n,(U_H^F/\epsilon_H)^{1/2}\}), \tag{10}$$

分别地，其中 $\epsilon_g,\epsilon_H\in(0,1)$ 是给定容差。然而，这一假设可能难以检验，因为这些试探点在方法终止前是未知的，而且原点与这些点之间的距离取决于 $\mathcal{O}(\epsilon_H^{-1})$ 中的容差 ϵ_H（见 [56, 引理 3]）。此外，如式 (10) 所示，[56] 中的牛顿共轭梯度法的迭代复杂度和运算复杂度均以三次方依赖于 L_H^F。请注意，L_H^F 有时可能非常大。例如，算法 2 中出现的增广拉格朗日子问题满足$L_H^F=\mathcal{O}(\epsilon_1^{-2})$ 或 $\mathcal{O}(\epsilon_1^{-1})$，其中 $\epsilon_1\in(0,1)$ 是问题 (1) 的给定容差（见第 4 节）。从理论复杂度的角度看，对 L_H^F 的三次方依赖使得这类牛顿共轭梯度法不适合作为增广拉格朗日子问题求解器。

在本节余下部分，我们提出一种牛顿共轭梯度法的变体 [56, 算法 3]，并证明在假设 3.1 下，该方法的迭代复杂度和运算复杂度为

$$\mathcal{O}((L_H^F)^2\max\{\epsilon_g^{-2}\epsilon_H,\epsilon_H^{-3}\})\quad\text{and}\quad\widetilde{\mathcal{O}}((L_H^F)^2\max\{\epsilon_g^{-2}\epsilon_H,\epsilon_H^{-3}\}\min\{n,(U_H^F/\epsilon_H)^{1/2}\}), \tag{11}$$

分别用于求解问题 (7) 的随机 (ϵ_g,ϵ_H)-二阶驻点。这些复杂度显著优于 [56] 中牛顿共轭梯度法达到的 (10) 中的复杂度。事实上，(11) 中的复杂度对 L_H^F 呈二次依赖，而 (10) 中的复杂度对 L_H^F 呈三次依赖。此外，可以验证，它们对 ϵ_g 和 ϵ_H 的依赖阶数与 (10) 中给出的相比有所改善或保持不变。

3.1 牛顿共轭梯度法的主要组成部分

本小节简要讨论 [56]，中牛顿共轭梯度法的两个主要组成部分；下一小节将利用它们提出该方法的一种变体，以求解问题 (7) 的近似随机二阶驻点。

[56] 中的牛顿共轭梯度法的第一个主要组成部分是截断共轭梯度方法 [56, 算法 1]，它是一种经过修改的共轭梯度方法，用于求解可能是不定的线性方程组

$$(H + 2\varepsilon I)d = -g, \tag{12}$$

其中，$0 \neq g \in \mathbb{R}^n$、$\varepsilon > 0$和 $H \in \mathbb{R}^{n\times n}$ 是对称矩阵。该截断共轭梯度方法会在有限次迭代内终止。它输出一个近似解 d 使其满足 $\|(H+2\varepsilon I)d+g\| \leq \widehat{\zeta}\|g\|$ 和对于某个 $\widehat{\zeta} \in [0,1)$，或者输出一个具有 $d^T H d \geq -\varepsilon\|d\|^2$ 的、足够负的曲率方向 d。[56] 中的牛顿共轭梯度法的第二个主要组成部分是一个最小特征值预言机，它要么产生一个满足 $\|v\| = 1$ 和 $v^T H v \leq -\varepsilon/2$的 H 的足够负的曲率方向 v，要么以高概率证明 $\lambda_{\min}(H) \geq -\varepsilon$ 成立。为便于参考，我们分别在附录 A 和 B 中的算法 3 和 4 介绍这两个组成部分。

算法 1 求解问题 (7) 的牛顿共轭梯度法

Input：容差 ϵ_g、ϵ_H $\in (0,1)$，回溯比率 $\theta \in (0,1)$，初始点 u^0，共轭梯度精度参数 $\zeta \in (0,1)$，线搜索参数 $\eta \in (0,1)$，概率参数 $\delta \in (0,1)$。

设置 $x^0 = u^0$；

对于 $t = 0, 1, 2, \ldots$ 执行

若 $\|\nabla F(x^t)\| > \epsilon_g$ 则

调用算法 3，输入 $H = \nabla^2 F(x^t)$、$\varepsilon = \epsilon_H$、$g = \nabla F(x^t)$、精度参数 ζ 和 $U = 0$，得到输出 d，d_类型；

若 d_类型=负曲率，则

$$d^t \leftarrow -\operatorname{sgn}(d^T \nabla F(x^t)) \frac{|d^T \nabla^2 F(x^t) d|}{\|d\|^3} d; \tag{13}$$

否则 {d_类型=SOL}

$$d^t \leftarrow d; \tag{14}$$

结束 if

转到线搜索；

else

调用算法 4，输入 $H = \nabla^2 F(x^t)$、$\varepsilon = \epsilon_H$ 和概率参数 δ；

如果算法 4 证实 $\lambda_{\min}(\nabla^2 F(x^t)) \geq -\epsilon_H$ 则

输出 x^t 并终止；

否则 {由算法 4}返回的充分负曲率方向 v

设置 d_类型为=负曲率，并且

$$d^t \leftarrow -\operatorname{sgn}(v^T \nabla F(x^t)) |v^T \nabla^2 F(x^t) v| v; \tag{15}$$

转到线搜索；

结束 if

结束 if

线搜索：

若 d_类型为=SOL，则

求 $\alpha_t = \theta^{j_t}$，其中 j_t 是满足以下条件的最小非负整数 j

$$F(x^t + \theta^j d^t) < F(x^t) - \eta \epsilon_H \theta^{2j} \|d^t\|^2; \tag{16}$$

否则 {d_类型为=负曲率}

求 $\alpha_t = \theta^{j_t}$，其中 j_t 是满足以下条件的最小非负整数 j

$$F(x^t + \theta^j d^t) < F(x^t) - \eta \theta^{2j} \|d^t\|^3/2; \tag{17}$$

结束 if

$x^{t+1} = x^t + \alpha_t d^t$；

结束循环

3.2 问题 (7) 的牛顿共轭梯度法

本小节提出算法 1 中的牛顿共轭梯度法，该方法是用于求解问题 (7) 的近似随机二阶驻点的牛顿共轭梯度法 [56, 算法 3]，的一种变体。

我们的牛顿共轭梯度法（算法 1）遵循与 [56, 算法 3] 相同的框架。具体而言，在每次迭代中，如果当前迭代点处 F 的梯度不够小，则调用截断共轭梯度方法（算法 3）求解阻尼牛顿系统，以获得一个下降方向；随后沿该方向进行线搜索，从而使 F 充分下降。否则，当前迭代点已经是 (7) 的近似一阶驻点，此时调用最小特征值预言机（算法 4）：它要么为 F 产生一个足够负曲率的方向，并通过随后沿该方向进行的线搜索使 F 充分下降；要么以高概率确认当前迭代点是 (7) 的近似二阶驻点，并终止算法。有关该框架的更多细节，请参见 [56]。

尽管采用相同的框架，我们的牛顿共轭梯度法与 [56, 算法 3] 使用了不同的线搜索准则。具体而言，我们的牛顿共轭梯度法采用了源自 [59], 的混合线搜索准则，该准则结合了二次下降准则 (16) 和三次下降准则 (17)。具体来说，当搜索方向属于“SOL”类型时，采用二次下降准则 (16)；而当搜索方向属于“NC”类型时，则采用三次下降准则 (17)。[3] 相比之下，[56] 中的牛顿共轭梯度法始终采用三次下降准则，不考虑搜索方向的类型。从下文的定理 3.1 可以看出，我们的牛顿共轭梯度法实现了 (11) 中给出的迭代复杂度和运算复杂度；就对 L_H^F的依赖阶而言，这优于 [56, 算法 3] 所实现的 (10) 中的复杂度，同时对 ϵ_g 和 ϵ_H 的依赖阶与 (10) 中给出的结果相比有所改善或保持不变。因此，从理论复杂度的角度看，作为第 4 节提出的增广拉格朗日方法的增广拉格朗日子问题求解器，我们的牛顿共轭梯度法比 [56, 算法 3] 更具吸引力。

以下定理给出了算法 1 的迭代和运算复杂度，其证明将在第 6.1 节中给出。

定理 3.1. 假设假设 3.1 成立。令

$$T_1 := \left\lceil \frac{F_{\rm hi} - F_{\rm low}}{\min\{c_{\rm sol}, c_{\rm nc}\}} \max\{\epsilon_g^{-2}\epsilon_H, \epsilon_H^{-3}\} \right\rceil + \left\lceil \frac{F_{\rm hi} - F_{\rm low}}{c_{\rm nc}} \epsilon_H^{-3} \right\rceil + 1,\ T_2 := \left\lceil \frac{F_{\rm hi} - F_{\rm low}}{c_{\rm nc}} \epsilon_H^{-3} \right\rceil + 1, \tag{18}$$

其中，$F_{\rm hi} = F(u^0)$、$F_{\rm low}$ 在（9）中给出，并且

$$c_{\rm sol} := \eta \min\left\{ \left[\frac{4}{4 + \zeta + \sqrt{(4+\zeta)^2 + 8L_H^F}} \right]^2, \left[\frac{\min\{6(1-\eta), 2\}\theta}{L_H^F} \right]^2 \right\}, \tag{19}$$

$$c_{\rm nc} := \frac{\eta}{16} \min\left\{ 1, \left[\frac{\min\{3(1-\eta), 1\}\theta}{L_H^F} \right]^2 \right\}. \tag{20}$$

则以下结论成立。

（i）算法 1 中调用算法 4 的总次数至多为 T_2。

（ii）算法 1 中调用算法 3 的总次数至多为 T_1。

（iii）（迭代复杂度）算法 1 至多经过 $T_1 + T_2$ 次迭代即可终止，且

$$T_1 + T_2 = \mathcal{O}((F_{\rm hi} - F_{\rm low})(L_H^F)^2 \max\{\epsilon_g^{-2}\epsilon_H, \epsilon_H^{-3}\}). \tag{21}$$

此外，其输出 x^t 确定性地满足 $\|\nabla F(x^t)\| \le \epsilon_g$，并且以至少 $1-\delta$ 的概率满足 $\lambda_{\min}(\nabla^2 F(x^t)) \ge -\epsilon_H$，其中 $0 \le t \le T_1 + T_2$ 为某个值。

（iv）（运算复杂度）算法 1 至多需要

$$\widetilde{\mathcal{O}}((F_{\rm hi} - F_{\rm low})(L_H^F)^2 \max\{\epsilon_g^{-2}\epsilon_H, \epsilon_H^{-3}\} \min\{n, (U_H^F/\epsilon_H)^{1/2}\})$$

次矩阵-向量乘积，其中 U_H^F 由式 (9) 给出。

[3] SOL 和 NC 分别代表“近似解”和“负曲率”。

4 求解问题 (1) 的牛顿共轭梯度型增广拉格朗日方法

本节提出一种基于牛顿共轭梯度法的增广拉格朗日方法，用于在任意给定容差 ϵ_1、$\epsilon_2 \in (0, 1)$ 下求解问题 (1) 的随机 (ϵ_1, ϵ_2)-二阶驻点。在继续讨论之前，我们对问题 (1) 作出一些额外假设。

假设 4.1. (a) 已知问题 (1) 的一个 $\epsilon_1/2$-近似可行点 z_{ϵ_1}，即满足 $\|c(z_{\epsilon_1})\| \leq \epsilon_1/2$。

(b) 存在常数 $f_{\rm hi}$、$f_{\rm low}$ 和 $\gamma > 0$，它们与 ϵ_1 和 ϵ_2 无关，使得

$$f(z_{\epsilon_1}) \leq f_{\rm hi}, \tag{22}$$
$$f(x) + \gamma \|c(x)\|^2/2 \geq f_{\rm low}, \quad \forall x \in \mathbb{R}^n, \tag{23}$$

其中 z_{ϵ_1} 由 (a) 给出。

(c) 存在某些 δ_f、$\delta_c > 0$，使得集合

$$\mathcal{S}(\delta_f, \delta_c) := \{x : f(x) \leq f_{\rm hi} + \delta_f,\ \|c(x)\| \leq 1 + \delta_c\} \tag{24}$$

是紧的，其中 $f_{\rm hi}$ 如上所述。此外，$\nabla^2 f$ 和 $\nabla^2 c_i$、$i = 1$、2、.....、m 在 $\mathcal{S}(\delta_f, \delta_c)$ 的一个凸开邻域 $\Omega(\delta_f, \delta_c)$ 内利普希茨连续。

现在我们对假设 4.1 作一些说明。

注 4.1. (i) *[31, 37, 49, 60]*中考虑过与假设 4.1(a) 非常相似的假设。通过施加假设 4.1(a)，我们将研究范围限制在这样的问题 (1) 上：可以通过成本低廉的步骤找到一个 $\epsilon_1/2$-近似可行点 z_{ϵ_1}。这类问题的一个例子是，存在 v^0 使得 $\{x : \|c(x)\| \leq \|c(v^0)\|\}$ 为紧集，$\nabla^2 c_i, 1 \leq i \leq m$在该集合的凸邻域上是 Lipschitz 连续的，并且该集合上满足 LICQ。事实上，对于这类问题实例，通过将我们的牛顿共轭梯度法（算法 1）应用于问题 $\min_{x\in\mathbb{R}^n} \|c(x)\|^2$，即可计算出满足 $\|c(z_{\epsilon_1})\| \leq \epsilon_1/2$ 的点 z_{ϵ_1}。由定理 3.1 可知，算法 1 找到此类 z_{ϵ_1} 所需的迭代复杂度和运算复杂度分别为 $\mathcal{O}(\epsilon_1^{-3/2})$ 和 $\widetilde{\mathcal{O}}(\epsilon_1^{-3/2} \min\{n, \epsilon_1^{-1/4}\})$；与我们的增广拉格朗日方法的复杂度相比（见下文定理 4.4 和 4.5），这些复杂度可以忽略不计。另一个例子是，当标准误差界条件 $\|c(x)\|^2 = \mathcal{O}(\|\nabla(\|c(x)\|^2)\|^\nu)$在某个 $\nu > 0$的水平集上成立时，可以通过对问题 $\min_{x\in\mathbb{R}^n} \|c(x)\|^2$ 应用梯度法来找到上述 z_{ϵ_1}（例如，参见 *[46, 58]*）。此外，下文提出的基于牛顿共轭梯度法的增广拉格朗日方法（算法 2）是一种旨在寻找二阶驻点的二阶方法。通常，它比一阶方法开销更大。为了在实践中充分发挥这类增广拉格朗日方法的作用，自然可以先运行一阶方法，得到一个 $\epsilon_1/2$-一阶驻点 z_{ϵ_1}，然后将 z_{ϵ_1} 作为 $\epsilon_1/2$- 近似可行点来运行增广拉格朗日方法。因此，只要能通过一阶方法找到问题 (1) 的一个 $\epsilon_1/2$- 一阶驻点，假设 4.1(a) 在实践中便可满足。

(ii) 假设 4.1(b) 较为温和。特别地，若对于所有满足 x with$\|c(x)\| \leq 1$ 的 $f(x)$，$\leq f_{\rm hi}$ holds，则假设 (22) 成立；这一条件在 *[60,* 假设 *3]* 中有所采用。若问题 (1) 有一个已知可行点，该假设也成立；在设计用于求解非凸约束优化的增广拉格朗日方法时，通常会采用这一条件（例如，见 *[49, 31, 48, 37]*）。此外，假设 (23) 意味着当相应的罚参数足够大时，二次罚函数有下界；在研究用于求解问题 (1) 的二次罚函数和增广拉格朗日方法时，通常会用到这一性质（例如，见 *[40, 37, 60, 43]*）。显然，当 $\inf_{x\in\mathbb{R}^n} f(x) > -\infty$ 时，可以看出 (23) 对任意 $\gamma > 0$ 都成立。一般而言，识别 γ 的一种可能方法是，应用文献中提出的不可行性检测技术（例如，*[20, 19, 6]*），检查水平集 $\{x : f(x)+\gamma\|c(x)\|^2/2 \leq \tilde{f}_{\rm low}\}$ 在某个足够小的 $\tilde{f}_{\rm low}$ 下是否可行。注意，若该水平集对于某个 $\tilde{f}_{\rm low}$ 不可行，则意味着对于给定的 γ 和 $f_{\rm low} = f_{\rm low}$，(23) 成立。

(iii) 假设 4.1(c) 的限制性不强。事实上，如果 f 或 $f(\cdot)+\gamma\|c(\cdot)\|^2/2$ 是水平有界的，则集合 $\mathcal{S}(\delta_f,\delta_c)$ 是紧的。后一水平有界性假设在研究增广拉格朗日方法时通常会采用（例如，见 *[37, 60]*），且它比我们的假设更强。

接下来，我们在算法 2 中提出一种基于牛顿共轭梯度法的增广拉格朗日方法，在假设 4.1 下求解问题 (1) 的随机 (ϵ_1,ϵ_2)-二阶驻点。该方法不直接求解问题 (1)，而是求解以下扰动问题：

$$\min_{x\in\mathbb{R}^n} \ f(x) \quad \text{s.t. } \tilde{c}(x) := c(x) - c(z_{\epsilon_1}) = 0, \tag{25}$$

其中，z_{ϵ_1} 由假设 4.1(a) 给出。具体而言，在第 k 次迭代中，该方法应用牛顿共轭梯度法（算法 1），求解与 (25) 对应的增广拉格朗日子问题，以找到其近似随机二阶驻点 x^{k+1}：

$$\min_{x\in\mathbb{R}^n} \left\{\widetilde{\mathcal{L}}(x,\lambda^k,\rho_k) := f(x) + (\lambda^k)^T\tilde{c}(x) + \rho_k\|\tilde{c}(x)\|^2/2\right\} \tag{26}$$

使得 $\widetilde{\mathcal{L}}(x^{k+1},\lambda^k;\ \rho_k)$ 低于某个阈值（见 (27) 和 (28)），其中 λ^k 是截断拉格朗日乘子，即将标准乘子估计 $\tilde{\lambda}^k$ 投影到欧几里得球后得到的乘子（见算法 2 的步骤 6）。随后，按照算法 2 步骤 4 中所述的经典方案更新标准乘子估计 $\tilde{\lambda}^{k+1}$。最后，根据约束违反度的改善情况自适应地更新罚参数 ρ_{k+1}（见算法 2 的步骤 7）。文献中经常采用这种实用的更新方案（例如，见 [7, 2, 31]）。

需要指出的是，截断拉格朗日乘子序列 $\{\lambda^k\}$ 用于算法 2 的增广拉格朗日子问题中，并且是有界的；而标准拉格朗日乘子序列 $\{\tilde{\lambda}^k\}$ 用于经典增广拉格朗日方法的相应子问题中，可能是无界的。因此，算法 2 可视为一种带保护机制的增广拉格朗日方法。文献中已经使用截断拉格朗日乘子来设计一些增广拉格朗日方法 [2, 11, 42, 13], ，并且它们将在后续对算法 2 的复杂度分析中发挥关键作用。

算法 2 基于牛顿共轭梯度法的增广拉格朗日方法求解问题 (1)

设 γ 由假设 4.1 给出。输入：ϵ_1、$\epsilon_2\in(0,1)$、$\Lambda>0$、$x^0\in\mathbb{R}^n$、$\lambda^0\in\mathcal{B}_\Lambda$、$\rho_0>2\gamma$、$\alpha\in(0,1)$、$r>1$、$\delta\in(0,1)$，以及由假设 4.1 给出的 z_{ϵ_1}。1: 设置 $k=0$。2: 设置 $\tau_k^g=\max\{\epsilon_1, r^k{}_{\log}\epsilon_1/{}_{\log}{}^2\}$，并设置 $\tau_k^H=\max\{\epsilon_2, r^k{}_{\log}\epsilon_2/{}_{\log}{}^2\}$。3: 调用算法 1，利用 $\epsilon_g=\tau_k^g$、$\epsilon_H=\tau_k^H$ 和 $u^0=x_{\text{init}}^k$，求得问题 $\min_{x\in\mathbb{R}^n}\widetilde{\mathcal{L}}(x,\lambda^k;\ \rho_k)$ 的一个近似解 x^{k+1}，使得

$$\widetilde{\mathcal{L}}(x^{k+1},\lambda^k;\rho_k)\le f(z_{\epsilon_1}),\ \|\nabla_x\widetilde{\mathcal{L}}(x^{k+1},\lambda^k;\rho_k)\|\le\tau_k^g, \tag{27}$$

$$\lambda_{\min}(\nabla_{xx}^2\widetilde{\mathcal{L}}(x^{k+1},\lambda^k;\rho_k))\ge-\tau_k^H \text{ with probability at least } 1-\delta, \tag{28}$$

其中

$$x_{\text{init}}^k=\begin{cases} z_{\epsilon_1} & \text{if } \widetilde{\mathcal{L}}(x^k,\lambda^k;\rho_k)>f(z_{\epsilon_1}),\\ x^k & \text{otherwise,}\end{cases}\qquad \text{for } k\ge0. \tag{29}$$

4: 设置 $\tilde{\lambda}^{k+1}=\lambda^k+\rho_k\tilde{c}(x^{k+1})$。5: 如果 $\tau_k^g\le\epsilon_1$、$\tau_k^H\le\epsilon_2$ 和 $\|c(x^{k+1})\|\le\epsilon_1$，则输出 $(x^{k+1},\tilde{\lambda}^{k+1})$ 并终止。6: 设置 $\lambda^{k+1}=\Pi_{\mathcal{B}_\Lambda}(\tilde{\lambda}^{k+1})$。7: 如果 $k=0$ 或 $\|\tilde{c}(x^{k+1})\|>\alpha\|\tilde{c}(x^k)\|$，则设置 $\rho_{k+1}=r\rho_k$。否则，设置 $\rho_{k+1}=\rho_k$。8: 设置 $k\leftarrow k+1$，并转到步骤 2。

注4.2. (i) 注意，算法 2 的初始点 x_{init}^0 可以不同于 z_{ϵ_1}，而且可能相当不可行，尽管 z_{ϵ_1} 是问题 (1) 的近似可行点。此外，z_{ϵ_1} 用于确保算法 2 收敛。具体而言，如果算法遇到“严重不可行点” x^k，即满足$\widetilde{\mathcal{L}}(x^k,\lambda^k;\ \rho_k)>f(z_{\epsilon_1})$，则会用 z_{ϵ_1}（见 (29)）取代它，从而防止迭代点 $\{x^k\}$ 收敛到不可行点。然而，当 k 不大时，x^k 可能相当不可行。因此，算法 2 与漏斗型或两阶段型算法有显著区别；在后者中，会先找到一个近似可行点

在阶段 1 中，然后在阶段 2 中寻求近似驻点性，同时始终保持近似可行（例如，见 *[9, 16, 26, 27, 28, 29, 30, 36]*）。

(ii) 在算法 2 中选择 ρ_0 主要是为了简化复杂度分析。不过，在实践中它可能过大，从而导致增广拉格朗日子问题的条件数很差。为了提高算法 2 在实践中的效率，可以考虑对其进行修改：先选择较小的初始罚参数，然后用一阶方法求解后续的增广拉格朗日子问题，直到找到问题 (1) 的一个 ϵ_1-一阶驻点 $\hat{}\hat{}x$ 及其对应的拉格朗日乘子 λ；最后，执行算法 2 中所述的步骤，但将 $x^0 = \hat{}x$ 和 $\lambda^0 = \Pi_{\mathcal{B}_\Lambda}(\hat{}\lambda)$ 作为参数。

在分析算法 2 的复杂度之前，我们首先说明，只要恰当选择 ρ_0，该算法就是定义良好的。具体而言，我们将证明，当 ρ_0 足够大时，可以对增广拉格朗日子问题 $\min_{x\in\mathbb{R}^n} \widetilde{\mathcal{L}}(x, \lambda^k; \rho_k)$ 应用牛顿共轭梯度法（算法 1），并以 $x^k_{\rm init}$ 为初始点，求得满足 (27) 和 (28) 的 x^{k+1}。为此，我们先根据 (22)、(25)、(26) 和 (29) 注意到

$$\widetilde{\mathcal{L}}(x^k_{\rm init}, \lambda^k; \rho_k) \le \max\{\widetilde{\mathcal{L}}(z_{\epsilon_1}, \lambda^k; \rho_k), f(z_{\epsilon_1})\} = f(z_{\epsilon_1}) \le f_{\rm hi}. \tag{30}$$

根据上述观察，我们将在下一个引理中证明，当 ρ_0 足够大时，$\widetilde{\mathcal{L}}(\cdot, \lambda^k; \rho_k)$ 有下界，且其某个水平集有界；证明推迟到第 6.2 节。

引理 4.1. 假设假设 4.1 成立。设 (λ^k, ρ_k) 是针对某个 $k \ge 0$，由算法 2 在第 k 次迭代中生成的，且 $\mathcal{S}(\delta_f, \delta_c)$ 和 $x^k_{\rm init}$ 分别由式 (24) 和 (29) 定义，并设 $f_{\rm hi}$、$f_{\rm low}$、δ_f 和 δ_c 如假设 4.1 所给定。假设 ρ_0 足够大，使得 $\delta_{f,1} \le \delta_f$ 且 $\delta_{c,1} \le \delta_c$，其中

$$\delta_{f,1} := \Lambda^2/(2\rho_0) \quad and \quad \delta_{c,1} := \sqrt{\frac{2(f_{\rm hi} - f_{\rm low} + \gamma)}{\rho_0 - 2\gamma} + \frac{\Lambda^2}{(\rho_0 - 2\gamma)^2}} + \frac{\Lambda}{\rho_0 - 2\gamma}. \tag{31}$$

则以下结论成立。

(i) $\{x : \widetilde{\mathcal{L}}(x, \lambda^k; \rho_k) \le \widetilde{\mathcal{L}}(x^k_{\rm init}, \lambda^k; \rho_k)\} \subseteq \mathcal{S}(\delta_f, \delta_c)$.

(ii) $\inf_{x\in\mathbb{R}^n} \widetilde{\mathcal{L}}(x, \lambda^k; \rho_k) \ge f_{\rm low} - \gamma - \Lambda\delta_c$.

利用引理 4.1，我们可以验证，以 $u^0 = x^k_{\rm init}$ 为初始点的牛顿共轭梯度法（算法 1）能够求得增广拉格朗日子问题 $\min_{x\in\mathbb{R}^n} \widetilde{\mathcal{L}}(x, \lambda^k; \rho_k)$ 的一个近似解 x^{k+1}，使其满足式 (27) 和 (28)。具体而言，令 $F(\cdot) = \widetilde{\mathcal{L}}(\cdot, \lambda^k; \rho_k)$ 和 $u^0 = x^k_{\rm init}$。由这些条件和引理 4.1 可知，$\{x\colon F(x) \le F(u^0)\} \subseteq \mathcal{S}(\delta_f, \delta_c)$。由此以及假设 4.1(c) 可知，水平集 $\{x\colon F(x) \le F(u^0)\}$ 是紧集，且 $\nabla^2 F$ 在 $\{x\colon F(x) \le F(u^0)\}$ 的某个凸开邻域上利普希茨连续。因此，这样的 F 和 u^0 满足假设 3.1。基于这一点以及第 3 节中的讨论，可以得出结论：以 $u^0 = x^k_{\rm init}$ 为初始点的算法 1 适用于增广拉格朗日子问题 $\min_{x\in\mathbb{R}^n} \widetilde{\mathcal{L}}(x, \lambda^k; \rho_k)$。此外，根据定理 3.1，该算法在 $(\epsilon_g, \epsilon_H) = (\tau^g_k, \tau^H_k)$ 时可生成一个满足式 (28) 以及式 (27) 中第二个关系的点 x^{k+1}。另外，由于该算法具有下降性质，且其初始点为 $x^k_{\rm init}$，因此其输出 x^{k+1} 必须满足 $\widetilde{\mathcal{L}}(x^{k+1}, \lambda^k; \rho_k) \le \widetilde{\mathcal{L}}(x^k_{\rm init}, \lambda^k; \rho_k)$。结合式 (30) 可得 $\widetilde{\mathcal{L}}(x^{k+1}, \lambda^k; \rho_k) \le f(z_{\epsilon_1})$，因而 x^{k+1} 也满足式 (27) 中的第一个关系。

上述讨论　　由此得出以下关于算法 2 定义良好的结论。

定理 4.1。在与引理 4.1 相同的设置下，将牛顿共轭梯度法（算法 1）应用于增广拉格朗日子问题 $\min_{x\in\mathbb{R}^n} \widetilde{\mathcal{L}}(x, \lambda^k; \rho_k)$，并满足 $u^0 = x^k_{\rm init}$，可找到满足 (27) 和 (28) 的点 x^{k+1}。

以下定理刻画了算法 2 的输出。其证明推迟至第 6.2 节。

定理 4.2. 假设假设 4.1 成立，且 ρ_0 足够大，使得 $\delta_{f,1} \le \delta_f$ 和$\delta_{c,1} \le \delta_c$，其中 $\delta_{f,1}$ 和 $\delta_{c,1}$ 在 (31) 中定义。若算法 2 在某次迭代 k 时终止，则 x^{k+1}是问题 (1) 的确定性 ϵ_1-一阶驻点；此外，以至少 $1-\delta$ 的概率，它是问题 (1) 的 (ϵ_1, ϵ_2)-二阶驻点。

备注 4.3. 从该定理可以看出，算法 2 的输出是问题 (1) 的随机 (ϵ_1, ϵ_2)-二阶驻点。尽管如此，我们可以很容易地修改算法 2，以寻找其他近似解。例如，如果只对寻找问题 (1) 的 ϵ_1-一阶驻点感兴趣，可以从算法 2 中移除条件 (28)。此外，如果目标是寻找确定性的 (ϵ_1, ϵ_2)-二阶驻点，可以分别用 $\lambda_{\min}(\nabla^2_{xx}\widetilde{\mathcal{L}}(x^{k+1}, \lambda^k;\ \rho_k)) \geq -\tau_k^H$和其确定性对应形式替换条件 (28) 和算法 1。在条件 (28) 中要求高概率，是为了能够推导算法 2 的运算复杂度，并以矩阵向量积的次数衡量。

在本节余下部分，我们研究算法 2 的最坏情况复杂度。由于我们的方法包含两个嵌套循环，具体而言，外层循环由增广拉格朗日方法执行，内层循环由牛顿共轭梯度法执行以求解增广拉格朗日子问题，因此我们考虑算法 2 的以下复杂度度量。

- 外层迭代复杂度，用于衡量算法 2 的外层迭代次数；
- 总内层迭代复杂度，用于衡量算法 2 中执行的牛顿共轭梯度法的迭代总次数；
- 运算复杂度，用于衡量算法 2 中计算的、涉及增广拉格朗日函数海森矩阵的矩阵-向量乘积总次数。

4.1 算法 2 的外层迭代复杂度

在本小节中，我们建立算法 2 的外层迭代复杂度。为方便记号，我们将算法 2 中出现的 (τ_k^g, τ_k^H) 改写为

$$(\tau_k^g, \tau_k^H) = (\max\{\epsilon_1, \omega_1^k\}, \max\{\epsilon_2, \omega_2^k\}) \ \text{ with } \ (\omega_1, \omega_2) := (r^{\log \epsilon_1/\log 2}, r^{\log \epsilon_2/\log 2}), \tag{32}$$

其中，ϵ_1、ϵ_2 和 r 是算法 2 的输入参数。由于 $r > 1$ 且 ϵ_1、$\epsilon_2 \in (0, 1)$，不难验证 ω_1、$\omega_2 \in (0, 1)$。此外，我们引入以下量，后文将频繁用到：

$$K_{\epsilon_1} := \left\lceil \min\{k \geq 0 : \omega_1^k \leq \epsilon_1\} \right\rceil = \lceil \log \epsilon_1 / \log \omega_1 \rceil . \tag{33}$$

根据 (32)、(33) 以及以下事实

$$\log \epsilon_1 / \log \omega_1 = \log \epsilon_2 / \log \omega_2 = \log 2 / \log r, \tag{34}$$

我们可知，对于所有 $k \geq K_{\epsilon_1}$，$(\tau_k^g,\ \tau_k^H) = (\epsilon_1,\ \epsilon_2)$。结合算法 2 的终止准则可知，该算法至少运行 K_{ϵ_1} 次迭代，并在某个 $k \geq K_{\epsilon_1}$满足 $\|c(x^{k+1})\| \leq \epsilon_1$ 时终止。因此，要确定算法 2 的外层迭代复杂度，只需界定此类 k。算法 2 的外层迭代复杂度如下所示，其证明推迟至第 6.2 节。

定理 4.3. 假设假设 4.1成立，且 ρ_0 足够大，使得$\delta_{f,1} \leq \delta_f$ and$\delta_{c,1} \leq \delta_c$，其中$\delta_{f,1}$ 和$\delta_{c,1}$ 定义于式 (31) 中。令

$$\rho_{\epsilon_1} := \max\left\{8(f_{\text{hi}} - f_{\text{low}} + \gamma)\epsilon_1^{-2} + 4\Lambda\epsilon_1^{-1} + 2\gamma, 2\rho_0\right\}, \tag{35}$$

$$\overline{K}_{\epsilon_1} := \inf\{k \geq K_{\epsilon_1} : \|c(x^{k+1})\| \leq \epsilon_1\}, \tag{36}$$

其中 K_{ϵ_1} 定义于式 (33) 中，γ、f_{hi} 和 f_{low} 见假设 4.1。则$\overline{K}_{\epsilon_1}$ 有限，且算法 2 在第 $\overline{K}_{\epsilon_1}$ 次迭代时终止，并满足

$$\overline{K}_{\epsilon_1} \leq \left(\frac{\log(\rho_{\epsilon_1}\rho_0^{-1})}{\log r} + 1\right)\left(\left|\frac{\log(\epsilon_1(2\delta_{c,1})^{-1})}{\log \alpha}\right| + 2\right) + 1. \tag{37}$$

此外，对于 $0 \leq k \leq \overline{K}_{\epsilon_1}$，$\rho_k \leq r\rho_{\epsilon_1}$ 成立

注 4.4（ $\overline{K}_{\epsilon_1}$ 和 $\{\rho_k\}$的上界）。由定理 4.3 可知，算法 2 求解问题 (1) 的随机 (ϵ_1, ϵ_2)-二阶驻点所需的外层迭代次数为 $\overline{K}_{\epsilon_1} + 1$，至多为 $\mathcal{O}(|\log \epsilon_1|^2)$。此外，该算法生成的罚参数 $\{\rho_k\}$ 至多为 $\mathcal{O}(\epsilon_1^{-2})$。

4.2 算法 2 的内层迭代总数与运算复杂度

我们给出了算法 2 求解问题 (1) 的随机 (ϵ_1, ϵ_2)-二阶驻点所需的总内层迭代次数和运算复杂度，其证明见第 6.2 节。

定理 4.4. 假设假设 4.1 成立，且 ρ_0 足够大，使得 $\delta_{f,\,1} \le \delta_f$ 和 $\delta_{c,\,1} \le \delta_c$，其中 $\delta_{f,\,1}$ 和 $\delta_{c,\,1}$ 在 (31) 中定义。则以下结论成立。

(i) 算法 2 中算法 1 执行的迭代总数至多为 $\widetilde{\mathcal{O}}(\epsilon_1^{-4}\max\{\epsilon_1^{-2}\epsilon_2, \epsilon_2^{-3}\})$。如果进一步假设 c 是仿射的，则至多为 $\widetilde{\mathcal{O}}(\max\{\epsilon_1^{-2}\epsilon_2, \epsilon_2^{-3}\})$。

(ii) 算法 2 中算法 1 执行的矩阵-向量乘积总数至多为$\widetilde{\mathcal{O}}(\epsilon_1^{-4}\max\{\epsilon_1^{-2}\epsilon_2, \epsilon_2^{-3}\}\min\{n, \epsilon_1^{-1}\epsilon_2^{-1/2}\})$。如果进一步假设 c 是仿射的，则至多为$\widetilde{\mathcal{O}}(\max\{\epsilon_1^{-2}\epsilon_2, \epsilon_2^{-3}\}\min\{n, \epsilon_1^{-1}\epsilon_2^{-1/2}\})$。

注4.5. (i) 请注意，算法 2 的上述复杂度结果是在不假设任何约束品性条件（CQ）的情况下得出的。相比之下，文献 *[60]* 在广义 LICQ 条件下针对近端增广拉格朗日方法得到了类似的复杂度结果。据我们所知，我们的工作首次研究了在不要求约束品性条件的情况下，求解问题 (1) 的随机二阶驻点的复杂度。

(ii) 令 $(\epsilon_1, \epsilon_2) = (\epsilon, \sqrt{\epsilon})$，其中 $\epsilon \in (0, 1)$，可知算法 2 在不要求约束品性条件的情况下，求解问题 (1) 的随机 $(\epsilon, \sqrt{\epsilon})$-二阶驻点时，总内层迭代复杂度为 $\widetilde{\mathcal{O}}(\epsilon^{-11/2})$，运算复杂度为 $\widetilde{\mathcal{O}}(\epsilon^{-11/2}\min\{n, \epsilon^{-5/4}\})$。

4.3 En约束品性条件下算法 2 复杂度的提升

本小节研究在一项额外假设下算法 2 的复杂度：问题 (1) 满足下文将介绍的广义线性无关约束品性条件（GLICQ）。特别地，在 GLICQ 条件下，我们将得到算法 2 更优的总内层迭代复杂度和运算复杂度；当问题 (1) 含有非线性约束时，这些复杂度界显著优于定理 4.4 中的结果。此外，当 $(\epsilon_1, \epsilon_2) = (\epsilon, \sqrt{\epsilon})$，其中 $\epsilon \in (0, 1)$，我们的复杂度界也优于文献 [60] 中针对近端增广拉格朗日方法所得的结果。现在介绍问题 (1) 的 GLICQ 假设。

假设4.2（ 广义线性无关约束资格条件）。对于所有 $x \in \mathcal{S}(\delta_f, \delta_c)$， $\nabla c(x)$ 均列满秩，其中 $\mathcal{S}(\delta_f, \delta_c)$ 的定义见（24 ）.

注4.6。针对问题（1）， *[60,* 假设 2(ii)*]* 中施加了一个相关但不同的广义线性无关约束资格条件，该条件假设对于 $f(\cdot) + \gamma\|c(\cdot)\|^2/2$ 的水平集中的所有 x， $\nabla c(x)$ 均列满秩。不难验证，该假设通常比上述广义线性无关约束资格条件更强。

下面的定理表明，在假设4.2成立的条件下，定理 4.4 中给出的总内层迭代次数和运算复杂度结果可以得到显著改进，其证明推迟至第 6.2 节。

定理 4.5。假设假设 4.1 和 4.2成立，且 ρ_0 足够大，使得 $\delta_{f,\,1} \le \delta_f$ 和 $\delta_{c,\,1} \le \delta_c$，其中 $\delta_{f,\,1}$ 和 $\delta_{c,\,1}$ 定义见（31）。则以下结论成立。

(i) 算法 2 中算法 1 执行的迭代总次数至多为 $\widetilde{\mathcal{O}}(\epsilon_1^{-2}\max\{\epsilon_1^{-2}\epsilon_2, \epsilon_2^{-3}\})$。如果进一步假设 c 是仿射的，则至多为 $\widetilde{\mathcal{O}}(\max\{\epsilon_1^{-2}\epsilon_2, \epsilon_2^{-3}\})$。

(ii) 算法 2 中算法 1 执行的矩阵-向量乘积总次数至多为$\widetilde{\mathcal{O}}(\epsilon_1^{-2}\max\{\epsilon_1^{-2}\epsilon_2, \epsilon_2^{-3}\}\min\{n, \epsilon_1^{-1/2}\epsilon_2^{-1/2}\})$。如果进一步假设 c 是仿射的，则至多为$\widetilde{\mathcal{O}}(\max\{\epsilon_1^{-2}\epsilon_2, \epsilon_2^{-3}\}\min\{n, \epsilon_1^{-1/2}\epsilon_2^{-1/2}\})$。

注4.7。(i) 从定理 4.5 可以看出，当问题 (1) 含有非线性约束时，在满足 GLICQ 和其他一些适当假设的条件下，算法 2 达到的复杂度界显著优于定理 4.4 中在不要求约束品性条件时得到的界。

(ii) 令 $(\epsilon_1, \epsilon_2) = (\epsilon, \sqrt{\epsilon})$，其中 $\epsilon \in (0, 1)$。由此可见，当问题 (1) 含有非线性约束时，在满足 GLICQ 和其他一些适当假设的条件下，算法 2 的总内层迭代复杂度为 $\widetilde{\mathcal{O}}(\epsilon^{-7/2})$，运算复杂度为 $\widetilde{\mathcal{O}}(\epsilon^{-7/2}\min\{n, \epsilon^{-3/4}\})$。这远优于文献 *[60]* 中近端增广拉格朗日方法为求解问题 (1) 的随机 $(\epsilon,\sqrt{\epsilon})$-二阶驻点而达到的总内层迭代复杂度 $\widetilde{\mathcal{O}}(\epsilon^{-11/2})$ 和运算复杂度 $\widetilde{\mathcal{O}}(\epsilon^{-11/2}\min\{n, \epsilon^{-3/4}\})$；不过，该方法要求的 GLICQ 通常更强。

5 数值结果

我们进行了一些初步实验，以测试所提出方法（算法1和算法2）的性能，并分别将它们与 [56] 中的牛顿共轭梯度法和 [60],中的近端增广拉格朗日方法进行比较。所有算法均使用 Matlab 编写，所有计算均在一台台式机上完成，该计算机配备 3.79 GHz AMD 3900XT 12核处理器和 32 GB 随机存取存储器。

5.1 正则化鲁棒回归

在本小节中，我们考虑正则化鲁棒回归问题

$$\min_{x\in\mathbb{R}^n} \sum_{i=1}^{m} \phi(a_i^T x - b_i) + \mu\|x\|_4^4, \tag{38}$$

其中，对于任意 $p \geq 1$，$\phi(t) = t^2/(1+t^2)$，$\|x\|_p = (\sum_{i=1}^n |x_i|^p)^{1/p}$，且 $\mu > 0$。

对于每组三元组（n，m，μ），我们随机生成问题（38）的 10 个实例。具体而言，我们首先随机生成 a_i、$1 \leq i \leq m$，其中所有元素均独立地取自标准正态分布。然后，我们根据标准正态分布随机生成 b_i，并对 $i = 1$，……，m设置 $b_i = 2mb_i$。

我们的目标是通过算法 1 和 [56] 中的牛顿共轭梯度方法，为上述实例求解（$10^{-5}, 10^{-5/2}$）-二阶驻点，并比较两者的性能。为确保比较公平，我们对两种方法都使用一个能返回确定性输出的最小特征值预言机，从而保证它们都能输出一个近似二阶驻点。具体而言，我们使用 Matlab 子程序 [v,λ] = eigs(H,1,'smallestreal') 作为最小特征值预言机，以求出实对称矩阵 H的最小特征值 λ 及其对应的单位特征向量 v。此外，对于两种方法，我们均选择全 1 向量作为初始点，并设置$\theta = 0.8$、$\zeta = 0.5$和 $\eta = 0.2$。

算法 1 和 [56] 中的牛顿共轭梯度方法在上述随机生成实例上的计算结果列于表 2。具体而言，前 3 列分别列出了 n、m和 μ 的值。对于每组 (n, m, μ)，其余各列给出了 10 个随机实例的平均 CPU 时间（秒）、平均迭代次数和平均最终目标值。可以观察到，两种方法都能输出目标值相近的近似解，而我们的算法 1 在 CPU 时间方面显著优于 [56] 中的牛顿共轭梯度方法。这与我们的理论结果一致：在求解近似二阶驻点时，算法 1 关于海森矩阵的利普希茨常数的依赖具有比 [56] 中的牛顿共轭梯度方法更优的迭代复杂度。

5.2 球约束正则化稳健回归

In 小节：我们考虑球约束正则化稳健回归问题 m

$$\min_{x\in\mathbb{R}^n} \sum_{i=1}^{m} \phi(a_i^T x - b_i) + \mu\|x\|_4^4 \quad \text{s.t.} \quad \|x\|_2^2 = 1, \tag{39}$$

其中，对于任意 $p \geq 1$，$\phi(t) = t^2/(1+t^2)$，$\|x\|_p = (\sum_{i=1}^n |x_i|^p)^{1/p}$，而 $\mu > 0$ 是一个调节参数。对于每个三元组 (n, m, μ)，我们按照第 5.1 小节所述的相同方式，随机生成问题 (39) 的 10 个实例。

			目标值		迭代		CPU 时间（秒）	
n	m	μ	算法 1	牛顿共轭梯度	算法 1	牛顿共轭梯度	算法 1	牛顿共轭梯度
100	10	1	5.9	5.9	85.7	116.3	1.4	1.6
100	50	1	45.9	45.9	82.6	158.2	1.0	2.7
100	90	1	84.8	84.8	102.2	224.7	2.0	4.2
500	50	5	42.2	42.5	173.1	344.7	44.2	72.2
500	250	5	243.0	242.9	145.5	362.4	41.9	95.0
500	450	5	442.2	442.2	163.7	425.2	47.6	138.3
1000	100	10	90.1	90.4	162.5	361.0	110.8	259.0
1000	500	10	491.1	491.2	158.3	475.4	129.1	558.4
1000	900	10	891.1	891.1	193.5	300.7	187.0	298.5

表 2：问题 (38) 的数值结果

			目标值		可行性违反量 ($\times 10^{-4}$)		内部迭代总次数		CPU 时间（秒）	
n	m	μ	算法 2	Prox-AL	算法 2	Prox-AL	算法 2	Prox-AL	算法 2	Prox-AL
100	10	1	7.1	7.1	0.18	0.27	40.9	97.3	0.73	2.2
100	50	1	46.6	46.6	0.21	0.30	37.0	86.3	0.78	1.7
100	90	1	87.0	87.0	0.12	0.40	39.5	68.6	1.1	1.9
500	50	5	44.4	44.4	0.40	0.68	59.0	343.4	11.4	134.9
500	250	5	244.3	244.3	0.37	0.47	59.0	543.3	11.7	178.2
500	450	5	444.0	444.0	0.27	0.53	66.7	634.1	17.1	158.2
1000	100	10	92.8	92.8	0.28	0.42	95.0	2054.6	46.3	1516.8
1000	500	10	491.9	491.9	0.22	0.72	68.3	756.2	39.5	558.6
1000	900	10	893.4	893.4	0.19	0.37	81.8	1281.4	57.7	1099.6

表 3：问题 (39) 的数值结果

我们的目标是通过算法 2 和近端增广拉格朗日方法 [60, Algorithm 3]，为上述实例求解问题 (39) 的 ($10^{-4}, 10^{-2}$)-二阶驻点，并比较它们的性能。为公平比较，我们使用一个能为两种方法返回确定性输出的最小特征值预言机，从而确保它们都能输出近似二阶驻点。具体而言，我们使用 Matlab 子程序 [v,λ] = eigs(H,1,'smallestreal') 作为最小特征值预言机，以求得实对称矩阵 H 的最小特征值 λ 及其对应的单位特征向量 v。此外，对于两种方法，我们均将初始点设为 $z^0 = (1/\sqrt{n}, \ldots, 1/\sqrt{n})^T$，初始拉格朗日乘子设为 $\lambda^0 = 0$，其他参数设为

- $\Lambda = 100$, $\rho_0 = 10$, $\alpha = 0.25$和 $r = 10$ 用于算法 2;
- $\eta = 1$、 $q = 10$ 和 $T_0 = 2$ 用于近端增广拉格朗日方法（[60]）。

表 3 展示了算法 2 和 [60] 中的近端增广拉格朗日方法（简称 Prox-AL）在求解问题 (39) 时，对上述随机生成实例的计算结果。具体而言， n、 m和 μ的值分别列于前三列。对于每组三元组（n、 m、 μ），其余各列给出了 10 个随机实例的平均 CPU 时间（以秒为单位）、平均内层迭代总次数、平均最终目标值以及平均最终可行性违反量。可以观察到，就目标值和可行性违反量而言，两种方法均输出了质量相近的近似解；而在 [60]方面，就 CPU 时间而言，我们的算法 2 大幅优于近端增广拉格朗日方法。这印证了我们的理论发现：在寻找近似二阶驻点方面，算法 2 的运算复杂度显著优于 [60] 中的近端增广拉格朗日方法。

6 主要结果的证明

我们在第 3 节和第 4 节中给出主要结果的证明，包括定理 3.1、引理 4.1 以及定理 4.2、4.3、4.4 和 4.5。

6.1 第 3 节主要结果的证明

在本小节中，我们首先建立若干技术引理，然后利用它们证明定理 3.1。

从假设 3.1(b) 可以看出，对于所有 x 和 $y \in \Omega$,

$$\|\nabla F(y) - \nabla F(x) - \nabla^2 F(x)(y-x)\| \le L_H^F \|y-x\|^2/2, \tag{40}$$
$$F(y) \le F(x) + \nabla F(x)^T(y-x) + (y-x)^T \nabla^2 F(x)(y-x)/2 + L_H^F \|y-x\|^3/6. \tag{41}$$

下一个引理给出了算法 3 的输出的一些有用性质，其证明与 [56, 引理 3] 和 [54, 引理 7] 中的证明类似，因此此处省略。

引理 6.1. 假设假设 3.1成立，且方向 d^t 是算法 1 在某次迭代 t 中由算法 3 的输出 d 得到的，其类型由 d type 指定。则以下结论成立。

(i) 若 d_type=SOL，则 d^t 满足

$$\epsilon_H \|d^t\|^2 \le (d^t)^T \left(\nabla^2 F(x^t) + 2\epsilon_H I\right) d^t, \tag{42}$$
$$\|d^t\| \le 1.1 \epsilon_H^{-1} \|\nabla F(x^t)\|, \tag{43}$$
$$(d^t)^T \nabla F(x^t) = -(d^t)^T \left(\nabla^2 F(x^t) + 2\epsilon_H I\right) d^t, \tag{44}$$
$$\|(\nabla^2 F(x^t) + 2\epsilon_H I) d^t + \nabla F(x^t)\| \le \epsilon_H \zeta \|d^t\|/2. \tag{45}$$

(ii) 若 d_type=负曲率，则 d^t 满足 $(d^t)^T \nabla F(x^t) \le 0$ 并且

$$(d^t)^T \nabla^2 F(x^t) d^t / \|d^t\|^2 = -\|d^t\| \le -\epsilon_H. \tag{46}$$

下一个引理表明，当算法 1 中的搜索方向 d^t 类型为“SOL”时，线搜索步骤会使 F充分下降。

引理 6.2. 假设假设 3.1成立，且方向 d^t 是算法 1 在某次迭代 t 中由算法 3 的输出 d 得到的，其 d type 为=SOL。令 U_g^F和 $c_{\rm sol}$ 分别由 (9) 和 (19) 给出。则以下结论成立。

(i) 步长 α_t 定义良好，此外，

$$\alpha_t \ge \min\left\{1, \sqrt{\frac{\min\{6(1-\eta), 2\}}{1.1 L_H^F U_g^F}} \theta \epsilon_H \right\}. \tag{47}$$

(ii) 下一个迭代点 $x^{t+1} = x^t + \alpha_t d^t$ 满足

$$F(x^t) - F(x^{t+1}) \ge c_{\rm sol} \min\{\|\nabla F(x^{t+1})\|^2 \epsilon_H^{-1}, \epsilon_H^3\}. \tag{48}$$

证明。可以观察到，沿着算法 1 生成的迭代点，F是下降的（只要其定义良好）；结合 $x^0 = u^0$可得 $F(x^t) \le F(u^0)$，因此由 (9) 可知 $\|\nabla F(x^t)\| \le U_g^F$。此外，由于 d^t 是由算法 3 输出的 d 得到的，且 d 类型为 =SOL，因此可知 $\|\nabla F(x^t)\| > \epsilon_g$ 以及 (42)-(45) 对 d^t成立。此外，由 $\|\nabla F(x^t)\| > \epsilon_g$ 和 (45)，可以得出 $d^t \ne 0$。

我们先证明陈述 (i)。如果 (16) 对 $j = 0$成立，那么 $\alpha_t = 1$，这显然意味着 (47) 成立。现在假设 (16) 对 $j = 0$不成立。断言，对于所有违反 (16) 的 $j \ge 0$ ，都有

$$\theta^{2j} \ge \min\{6(1-\eta), 2\} \epsilon_H (L_H^F)^{-1} \|d^t\|^{-1}. \tag{49}$$

事实上，假设某个 $j \ge 0$ 违反了式 (16)。下面我们通过分别讨论两种情况，证明此时 j 满足式 (49)。

情况 1) $F(x^t + \theta^j d^t) > F(x^t)$。令 $\phi(\alpha) = F(x^t + \alpha d^t)$。则 $\phi(\theta^j) > \phi(0)$。此外，由于 $d^t \ne 0$, (42) 根据 and (44)，可得

$$\phi'(0) = \nabla F(x^t)^T d^t = -(d^t)^T (\nabla^2 F(x^t) + 2\epsilon_H I) d^t \le -\epsilon_H \|d^t\|^2 < 0.$$

利用这些结果，我们可以观察到，存在一个 ϕ 的局部极小值点 $\alpha^* \in (0, \theta^j)$，使得 $\phi'(\alpha^*) = \nabla F(x^t + \alpha^* d^t)^T d^t = 0$，且 $\phi(\alpha^*) < \phi(0)$，这意味着 $F(x^t + \alpha^* d^t) < F(x^t) \leq F(u^0)$。因此，(40) 对 $x = x^t$ 和 $y = x^t + \alpha^* d^t$ 成立。利用这一点、$0 < \alpha^* < \theta^j \leq 1$ 和 $\nabla F(x^t + \alpha^* d^t)^T d^t = 0$，我们得到

$$\begin{aligned}
&\tfrac{(\alpha^*)^2 L_H^F}{2}\|d^t\|^3 \overset{(40)}{\geq} \|d^t\|\|\nabla F(x^t + \alpha^* d^t) - \nabla F(x^t) - \alpha^* \nabla^2 F(x^t) d^t\| \\
&\geq (d^t)^T(\nabla F(x^t + \alpha^* d^t) - \nabla F(x^t) - \alpha^* \nabla^2 F(x^t) d^t) = -(d^t)^T \nabla F(x^t) - \alpha^* (d^t)^T \nabla^2 F(x^t) d^t \\
&\overset{(44)}{=} (1 - \alpha^*)(d^t)^T(\nabla^2 F(x^t) + 2\epsilon_H I) d^t + 2\alpha^* \epsilon_H \|d^t\|^2 \overset{(42)}{\geq} (1 + \alpha^*)\epsilon_H \|d^t\|^2 \geq \epsilon_H \|d^t\|^2,
\end{aligned}$$

结合 $d^t \neq 0$ 可得 $(\alpha^*)^2 \geq 2\epsilon_H (L_H^F)^{-1} \|d^t\|^{-1}$。利用这一点和 $\theta^j > \alpha^*$，我们得出此情形下 (49) 成立。

情形 2） $F(x^t + \theta^j d^t) \leq F(x^t)$。结合 $F(x^t) \leq F(u^0)$，可得对于 $x = x^t$ 和 $y = x^t + \theta^j d^t$，(41) 成立。于是，由于 j 不满足 (16)，我们得到

$$\begin{aligned}
&-\eta \epsilon_H \theta^{2j} \|d^t\|^2 \leq F(x^t + \theta^j d^t) - F(x^t) \overset{(41)}{\leq} \theta^j \nabla F(x^t)^T d^t + \frac{\theta^{2j}}{2}(d^t)^T \nabla^2 F(x^t) d^t + \frac{L_H^F}{6}\theta^{3j}\|d^t\|^3 \\
&\overset{(44)}{=} -\theta^j (d^t)^T(\nabla^2 F(x^t) + 2\epsilon_H I) d^t + \frac{\theta^{2j}}{2}(d^t)^T \nabla^2 F(x^t) d^t + \frac{L_H^F}{6}\theta^{3j}\|d^t\|^3 \\
&= -\theta^j \left(1 - \frac{\theta^j}{2}\right)(d^t)^T(\nabla^2 F(x^t) + 2\epsilon_H I) d^t - \theta^{2j}\epsilon_H \|d^t\|^2 + \frac{L_H^F}{6}\theta^{3j}\|d^t\|^3 \\
&\overset{(42)}{\leq} -\theta^j \left(1 - \frac{\theta^j}{2}\right)\epsilon_H \|d^t\|^2 - \theta^{2j}\epsilon_H \|d^t\|^2 + \frac{L_H^F}{6}\theta^{3j}\|d^t\|^3 \leq -\theta^j \epsilon_H \|d^t\|^2 + \frac{L_H^F}{6}\theta^{3j}\|d^t\|^3. \qquad (50)
\end{aligned}$$

回顾 $d^t \neq 0$。将 (50) 两边除以 $L_H^F \theta^j \|d^t\|^3/6$，并利用 $\eta, \theta \in (0, 1)$，可得

$$\theta^{2j} \geq 6(1 - \theta^j \eta)\epsilon_H (L_H^F)^{-1}\|d^t\|^{-1} \geq 6(1 - \eta)\epsilon_H (L_H^F)^{-1}\|d^t\|^{-1}.$$

因此，此情形下 (49) 也成立。

综合上述两种情形，我们得出对于任何不满足 (16) 的 $j \geq 0$，(49) 均成立。由此以及 $\theta \in (0, 1)$ 可知，所有不满足 (16) 的 $j \geq 0$ 都有上界。因此，与 (16) 相关的步长 α_t 是良定义的。接下来我们证明 (47)。根据算法 1 中 j_t 的定义可知，$j = j_t - 1$ 不满足 (16)，因此 (49) 对 $j = j_t - 1$ 成立。于是，由 (49) 结合 $j = j_t - 1$ 和 $\alpha_t = \theta^{j_t}$，可得

$$\alpha_t = \theta^{j_t} \geq \sqrt{\min\{6(1 - \eta), 2\}\epsilon_H (L_H^F)^{-1}}\, \theta \|d^t\|^{-1/2}, \qquad (51)$$

由此再结合 (43) 和 $\|\nabla F(x^t)\| \leq U_g^F$，可推出 (47)。这就证明了陈述 (i)　　.

接下来，我们通过考虑以下两种情况来证明陈述 (ii)。

情况 1) $\alpha_t = 1$。由此可知，(16) 对 $j = 0$ 成立。因此可得 $F(x^t + d^t) \leq F(x^t) \leq F(u^0)$，这意味着 (40) 对 $x = x^t$ 和 $y = x^t + d^t$成立。由此并结合 (45)，可得

$$\begin{aligned}
\|\nabla F(x^{t+1})\| = \|\nabla F(x^t + d^t)\| &\leq \|\nabla F(x^t + d^t) - \nabla F(x^t) - \nabla^2 F(x^t) d^t\| \\
&\quad + \|(\nabla^2 F(x^t) + 2\epsilon_H I) d^t + \nabla F(x^t)\| + 2\epsilon_H \|d^t\| \\
&\leq \tfrac{L_H^F}{2}\|d^t\|^2 + \tfrac{4+\zeta}{2}\epsilon_H \|d^t\|,
\end{aligned}$$

最后一个不等式由 (40) 和 (45) 得出。对上述不等式关于 $\|d^t\|$ 求解，并利用 $\|d^t\| > 0$这一事实，可得

$$\begin{aligned}
\|d^t\| &\geq \tfrac{-(4+\zeta)\epsilon_H + \sqrt{(4+\zeta)^2\epsilon_H^2 + 8L_H^F\|\nabla F(x^{t+1})\|}}{2L_H^F} \geq \tfrac{-(4+\zeta)\epsilon_H + \sqrt{(4+\zeta)^2\epsilon_H^2 + 8L_H^F\epsilon_H^2}}{2L_H^F}\min\{\|\nabla F(x^{t+1})\|/\epsilon_H^2, 1\} \\
&= \tfrac{4}{4+\zeta+\sqrt{(4+\zeta)^2+8L_H^F}}\min\{\|\nabla F(x^{t+1})\|/\epsilon_H, \epsilon_H\},
\end{aligned}$$

其中，第二个不等式由不等式 $-a + \sqrt{a^2 + bs} \geq (-a + \sqrt{a^2 + b})\min\{s, 1\}$ for all$a, b, s \geq 0$ 得出，这一点可分别对项 $-a + \sqrt{a^2 + b}$ 和 $-a + \sqrt{a^2 + bs}$ 进行有理化验证。由此，结合 $\alpha_t = 1$、(16) 和 (19)，可知 (48) 成立。

情形 2) $\alpha_t < 1$。因此， $j = 0$ 违反了 (16)，故 (49) 对 $j = 0$ 成立。现在，将 $j = 0$ 代入 (49)，得到 $\|d^t\| \geq \min\{6(1-\eta)\, 2\}\epsilon_H / L_H^F$，结合 (16) 和 (51) 可推出

$$F(x^t) - F(x^{t+1}) \geq \eta\epsilon_H\theta^{2j_t}\|d^t\|^2 \geq \eta\frac{\min\{6(1-\eta),2\}\epsilon_H^2}{L_H^F}\theta^2\|d^t\| \geq \eta\left[\frac{\min\{6(1-\eta),2\}\theta}{L_H^F}\right]^2\epsilon_H^3.$$

由此以及 (19)，可以看出在这种情况下 (48) 也成立。 □

下面的引理表明，当算法 1 中的搜索方向 d^t 属于“负曲率”类型时，线搜索得到的步长也会使 F 充分下降。

引理 6.3. 假设假设 3.1 成立，且方向 d^t 是算法 1 某次迭代中由算法 3 输出的类型为 =负曲率的 d ，或由算法 4 输出的 v 。-令 $c_{\rm nc}$ 由 (20) 定义，则以下结论成立。

(i) 步长 α_t 定义良好，且 $\alpha_t \geq \min\{1, \theta/L_H^F, 3(1-\eta)\theta/L_H^F\}$。

(ii) 下一个迭代点 $x^{t+1} = x^t + \alpha_t d^t$ 满足 $F(x^t) - F(x^{t+1}) \geq c_{\rm nc}\epsilon_H^3$。

证明。注意，沿着算法 1 生成的迭代点， F 是下降的（只要其定义良好）。利用这一点和 $x^0 = u^0$，可得 $F(x^t) \leq F(u^0)$。根据关于 d^t 的假设，可以从算法 1 看出， d^t 是 (13) 或 (15) 中给出的负曲率方向。此外，注意到算法 4 返回的向量 v 满足 $\|v\| = 1$。由此以及引理 6.1(ii)、(13) 和 (15)，可以观察到

$$\nabla F(x^t)^T d^t \leq 0, \quad (d^t)^T\nabla^2 F(x^t)d^t = -\|d^t\|^3 < 0. \tag{52}$$

我们首先证明陈述 (i)。如果 $j = 0$ 满足 (17)，那么 $\alpha_t = 1$，这显然意味着 $\alpha_t \geq \min\{1, \theta/L_H^F, 3(1-\eta)\theta/L_H^F\}$。现在假设 $j = 0$ 不满足 (17)。断言，对于所有违反 (17) 的 $j \geq 0$，都有

$$\theta^j \geq \min\{1/L_H^F, 3(1-\eta)/L_H^F\}. \tag{53}$$

事实上，假设某个 $j \geq 0$ 违反了 (17)。下面分两种情况，证明这样的 j 满足 (53)。

情况 1) $F(x^t + \theta^j d^t) > F(x^t)$。令 $\phi(\alpha) = F(x^t + \alpha d^t)$。则 $\phi(\theta^j) > \phi(0)$。此外，由 (52) 可知， s

$$\phi'(0) = \nabla F(x^t)^T d^t \leq 0, \quad \phi''(0) = (d^t)^T\nabla^2 F(x^t)d^t < 0.$$

利用这些结果，我们可以观察到，存在 ϕ 的一个局部极小值点 $\alpha^* \in (0, \theta^j)$，使得 $\phi(\alpha^*) < \phi(0)$，即 $F(x^t + \alpha^* d^t) < F(x^t)$。根据 ϕ 在 α^* 处的二阶最优性条件，有 $\phi''(\alpha^*) = (d^t)^T\nabla^2 F(x^t + \alpha^* d^t)d^t \geq 0$。由于 $F(x^t + \alpha^* d^t) < F(x^t) \leq F(u^0)$，因此 (8) 对 $x = x^t$ 和 $y = x^t + \alpha^* d^t$ 成立。利用这一点、(52) 中的第二个关系以及 $(d^t)^T\nabla^2 F(x^t + \alpha^* d^t)d^t \geq 0$，我们得到，在 (52) 和 $(d^t)^T\nabla^2 F(x^t + \alpha^* d^t)d^t \geq 0$ 中，我们得到

$$\begin{array}{rcl} L_H^F\alpha^*\|d^t\|^3 & \overset{(8)}{\geq} & \|d^t\|^2\|\nabla^2 F(x^t + \alpha^* d^t) - \nabla^2 F(x^t)\| \geq (d^t)^T(\nabla^2 F(x^t + \alpha^* d^t) - \nabla^2 F(x^t))d^t \\ & \geq & -(d^t)^T\nabla^2 F(x^t)d^t = \|d^t\|^3. \end{array} \tag{54}$$

由 (52) 可知 $d^t \neq 0$。进而由 (54) 得到 $\alpha^* \geq 1/L_H^F$，结合 $\theta^j > \alpha^*$ 可推出 $\theta^j > 1/L_H^F$。因此，在这种情况下，(53) 成立。

情形 2) $F(x^t + \theta^j d^t) \leq F(x^t)$。由此以及 $F(x^t) \leq F(u^0)$ 可知，对于 $x = x^t$ 和 $y = x^t + \theta^j d^t$，(41) 成立。由此，再根据 j 不满足 (17) 这一事实，可得

$$\begin{array}{rcl} -\frac{\eta}{2}\theta^{2j}\|d^t\|^3 & \leq & F(x^t + \theta^j d^t) - F(x^t) \overset{(41)}{\leq} \theta^j\nabla F(x^t)^T d^t + \frac{\theta^{2j}}{2}(d^t)^T\nabla^2 F(x^t)d^t + \frac{L_H^F}{6}\theta^{3j}\|d^t\|^3 \\ & \overset{(52)}{\leq} & -\frac{\theta^{2j}}{2}\|d^t\|^3 + \frac{L_H^F}{6}\theta^{3j}\|d^t\|^3, \end{array}$$

结合 $d^t \neq 0$ 可推出 $\theta^j \geq 3(1-\eta)/L_H^F$。因此，在这种情况下，(53) 也成立。

综合上述两种情况，我们得出，对于任何违反 (17) 的 $j \geq 0$ ，(53) 都成立。由此以及 $\theta \in (0, 1)$ 可知，所有违反 (17) 的 $j \geq 0$ 都有上界。因此，与 (17) 相关的步长 α_t 定义良好。接下来，我们推导 α_t 的下界。根据算法 1 中 j_t 的定义，注意到 $j = j_t - 1$ 违反 (17)，因此 (53) 对 $j = j_t - 1$ 成立。于是，由取 $j = j_t - 1$ 和 $\alpha_t = \theta^{j_t}$ 时的 (53)，可得 $\alpha_t = \theta^{j_t} \geq \min\{\theta/L_H^F, 3(1-\eta)\theta/L_H^F\}$，从而立即得到 $\alpha_t \geq \min\{1, \theta/L_H^F, 3(1-\eta)\theta/L_H^F\}$，即为所求。

接下来，我们通过考察以下两种情形来证明命题 (ii)。

情况 1) d^t 是由算法 3 的输出 d 得到的，其中 d 类型为=负曲率。由此，根据 (46) 可得$\|d^t\| \geq \epsilon_H$。结合 (17) 和命题 (i)，可知命题 (ii) 成立。

情形 2) d^t 是由算法 4 的输出 v 得到的。注意，根据算法 4，$\|v\| = 1$ 和$v^T \nabla^2 F(x^t) v \leq -\epsilon_H/2$，结合 (15) 可得 $\|d^t\| \geq \epsilon_H/2$。由此，结合 (17) 和命题 (i)，可知命题 (ii) 同样成立。

□

定理 3.1 的证明。为方便记号，我们令 $\{x^t\}_{t\in\mathbb{T}}$ 表示算法 1 生成的所有迭代点，其中 $\mathbb{T}$ 是从 0 开始的连续非负整数集合。注意，F 沿算法 1 生成的迭代点单调下降，结合 $x^0 = u^0$ 可得 $x^t \in \{x : F(x) \leq F(u^0)\}$。由此，根据 (9)，对所有 $t \in \mathbb{T}$，$\|\nabla^2 F(x^t)\| \leq U_H^F$成立。

(i) 反设算法 1 中算法 4 的调用总次数大于 T_2。由算法 1 和引理 6.3(ii) 可知，除最后一次调用外，这些调用中的每一次都会返回一个充分负的曲率方向，并且每次调用都会使 F 至少减少 $c_{\rm nc}\epsilon_H^3$。因此，

$$T_2 c_{\rm nc}\epsilon_H^3 \leq \sum_{t\in\mathbb{T}} [F(x^t) - F(x^{t+1})] \leq F(x^0) - F_{\rm low} = F_{\rm hi} - F_{\rm low},$$

这与 (18) 中 T_2 的定义矛盾。因此，定理 3.1 的陈述 (i) 成立。

(ii) 反设算法 1 中算法 3 的调用总次数大于 T_1。注意，若算法 3 在某次迭代 t 被调用，并生成满足 $\|\nabla F(x^{t+1})\| \leq \epsilon_g$的下一个迭代点 x^{t+1}，则算法 4 必须在下一次迭代 $t+1$被调用。结合这一点与定理 3.1 的陈述 (i)，可知此类迭代 t 的总数至多为 T_2。因此，算法 1 中调用算法 3 并生成满足 $\|\nabla F(x^{t+1})\| > \epsilon_g$ 的下一个迭代点 x^{t+1} 的迭代 t 总数至少为 $T_1 - T_2 + 1$。此外，对于这些迭代 t，由引理 6.2(ii) 和 6.3(ii) 可知，$F(x^t) - F(x^{t+1}) \geq \min\{c_{\rm sol}, c_{\rm nc}\} \min\{\epsilon_g^2\epsilon_H^{-1}, \epsilon_H^3\}$。由此可得

$$(T_1 - T_2 + 1) \min\{c_{\rm sol}, c_{\rm nc}\} \min\{\epsilon_g^2\epsilon_H^{-1}, \epsilon_H^3\} \leq \sum_{t\in\mathbb{T}} [F(x^t) - F(x^{t+1})] \leq F_{\rm hi} - F_{\rm low},$$

这与式 (18) 中 T_1 和 T_2 的定义矛盾。因此，定理 3.1 的陈述 (ii) 成立。

(iii) 注意，在算法 1 的每次迭代中，都会调用算法 3 或算法 4。由此以及定理 3.1 的陈述 (i) 和 (ii) 可知，算法 1 的迭代总次数至多为 $T_1 + T_2$。此外，由式 (19)、(20) 和 (18) 可得关系式 (21)。还可以看出，算法 1 的输出 x^t 确定性地满足 $\|\nabla F(x^t)\| \leq \epsilon_g$，并且以至少 $1 - \delta$ 的概率满足 $\lambda_{\min}(\nabla^2 F(x^t)) \geq -\epsilon_H$，其中某个 $0 \leq t \leq T_1 + T_2$；后者是由算法 4 保证的。至此，定理 3.1 的陈述 (ii) 得证。

(iv) 根据定理 A.1，取 $(H, \varepsilon) = (\nabla^2 F(x^t)\ \epsilon_H)$，并结合 $\|\nabla^2 F(x^t)\| \leq U_H^F$ 这一事实，可以看出，算法 3 每次以 $U = 0$ 为输入调用时所需的海森-向量积次数至多为 $\widetilde{\mathcal{O}}(\min\{n, (U_H^F/\epsilon_H)^{1/2}\})$。此外，根据定理 B.1，取 $(H, \varepsilon) = (\nabla^2 F(x^t)\ \epsilon_H)$，并结合 $\|\nabla^2 F(x^t)\| \leq U_H^F$ 以及 Lanczos 方法的每次迭代只需一次矩阵向量积这一事实，可以看出，算法 4 每次调用所需的海森-向量积次数也至多为$\widetilde{\mathcal{O}}(\min\{n, (U_H^F/\epsilon_H)^{1/2}\})$。基于这些观察以及定理 3.1 的陈述 (iii)，我们可知该定理的陈述 (iv) 成立。

□

6.2 第 4 节主要结果的证明

根据假设 4.1(a)，回顾可知 $\|c(z_{\epsilon_1})\| \le \epsilon_1/2 < 1$。由此，结合式 (23) 和式 (25) 中 $\tilde{c}$ 的定义，我们得到

$$f(x)+\gamma\|\tilde{c}(x)\|^2 \ge f(x)+\gamma\|c(x)\|^2/2-\gamma\|c(z_{\epsilon_1})\|^2 \ge f_{\rm low}-\gamma, \quad \forall x\in\mathbb{R}^n. \tag{55}$$

现在我们证明以下辅助引理，后文将频繁用到它。

引理 6.4. 假设假设 4.1 成立。设 γ、$f_{\rm hi}$ 和 $f_{\rm low}$ 如假设 4.1 所给定。假设 $\rho>2\gamma$、$\lambda\in\mathbb{R}^m$ 和 $x\in\mathbb{R}^n$ 满足

$$\widetilde{\mathcal{L}}(x,\lambda;\rho)\le f_{\rm hi}, \tag{56}$$

其中 $\mathcal{L}$ 定义于式 (26)。则以下结论成立。

(i) $f(x)\le f_{\rm hi}+\|\lambda\|^2/(2\rho)$。(ii) $\|\tilde{c}(x)\|\le\sqrt{2(f_{\rm hi}-f_{\rm low}+\gamma)/(\rho-2\gamma)+\|\lambda\|^2/(\rho-2\gamma)^2}+\|\lambda\|/(\rho-2\gamma)$。(iii) 若对某个 $\tilde{\delta}_f>0$，有 $\rho\ge\|\lambda\|^2/(2\tilde{\delta}_f)$，则 $f(x)\le f_{\rm hi}+\tilde{\delta}_f$。

(iv) 若

$$\rho\ge 2(f_{\rm hi}-f_{\rm low}+\gamma)\tilde{\delta}_c^{-2}+2\|\lambda\|\tilde{\delta}_c^{-1}+2\gamma \tag{57}$$

对于某个 $\delta_c>0$，则 $\|\tilde{c}(x)\|\le\delta_c$。

证明。（i）由 (56) 以及 (26) 中 $\mathcal{L}$ 的定义可知

$$f_{\rm hi}\ge f(x)+\lambda^T\tilde{c}(x)+\tfrac{\rho}{2}\|\tilde{c}(x)\|^2=f(x)+\tfrac{\rho}{2}\left\|\tilde{c}(x)+\tfrac{\lambda}{\rho}\right\|^2-\tfrac{\|\lambda\|^2}{2\rho}\ge f(x)-\tfrac{\|\lambda\|^2}{2\rho}.$$

因此，命题 (i) 成立。

（ii）根据 (55) 和 (56)，有

$$\begin{aligned} f_{\rm hi} &\overset{(56)}{\ge} f(x)+\lambda^T\tilde{c}(x)+\tfrac{\rho}{2}\|\tilde{c}(x)\|^2=f(x)+\gamma\|\tilde{c}(x)\|^2+\tfrac{\rho-2\gamma}{2}\left\|\tilde{c}(x)+\tfrac{\lambda}{\rho-2\gamma}\right\|^2-\tfrac{\|\lambda\|^2}{2(\rho-2\gamma)}\\ &\overset{(55)}{\ge} f_{\rm low}-\gamma+\tfrac{\rho-2\gamma}{2}\left\|\tilde{c}(x)+\tfrac{\lambda}{\rho-2\gamma}\right\|^2-\tfrac{\|\lambda\|^2}{2(\rho-2\gamma)}. \end{aligned}$$

由此可得 $\left\|\tilde{c}(x)+\frac{\lambda}{\rho-2\gamma}\right\|\le\sqrt{\frac{2(f_{\rm hi}-f_{\rm low}+\gamma)}{\rho-2\gamma}+\frac{\|\lambda\|^2}{(\rho-2\gamma)^2}}$，这说明命题 (ii) 成立。

(iii) 结论 (iii) 由结论 (i) 和 $\rho\ge\|\lambda\|^2/(2\tilde{\delta}_f)$ 立即得到。

(iv) 假设 (57) 成立。将 (57) 两边乘以 $\tilde{\delta}_c^2$ 并重新整理各项，可得

$$(\rho-2\gamma)\tilde{\delta}_c^2-2\|\lambda\|\tilde{\delta}_c-2(f_{\rm hi}-f_{\rm low}+\gamma)\ge 0.$$

回顾 $\rho>2\gamma$ 以及 $\delta_c>0$。对该不等式求解 δ_c 可得

$$\tilde{\delta}_c\ge\sqrt{2(f_{\rm hi}-f_{\rm low}+\gamma)/(\rho-2\gamma)+\|\lambda\|^2/(\rho-2\gamma)^2}+\|\lambda\|/(\rho-2\gamma),$$

结合结论 (ii) 可知 $\|\tilde{c}(x)\|\le\delta_c$。因此，结论 (iv) 成立。 □

引理 4.1 的证明。(i) 令 x 为任意一点，使得 $\widetilde{\mathcal{L}}(x,\lambda^k;\rho_k)\le\widetilde{\mathcal{L}}(x_{\rm init}^k,\lambda^k;\rho_k)$。由此，根据 (30) 可得 $\widetilde{\mathcal{L}}(x,\lambda^k;\rho_k)\le f_{\rm hi}$。由此，$\|\lambda^k\|\le\Lambda$, $\rho_k\ge\rho_0>2\gamma$, $\delta_{f,1}\le\delta_f$, $\delta_{c,1}\le\delta_c$，并结合引理 6.4 与 $(\lambda,\rho)=(\lambda^k,\rho_k)$，可得

$$\begin{aligned} &f(x)\le f_{\rm hi}+\|\lambda^k\|^2/(2\rho_k)\le f_{\rm hi}+\Lambda^2/(2\rho_0)=f_{\rm hi}+\delta_{f,1}\le f_{\rm hi}+\delta_f,\\ &\|\tilde{c}(x)\|\le\sqrt{\tfrac{2(f_{\rm hi}-f_{\rm low}+\gamma)}{\rho_k-2\gamma}+\tfrac{\|\lambda^k\|^2}{(\rho_k-2\gamma)^2}}+\tfrac{\|\lambda^k\|}{\rho_k-2\gamma}\le\sqrt{\tfrac{2(f_{\rm hi}-f_{\rm low}+\gamma)}{\rho_0-2\gamma}+\tfrac{\Lambda^2}{(\rho_0-2\gamma)^2}}+\tfrac{\Lambda}{\rho_0-2\gamma}=\delta_{c,1}\le\delta_c. \end{aligned} \tag{58}$$

此外，根据 (25) 中 $\tilde{c}$ 的定义以及 $\|c(z_{\epsilon_1})\|\le 1$，可知 $\|c(x)\|\le 1+\|\tilde{c}(x)\|$。结合上述不等式和 (24)，可得 $x\in\mathcal{S}(\delta_f,\delta_c)$。因此，引理 4.1 的陈述 (i) 成立。

(ii) 注意到 $\inf_{x\in\mathbb{R}^n}\widetilde{\mathcal{L}}(x,\lambda^k;\ \rho_k)\ =\inf_{x\in\mathbb{R}^n}\{\widetilde{\mathcal{L}}(x,\lambda^k;\ \rho_k):\ \widetilde{\mathcal{L}}(x,\lambda^k;\ \rho_k)\ \le\ \widetilde{\mathcal{L}}(x^k_{\rm init},\lambda^k;\ \rho_k)\}$。因此，要证明引理 4.1 的陈述 (ii)，只需证明

$$\inf_{x\in\mathbb{R}^n}\{\widetilde{\mathcal{L}}(x,\lambda^k;\rho_k):\widetilde{\mathcal{L}}(x,\lambda^k;\rho_k)\le\widetilde{\mathcal{L}}(x^k_{\rm init},\lambda^k;\rho_k)\}\ge f_{\rm low}-\gamma-\Lambda\delta_c. \tag{59}$$

为此，设 x 是满足 $\widetilde{\mathcal{L}}(x,\lambda^k;\ \rho_k)\ \le\ \widetilde{\mathcal{L}}(x^k_{\rm init},\lambda^k;\ \rho_k)$ 的任意一点。由 (58) 可知，$\|\tilde{c}(x)\|\le\delta_c$。据此，结合 $\|\lambda^k\|\le\Lambda$、$\rho_k>2\gamma$ 和 (55)，可得

$$\begin{aligned}&\widetilde{\mathcal{L}}(x,\lambda^k;\rho_k)=f(x)+\gamma\|\tilde{c}(x)\|^2+(\lambda^k)^T\tilde{c}(x)+\tfrac{\rho_k-2\gamma}{2}\|\tilde{c}(x)\|^2\\&\ge f(x)+\gamma\|\tilde{c}(x)\|^2-\Lambda\|\tilde{c}(x)\|\ge f_{\rm low}-\gamma-\Lambda\delta_c,\end{aligned}$$

因此，所需的 (59) 成立。 □

定理 4.2 的证明。假设算法 2 在某次迭代 k 处终止，即 $\tau_k^g\le\epsilon_1$、$\tau_k^H\le\epsilon_2$，且 $\|c(x^{k+1})\|\le\epsilon_1$ 成立。于是，根据 $\tau_k^g\le\epsilon_1$、$\tilde{\lambda}^{k+1}=\lambda^k+\rho_k\tilde{c}(x^{k+1})$、$\nabla\tilde{c}=\nabla c$ 以及 (27) 中的第二个关系，可得

$$\begin{aligned}\|\nabla f(x^{k+1})+\nabla c(x^{k+1})\tilde{\lambda}^{k+1}\|&=\|\nabla f(x^{k+1})+\nabla\tilde{c}(x^{k+1})(\lambda^k+\rho_k\tilde{c}(x^{k+1}))\|\\&=\|\nabla_x\widetilde{\mathcal{L}}(x^{k+1},\lambda^k;\rho_k)\|\le\tau_k^g\le\epsilon_1.\end{aligned}$$

因此，$(x^{k+1},\tilde{\lambda}^{k+1})$ 满足 (5) 中的第一个关系。此外，由 (28) 和 $\tau_k^H\le\epsilon_2$ 可知，$\lambda_{\min}(\nabla^2_{xx}\widetilde{\mathcal{L}}(x^{k+1},\lambda^k;\ \rho_k))\ \ge\ -\epsilon_2$，其成立概率至少为 $1-\delta$，从而对所有 $d\in\mathbb{R}^n$，$d^T\nabla^2_{xx}\widetilde{\mathcal{L}}(x^{k+1},\lambda^k;\ \rho_k)d\ge-\epsilon_2\|d\|^2$ 以至少 $1-\delta$ 的概率成立。利用这一点，结合 $\tilde{\lambda}^{k+1}=\lambda^k+\rho_k\tilde{c}(x^{k+1})$、$\nabla\tilde{c}=\nabla c$ 以及当 $1\le i\le m$ 时的 $\nabla^2\tilde{c}_i=\nabla^2c_i$，可知以下结论以至少 $1-\delta$ 的概率成立：

$$d^T\left(\nabla^2f(x^{k+1})+\sum_{i=1}^m\tilde{\lambda}_i^{k+1}\nabla^2c_i(x^{k+1})+\rho_k\nabla c(x^{k+1})\nabla c(x^{k+1})^T\right)d\ge-\epsilon_2\|d\|^2\ \ \forall d\in\mathbb{R}^n,$$

这意味着对于所有 $d\in\mathcal{C}(x^{k+1})$，$d^T(\nabla^2f(x^{k+1})\ +\sum_{i=1}^m\tilde{\lambda}_i^{k+1}\nabla^2c_i(x^{k+1}))d\ge-\epsilon_2\|d\|^2$，其中 $\mathcal{C}(\cdot)$ 定义于式 (4)。因此，$(x^{k+1},\tilde{\lambda}^{k+1})$ 以至少 $1-\delta$ 的概率满足式 (6)。结合 $\|c(x^{k+1})\|\le\epsilon_1$，我们得出结论：$x^{k+1}$ 是问题 (1) 的确定性 ϵ_1-一阶驻点，并且以至少 $1-\delta$ 的概率是问题 (1) 的 (ϵ_1,ϵ_2)-二阶驻点。因此，定理 4.2 成立。 □

定理 4.3 的证明。由式 (35) 可得 $\rho_{\epsilon_1}\ge2\rho_0$。据此，有

$$K_{\epsilon_1}\overset{(33)}{=}\lceil\log\epsilon_1/\log\omega_1\rceil\overset{(32)}{=}\lceil\log2/\log r\rceil\ \le\ \log(\rho_{\epsilon_1}\rho_0^{-1})/\log r+1. \tag{60}$$

注意到，随着 k 增大，$\{\rho_k\}$ 要么保持不变，要么按比例 r 增大。由此以及 (60)，可知

$$\max_{0\le k\le K_{\epsilon_1}}\rho_k\le r^{K_{\epsilon_1}}\rho_0\overset{(60)}{\le}r^{\frac{\log(\rho_{\epsilon_1}\rho_0^{-1})}{\log r}+1}\rho_0=r\rho_{\epsilon_1}. \tag{61}$$

此外，注意到 $\rho_k>2\gamma$ 和 $\|\lambda^k\|\le\Lambda$。利用这些结果、(22)、(27) 中的第一个关系，以及对 $(x,\lambda,\rho)=(x^{k+1},\lambda^k,\rho_k)$ 应用引理 6.4(ii)，可得

$$\|\tilde{c}(x^{k+1})\|\le\sqrt{\tfrac{2(f_{\rm hi}-f_{\rm low}+\gamma)}{\rho_k-2\gamma}+\tfrac{\|\lambda^k\|^2}{(\rho_k-2\gamma)^2}}+\tfrac{\|\lambda^k\|}{\rho_k-2\gamma}\le\sqrt{\tfrac{2(f_{\rm hi}-f_{\rm low}+\gamma)}{\rho_k-2\gamma}+\tfrac{\Lambda^2}{(\rho_k-2\gamma)^2}}+\tfrac{\Lambda}{\rho_k-2\gamma}. \tag{62}$$

此外，根据 $\|c(z_{\epsilon_1})\|\le\epsilon_1/2$ 以及 (25) 中 $\tilde{c}$ 的定义，可观察到

$$\|c(x^{k+1})\|\le\|\tilde{c}(x^{k+1})\|+\|c(z_{\epsilon_1})\|\le\|\tilde{c}(x^{k+1})\|+\epsilon_1/2. \tag{63}$$

现在证明 $\overline{K}_{\epsilon_1}$ 是有限的。反设 $\overline{K}_{\epsilon_1}$ 是无限的。由此以及 (36) 可得，对所有 $k\ge K_{\epsilon_1}$，$\|c(x^{k+1})\|>\epsilon_1$；结合 (63)，可推出对所有$k\ge K_{\epsilon_1}$，$\|\tilde{c}(x^{k+1})\|>\epsilon_1/2$。进而可知，对无穷多个 k，必有 $\|\tilde{c}(x^{k+1})\|>\alpha\|\tilde{c}(x^k)\|$。利用这一点以及 $\{\rho_k\}$上的更新方案，可推得对无穷多个 k，$\rho_{k+1}=r\rho_k$ 成立；再结合

$\{\rho_k\}$ 的单调性可知，随着 $k\to\infty$，$\rho_k\to\infty$。由此以及 (62) 可见，随着 $k\to\infty$，$\|\tilde{c}(x^{k+1})\|\to 0$，这与对所有 $k\geq K_{\epsilon_1}$，$\|\tilde{c}(x^{k+1})\|>\epsilon_1/2$ 成立这一事实矛盾。因此，$\overline{K}_{\epsilon_1}$ 是有限的。此外，由 (32)、(33) 和 (34) 可知，对所有 $k\geq K_{\epsilon_1}$，$(\tau_k^g,\tau_k^H)=(\epsilon_1,\epsilon_2)$。结合算法 2 的终止准则以及 $\overline{K}_{\epsilon_1}$ 的定义可知，算法 2 必须在第 $\overline{K}_{\epsilon_1}$ 次迭代时终止。

接下来，我们通过下面的两种情形说明，对于 $0\leq k\leq\overline{K}_{\epsilon_1}$，(37) 和 $\rho_k\leq r\rho_{\epsilon_1}$ 成立。

情况 1) $\|c(x^{K_{\epsilon_1}+1})\|\leq\epsilon_1$。由此结合 (36) 可知 $\overline{K}_{\epsilon_1}=K_{\epsilon_1}$，再结合 (60) 和 (61) 可得，对于 $0\leq k\leq\overline{K}_{\epsilon_1}$，(37) 和 $\rho_k\leq r\rho_{\epsilon_1}$ 成立。

情况 2) $\|c(x^{K_{\epsilon_1}+1})\|>\epsilon_1$。由此结合 (36) 可知 $\overline{K}_{\epsilon_1}>K_{\epsilon_1}$，并且对于所有 $K_{\epsilon_1}\leq k\leq\overline{K}_{\epsilon_1}-1$，$\|c(x^{k+1})\|>\epsilon_1$，结合 (63) 可得

$$\|\tilde{c}(x^{k+1})\|>\epsilon_1/2,\quad \forall K_{\epsilon_1}\leq k\leq\overline{K}_{\epsilon_1}-1. \tag{64}$$

由此可知，根据 $\|\lambda^k\|\leq\Lambda$、(22)、(27) 中的第一个关系以及引理 6.4(iv)，取 $(x,\lambda,\rho,\tilde{\delta}_c)=(x^{k+1},\lambda^k,\rho_k,\epsilon_1/2)$，可得

$$\begin{aligned}\rho_k&<8(f_{\rm hi}-f_{\rm low}+\gamma)\epsilon_1^{-2}+4\|\lambda^k\|\epsilon_1^{-1}+2\gamma\\&\leq 8(f_{\rm hi}-f_{\rm low}+\gamma)\epsilon_1^{-2}+4\Lambda\epsilon_1^{-1}+2\gamma\overset{(35)}{\leq}\rho_{\epsilon_1},\quad \forall K_{\epsilon_1}\leq k\leq\overline{K}_{\epsilon_1}-1.\end{aligned} \tag{65}$$

结合这一关系、(61) 以及事实 $\rho_{\overline{K}_{\epsilon_1}}\leq r\rho_{\overline{K}_{\epsilon_1}-1}$，我们得出 $\rho_k\leq r\rho_{\epsilon_1}$ 对于 $0\leq k\leq\overline{K}_{\epsilon_1}$ 成立。剩下需要证明 (37) 成立。为此，令

$$\mathbb{K}=\{k:\rho_{k+1}=r\rho_k,K_{\epsilon_1}\leq k\leq\overline{K}_{\epsilon_1}-2\}.$$

由 (65) 以及 ρ_k 的更新方案可得

$$r^{|\mathbb{K}|}\rho_{K_{\epsilon_1}}=\max_{K_{\epsilon_1}\leq k\leq\overline{K}_{\epsilon_1}-1}\{\rho_k\}\leq\rho_{\epsilon_1},$$

结合 $\rho_{K_{\epsilon_1}}\geq\rho_0$，可得

$$|\mathbb{K}|\leq\log(\rho_{\epsilon_1}\rho_{K_{\epsilon_1}}^{-1})/\log r\leq\log(\rho_{\epsilon_1}\rho_0^{-1})/\log r. \tag{66}$$

令 $\{k_1,k_2,\ldots,k_{|\mathbb{K}|}\}$ 表示按升序排列的 $\mathbb{K}$ 中所有元素，并令 $k_0=K_{\epsilon_1}$ 和 $k_{|\mathbb{K}|+1}=K_{\epsilon_1}-1$。接下来，我们推导 $j=0$ 的 $k_{j+1}-k_j$ 上界，其中 $j=0,1,\ldots,|\mathbb{K}|$。根据 $\mathbb{K}$ 的定义，可知对于 $k_j<k$，$k'\leq k_{j+1}$，有 $\rho_k=\rho_{k'}$。利用这一点以及 ρ_k 的更新方案，我们推得

$$\|\tilde{c}(x^{k+1})\|\leq\alpha\|\tilde{c}(x^k)\|,\quad \forall k_j<k<k_{j+1}. \tag{67}$$

另一方面，由 (31)、(62) 和 $\rho_k\geq\rho_0$，可得 $\|\tilde{c}(x^{k+1})\|\leq\delta_{c,1}$ 对于 $0\leq k\leq\overline{K}_{\epsilon_1}$。由此结合 (64)，可知

$$\epsilon_1/2<\|\tilde{c}(x^{k+1})\|\leq\delta_{c,1},\quad \forall K_{\epsilon_1}\leq k\leq\overline{K}_{\epsilon_1}-1. \tag{68}$$

现在注意到，要么 $k_{j+1}-k_j=1$，要么 $k_{j+1}-k_j>1$。在后一种情况下，我们可以对 $k=k_{j+1}-1$、……、k_j+1 应用 (67)，并结合 (68)，从而推出

$$\epsilon_1/2<\|\tilde{c}(x^{k_{j+1}})\|\leq\alpha\|\tilde{c}(x^{k_{j+1}-1})\|\leq\cdots\leq\alpha^{k_{j+1}-k_j-1}\|\tilde{c}(x^{k_j+1})\|\leq\alpha^{k_{j+1}-k_j-1}\delta_{c,1},\quad \forall j=0,1,\ldots,|\mathbb{K}|.$$

结合这两种情况，可得

$$k_{j+1}-k_j\leq|\log(\epsilon_1(2\delta_{c,1})^{-1})|/\log\alpha|+1,\quad \forall j=0,1,\ldots,|\mathbb{K}|. \tag{69}$$

将这些不等式求和，并利用 (60)、(66)、$k_0=K_{\epsilon_1}$ 和 $k_{|\mathbb{K}|+1}=\overline{K}_{\epsilon_1}-1$，可得

$$\begin{aligned}\overline{K}_{\epsilon_1}&=1+k_{|\mathbb{K}|+1}=1+k_0+\textstyle\sum_{j=0}^{|\mathbb{K}|}(k_{j+1}-k_j)\overset{(69)}{\leq}1+K_{\epsilon_1}+(|\mathbb{K}|+1)\left(\left|\frac{\log(\epsilon_1(2\delta_{c,1})^{-1})}{\log\alpha}\right|+1\right)\\&\leq 2+\frac{\log(\rho_{\epsilon_1}\rho_0^{-1})}{\log r}+\left(\frac{\log(\rho_{\epsilon_1}\rho_0^{-1})}{\log r}+1\right)\left(\left|\frac{\log(\epsilon_1(2\delta_{c,1})^{-1})}{\log\alpha}\right|+1\right)=1+\left(\frac{\log(\rho_{\epsilon_1}\rho_0^{-1})}{\log r}+1\right)\left(\left|\frac{\log(\epsilon_1(2\delta_{c,1})^{-1})}{\log\alpha}\right|+2\right),\end{aligned}$$

其中，第二个不等式由 (60) 和 (66) 得出。因此，在这种情况下，(37) 也成立。 □

接下来我们证明定理 4.4。在继续之前，我们先介绍一些即将用到的记号。令 $L_{k,H}$ 表示 $\nabla^2_{xx}\widetilde{\mathcal{L}}(x,\lambda^k;\ \rho_k)$ 在 $\mathcal{S}(\delta_f,\delta_c)$ 的凸开邻域 $\Omega(\delta_f,\delta_c)$ 上的利普希茨常数，其中 $\mathcal{S}(\delta_f,\delta_c)$ 定义于 (24) 中，并令 $U_{k,H}=\sup_{x\in\mathcal{S}(\delta_f,\delta_c)}\|\nabla^2_{xx}\widetilde{\mathcal{L}}(x,\lambda^k;\ \rho_k)\|$。由 (25) 和 (26) 可知，

$$\nabla^2_{xx}\widetilde{\mathcal{L}}(x,\lambda^k;\rho_k)=\nabla^2 f(x)+\sum_{i=1}^m\lambda_i^k\nabla^2 c_i(x)+\rho_k\left(\nabla c(x)\nabla c(x)^T+\sum_{i=1}^m\tilde{c}_i(x)\nabla^2 c_i(x)\right). \tag{70}$$

由此，$\|\lambda^k\|\le\Lambda$、$\tilde{c}$的定义以及 $\nabla^2 f$ 和 $\nabla^2 c_i$'的利普希茨连续性（见假设 4.1(c)）可知，存在一些常数 L_1、L_2、U_1 和 U_2，它们仅依赖于 f、c、Λ、δ_f 和 δ_c，使得

$$L_{k,H}\le L_1+\rho_k L_2,\quad U_{k,H}\le U_1+\rho_k U_2. \tag{71}$$

定理 4.4 的证明。令 T_k 和 N_k 分别表示算法 2 的第 k 次外层迭代中算法 1 执行的迭代次数和矩阵-向量乘积次数。由定理 4.3 可知，算法 2 中算法 1 执行的迭代总次数和矩阵-向量乘积总次数分别为$\sum_{k=0}^{\overline{K}_{\epsilon_1}}T_k$ 和 $\sum_{k=0}^{\overline{K}_{\epsilon_1}}N_k$。此外，由式（35）和定理 4.3 可知，$\rho_{\epsilon_1}=\mathcal{O}(\epsilon_1^{-2})$ 且$\rho_k\le r\rho_{\epsilon_1}$，由此得到 $\rho_k=\mathcal{O}(\epsilon_1^{-2})$。

我们首先断言，对于任意 $k\ge 0$，均有 $(\tau_k^g)^2/\tau_k^H\ge\min\{\epsilon_1^2/\epsilon_2,\epsilon_2^3\}$。事实上，令 $\bar{t}=\log\epsilon_1/\log\omega_1$，并且对所有 $t\in\mathbb{R}$，$\psi(t)=\max\{\epsilon_1,\omega_1^t\}^2/\max\{\epsilon_2,\omega_2^t\}$。由此可从 (34) 得出 $\omega_1^{\bar{t}}=\epsilon_1$ 和 $\omega_2^{\bar{t}}=\epsilon_2$。结合这一点与 ω_1、$\omega_2\in(0,1)$，可以看出，若 $t\le\bar{t}$，则 $\psi(t)=(\omega_1^2/\omega_2)^t$；否则，$\psi(t)=\epsilon_1^2/\epsilon_2$。再结合 $\epsilon_2\in(0,1)$，可得

$$\min_{t\in[0,\infty)}\psi(t)=\min\{\psi(0),\psi(\bar{t})\}=\min\{1,\epsilon_1^2/\epsilon_2\}\ge\min\{\epsilon_1^2/\epsilon_2,\epsilon_2^3\},$$

由此结合 (32) 可得，对于所有 $k\ge 0$，均有 $(\tau_k^g)^2/\tau_k^H=\psi(k)\ge\min\{\epsilon_1^2/\epsilon_2,\epsilon_2^3\}$。

(i) 根据引理 4.1(i) 以及 $\Omega(\delta_f,\delta_c)$ 和 $L_{k,H}$ 的定义，我们可知，$L_{k,H}$ 是 $\nabla^2_{xx}\widetilde{\mathcal{L}}(x,\lambda^k;\ \rho_k)$ 在 $\{x$ 的某个凸开邻域上的利普希茨常数：$\widetilde{\mathcal{L}}(x,\lambda^k;\ \rho_k)\le\widetilde{\mathcal{L}}(x^k_{\rm init},\lambda^k;\ \rho_k)\}$。此外，回顾引理 4.1(ii)，有 $\inf_{x\in\mathbb{R}^n}\widetilde{\mathcal{L}}(x,\lambda^k;\ \rho_k)\ge f_{\rm low}-\gamma-\Lambda\delta_c$。由此，根据 $\widetilde{\mathcal{L}}(x^k_{\rm init},\lambda^k;\ \rho_k)\le f_{\rm hi}$（见 (30)）以及定理 3.1(iii)，令 $(F_{\rm hi},F_{\rm low},L_H^F,\epsilon_g,\epsilon_H)=(\widetilde{\mathcal{L}}(x^k_{\rm init},\lambda^k;\ \rho_k)\ f_{\rm low}-\gamma-\Lambda\delta_c,L_{k,H},\tau_k^g,\tau_k^H)$，可得

$$\begin{aligned}T_k&=\mathcal{O}((f_{\rm hi}-f_{\rm low}+\gamma+\Lambda\delta_c)L_{k,H}^2\max\{(\tau_k^g)^{-2}\tau_k^H,(\tau_k^H)^{-3}\})\\&\overset{(71)}{=}\mathcal{O}(\rho_k^2\max\{(\tau_k^g)^{-2}\tau_k^H,(\tau_k^H)^{-3}\})=\mathcal{O}(\epsilon_1^{-4}\max\{\epsilon_1^{-2}\epsilon_2,\epsilon_2^{-3}\}),\end{aligned} \tag{72}$$

其中，最后一个等式由 $(\tau_k^g)^2/\tau_k^H\ge\min\{\epsilon_1^2/\epsilon_2,\epsilon_2^3\}$, $\tau_k^H\ge\epsilon_2$ 和 $\rho_k=\mathcal{O}(\epsilon_1^{-2}$ 得出).

接下来，若 $c(x)=Ax-b$ 对某些 $A\in\mathbb{R}^{m\times n}$ 和 $b\in\mathbb{R}^m$ 成立，则 $\nabla c(x)=A^T$ 且 $\nabla^2 c_i(x)=0$ 对于$1\le i\le m$成立。由此以及 (70)，可得 $L_{k,\ H}=\mathcal{O}(1)$。利用这一点并采用与 (72) 类似的论证，可得 $T_k=\mathcal{O}(\max\{\epsilon_1^{-2}\epsilon_2,\epsilon_2^{-3}\})$。由此，结合 (72) 和 $\overline{K}_{\epsilon_1}=\mathcal{O}(|\log\epsilon_1|^2)$（见注 4.4），可知定理 4.4 的陈述 (i) 成立。

(ii) 根据引理 4.1(i) 以及 $U_{k\text{的定义},\ H}$，可以看出

$$U_{k,H}\ge\sup_{x\in\mathbb{R}^n}\{\|\nabla^2_{xx}\widetilde{\mathcal{L}}(x,\lambda^k;\rho_k)\|:\widetilde{\mathcal{L}}(x,\lambda^k;\rho_k)\le\widetilde{\mathcal{L}}(x^k_{\rm init},\lambda^k;\rho_k)\}.$$

利用这一点，$\widetilde{\mathcal{L}}(x^k_{\rm init},\lambda^k;\ \rho_k)\le f_{\rm hi}$ 以及定理 3.1(iv) 与 $(F_{\rm hi},F_{\rm low},L_H^F,U_H^F,\epsilon_g,\epsilon_H)=(\widetilde{\mathcal{L}}(x^k_{\rm init},\lambda^k;\ \rho_k)\ f_{\rm low}-\gamma-\Lambda\delta_c,L_{k,H},U_{k,H},\tau_k^g,\tau_k^H)$，可得

$$\begin{aligned}N_k&=\widetilde{\mathcal{O}}((f_{\rm hi}-f_{\rm low}+\gamma+\Lambda\delta_c)L_{k,H}^2\max\{(\tau_k^g)^{-2}\tau_k^H,(\tau_k^H)^{-3}\}\min\{n,(U_{k,H}/\tau_k^H)^{1/2}\})\\&\overset{(71)}{=}\widetilde{\mathcal{O}}(\rho_k^2\max\{(\tau_k^g)^{-2}\tau_k^H,(\tau_k^H)^{-3}\}\min\{n,(\rho_k/\tau_k^H)^{1/2}\})\\&=\widetilde{\mathcal{O}}(\epsilon_1^{-4}\max\{\epsilon_1^{-2}\epsilon_2,\epsilon_2^{-3}\}\min\{n,\epsilon_1^{-1}\epsilon_2^{-1/2}\}),\end{aligned} \tag{73}$$

其中，最后一个等式由 $(\tau_k^g)^2/\tau_k^H\ge\min\{\epsilon_1^2/\epsilon_2,\epsilon_2^3\}$, $\tau_k^H\ge\epsilon_2$ 得出，且 $\rho_k=\mathcal{O}(\epsilon_1^{-2})$。

另一方面，若假设 c 是仿射的，则由上述讨论可知 $L_{k,\ H}=\mathcal{O}(1)$。利用这一点、$U_{k,\ H}\le U_1+\rho_k U_2$，以及与 (73) 类似的论证，可得 $N_k=\widetilde{\mathcal{O}}(\max\{\epsilon_1^{-2}\epsilon_2,\epsilon_2^{-3}\}\min\{n,\epsilon_1^{-1}\epsilon_2^{-1/2}\})$。由此，结合 (73) 和 $\overline{K}_{\epsilon_1}=\mathcal{O}(|\log\epsilon_1|^2)$（见注 4.4），我们得出定理 4.4 的陈述 (ii) 成立。

□

接下来，我们给出定理 4.5 的证明。为此，我们首先根据假设 4.1(c) 和 4.2 注意到，存在 $U_g^f > 0$、$U_g^c > 0$ 和 $\sigma > 0$，使得

$$\|\nabla f(x)\| \le U_g^f, \quad \|\nabla c(x)\| \le U_g^c, \quad \lambda_{\min}(\nabla c(x)^T \nabla c(x)) \ge \sigma^2, \quad \forall x \in \mathcal{S}(\delta_f, \delta_c). \tag{74}$$

接下来，我们将建立几个稍后会用到的技术引理。

引理 6.5. 假设假设 4.1 和 4.2 成立，并且 ρ_0 足够大，使得 $\delta_{f,\,1} \le \delta_f$且 $\delta_{c,\,1} \le \delta_c$，其中 $\delta_{f,\,1}$ 和 $\delta_{c,\,1}$ 在 (31) 中定义。令 $\{(x^k, \lambda^k, \rho_k)\}$ 由算法 2 生成。假设

$$\rho_k \ge \max\{\Lambda^2(2\delta_f)^{-1}, 2(f_{\rm hi} - f_{\rm low} + \gamma)\delta_c^{-2} + 2\Lambda\delta_c^{-1} + 2\gamma, 2(U_g^f + U_g^c\Lambda + 1)(\sigma\epsilon_1)^{-1}\} \tag{75}$$

对于某个 $k \ge 0$，其中 γ、$f_{\rm hi}$、$f_{\rm low}$、δ_f 和 δ_c 由假设 4.1 给出，而 U_g^f、U_g^c 和 σ 由 (74) 给出。则有 $\|c(x^{k+1})\| \le \epsilon_1$。

证明. 由 (75) 和 $\|\lambda^k\| \le \Lambda$ （见算法 2 的第 6 步），可知

$$\rho_k \ge \max\{\|\lambda^k\|^2(2\delta_f)^{-1}, 2(f_{\rm hi} - f_{\rm low} + \gamma)\delta_c^{-2} + 2\|\lambda^k\|\delta_c^{-1} + 2\gamma\}.$$

利用这一点、(22)、(27) 中的第一个关系式，以及引理 6.4(iii) 和 (iv)（其中 $(x, \lambda, \rho, \tilde\delta_f, \tilde\delta_c) = (x^{k+1}, \lambda^k, \rho_k, \delta_f, \delta_c)$），可得 $f(x^{k+1}) \le f_{\rm hi} + \delta_f$和 $\|\tilde c(x^{k+1})\| \le \delta_c$。此外，回顾 $\|c(z_{\epsilon_1})\| \le 1$ 以及 (25) 中 $\tilde c$的定义，可知 $\|c(x^{k+1})\| \le 1 + \|\tilde c(x^{k+1})\|$。结合 (24)，这些结果表明 $x^{k+1} \in \mathcal{S}(\delta_f, \delta_c)$。进而由 (74) 可得 $\|\nabla f(x^{k+1})\| \le U_g^f$、$\|\nabla c(x^{k+1})\| \le U_g^c$，以及 $\lambda_{\min}(\nabla c(x^{k+1})^T \nabla c(x^{k+1})) \ge \sigma^2$。根据$\|\nabla f(x^{k+1})\| \le U_g^f$、$\|\nabla c(x^{k+1})\| \le U_g^c$、$\tau_k^g \le 1$、$\|\lambda^k\| \le \Lambda$、(25) 和 (27)，可得

$$\begin{aligned}\rho_k\|\nabla c(x^{k+1})\tilde c(x^{k+1})\| &\le \|\nabla f(x^{k+1}) + \nabla c(x^{k+1})\lambda^k\| + \|\nabla_x \widetilde{\mathcal{L}}(x^{k+1}, \lambda^k; \rho_k)\| \\ &\overset{(27)}{\le} \|\nabla f(x^{k+1})\| + \|\nabla c(x^{k+1})\|\|\lambda^k\| + \tau_k^g \le U_g^f + U_g^c\Lambda + 1.\end{aligned} \tag{76}$$

此外，注意到 $\lambda_{\min}(\nabla c(x^{k+1})^T \nabla c(x^{k+1})) \ge \sigma^2$ 意味着 $\nabla c(x^{k+1})^T \nabla c(x^{k+1})$ 可逆。利用这一事实和 (76)，我们得到

$$\begin{aligned}\|\tilde c(x^{k+1})\| &\le \|(\nabla c(x^{k+1})^T \nabla c(x^{k+1}))^{-1} \nabla c(x^{k+1})^T\|\|\nabla c(x^{k+1})\tilde c(x^{k+1})\| \\ &= \lambda_{\min}(\nabla c(x^{k+1})^T \nabla c(x^{k+1}))^{-\frac{1}{2}}\|\nabla c(x^{k+1})\tilde c(x^{k+1})\| \overset{(76)}{\le} (U_g^f + U_g^c\Lambda + 1)/(\sigma\rho_k).\end{aligned} \tag{77}$$

我们还从 (75) 得出 $\rho_k \ge 2(U_g^f + U_g^c\Lambda + 1)(\sigma\epsilon_1)^{-1}$，结合 (77) 可证明 $\|\tilde c(x^{k+1})\| \le \epsilon_1/2$。将其与 (25) 中 $\tilde c$ 的定义以及 $\|c(z_{\epsilon_1})\| \le \epsilon_1/2$ 结合起来，我们得出所需的 $\|c(x^{k+1})\| \le \epsilon_1$ 成立。

□

Th下一个引理为 $\{\rho_k\}$ 提供了一个比定理 4.3 中更紧的上界。

引理 6.6。假设假设 4.1 和 4.2 成立，且 ρ_0 足够大，使得 $\delta_{f,1} \le \delta_f$ 且 $\delta_{c,1} \le \delta_c$，其中 $\delta_{f,1}$ 和 $\delta_{c,1}$ 在 (31) 中定义。设 $\{\rho_k\}$ 由算法 2 生成，并且

$$:\overset{\rho_{\epsilon_1}}{=}\max\{\Lambda^2(2\delta_f)^{-1}, 2(f_{\rm hi} - f_{\rm low} + \gamma)\delta_c^{-2} + 2\Lambda\delta_c^{-1} + 2\gamma, 2(U_g^f + U_g^c\Lambda + 1)(\sigma\epsilon_1)^{-1}, 2\rho_0\} \tag{78}$$

其中，γ、$f_{\rm hi}$、$f_{\rm low}$、δ_f和 δ_c 由假设 4.1给出，而 U_g^f、U_g^c和 σ 由(74)给出。于是，对$0 \le k \le \overline{K}_{\epsilon_1}$，$\rho_k \le r\tilde\rho_{\epsilon_1}$ 成立，其中 $\overline{K}_{\epsilon_1}$ 定义于(36)。

证明。由(78)可得 $\tilde\rho_{\epsilon_1} \ge 2\rho_0$。由此并采用与(60)类似的论证，可得 $K_{\epsilon_1} \le \log(\tilde\rho_{\epsilon_1}\rho_0^{-1})/\log\, r + 1$，其中 K_{ϵ_1}定义于(33)。利用这一结果、针对 $\{\rho_k\}$的更新方案以及与(61)类似的论证，可得

$$\max_{0 \le k \le K_{\epsilon_1}} \rho_k \le r\tilde\rho_{\epsilon_1}. \tag{79}$$

若 $\|c(x^{K_{\epsilon_1}+1})\| \le \epsilon_1$，则由(36)可得$\overline{K}_{\epsilon_1} = K_{\epsilon_1}$，结合(79)可推出 $\rho_k \le r\tilde\rho_{\epsilon_1}$对 $0 \le k \le \overline{K}_{\epsilon_1}$成立。另一方面，若 $\|c(x^{K_{\epsilon_1}+1})\| > \epsilon_1$，则由(36)可得 $\|c(x^{k+1})\| > \epsilon_1$ 对$K_{\epsilon_1} \le k \le \overline{K}_{\epsilon_1} - 1$成立。结合引理 6.5和(78)，可得对所有 $K_{\epsilon_1} \le k \le \overline{K}_{\epsilon_1} - 1$，

$$\rho_k < \max\{\Lambda^2(2\delta_f)^{-1}, 2(f_{\rm hi} - f_{\rm low} + \gamma)\delta_c^{-2} + 2\Lambda\delta_c^{-1} + 2\gamma, 2(U_g^f + U_g^c\Lambda + 1)(\sigma\epsilon_1)^{-1}\} \overset{(78)}{\le} \tilde\rho_{\epsilon_1}.$$

由此，结合 (79) 和 $\rho_{\overline{K}_{\epsilon_1}} \le r\rho_{\overline{K}_{\epsilon_1}-1}$，我们还可得出 $\rho_k \le r\tilde\rho_{\epsilon_1}$对 $0 \le k \le \overline{K}_{\epsilon_1}$成立。 □

定理 4.5 的证明。由 (78) 和引理 6.6 可知，$\tilde{\rho}_{\epsilon_1} = \mathcal{O}(\epsilon_1^{-1})$ 且 $\rho_k \leq r\tilde{\rho}_{\epsilon_1}$，由此得到$\rho_k = \mathcal{O}(\epsilon_1^{-1})$。因此，结合这一结果以及定理 4.4 的证明中相同的论证，只需将 $\rho_k = \mathcal{O}(\epsilon_1^{-2})$ 替换为 $\rho_k = \mathcal{O}(\epsilon_1^{-1})$，即可得出定理 4.5 的结论。□

7 未来工作

这项工作还有若干值得开展的后续研究。首先，将我们的增广拉格朗日方法扩展到求解非凸优化问题在不等式约束或更一般约束下的近似二阶驻点，会很有意义。事实上，对于带不等式约束的非凸优化问题，可以利用平方松弛变量将其重新表述为等式约束问题（例如，参见 [7]）。可以证明，后一个问题的二阶驻点会导出原问题的弱二阶驻点；此外，如果原问题满足线性无关约束规格，那么后一个问题也满足该条件。因此，将我们的增广拉格朗日方法应用于等价的等式约束问题，以求得不等式约束问题的近似弱二阶驻点，前景可期。其次，值得研究能否在较弱的约束品性条件下推导第 4.3 节中的增强复杂度结果（例如，参见 [5]）。第三，我们的增广拉格朗日方法建立在一个较强的假设之上，即已知问题的一个近似可行解。如果能通过修改该方法（可能借助不可行性检测技术，例如参见 [19]）来消除这一假设，该方法便可适用于更广泛的问题。最后，开展更多数值研究将有助于从实际应用的角度进一步改进我们的增广拉格朗日方法。

参考文献

[1] N. Agarwal, Z. Allen-Zhu, B. Bullins, E. Hazan, and T. Ma, 比梯度下降更快地寻找近似局部极小点，载于《第 49 届 ACM SIGACT 年度计算理论研讨会论文集》，2017年，第 1195–1199 页。[2] R. Andreani, E. G. Birgin, J. M. Martínez, and M. L. Schuverdt, 关于具有一般下层约束的增广拉格朗日方法，SIAM 优化杂志，18 (2008 年)，第 1286–1309 页。[3] R. Andreani, G. Haeser, and J. M. Martínez, 关于光滑约束优化的序列最优性条件，优化，60 (2011 年)，第 627–641 页。[4] R. Andreani, G. Haeser, A. Ramos, and P. J. Silva, 一种与优化算法收敛相关的二阶序列最优性条件，IMA 数值分析杂志，37 (2017 年)，第 1902–1929 页。[5] R. Andreani, G. Haeser, M. L. Schuverdt, and P. J. Silva, 两种新的弱约束资格条件及其应用，SIAM 优化杂志，22 (2012 年)，第 1109–1135 页。[6] P. Armand and N. N. Tran, 一种具有快速不可行性检测能力的等式约束优化增广拉格朗日方法，《优化理论与应用杂志》，181 (2019年)，第 197–215 页。[7] D. P. Bertsekas,《非线性规划》，Athena Scientific，1999 年。[8] W. Bian, X. Chen, and Y. Ye, 非 Lipschitz 非凸极小化内点算法的复杂度分析，《数学规划》，149 (2015 年)，第 301–327 页。[9] E. G. Birgin, J. Gardenghi, J. M. Martínez, S. A. Santos, and P. L. Toint, 使用未缩放 KKT 条件和高阶模型进行非线性约束优化的评估复杂度，SIAM 优化杂志，26 (2016 年)，第 951–967 页。[10] E. G. Birgin, G. Haeser, and A. Ramos, 具有约束子问题且收敛到二阶驻点的增广拉格朗日方法，《计算优化与应用》，69 (2018 年)，第 51–75 页。[11] E. G. Birgin and J. M. Martínez,《约束优化的实用增广拉格朗日方法》，SIAM，2014年。

[12] E. G. Birgin and J. M. Martínez, 在无约束优化中使用带三次下降条件的二次正则化，SIAM 优化杂志，27 (2017 年)，第 1049–1074 页。[13] E. G. Birgin and J. M. Martínez, 一种增广拉格朗日算法的复杂度与性能，优化方法与软件，35 (2020年)，第 885–920 页。
[14] J. F. Bonnans and G. Launay, 对位移进行惩罚的序列二次规划，SIAM 优化杂志，5 (1995年)，第 792–812 页。[15] N. Boumal, V. Voroninski, and A. S. Bandeira, 非凸 Burer-Monteiro 方法适用于光滑半定规划，载于《神经信息处理系统进展》，第 29 卷，2016 年，第 2757–2765 页。
[16] L. F. Bueno and J. M. Martínez, 约束优化非精确恢复方法的复杂度，SIAM 优化杂志，30 (2020年)，第 80–101 页。[17] S. Burer and R. D. C. Monteiro, 一种通过低秩分解求解半定规划的非线性规划算法，《数学规划》，95 (2003年)，第 329–357 页。[18] S. Burer and R. D. C. Monteiro, 低秩半定规划中的局部极小值与收敛，《数学规划》，103 (2005年)，第 427–444 页。[19] J. V. Burke, F. E. Curtis, and H. Wang, 一种具有快速不可行性检测功能的序列二次优化算法，SIAM 优化杂志，24 (2014年)，第 839–872 页。[20] R. H. Byrd, F. E. Curtis, and J. Nocedal, 非线性优化中的不可行性检测与序列二次规划方法，SIAM 优化杂志，20 (2010年)，第 2281–2299 页。[21] R. H. Byrd, R. B. Schnabel, and G. A. Shultz, 一种非线性约束优化信赖域算法，SIAM 数值分析杂志，24 (1987年)，第 1152–1170 页。[22] Y. Carmon and J. C. Duchi, 梯度下降可求得三次正则化非凸牛顿步，SIAM 优化杂志，29 (2019年)，第 2146–2178 页。[23] Y. Carmon, J. C. Duchi, O. Hinder, and A. Sidford, “在证明有罪之前，皆为凸”：非凸函数上梯度下降的无维数依赖加速，载于《国际机器学习会议》，PMLR，2017年，第 654–663 页。[24] Y. Carmon, J. C. Duchi, O. Hinder, and A. Sidford, 非凸优化的加速方法，SIAM 优化杂志，28 (2018 年)，第 1751–1772 页。[25] C. Cartis, N. I. Gould, and P. L. Toint, 无约束优化的自适应三次正则化方法。第二部分：最坏情况下的函数与导数评估复杂度，《数学规划》，130 (2011 年)，第 295–319 页。[26] C. Cartis, N. I. Gould, and P. L. Toint, 潜在秩亏非线性最小二乘问题的三次正则化方法评估复杂度及其与约束非线性优化的相关性，SIAM 优化杂志，23 (2013年)，第 1553–1574 页。[27] C. Cartis, N. I. Gould, and P. L. Toint, 约束非线性优化中寻找一阶临界点的复杂度，《数学规划》，144 (2014年)，第 93–106 页。[28] C. Cartis, N. I. Gould, and P. L. Toint, 使用二阶方法进行约束非线性最小二乘及一般约束非线性优化的评估复杂度，SIAM 数值分析杂志，53 (2015 年)，第 836–851 页。[29] C. Cartis, N. I. Gould, and P. L. Toint, 使用缩放 KKT 条件、高阶模型和临界性度量 χ 进行光滑约束非线性优化的评估复杂度界，载于《逼近与优化》，Springer，2019年，第 5–26 页。

[30] C. Cartis, N. I. Gould, and P. L. Toint, 一阶至三阶及更高阶最优性：约束非凸优化中的刻画与评估复杂度，J. Complex., 53 (2019), pp. 68–94.[31] X. Chen, L. Guo, Z. Lu, and J. J. Ye, 非 Lipschitz 非凸规划的增广拉格朗日方法，SIAM J. Numer. Anal., 55 (2017), pp. 168–193.
[32] D. Cifuentes and A. Moitra, Burer-Monteiro 方法的多项式时间保证，arXiv preprint arXiv:1912.01745, (2019).[33] T. F. Coleman, J. Liu, and W. Yuan, 一种新的等式约束优化信赖域算法，Comput. Optim. Appl., 21 (2002), pp. 177–199.[34] F. E. Curtis, D. P. Robinson, C. W. Royer, and S. J. Wright, 具有强二阶复杂度保证的非凸优化信赖域 Newton-CG，SIAM J Optim., 31 (2021), pp. 518–544.[35] F. E. Curtis, D. P. Robinson, and M. Samadi, 一种用于非凸优化的信赖域算法，其最坏情况迭代复杂度为 $\mathcal{O}(\epsilon^{-3/2})$，Math. Program., 162 (2017), pp. 1–32.
[36] F. E. Curtis, D. P. Robinson, and M. Samadi, 等式约束优化信赖漏斗算法的复杂度分析，SIAM J. Optim., 28 (2018), pp. 1533–1563.[37] G. N. Grapiglia and Y. Yuan, 关于非凸优化增广拉格朗日方法的复杂度，IMA J. Numer. Anal., 41 (2021), pp. 1508–1530.[38] G. Haeser, H. Liu, and Y. Ye, 线性约束优化的最优性条件与复杂度分析：边界处不可微的情形，Math. Program., (2019), pp. 1–37.
[39] M. R. Hestenes, 乘子法与梯度法，J. Optim. Theory Appl., 4 (1969), pp. 303–320.[40] M. Hong, D. Hajinezhad, and M.-M. Zhao, Prox-PDA：用于网络上的快速分布式非凸优化与学习的近端原始-对偶算法，载于国际机器学习会议，PMLR, 2017, pp. 1529–1538.[41] C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan, 如何高效逃离鞍点，载于国际机器学习会议，PMLR, 2017, pp. 1724–1732.[42] C. Kanzow and D. Steck, 标准增广拉格朗日方法与带保护机制的增广拉格朗日方法比较示例，Oper. Res. Lett., 45 (2017), pp. 598–603.[43] W. Kong, J. G. Melo, and R. D. C. Monteiro, 求解线性约束非凸复合规划的二次罚函数加速非精确近端点法的复杂度，SIAM J. Optim., 29 (2019), pp. 2566–2593.[44] J. Kuczyński and H. Woźniakowski, 通过幂算法和 Lanczos 算法结合随机初始值估计最大特征值，SIAM J. Matrix Anal. Appl., 13 (1992), pp. 1094–1122.
[45] Z. Li, P.-Y. Chen, S. Liu, S. Lu, and Y. Xu, 约束非凸优化的收敛速率改进型非精确增广拉格朗日方法，载于国际人工智能与统计会议，PMLR, 2021, pp. 2170–2178.[46] S. Lu, 一种用于非凸函数约束优化的单循环梯度下降与扰动上升算法，载于国际机器学习会议，PMLR, 2022, pp. 14315–14357.[47] S. Lu, M. Razaviyayn, B. Yang, K. Huang, and M. Hong, 在光滑非凸线性约束优化问题中高效寻找二阶驻点，神经信息处理系统进展，33 (2020), pp. 2811–2822。

[48] Z. Lu and X. Li, 通过部分正则化实现稀疏恢复：模型、理论与算法，Math. Oper. Res., 43 (2018), pp. 1290–1316.[49] Z. Lu and Y. Zhang, 稀疏主成分分析的增广拉格朗日方法，Math. Program., 135 (2012), pp. 149–193.[50] J. M. Martínez and M. Raydan, 无约束最小化中变范数信赖域方法的三次正则化对应方法，J. Glob. Optim., 68 (2017), pp. 367–385.[51] J. G. Melo, R. D. Monteiro, and W. Kong, 基于经典拉格朗日函数和完整拉格朗日乘子更新的内层加速非精确近端增广拉格朗日方法的迭代复杂度，arXiv preprint arXiv:2008.00562, (2020).
[52] Y. Nesterov and B. T. Polyak, 牛顿法的三次正则化及其全局性能，Math. Program., 108 (2006), pp. 177–205.[53] J. Nocedal and S. J. Wright, 数值优化，Springer, 第 2 版，2006.
[54] M. O'Neill and S. J. Wright, 具有复杂度保证的有界约束优化对数障碍牛顿共轭梯度法，IMA J. Numer. Anal., 41 (2021), pp. 84–121.[55] R. T. Rockafellar, 拉格朗日乘子与最优性，SIAM revie w, 35 (1993), pp. 183–238.[56] C. W. Royer, M. O'Neill, and S. J. Wright, 一种具有复杂度保证的光滑无约束优化牛顿共轭梯度算法，Math. Program., 180 (2020), pp. 451–488.
[57] C. W. Royer and S. J. Wright, 光滑非凸优化二阶线搜索算法的复杂度分析，SIAM J. Optim., 28 (2018), pp. 1448–1477.[58] M. F. Sahin, A. Eftekhari, A. Alacaoglu, F. Latorre, and V. Cevher, 一种带非线性约束的非凸优化非精确增广拉格朗日框架，神经信息处理系统进展，32 (2019).[59] Y. Xie and S. J. Wright, 有界约束优化投影牛顿法的复杂度，arXiv preprint arXiv:2103.15989, (2021).[60] Y. Xie and S. J. Wright, 带非线性等式约束的非凸优化近端增广拉格朗日法的复杂度，J. Sci. Comput., 86 (2021), pp. 1–30.[61] L. Yang, D. Sun, and K. C. Toh, SDPNAL+: 一种用于带非负约束半定规划的主化半光滑牛顿共轭梯度增广拉格朗日方法，Math. Program. Comput., 7 (2015), pp. 331–366.[62] X. Zhao, D. Sun, and K. C. Toh, 一种用于半定规划的牛顿共轭梯度增广拉格朗日方法，SIAM J. Optim., 20 (2010), pp. 1737–1765。

附录

一种截断共轭梯度法

本部分介绍 [56, 算法 1] 提出的截断共轭梯度方法，该方法用于求解线性方程组 (12) 的近似解，或寻找其对应矩阵 H 的充分负曲率方向；这一方法已在第 3.1 节中简要讨论。其详细内容见 [56, 第 3.1] 节。

下面的定理给出了算法 3 的迭代复杂度。

定理 A.1（算法 3 的迭代复杂度）。考虑将输入 $U = 0$用于线性方程组 (12)，其中 $g \neq 0$、$\varepsilon > 0$ 和 H 是一个 $n \times n$ 对称矩阵。则算法 3 的迭代次数为 $\widetilde{\mathcal{O}}(\min\{n, \sqrt{\|H\|/\varepsilon}\})$。

算法 3 截断共轭梯度法

Inputs: 对称矩阵 $H \in \mathbb{R}^{n\times n}$、向量 $g \neq 0$、阻尼参数 $\varepsilon \in (0,1)$、期望相对精度 $\zeta \in (0,1)$。
Optional input: 标量 $U \geq 0$（如果未提供，则设为 0）。
Outputs: d 类型，d。
Secondary outputs: U、κ、$\widehat{\zeta}$、τ 和 T 的最终值。
Set

$$\bar{H} := H + 2\varepsilon I, \quad \kappa := \frac{U+2\varepsilon}{\varepsilon}, \quad \widehat{\zeta} := \frac{\zeta}{3\kappa}, \quad \tau := \frac{\sqrt{\kappa}}{\sqrt{\kappa}+1}, \quad T := \frac{4\kappa^4}{(1-\sqrt{\tau})^2},$$

$y^0 \leftarrow 0$、$r^0 \leftarrow g$、$p^0 \leftarrow -g$、$j \leftarrow 0$.
如果 $(p^0)^T \bar{H} p^0 < \varepsilon \|p^0\|^2$ 则
　设置 $d \leftarrow p^0$，并以 d 类型 = 负曲率; 终止
否则如果 $\|Hp^0\| > U\|p^0\|$ 则
　设置 $U \leftarrow \|Hp^0\|/\|p^0\|$，并更新 $\kappa, \widehat{\zeta}$、τ、T相应地;
结束 if
当 TRUE 时执行
　$\alpha_j \leftarrow (r^j)^T r^j / (p^j)^T \bar{H} p^j$; {开始标准共轭梯度操作}
　$y^{j+1} \leftarrow y^j + \alpha_j p^j$;
　$r^{j+1} \leftarrow r^j + \alpha_j \bar{H} p^j$;
　$\beta_{j+1} \leftarrow \|r^{j+1}\|^2 / \|r^j\|^2$;
　$p^{j+1} \leftarrow -r^{j+1} + \beta_{j+1} p^j$; {结束标准 CG 操作}
　$j \leftarrow j+1$;
　如果 $\|Hp^j\| > U\|p^j\|$，则
　　设置 $U \leftarrow \|Hp^j\|/\|p^j\|$ 并更新 κ, $\widehat{\zeta}$、τ、T 相应地;
　结束如果
　如果 $\|Hy^j\| > U\|y^j\|$，则
　　设置 $U \leftarrow \|Hy^j\|/\|y^j\|$ 并更新 κ, $\widehat{\zeta}$、τ、T相应地;
　结束如果
　如果 $\|Hr^j\| > U\|r^j\|$，则
　　设置 $U \leftarrow \|Hr^j\|/\|r^j\|$ 并更新 κ, $\widehat{\zeta}$、τ、T相应地;
　结束条件判断
　如果 $(y^j)^T \bar{H} y^j < \varepsilon \|y^j\|^2$ 则
　　设置 $d \leftarrow y^j$ 并以 d 型 = 负曲率终止;
　否则如果 $\|r^j\| \leq \widehat{\zeta}\|r^0\|$ 则
　　设置 $d \leftarrow y^j$，并以 d 类型 = SOL 终止;
　否则，如果 $(p^j)^T \bar{H} p^j < \varepsilon \|p^j\|^2$，则
　　设置 $d \leftarrow p^j$，并以 d 类型 = 负曲率终止;
　否则，如果 $\|r^j\| > \sqrt{T}\tau^{j/2}\|r^0\|$，则
　　按照上述主循环中的方法计算 α_j、y^{j+1};
　　求出 $i \in \{0,\ldots,j-1\}$，使得

$$(y^{j+1}-y^i)^T \bar{H}(y^{j+1}-y^i) < \varepsilon \|y^{j+1}-y^i\|^2;$$

　　设置 $d \leftarrow y^{j+1}-y^i$ 并以 d 类型 = 负曲率终止;
　结束 if
结束 while

证明。由 [56, 引理 1], 可知，算法 3 的迭代次数至多为 $\min\{n, J(U,\varepsilon,\zeta)\}$，其中 $J(U,\varepsilon,\zeta)$ 是满足 $\sqrt{T}\tau^{J/2} \leq \widehat{\zeta}$ 的最小整数 J，且 U、$\widehat{\zeta}$、T 和 τ 是算法 3 返回的值。此外，第 3.1] 节中已证明 $J(U,\varepsilon,\zeta) \leq \left\lceil \left(\sqrt{\kappa}+\frac{1}{2}\right) \ln \left({}^{144}{}_{(\sqrt{\kappa}+1)\zeta^2} \quad {}^{2}\kappa^{6} \right) \right\rceil$，其中 $\kappa = \mathcal{O}(U/\varepsilon)$ 是算法 3 的一个输出。于是可知 $J(U,\varepsilon,\zeta) = \widetilde{\mathcal{O}}(\sqrt{U/\varepsilon})$。注意算法 3 中的输出 $U \leq \|H\|$。综合以上结果，即可得到所需结论。 □

B 基于随机化 Lanczos 的最小特征值预言机

本部分介绍 [56, 第 3.2], 节提出的随机化 Lanczos 方法，该方法可用作算法 1 的最小特征值预言机。正如第 3.1 节中简要讨论的，该预言机以高概率输出 H 的一个足够负的曲率方向，或者给出一份证书，证明 H 近似为半正定。有关该方法的更多动机和解释，请参见 [56, 第 3.2] 节。

下面的定理证明，算法 4 是算法 1 的合适的最小特征值预言机。其证明与 [56, 引理 2] 完全相同，因此省略。

算法 4 基于随机化 Lanczos 的最小特征值预言机

输入：对称矩阵 $H \in \mathbb{R}^{n\times n}$、容差 $\varepsilon > 0$ 和概率参数 $\delta \in (0,1)$。输出：一个满足 $v^T H v \le -\varepsilon/2$ 和 $\|v\| = 1$ 的充分负曲率方向 v；或者一个证书，证明 $\lambda_{\min}(H) \ge -\varepsilon$ 的概率至少为 $1-\delta$。从单位球面上均匀随机生成一个向量作为初始向量，应用 Lanczos 方法 [44] 估计 $\lambda_{\min}(H)$，并至多运行

$$N(\varepsilon,\delta) := \min\left\{ n, 1 + \left\lceil \frac{\ln(2.75n/\delta^2)}{2}\sqrt{\frac{\|H\|}{\varepsilon}} \right\rceil \right\} \tag{80}$$

迭代。如果在某次迭代中找到满足 $v^T H v \le -\varepsilon/2$ 的单位向量 v，则立即终止并返回 v。

定理 B.1（算法 4 的迭代复杂度）。考虑以容差 $\varepsilon > 0$、概率参数 $\delta \in \in (0,1)$ 和对称矩阵 $H \in \mathbb{R}^{n\times n}$ 作为输入的算法 4。则该算法要么找到满足 $v^T H v \le -\varepsilon/2$ 和 $\|v\| = 1$ 的充分负曲率方向 v，要么以至少 $1-\delta$ 的概率，在至多 $N(\varepsilon,\delta)$ 次迭代内证明 $\lambda_{\min}(H) \ge -\varepsilon$ 成立，其中 $N(\varepsilon,\delta)$ 在式 (80) 中定义。

请注意，算法 4 需要 $\|H\|$。一般而言，当 n 较大时，计算 $\|H\|$ 的成本可能很高。尽管如此，仍可通过随机化方案高效地估计 $\|H\|$，并获得较高置信度（例如，参见附录 B3] 中的讨论 [56, ）。