



学校代码：10272

学 号：2021310197

上海财经大学

SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

博士学位论文

DOCTORAL DISSERTATION

论文题目：齐次二阶优化方法：模型，算法及应用

作者姓名：张楚文

院(系所) 信息管理与工程学院

专 业 管理科学与工程

指导教师 葛冬冬

完成日期 2025 年 4 月 11 日

## 学位论文原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品成果。对本人的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。


论文作者签名：，日期：2025 年 5 月 15 日

## 学位论文版权使用授权书 (博士学位论文用)

本人完全了解上海财经大学关于收集、保存、使用学位论文的规定，即：按照有关要求提交学位论文的印刷本和电子版本。上海财经大学有权保留并向国家有关部门或机构送交本论文的复印件和扫描件，允许论文被查阅和借阅。本人授权上海财经大学可以将学位论文的全部或部分内容编入有关数据库进行检索和传播，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。

☒ 本年度公开。

☐ 延迟一年公开。延迟理由：\_\_\_\_\_

论文作者签名：，

导师签名：

日期：2025 年 5 月 15 日

日期： 年 月 日

## 摘要

二阶优化方法是一类强大的优化算法，相较于一阶方法不仅能够实现更快的收敛速度，而且对问题的条件数不那么敏感。

本论文提出了一个基于“齐次模型”二阶优化框架。该框架最早出现在我关于“齐次二阶下降法”(HSODM)<sup>[176]</sup>的论文中，该方法考虑了一个梯度—海森的聚合矩阵以及对应的  $n + 1$  维提升问题。“齐次化”的优点在于每次迭代仅需计算聚合矩阵的最左特征向量（而不是计算线性方程组）。因此，与其他二阶方法相比，该算法是一个易于实现的单循环方法。

接下来本论文讨论了使用齐次模型及其相应框架的优势。具体而言，我们展示了在病态问题中，求解齐次模型与求解线性方程相比，使用 Krylov 子空间方法时能够获得更好的条件数；若采用 Lanczos 方法，则条件数始终有界。我们给出了条件数的估计，并证明了该估计的紧致性。

从这些理解出发，接下来将原始的“齐次二阶下降法”抽象为一个“齐次二阶下降框架”(HSODF)，在该框架下，可以设计出各类基于齐次模型及特征值问题的二阶优化方法。比如，本文提出了一种同伦方法来求解结构化凸优化问题，并证明该方法在一些退化问题中仍能具有线性收敛速度。

论文的最后部分致力于将所提出的框架应用于内点法。我们开发了一种基于齐次模型的内点法，用以求解 Fisher 交换市场模型。我们对常数替代弹性 (CES) 效用函数及其最优响应映射的二阶性质进行了分析，据我所知，这些分析是第一次出现。通过充分利用 Fisher 市场的低秩特性，我们开发了一种“无分解”内点方法，其迭代代价与一阶方法（如 tâtonnement 方法）类似，数值实验可以证明该方法的优越性。

**关键词：** 齐次模型；二阶优化；凸优化；非凸优化；数值优化；内点法；交换市场。

## Abstract

Second-order optimization methods are a class of powerful optimization algorithms that are favorable over first-order methods because they can achieve faster convergence rates compared to first-order methods, and they are less sensitive to conditioning of the problems. This thesis is devoted to a novel second-order optimization framework based on the “Homogeneous Models”.

We begin our construction by introducing the original “Homogeneous Second-Order Descent Method” (HSODM)<sup>[176]</sup> that considers the gradient-Hessian aggregated matrix and a corresponding  $n+1$  dimensional lifted problem. The merit of “homogenization” is that only the leftmost eigenvector of the aggregated matrix is computed at each iteration. Therefore, the algorithm is a single-loop method that is easy to implement compared with other second-order methods.

We discuss the benefits of using homogeneous models and the corresponding framework. Specifically, we show that in ill-conditioned problems, homogeneous models can be solved with better conditioning compared to solving a linear equation by Krylov subspace methods. The condition number is always bounded if using a Lanczos method. The idea is extended beyond the original line-search HSODM and we propose a “Homogeneous Second-Order Descent Framework” (HSODF) that provides a unified framework for developing second-order optimization methods. A homotopy method is also proposed to solve the structured convex optimization problems.

The last part of the thesis is devoted to the application of the proposed framework to interior-point methods (IPMs). We develop an IPM based on homogeneous models to solve Fisher exchange market models. An analysis of second-order properties of the constant elasticities of substitution (CES) utilities and their best-response mappings are provided and appear to be novel. By properly exploiting the low-rank properties of Fisher market, we develop a “factorization-free” IPM that has the same iteration cost as the first-order methods such the tâtonnement methods.

**Keywords: Homogeneous Models; Second-Order Optimization; Convex Optimization; Nonconvex Optimization; Numerical Optimization; Interior-Point Methods; Exchange Market.**

## 致谢

2019 年，我在疫情期间不幸被封控在了湖北老家。当年夏天，膝盖就不幸受伤，不得不进行韧带手术。正是这些人生中发生的诸多意外，才给了我继续求学的契机。

首先，我应该感谢我在交大，也是 UT Austin 的学妹袁媛，如果没有她的介绍，我或许根本不可能加入杉数，更不会想到将来会有这样的机会继续自己的学术理想。她不仅在工作上成为了我最可靠的伙伴，更在生活上给了我极大帮助。也正是这样的机会，使我结识了很多志同道合，意气相投的朋友，主要是张澍民、苏广俊、黄翔、陶杨懿、林甜甜、仲思潼、李硕森（还有很多）。

我非常有幸在读博期间，认识了很多才华横溢的同学，尤其是同在葛冬冬老师课题组朝夕相处的伙伴，主要是杨静远、刘天浩、刘劲松、高文智、梁阔、赵相婕、浦善文、张碗钰，这种良好的学术环境让我受益匪浅。我非常有幸能与其中很多人一起合作，比如南航项目上，与静远从零开始学习随机规划，最后论文发表在了 EJOR 上；也要感谢在芝大商学院访问期间她对我的诸多帮助。还有和善文、天浩、相婕一起讨论铁路规划问题，最后文章发表在了 IEEE TPAMI 上。与劲松和天浩一起写出了 GPU 架构的线性规划一阶求解器，这个工作最终成为 COPT 的一部分，同时我们的开源工作受到很多工业界和学术界的关注（英伟达的优化求解器就是基于这个工作）。与文智以及碗钰主要是讨论一些理论问题，由于大家研究的内容不尽相同，这些讨论不经意间开拓了我的视野。在葛老师课题组外，我还结识了两位长期合作者，何畅和姜云天。我们在江波老师的指导下，在非线性优化领域发现了很多有意思课题（我们的合作论文已经快 10 篇了）。

我要感谢江波老师，他是我在非线性优化领域的引路人。他在学术写作、科研选题上给了我很多帮助，他的奉献精神和学者之风教会我应该如何做一名合格的科研工作者。我要感谢杉数的 CTO 王子卓教授，是他和葛老师推荐我跟随叶荫宇老师求学。王老师鼓励我在入学前就提前接触一些课题，并在百忙之中不厌其烦地帮我修改手稿。正是他帮助让我完成了一名算法工程师到一名学术研究者的转变。我非常感谢我的导师葛冬冬教授，感谢他读博期间的慷慨资助，给我提供最好的工作和学习机会，使我有机会与世界最顶尖学者交流合作。我要感谢他不断催促我参加各类学术会议，形成良好的科研规划。我要感谢他给予我自由探索的学术环境，让我可以接触很多极具挑战性的业界难题，并不断鞭策我追求卓越。这些问题不经意间锻炼了我的技术能力，启发我对优化理论的兴趣。最后，我要感谢我的另一位导师叶荫宇教授，我非常有幸在过去五年的时间内受到他的指导。作为运筹学的泰斗，是他的论文在塑造我的研究风格方面起到了关键作用。他的传奇经历、永不停歇的学术热情以及对科研的执着追求，都深深影响了我。

我要感谢我的父母，以及岳父岳母，感谢他们的理解、耐心和支持。最后，我要感谢我的妻子

明雨，她是我精神的依靠，正是她无微不至的关怀和无私的奉献，才让我可以心无旁骛地追求自己的学术理想，让这段美妙的旅程得以延续。

2025 年 5 月 14 日

攻读博士学位期间取得的研究成果

已发表(包括已接受待发表)的论文，以及已投稿、或已成文打算投稿、或拟成文投稿的论文情况(只填写与学位论文内容相关的部分)：

序号	作者(全体作者，按顺序排列)	题目	发表或投稿刊物名称、级别	发表的卷期、年月、页码	相当于学位论文的哪一部分(章、节)	被索引收录情况
1	Zhang Chuwen, Ge Dongdong, He Chang, Jiang Bo, Jiang Yuntian, Ye Yinyu	A Homogeneous Second-Order Descent Method for Nonconvex Optimization	A, Mathematics of Operations Research	已接收	第 3 章	
2	He Chang, Jiang Yuntian, Zhang Chuwen, Ge Dongdong, Jiang Bo, Ye Yinyu	Homogeneous Second-Order Descent Framework: A Fast Alternative to Newton-Type Methods	A, Mathematical Programming	已接收	第 4 章 至 第 6 章	
3	Zhang Chuwen, Chang He, Dongdong Ge, Jiang Bo, Yinyu Ye	Price Updates by a Homogeneous Interior-Point Method	拟投稿		第 7 章	

# 目 录

摘要 .....	3
Abstract .....	4
致谢 .....	5
攻读博士学位期间取得的研究成果 .....	7
插图目录 .....	12
表格目录 .....	13
第一章 绪论 .....	1
1.1 引言 .....	1
1.2 研究内容 .....	4
第二章 文献综述 .....	6
2.1 复杂度理论 .....	6
2.2 优化算法的评价体系 .....	7
2.3 二阶算法及其复杂度理论 .....	9
2.4 二阶算法在凸优化及结构化问题中的理论分析 .....	11
2.5 二阶子问题的求解方法 .....	12
2.6 Lanczos 方法以及随机求解技术 .....	12
2.6.1 Lanczos 方法的误差界及基本概念 .....	12
2.6.2 Kaniel–Paige 收敛理论 .....	14
2.6.3 Kuczyński 估计 .....	16
第三章 齐次二阶下降法 .....	21
3.1 介绍 .....	21
3.1.1 本章贡献 .....	22
3.2 齐次二次模型与二阶下降方法 .....	23
3.2.1 齐次化的动机 .....	23
3.2.2 方法概述 .....	24
3.2.3 齐次二次模型的初步分析 .....	25
3.3 全局收敛速率 .....	29
3.3.1 关于 $\ d_k\ $ 大值情况的分析 .....	29



3.3.2	关于 $\ d_k\ $ 小值情况的分析	35
3.3.3	全局收敛性	37
3.4	局部收敛率	38
3.5	非精确 HSODM	41
3.5.1	Lanczos 方法概述	41
3.5.2	非精确 HSODM 的概述	43
3.5.3	带有偏随机化的定制 Lanczos 方法	46
3.5.4	小值情况下的非精确 HSODM	49
3.5.5	非精确 HSODM 的全局收敛性分析	51
3.6	数值实验	53
3.6.1	实现细节	54
3.6.2	CUTEst 数据集中的无约束问题	54
3.7	结论	56
	本章附录	57
3.A	本章附加证明	57
3.A.1	对引理 3.26 的证明	57
3.A.2	对引理 5.30 的证明	57
3.A.3	对定理 3.31 的证明	58
3.B	CUTEst 数据集的详细结果	58
第四章	广义齐次模型及下降框架	76
4.1	简介	76
4.2	齐次二阶下降框架	77
4.2.1	HSODF 的动机	78
4.2.2	原始-对偶解的刻画	82
4.2.3	在二阶算法中使用 GHM	87
	本章附录	89
4.A	第 4.2.1 节的主要证明	89
4.A.1	技术引理	89
4.A.2	基本结果	90
4.A.3	对定理 4.2 的证明	92

4.B 第 4.2.2 节的主要证明	93
4.B.1 对引理 4.5 的证明	93
4.B.2 对引理 4.8 的证明	94
4.B.3 对引理 4.10 的证明	94
4.B.4 对引理 4.12 的证明	95
4.C 第 4.2.3 节的主要证明	96
4.C.1 对定理 4.14 的证明	96
4.C.2 对定理 4.15 的证明	97
第五章 自适应的齐次二阶法	99
5.1 方法概述	99
5.2 收敛性分析	102
5.2.1 全局收敛	102
5.2.2 局部收敛性	107
5.2.3 关于困难情况的讨论	109
5.3 齐次框架中的二分法	113
5.3.1 基于 $h_k$ 的二分法	113
5.4 数值实验	117
5.4.1 CUTEst 基准测试	117
本章附录	119
5.A 基于非精确特征值的自适应齐次二阶法	119
5.A.1 收敛性分析	121
5.A.2 二分法的复杂度	123
5.A.3 处理困难情况	126
第六章 齐次同伦法	131
6.1 自协 Lipschitz 函数与同伦模型	131
6.1.1 自协 Lipschitz 函数的基本性质	131
6.1.2 同伦 HSODM	136
6.2 收敛性分析	137
6.2.1 近似居中条件的性质	137
6.2.2 GHM 的基本特性	138

6.3 数值实验 .....	144
6.3.1 $\ell_2$ 正则化逻辑回归 .....	144
第七章 齐次框架在交换市场中的应用 .....	147
7.1 介绍 .....	147
7.1.1 研究动机 .....	148
7.1.2 相关工作 .....	149
7.2 方法概述 .....	149
7.2.1 消费者理论的预备知识 .....	149
7.2.2 基于齐次模型的 Barrier 算法 .....	150
7.2.3 通过特征值问题更新迭代 .....	152
7.3 对 CES 效用函数的新分析 .....	154
7.3.1 对角线加秩一 (Diagonal plus rank-one, DR1) 近似 .....	156
7.4 价格更新算法的收敛性分析 .....	158
7.5 数值实验 .....	161
本章附录 .....	164
7.A CES 效用函数下自协性质的证明 .....	164
7.A.1 证明 引理 7.7 .....	165
7.A.2 证明 引理 7.8 .....	165
7.A.3 证明 引理 7.9 .....	165
7.A.4 证明 引理 7.11 .....	167
7.A.5 证明 定理 7.12 .....	168
7.B 齐次模型的相关证明 .....	170
7.B.1 证明 定理 7.4 .....	170
7.B.2 证明 推论 7.5 .....	171
第八章 总结与展望 .....	172
参考文献 .....	173

## 插图目录

1-1	论文结构 . . . . .	5
2-1	验证 $f(t) = \frac{1}{\log(t+\sqrt{t^2-1})} - \sqrt{\frac{2}{t-1}} \leq 0$ . . . . .	16
3-1	用于 CUTEst 问题的二阶方法性能曲线。(a) 报告了迭代次数的性能。(b) 包括了梯度计算的结果；仅包含使用 Krylov 子空间的方法。 . . . . .	55
4.2.1	$\kappa(H_k + \epsilon_N I)$ 和 $\kappa_L(F_k)$ 在退化情况下的比较。 . . . . .	80
4.2.2	计算扰动 Hilbert 矩阵和 GHM 的牛顿型方向的结果。蓝色到绿色的条形图表示 Krylov 迭代次数。图例中显示的数字对应于不同的正则化 $\epsilon_N$ 。 . . . . .	81
4.2.3	当 $g_k \perp \mathcal{S}_1(H_k)$ 时的“扰动”情形示意图，其中 $\phi_k = g_k + \varepsilon \cdot u_1, u_1 \in \mathcal{S}_1(H_k)$ 以提供更清晰的展示。图中标注了拐点 $\overline{\delta_k^{\text{cvx}}}$ 和 $\tilde{\alpha}_1$ 。 . . . . .	86
5.4.1	CUTEst 问题的性能概况。 . . . . .	118
6.3.1	$\ell_2$ 正则化逻辑回归问题上不同 SOM 方法的性能表现。 . . . . .	145
6.3.2	$\ell_2$ 正则化逻辑回归问题上 Homotopy-HSODM 采用热启动策略的性能表现。 . . . . .	146
7.3.1	估计 $\mathbf{P}\nabla^2 f(\mathbf{p})\mathbf{P}$ 的误差随玩家数目 $ \mathcal{I} $ 的变化 . . . . .	157
7.5.1	算法 7-1 在不同弹性系数下的比较，其中 . . . . .	162
7.5.2	不同算法在 $m = 200, n = 100$ 时的收敛情况 (迭代数) . . . . .	162
7.5.3	不同算法在 $m = 200, n = 100$ 时的收敛情况 (时间) . . . . .	163

## 表格目录

3-1 几种二阶算法的简要比较。这里， $p \in (0, 1)$ 表示随机 Lanczos 方法的失败概率。在最后一列中，我们使用“E”表示极端特征值问题，使用“N”表示牛顿型方程。 . . . . .	23
3-2 不同算法在 CUTEst 数据集上的 SGM 性能。注意 $\bar{t}_G, \bar{k}_G$ 是缩放后的几何平均值 (分别以 1 秒和 50 次迭代为单位缩放)。若某实例求解失败，其迭代次数和求解时间设为 20,000。 . . . . .	55
3.B.1 Abbreviations of the Methods . . . . .	58
3.B.2 CUTEst 数据集的完整结果，迭代和时间 . . . . .	59
3.B.3 CUTEst 数据集的完整结果，函数值和梯度范数 . . . . .	69
4.2.1 线性最小二乘问题中计算一个牛顿型方向或 GHM 的数据集详情及 Krylov 迭代次数的平均值。 . . . . .	83
4.2.2 利用齐次框架通过自适应 $\delta_k, \phi_k$ 来恢复或替代其他二阶方法。内循环复杂度 $\mathcal{T}_k$ 表示与外部迭代 $k$ 相关的内部迭代次数上界。 . . . . .	88
5.4.1 CUTEst 基准测试结果汇总 (81 个实例)。 . . . . .	118
7.5.1 不同问题规模下，算法 7-1 的运行时间。* 表示在 100.0 秒内未收敛到容差，括号内表示停机时的梯度范数大小。 . . . . .	163



## 第一章 绪论

## 第一节 引言

数学优化问题是一类考虑对有限资源进行规划,并使得某种目标函数最优的重要数学问题。有限维线性空间  $\mathbb{E}$  中的优化问题可以写成非常一般的形式:  $\min_x f(x)$ ,  $x \in \mathcal{X}$ , 其中  $\mathcal{X} \subseteq \mathbb{E}$ ,  $f: \mathbb{E} \mapsto \mathbb{R}$  是目标函数。现代优化理论起源于二十世纪,尤其是运筹学诞生之后,学术界和工业界不断发现各种富有实际意义的优化问题。优化问题的类型丰富,按照目标函数的性质和约束条件来分,可以分为有约束、无约束,凸、非凸,线性、非线性,光滑、非光滑等等。其中,线性规划可能是最早被关注的问题。对线性规划较为正式的研究起于 1930 年代后期,苏联数学家 Leonid Kantorovich 和美国经济学家 Wassily Leontief 分别在制造计划和经济学领域首次认真尝试了线性规划方法的应用。至二十世纪中叶,三位数学家 Harold Kuhn, Albert Tucker, 和 William Karush 发现了著名的 KKT 条件,使得人们开始关注非线性的优化目标或者资源约束,这类问题广泛出现在化工,经济学,统计学等领域。时至今日,优化问题仍然在各类科学、工程、经济、金融、统计学等领域发挥着重要作用。譬如训练多层深度神经网络,优化供应链,航空调度,强化学习等等。

早在计算机和运筹学出现之前,数学家已经对基于导数的优化方法作了基本阐述。为方便深入地讨论算法,考虑  $f: \mathbb{E} \mapsto \mathbb{R}$  是  $p$ -阶光滑函数。当函数  $f(x)$  连续  $p$  次可导时,我们可以利用导数信息设计算法。事实上,早在计算机和运筹学出现之前,数学家已经发现了基于导数的优化方法。比如,当  $p = 1$  时,可以计算梯度  $\nabla f(x) = \frac{d}{dx} f(x)$ ,柯西 (Augustin-Louis Cauchy) 在 1847 年已经对梯度法有了比较深入的阐述<sup>[30,107]</sup>。有趣的是,基于二阶导数  $\nabla^2 f(x) = \frac{d}{dx} \nabla f(x)$  的求解方法出现的时间更早。牛顿 (Isaac Newton) 早在 1669 年就提出了牛顿法<sup>[129]</sup>,最早用于求解非线性方程。然而,虽然牛顿给出了基本思想,但他的方法不同于我们日常所知的现代牛顿法。他只将该方法应用于多项式,没有明确地将该方法与导数联系起来。随后,Joseph Raphson<sup>[143]</sup> 在 1690 年提出了简化了这种思想,并归纳成可重复使用的迭代表达式。如果函数具有更高阶的导数信息 —  $p \geq 3$  — 则可以利用泰勒展开式构造更高阶的优化算法,这类方法一般称为张量优化算法<sup>[88,89,126]</sup>,这里我们不再详细介绍。这些方法是当代优化方法的理论基石。

可以列举如下常见的几个无约束优化问题:

**例 1.1** ( $\ell_2$  正则化逻辑回归). 在二分类问题中,输出变量取值于离散空间。对于二分类问题,预测变量只有两个取值,即  $\{-1, 1\}$ 。我们假设有  $N$  组数据对  $a_i \in \mathbb{R}^n, b_i \in \{-1, 1\}, i = 1, \dots, N$ 。  $\ell_2$  正则化逻辑回归可以写为如下的优化问题:

$$\min_{x \in \mathbb{R}^n} \sum_{i=1}^N \ln \left( 1 + \exp \left( -b_i \cdot a_i^T x \right) \right) + \gamma/2 \|x\|_2^2.$$

该问题是一个凸的非线性优化问题，其中  $\mathcal{X} = \mathbb{R}^n$ 。

**例 1.2 (策略优化问题)**. 在强化学习中，我们常常考虑求解一个复杂环境中的随机策略，记为  $\pi$ 。一般地，我们需要考虑状态空间  $\mathcal{S}$  和动作集合  $\mathcal{A}$ ，在一次决策过程中，根据策略  $\pi$  将形成了一个长度不定的随机决策路径，称为  $\tau$ 。如果策略由一个带参函数  $\pi_\theta$  表示，那么对于奖励函数  $r$  的优化问题可以写作：

$$\begin{aligned} \pi_\theta(a|s) : \mathcal{S} \times \mathcal{A} &\rightarrow [0, 1] \\ \max_{\theta \in \mathbb{R}^d} L(\theta) &:= \mathbb{E}_{\tau \sim \pi_\theta} [r(\tau)] \end{aligned}$$

以上都是当下数据科学中非常流行的问题。尽管这些问题是无约束的 – 如  $\mathcal{X} = \mathbb{R}^n$  – 但时至大数据时代， $n$  的维度往往变得很大，这使得这些问题在计算上变得非常困难。我们随后将进一步理解这些问题的计算难度。在此之前，我们介绍一些  $\mathcal{X}$  为闭凸集的情况。比如线性约束的约束集合：

$$\mathcal{X} = \{x \in \mathbb{E} : a(x) = 0, x \in \mathcal{K}\},$$

其中  $a(x) = Ax - b$  为仿射函数， $A : \mathbb{E} \mapsto \mathbb{R}^m$  为线性算子， $b \in \mathbb{R}^m$ ， $\mathcal{K} \subseteq \mathbb{E}$  是一个闭凸锥。常见的情况如  $\mathcal{K} = \mathbb{R}_+^n$ ，即非负象限； $\mathcal{K} = \{x \in \mathbb{R}^{n+1} : x_1 \geq \|x_{[2:n+1]}\|\}$ ，即 Lorentz 锥。这类问题在日常生活中非常常见，比如在经济学中，Fisher 交换市场问题就是一个经典的例子。

**例 1.3 (Fisher 交换市场问题)**. 在经济学中，Fisher 交换市场问题是一个经典的优化问题，更广的模型称为 Arrow-Debreu 模型<sup>[11,158]</sup>。考虑一个市场，其中有  $m$  个消费者及  $n$  种可分的商品 (divisible goods)，不妨假设每种商品的总量为 1。每个消费者  $i$  具有一个初始财富  $w_i$  以及一个效用函数  $u_i$ 。每个消费者  $i$  的需求为  $\mathbf{x}_i \in \mathbb{R}_+^n$ ，并获得效用  $u_i(\mathbf{x}_i)$ 。一般我们假设效用函数是凹函数。假设市场中的商品价格为  $\mathbf{p} \in \mathbb{R}_+^n$ ，则消费者  $i$  的预算约束写作  $\langle \mathbf{p}, \mathbf{x}_i \rangle \leq w_i$ 。市场的均衡定义为这样的需求-价格对， $(\mathbf{x}_1^*, \dots, \mathbf{x}_m^*, \mathbf{p}^*)$ ，即在某价格  $\mathbf{p}$  时，每个参与者  $i$  均达到最大化效用 (Utility Maximization)，且市场上的商品需求等于供给。

$$\mathbf{x}_i^* \in \arg \min_{\mathbf{x}_i \in \mathbb{R}_+^n} -u_i(\mathbf{x}_i), \text{ s.t. } \langle \mathbf{p}, \mathbf{x}_i \rangle \leq w_i, \quad \forall i = 1, \dots, m, \quad (1-1a)$$

$$\sum_{i=1}^m x_{i,j}^* = 1, \quad \forall j = 1, \dots, n. \quad (1-1b)$$

直接求解以上问题具有很高的难度，一般需要利用变分不等式技术<sup>[60]</sup>，因此一个更简单的方法是优化 Eisenberg-Gale (EG) 势函数：

$$\max \Psi(\mathbf{x}_1, \dots, \mathbf{x}_m) := \sum_i w_i \log(u_i(\mathbf{x}_i)) \quad (1-2a)$$

$$\text{s.t. } \sum_i \mathbf{x}_i \leq \mathbf{1}, \quad (1-2b)$$

$$\mathbf{x}_i \in \mathbb{R}_+^n, \forall i = 1, \dots, m. \quad (1-2c)$$



Eisenberg and Gale<sup>[57]</sup> 证明了以上凸问题的解可以构造出 Fisher 模型的均衡。

在互联网蓬勃发展的时代，线上的交易市场问题往往具有非常大的规模，以至于消费者的数量可以达到百万级，甚至千万级。为了大数据时代带来的规模问题，比较直接的方法是基于梯度及次梯度的一阶算法。这些方法已经在一些主流的开源软件中被高效地实现。比如深度学习，强化学习等领域，一阶方法一般都内置在各类软件平台上，如著名的 Google TensorFlow<sup>[1]</sup> 和 PyTorch 库<sup>[140]</sup>。一阶方法的优势是**实现简单，单次迭代的复杂度较低，容易实现并行化**等。但同时具有明显的缺陷：比如收敛到的解（一阶稳定点）效果欠佳，**迭代复杂度高，比较依赖问题的条件数**等。然而，除了**维度问题，优化问题还具有更棘手的、各式各样的难点**。

比如 例 1.1，虽然目标函数是凸函数，但该问题往往具有很强的退化 (Degenerate, 有的文献亦称 Singular) 特点 (特别是  $\gamma$  较小时)。一种情况是特征数远大于样本数，即  $n \gg N$ ；另一种情况是每组数据对  $(a_i, b_i)$  的非零元过少。这两种情况都容易导致  $a_i$  形成的矩阵  $A = [a_1, \dots, a_N] \in \mathbb{R}^{n \times N}$  退化，使得问题的条件数  $\kappa(A) = \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)}$  过大，这对于一阶算法非常不友好。

又比如，在 例 1.3 中，我们可以证明 Eisenberg-Gale 对偶问题的势函数是无限次可微的，但该函数的梯度、Hessian 信息并不具有全局 Lipschitz 连续性，更好反映函数特点的性质是自协性 (Self-Concordance)。这些问题都比较适合利用牛顿类算法 (二阶算法) 解决。

除此之外，对于有约束的问题，由于二阶算法可以收敛到二阶稳定点，同时在局部具有更快的收敛速度，这意味着二阶算法能够求得精度更高的解。如线性规划问题，(基于牛顿法)的内点法可以高效地求到精度达  $10^{-8}$  的解，对于商业求解器，往往可以达到  $10^{-12}$  的精度<sup>[2,66]</sup>；与之相比，一阶算法的有效精度往往停留在  $10^{-4}$ <sup>[47,131,133]</sup>。这种精度的差距在一些大型工程系统的实际问题中是难以接受的。比如，在解的可行性要求很高的情况下 (如发电机启停规划<sup>[13]</sup>)，这要求优化算法的原始可行性 (对偶梯度) 求解精度较高。又如，在一些目标函数值较大，又无法简单缩放的问题中，对解的最优性要求很高，如一些供应链优化问题<sup>[141]</sup>，这些目标函数往往指代企业的运营成本，即使与最优解仅有 1% 的差距，这样的解也无法真正落地使用。在线性规划问题上，现有研究尚可通过线性可行系统的锐度 (sharpness) 等误差界性质 (error bound)<sup>[8,9]</sup>，结合 GPU 硬件，得到较大幅度的提升<sup>[108,109]</sup>。但对于非线性规划问题，当前更为稳定的方法仍然是基于信赖域、Filter 线搜索等技术的牛顿类算法<sup>[22,160]</sup>，这类算法在 GPU 上也已有一些加速的效果<sup>[134,135,150]</sup>。

由于二阶算法在数学性质、实践效果上具有非常迷人的优点，但其在理论上、计算上尚有一些技术问题亟待解决，如何完善二阶算法的理论和计算仍然是一个非常前沿性的研究难点。在本论文中，我们的目标是开发高效的**二阶优化方法**，这些方法具有全局迭代复杂度保证。我们着重探索二阶算法在大规模问题上的改进措施，提出一些新的二阶算法。在理论上，我们证明这些方法的有效性。在工程实现上，我们将广泛地从各个领域，挖掘二阶算法的可用场景，在数值上提升其求解

效率。

### 第二节 研究内容

本论文的主要内容可以分为如下几个部分：

- (a) 第2章 主要介绍一些预备知识及概念，包括一些优化问题的复杂度概念，以及一些基础的数值线性代数中的结论。考虑到篇幅，我们不对论文中涉及的凸分析、最优性条件、稳定点等概念进行详细介绍，这些定义我们主要参考文献<sup>[111,125]</sup>。只有极少的情况下，我们需要借助非光滑分析中的结论<sup>[145]</sup>。对于内点法的预备知识，我们将在所涉及的章节作简要介绍，力求简洁，这些结论现在已经非常成熟，可以参考文献<sup>[14,127,165]</sup>。
- (b) 在第3章中，我们提出一种新的二阶算法：齐次二阶下降法 (Homogeneous Second-Order Descent Method, HSODM)。这个方法主要受半正定规划的启发，利用齐次化技术，将非齐次的局部二次估计转化为齐次问题(我们称为 Ordinary Homogeneous Model, OHM)；这意味着，原先的线性方程组可以利用对称特征值问题进行求解。这个思路的出发点主要是结合数值线性代数中负曲率的随机求解技术<sup>[103]</sup> (复杂度为  $O(n^2\epsilon^{-0.25})$ )，以替换更为昂贵的牛顿步。后者如果使用矩阵分解，需要  $O(n^3)$  的复杂度<sup>注1</sup>。在现有文献中<sup>[23,44,147]</sup>，负曲率技术只能作为一个子模块运行，为保证整体运行效率，这些方法需要反复在特征值问题和牛顿步之间切换。与之相比，齐次二阶法是第一个完全利用负曲率运行的算法，可以大大降低工程难度。

本章主要依据如下文章：

**Zhang C**, Ge D, He C, Jiang B, Jiang Y, Xue C, Ye Y. A Homogeneous Second-Order Descent Method for Nonconvex Optimization. *Mathematics of Operations Research* (2025, to appear).

- (c) 在第4章中，我们进一步讨论齐次模型的拓展性，我们发现，齐次化技术不仅可以用于非凸优化，还可以以此设计一个广义的二阶算法框架，我们称之为 Homogeneous Second-Order Descent Framework (HSODF)。在这个框架下，各类牛顿算法中的线性方程组可以被替换为齐次问题。为此，我们提出广义齐次模型 (Generalized Homogeneous Model, GHM)，对其特征值及参数控制方法进行详尽的分析。我们证明求解齐次问题所需的条件数与牛顿步的条件数是不同的，在退化情况下，齐次问题条件数远远小于原问题条件数；且总是有界的。
- (d) 在第5章中，基于广义齐次模型及下降框架，我们对原来的 HSODM 进行改进，我们称之为自适应 (adaptive) 的 HSODM，这个方法不需要预先对 Lipschitz 常数进行估计，同时可以推广到凸优化中。相应地，我们引入一个辅助的二分法，用来对齐次模型的参数进行调节。这个方法

<sup>注1</sup> 有一些迭代方法可以在  $O(n^2)$  量级下运行，我们随后在第4章中会详细讨论。

具有基本的  $O(\log(\frac{1}{\epsilon}))$  的迭代复杂度<sup>注2</sup>。值得注意的是，这个思路可以支持 HSODF 框架下的任意搜索过程。

- (e) 在第6章中, 我们考虑一种同伦 (Homotopy) 算法<sup>[127,144,154]</sup>, 用来计算一些退化现象非常严重的情况 (比如极稀疏的例1.1). 谓之齐次同伦法 (Homotopy HSODM). 对于该类问题, 我们证明了一个某种自协条件, 使得齐次同伦法可以在退化的凸问题中具有全局线性收敛速度; 这些问题不是强凸的, 甚至不是严格凸的。

以上第4章至第6章基于如下文章:

He C, Jiang Y, **Zhang C**, Ge D, Jiang B, Ye Y. Homogeneous Second-Order Descent Framework: A Fast Alternative to Newton-Type Methods. Mathematical Programming (2025, to appear).

- (f) 在第7章中, 我们考虑将齐次化技术应用于 Fisher 交换市场问题。我们证明了 Fisher 交换市场问题的最优响应映射是自协的, 并利用该性质设计了一种基于齐次模型的内点法, 并证明该方法的收敛性, 据我们所知, 这些性质在文献中是首次出现。同时, 由于 Fisher 交换市场问题的低秩特性, 我们证明, 该问题<sup>注3</sup>的 Hessian 矩阵可以被一个对角阵 + 秩一矩阵<sup>注4</sup>估计, 在高概率下成立, 我们证明所需玩家数是  $O(\frac{1}{\epsilon})$ . 因此, 我们设计了一种“无需矩阵分解”内点法, 可以显式地得到逆矩阵, 其迭代代价与一阶方法相同。

以上第7章基于如下拟投稿文章:

**Zhang C**, He C, Jiang B, Ge D, Ye Y (2025) Price Updates by a Homogeneous Interior-Point Method

章节间的关系可以用如下框图表示:

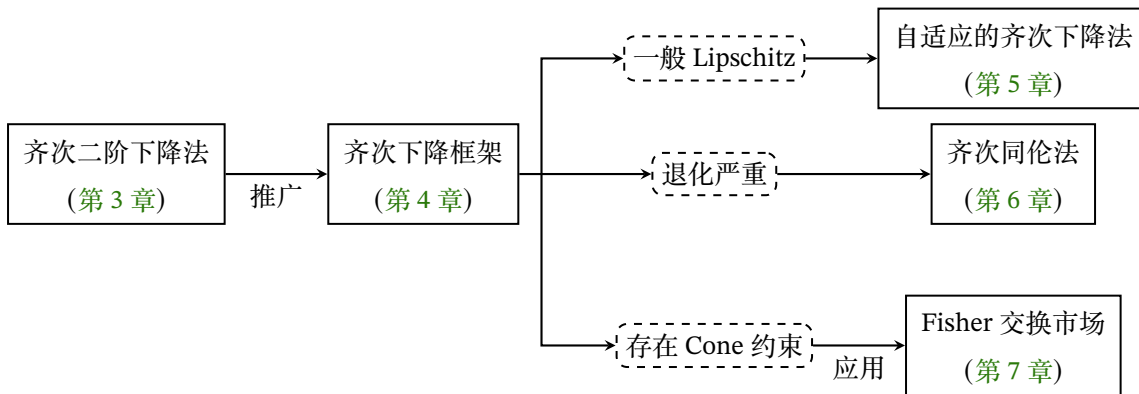


图 1-1 论文结构

最后, 论文的总结与展望见第8章。

<sup>注2</sup>显然, 最简单的二分法也具有  $O(\log(\frac{1}{\epsilon}))$  的复杂度

<sup>注3</sup>准确的说, 是 Eisenberg-Gale 势函数的对偶问题

<sup>注4</sup>故称 DR1 (Diagonal + Rank-1) 估计

## 第二章 文献综述

## 第一节 复杂度理论

复杂度理论是算法理论的基石。该理论的目标有两个：一是建立衡量各种算法有效性的标准（从而能够使用这些标准来比较算法），二是评估各种问题的固有难度。复杂度理论可以参考一些经典教材，如<sup>[136]</sup>。限于优化算法讨论，我们主要参考了文献<sup>[14,76,165]</sup>。

对于优化算法而言，为做好复杂度分析，我们需要定义问题模型  $\mathcal{P}$  和算法模型  $\mathcal{A}$ 。复杂度分析的目标不仅仅是考虑算法在单个算例（即对问题  $\mathcal{P}$  赋予具体的数据  $X$  后的某个具体  $p$ , instance）的表现，而是对一类具有共性的问题进行讨论，如：无约束的优化问题，目标函数满足一阶光滑条件；凸优化问题，约束是一组仿射函数；离散时间的马尔科夫决策问题，折算因子 (discount factor) 小于 1 等等。算法复杂度指的是算法  $\mathcal{A}$  在  $\mathcal{P}$  这一类问题上的普遍表现。同时，我们并不关心某个特定编程语言实现的程序在特定计算机上的执行时间，因为其中涉及太多偶然因素。粗略地说，进行复杂性分析需要确定以下三点：输入规模 (input size) 的定义，基本运算 (basic operations) 的集合，以及每种基本运算的成本 (cost)。后两项共同决定了计算的成本。基本运算的选择通常较为简单。一般的优化算法自然适用于以下运算集合： $\{+, -, \times, /, \leq\}$  即四种算术运算和比较运算。相比之下，输入规模和运算成本的选择则更为微妙，取决于算法处理的数据类型。一些数据类型可以存储在固定大小的内存中，而另一些数据的存储大小则因数据本身的取值而变化。一般分为两种，固定大小数 (Fixed-Precision Data): 通常以 32 位或 64 位存储。对于这类数据，每个元素的大小通常设为 1，即单位大小 (unit size); 可变大小数据 (Variable-Precision Data): 如对整数而言，其存储所需的位数大致等于其绝对值的对数，即“位大小” (bit-size); 类似地，有理数的存储需求也依赖于其分子和分母的位大小。设  $X = (x_1, \dots, x_n) \in \mathbb{R}^n$  表示某个为长度为  $n$  的数据向量：

- (a) 若  $X$  属于固定大小数据，则定义输入规模： $\text{size}(X) = n$ 。
- (b) 若  $X$  属于可变大小数据，则定义输入规模： $\text{size}(X) = \sum_{i=1}^n \text{bit-size}(x_i)$ 。

类似的考虑也适用于算术运算的成本。当运算涉及两个单位大小的数时，计算成本记为 1。在位大小的情况下：乘法与除法的成本是其操作数位大小的乘积。加法、减法和比较的成本是其操作数位大小的最大值。将整数或有理数数据及其相应的位大小和位成本考虑在内，通常称为图灵计算模型 (Turing model of computation)<sup>[136]</sup>，这种方式一般称为 size-based，需要考虑输入每个元素的位大小。现在更普遍的是对于理想化实数，假设每个输入具有单位大小和单位成本，我们只关心基本运算的次数，被称为算术计算模型 (Arithmetic model of computation)，亦称实数模型 (Real number model)，比如 BSS 计算模型<sup>[18]</sup>。



不论是对 Turing 模型还是算术模型，我们都可以定义所谓的“多项式时间算法”。我们可以通过高斯消元法简单比较。该算法需要对实数进行多项式次算术运算，因此在实数模型中它是多项式的；但是，中间计算中使用的数字可能呈指数增长，因此它在图灵机模型中的运行时间是指数的。在优化问题中，尤其是非线性优化问题，很难保证问题数据和解是有理数<sup>注1</sup>。因此以下我们讨论算数模型，如下定义多项式时间较为适宜。对问题  $\mathcal{P}$ ，算法  $\mathcal{A}$  的复杂度记为关于问题类型，数据大小，问题容差的函数

$$T_{\mathcal{A}}(\mathcal{P}, X, \epsilon).$$

其中  $X$  依旧表示输入数据， $\epsilon$  表示某种“问题容差”，即算法输出结果与“解”的误差。如果对于任意  $X$ ， $\epsilon > 0$ ， $T_{\mathcal{A}}$  的上界是关于  $\text{size}(X), \log(\frac{1}{\epsilon})$  的多项式，则我们称算法  $\mathcal{A}$  是多项式时间算法 (Polynomial-Time Algorithm)。如果与  $\epsilon$  无关，则称算法  $\mathcal{A}$  是强多项式时间算法 (Strongly Polynomial-Time Algorithm)。如果是关于  $\text{size}(X), \frac{1}{\epsilon}$  的多项式，则称算法  $\mathcal{A}$  是一个 FPTAS (Fully Polynomial-Time Approximation Scheme)。如果只能保证对固定的  $\epsilon$ ， $T_{\mathcal{A}}$  是关于  $\text{size}(X)$  的多项式，则称算法  $\mathcal{A}$  是一个 PTAS (Polynomial-Time Approximation Scheme)。在优化问题的概念下，这种考虑复杂度的方式一般认为是算术复杂度 (Arithmetic Complexity)。这种定义直接与复杂度理论相关。

## 第二节 优化算法的评价体系

对于一般优化问题，我们无法保证可以得到“最优解”，因此我们对“解”的定义需要放宽，所谓问题容差  $\epsilon$  也需要相应的定义。这个概念的含义是比较宽泛的，比如，对凸优化问题，可以定义为  $f(x) - f(x^*) \leq \epsilon$ 。一般的问题，我们只能考虑的解是满足某种“稳定性条件”的点。比如对于光滑优化问题，我们考虑的解是满足一阶稳定点条件或二阶稳定点条件的点。对非光滑问题，则需要 Clarke 稳定点。

分析优化算法时，我们一般需要先验证算法是否具有全局收敛性，以证明算法在某种程度上是“良定义”的，可以输出正确的结果。这种分析方式出现在较为经典的文献中，比如利用 Zoutendijk 条件分析一个无穷的算法序列  $\{x_k\}_{k=0}^{\infty}$  的收敛性<sup>[130]</sup>。然而只考虑渐进的收敛性并不足以作为比较不同算法效率的标准<sup>[119]</sup>。在 1950-1960 年代，在非线性优化领域，人们开始采用渐进收敛速度来评价算法的优良程度<sup>[111,130]</sup>，这种范式也广泛应用在数值线性代数<sup>[71]</sup>中，现在仍然是优化算法重要的评价标准。渐进收敛速度主要的评价指标是 Q-收敛速度和 R-收敛速度<sup>[130]</sup>。比如，梯度法最广为人知的性质是当算法序列接近于局部最优解时，可以证明梯度法的局部收敛速度是线性的，具体的速度与当前点邻域的二阶信息的一个比率有关： $(\frac{\kappa-1}{\kappa+1})^2$ ，其中  $\kappa$  为局部最优点 Hessian 矩阵的条件数。该比率历史上被称为 Kantorovich 标准速率 (Canonical Rate)<sup>[98,110,111]</sup>。其他更复杂的优化算法，甚至是用于约束问题上的投影梯度法，罚函数法，对偶乘子法等方法的局部收敛速度，均可以依据一

<sup>注1</sup>很容易构造出二次规划的解是无理数。

阶算法的标准速率理论推广得到。局部收敛理论的一大缺点是无法评价算法前期远离最优解时的表现<sup>注2</sup>。更多时候，我们需要讨论全局收敛速度 (Global Rate of Convergence)。另外，也无法评价一些具有组合性质的算法，如单纯形法，动态规划等。这时就可以利用前述复杂度理论来评价算法的运行效率。大多数时候，**我们对算法的全局效率(用复杂度来衡量)和局部效率(用局部收敛速度来衡量)都感兴趣。**

以线性规划为例，继 1947 年 Dantzig 提出单纯形法以后，人们就开始开始关注它的运行效率。然而，Klee and Minty<sup>[101]</sup> 构建出一类经典的反例(称为 Klee-Minty cubes)，证明了经典的单纯形法具有指数级的算法复杂度。然而，人们发现单纯形法的实际表现非常优越，却没有一套理论去解释它。部分解释是现实中充满了网络流类的问题，且网络单纯性法单纯形法<sup>[6]</sup> 有一些强多项式的结果<sup>[10,70]</sup>。随后学者发现在马尔科夫决策问题上，也是强多项式的<sup>[171]</sup>。这些结果不仅极大推动了单纯形法的研究，也印证了复杂度理论的合理性。1979 年，俄罗斯数学家 Khachiyan 发现了第一种求解一般线性规划的多项式时间算法<sup>[76,100]</sup>。然而，Khachiyan 的算法(称为椭球法)在实际应用时却比单纯形法慢得多，其表现非常接近于证明出的最坏情况。Traub and Woniakowski<sup>[155]</sup> 给出了一个例子，表明椭球法在算术模型中复杂度没有上界。在 20 世纪 80 年代中期，Karmarkar 的开创性论文<sup>[99]</sup> 开启了线性规划的新纪元<sup>[122]</sup>。这篇论文提出了一种新的多项式时间的算法(Karmarkar's Projective Method)，后来被简称之为内点法。它的重要性不仅是提出了一个多项式时间的线性规划算法，也不仅仅是优雅简约的复杂度分析；更重要的是，Karmarkar 内点法的理论结果完美匹配上它在各种问题中数值表现。这同时改变了非线性优化研究的风格和方向。此后，为新方法提供复杂性分析变得越来越普遍，复杂度分析甚至与计算实验中的表现同等重要。后来的研究者发现，内点法不仅是全局线性的，而且在局部还具有二次收敛速度<sup>[174]</sup>。

随着优化场景的不断扩充，优化算法的复杂度分析也形成了很多变式。在一些场景中，获取函数值、获取梯度信息的算术复杂度无法得知，人们可能只关心算法计算函数值，计算梯度，或者计算 Hessian 矩阵的次数 (Function, Gradient, Hessian Evaluations)，这种计数方法(又称 Evaluation Complexity)，在当代优化的文献中非常常见<sup>[29]</sup>。更有甚者，有时人们甚至直接考虑子模块的求解次数，而完全忽略子问题的复杂度；比如条件梯度法的分析中，一般直接认为的子问题的复杂度是可控的，直接将求解的次数当做是这种计数目标<sup>[55]</sup>；这种分析方式也适用于随机规划问题，因为这类问题一般需要反复求解线性规划子问题(或一些凸的分隔问题)<sup>[104]</sup>。除了分析最坏情况复杂度 (Worse-Case Complexity)，也有一类分析平均情况复杂度 (Average-Case) 的结果，比如内点法专著<sup>[165]</sup> 采用了一整个章节介绍平均情况复杂度的分析，这里不做赘述。近年来，随着随机优化方法的兴起，由于梯度信息和二阶信息往往需要通过采样来得到，很多学者不仅关心算法的迭代次数，也对算法的随机采样次数有所要求，这种分析角度也称为采样复杂度 (Sample Complexity)<sup>[61,114]</sup>。近

<sup>注2</sup>甚至在传统观点上，人们认为局部收敛速度快与全局收敛性是难以同时保证的。<sup>[130]</sup>

年来,相较于绝对误差,也有学者开始关心相对误差 (Relative Scale) 的收敛速度,目前还是一个蓬勃发展的领域<sup>[123]</sup>.

总的来说,优化问题是非常难的问题,对此应有基本认识<sup>[138]</sup>.对于一般线性规划,目前尚未发现算术模型下的强多项式时间算法.对于非线性优化,假设可行集是紧凑的,即使验证非凸二次的一个给定点是否是局部最小值都是 NP 难的,虽然这并不排除可以找到(另)一个可轻松验证为局部最小值的点的可能性.那么也不难想象,对于可行集无界的二次问题,检查 KKT 点的存在性是 NP 难的.具有一个负特征值的二次规划是 NP 难的<sup>[137]</sup>.

### 第三节 二阶算法及其复杂度理论

我们回到无约束优化问题:

$$\min_{x \in \mathbb{R}^n} f(x), \quad (2-1)$$

其中  $f: \mathbb{R}^n \mapsto \mathbb{R}$  是两次连续可微且  $f_{\inf} := \inf f(x) > -\infty$ . 允许给定容差  $\epsilon > 0$ , 假设我们的目标是找到  $x \in \mathbb{R}^n$ , 满足  $\epsilon$ -近似二阶稳定性条件 (Second-Order Stationary Point, 简称 SOSP):

$$\|\nabla f(x)\| \leq O(\epsilon) \quad (2-2a)$$

$$\lambda_1(\nabla^2 f(x)) \geq \Omega(-\sqrt{\epsilon}). \quad (2-2b)$$

其中  $\lambda_1(A)$  表示  $A$  的最小特征值. 由于二阶稳定点问题 (2-1) 是优化理论中的基本问题. 相应地, 一般我们将只满足 (2-2a) 的点称为一阶稳定点 (First-Order Stationary Point, FOSP).

由于论文主要讨论二阶方法, 我们不对一阶算法的复杂度理论结果作详细阐述, 详细结果可以参考<sup>[105,125]</sup>. 这里只作简要介绍. 当采用梯度法 (一阶算法) 求解以上问题时, 我们计算一系列迭代点  $\{x_k\}_{k=0}^{\infty}$ ,

$$x_{k+1} = x_k - \eta_k g_k,$$

其中  $\eta_k, g_k$  分别为步长和  $x_k$  处的梯度. 一阶方法一般认为无法得到二阶稳定点<sup>注3</sup>, 只能满足一阶稳定条件(2-2a). 在梯度法的分析中, 我们一般假设梯度是  $L$ -Lipschitz 连续的:

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \forall x, y.$$

我们知道, 当  $f$  为非凸时, 在标准的  $L$ -Lipschitz 连续梯度条件下, 可以证明梯度下降法 (GD) 的复杂度为  $O(\epsilon^{-2})$ <sup>[125]</sup>. 再将梯度法每次迭代的代价——一般为  $O(n^2)$  考虑在内, 其算术复杂度可以写作  $O(n^2 \cdot \epsilon^{-2})$ . 在凸问题中, 利用 Nesterov 加速技术, 复杂度可以提高到  $O(\epsilon^{-1/2})$ .

与之相比, 牛顿法利用梯度和 Hessian 矩阵执行以下迭代,

$$x_{k+1} = x_k - H_k^{-1} g_k,$$

注3——部分包含“鞍点跳出”机制的一阶算法需要利用曲率信息, 可以认为不是一般意义下的一阶算法。

其中  $g_k = \nabla f(x_k)$  和  $H_k = \nabla^2 f(x_k)$ . 如前面介绍, 牛顿法一般指代起源于 1669 年的 Newton-Raphson 法<sup>[129,143]</sup>, 起初是作为一种方程求根算法。后来对于一般的广义方程和包含问题 (Generalized equation, inclusion), 一般称为 Josephy-Newton 法<sup>[96]</sup>. 如果限制在优化问题中, 牛顿步也可以看作是局部二次模型的最小值,

$$m_k(d) = f(x_k) + g_k^T d + \frac{1}{2} d^T H_k d. \quad (2-3)$$

它一直是解决光滑问题最强大的方法之一, 特别是由于它能够稳定地解决病态问题, 同时非退化最优解附近具有二次收敛速度<sup>[130]</sup>, 即使用拟牛顿法估计 Hessian 矩阵, 在一些标准条件下, 也能达到超线性的收敛速度。牛顿法的在无约束问题中的二次收敛性是如此重要, 以至于这种性质可以作为基础推广到其他更复杂的算法中。如信赖域法<sup>[38]</sup>, 可以证明只要信赖域的更新方法合理<sup>[118]</sup>, 当算法接近于局部最优解后, 信赖域约束就会失效, 算法此时可以匹配牛顿法的二次收敛性。在内点法理论中, Potra and Ye<sup>[142]</sup>, Ye et al.<sup>[174]</sup> 证明了即使没有非退化假设, 一些基于牛顿步的内点法 (如 Mizuno—Todd—Ye<sup>[116]</sup>) 仍然同时具有多项式复杂度和局部二次收敛速度。

在二阶算法的分析中, 一般假设 Hessian 矩阵是  $M$ -Lipschitz 连续的, 即:

$$\|H(x) - H(y)\| \leq M\|x - y\|, \forall x, y.$$

令人惊讶的是在一般非凸的函数中, 尽管一些具有全局收敛性的牛顿类算法, 如信赖域法, 在实践中仍然表现出色, 经典的信赖法的二阶算法在理论上最坏情况复杂度为  $O(\epsilon^{-2})$ <sup>[26]</sup>, 且无法被提高, 这与梯度法有相似的运行效率。改进这个复杂度分析的结果并非易事。为了提高二阶算法的分析复杂度, 近年来有一部分学者提出了一些改进的二阶算法, 使得二阶法的分析复杂度提高到  $O(\epsilon^{-3/2})$ . 主要包括 Nesterov and Polyak<sup>[128]</sup>, Cartis et al.<sup>[27, 28]</sup>, Grapiglia et al.<sup>[74]</sup>, 及 Ye<sup>[168]</sup> 中提供的算法技术。这表明二阶算法能在理论上具有更好的全局复杂度。我们做一些进一步阐述。

2006 年, Nesterov and Polyak<sup>[128]</sup> 提出使用三次正则的牛顿法 (Cubic Regularized Newton Method), 可以将二阶算法的分析复杂度  $O(\epsilon^{-3/2})$ , 这个方法可以追溯到 1981 年 Griewank<sup>[75]</sup> 的一个技术报告中, 但当时并没有成熟的复杂度理论。随后 Cartis et al.<sup>[27][28]</sup> 在两篇论文中详细阐述了更为实际的三次正则牛顿法, 讨论如何利用非精确条件达到原文中的收敛性<sup>[128]</sup>, 计算效果非常喜人。基于三次正则的子问题, Nesterov<sup>[121]</sup> 提出了对于凸问题的加速版本, 将复杂度进一步提高到  $O(\epsilon^{-1/3})$ .

但由于三次正则子问题的计算难度较大, 实际上在非线性优化的软件中, 仍然是信赖域法、二次正则类的牛顿法应用更广泛。相比三次正则法, 这些牛顿类算法的理论分析更具有挑战性。第一个具有  $O(\epsilon^{-3/2})$  分析复杂度的信赖域法来源于叶荫宇教授的授课讲义<sup>[168]</sup>, 这个方法需要固定信赖域半径为  $\sqrt{\epsilon}$ . 直到 2017 年, Curtis et al.<sup>[41]</sup> 才提出了第一个真正意义上“动态调节”的信赖域法, 解答了这个长久的开放问题。随后, Royer and Wright<sup>[147]</sup> 提出了一个基于线搜索的  $O(\epsilon^{-3/2})$  的二阶



算法，证明了利用线搜索也能达到相似的效果。但实际上这两个方法<sup>[41,147]</sup>的流程都极为复杂，算法中包含了很多复杂的子模块，对于工程实践并不友好。Cartis et al.<sup>[29]</sup> 最近的非凸优化专著中，对于这些算法作了较为详细的总结。

#### 第四节 二阶算法在凸优化及结构化问题中的理论分析

即使在非凸问题中，利用结构特征，二阶算法完全可能超过一般  $O(\epsilon^{-3/2})$  或者  $O(\epsilon^{-2})$  的理论结果。一个较为著名的结论是非凸的信赖域子问题 (Trust-region Subproblem, TRS). 可以证明 TRS 具有隐凸性<sup>[161]</sup>. 早在上世纪 90 年代初, Vavasis and Zippel<sup>[157]</sup>, Ye<sup>[163]</sup> 分别证明了 TRS 是多项式可解的, 其复杂度可以提高到  $O(\log \log(1/\epsilon))$ . 对于一般的线性约束的非凸二次规划, Ye<sup>[166]</sup> 证明了利用二阶算法其复杂度为  $O(\epsilon^{-1})$ , 据我们所知, 对于估计 KKT 点, 这是第一个打破了平凡的  $O(\epsilon^{-2})$  复杂度的结果。在 21 世纪初, 有学者证明对一类压缩感知问题, 其复杂度高于一般认知下的  $O(\epsilon^{-2})$  (甚至有可能是  $O(\log(\epsilon^{-1}))$ ), 比较有代表性的工作包括<sup>[32,64,65]</sup>. 即使对于不满足 Lipschitz 条件的问题, 也可以证明不平凡的结果<sup>[31,54,78,82]</sup>.

二阶算法在凸优化中的表现也得到了广泛的研究。比较有代表性的如 Nesterov and Polyak<sup>[128]</sup> 提出的三次正则牛顿法 (Cubic Regularization, CR), 以及利用 Nesterov 估计序列 (Estimating Sequence) 加速技术针对凸优化的改进方法<sup>[121]</sup>. 上述两个工作可以分别将二阶方法在一般性的满足二阶 Lipschitz 连续条件的凸优化问题复杂度提高到  $O(\epsilon^{-1/2})$  和  $O(\epsilon^{-1/3})$ . 利用 Monteiro-Svaitor 框架<sup>[117]</sup>, 这个结果可以进一步提高到  $O(\epsilon^{-1/3.5})$ . 但算法将额外需要一个线搜索环节。

很多实践结果表明, 加速二阶算法的实际表现往往不尽如人意, 其全局收敛性的优势没有很强的体现<sup>[24]</sup>. 这启发人们进一步提高二阶算法, 尤其是在凸优化问题中的表现。2021 年, Mishchenko<sup>[115]</sup> 发现, 可以将正则项设为  $\|\nabla f(x)\|^{1/2}$ , 首次证明了“二次”正则化的牛顿法也具有  $O(\epsilon^{-1/2})$  的复杂度, 这个结果以往被认为是三次正则化技术的结果; 实际上, Mishchenko 的算法框架隐含了一个线性收敛的子列。随后, Doikov et al.<sup>[51]</sup> 结合该技术提出一种统一的二阶算法框架, 利用“凸函数的三阶信息并不重要”的直觉, 统一对 Hölder 连续的函数作分析, 同时囊括了一阶、二阶、三阶 Lipschitz 连续假设下的分析结果, 文章中汇报的数值结果也非常喜人。Jiang et al.<sup>[92]</sup> 提出一种信赖域法, 首次证明了信赖域法在凸优化问题中的复杂度。

除次之外, 另一类分析聚焦于更为特殊的问题。如在内点法中, 牛顿法的全局线性速度主要基于一类自协函数 (Self-Concordant Functions), 其中就包括对数障碍罚函数 (Logarithmic Barriers), (分析中心的) 势函数 (Potential function of Analytic Centers) 等<sup>[127]</sup>. 从自协函数理论的提出以来, 不断有学者尝试推广这类函数的性质, 近几年比较有名的工作是“广义自协函数”<sup>[153]</sup>. 这套理论随后被推广到了条件梯度法<sup>[55]</sup> (又称 Frank-Wolfe 法) 及变分问题中的单调算子中<sup>[154]</sup>. 在半光滑和非光滑问题中, 也有利用误差界进行分析的文献, 取得了极好的效果。如文献<sup>[175]</sup> 利用 Luo-Tseng 误差

界可以证明一类邻近牛顿法具有超线性收敛速度。随机优化问题中的二阶算法也是一个热点问题，见综述<sup>[19]</sup>。比较重要的方法是随机拟牛顿算法的发展<sup>[21]</sup>，以及最近应用到深度学习中的拟牛顿算法<sup>[12]</sup>等，这里不再赘述。

## 第五节 二阶子问题的求解方法

二阶算法中，如何高效求解子问题是一个核心课题。在二阶算法中，子问题一般是某种信赖域子问题的变种，形如，

$$d_k = \arg \min_{d \in \mathbb{R}^n} f(x_k) + g_k^T d + \frac{1}{2} d^T H_k d + \gamma_k \|d\|^p. \quad (2-4)$$

在信赖域法中， $\gamma_k$  可以认为是信赖域约束的对偶变量，此时取  $p = 2$ 。在三次正则牛顿法中，取  $p = 3$ 。在一般的二次正则牛顿法中，一般取  $p = 2$ ， $\gamma_k$  是一个自由调节的常数。一般地，形同(2-4)的子问题在理论和实际上不要求到精确解，只需要满足一定的正则条件，如 Dennis-Moré<sup>[118]</sup> 条件等<sup>[27,41]</sup>。

从计算上，求解二阶子问题的方法可以分为两种，一种需要矩阵分解，另一种依靠一类迭代算法求解。矩阵分解法主要依赖于 Cholesky 或 LDL 分解，这种方法主要依赖于强大的数值计算软件，如 BLAS<sup>[17]</sup>，LAPACK<sup>[7]</sup>。另一类方法主要依赖于方程组的迭代方法，尤其是 Krylov 子空间方法<sup>[71]</sup>，参见文献<sup>[27,29]</sup>。最简单的 Krylov 子空间方法称为共轭梯度法 (Conjugate Gradient Method)，可以证明，共轭梯度法在严格正定系统中是有限步收敛的。近年来，有学者发现 Krylov 类的方法，尤其是 Lanczos Method<sup>[71]</sup> 在某些条件下，求解 TRS (或广义 TRS) 具有线性收敛速度<sup>[73,178,179]</sup>。值得注意的是，Rojas et al.<sup>[146]</sup> 意识到 TRS 可以使用一个对称特征值问题进行求解，这个想法随后被<sup>[3,4]</sup> 推广。近年来，也有用二次特征值问题，广义特征值问题求解三次子问题的做法，如文献<sup>[86]</sup>。也有学者利用一阶算法求解子问题，并证明算法相对与输入数据的非零元是“线性时间”的，如<sup>[80,91]</sup>。

## 第六节 Lanczos 方法以及随机求解技术

这里回顾和总结 Lanczos 方法在求解对称实矩阵  $A \in \mathcal{S}^n$  特征对 (eigenpair) 方面的标准结果。不妨假定特征值按以下顺序排列：

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n. \quad (2-5)$$

不失一般性，我们假定  $\lambda_1 > \lambda_2$ ；否则，若存在重数，可直接选取下一个特征值。

### 2.6.1 Lanczos 方法的误差界及基本概念

特征值估计的误差可以有多种理解，我们简述之。

- **标准相对误差与谱间距。**标准理论<sup>[149]</sup> 对实对称矩阵第  $i$  个特征值估计了如下的算术运算，使得

近似特征值  $\xi$  满足 **相对误差与谱间距**:

$$0 \leq \lambda_1 - \xi \leq (\lambda_1 - \lambda_n) \varepsilon, \quad (2-6)$$

其中  $\varepsilon > 0$  为 **相对误差**,  $\xi$  (即 Ritz 值) 来源于当次 Lanczos 法的迭代  $k$ 。

- **Kuczyński 的相对误差**。另一方面, Kuczyski and Woniakowski<sup>[103]</sup> 采用不同的误差定义:

$$\left| \frac{\xi(A) - \lambda_1(A)}{\lambda_1(A)} \right| \leq \varepsilon. \quad (2-7)$$

(2-7) 和 (2-6) 可通过以下不等式联系起来:

$$\frac{\lambda_1(A) - \xi(A)}{\lambda_1(A) - \lambda_n(A)} = \frac{\lambda_1(A) - \xi(A)}{\lambda_1(A - \lambda_n(A) \cdot I)} \quad (2-8)$$

$$= \frac{\lambda_1(A - \lambda_n(A)I) - \xi(A - \lambda_n(A)I)}{\lambda_1(A - \lambda_n(A)I)}. \quad (2-9)$$

这表明 (2-6) 等价于 Kuczyński 估计中的  $\lambda_1(A - \lambda_n(A)I)$ 。由于 Lanczos 方法具有尺度不变性, 所得结果与 (2-6) 完全一致。

- **最小特征值**。让我们对最小特征值进行估计。该问题等价于计算  $-A$  的最大特征值, 即  $\lambda_n(A) = -\lambda_1(-A)$ 。设  $-\xi$  近似  $\lambda_1(-A)$ , 根据 (2-6), 误差满足:

$$\lambda_1(-A) - (-\xi) \leq (\lambda_1(-A) - \lambda_n(-A)) \varepsilon. \quad (2-10)$$

注意, 原始理论假设  $A \geq 0$ , 而  $-A$  显然不满足该条件。我们通过某个  $M \geq \lambda_1(A)$  进行缩放, 使上述不等式等价于:

$$\lambda_1(M \cdot I - A) - (M \cdot I - \xi) \leq (\lambda_1(M \cdot I - A) - \lambda_n(M \cdot I - A)) \varepsilon. \quad (2-11)$$

这进一步等价于:

$$\begin{aligned} \xi - \lambda_n(A) &\leq (\lambda_1(-A) - \lambda_n(-A)) \varepsilon \\ &\leq (\lambda_1(A) - \lambda_n(A)) \varepsilon. \end{aligned} \quad (2-12)$$

由于  $A$  和  $-A$  (或  $M \cdot I - A$ ) 的谱间距相同, 上述结果成立。

- **用诱导范数  $\|A\|$  进行恒量**。当  $A$  由 Hessian 矩阵给出时, 可能是非正定的, 此时 Lipschitz 连续性成立。通常, 我们可以考虑  $\|A\|$ , 即  $\ell_2$  诱导范数。由于所有上述结果均具有尺度不变性, 相同的界仍然成立; 因此, 可考虑  $A - \lambda_1(A)I$  并使用 (2-12)。另一方面, 我们可以用  $\|A\|$  代替  $\lambda_1(A) - \lambda_n(A)$ :

$$\|A\| = \sup_{\|x\|=1} \|Ax\| = \max\{|\lambda_1(A)|, |\lambda_n(A)|\}. \quad (2-13)$$

从而:

$$\lambda_1(A) - \lambda_n(A) \leq 2 \max\{|\lambda_1(A)|, |\lambda_n(A)|\} \leq 2\|A\|. \quad (2-14)$$

因此，误差估计满足：

$$\frac{\xi - \lambda_n(A)}{2\|A\|} \leq \frac{\xi - \lambda_n(A)}{\lambda_1(A) - \lambda_n(A)} \leq \varepsilon. \quad (2-15)$$

由此可得：

$$\xi - \lambda_n(A) \leq 2\varepsilon\|A\|. \quad (2-16)$$

- **绝对误差。**以上所有结果均基于相对误差。对于绝对误差，例如：

$$|\lambda_1 - \xi| \leq \epsilon, \quad (2-17)$$

等价于：

$$\varepsilon = \frac{\epsilon}{\lambda_1(A) - \lambda_n(A)}. \quad (2-18)$$

### 2.6.2 Kaniel–Paige 收敛理论

经典复杂度依赖于与特征值分布相关的界，即特征值谱。以下定理可在标准特征值教材中找到，如文献<sup>[149]</sup>的 Theorem 6.4.

#### 定理 2.1

设  $\lambda_i$  和  $\lambda_i^{(k)}$  分别为第  $i$  个精确和近似特征值，则它们满足以下双重不等式：

$$0 \leq \lambda_i - \lambda_i^{(k)} \leq (\lambda_1 - \lambda_n) \left( \frac{\kappa_i^{(k)} \tan \angle(b, v_i)}{C_{m-i}(1 + 2\gamma_i)} \right)^2$$

其中  $b$  为初始向量，

$$v_i \in \mathcal{S}_i(A), \gamma_i = \frac{\lambda_i - \lambda_{i+1}}{\lambda_{i+1} - \lambda_n},$$

且  $\kappa_i^{(k)}$  由以下递推关系给出：

$$\kappa_1^{(k)} \equiv 1, \quad \kappa_i^{(k)} = \prod_{j=1}^{i-1} \frac{\lambda_j^{(k)} - \lambda_n}{\lambda_j^{(k)} - \lambda_i}, \quad i > 1.$$

其中， $C(\cdot)$  为第一类 Chebyshev 多项式。

#### 定义 2.2: 第一类 Chebyshev 多项式

第  $k$  阶第一类 Chebyshev 多项式定义为：

$$C_k(t) = \cos[k \cos^{-1}(t)], \quad \forall t, -1 \leq t \leq 1.$$

由以上定义，从而递推关系为：

$$C_{k+1}(t) = 2t \cdot C_k(t) - C_{k-1}(t).$$

当  $|t| > 1$  时, 我们使用以下扩展公式:

$$C_k(t) = \cosh \left[ k \cosh^{-1}(t) \right], \quad |t| \geq 1.$$

这一公式可通过复变量方法推导, 利用  $\cos \theta = (e^{i\theta} + e^{-i\theta})/2$ , 从而得到:

$$C_k(t) = \frac{1}{2} \left[ \left( t + \sqrt{t^2 - 1} \right)^k + \left( t - \sqrt{t^2 - 1} \right)^k \right].$$

因此可得下界估计:

$$C_k(t) \geq \frac{1}{2} \left( t + \sqrt{t^2 - 1} \right)^k := \underline{C}_k(t), \quad \forall |t| \geq 1. \quad (2-19)$$

基于上述结论, 对极端情况 ( $i \in \{1, n\}$ ), 我们有以下推论:

### 推论 2.3

设  $i = 1$ , 则最大特征值的近似满足:

$$0 \leq \lambda_1 - \lambda_1^{(k)} \leq (\lambda_1 - \lambda_n) \left( \frac{\tan \angle(b, v_1)}{C_{k-1}(1 + 2\gamma_1)} \right)^2, \quad \gamma_1 = \frac{\lambda_1 - \lambda_2}{\lambda_2 - \lambda_n}.$$

注意, 对于所有  $i \geq 1$ , 均有  $1 + 2\gamma_i \geq 1$ 。我们引入以下定理:

### 定理 2.4: 求解最大特征值的复杂度

设  $i = 1$ , 则使

$$0 \leq \lambda_1 - \lambda_1^{(k)} \leq (\lambda_1 - \lambda_n) \varepsilon$$

成立所需的迭代次数满足:

$$K_\varepsilon = 1 + \left\lceil \sqrt{\frac{\lambda_2 - \lambda_n}{\lambda_1 - \lambda_2}} \log \left( \frac{2 \sin \angle(b, v_1)}{\sqrt{\varepsilon} \cdot \langle b, v_1 \rangle} \right) \right\rceil.$$

进一步地, 使

$$0 \leq \lambda_1 - \lambda_1^{(k)} \leq \epsilon$$

成立的迭代次数, 即  $\varepsilon = (\lambda_1 - \lambda_n)^{-1}\epsilon$ , 满足:

$$K_\epsilon^{\text{abs}} = 1 + \left\lceil \sqrt{\frac{\lambda_2 - \lambda_n}{\lambda_1 - \lambda_2}} \log \left( \frac{2 \sin \angle(b, v_1) \cdot \sqrt{\lambda_1 - \lambda_n}}{\sqrt{\epsilon} \cdot \langle b, v_1 \rangle} \right) \right\rceil.$$

*Proof.* 需找到满足下列条件的  $k$ :

$$\left( \frac{\tan \angle(b, v_1)}{C_{k-1}(1 + 2\gamma_1)} \right)^2 \leq \varepsilon.$$

观察若满足:

$$\varepsilon^{-1/2} |\tan \angle(b, v_1)| \leq \frac{1}{2} \left( t + \sqrt{t^2 - 1} \right)^k, \quad (2-20)$$

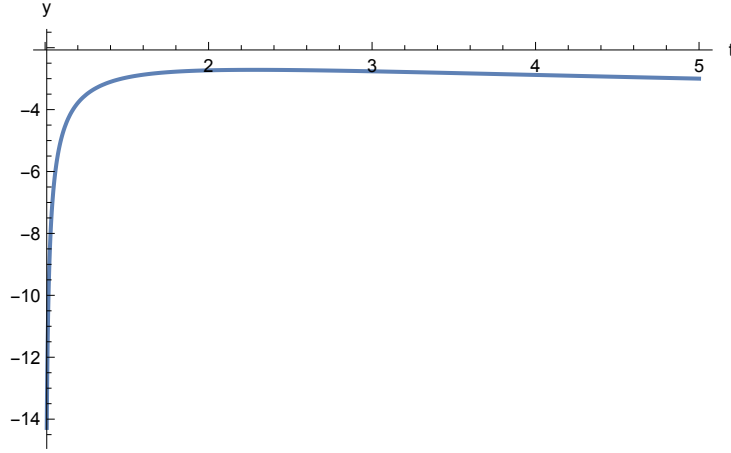


图 2-1 验证  $f(t) = \frac{1}{\log(t + \sqrt{t^2 - 1})} - \sqrt{\frac{2}{t-1}} \leq 0$ .

则:

$$\begin{aligned} \varepsilon^{-1/2} |\tan \angle(b, v_1)| &\leq \frac{1}{2} \left(t + \sqrt{t^2 - 1}\right)^k \\ &\leq \underline{C}_{k-1} (1 + 2\gamma_1) \leq C_{k-1} (1 + 2\gamma_1), \end{aligned} \quad (2-21)$$

进一步推出:

$$\frac{\tan \angle(b, v_1)}{C_{k-1} (1 + 2\gamma_1)} \leq \sqrt{\varepsilon}. \quad (2-22)$$

注意到,

$$\left(t + \sqrt{t^2 - 1}\right)^{k-1} \geq 2 \cdot |\tan \angle(b, v_1)| / \sqrt{\varepsilon}, \quad \text{其中 } t = 1 + 2\gamma_1. \quad (2-23)$$

$$\Rightarrow k \geq 1 + \log \left( \frac{2 |\tan \angle(b, v_1)|}{\sqrt{\varepsilon}} \right) / \log \left( t + \sqrt{t^2 - 1} \right). \quad (2-24)$$

数值计算 (见上图) 可验证 图 2-1 (我们省略证明):

$$\log \left( t + \sqrt{t^2 - 1} \right)^{-1} \leq \sqrt{2/(t-1)} = \sqrt{\gamma_1^{-1}} = \sqrt{\frac{\lambda_2 - \lambda_n}{\lambda_1 - \lambda_2}},$$

因此:

$$k \leq 1 + \sqrt{\frac{\lambda_2 - \lambda_n}{\lambda_1 - \lambda_2}} \log \left( \frac{2 \tan \angle(b, v_1)}{\sqrt{\varepsilon}} \right).$$

由于:

$$\tan \angle(b, v_1) = \frac{\sin \angle(b, v_1)}{\cos \angle(b, v_1)} = \frac{\sin \angle(b, v_1)}{\langle b, v_1 \rangle},$$

证毕。 □

### 2.6.3 Kuczyński 估计

由于 Lanczos 方法通过 Krylov 子空间演化, 众所周知, 该方法的收敛性取决于特征值分布及角度  $\angle(b, v_1)$ ; 若二者正交, 则方法失效。Kuczyski and Woniakowski<sup>[103]</sup> 证明了以下结论: (a) 对于

某个固定的  $b \in \mathbb{R}^n$ ，不存在任何 Krylov 类型算法能够对所有  $A \in \mathcal{S}_n$  近似  $v_1 \in \mathcal{S}_1$ ；(b) 若使用  $k$  维子空间（即  $k$  次迭代），则在少于  $n - 1$  次迭代内无法近似  $v_1$ 。其核心思想是利用概率方法绕过  $b \perp \mathcal{S}_1$  的情况，即  $b \not\perp \mathcal{S}_1$  成立 a.s.

我们使用原始表述，考虑

$$\left| \frac{\xi(A) - \lambda_1(A)}{\lambda_1(A)} \right| \leq \varepsilon.$$

根据 Lanczos 方法的特性，上述式子等价于

$$\frac{\lambda_1(A) - \xi(A)}{\lambda_1(A)} \leq \varepsilon.$$

令  $\lambda_1^{(k)}$  表示第  $k$  次迭代得到的 Ritz 近似值。对于某个分布  $\mu$ ，定义两种误差度量：

$$e_{\text{avg}}^{(k)} = \int_{\|b\|=1} \left| \frac{\lambda_1^{(k)} - \lambda_1(A)}{\lambda_1(A)} \right| \mu(db),$$

$$\mathbb{P}(\lambda_1^{(k)}, A, k, \varepsilon) = \mu \left\{ b \in \mathbb{R}^n : \|b\| = 1, \left| \frac{\xi(A, b, k) - \lambda_1(A)}{\lambda_1(A)} \right| > \varepsilon \right\}.$$

### 2.6.3.1 期望误差

我们首先分析期望概念下的误差。

#### 定理 2.5

设 Lanczos 方法在  $A$  上运行，初始向量  $b \sim \mu$ .

(a) 对于任意对称正定矩阵  $A$ ，令  $d$  表示其不同特征值的个数。则当  $k \geq d$  时，

$$e_{\text{avg}}^{(k)} = 0.$$

当  $k \in [4, m - 1]$  时，

$$e_{\text{avg}}^{(k)} \leq 0.103 \left( \frac{\ln(n(k-1)^4)}{k-1} \right)^2 \leq 2.575 \left( \frac{\ln n}{k-1} \right)^2.$$

(b) 对于任意对称正定矩阵  $A$ ，设  $\lambda_2$  和  $\lambda_n$  分别为  $A$  的第二大和最小特征值，则

$$e_{\text{avg}}^{(k)} \leq 2.589 \sqrt{n} \left( \frac{1 - \sqrt{(\lambda_1 - \lambda_2) / (\lambda_1 - \lambda_n)}}{1 + \sqrt{(\lambda_1 - \lambda_2) / (\lambda_1 - \lambda_n)}} \right)^{k-1}.$$

我们将其转化为收敛率：

**推论 2.6**

设  $K$  使得  $e_{\text{avg}}^{(K)} \leq \varepsilon$ , 则 (a) 给出

$$K \geq 1 + \left\lceil 1.605 \cdot \varepsilon^{-1/2} \ln n \right\rceil.$$

而 (b) 令

$$\kappa_L = \frac{\lambda_1 - \lambda_n}{\lambda_1 - \lambda_2},$$

则

$$K = 1 + \left\lceil \frac{1}{2} \sqrt{\kappa_L} \log \left( \frac{2.589 \sqrt{n}}{\varepsilon} \right) \right\rceil.$$

*Proof.* 证明 (b) 部分。令  $\nu = \sqrt{\kappa_L^{-1}}$ , 则

$$\begin{aligned} 2.589 \sqrt{n} \left( \frac{1 - \nu}{1 + \nu} \right)^{k-1} &\leq \varepsilon, \\ \Rightarrow k &\geq 1 + \log \left( \frac{2.589 \sqrt{n}}{\varepsilon} \right) / \log \left( \frac{1 + \nu}{1 - \nu} \right) \\ &= 1 + \log \left( \frac{2.589 \sqrt{n}}{\varepsilon} \right) / \log \left( 1 + \frac{2\nu}{1 - \nu} \right). \end{aligned}$$

由于

$$\begin{aligned} \log(1 + 1/t) &\geq \frac{1}{1/2 + t}, \\ \Rightarrow \log \left( 1 + \frac{2\nu}{1 - \nu} \right) &\geq \frac{1}{1/2 + \frac{1-\nu}{2\nu}} = 2\nu, \end{aligned}$$

因此可得:

$$K = 1 + \left\lceil \frac{\sqrt{\kappa_L}}{2} \log \left( \frac{2.589 \sqrt{n}}{\varepsilon} \right) \right\rceil.$$

□

### 2.6.3.2 高概率复杂度

我们现在考虑概率复杂度, 证明误差估计以高概率  $1 - \delta$  成立, 即满足:

$$\mathbb{P}(\lambda_1^{(k)}, A, k, \varepsilon) \leq \delta.$$

**定理 2.7**

设  $\varepsilon \in [0, 1)$ ,

(a) 对于任意对称正定矩阵  $A$ , 令  $m$  为  $A$  的不同特征值个数。则当  $k \geq m$  时,

$$\mathbb{P}(\lambda_1^{(k)}, A, k, \varepsilon) = 0.$$



对于任意  $k$ , 有

$$\mathbb{P}(\lambda_1^{(k)}, A, k, \varepsilon) \leq 1.648\sqrt{n}e^{-\sqrt{\varepsilon}(2k-1)}.$$

(b) 设  $\lambda_2$  和  $\lambda_n$  分别为  $A$  的第二大和最小特征值, 则

$$\mathbb{P}(\lambda_1^{(k)}, A, k, \varepsilon) \leq 1.648\sqrt{\frac{n}{\varepsilon}} \left( \frac{1 - \sqrt{(\lambda_1 - \lambda_2)/(\lambda_1 - \lambda_n)}}{1 + \sqrt{(\lambda_1 - \lambda_2)/(\lambda_1 - \lambda_n)}} \right)^{k-1}.$$

类似于期望误差, 我们可以推导收敛速率:

### 推论 2.8

设  $A \geq 0$ , 我们希望找到某个  $K$ , 使得误差估计

$$\frac{\lambda_1(A) - \xi(A)}{\lambda_1(A)} \leq \varepsilon$$

在概率  $1 - \delta$  下成立, 其中  $\delta \in (0, 1)$ 。则 (a) 给出:

$$\begin{aligned} K &= 1 + \left\lceil \frac{1}{2}\varepsilon^{-1/2} \log \left( \frac{1.648 \cdot n}{\delta^2} \right) \right\rceil \\ &\approx 1 + \left\lceil \frac{1}{4}\varepsilon^{-1/2} \log \left( \frac{n}{\delta^2} \right) \right\rceil. \end{aligned}$$

对于 (b), 令

$$\kappa_L = \frac{\lambda_1 - \lambda_n}{\lambda_1 - \lambda_2},$$

则

$$\begin{aligned} K &= 1 + \left\lceil \frac{\sqrt{\kappa_L}}{2} \log \left( \frac{1.648\sqrt{n}}{\delta\sqrt{\varepsilon}} \right) \right\rceil \\ &\approx 1 + \left\lceil \frac{\sqrt{\kappa_L}}{2} \log \left( \frac{2.716 \cdot n}{\delta^2 \cdot \varepsilon} \right) \right\rceil \\ &\approx 1 + \left\lceil \sqrt{\kappa_L} \log \left( \frac{n}{\delta^2 \cdot \varepsilon} \right) \right\rceil. \end{aligned}$$

为完整起见, 我们提供 (b) 部分的证明:

*Proof.* 令  $\nu = \sqrt{\kappa_L^{-1}}$ , 则

$$1.648\sqrt{\frac{n}{\varepsilon}} \left( \frac{1 - \nu}{1 + \nu} \right)^{k-1} \leq \delta.$$

于是,

$$k \geq 1 + \log \left( \frac{1.648\sqrt{n}}{\delta\sqrt{\varepsilon}} \right) / \log \left( \frac{1 + \nu}{1 - \nu} \right).$$

因此可得：

$$K = 1 + \left\lceil \frac{\sqrt{\kappa_L}}{2} \log \left( \frac{1.648n}{\delta\varepsilon} \right) \right\rceil.$$

□

以上结论将在非精确的齐次二阶法中被反复用到，为简洁起见，我们后续将不再重复这些结论。

### 第三章 齐次二阶下降法

#### 第一节 介绍

我们考虑以下无约束优化问题：

$$\min_{x \in \mathbb{R}^n} f(x), \quad (3-1)$$

其中,  $f: \mathbb{R}^n \mapsto \mathbb{R}$  为二阶连续可微函数, 且  $f_{\inf} := \inf f(x) > -\infty$ 。给定容差  $\epsilon > 0$ , 目标是找到满足以下条件的  $\epsilon$ -近似二阶稳定点 (SOSP)  $x$ 。其中,  $\lambda_1(A)$  表示矩阵  $A$  的最小特征值。对于非凸函数  $f$ , 已有研究表明, 在标准  $L$ -Lipschitz 连续梯度条件下, 梯度下降 (GD) 方法可在  $O(\epsilon^{-2})$  次迭代内找到满足条件 (2-2a) 的  $\epsilon$ -近似一阶稳定点。若进一步要求满足二阶条件 (2-2b), 一阶方法可能失效, 通常我们切换到二阶方法, 例如信赖域法<sup>[38]</sup>。

在每次迭代中, 牛顿类方法通常在当前迭代点  $x_k$  构造二阶近似模型, 然后计算更新方向  $d_k$ 。例如, 牛顿法采用以下二次近似：

$$d_k = \arg \min_{d \in \mathbb{R}^n} m_k(d) := g_k^T d + \frac{1}{2} d^T H_k d, \quad (3-2)$$

其中,  $g_k = \nabla f(x_k)$ ,  $H_k = \nabla^2 f(x_k)$ 。尽管牛顿法在实践中表现优异, 但 Cartis et al.<sup>[25]</sup> 显示, 牛顿法在非凸情况下的最坏复杂度与梯度下降法相同, 为  $O(\epsilon^{-2})$ 。因此, 需采用高级技术以提高牛顿法的收敛性能。Nesterov and Polyak<sup>[128]</sup> 引入了三次正则化 (CR), 考虑以下子问题：

$$d_k^{\text{CR}} = \arg \min_d m_k^{\text{CR}}(d) := g_k^T d + \frac{1}{2} d^T H_k d + \frac{\sigma_k}{3} \|d\|^3, \quad (3-3)$$

其中  $\sigma_k > 0$ 。他们证明, 三次正则化牛顿法的迭代复杂度提升为  $O(\epsilon^{-3/2})$ 。Cartis et al.<sup>[27, 28]</sup> 随后提出了自适应和不精确的三次正则化 (ARC) 版本, 具有相同的复杂度。在三次正则化出现之前, 信赖域 (TR) 方法是一种广泛使用的经典算法。它基于与牛顿法相同的模型函数计算更新方向, 但将其限制在预设的信赖域半径  $\Delta_k$  内, 并在相应的接受率  $\rho_k$  超过某一阈值时接受更新<sup>[38]</sup>：

$$d_k^{\text{TR}} = \arg \min_{\|d\| \leq \Delta_k} m_k(d), \quad (3-4a)$$

$$\rho_k := \frac{f(x_k + d_k) - f(x_k)}{m_k(d_k) - m_k(0)}. \quad (3-4b)$$

然而, 通过这种方式建立改进的  $O(\epsilon^{-3/2})$  迭代复杂度更加困难。据我们所知, Ye<sup>[168, 169]</sup> 提出了最早的基于固定半径策略的  $O(\epsilon^{-3/2})$  信赖域方法。最近, Curtis et al.<sup>[41]</sup> 指出, 基于 (3-4) 的经典信赖域方法因使用经典的  $\rho_k$  接受规则和线性更新半径而未能满足实现  $O(\epsilon^{-3/2})$  复杂度所需的充分下降性质。

为了解决这一问题, 他们开发了一种名为 TRACE 的算法<sup>[40, 41]</sup>, 该算法通过调整  $\Delta_k$  的扩张和收缩规则实现了预期的复杂度结果, 这种规则基于  $\|d_k^{\text{TR}}\|$  式 (3-4a) 的对偶解之间的非线性关系。

这一复杂度界也可以通过 Royer and Wright<sup>[147]</sup> 提出的线搜索 Newton CG 框架实现。其算法根据 Hessian 矩阵  $H_k$  的最小特征值在牛顿步和正则化牛顿步之间交替，并根据<sup>[27,41]</sup> 中类似的接受规则选择步长。然而，上述所有方法都需要求解牛顿系统，其计算成本通常为  $O(n^3)$ 。通过使用负曲率模块 (Negative Curvature Oracle) 和共轭梯度方法，可以找到具有更好复杂度的不精确解。在这种意义上，许多经典算法都有改进的空间<sup>[44,147,148]</sup>。

### 3.1.1 本章贡献

受二次规划的半正定松弛技术的启发，我们提出了局部二次近似  $m_k(d)$  的齐次化版本。我们证明了所得问题本质上是一个特征值问题，可以通过随机启动的 Lanczos 算法<sup>[103]</sup> 来求解，在高概率下具有复杂度  $\tilde{O}(n(n+1)\epsilon^{-1/4})$  的保证。我们证明了齐次化矩阵的最左特征值始终为负；即使原始 Hessian 矩阵 (接近) 正半定，**齐次化负曲率** 仍然存在。类似于梯度下降法通过沿负梯度方向移动以达到一阶稳定点，我们可以通过沿齐次化负曲率对应的方向移动来达到二阶稳定点。

其次，我们提出了一种新的二阶方法，称为齐次二阶下降方法 (Homogeneous Second-Order Descent Method, HSODM) (算法 3-1)，其子问题采用齐次二次模型。我们提供了两种步长策略来利用齐次化负曲率，包括固定半径策略和简单的回溯线搜索方法。我们的方法实现了更优的迭代复杂度  $O(\epsilon^{-3/2})$ ，相比于标准信赖域方法<sup>[42]</sup> 和基于负曲率的方法<sup>[39]</sup> 的  $O(\epsilon^{-2})$  复杂度，更快地收敛到二阶稳定点 (SOSP)。考虑到子问题，它需要  $\tilde{O}((n+1)^2\epsilon^{-7/4})$  次算术操作。与<sup>[5,23,94,147]</sup> 的方法相比，HSODM 仅依赖于齐次化模型，不需要在不同的子程序之间切换。该算法形式简单而优雅，对实践者非常友好。为了清晰地比较，我们提供了以下表 3-1，其中包含了具有最先进复杂度结果的算法。请注意，ARC<sup>[27,28]</sup> 和 TRACE<sup>[40,41]</sup> 分别需要牛顿型方程，来自三次正则化问题和信赖域子问题。这两者都可以通过应用矩阵分解 ( $O(n^3)$ ) 和适当的参数搜索过程 ( $O(n^2 \log(1/\epsilon))$ ) 来解决。为了包含不精确的子问题解，文献<sup>[44,147,148]</sup> 中的方法在用于线性方程的共轭梯度法和用于极端特征值问题的随机 Lanczos 方法之间切换，从而将复杂度率提高到  $\tilde{O}(n^2\epsilon^{-1/4})$ 。对于 HSODM，仅需要极端特征值问题。**据我们所知，HSODM 是第一个仅需要依赖特征值计算的二阶方法。**

最后，所提出方法的数值结果也令人鼓舞。具体来说，HSODM 的两个变体在 CUTEst 数据集中优于标准的二阶方法，包括经典的信赖域方法和三次正则化牛顿方法。

本章中，我们使用如下的记号。

令  $\|\cdot\|$  为  $\mathbb{R}^n$  空间中的标准欧几里得范数。记  $B(x, R) = \{y : \|y - x\| \leq R\}$  为以  $x$  为中心、半径为  $R$  的闭球。对于矩阵  $A \in \mathbb{R}^{n \times n}$ ， $\|A\|$  表示其诱导的  $\ell_2$  范数， $\lambda_1(A), \lambda_2(A), \dots, \lambda_{\max}(A)$  表示其按升序排列的不同特征值。对于  $n > 0$ ， $I_n$  表示  $n$  维单位矩阵；如果上下文清楚，我们将省略  $n$ 。在某次迭代  $x_k$  时，我们简记  $g_k = \nabla f(x_k)$  和  $H_k = \nabla^2 f(x_k)$ 。我们按照通常的意义使用渐近符

表 3-1 几种二阶算法的简要比较。这里， $p \in (0, 1)$  表示随机 Lanczos 方法的失败概率。在最后一列中，我们使用“E”表示极端特征值问题，使用“N”表示牛顿型方程。

算法	迭代复杂度	子问题复杂度	Oracle(s)
ARC <sup>[27]</sup>	$O(\epsilon^{-3/2})$	$O(n^3 + n^2 \log(1/\epsilon))$	N
TRACE <sup>[40,41]</sup>	$O(\epsilon^{-3/2})$	$O(n^3 + n^2 \log(1/\epsilon))$	N
<sup>[44]</sup> Algorithm 4.1	$O(\epsilon^{-3/2})$	$O(n^2 \epsilon^{-1/4} \log(n/p\epsilon))$	N & E
Newton-CG <sup>[147,148]</sup>	$O(\epsilon^{-3/2})$	$O(n^2 \epsilon^{-1/4} \log(n/p\epsilon))$	N & E
HSODM	$O(\epsilon^{-3/2})$	$O((n+1)^2 \epsilon^{-1/4} \log(n(n+1)/p\epsilon))$	E

号  $O$ ,  $\Omega$ ,  $\Theta$ , 而  $\tilde{O}$  隐藏了相对于  $\epsilon$  的对数项。特别地，给定两个常数  $A$  和  $B$ ，如果存在常数  $c > 0$  使得  $A \leq c \cdot B$ ，我们说  $A = O(B)$ ；如果存在常数  $c > 0$  使得  $A \geq c \cdot B$ ，我们说  $A = \Omega(B)$ 。如果  $A = O(B)$  且  $A = \Omega(B)$ ，我们说  $A = \Theta(B)$ 。我们使用  $[a; b]$  (分别为垂直拼接) 和  $[a, b]$  (分别为水平拼接) 来表示数组或数字的拼接。对于向量  $a \in \mathbb{R}^n$  和  $0 \leq j \leq n$ ，我们记  $a_{[1:j]}$  为  $a$  的前  $j$  个元素。

本章的其余部分组织如下。在 第 3.2 节 中，我们简要描述了基于齐次二次模型的方法。通过将齐次模型作为特征值问题求解，相应的 HSODM 在 算法 3-1 中被引入。在 第 3.3 节 和 第 3.4 节 中，我们分析了 HSODM 的全局和局部收敛性。我们的结果表明，HSODM 对于  $O(\epsilon^{-3/2})$  的  $\epsilon$ -近似二阶稳定点具有全局复杂度。如果不提前终止算法，它以局部二次速率收敛。我们在 第 3.5 节 中解决了 HSODM 的不精确性，介绍了一种具有偏置初始化的 Lanczos 方法，以利用 Ritz 近似于齐次曲率。在 第 3.6 节 中，我们通过在 CUTEst 基准测试中与其他标准二阶方法进行比较，展示了我们方法的有效性，并提供了丰富的计算结果。

## 第二节 齐次二次模型与二阶下降方法

### 3.2.1 齐次化的动机

许多非凸优化方法利用 Hessian 矩阵的负曲率。特别地，给定一个迭代点  $x_k$ ，对于某个容差  $\epsilon > 0$ ，通常需要确定是否存在  $\xi_k \in \mathbb{R}^n$  满足

$$\mathcal{R}_k(\xi_k) := \frac{\xi_k^T H_k \xi_k}{\|\xi_k\|^2} \leq -\sqrt{\epsilon}, \quad (3-5)$$

这暗示了  $\lambda_{\min}(H_k) \leq -\sqrt{\epsilon}$ 。在计算上，可以使用随机 Lanczos 方法，以  $\tilde{O}(n^2 \cdot \epsilon^{-1/4})$  的算术操作成本找到这样的方向  $\xi_k$ <sup>[103]</sup>。如果将该方向与合适的步长  $\eta$  配合使用，则在二阶 Lipschitz 连续性下，函数值将减少  $\Omega(\epsilon^{3/2})$ 。这种性质在基于负曲率的一阶方法中被广泛应用<sup>[23,94]</sup>。然而，如果条件 (3-5) 不成立，则必须切换到其他子程序<sup>[5,23,94,147]</sup>，这使得迭代过程变得复杂，从而难以高效实现和参数调优。

为解决这一问题，我们对  $x_k$  处的二阶泰勒展开式 (3-2) 应用了齐次化技巧 (例如，参见 [152,173])：

$$t^2 \left( m_k(d) - \frac{1}{2}\delta \right) = t^2 \left( g_k^T(v/t) + \frac{1}{2}(v/t)^T H_k(v/t) - \frac{1}{2}\delta \right) \quad (d := v/t) \quad (3-6)$$

$$= \frac{1}{2} \begin{bmatrix} v \\ t \end{bmatrix}^T \begin{bmatrix} H_k & g_k \\ g_k^T & -\delta \end{bmatrix} \begin{bmatrix} v \\ t \end{bmatrix}, \quad F_k := \begin{bmatrix} H_k & g_k \\ g_k^T & -\delta \end{bmatrix}. \quad (3-7)$$

第二个方程被称为**齐次二次模型**。**齐次矩阵**  $F_k$  的一个优良性质是：即使  $H_k$  是正定的， $F_k$  仍然是不定的，因此可以从这个  $(n+1)$  维提升矩阵中计算出“齐次负曲率”。为了与 (3-5) 中给出的 Rayleigh 商建立联系，我们施加一个球约束  $\|[v; t]\| \leq 1$ ，从而 (3-7) 是有界的。此外，如果取  $d = v/t$ ，则齐次模型与缩放了  $t^2$  的二阶近似 (3-2) 在某个常数 (即  $-\delta/2$ ) 范围内是等价的。

### 3.2.2 方法概述

我们在 算法 3-1 中给出了 HSODM 算法。本文其余部分将讨论在迭代中使用“齐次化”矩阵的方法。我们正式定义齐次二次模型如下：给定一个迭代点  $x_k \in \mathbb{R}^n$ ，令  $\psi_k(v, t; \delta)$  为齐次二次模型，

$$\psi_k(v, t; \delta) := \begin{bmatrix} v \\ t \end{bmatrix}^T \begin{bmatrix} H_k & g_k \\ g_k^T & -\delta \end{bmatrix} \begin{bmatrix} v \\ t \end{bmatrix}, \quad v \in \mathbb{R}^n, t \in \mathbb{R}, \quad (3-8)$$

其中  $\delta \geq 0$  是一个预定义常数。在每次迭代中，HSODM 在当前迭代点  $x_k$  处最小化模型，即，

$$\min_{\|[v; t]\| \leq 1} \psi_k(v, t; \delta). \quad (3-9)$$

将问题 (3-9) 的最优解记为  $[v_k; t_k]$ 。由于子问题 (3-9) 本质上是一个特征值问题， $[v_k; t_k]$  是对应于  $F_k$  最小特征值的特征向量。因此，我们可以使用特征向量求解程序来解决该子问题，参见 [103]。

在求解 (3-9) 后，我们基于最优解  $[v_k; t_k]$  构造一个下降方向  $d_k$ ，并仔细选择步长  $\eta_k$  以确保足够的下降。根据 (3-6)， $d_k = v_k/t_k$  是一个自然选择。然而，极端情况下  $t_k = 0$  可能导致  $d_k$  趋向于无穷大。从直观上看，如果  $|t_k|$  足够小，这意味着 Hessian 矩阵  $H_k$  在齐次模型中占主导作用，因此我们直接选择截断方向  $v_k$  (见 行 8)。否则，预定义参数  $-\delta$  将变得重要，我们选择  $v_k/t_k$  作为下降方向 (见 行 10)。我们使用  $\sqrt{1/(1+\Delta^2)}$  和  $\nu$  作为  $|t_k|$  的阈值来确定其是否足够小。对于步长规则，我们提供了两种选择步长的策略：第一种是使用线搜索确定  $\eta_k$ ，第二种是采用固定半径信赖域方法的思想 [111,177]，使得  $\|\eta_k d_k\| = \Delta$ ，其中  $\Delta$  是某个预先确定的常数。通过迭代执行该子程序，我们的算法将收敛到  $\epsilon$ -近似的 SOSP。

**算法 3-1:** 齐次二阶下降方法 (HSODM)

---

**Data:** 初始点  $x_1, \nu \in (0, 1/2), \Delta = \Theta(\sqrt{\epsilon})$

```

1 for  $k = 1, 2, \dots$  do
2   求解子问题 (3-9), 并获得解  $[v_k; t_k]$ ;
3   if  $|t_k| > \sqrt{1/(1 + \Delta^2)}$  then                                     // 小值情况
4      $d_k \leftarrow v_k/t_k$ ;
5     更新  $x_{k+1} \leftarrow x_k + d_k$ ;
6     (提前) 终止 (或设置  $\delta = 0$  后继续);
7   if  $|t_k| \geq \nu$  then                                               // 大值情况 (a)
8      $d_k \leftarrow v_k/t_k$ 
9   else                                                                 // 大值情况 (b)
10     $d_k \leftarrow \text{sign}(-g_k^T v_k) \cdot v_k$ 
11    使用固定半径策略或线搜索策略选择步长  $\eta_k$  (参见算法 3-2);
12    更新  $x_{k+1} \leftarrow x_k + \eta_k \cdot d_k$ ;
```

---

### 3.2.3 齐次二次模型的初步分析

在本小节中, 我们对齐次二次模型进行一些初步分析。首先, 我们研究 Hessian  $H_k$  和  $F_k$  的最小特征值与扰动参数  $\delta$  之间的关系。随后, 我们给出问题 (3-9) 的最优性条件, 并基于这些条件提供一些有用的结果。

**引理 3.1:** 关于  $\lambda_1(F_k)$ 、 $\lambda_1(H_k)$  和  $\delta$  的关系

设  $\lambda_1(H_k)$  和  $\lambda_1(F_k)$  分别是  $H_k$  和  $F_k$  的最小特征值。用  $\mathcal{S}_{\lambda_1}$  表示  $\lambda_1(H_k)$  对应的特征空间。

如果  $g_k \neq 0$  且  $H_k \neq 0$ , 则以下结论成立:

- (1)  $\lambda_1(F_k) < -\delta$  且  $\lambda_1(F_k) \leq \lambda_1(H_k)$ ;
- (2) 仅当  $\lambda_1(H_k) < 0$  且  $g_k \perp \mathcal{S}_{\lambda_1}$  时,  $\lambda_1(F_k) = \lambda_1(H_k)$ 。

*Proof.* 我们首先证明结论 (1)。根据 Cauchy 交错定理<sup>[139]</sup>, 我们立即得到  $\lambda_1(F_k) \leq \lambda_1(H_k)$ 。现在我们需要证明  $\lambda_1(F_k) < -\delta$ 。只需证明矩阵  $F_k + \delta I$  存在一个负特征值。

考虑方向  $[-\eta g_k; t]$ , 其中  $\eta, t > 0$ . 定义以下关于  $(\eta, t)$  的函数:

$$\begin{aligned} f(\eta, t) &:= \begin{bmatrix} -\eta g_k \\ t \end{bmatrix}^T (F_k + \delta I) \begin{bmatrix} -\eta g_k \\ t \end{bmatrix}, \\ &= \eta^2 g_k^T (H_k + \delta I) g_k - 2\eta t \|g_k\|^2. \end{aligned}$$

对于任意固定的  $t > 0$ , 有

$$f(0, t) = 0 \quad \text{且} \quad \frac{\partial f(0, t)}{\partial \eta} = -2t \|g_k\|^2 < 0.$$

因此, 对于足够小的  $\eta > 0$ , 成立  $f(\eta, t) < 0$ , 这表明  $[-\eta g_k; t]$  是一个负曲率方向。由此可得  $\lambda_1(F_k) < -\delta$ 。对于结论 (2) 的证明, 与<sup>[146]</sup>中定理 3.1 的证明类似, 因此为简洁起见, 此处省略。□

**引理 3.1** 表明, 通过调整扰动参数  $\delta$ , 可以控制齐次矩阵  $F_k$  的最小特征值。这有助于找到更好的方向来降低目标函数的值。同时需要注意,  $g_k \perp \mathcal{S}_{\lambda_1}$  的情况通常被认为是解决信赖域子问题中的一个难点。然而, 在我们的收敛性分析中, 这一挑战不会对 HSODM 构成障碍。接下来我们将证明, 在此情形下, 函数值具有足够的下降。因此, 由于不存在难点情况, HSODM 的子问题比信赖域子问题更易于求解。

我们指出, **引理 3.1** 是<sup>[146]</sup>中引理 3.3 的简化版本, 其中作者更详细地分析了扰动参数  $\delta$  与齐次矩阵  $F_k$  的特征值对之间的关系。然而, 不同之处在于, 他们尝试通过齐次化技巧来求解信赖域子问题, 而我们的目标是寻求一个好的方向以减少函数值。此外, 如果使用齐次模型, 我们可以证明 HSODM 拥有最优的  $O(\epsilon^{-3/2})$  迭代复杂度。然而, 如果将齐次化技巧用于求解信赖域子问题(如<sup>[146]</sup>), 仍然需要类似 Curtis et al.<sup>[41]</sup>中的框架来保证相同的收敛性质。此外, 在该框架的每次迭代中, 还需要解决一系列齐次问题。

在以下引理中, 我们基于标准信赖域子问题的最优性条件, 刻画了问题 (3-9) 的最优解  $[v_k; t_k]$ 。

### 引理 3.2: 最优性条件

$[v_k; t_k]$  是子问题 (3-9) 的最优解, 当且仅当存在一个对偶变量  $\theta_k > \delta \geq 0$ , 使得

$$\begin{bmatrix} H_k + \theta_k \cdot I & g_k \\ g_k^T & -\delta + \theta_k \end{bmatrix} \geq 0, \quad (3-10)$$

$$\begin{bmatrix} H_k + \theta_k \cdot I & g_k \\ g_k^T & -\delta + \theta_k \end{bmatrix} \begin{bmatrix} v_k \\ t_k \end{bmatrix} = 0, \quad (3-11)$$

$$\|[v_k; t_k]\| = 1. \quad (3-12)$$

此外,  $-\theta_k$  是扰动后的齐次矩阵  $F_k$  的最小特征值, 即  $-\theta_k = \lambda_1(F_k)$ 。



*Proof.* 根据标准信赖域子问题的最优性条件,  $[v_k; t_k]$  是最优解, 当且仅当存在一个对偶变量  $\theta_k \geq 0$ , 使得

$$\begin{bmatrix} H_k + \theta_k \cdot I & g_k \\ g_k^T & -\delta + \theta_k \end{bmatrix} \geq 0, \begin{bmatrix} H_k + \theta_k \cdot I & g_k \\ g_k^T & -\delta + \theta_k \end{bmatrix} \begin{bmatrix} v_k \\ t_k \end{bmatrix} = 0, \text{ 且 } \theta_k \cdot (\|[v_k; t_k]\| - 1) = 0$$

结合 [引理 3.1](#), 我们有  $\lambda_1(F_k) < -\delta \leq 0$ 。因此,  $\theta_k \geq -\lambda_1(F_k) > \delta \geq 0$ , 进一步可得  $\|[v_k; t_k]\| = 1$ 。此外, 由 (3-11), 我们得到

$$\begin{bmatrix} H_k & g_k \\ g_k^T & -\delta \end{bmatrix} \begin{bmatrix} v_k \\ t_k \end{bmatrix} = -\theta_k \begin{bmatrix} v_k \\ t_k \end{bmatrix}.$$

将上述方程左乘  $[v_k; t_k]^T$ , 我们有

$$\min_{\|[v; t]\| \leq 1} \psi_k(v, t; \delta) = -\theta_k$$

注意, 由 (3-12), 问题 (3-9) 的最优值等价于  $F_k$  的最小特征值, 即  $\lambda_1(F_k)$ 。因此,  $-\theta_k = \lambda_1(F_k)$ 。证明完毕。  $\square$

根据上述最优性条件, 我们可以推导以下推论。

### 推论 3.3

[引理 4.1](#) 中的方程 (3-11) 可以重写为:

$$(H_k + \theta_k I)v_k = -t_k g_k \quad \text{和} \quad g_k^T v_k = t_k(\delta - \theta_k) \quad (3-13)$$

此外,

(1) 如果  $t_k = 0$ , 则有

$$(H_k + \theta_k I)v_k = 0 \quad \text{和} \quad g_k^T v_k = 0 \quad (3-14)$$

这意味着  $(-\theta_k, v_k)$  是 Hessian 矩阵  $H_k$  的特征值对。

(2) 如果  $t_k \neq 0$ , 则有

$$g_k^T d_k = \delta - \theta_k \quad \text{和} \quad (H_k + \theta_k \cdot I)d_k = -g_k \quad (3-15)$$

其中  $d_k = v_k/t_k$ 。

上述推论是 [引理 4.1](#) 的直接应用, 因此在本文中省略其证明。

### 推论 3.4: 方向 $v_k$ 的非平凡性

如果  $g_k \neq 0$ , 则有  $v_k \neq 0$ 。

*Proof.* 采用反证法。假定  $v_k = 0$ 。由 [推论 3.3](#) 中的方程 (3-13), 我们有  $t_k g_k = 0$ 。由于  $g_k \neq 0$ , 进一

步得到  $t_k = 0$ 。然而，这与最优性条件中的  $\| [v_k; t_k] \| = 1$  矛盾。因此，我们有  $v_k \neq 0$ 。

□

该推论表明，总能找到一个非平凡方向  $v_k$ ，因此 算法 3-1 不会停滞。

#### 推论 3.5

对于符号函数值  $\text{sign}(-g_k^T v_k)$ ，总有  $\text{sign}(-g_k^T v_k) \cdot t_k = |t_k|$ 。

*Proof.* 根据最优性条件方程 (3-13) 的第二个方程，并且  $\delta < \theta_k$ ，我们有

$$\text{sign}(-g_k^T v_k) = \text{sign}(t_k),$$

因此

$$\text{sign}(-g_k^T v_k) \cdot t_k = \text{sign}(t_k) \cdot t_k = |t_k|$$

证明完毕。

□

作为副产品，我们还得到以下结果。

#### 推论 3.6: 平凡情况, $g_k = 0$

假定  $g_k = 0$ ，则以下结论成立：

- (1) 如果  $\lambda_1(H_k) > -\delta$ ，则  $t_k = 1$ 。
- (2) 如果  $\lambda_1(H_k) < -\delta$ ，则  $t_k = 0$ 。

*Proof.* 当  $g_k = 0$  时，齐次矩阵  $F_k = [H_k, 0; 0, -\delta]$ ，子问题 (3-9) 为

$$\min_{\| [v; t] \| \leq 1} \psi_k(v, t; \delta) = v^T H_k v - t^2 \cdot \delta$$

我们首先通过反证法证明结论 (1)。假定  $t_k \neq 1$ ，则由方程 (3-12)，可得  $v_k \neq 0$ 。因此，

$$\psi_k(v_k, t_k; \delta) = (v_k)^T H_k v_k - t_k^2 \cdot \delta > -\delta = \psi_k(0, 1; \delta), \quad (3-16)$$

其中，不等式成立是由于  $(v_k)^T H_k v_k \geq \lambda_1(H_k) \|v_k\|^2 > -\delta \|v_k\|^2$ 。方程 (3-16) 与  $(v_k, t_k)$  的最优性矛盾，因此  $t_k = 1$ 。结论 (2) 可以用相同的方法证明，这里省略。

□

### 第三节 全局收敛速率

在本节中，我们分析所提出的 HSODM 的收敛速率。为便于分析，我们分别针对  $\|d_k\|$  的大值和小值情况提出两个构建块。在  $\|d_k\|$  的大值情况下，我们证明，在仔细选择扰动参数  $\delta$  后，每次迭代的函数值至少减少  $\Omega(\epsilon^{3/2})$ 。而在后者情况下，我们证明下一次迭代点  $x_{k+1}$  已经是一个  $\epsilon$ -近似 SOSp，因此算法可以终止。本文始终假定以下标准条件成立。

#### 假设 3.7

假定  $f$  的 Hessian 在包含所有迭代点  $x_k$  的一个开凸集  $X$  上是  $M$ -Lipschitz 连续的，即存在某个  $M > 0$ ，满足

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq M\|x - y\|, \forall x, y \in X, \quad (3-17)$$

并且 Hessian 矩阵有界，

$$\|\nabla^2 f(x_k)\| \leq U_H, \forall k \geq 0, \quad (3-18)$$

其中  $U_H > 0$ 。

我们还引入以下引理作为准备。

#### 引理 3.8: Nesterov<sup>[125]</sup>

如果  $f: \mathbb{R}^n \mapsto \mathbb{R}$  满足 假设 3.7，则对于所有  $x, y \in \mathbb{R}^n$ ，有：

$$\|\nabla f(y) - \nabla f(x) - \nabla^2 f(x)(y - x)\| \leq \frac{M}{2}\|y - x\|^2, \quad (3-19a)$$

$$\left| f(y) - f(x) - \nabla f(x)^T(y - x) - \frac{1}{2}(y - x)^T \nabla^2 f(x)(y - x) \right| \leq \frac{M}{6}\|y - x\|^3. \quad (3-19b)$$

#### 3.3.1 关于 $\|d_k\|$ 大值情况的分析

在 HSODM 中，我们将  $\|d_k\|$  的大值情况定义为其范数大于信赖域半径  $\Delta$  的情况，即  $\|d_k\| > \Delta$ 。注意，在  $\nu \leq |t_k| \leq \sqrt{1/(1+\Delta^2)}$  的情况下，我们有  $\|d_k\| = \|v_k\|/|t_k| = \sqrt{1-|t_k|^2}/|t_k| \geq \Delta$ 。此外，当  $|t_k| \leq \nu$  且  $\nu \in (0, 1/2)$  时， $\|d_k\| = \|v_k\| = \sqrt{1-|t_k|^2} \geq \sqrt{3}/2 \geq \Delta = \Theta(\sqrt{\epsilon})$ 。因此，我们分别将这两种情况称为 算法 3-1 中的大值情况 (a) 和 (b)。在这种情况下，齐次方向可以是  $d_k = \text{sign}(-g_k^T v_k) \cdot v_k$  或  $d_k = v_k/t_k$ 。以下讨论表明，这两种步长选择策略均可导致足够的函数值下降。固定半径策略的分析更为简洁明了，但主要作为理论结果。相比之下，尽管线搜索步长选择策略的分析稍显复杂，但其更为实际。

## 3.3.1.1 固定半径策略

对于固定半径策略，下一次迭代  $x_{k+1}$  被约束满足  $\|x_{k+1} - x_k\| = \Delta$ ，因此步长被选择为  $\Delta/\|d_k\|$ 。首先，我们考虑  $|t_k| < \nu$  且  $d_k = \text{sign}(-g_k^T v_k) \cdot v_k$  的情形。我们指出，此特定情形包含了信赖域方法中的所谓“难点情况”( $t_k = 0$ )<sup>[146]</sup>。当  $t_k = 0$  时，推论 3.3 表明  $(-\theta_k, v_k)$  是 Hessian  $H_k$  的特征值对，且由于  $-\theta_k < -\delta \leq 0$ ， $v_k$  是一个充分负的曲率方向。因此，沿着  $v_k$  的方向以合适的步长移动将始终减少函数值<sup>[23]</sup>。我们首先给出一个适用于  $|t_k| < \nu$  情况的引理，可将其视为广义下降引理。

## 引理 3.9

假定假设 3.7 成立，且设置  $\nu \in (0, 1/2)$ 。如果  $|t_k| < \nu$ ，令  $d_k = \text{sign}(-g_k^T v_k) \cdot v_k$  且  $\eta_k = \Delta/\|d_k\|$ ，则有

$$f(x_{k+1}) - f(x_k) \leq -\frac{\Delta^2}{2}\delta + \frac{M}{6}\Delta^3 \quad (3-20)$$

*Proof.* 当  $d_k = \text{sign}(-g_k^T v_k) \cdot v_k$  时，由推论 3.3 中的最优性条件 (3-13) 和推论 3.5，我们有

$$d_k^T H_k d_k = -\theta_k \|d_k\|^2 - t_k^2 \cdot (\delta - \theta_k) \quad \text{和} \quad g_k^T d_k = |t_k| \cdot (\delta - \theta_k) \quad (3-21)$$

由于  $\eta_k = \Delta/\|d_k\| \in (0, 1)$ ，则  $\eta_k - \eta_k^2/2 \geq 0$ ，进一步有

$$\left(\eta_k - \frac{\eta_k^2}{2}\right) \cdot (\delta - \theta_k) \leq 0 \quad (3-22)$$

根据  $\nabla^2 f(x)$  的  $M$ -Lipschitz 连续性质，我们有

$$\begin{aligned} f(x_{k+1}) - f(x_k) &= f(x_k + \eta_k d_k) - f(x_k) \\ &\leq \eta_k \cdot g_k^T d_k + \frac{\eta_k^2}{2} \cdot d_k^T H_k d_k + \frac{M}{6} \eta_k^3 \|d_k\|^3 \\ &= \eta_k \cdot |t_k| \cdot (\delta - \theta_k) - \frac{\eta_k^2}{2} \cdot \theta_k \|d_k\|^2 - \frac{\eta_k^2}{2} \cdot t_k^2 \cdot (\delta - \theta_k) + \frac{M}{6} \eta_k^3 \|d_k\|^3 \end{aligned} \quad (3-23a)$$

$$\leq \eta_k \cdot t_k^2 \cdot (\delta - \theta_k) - \frac{\eta_k^2}{2} \cdot \theta_k \|d_k\|^2 - \frac{\eta_k^2}{2} \cdot t_k^2 \cdot (\delta - \theta_k) + \frac{M}{6} \eta_k^3 \|d_k\|^3 \quad (3-23b)$$

$$\begin{aligned} &= \left(\eta_k - \frac{\eta_k^2}{2}\right) \cdot t_k^2 \cdot (\delta - \theta_k) - \frac{\eta_k^2}{2} \cdot \theta_k \|d_k\|^2 + \frac{M}{6} \eta_k^3 \|d_k\|^3 \\ &\leq -\theta_k \cdot \frac{\Delta^2}{2} + \frac{M}{6} \Delta^3 \end{aligned} \quad (3-23c)$$

$$\leq -\frac{\Delta^2}{2}\delta + \frac{M}{6}\Delta^3, \quad (3-23d)$$

其中，(3-23a) 由 (3-21) 得到，(3-23b) 由于  $|t_k| < \nu < 1$  且  $\delta - \theta_k < 0$  成立。不等式 (3-23c) 由 (3-22) 和  $\eta_k = \Delta/\|d_k\|$  得到。□

现在我们讨论  $|t_k| \geq \nu$  的情况，并令更新方向  $d_k = v_k/t_k$ 。当  $\|d_k\|$  足够大，即  $\|d_k\| > \Delta$ ，我们可以通过以下引理得到相同的函数值下降。

**引理 3.10**

假定 [假设 3.7](#) 成立，且设置  $\nu \in (0, 1/2)$ 。如果  $|t_k| \geq \nu$  且  $\|v_k/t_k\| > \Delta$ ，令  $d_k = v_k/t_k$  且  $\eta_k = \Delta/\|d_k\|$ ，则有

$$f(x_{k+1}) - f(x_k) \leq -\frac{\Delta^2}{2}\delta + \frac{M}{6}\Delta^3 \quad (3-24)$$

*Proof.* 当  $t_k \neq 0$  时，由 [推论 3.3](#) 中的方程 (3-15)，我们有

$$d_k^T H_k d_k = -g_k^T d_k - \theta_k \|d_k\|^2 \quad \text{和} \quad g_k^T d_k = \delta - \theta_k \leq 0 \quad (3-25)$$

由于  $\eta_k = \Delta/\|d_k\| \in (0, 1)$ ，则  $\eta_k - \eta_k^2/2 \geq 0$ ，进一步有

$$\left( \eta_k - \frac{\eta_k^2}{2} \right) \cdot g_k^T d_k \leq 0 \quad (3-26)$$

根据  $\nabla^2 f(x)$  的  $M$ -Lipschitz 连续性质，我们有

$$\begin{aligned} f(x_{k+1}) - f(x_k) &= f(x_k + \eta_k d_k) - f(x_k) \\ &\leq \eta_k \cdot g_k^T d_k + \frac{\eta_k^2}{2} \cdot d_k^T H_k d_k + \frac{M}{6} \eta_k^3 \|d_k\|^3 \\ &= \left( \eta_k - \frac{\eta_k^2}{2} \right) \cdot g_k^T d_k - \theta_k \cdot \frac{\eta_k^2}{2} \|d_k\|^2 + \frac{M}{6} \eta_k^3 \|d_k\|^3 \end{aligned} \quad (3-27a)$$

$$\leq -\theta_k \cdot \frac{\eta_k^2}{2} \|d_k\|^2 + \frac{M}{6} \eta_k^3 \|d_k\|^3 \quad (3-27b)$$

$$\leq -\frac{\Delta^2}{2}\delta + \frac{M}{6}\Delta^3, \quad (3-27c)$$

其中，(3-27a) 由方程 (3-25) 得到，(3-27b) 由方程 (3-26) 得到，而在 (3-27c) 中，我们用  $\eta_k = \Delta/\|d_k\|$  替换  $\eta_k$  并使用  $\theta_k \geq \delta$ 。□

### 3.3.1.2 线搜索策略

对于线搜索策略，我们利用回溯子程序来确定步长  $\eta_k$ ，以确保产生足够的函数值下降。该子程序的细节如下所示。

**算法 3-2:** 回溯线搜索

**Data:** 当前迭代点  $x_k$ , 方向  $d_k$ , 初始步长  $\eta_k = 1$ ,  $\gamma > 0$ ,  $\beta \in (0, 1)$

```

1 For  $j = 0, 1, 2, \dots$  do:
2   计算减少量  $D_k := f(x_k) - f(x_k + \eta_k d_k)$ ;
3   If  $D_k \geq \gamma \eta_k^3 \|d_k\|^3 / 6$  then:
4     Break;
5   Else:
6     更新  $\eta_k := \beta \cdot \eta_k$ ;
7 Output: 步长  $\eta_k$ .

```

类似地, 我们推导了采用线搜索策略时的下降引理, 并进一步对线搜索程序所需的迭代次数进行了上界估计。对于  $|t_k| < \nu$  和  $|t_k| \geq \nu$  的情况, 我们得到了以下两个描述足够下降性质的引理。

**引理 3.11**

假定 [假设 3.7](#) 成立, 且设置  $\nu \in (0, 1/2)$ 。如果  $|t_k| < \nu$ , 令  $d_k = \text{sign}(-g_k^T v_k) \cdot v_k$ 。回溯线搜索以  $\eta_k = \beta^{j_k}$  终止, 且  $j_k$  的上界为

$$j_N := \left\lceil \log_{\beta} \left( \frac{3\delta}{M + \gamma} \right) \right\rceil,$$

且与步长  $\eta_k$  相关的函数值满足

$$f(x_{k+1}) - f(x_k) \leq -\min \left\{ \frac{\sqrt{3}\gamma}{16}, \frac{9\gamma\beta^3\delta^3}{2(M + \gamma)} \right\} \quad (3-28)$$

*Proof.* 假定回溯线搜索在  $\eta_k = 1$  处终止, 则有

$$f(x_k + \eta_k d_k) - f(x_k) \leq -\frac{\gamma}{6} \eta_k^3 \|d_k\|^3 = -\frac{\gamma}{6} \|v_k\|^3 \leq -\frac{\sqrt{3}\gamma}{16},$$

其中最后一个不等式源于  $\|v_k\| = \sqrt{1 - |t_k|^2} \geq \sqrt{1 - \nu^2} \geq \sqrt{3}/2$ 。假定算法在某次迭代  $j \geq 0$  未终止,

且第 4 行的条件未满足, 即  $D_k < \frac{\gamma}{6}\beta^{3j}\|d_k\|^3 = \frac{\gamma}{6}\beta^{3j}\|v_k\|^3$ 。类似于引理 3.9 的证明, 我们有

$$\begin{aligned}
 -\frac{\gamma}{6}\beta^{3j}\|v_k\|^3 &< f(x_k + \beta^j d_k) - f(x_k) \\
 &\leq \beta^j \cdot g_k^T d_k + \frac{\beta^{2j}}{2} \cdot d_k^T H_k d_k + \frac{M}{6}\beta^{3j}\|d_k\|^3 \\
 &= \beta^j \cdot |t_k| \cdot (\delta - \theta_k) - \frac{\beta^{2j}}{2} \cdot \theta_k \|v_k\|^2 - \frac{\beta^{2j}}{2} \cdot t_k^2 \cdot (\delta - \theta_k) + \frac{M}{6}\beta^{3j}\|v_k\|^3 \\
 &\leq \beta^j \cdot t_k^2 \cdot (\delta - \theta_k) - \frac{\beta^{2j}}{2} \cdot \theta_k \|v_k\|^2 - \frac{\beta^{2j}}{2} \cdot t_k^2 \cdot (\delta - \theta_k) + \frac{M}{6}\beta^{3j}\|v_k\|^3 \\
 &= \left( \beta^j - \frac{\beta^{2j}}{2} \right) \cdot t_k^2 \cdot (\delta - \theta_k) - \frac{\beta^{2j}}{2} \cdot \theta_k \|v_k\|^2 + \frac{M}{6}\beta^{3j}\|v_k\|^3 \\
 &\leq -\frac{\beta^{2j}}{2} \cdot \theta_k \|v_k\|^2 + \frac{M}{6}\beta^{3j}\|v_k\|^3 \\
 &\leq -\frac{\beta^{2j}}{2} \cdot \delta \|v_k\|^2 + \frac{M}{6}\beta^{3j}\|v_k\|^3
 \end{aligned} \tag{3-29}$$

因此,  $\beta^j > \frac{3\delta}{(M+\gamma)\|v_k\|}$ , 进一步推出

$$j < \log_\beta \left( \frac{3\delta}{(M+\gamma)\|v_k\|} \right)$$

但由于  $\|v_k\| \leq 1$ , 因此  $j_N := \left\lceil \log_\beta \left( \frac{3\delta}{M+\gamma} \right) \right\rceil \geq \log_\beta \left( \frac{3\delta}{(M+\gamma)\|v_k\|} \right)$ 。

这意味着当  $j = j_N$  时, 不等式 (3-29) 不成立, 因此第 4 行的条件在此情况下得以满足。因此, 回溯子程序的迭代次数  $j_k$  的上界为  $j_N$ , 且函数值下降满足

$$\begin{aligned}
 f(x_k + \eta_k d_k) - f(x_k) &\leq -\frac{\gamma}{6}\beta^{3j_k}\|v_k\|^3 \\
 &= -\frac{\gamma\beta^3}{6}\beta^{3(j_k-1)}\|v_k\|^3 \\
 &\leq -\frac{9\gamma\beta^3\delta^3}{2(M+\gamma)^3},
 \end{aligned}$$

其中最后一个不等式来自  $\beta^{j_k-1} \geq \frac{3\delta}{(M+\gamma)\|v_k\|}$ 。 □

### 引理 3.12

假定假设 3.7 成立, 且设置  $\nu \in (0, 1/2)$ 。如果  $|t_k| \geq \nu$  且  $\|v_k/t_k\| > \Delta$ , 令  $d_k = v_k/t_k$ 。回溯线搜索以  $\eta_k = \beta^{j_k}$  终止, 且  $j_k$  的上界为

$$j_N := \left\lceil \log_\beta \left( \frac{3\delta\nu}{M+\gamma} \right) \right\rceil,$$

且与步长  $\eta_k$  相关的函数值满足

$$f(x_{k+1}) - f(x_k) \leq -\min \left\{ \frac{\gamma\Delta^3}{6}, \frac{9\gamma\beta^3\delta^3}{2(M+\gamma)^3} \right\}. \tag{3-30}$$

*Proof.* 类似地, 假定回溯线搜索在  $\eta_k = 1$  处终止, 则有

$$\begin{aligned} f(x_k + \eta_k d_k) - f(x_k) &\leq -\frac{\gamma}{6} \eta_k^3 \|d_k\|^3 \\ &\leq -\frac{\gamma}{6} \Delta^3, \end{aligned}$$

其中最后一个不等式来自  $\|d_k\| > \Delta$ 。如果  $\eta_k = 1$  未导致足够的函数值下降, 则对于任意满足第 4 行条件未满足的  $j \geq 0$ , 我们有

$$\begin{aligned} -\frac{\gamma}{6} \beta^{3j} \|d_k\|^3 &< f(x_k + \beta^j d_k) - f(x_k) \\ &\leq \beta^j \cdot g_k^T d_k + \frac{\beta^{2j}}{2} \cdot d_k^T H_k d_k + \frac{M}{6} \beta^{3j} \|d_k\|^3 \\ &= (\beta^j - \frac{\beta^{2j}}{2}) \cdot (\delta_k - \theta_k) - \frac{\beta^{2j}}{2} \theta_k \|d_k\|^2 + \frac{M}{6} \beta^{3j} \|d_k\|^3 \\ &\leq -\frac{\beta^{2j}}{2} \delta \|d_k\|^2 + \frac{M}{6} \beta^{3j} \|d_k\|^3 \end{aligned} \tag{3-31}$$

因此,  $\beta^j \geq \frac{3\delta}{(M+\gamma)\|d_k\|}$  并且进一步推出

$$j < \log_\beta \left( \frac{3\delta}{(M+\gamma)\|d_k\|} \right)$$

注意到

$$\|d_k\| = \|v_k\|/|t_k| = \frac{\sqrt{1-|t_k|^2}}{|t_k|} \leq \frac{1}{\nu}, \tag{3-32}$$

且  $j_N := \left\lceil \log_\beta \left( \frac{3\delta\nu}{M+\gamma} \right) \right\rceil \geq \log_\beta \left( \frac{3\delta}{(M+\gamma)\|d_k\|} \right)$ 。这意味着当  $j = j_N$  时, 不等式 (3-31) 不成立, 因此第 4 行的条件在此情况下得以满足。回溯子程序的迭代次数  $j_k$  的上界为  $j_N$ , 且函数值下降满足

$$\begin{aligned} f(x_k + \eta_k d_k) - f(x_k) &\leq -\frac{\gamma}{6} \beta^{3j_k} \|d_k\|^3 \\ &= -\frac{\gamma\beta^3}{6} \beta^{3(j_k-1)} \|d_k\|^3 \\ &\leq -\frac{9\gamma\beta^3\delta^3}{2(M+\gamma)^3}, \end{aligned}$$

其中最后一个不等式源于  $\beta^{j_k-1} \geq \frac{3\delta}{(M+\gamma)\|d_k\|}$ 。 □

结合上述两个引理, 我们现在总结出带有回溯线搜索的齐次负曲率的一致下降性质。

### 推论 3.13

假定 [假设 3.7](#) 成立, 并设置  $\nu \in (0, 1/2)$ 。令回溯线搜索参数  $\beta, \gamma$  满足  $\beta \in (0, 1)$  且  $\gamma > 0$ 。则在每次外部迭代之后, 函数值下降满足

$$f(x_{k+1}) - f(x_k) \leq -\min \left\{ \frac{\sqrt{3}\gamma}{16}, \frac{9\gamma\beta^3\delta^3}{2(M+\gamma)}, \frac{\gamma\Delta^3}{6}, \frac{9\gamma\beta^3\delta^3}{2(M+\gamma)^3} \right\}.$$

回溯线搜索的内部迭代次数最多为

$$j_N \leq \max \left\{ \left\lceil \log_\beta \left( \frac{3\delta}{M+\gamma} \right) \right\rceil, \left\lceil \log_\beta \left( \frac{3\delta\nu}{M+\gamma} \right) \right\rceil \right\} = \left\lceil \log_\beta \left( \frac{3\delta\nu}{M+\gamma} \right) \right\rceil.$$



**Remark 3.14**

**推论 3.13** 的一个有趣含义是目标函数值的下降量几乎不受截断参数  $\nu$  的选择影响。 $\nu$  的选择仅影响回溯线搜索的迭代次数，该迭代次数为  $O(\log_{\beta}(\delta\nu))$ 。然而，不建议选择较小的  $\nu$ ，因为这会随着  $\beta < 1$  增大线搜索的复杂度。

### 3.3.2 关于 $\|d_k\|$ 小值情况的分析

在本小节中，我们讨论  $\|d_k\| \leq \Delta$  的小值情况。注意，当  $|t_k| \geq \sqrt{1/(1+\Delta)}$  时，有  $\|d_k\| = \|v_k\|/|t_k| = \sqrt{1-|t_k|^2}/|t_k| \leq \Delta$ ，这验证了在 **algorithm 3-1** 中将其称为小值情况的合理性。在这种情况下，我们证明下一次迭代  $x_{k+1} = x_k + d_k$  已经是一个  $\epsilon$ -近似 SOSP。因此，在小值情况下，算法可以在一次迭代后终止。为了证明这一结果，我们首先对  $\|g_k\|$  提供一个上界作为准备。

**引理 3.15**

假定 **假设 3.7** 成立。如果  $g_k \neq 0$  且  $\|d_k\| \leq \Delta \leq \sqrt{2}/2$ ，则有

$$\|g_k\| \leq 2(U_H + \delta)\Delta. \quad (3-33)$$

*Proof.* 由 **引理 3.1** 可知， $\theta_k - \delta > 0$ 。此外，由 **推论 3.3** 中的方程 (3-15)，我们可以给出  $\theta_k - \delta$  的上界，即

$$\theta_k - \delta = -g_k^T d_k \leq \|g_k\| \|d_k\| \leq \Delta \|g_k\| \quad (3-34)$$

定义  $h(t) = t^2 + (g_k^T H_k g_k / \|g_k\|^2 + \delta)t - \|g_k\|^2$ 。显然，方程  $h(t) = 0$  必有两个符号相反的实根。令其正根为  $t_2$ 。由于  $\theta_k - \delta > 0$ ，我们有  $\theta_k - \delta \geq t_2$ 。因此，必须满足

$$h(\Delta \|g_k\|) = \Delta^2 \|g_k\|^2 + \left( \frac{g_k^T H_k g_k}{\|g_k\|^2} + \delta \right) \Delta \|g_k\| - \|g_k\|^2 \geq 0$$

经过一些代数运算，我们得到

$$\begin{aligned} \|g_k\| &\leq \frac{(g_k^T H_k g_k / \|g_k\|^2 + \delta) \Delta}{1 - \Delta^2} \\ &\leq \frac{(U_H + \delta) \Delta}{1 - \Delta^2} \\ &\leq 2(U_H + \delta) \Delta \end{aligned} \quad (3-35)$$

第二个不等式源于  $H_k \leq U_H I$ ，这意味着  $g_k^T H_k g_k / \|g_k\|^2 \leq U_H$ 。最后一个不等式由  $\Delta \leq \sqrt{2}/2$  推导出。□

以下引理表明， $x_{k+1}$  处梯度的范数有上界，同时 Hessian 的最小特征值有下界。

**引理 3.16**

假定 [假设 3.7](#) 成立。如果  $g_k \neq 0$  且  $\|d_k\| \leq \Delta$ ，则令  $\eta_k = 1$ ，我们有

$$\|g_{k+1}\| \leq 2(U_H + \delta)\Delta^3 + \frac{M}{2}\Delta^2 + \delta\Delta, \quad (3-36)$$

$$H_{k+1} \geq -(2(U_H + \delta)\Delta^2 + M\Delta + \delta)I. \quad (3-37)$$

*Proof.* 我们首先证明 (3-36)。根据 [推论 3.3](#) 中的最优性条件 (3-15)，有

$$H_k d_k + g_k = -\theta_k d_k,$$

结合 (3-34)，得到

$$\theta_k \|d_k\| \leq (\delta + \Delta \|g_k\|) \|d_k\|$$

因此，可以得到

$$\|H_k d_k + g_k\| = \theta_k \|d_k\| \leq \delta\Delta + \|g_k\|\Delta^2 \quad (3-38)$$

接下来我们对  $\|g_{k+1}\|$  的范数进行上界估计，得到

$$\begin{aligned} \|g_{k+1}\| &\leq \|g_{k+1} - H_k d_k - g_k\| + \|H_k d_k + g_k\| \\ &\leq \frac{M}{2} \|d_k\|^2 + \delta\Delta + \|g_k\|\Delta^2 \end{aligned} \quad (3-39a)$$

$$\begin{aligned} &\leq \frac{M}{2} \Delta^2 + \delta\Delta + 2(U_H + \delta)\Delta \cdot \Delta^2 \\ &= 2(U_H + \delta)\Delta^3 + \frac{M}{2}\Delta^2 + \delta\Delta, \end{aligned} \quad (3-39b)$$

其中，(3-39a) 由于  $\nabla^2 f(x)$  的  $M$ -Lipschitz 连续性以及方程 (3-38) 成立，(3-39b) 由 [引理 3.15](#) 得到。

现在我们证明 (3-37)。注意到 [引理 4.1](#) 中的最优性条件 (3-10) 表明

$$H_k + \theta_k \cdot I \geq 0$$

结合 (3-34) 和 (3-35)，进一步得到

$$\begin{aligned} H_k &\geq -\theta_k I \geq -(\Delta \|g_k\| + \delta)I \\ &\geq -2(U_H + \delta)\Delta^2 I - \delta I \end{aligned} \quad (3-40)$$

为了对  $H_{k+1}$  进行界定，我们有

$$H_{k+1} \geq H_k - \|H_{k+1} - H_k\|I \geq H_k - M\|d_k\|I \geq H_k - M\Delta I, \quad (3-41)$$

其中，第二个不等式由  $\nabla^2 f(x)$  的  $M$ -Lipschitz 连续性得到，最后一个不等式由  $\|d_k\| \leq \Delta$  推导出。

结合 (3-40)，我们得到

$$H_{k+1} \geq -2(U_H + \delta)\Delta^2 I - \delta I - M\Delta I. \quad (3-42)$$

证明完毕。  $\square$

### 3.3.3 全局收敛性

综合上述结果，我们在 [定理 3.17](#) 和 [定理 3.18](#) 中分别给出 HSODM 在固定半径和回溯线搜索策略下的形式化全局收敛结果。结果表明，通过适当选择扰动参数  $\delta$  和半径  $\Delta$ ，HSODM 能以  $O(\epsilon^{-3/2})$  的迭代复杂度找到  $\epsilon$ -近似的 SOSP。

#### 定理 3.17

假定 [假设 3.7](#) 成立。令  $\delta = \sqrt{\epsilon}$ ， $\Delta = 2\sqrt{\epsilon}/M$  且  $\nu \in (0, 1/2)$ ，则采用固定半径策略的 HSODM 最多在  $O(\epsilon^{-3/2})$  步内终止，并且下一步迭代  $x_{k+1}$  是一个 SOSP。

*Proof.* 由于取  $\delta = \sqrt{\epsilon}$  且  $\Delta = 2\sqrt{\epsilon}/M$ ，根据 [引理 3.9](#) 和 [引理 3.10](#)，我们立即得到在大步长情况下函数值至少减少  $\Omega(\epsilon^{3/2})$ ，即

$$f(x_{k+1}) - f(x_k) \leq -\frac{2}{3M^2}\epsilon^{3/2}$$

当算法终止时，根据 [引理 3.16](#)，我们有

$$\begin{aligned} \|g_{k+1}\| &\leq 2(U_H + \delta)\Delta^3 + \frac{M}{2}\Delta^2 + \delta\Delta \\ &\leq \frac{16U_H\epsilon^{3/2} + 16\epsilon^2}{M^3} + \frac{4\epsilon}{M} \leq O(\epsilon) \end{aligned} \quad (3-43)$$

和

$$\begin{aligned} \lambda_1(H_{k+1}) &\geq -(2(U_H + \delta)\Delta^2 + M\Delta + \delta) \\ &\geq -\left(\frac{8U_H\epsilon + 8\epsilon^{3/2}}{M^2} + 3\sqrt{\epsilon}\right) \geq \Omega(-\sqrt{\epsilon}) \end{aligned} \quad (3-44)$$

因此，下一步迭代  $x_{k+1}$  已经是一个 SOSP。

注意，目标函数值的总减少量不能超过  $f(x_1) - f_{\inf}$ 。因此，大步长情况下的迭代次数上界为

$$O\left(\frac{3M^2}{2}(f(x_1) - f_{\inf})\epsilon^{-3/2}\right)$$

这也是算法的迭代复杂度。  $\square$

#### 定理 3.18

假定 [假设 3.7](#) 成立。令  $\delta = \sqrt{\epsilon}$ ， $\Delta = 2\sqrt{\epsilon}/M$  且  $\nu \in (0, 1/2)$ ，并且回溯线搜索参数  $\beta, \gamma$  满足  $\beta \in (0, 1)$  且  $\gamma > 0$ 。则采用回溯线搜索的 HSODM 最多在  $O(\epsilon^{-3/2} \log_{\beta}(\epsilon))$  步内终止，并且

下一步迭代  $x_{k+1}$  是一个 SOSP。特别地, 迭代次数上界为

$$O\left(\max\left\{\frac{2(M+\gamma)}{9\gamma\beta^3}, \frac{3M^3}{4\gamma}, \frac{2(M+\gamma)^3}{9\gamma\beta^3}\right\}\left\lceil\log_\beta\left(\frac{3\sqrt{\epsilon}\nu}{M+\gamma}\right)\right\rceil(f(x_1) - f_{\inf})\epsilon^{-3/2}\right)$$

*Proof.* 由于取  $\delta = \sqrt{\epsilon}$  和  $\Delta = 2\sqrt{\epsilon}/M$ , 根据 [推论 3.13](#), 我们立即得到在大步长情况下函数值至少减少  $\Omega(\epsilon^{3/2})$ , 即

$$\begin{aligned} f(x_{k+1}) - f(x_k) &\leq -\min\left\{\frac{\sqrt{3}\gamma}{16}, \frac{9\gamma\beta^3\delta^3}{2(M+\gamma)}, \frac{\gamma\Delta^3}{6}, \frac{9\gamma\beta^3\delta^3}{2(M+\gamma)^3}\right\} \\ &\leq -\min\left\{\frac{9\gamma\beta^3}{2(M+\gamma)}, \frac{4\gamma}{3M^3}, \frac{9\gamma\beta^3}{2(M+\gamma)^3}\right\}\epsilon^{3/2}, \end{aligned}$$

并且回溯线搜索的内部迭代次数最多为

$$j_N \leq \left\lceil\log_\beta\left(\frac{3\delta\nu}{M+\gamma}\right)\right\rceil = \left\lceil\log_\beta\left(\frac{3\sqrt{\epsilon}\nu}{M+\gamma}\right)\right\rceil$$

当算法终止时, 类似于 (3-43) 和 (3-44), 我们有

$$\|g_{k+1}\| \leq O(\epsilon) \quad \text{和} \quad \lambda_1(H_{k+1}) \geq \Omega(-\sqrt{\epsilon})$$

因此, 下一步迭代  $x_{k+1}$  已经是一个 SOSP。

注意, 目标函数值的总减少量不能超过  $f(x_1) - f_{\inf}$ 。因此, 大步长情况下的迭代次数上界为

$$O\left(\max\left\{\frac{2(M+\gamma)}{9\gamma\beta^3}, \frac{3M^3}{4\gamma}, \frac{2(M+\gamma)^3}{9\gamma\beta^3}\right\}\left\lceil\log_\beta\left(\frac{3\sqrt{\epsilon}\nu}{M+\gamma}\right)\right\rceil(f(x_1) - f_{\inf})\epsilon^{-3/2}\right)$$

由于  $\beta < 1$ , 证明完毕。  $\square$

由于  $\delta = \sqrt{\epsilon}$ , 可以看出相比固定半径策略, 回溯线搜索版本额外引入了  $O(\log_\beta \epsilon)$  的开销。在实际中, 回溯线搜索版本可以选择远大于  $\Delta$  的步长, 因此收敛速度更快。这一优点可以在 [第 3.6 节](#) 中观察到。

#### 第四节 局部收敛率

在本节中, 我们对 HSODM 的局部收敛性进行分析。特别地, 当  $x_k$  足够接近一个  $\epsilon$ -近似的二阶稳定点 (SOSP)  $x^*$  时, 我们将证明步长  $\eta_k$  始终等于 1, 并且不需要进行线搜索过程。因此, 通过对后续迭代设置扰动参数  $\delta = 0$ , HSODM 实现了局部二次收敛率。

为了便于局部收敛性分析, 我们先引入标准假定 [\[38,125,130\]](#)。

##### 假设 3.19

假定 HSODM 收敛于一个严格的局部最优点  $x^*$ , 其满足  $\nabla f(x^*) = 0$  且  $\nabla^2 f(x^*) > 0$ 。

**Remark 3.20**

从 [假设 3.19](#) 我们可以立即得出, 存在一个半径为  $R > 0$  和  $\mu > 0$  的小邻域, 使得

$$\forall x \in B(x^*, R) \Rightarrow \nabla^2 f(x) \geq \mu \cdot I \quad (3-45)$$

换句话说, 对于足够大的  $k$ ,  $x_k$  会进入  $x^*$  的邻域, 从而  $H_k$  和  $H_k + \theta_k I$  都是非奇异的。

为了证明局部收敛率, 我们需要以下辅助结果作为准备。

**推论 3.21**

假定 [假设 3.19](#) 成立, 则当  $k$  足够大时, 有  $t_k \neq 0$ 。

*Proof.* 我们通过反证法证明。假定  $t_k = 0$ , 则根据 [推论 3.3](#),  $(-\theta_k, v_k)$  是  $H_k$  的特征对, 意味着

$$\lambda_1(H_k) \leq -\theta_k$$

回顾 [引理 4.1](#), 我们有  $\theta_k > 0$ , 因此  $\lambda_1(H_k) < 0$ 。

这与  $H_k > 0$  矛盾, 因此证明完成。 □

以下引理表明, 对于足够大的  $k$ , HSODM 生成的步长  $d_k$  最终会落入小步长情况。因此, 我们选择  $\eta_k = 1$  并通过  $x_{k+1} = x_k + d_k$  更新迭代, 如 [第 3.3.2 节](#) 中所示。需要注意的是, 这类类似于经典牛顿信赖域方法中的情况 (参见 <sup>[130]</sup> Theorem 4.9), 其中更新渐进地与纯牛顿步相似。

**引理 3.22**

对于足够大的  $k$ , 我们有  $\|d_k\| \leq \Delta$ 。

*Proof.* 由于  $t_k \neq 0$ , 根据 [推论 3.3](#) 中的方程 (3-15), 我们有

$$d_k = -(H_k + \theta_k I)^{-1} g_k,$$

进一步可得

$$\begin{aligned} \|d_k\| &\leq \|(H_k + \theta_k I)^{-1}\| \|g_k\| \\ &\leq \frac{\|g_k\|}{\mu + \theta_k} \leq \frac{\|g_k\|}{\mu} \end{aligned} \quad (3-46)$$

以上不等式成立是因为  $H_k \geq \mu I$  且  $\theta_k > 0$ 。注意, 根据 [假设 3.19](#),  $\|g_k\| \rightarrow 0$  当  $k \rightarrow \infty$ 。因此, 存在一个足够大的  $K \geq 0$ , 使得

$$\|g_k\| \leq \Delta \mu, \forall k \geq K \quad (3-47)$$

结合 (3-46), 我们得出  $\|d_k\| \leq \Delta$  成立。 □

在局部阶段，我们设置扰动参数  $\delta = 0$  并求解

$$\min_{\| [v; t] \| \leq 1} \psi_k(v, t; 0) := \begin{bmatrix} v \\ t \end{bmatrix}^T \begin{bmatrix} H_k & g_k \\ g_k^T & 0 \end{bmatrix} \begin{bmatrix} v \\ t \end{bmatrix}. \quad (3-48)$$

我们还用  $[v_k; t_k]$  表示 (3-48) 的最优解。基于以上结果，我们准备证明以下定理。

**定理 3.23**

假定 [假设 3.7](#) 和 [假设 3.19](#) 成立。当  $k$  足够大时，HSODM 以二次速度收敛到  $x^*$ ，即

$$\|x_{k+1} - x^*\| \leq \left( \frac{M}{\mu} + \frac{\Delta(MR + \mu)^2}{\mu^2(1 - \Delta^2)^2} \right) \|x_k - x^*\|^2$$

其中  $R$  定义如 (3-45)。

*Proof.* 由 [推论 3.21](#)，我们有  $t_k \neq 0$ 。由于我们取  $\delta = 0$ ，根据 [推论 3.3](#) 中的方程 (3-15)，我们有

$$g_k^T d_k = -\theta_k \quad \text{且} \quad (H_k + \theta_k I) d_k = -g_k,$$

这意味着

$$\begin{aligned} \|H_k^{-1} g_k + d_k\| &= \|-\theta_k H_k^{-1} d_k\| \\ &\leq \|H_k^{-1}\| \cdot |\theta_k| \|d_k\| \\ &\leq \frac{1}{\mu} \|g_k\| \|d_k\| \end{aligned} \quad (3-49)$$

由 [引理 3.22](#)，我们有  $x_{k+1} = x_k + d_k$ 。因此，

$$\begin{aligned} \|x_{k+1} - x^*\| &= \|x_k + d_k + H_k^{-1} g_k - H_k^{-1} g_k - x^*\| \\ &\leq \|x_k - H_k^{-1} g_k - x^*\| + \|H_k^{-1} g_k + d_k\| \\ &\leq \frac{M}{\mu} \|x_k - x^*\|^2 + \frac{1}{\mu} \|g_k\| \|d_k\|^2 \end{aligned} \quad (3-50a)$$

$$\leq \frac{M}{\mu} \|x_k - x^*\|^2 + \Delta \|d_k\|^2 \quad (3-50b)$$

其中 (3-50a) 因标准牛顿方法分析<sup>[130]</sup> 和方程 (3-49) 而成立，而 (3-50b) 则源于 [引理 3.22](#) 中的  $\|g_k\| \leq \Delta\mu$ 。进一步计算可得

$$\begin{aligned} \|d_k\| &= \|x_{k+1} - x^* - (x_k - x^*)\| \\ &\leq \|x_{k+1} - x^*\| + \|x_k - x^*\| \\ &\leq \frac{M}{\mu} \|x_k - x^*\|^2 + \|x_k - x^*\| + \Delta \|d_k\|^2 \\ &\leq \frac{MR}{\mu} \|x_k - x^*\| + \|x_k - x^*\| + \Delta^2 \|d_k\|, \end{aligned}$$

最后我们有

$$\|x_{k+1} - x^*\| \leq \frac{M}{\mu} \|x_k - x^*\|^2 + \Delta \|d_k\|^2$$

以上即为所需证明。  $\square$

## 第五节 非精确 HSODM

上述分析依赖于精确求解子问题 (3-9)，这需要进行矩阵分解，涉及  $O((n+1)^3)$  的算术操作。在本节中，我们提出了一种非精确 HSODM (见 算法 3-3)，该方法在每次迭代中利用 Lanczos 方法 (见 算法 3-4) 来近似求解 (3-9)。随后，我们基于  $F_k$  的 Ritz 对而非其精确的最左特征对构造迭代点。我们将证明该方法在概率意义下的最坏情况算术操作复杂度为  $\tilde{O}((n+1)^2 \epsilon^{-7/4})$ ，即对  $n$  的依赖性更小。为了方便叙述，我们简要回顾 Lanczos 方法的一些标准性质，并给出其误差界，其中一些性质在第 2.6.1 节中已经给出，较为熟悉的读者可以跳过先前内容。

### 3.5.1 Lanczos 方法概述

在深入细节之前，我们简要介绍 Lanczos 方法，它用于计算对称矩阵  $A \in \mathbb{R}^{n \times n}$  的极值特征值。在第  $j$  次迭代时，Lanczos 方法从  $j$  阶 Krylov 子空间  $\mathcal{K}(j; A, q_1) := \text{span}\{q_1, Aq_1, \dots, A^{j-1}q_1\}$  构造一个正交基  $Q_j = [q_1, q_2, \dots, q_j] \in \mathbb{R}^{n \times j}$ ，同时保持  $T_j = Q_j^T A Q_j$  为三对角矩阵。以下引理给出了 Lanczos 方法的一些标准性质。

#### 引理 3.24: Lanczos 方法的基本性质<sup>[71]</sup>

对任意对称矩阵  $A \in \mathbb{R}^{n \times n}$ ，令  $q_1 \in \mathbb{R}^n$  且  $\|q_1\| = 1$ 。假定 Lanczos 方法运行至第  $J = \text{rank}(\mathcal{K}(n; A, q_1))$  次迭代，则以下结论成立：

- (1) 对任意  $j = 1, 2, \dots, J$ ，令  $Q_j = [q_1, q_2, \dots, q_j]$  为  $\mathcal{K}(j; A, q_1)$  的正交基，则有

$$A Q_j = Q_j T_j + \xi_j (1_j)_{[1:j]}^T \quad \text{且} \quad Q_j \perp \xi_j,$$

其中  $T_j = Q_j^T A Q_j$  为三对角矩阵， $1_j \in \mathbb{R}^n$  为单位矩阵  $I_n$  的第  $j$  列， $\xi_j$  为残差向量。

- (2) 假定  $Y_j = Q_j S_j$  为 Lanczos 方法第  $j$  次 Krylov 迭代计算所得，且满足实 Schur 分解  $S_j^T T_j S_j = \Gamma_j$ 。令  $\gamma_i$  为  $\Gamma_j$  对角线上第  $i$  个元素， $y_i$  为  $Y_j$  的第  $i$  列向量，则以下误差估计成立：

$$A y_i - \gamma_i y_i = (1_j)_{[1:j]}^T S_j (1_i)_{[1:j]} \cdot \xi_j := s_{ji} \cdot \xi_j \quad \text{其中} \quad |s_{ji}| < 1 \quad \text{且} \quad y_i \perp \xi_j, \quad \forall i \leq j.$$

我们称  $(\gamma_i, y_i)$  为第  $i$  个 Ritz 对。

在本文的其余部分，我们有时会为了简洁省略索引  $[1:j]$ 。这隐含了矩阵-向量运算在大小上的兼容性。下面令  $[v_k; t_k]$  表示近似解。仍然令  $-\theta_k = \lambda_1(F_k)$  表示  $F_k$  的最小特征值，并用  $\chi_k$  表示其

特征向量。

**定理 3.25: 近似解的性质**

假定 Lanczos 方法用于近似求解 (3-9)，并返回一个 Ritz 对  $(-\gamma_k, [v_k; t_k])$ 。则有

$$\begin{bmatrix} H_k & g_k \\ g_k^T & -\delta \end{bmatrix} \begin{bmatrix} v_k \\ t_k \end{bmatrix} + \gamma_k \begin{bmatrix} v_k \\ t_k \end{bmatrix} = \begin{bmatrix} r_k \\ \sigma_k \end{bmatrix}, \quad (3-51a)$$

$$r_k^T v_k + \sigma_k \cdot t_k = 0. \quad (3-51b)$$

其中  $[r_k; \sigma_k] \in \mathbb{R}^n \times \mathbb{R}$  被称为 Ritz 误差。

上述定理是 引理 3.24 第 (2) 点的直接应用。由于  $(-\gamma_k, [v_k; t_k])$  只是一个近似解，我们引入一些误差估计  $e_k > 0$ ，使得  $|\theta_k - \gamma_k| \leq e_k$ 。在 Lanczos 方法中， $-\gamma_k$  总是  $-\theta_k$  的一个高估值<sup>[71]</sup>，因此我们终止条件为  $\theta_k - e_k \leq \gamma_k \leq \theta_k$ 。以下引理给出了关于指定误差  $e_k$  的复杂度估计。

**引理 3.26: Lanczos 方法的复杂度**

假定 Lanczos 方法用于近似求解 (3-9)，并返回一个 Ritz 对  $(-\gamma_k, [v_k; t_k])$ ，满足  $\theta_k - e_k \leq \gamma_k \leq \theta_k$  对某些  $e_k > 0$ 。则所需的迭代次数可由以下任一公式上界：

(1)

$$1 + \left\lceil 2 \sqrt{\frac{\|F_k\|}{e_k}} \log \left( \frac{16 \|F_k\|}{e_k (q_1^T \chi_k)^2} \right) \right\rceil, \quad (3-52)$$

其中  $(-\theta_k, \chi_k)$  是  $F_k$  的精确最左特征对<sup>[103,147]</sup>；

(2)

$$1 + \left\lceil \sqrt{\frac{2 \|F_k\|}{\lambda_2(F_k) - \lambda_1(F_k)}} \log \left( \frac{8 \|F_k\|}{e_k (q_1^T \chi_k)^2} \right) \right\rceil, \quad (3-53)$$

其中  $\lambda_2(F_k)$  是  $F_k$  的第二小特征值，满足  $\lambda_2(F_k) - \lambda_1(F_k) > 0$ <sup>[103]</sup>。

我们还注意到，Lanczos 方法具有有限收敛性。最后，我们将 Ritz 误差与所需的精度  $e_k$  联系起来。

**引理 3.27**

假定 假设 3.7 成立，并且  $F_k$  如 (3-8) 中所定义，则有

$$\|F_k\| \leq \max\{U_H, \delta\} + \|g_k\|. \quad (3-54)$$



如果令  $\varsigma_k := \lambda_2(F_k) - \lambda_1(F_k) > 0$ ，则对于 (3-51) 中的  $[r_k; \sigma_k]$ ，存在  $\tau_k \in [0, 1]$  使得

$$\|[r_k; \sigma_k]\| \leq \tau_k e_k + 2(\max\{U_H, \delta\} + \|g_k\|) \sqrt{\frac{e_k}{\varsigma_k}}. \quad (3-55)$$

我们将 引理 3.26 和 引理 5.30 的证明推迟到附录中，因为这些结果主要与线性代数相关。

### 3.5.2 非精确 HSODM 的概述

现在，我们准备介绍非精确 HSODM，见 算法 3-3。该算法遵循精确 HSODM 的基本思想，但使用 Lanczos 方法近似求解 式 (3-9)。非精确性在建立相应的收敛性结果时带来了若干挑战。首先，由于 Ritz 对中的  $\gamma_k$  是一个非精确的对偶变量，无法保证  $\gamma_k$  超过  $\delta$ ，这可能导致不足够的下降性质。其次，在小值情形 (当  $t_k > \sqrt{1/(1+\Delta^2)}$ ) 下，较大的 Ritz 误差可能阻止下一次迭代  $x_{k+1}$  成为 SOSP，即便我们通过  $x_{k+1} = x_k + d_k$  进行更新。

为克服第一个挑战，我们提出了一种自定义的带偏随机化的 Lanczos 方法 (算法 3-4)，该方法以高概率保证  $\gamma_k$  总是不小于  $\delta$  (参见 定理 3.31 和 定理 3.32)。针对第二个挑战，我们讨论  $\|r_k\|$  的量级。如果  $\|r_k\|$  足够小，我们可以安全地断定  $x_{k+1} = x_k + d_k$  已经是一个 SOSP (引理 3.35)。否则，我们增加扰动参数  $\delta$  并求解子问题 式 (3-9)。通过对谱的精细分析，我们证明了齐次矩阵  $F_k$  的特征值间隔  $\varsigma_k$  足够大 (例如  $\Omega(\sqrt{\epsilon})$ )。这表明可以在指示的误差上追求更高的精度 (见 行 10)。

**算法 3-3:** 非精确齐次二阶下降法 (Inexact HSODM)

**Input:** 初始点  $x_1$ ,  $\nu \in (1/4, 1/2)$ ,  $\Delta = \sqrt{\epsilon}/M$ ,  $\epsilon > 0$ .

```

1 for  $k = 1, 2, \dots$  do
2   设置  $\delta \leftarrow \sqrt{\epsilon}$ ,  $e_k \leftarrow \sqrt{\epsilon}$ ,  $J_{\max} \leftarrow n + 1$ ;
3   使用 算法 3-4, 输入  $(\delta, e_k, J_{\max})$ , 得到 Ritz 对  $(\gamma_k, [v_k; t_k])$  和 Ritz 误差  $[r_k; \sigma_k]$ ;
4   if  $|t_k| > \sqrt{1/(1 + \Delta^2)}$  then                                     // 小值情形
5       if  $\|r_k\| \leq 2\epsilon$  then
6           设置  $d_k \leftarrow v_k/t_k$ ;
7           更新  $x_{k+1} \leftarrow x_k + d_k$ ;
8           (提前) 终止 (或设置  $\delta = 0$  并继续);
9       else
10          设置  $\delta \leftarrow 3\sqrt{\epsilon} + 2\|g_k\|\Delta + (U_H + \gamma_k)\Delta^2$ ,  $e_k = \min \left\{ \epsilon, \frac{\epsilon_{\text{abs}}}{4(U_H + U_g)^2} \right\}$ ;
11          返回 行 3;
12  if  $|t_k| \geq \nu$  then                                               // 大值情形 (a)
13      设置  $d_k \leftarrow v_k/t_k$ ;
14  else                                                                // 大值情形 (b)
15      设置  $d_k \leftarrow \text{sign}(-g_k^T v_k) \cdot v_k$ ;
16  使用固定半径策略选择步长  $\eta_k$ ;
17  更新  $x_{k+1} \leftarrow x_k + \eta_k \cdot d_k$ ;
    
```

在本小节接下来的部分中, 我们分析非精确 HSODM 中大值情形 (a) 和 (b) 下的下降性质 (见 [算法 3-3](#) 的 [行 12](#) 和 [行 14](#))。它们的分析与精确 HSODM 中的类似, 我们的分析表明非精确性确实对收敛性分析带来了一些障碍。

**引理 3.28: 大值情形 (a)**

假定 [假设 3.7](#) 成立并设置  $\nu \in (1/4, 1/2)$ 。如果  $|t_k| \geq \nu$  且  $\|v_k/t_k\| \geq \Delta$ , 令  $d_k = v_k/t_k$  且  $\eta_k = \Delta/\|d_k\|$ , 则有

$$f(x_{k+1}) - f(x_k) \leq \left( \eta_k - \frac{1}{2}\eta_k^2 \right) (\delta - \gamma_k) + 4|\sigma_k| - \frac{\gamma_k}{2}\Delta^2 + \frac{M}{6}\Delta^3.$$

*Proof.* 根据 [\(3-51a\)](#) 和  $d_k = v_k/t_k$ , 我们有

$$d_k^T H_k d_k + g_k^T d_k = -\gamma_k \|d_k\|^2 + \frac{r_k^T v_k}{t_k^2},$$

$$g_k^T d_k = -\gamma_k + \delta + \frac{\sigma_k}{t_k}.$$

因此, 我们得到

$$\begin{aligned} f(x_{k+1}) - f(x_k) &= f(x_k + \eta_k \cdot d_k) - f(x_k) \\ &\leq \eta_k \cdot g_k^T d_k + \frac{\eta_k^2}{2} \cdot d_k^T H_k d_k + \frac{M\eta_k^3}{6} \cdot \|d_k\|^3 \\ &= \eta_k \cdot g_k^T d_k + \frac{1}{2}\eta_k^2 \left( \frac{r_k^T v_k}{t_k^2} - g_k^T d_k - \gamma_k \|d_k\|^2 \right) + \frac{M\eta_k^3}{6} \cdot \|d_k\|^3 \\ &= \left( \eta_k - \frac{1}{2}\eta_k^2 \right) \left( \frac{\sigma_k}{t_k} + \delta - \gamma_k \right) + \frac{\eta_k^2}{2} \left( \frac{r_k^T v_k}{t_k^2} \right) - \frac{\gamma_k}{2} \Delta^2 + \frac{M}{6} \Delta^3 \\ &= \left( \eta_k - \frac{1}{2}\eta_k^2 \right) (\delta - \gamma_k) - (\eta_k^2 - \eta_k) \frac{\sigma_k}{t_k} - \frac{\gamma_k}{2} \Delta^2 + \frac{M}{6} \Delta^3. \end{aligned}$$

最后一个等式成立是因为式 (3-51b)。由于  $\eta_k \in (0, 1)$ ,  $|t_k| \geq \nu$  且  $\nu \geq 1/4$ , 因此有

$$-(\eta_k^2 - \eta_k) \frac{\sigma_k}{t_k} \leq \left| \frac{\sigma_k}{\nu} \right| \leq 4|\sigma_k|.$$

最终, 我们得出

$$f(x_{k+1}) - f(x_k) \leq \left( \eta_k - \frac{1}{2}\eta_k^2 \right) (\delta - \gamma_k) + 4|\sigma_k| - \frac{\gamma_k}{2} \Delta^2 + \frac{M}{6} \Delta^3.$$

□

#### 引理 3.29: 大值情形 (b)

假定假设 3.7 成立并设置  $\nu \in (1/4, 1/2)$ 。如果  $|t_k| \leq \nu$ , 令  $d_k = \text{sign}(-g_k^T v_k) \cdot v_k$  且  $\eta_k = \Delta/\|d_k\|$ , 则有

$$f(x_{k+1}) - f(x_k) \leq |\sigma_k| - \frac{\gamma_k}{2} \Delta^2 + \frac{M}{6} \Delta^3.$$

*Proof.* 从 (3-51a), 我们得到

$$\begin{aligned} v_k^T H_k v_k &= r_k^T v_k - \gamma_k \|v_k\|^2 - t_k g_k^T v_k, \\ g_k^T v_k &= \sigma_k + t_k \cdot (\delta - \gamma_k). \end{aligned}$$

因此, 有

$$\begin{aligned} f(x_{k+1}) - f(x_k) &= f(x_k + \eta_k \cdot d_k) - f(x_k) \\ &\leq \eta_k \cdot g_k^T d_k + \frac{\eta_k^2}{2} \cdot d_k^T H_k d_k + \frac{M\eta_k^3}{6} \cdot \|d_k\|^3 \\ &= \eta_k \cdot \text{sign}(-g_k^T v_k) g_k^T v_k + \frac{1}{2}\eta_k^2 (v_k)^T H_k v_k + \frac{M}{6}\eta_k^3 \|v_k\|^3 \\ &= -\eta_k \cdot |g_k^T v_k| + \frac{1}{2}\eta_k^2 r_k^T v_k - \frac{1}{2}\eta_k^2 t_k g_k^T v_k - \frac{1}{2}\eta_k^2 \gamma_k \|v_k\|^2 + \frac{M}{6}\eta_k^3 \|v_k\|^3 \end{aligned}$$

$$\begin{aligned} &\leq -\eta_k \cdot |g_k^T v_k| + \frac{1}{2}\eta_k^2 r_k^T v_k + \frac{1}{2}\eta_k^2 |t_k| |g_k^T v_k| - \frac{1}{2}\eta_k^2 \gamma_k \|v_k\|^2 + \frac{M}{6}\eta_k^3 \|v_k\|^3 \\ &= -\frac{1}{2}\eta_k^2 t_k \sigma_k - \left( \eta_k - \frac{1}{2}\eta_k^2 |t_k| \right) |g_k^T v_k| - \frac{\gamma_k}{2}\Delta^2 + \frac{M}{6}\Delta^3, \end{aligned}$$

其中最后一个等式利用了 (3-51b) 和  $\eta_k \|v_k\| = \eta_k \|d_k\| = \Delta$ 。由于  $\eta_k < 1$  且  $|t_k| \leq \nu < 1$ ，因此  $\eta_k^2 |t_k| \leq \eta_k < 1$ ，从而

$$f(x_{k+1}) - f(x_k) \leq |\sigma_k| - \frac{\gamma_k}{2}\Delta^2 + \frac{M}{6}\Delta^3.$$

□

上述两个引理说明了 Ritz 误差  $[r_k; \sigma_k]$  和不精确对偶变量  $\gamma_k$  如何妨碍下降性质的成立。为了确保非精确 HSODM 的收敛，Lanczos 方法需要保证  $\gamma_k \geq \delta$  并提供足够小的 Ritz 误差。然而，传统的随机初始 Lanczos 方法<sup>[103]</sup>无法满足这一需求。在下一小节中，我们提出了一种带有偏随机化的定制 Lanczos 方法以克服这一挑战，该方法本身可能也具有独立的意义。

我们以一个假定作为本小节的结尾，该假定在二阶算法分析中被广泛采用<sup>[27,147]</sup>。

#### 假设 3.30

假定存在一个与  $k$  无关的常数  $U_g > 0$ ，使得  $\|\nabla f(x_k)\| \leq U_g, \forall k \geq 1$ 。

由于非精确 HSODM 是单调的（这一性质将在 [定理 3.38](#) 中得到证明），上述假定在子水平集  $\{x : f(x) \leq f(x_1)\}$  是紧集的情况下很容易满足。根据 [引理 5.30](#)，这一假定表明  $U_H + U_g$  可以作为  $\|F_k\|$  的上界，这对于建立 [定理 3.31](#) 中定制 Lanczos 方法的性质是必要的。

### 3.5.3 带有偏随机化的定制 Lanczos 方法

在本小节中，我们开发了一种带有偏随机化的 Lanczos 方法，使其能够实现类似于精确 HSODM 的收敛行为。该方法的核心在于对初始向量  $q_1$  的偏随机化（[行 2 in 算法 3-4](#)）。基本思想是对  $q_1$  的最后一项赋予更大的权重。具体来说，我们首先从标准正态分布  $\mathcal{N}(0, 1)$  中独立采样  $b_i, i = 1, \dots, n+1$ ，然后将最后一项  $b_{n+1}$  乘以一个大常数  $\Psi_k$ 。令  $b = [b_1, \dots, b_n, \Psi_k \cdot b_{n+1}]^T$ ，并选择归一化后的向量  $q_1 := b/\|b\|$  作为 Lanczos 方法的初始向量。

**算法 3-4:** 带有偏随机化的 Lanczos 方法

**Input:** 迭代点  $x_k, g_k, H_k; \delta > 0, p \in (\exp(-n), 1), e_k > 0, J_{\max} \geq 0$

- 1 **初始化:** 从标准正态分布  $\mathcal{N}(0, 1)$  独立采样  $b_1, b_2, \dots, b_{n+1}$ ;
- 2 设置  $\Psi_k$  由 (3-61) 给出, 令  $b := [b_1, \dots, b_n, \Psi_k \cdot b_{n+1}]^T$ , 并令  $q_1 = b/\|b\|$ ;
- 3 构造  $F_k$ , 其中  $\chi_k$  为其精确的最小特征向量, 设
 
$$J_m = \min \left\{ J_{\max}, 1 + \sqrt{\frac{2\|F_k\|}{e_k}} \log \left( \frac{8}{e_k(q_1^T \chi_k)^2} \right) \right\};$$
- 4 **while**  $j = 1, \dots, J_m$  **do**
  - 5     计算  $F_k Q_j = Q_j T_j + \xi_j(1_j)_{[1:j]}^T$ ;
  - 6     **if**  $\|\xi_j\| \leq \epsilon$  **then**
    - 7         终止;
  - 8      $j \leftarrow j + 1$ ;
- 9 计算  $T_j$  的 Schur 分解, 使得  $S_j^T T_j S_j = \Gamma_j$ ;
- 10 计算 Ritz 近似  $(-\gamma_k, [v_k; t_k])$ ;
- 11 **return**  $(-\gamma_k, [v_k; t_k])$  和相应的 Ritz 误差  $[r_k; \sigma_k]$

为了便于理论分析, 由于  $\|q_1\| = 1$ , 可以将其重写为  $q_1 := \sqrt{1 - \alpha^2} \cdot [u; 0] + \alpha \cdot [0; 1] \in \mathbb{R}^{n+1}$ , 其中  $u \in \mathbb{R}^n$  且  $\|u\| = 1$ 。以下定理表明, 当  $|\alpha|$  超过某个阈值时, 不等式  $\gamma_k \geq \delta$  能够得到保证。令人惊讶的是, Ritz 误差最后一项的大小也可以通过  $|\alpha|$  加以控制。

**定理 3.31**

假定 [假设 3.7](#) 和 [假设 3.30](#) 成立。对于齐次矩阵  $F_k$ , 假定 Lanczos 方法以初始向量  $q_1 := \sqrt{1 - \alpha^2} \cdot [u; 0] + \alpha \cdot [0; 1] \in \mathbb{R}^{n+1}$  开始, 其中  $u \in \mathbb{R}^n$  且  $\|u\| = 1$ , 则对于任何  $|\alpha| \geq 1/2$ , 以下结论成立:

- (1) 在第  $j$  次迭代后 ( $j \geq 2$ ), Lanczos 向量  $q_j = [\ell_j; \beta_j] \in \mathbb{R}^n \times \mathbb{R}$  的最后一项满足  $|\beta_j| \leq 2\sqrt{1 - \alpha^2}$ 。
- (2) 在第  $j$  次迭代后 ( $j \geq 4$ ), Ritz 误差  $[r_k; \sigma_k]$  的最后一项满足

$$|\sigma_k| \leq U_\sigma \sqrt{1 - \alpha^2}, \quad (3-58)$$

其中  $U_\sigma$  是一个与  $k$  无关的常数:

$$U_\sigma := \sqrt{(U_H + U_g)^2 + (\delta + U_g)^2} \sqrt{U_g^2 + \delta^2} + 4\sqrt{n}(U_g + \max\{U_H, \delta\}) \quad (3-59)$$

(3) 如果满足

$$\alpha \cdot g_k^T u \leq 0 \quad \text{且} \quad |\alpha| \geq \frac{U_H + \delta}{\sqrt{(U_H + \delta)^2 + 4(g_k^T u)^2}}, \quad (3-60)$$

则不精确对偶变量  $\gamma_k$  足够大, 即  $\gamma_k \geq \delta$ 。

基于上述定理, 我们进一步展示如何通过适当选择  $\Psi_k$  使得 算法 3-4 达到预期目的。

#### 定理 3.32

假定 假设 3.7 和 假设 3.30 成立。在 算法 3-4 的偏随机初始化(行 2)中, 选择  $\Psi_k$  使得

$$\Psi_k = \frac{\sqrt{10n}}{\sqrt{\pi}p} \cdot \max \left\{ \frac{16M^2 U_\sigma}{\epsilon^2}, \sqrt{1 + \frac{(U_H + \delta)^2}{2p^2 \pi \|g_k\|^2}}, \frac{2}{\sqrt{3}} \right\}, \quad (3-61)$$

其中  $U_\sigma$  的定义如 (3-59) 所示。记  $F_k$  的精确最小特征向量为  $\chi_k = [\chi_{k,1}, \dots, \chi_{k,n+1}]$ , 则对于任何常数  $p \in (\exp(-n), 1)$  和  $\epsilon > 0$ , 以至少  $1 - p$  的概率满足以下不等式:

$$(q_1^T \chi_k)^2 \geq \min \left\{ \frac{\epsilon^4}{256M^4 U_\sigma^2}, \left( 1 + \frac{(U_H + \delta)^2}{2p^2 \pi \|g_k\|^2} \right)^{-1}, \frac{3}{4} \right\} \cdot \frac{\pi^2 p^4 \sum_{i=1}^n \chi_{k,i}^2}{100n(n+1)} + \frac{p^2 \pi \chi_{k,n+1}^2}{10(n+1)}, \quad (3-62)$$

$$|\sigma_k| \leq \frac{\epsilon^2}{16M^2} \quad \text{且} \quad |\alpha| \geq \frac{U_H + \delta}{\sqrt{(U_H + \delta)^2 + 4(g_k^T b_{[1:n]})^2}} \quad (3-63)$$

我们将上述两个定理的证明推迟到附录, 因为它们涉及较为复杂的技术细节。这些定理表明, 通过偏随机化可以以高概率保证  $\sigma_k$  足够小。由于正态分布的对称性, 通过改变  $b_{[1:n]}$  的符号, 可以确保  $\alpha \cdot g_k^T b_{[1:n]} \leq 0$ , 从而保证不精确对偶变量  $\gamma_k$  满足  $\gamma_k \geq \delta$ 。此外, 我们证明了  $(q_1^T \chi_k)^2$  能够远离 0, 通常达到 (3-62) 中的第一项 (即  $\Omega(\epsilon^4/n(n+1))$ ); 这为后续方法的复杂度分析提供了可能性。

#### Remark 3.33: 注

算法 3-4 可能依赖于  $\|F_k\|$  的先验信息。在技术上, 可以通过使用<sup>[148]</sup> Algorithm 5 中的边界估计对 算法 3-4 进行稍微改进, 在这种情况下,  $\|F_k\|$  可以通过某个  $\hat{F}_k$  估计满足

$$\|F_k\| \in [\hat{F}_k/2, \hat{F}_k], \quad (3-64)$$

在前  $O(\log(n))$  次迭代中以高概率成立 (见<sup>[148]</sup> Lemma 10)。因此可以消除对  $\|F_k\|$  先验信息的依赖 (行 3 in 算法 3-4), 代价是一次试运行。

在以下推论中, 我们证明通过偏随机初始化的定制 Lanczos 方法可以在大值情况下实现充分的下降。

**推论 3.34**

假定 [假设 3.7](#) 和 [假设 3.30](#) 成立。如果运行 [算法 3-4](#) 并设置参数  $e_k = \delta = \sqrt{\epsilon}$  和  $\Delta = \frac{\sqrt{\epsilon}}{M}$ 。那么对于任意  $\epsilon > 0$  和  $p \in (\exp(-n), 1)$ ，在两种大值情况下，以至少  $1 - p$  的概率成立：

$$f(x_{k+1}) - f(x_k) \leq -\frac{\delta}{4}\Delta^2 + \frac{M}{6}\Delta^3.$$

*Proof.* 根据 [定理 3.31](#) 和 [定理 3.32](#)，以至少  $1 - p$  的概率，有

$$|\sigma_k| \leq \frac{\epsilon^2}{16M^2} \leq \frac{\epsilon^{3/2}}{16M^2} = \frac{\delta}{16}\Delta^2 \quad \text{且} \quad \gamma_k \geq \delta$$

因此，大步情况 (a) ([引理 3.28](#)) 表明：

$$f(x_{k+1}) - f(x_k) \leq 4|\sigma_k| - \frac{\gamma_k}{2}\Delta^2 + \frac{M}{6}\Delta^3$$

结合大步情况 (b) ([引理 3.29](#))：

$$f(x_{k+1}) - f(x_k) \leq |\sigma_k| - \frac{\gamma_k}{2}\Delta^2 + \frac{M}{6}\Delta^3,$$

我们有：

$$\begin{aligned} f(x_{k+1}) - f(x_k) &\leq 4|\sigma_k| - \frac{\gamma_k}{2}\Delta^2 + \frac{M}{6}\Delta^3 \\ &\leq \frac{\delta}{4}\Delta^2 - \frac{\delta}{2}\Delta^2 + \frac{M}{6}\Delta^3 = -\frac{\delta}{4}\Delta^2 + \frac{M}{6}\Delta^3. \end{aligned}$$

此证。 □

### 3.5.4 小值情况下的非精确 HSODM

对于小值情况，与之前相同，当  $|t_k| \geq \nu$  且  $d_k = v_k/t_k$  时发生。在这种情况下，我们可以证明当前迭代点  $x_k$  处的 Hessian 矩阵近似正半定。基于此，[算法 3-3](#) 会测试 Ritz 误差  $r_k$  是否足够小。如果不满足条件，则增加扰动参数  $\delta$  并通过 [算法 3-4](#) 重新计算 Ritz 对。在这种情况下，我们可以证明齐次矩阵  $F_k$  的特征值间隔 (eigengap) 满足  $\Omega(\sqrt{\epsilon})$  的下界。以下引理总结了这些结果。

**引理 3.35: 小值情况**

假定 [假设 3.7](#) 和 [假设 3.30](#) 成立。如果  $|t_k| > \sqrt{1/(1+\Delta^2)}$  且运行 [算法 3-4](#)，参数设置为  $e_k = \delta = \sqrt{\epsilon}$  和  $\Delta = \frac{\sqrt{\epsilon}}{M}$ ，其中  $\epsilon \leq \min\{(2MU_g/(2U_H + U_g))^2, 3M^2, 1\}$ ，则以下结果成立：

(1) 对于任意  $p \in (\exp(-n), 1)$ ，有

$$\lambda_1(H_k) \geq -2\delta - 2\|g_k\|\Delta - (U_H + \gamma_k)\Delta^2 \geq -2\left(1 + \frac{2U_g}{M}\right)\sqrt{\epsilon}$$

的概率至少为  $1 - p$ 。

(2) 如果 Ritz 误差  $r_k$  满足  $\|r_k\| \leq 2\epsilon$  ([行 7](#))，则下一次迭代点  $x_{k+1} = x_k + d_k$  已经是  $\epsilon$ -近似

SOSP。

(3) 否则，当重新设置  $\delta = 3\sqrt{\epsilon} + 2\|g_k\|\Delta + (U_H + \gamma_k)\Delta^2$  (行 10) 时，齐次矩阵的特征值间隔满足  $\varsigma_k = \lambda_2(F_k) - \lambda_1(F_k) \geq \sqrt{\epsilon}$ 。

*Proof.* 根据 (3-51)，我们有

$$-\gamma_k = -\delta t_k^2 + 2t_k g_k^T v_k + v_k^T H_k v_k$$

对其重新整理得到

$$\begin{aligned} (\gamma_k - \delta)t_k^2 &= -2t_k g_k^T v_k - \left( \gamma_k + \frac{v_k^T H_k v_k}{\|v_k\|^2} \right) \|v_k\|^2 \\ &\leq 2t_k \sqrt{1 - t_k^2} \|g_k\| - \left( \gamma_k + \frac{v_k^T H_k v_k}{\|v_k\|^2} \right) \|v_k\|^2 \\ &\leq 2t_k \sqrt{1 - t_k^2} \|g_k\| - (\gamma_k + \lambda_1(H_k)) (1 - t_k^2), \end{aligned}$$

其中第一等式利用了  $\|v_k\|^2 + t_k^2 = 1$ 。这进一步意味着

$$\begin{aligned} \gamma_k - \delta &\leq 2\Delta \|g_k\| + |\lambda_1(H_k) + \gamma_k| \Delta^2 \\ &\leq 2\Delta \|g_k\| + (U_H + \gamma_k) \Delta^2, \end{aligned}$$

其中第一不等式由  $\Delta \geq \sqrt{1 - t_k^2}/t_k$  推出。

注意到  $H_k + \theta_k I \geq 0$  且  $\theta_k \leq \gamma_k + e_k = \gamma_k + \delta$ ，进一步得到

$$\begin{aligned} \lambda_1(H_k) + \theta_k &\geq 0 \\ \Rightarrow \lambda_1(H_k) + 2\delta + 2\|g_k\|\Delta + (U_H + \gamma_k)\Delta^2 &\geq 0 \end{aligned}$$

由于  $\delta = \sqrt{\epsilon}$ ， $\Delta = \sqrt{\epsilon}/M$ ， $\|g_k\| \leq U_g$  且  $\gamma_k \leq \|F_k\| \leq U_H + U_g$ ，我们得到

$$\begin{aligned} \lambda_1(H_k) &\geq -2\sqrt{\epsilon} - \frac{2\|g_k\|}{M} \sqrt{\epsilon} - \frac{(U_H + \gamma_k)\epsilon}{M^2} \\ &\geq -2\sqrt{\epsilon} - \frac{2U_g}{M} \sqrt{\epsilon} - \frac{(2U_H + U_g)\epsilon}{M^2} \\ &\geq -2 \left( 1 + \frac{2U_g}{M} \right) \sqrt{\epsilon}, \end{aligned}$$

其中最后一个不等式成立是因为  $\epsilon \leq (2MU_g/(2U_H + U_g))^2$ 。

对于  $\|r_k\| \leq 2\epsilon$  的情况，由于  $\|d_k\| = \|v_k/t_k\| \leq \Delta$ ，类似于 引理 3.16 的推导可以得到



$\lambda_1(H_{k+1}) \geq \Omega(-\sqrt{\epsilon})$ 。现在考察  $\|g_{k+1}\|$  的值。根据二阶 Lipschitz 连续性, 我们有

$$\begin{aligned}\|g_{k+1}\| &\leq \|g_{k+1} - g_k - H_k d_k\| + \|g_k + H_k d_k\| \\ &\leq \frac{M}{2} \|d_k\|^2 + \|g_k + H_k d_k\| \\ &= \frac{M}{2} \|d_k\|^2 + \|r_k/t_k - \gamma_k d_k\| \\ &\leq \frac{M}{2} \Delta^2 + \nu \|r_k\| + |\gamma_k| \Delta\end{aligned}$$

重置  $\delta := 3\sqrt{\epsilon} + 2\|g_k\|\Delta + (U_H + \gamma_k)\Delta^2$ , 并结合 Cauchy 间隔定理, 可得特征值间隔下界  $\varsigma_k = \lambda_2(F_k) - \lambda_1(F_k) \geq \sqrt{\epsilon}$ , 此证。  $\square$

**剩余问题是描述增加扰动参数的场景 (行 10)**。对于重新计算的 Ritz 对  $[v_k; t_k]$ , 如果它落入大值情况, 则函数值会减少, 并进入下一次迭代。关键在于, 如果  $[v_k; t_k]$  再次落入小值情况, 则必须有  $\|r_k\| \leq 2\epsilon$  成立, 这表明  $x_{k+1} = x_k + d_k$  是  $\epsilon$ -近似 SOSP。以下定理形式化地说明了这一点。

#### 引理 3.36

假定 [假设 3.7](#) 和 [假设 3.30](#) 成立, 并在 [算法 3-3](#) 的 [行 10](#) 中重新设置

$$\delta = 3\sqrt{\epsilon} + 2\|g_k\|\Delta + (U_H + \gamma_k)\Delta^2, \quad e_k = \min \left\{ \epsilon, \frac{\epsilon^{\frac{5}{2}}}{4(U_H + U_g)^2} \right\}, \quad \text{且} \quad \Delta = \frac{\sqrt{\epsilon}}{M}.$$

对于任意  $0 < \epsilon < 1$  和  $p \in (\exp(-n), 1)$ , 若  $|t_k| > \sqrt{1/(1 + \Delta^2)}$ , 则  $\|r_k\| \leq 2\epsilon$  的概率至少为  $1 - p$ 。

*Proof.* 注意到对于增大的  $\delta$ , 由 [引理 3.35](#) 可知  $\varsigma_k = \lambda_2(F_k) - \lambda_1(F_k) \geq \sqrt{\epsilon}$ 。根据 [引理 5.30](#), 有

$$\begin{aligned}\|r_k\| &\leq \tau_k e_k + 2(\max\{U_H, \delta\} + \|g_k\|) \sqrt{\frac{e_k}{\varsigma_k}} \\ &\leq \tau_k e_k + 2(U_H + U_g) \sqrt{\frac{e_k}{\sqrt{\epsilon}}} \\ &\leq 2\epsilon,\end{aligned}$$

其中最后一个不等式成立是因为  $\tau_k < 1$  (参见 [引理 5.30](#) 和 (5-69))。  $\square$

### 3.5.5 非精确 HSODM 的全局收敛性分析

最后, 我们分析 Lanczos 方法的复杂度。

**推论 3.37: 关于算法 3-4 的复杂度**

假定假设 3.7 和假设 3.30 成立。当算法 3-4 在非精确 HSODM 的行 3 中被调用时，对于任意常数  $p \in (\exp(-n), 1)$ ，其完成一次调用所需的迭代次数以至少  $1 - p$  的概率被以下表达式上界：

$$O \left( \sqrt{\|F_k\|} \epsilon^{-1/4} \log \left( \frac{n(n+1)}{p\epsilon} \right) \right).$$

*Proof.* 回顾定理 3.32 表明内积  $q_1^T \chi_k > 0$  的概率至少为  $1 - p$ ，这促进了在引理 3.26 中复杂度结果的应用。具体来说，我们知道  $(q_1^T \chi_k)^2$  远离零，并且通常取 (3-62) 中的第一项，其量级为  $\Omega(\epsilon^4/n(n+1))$ ，因为当  $\epsilon < 1$  很小时，(3-62) 中的第二项几乎为常量（类似于最后一项）。

注意，在非精确 HSODM 的某次迭代  $k$  中调用算法 3-4 时，只会发生以下两种情况之一。第一种情况，我们设置  $e_k = \sqrt{\epsilon}$ 。根据 (3-52)，最坏情况的复杂度为：

$$O \left( \sqrt{\|F_k\|} \epsilon^{-1/4} \log \left( \frac{n(n+1)}{p\epsilon} \right) \right).$$

第二种情况，我们将  $\delta$  设置为更大的值（见行 10），并根据引理 3.35，我们知道

$$\varsigma_k = \lambda_2(F_k) - \lambda_1(F_k) \geq \sqrt{\epsilon}.$$

因此，我们可以使用更高的精度，同时通过 (3-53) 保持复杂度在相同的量级。  $\square$   $\square$

综上所述，我们证明了在任何情况下，算法 3-3 中的 Lanczos 方法都保证在  $\tilde{O}(\epsilon^{-1/4})$  次迭代内终止。然而，与 [147, 148] 中提出的复杂度结果相比，这些结果依赖于  $\|H_k\|$  并可以由  $U_H$  限制，而我们的方法需要  $\|F_k\|$  的大小，其上界为  $U_H + U_g$ 。

在以下定理中，我们证明了非精确 HSODM 的算术复杂度。

**定理 3.38: 非精确 HSODM 的复杂度**

假定假设 3.7 和假设 3.30 成立。对于任意常数  $p \in (\exp(-n), 1)$ ，非精确 HSODM (算法 3-3) 在以下次数内终止：

$$K = 12(f(x_1) - f_{\inf})M^2\epsilon^{-3/2},$$

并返回满足以下条件的迭代点  $x_{k+1}$ ：

$$\|g_{k+1}\| \leq O(\epsilon) \quad \text{和} \quad \lambda_1(H_{k+1}) \geq \Omega(-\sqrt{\epsilon}),$$

其概率至少为  $(1 - p)^{2K}$ 。此外，算法 3-3 所需的算术操作数上界为：

$$O \left( (n+1)^2 \epsilon^{-7/4} (f(x_1) - f_{\inf}) M^2 \sqrt{U_H + U_g} \log \left( \frac{n(n+1)}{p\epsilon} \right) \right).$$

*Proof.* 对于 算法 3-3 中的两种大值情况, 推论 3.34 表明, 通过选择  $\delta = \sqrt{\epsilon}$  和  $\Delta = \frac{\sqrt{\epsilon}}{M}$ , 函数值的减少至少为:

$$f(x_{k+1}) - f(x_k) \leq -\frac{\delta}{4}\Delta^2 + \frac{M}{6}\Delta^3 = -\frac{\epsilon^{3/2}}{12M^2}.$$

根据 引理 3.35 和 引理 3.36, 在小值情况下, 算法将终止于一个  $\epsilon$ -近似的 SOSP, 或者回到大值情况。

因此, 我们可以得出, 在到达  $\epsilon$ -近似 SOSP 之前, 迭代次数最多为  $K = 12(f(x_1) - f_{\inf})M^2\epsilon^{-3/2}$ 。在每次迭代中, 如果是大值情况, 则需要调用一次 算法 3-4; 否则需要重置参数 (见 行 10)。在这种情况下, 我们可能进入大值情况并继续, 或者再次进入小值情况。而后一种情况表明  $\|r_k\| \leq 2\epsilon$ , 如 引理 3.36 所示, 这将终止算法。总之, 每次迭代最多需要两次 算法 3-4 的调用, 高概率下可以保证。

由于 Lanczos 方法成功的概率至少为  $1 - p$ , 在  $K$  次迭代中, 不会发生 算法 3-4 的错误终止, 其概率至少为  $(1 - p)^{2K}$ 。结合这些结果和 推论 3.37, 可以得出算术操作的复杂度。

我们首先注意到, 对于满足  $p < 1/2K$  的某些  $p$ ,  $(1-p)^{2K} \geq 1 - 2Kp$  成立。回忆  $p \in (\exp(-n), 1)$ , 当  $n \geq \Omega(-\log \epsilon)$  时, 这一条件很容易满足。例如, 当  $\epsilon = 10^{-8}$  时, 有  $n \approx 20$ 。因此, 定理中“至少  $(1 - p)^{2K}$  的概率”为条件可以替换为“至少  $1 - 2Kp$  的概率”, 同时保持其信息性。

由于我们的算法需要对维度为  $(n + 1)$  的齐次矩阵进行算术操作, 其在维度上的依赖性和与特征值过程相关的复杂度 (推论 3.37) 相较于早期的二阶算法 (如 [5,23,44,147,148]) 略逊一筹。在 Lipschitz 常数方面, 我们对 Hessian 的 Lipschitz 常数  $M$  的依赖性相比 [5,23] 中的结果较弱, 因为我们的算法没有显式地包含该常数; 相反, 它仅在建立整体计算复杂度时被调用。

尽管如此, 我们的算法 HSODM 以其简洁性和统一性为特征, 每次迭代仅需要特征值过程。具体而言, 在 Hessian 矩阵退化时, 它的计算效率优于牛顿类型方法。此外, 下一节还展示了 HSODM 的实际性能具有很大的潜力。□

## 第六节 数值实验

本节中, 我们展示了 HSODM 在几类非凸优化问题上的计算结果。我们包括了 CUTEst 问题 [72], 它被广泛用作评估非线性问题算法性能的标准数据集。由于 HSODM 属于二阶方法的范畴, 我们重点与牛顿信赖域方法和自适应三次正则牛顿方法 [27] 进行比较。我们的 Julia 实现 [15] 可在 <https://github.com/bzhangcw/DRSOM.jl> 获得。所有实验均在 macOS 系统的台式机上运行, 该系统配备 3.2 GHz 的 6 核 Intel Core i7 处理器。

### 3.6.1 实现细节

除了 HSODM 的原始形式 (见 [算法 3-1](#))，我们还加入了一些实际实现的技术。

首先，我们注意到实际的 HSODM 实现可能不会显式构造 Hessian 矩阵  $H_k$ 。在计算  $F_k \cdot [v; t]$  时，其中  $v \in \mathbb{R}^n, t \in \mathbb{R}$ ，我们有：

$$F_k \cdot \begin{bmatrix} v \\ t \end{bmatrix} = \begin{bmatrix} H_k \cdot v + t \cdot g_k \\ g_k^T v - t \cdot \delta \end{bmatrix}.$$

基于上述公式，我们提供了一种利用 Hessian 向量积  $H_k v$  的矩阵无关实现，与其他不精确牛顿类型方法类似<sup>[27,43]</sup>。

尽管理论分析采用了回溯线搜索算法，但在实际中，同类方法可以结合任何定义良好的线搜索算法。在我们的实现中，我们使用了 Hager-Zhang 线搜索算法，并采用默认参数设置<sup>[79]</sup>。对于特征值问题，我们使用 Lanczos 方法以设定的公差 ( $10^{-6}$ ) 求解齐次子问题。由于这些方法已由一些高效的 Julia 包提供，我们直接使用了 LineSearches.jl<sup>[97]</sup> 中的线搜索算法，以及 KrylovKit.jl<sup>[77]</sup> 中的 Lanczos 方法。对于超参数的设置，我们选取  $\delta = -\sqrt{\epsilon}$ ， $\nu = 0.01$ ，以及  $\Delta = 10^{-4}$ 。

**基准算法** Orban and Siqueira<sup>[132]</sup> 提供了由 JuliaSmoothOptimizers 组织开发的高效 Julia 包。这些包实现了基于 Steihaug-Toint 共轭梯度方法的牛顿信赖域方法 (Newton-TR-STCG) 和自适应三次正则化 (ARC)，并包含必要的子例程和技术，如子问题求解和 Krylov 方法。最新的数值结果已在<sup>[53]</sup>中报告。我们使用 Orban and Siqueira<sup>[132]</sup> 的原始实现及其默认设置。

### 3.6.2 CUTEst 数据集中的无约束问题

接下来，我们展示 CUTEst 数据集中选定子集的结果。为了全面比较，我们提供了 HSODM 使用明确的 Hessian 矩阵的结果，称为 HSODM，以及通过 Hessian-向量积辅助的版本 (HSODM-HVP)。对所有测试算法，我们设定迭代限制为 20,000，终止准则为  $\|\nabla f(x_k)\| \leq 10^{-5}$ ；如果未满足此条件，则视为失败。我们聚焦于变量数  $n \in [4, 5000]$  的无约束问题。

对于 CUTEst 中的每个问题，如果它具有不同的参数，我们选取所有符合条件的实例。最终，我们共有 200 个实例，其中部分实例无法被任何方法求解。完整结果可见 [表 3.B.2](#) 和 [表 3.B.3](#)。

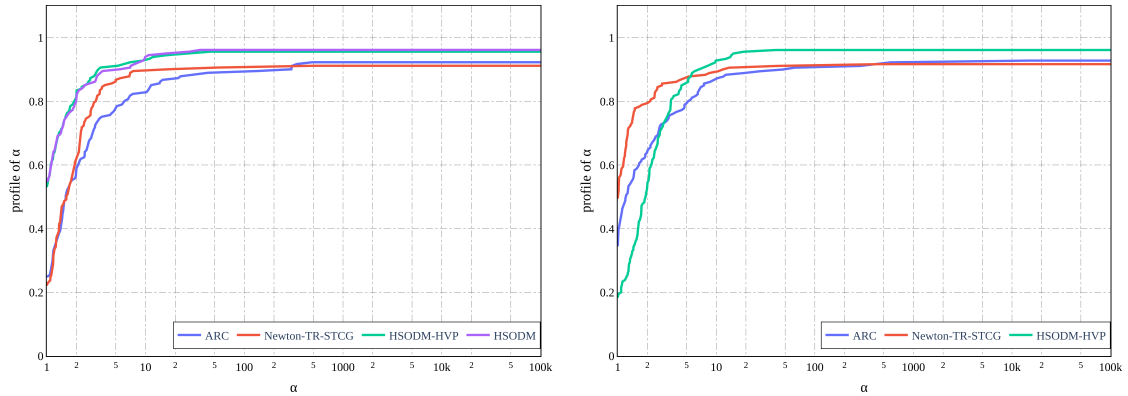
**算法的总体比较** 以下 [表 3-2](#) 总结了测试算法的表现。在该表中，我们用  $\mathcal{R}$  表示成功实例的数量。此外，我们基于缩放几何平均值 (SGM) 计算性能统计数据，包括  $\bar{t}_G, \bar{k}_G, \bar{k}_G^f, \bar{k}_G^g, \bar{k}_G^H$ ，分别表示平均运行时间、平均迭代次数、平均函数计算次数、平均梯度计算次数和平均 Hessian 计算次数。运行时间以 1 秒为单位缩放，其他指标按每 50 次计算或迭代缩放。

需要注意的是，三次正则化方法 ARC、牛顿信赖域方法 Newton-TR-STCG 和 HSODM-HVP 使用 Hessian-向量积，因此  $\bar{k}_G^H = 0$ ，而梯度计算次数  $\bar{k}_G^g$  实际上包括了 Hessian-向量积的次数。

表 3-2 不同算法在 CUTEst 数据集上的 SGM 性能。注意  $\bar{t}_G, \bar{k}_G$  是缩放后的几何平均值 (分别以 1 秒和 50 次迭代为单位缩放)。若某实例求解失败，其迭代次数和求解时间设为 20,000。

方法	$\mathcal{K}$	$\bar{t}_G$	$\bar{k}_G$	$\bar{k}_G^f$	$\bar{k}_G^g$	$\bar{k}_G^H$
Newton-TR-STCG	165.00	6.14	170.44	170.44	639.64	0.00
ARC	167.00	5.32	185.03	185.03	888.35	0.00
HSODM-HVP	173.00	4.79	111.24	200.60	787.32	0.00
HSODM	174.00	4.86	113.30	197.46	256.20	111.28

除了通过 SGM 衡量的指标，我们还使用文献<sup>[52]</sup>中定义的基于迭代次数的性能曲线。本质上，算法在图 3-1 中点  $\alpha$  的性能曲线表示在竞争者中最佳迭代次数的  $2^\alpha$  倍范围内成功求解的概率。



(a) 基于迭代次数的性能

(b) 基于梯度计算的性能

图 3-1 用于 CUTEst 问题的二阶方法性能曲线。(a) 报告了迭代次数的性能。(b) 包括了梯度计算的结果；仅包含使用 Krylov 子空间的方法。

这些初步实现的结果表明，HSODM 和 HSODM-HVP 在平均性能上优于标准的二阶方法，包括 Newton-TR-STCG 和 ARC。在竞争算法中，HSODM-HVP 和 HSODM 在  $\bar{k}_G, \bar{t}_G$  方面的迭代复杂度和运行时间表现更佳。HVP 变体 HSODM-HVP 在梯度评估次数上与 ARC 相当。由于 HSODM 需要更少的迭代，因此似乎需要更多的梯度评估。由于线搜索带来的额外开销，需要更多的函数评估。

有趣的是，在某些实例中，HSODM 和 HSODM-HVP (如 EXTROSNB)、Newton-TR-STCG (如 ARGGLINC) 以及 ARC (如 OSCIGRAD) 都表现最佳。

在性能曲线方面，可以看到 HSODM 和 HSODM-HVP 在迭代次数上具有优势。Newton-TR-STCG 在其成功的实例中梯度评估表现最佳。HSODM-HVP 由于使用了维度略大的  $n+1$  系统，因此需要更多的梯度评估。然而，这种劣势在实践中似乎是轻微的。

#### 第七节 结论

本文介绍了一种齐次的二阶下降方法 (Homogenized Second-Order Descent Method, HSODM), 其全局复杂度在一定范围广泛的二阶方法中是最优的 (见 <sup>[29]</sup>)。HSODM 使用了针对二阶泰勒展开的齐次技巧，从而将结果转化为一个可以通过求解特征值问题解决的齐次二次形式。我们证明了齐次的思想在凸优化和非凸优化情况下都是定义良好的，其中负曲率方向总是存在的。通过一直使用该模型，可以在小步长时安全地停下来，从而获得  $\epsilon$ -近似的二阶稳定点 (SOSP)，而无需切换到其他方法。

我们在 CUTEst 基准中的非线性优化问题上对 HSODM 进行了全面实验。HSODM 的两个变体在这些实验中显示了令人鼓舞的结果。未来的一个研究方向是将该方法用于约束优化问题。

## 本章附录

### 第一节 本章附加证明

#### 3.A.1 对引理 3.26 的证明

我们在此提供概要性证明，因为结果是<sup>[93]</sup>的复杂度估计和<sup>[147]</sup>的引理 9 的结合。考虑正半定矩阵  $F'_k := \|F_k\|I - F_k$ ，将  $\epsilon := \frac{e_k}{2\|F_k\|}$  代入<sup>[93]</sup>的复杂度结果，Lanczos 方法返回的估计  $\gamma_{\max}(F'_k)$  满足

$$\gamma_{\max}(F'_k) \geq \left(1 - \frac{e_k}{2\|F_k\|}\right) \lambda_{\max}(F'_k),$$

如果从向量  $q_1$  开始并最多运行

$$1 + 2\sqrt{\frac{\|F_k\|}{e_k}} \log\left(\frac{16\|F_k\|}{e_k(q_1^T \chi_k)^2}\right)$$

次迭代(无间隙版本)。由于  $\gamma_{\max}(F'_k) = \|F_k\| - \gamma_k$  且  $\lambda_{\max}(F'_k) = \|F_k\| - \lambda_1(F_k)$ ，根据<sup>[147]</sup>中引理 9 的相同论证，我们得出

$$\gamma_k \leq \lambda_1(F_k) + e_k.$$

间隙相关版本的结果可以类似地建立，因此我们在此省略。

#### 3.A.2 对引理 5.30 的证明

对于第一部分，注意到

$$\|F_k\| = \max_{\|[v;t]\|=1} \begin{bmatrix} v \\ t \end{bmatrix}^T \begin{bmatrix} H_k & g_k \\ g_k^T & -\delta \end{bmatrix} \begin{bmatrix} v \\ t \end{bmatrix},$$

其上界为

$$\|F_k\| \leq \max\{U_H, \delta\} + \|g_k\|,$$

完成了第一部分证明。对于第二部分，将 (3-51) 两边同时乘以  $[v_k; t_k]$ ，得到

$$[v_k; t_k]^T F_k [v_k; t_k] + \gamma_k = 0.$$

因为  $[v_k; t_k]$  是单位向量，我们可以写成  $[v_k; t_k] = \tau_k \cdot \chi_k + s$ ，其中  $\tau_k \in [0, 1]$ ，且  $s \perp \chi_k$  满足  $\tau_k^2 + \|s\|^2 = 1$ 。代入后得到

$$-\theta_k + e_k \geq -\gamma_k = -\theta_k \tau_k^2 + s^T F_k s \geq -\theta_k \tau_k^2 + (-\theta_k + \varsigma_k) \|s\|^2,$$

从而推导出

$$\|s\|^2 \leq \frac{e_k}{\varsigma_k}.$$

将其代入 (3-51), 我们有

$$[r_k; \sigma_k] = F_k[v_k; t_k] + \gamma_k[v_k; t_k],$$

其范数上界为

$$\|r_k\| \leq \tau_k e_k + 2(\max\{U_H, \delta\} + \|g_k\|) \sqrt{\frac{e_k}{\varsigma_k}}.$$

此证。

### 3.A.3 对 定理 3.31 的证明

对于第 (1) 部分, 基于 Lanczos 方法的机制, 对于任意正交基  $q_j = [\ell_j; \beta_j]$  且  $j \geq 2$ , 有  $q_j \perp q_1$ 。因此

$$\beta_j \alpha = -\ell_j^T u \sqrt{1 - \alpha^2},$$

推导出

$$|\beta_j| \leq \frac{\sqrt{1 - \alpha^2} \|\ell_j\| \|u\|}{|\alpha|} \leq 2\sqrt{1 - \alpha^2},$$

适用于  $|\alpha| \geq 1/2$ 。

对于第 (2) 部分, 记  $\zeta_{n+1} = F_k 1_{n+1}$  和  $y = \zeta_{n+1} - (\zeta_{n+1}^T q_1) \cdot q_1 - (\zeta_{n+1}^T q_2) \cdot q_2$ 。因此,

$$|\xi_{j,n+1}| \leq \sqrt{1 - \alpha^2} U_\sigma = O(\sqrt{1 - \alpha^2}).$$

利用 Ritz 逼近的性质, 可以推出

$$|\sigma_k| \leq |\xi_{j,n+1}| \leq \sqrt{1 - \alpha^2} U_\sigma.$$

对于第 (3) 部分, 由 Krylov 子空间的平移不变性, 可以证明  $q_1^T F_k q_1 \leq -\delta$ , 完成了证明。

## 第二节 CUTEst 数据集的详细结果

为简洁起见, 我们使用 表 3.B.1 中的缩写。

表 3.B.1 Abbreviations of the Methods

name	abbreviation
ARC	A
HSODM	H
HSODM-HVP	Hv
Newton-TR-STCG	N



表 3.B.2 CUTEst 数据集的完整结果，迭代和时间

name	n	k				$k^g$				t			
		A	H	Hv	N	A	H	Hv	N	A	H	Hv	N
ARGLINA	200	5	3	3	3	14	12	16	8	3.5e-03	9.4e-01	3.5e+00	1.8e-03
ARGLINB	200	3	4963	-	3	8	28756	-	8	1.6e-03	2.0e+02	0.0e+00	1.6e-03
ARGLINC	200	32	5049	5	3	897590	55465	52	8	2.0e+02	2.0e+02	1.2e+02	1.6e-03
ARGTRIGLS	200	7	13	13	7	964	57	4006	963	7.4e-01	7.8e-01	3.6e+00	6.5e-01
ARWHEAD	1000	7	6	6	7	25	24	41	25	1.2e-03	4.0e-03	3.7e-02	1.9e-03
	100	7	6	6	7	22	24	41	22	3.5e-03	6.1e-02	3.1e-02	4.1e-03
BDQRTIC	1000	12	11	11	12	110	50	185	110	1.6e-02	1.1e-02	1.1e-01	5.6e-03
	100	13	14	14	13	114	66	216	116	1.9e-02	1.5e-01	1.0e-01	1.9e-02
BOXPOWER	1000	4	5	5	4	14	18	34	14	7.0e-04	1.0e-03	3.4e-02	7.5e-04
	10	5	7	7	7	18	32	52	26	3.4e-03	9.2e-02	3.9e-02	4.9e-03
BOX	1000	18	9	9	18	86	41	74	86	2.3e-03	2.0e-03	5.1e-02	1.2e-03
	10	16	45	37	14	82	240	355	68	6.5e-03	4.3e-01	7.2e-01	6.8e-03
BROWNAL	1000	4	6	5	4	13	25	33	14	6.8e-03	2.8e-01	3.0e-02	3.3e-03
	200	4	4	4	5	12	16	24	15	5.7e-02	2.0e+01	1.4e-01	6.9e-02
BROYDN3DLS	1000	6	7	7	6	50	27	116	50	2.1e-03	3.0e-03	4.7e-02	1.6e-03
	50	7	10	10	6	55	45	204	46	5.6e-03	1.1e-01	7.5e-02	6.5e-03
BROYDN7D	500	102	15	15	36	551	77	477	227	2.1e-02	1.3e-02	1.0e-01	9.6e-03
	50	619	155	162	89	2857	843	8223	552	2.9e-01	7.0e-01	1.9e+00	1.6e-01
BROYDNBDLS	1000	16	10	10	17	196	43	209	193	5.9e-03	1.1e-02	6.3e-02	7.3e-03
	50	12	14	14	26	176	78	285	215	3.8e-02	1.9e-01	2.0e-01	4.4e-02
BRYEND	1000	16	10	10	17	196	43	209	193	1.4e-02	1.1e-02	6.2e-02	6.7e-03
	50	12	14	14	26	176	73	285	215	3.8e-02	2.0e-01	2.0e-01	5.5e-02

Continued on next page

表 3.B.2 CUTEst 数据集的完整结果，迭代和时间

name	n	k				$k^g$				t			
		A	H	Hv	N	A	H	Hv	N	A	H	Hv	N
CHAINW00	1000	97	40	40	85	425	203	432	435	2.1e-02	5.0e-03	1.9e-01	9.2e-03
	4	13192	585	676	2124	136952	2966	72418	22262	1.8e+01	7.0e+00	1.9e+01	1.8e+02
CHNROSNB	25	208	36	36	72	1406	182	948	623	4.0e-02	1.5e-02	2.0e-01	1.1e-02
CHNRSNB	25	232	39	39	88	1648	183	1092	801	5.3e-02	1.8e-02	2.2e-01	1.4e-02
COSINE	1000	9	9	12	10	38	39	197	43	2.5e-03	6.0e-03	7.1e-02	2.5e-03
	100	9	8	8	26	32	38	88	149	4.7e-03	8.2e-02	5.1e-02	3.0e-02
CRAGGLVY	1000	14	12	12	14	259	54	328	259	6.3e-03	4.0e-03	8.4e-02	9.6e-03
	50	15	15	15	14	208	73	390	176	4.5e-02	1.5e-01	1.8e-01	4.6e-02
CURLY10	1000	39	19	19	22	829	106	2660	646	2.0e-02	1.3e-01	4.0e-01	1.4e-02
	100	46	57	58	18	35411	298	44392	5459	2.9e+00	4.0e+01	9.4e+00	4.3e-01
CURLY20	1000	70	16	16	20	958	92	1445	503	2.8e-02	1.1e-01	2.8e-01	1.4e-02
	100	77	41	42	17	32093	211	37633	6745	4.2e+00	1.3e+02	1.0e+01	7.5e-01
CURLY30	1000	51	42	41	24	25457	214	38609	5737	4.5e+00	2.2e+02	1.3e+01	1.0e+00
DIXMAANA	3000	7	6	6	8	30	28	48	31	1.4e-03	5.0e-03	3.9e-02	2.2e-03
	90	8	7	7	8	33	35	60	31	1.5e-02	4.7e-01	5.2e-02	1.3e-02
DIXMAANB	3000	8	5	5	11	45	22	40	48	3.3e-03	4.0e-03	3.4e-02	3.1e-03
	90	9	6	6	9	32	28	51	33	1.5e-02	4.2e-01	4.5e-02	1.5e-02
DIXMAANC	3000	9	6	6	13	47	26	52	66	1.7e-03	5.0e-03	3.8e-02	4.0e-03
	90	9	6	6	10	35	28	49	38	1.6e-02	4.2e-01	4.5e-02	1.8e-02
DIXMAAND	3000	10	6	6	15	52	26	50	72	2.2e-03	6.0e-03	3.6e-02	4.5e-03
	90	9	6	6	11	36	29	50	41	1.7e-02	4.3e-01	4.8e-02	1.8e-02
DIXMAANE	3000	11	11	11	10	141	52	214	112	5.0e-03	1.2e-02	7.2e-02	2.6e-03

Continued on next page

表 3.B.2 CUTEst 数据集的完整结果，迭代和时间

name	n	k				$k^g$				t			
		A	H	Hv	N	A	H	Hv	N	A	H	Hv	N
	90	11	34	34	13	402	172	884	413	1.7e-01	2.7e+00	7.3e-01	1.5e-01
DIXMAANF	3000	13	9	9	19	157	41	170	167	4.7e-03	7.0e-03	6.1e-02	8.3e-03
	90	15	19	19	25	529	94	708	646	2.3e-01	1.7e+00	6.0e-01	3.1e-01
DIXMAANG	3000	14	9	9	17	172	41	159	163	1.8e-02	1.0e-02	5.8e-02	7.1e-03
	90	15	17	17	30	505	84	671	597	2.2e-01	1.5e+00	5.7e-01	2.5e-01
DIXMAANH	3000	14	9	9	25	153	41	166	228	5.6e-03	1.1e-02	5.9e-02	1.1e-02
	90	16	16	16	30	458	79	665	613	2.0e-01	1.4e+00	6.1e-01	2.3e-01
DIXMAANI	3000	12	22	23	14	400	108	1026	365	7.2e-03	3.8e-02	1.9e-01	1.2e-02
	90	12	85	143	13	5266	445	14134	5761	2.2e+00	1.3e+01	9.3e+00	2.1e+00
DIXMAANJ	3000	28	13	13	28	580	62	639	658	1.5e-02	2.9e-02	1.4e-01	1.9e-02
	90	54	32	62	36	6582	164	10653	5190	2.9e+00	1.3e+01	9.2e+00	2.1e+00
DIXMAANK	3000	25	13	13	25	528	62	687	518	2.2e-02	2.8e-02	1.4e-01	1.7e-02
	90	57	30	29	40	8987	154	3848	4719	3.9e+00	1.4e+01	9.3e+00	1.8e+00
DIXMAANL	3000	27	15	15	27	642	72	860	437	1.8e-02	3.0e-02	1.6e-01	1.5e-02
	90	87	29	40	64	9784	150	8847	10464	4.3e+00	1.5e+01	4.9e+00	4.9e+00
DIXMAANM	3000	9	30	30	12	341	149	1121	266	1.1e+00	4.0e-02	2.5e-01	6.9e-01
	90	11	108	297	13	12793	578	20134	5836	6.3e+00	1.3e+01	1.1e+01	2.8e+00
DIXMAANN	3000	15	22	22	18	601	108	839	710	1.5e-02	2.1e-02	1.9e-01	2.1e-02
	90	75	64	48	31	18601	332	2758	6956	8.1e+00	1.8e+01	8.5e+00	2.6e+00
DIXMAANO	3000	15	22	22	19	525	108	825	537	1.0e-02	3.8e-02	1.9e-01	1.5e-02
	90	79	59	171	28	18310	308	13239	6838	7.9e+00	2.0e+01	8.5e+00	3.1e+00
DIXMAANP	3000	18	23	23	25	620	113	922	735	1.4e-02	3.5e-02	2.3e-01	2.2e-02

Continued on next page

表 3.B.2 CUTEst 数据集的完整结果，迭代和时间

name	n	k				$k^g$				t			
		A	H	Hv	N	A	H	Hv	N	A	H	Hv	N
	90	79	65	91	33	18703	336	11944	5748	8.1e+00	2.8e+01	9.2e+00	2.6e+00
DIXON3DQ	1000	9	32	33	3	653	161	1479	166	7.5e-02	4.1e-02	2.4e-01	2.5e-03
	100	11	179	226	5	6407	963	28291	2036	2.9e-01	5.6e+00	4.1e+00	8.4e-02
DQDRTIC	1000	5	9	9	4	25	41	91	16	6.1e-03	4.0e-03	5.1e-02	9.8e-04
	50	5	8	8	7	23	36	79	29	3.1e-03	7.6e-02	4.3e-02	4.4e-03
DQRTIC	1000	24	9	9	25	200	51	116	173	4.9e-03	4.0e-03	5.2e-02	4.2e-03
	50	30	15	15	34	252	107	192	252	1.6e-02	1.3e-01	8.1e-02	1.8e-02
EDENSCH	2000	12	12	12	17	76	59	156	98	2.1e-03	5.0e-03	6.6e-02	3.6e-03
	36	13	11	11	15	71	58	149	82	2.1e-02	3.5e-01	9.4e-02	2.4e-02
EIGENALS	2550	10	10	7	9	32	77	50	31	1.2e-03	2.0e-03	5.6e-02	1.3e-03
	6	100	58	135	118	3914	986	12184	5545	2.5e+01	2.1e+02	2.0e+02	3.5e+01
EIGENBLS	2550	10	10	8	11	43	71	62	44	1.4e-03	2.0e-03	4.7e-02	1.6e-03
	6	2045	65	126	451	28271	355	27093	23072	2.0e+02	2.0e+02	2.0e+02	2.0e+02
EIGENCLS	2652	93	14	14	24	625	66	381	239	2.2e-02	1.3e-02	8.7e-02	9.2e-03
	30	2107	70	160	646	26158	701	25300	21199	2.0e+02	2.1e+02	2.0e+02	2.0e+02
ENGVAL1	1000	9	8	8	9	55	36	96	55	5.6e-03	5.0e-03	4.6e-02	2.2e-03
	50	-	8	8	10	-	34	99	53	-	8.4e-02	5.2e-02	9.0e-03
ERRINROS	25	117	81	82	82	880	437	1837	1081	3.8e-02	2.6e-02	3.9e-01	1.4e-02
ERRINRSM	25	314	282	-	202	2619	1514	-	3302	8.4e-02	7.4e-02	-	3.1e-02
EXTROSNB	1000	3901	9	8	4092	47796	85	165	72504	1.7e+00	8.0e-03	1.5e-01	1.9e+02
	100	3859	2346	274	790	47346	10621	4966	10901	3.9e+00	2.1e+01	1.9e+01	9.8e+02
FLETBV3M	1000	31	1	1	5	124	0	2	17	4.9e-03	0.0e+00	0.0e+00	9.5e-04

Continued on next page

表 3.B.2 CUTEst 数据集的完整结果，迭代和时间

name	n	k				$k^g$				t			
		A	H	Hv	N	A	H	Hv	N	A	H	Hv	N
	10	13	5	5	9	38	30	39	26	8.4e-03	5.3e-02	3.0e-02	6.7e-03
FLETGBV2	1000	4	4	4	2	23	16	30	9	1.1e-03	1.0e-03	2.9e-02	4.8e-04
	10	10	19	19	2	3514	96	4991	505	4.5e-01	1.7e+00	1.5e+00	6.4e-02
FLETGBV3	1000	341	1	1	7	1167	0	2	27	4.8e-02	0.0e+00	0.0e+00	1.3e-03
	10	20001	16946	14431	2985	60002	440053	418000	11921	1.1e+01	2.0e+02	2.0e+02	2.8e+02
FLETGBV	1000	10	114	114	7	83	792	1283	32	1.0e-02	1.6e-02	5.3e-01	1.2e-03
	10	20001	18174	16120	2581	80073	363037	386434	10306	2.8e+01	2.0e+02	2.0e+02	2.2e+02
FLETCHCR	1000	469	163	164	368	5366	776	4972	3942	4.2e-01	7.8e-02	9.8e-01	6.7e-02
	100	4669	1526	1526	761	54891	7349	50826	8458	5.8e+00	1.4e+01	1.6e+01	2.3e+02
FMINSRF2	16	17	17	17	21	156	82	248	111	6.0e-03	4.0e-03	9.3e-02	3.5e-03
	961	65	131	104	211	2354	666	6031	1096	5.6e-01	1.8e+00	2.2e+00	3.0e-01
FMINSURF	16	17	12	12	23	127	56	154	89	3.0e-03	3.0e-03	6.7e-02	3.4e-03
	961	71	71	74	197	1943	337	2176	1038	4.8e-01	1.2e+01	3.0e+00	2.4e-01
FREUROTH	1000	15	11	11	13	71	52	138	64	3.9e-03	6.0e-03	6.5e-02	3.0e-03
	50	15	15	15	10	73	73	191	55	1.2e-02	1.4e-01	1.1e-01	1.1e-02
GENHUMPS	1000	20001	229	331	3888	99259	1820	6401	15052	6.3e+00	2.2e-02	1.6e+00	2.5e+01
	10	20001	13759	8959	1386	108421	120667	557579	4838	2.8e+01	1.6e+02	2.0e+02	1.2e+03
GENROSE	100	844	76	74	175	6581	520	3553	1399	3.3e-01	5.9e-02	5.7e-01	3.4e-02
	500	3823	353	359	836	32215	2426	17279	6761	1.8e+00	1.2e+00	3.5e+00	4.2e-01
HILBERTA	6	6	11	9	3	31	52	89	12	1.2e-03	1.0e-03	5.0e-02	6.2e-04
HILBERTB	5	5	5	5	4	18	20	35	14	7.8e-04	1.0e-03	3.4e-02	7.2e-04
INDEFM	1000	20001	20000	20000	15925	81623	1039726	1102165	47857	6.5e+00	9.1e+00	1.1e+02	2.3e+02

Continued on next page

表 3.B.2 CUTEst 数据集的完整结果，迭代和时间

name	n	k				$k^g$				t			
		A	H	Hv	N	A	H	Hv	N	A	H	Hv	N
INDEF	50	20001	14992	11385	758	89617	776745	632944	2521	1.0e+01	2.0e+02	2.0e+02	5.3e+02
	1000	50	53	54	159	212	273	738	595	7.5e-03	1.6e-02	2.6e-01	2.0e-02
	50	39	63	88	194	194	351	935	823	2.9e-02	6.0e-01	9.5e-01	1.4e-01
INTEQNELS	102	4	5	5	4	16	19	38	16	2.8e-03	3.6e-02	2.7e-02	3.0e-03
	502	4	5	5	4	16	19	35	15	5.2e-02	3.0e+00	1.4e-01	4.8e-02
JIMACK	1521	6239	50	46	52	141701	488	37171	2969	2.0e+02	8.5e+00	4.5e+01	2.9e+00
	81	101	45	8	24	6236	612	7119	6784	2.0e+02	2.3e+02	2.3e+02	2.3e+02
LIARWHD	1000	11	12	12	11	42	58	99	42	1.6e-03	5.0e-03	6.4e-02	2.0e-03
	36	13	20	20	13	50	102	178	49	7.1e-03	2.0e-01	1.1e-01	8.6e-03
MANCINO	50	8	6	6	9	36	27	58	37	2.4e-02	2.5e-02	6.1e-02	2.4e-02
MODBEALE	10	11	11	11	12	94	47	149	86	3.4e-03	3.0e-03	6.3e-02	2.3e-03
MOREBV	2000	10	20	20	7754	132	100	390	236687	6.8e-02	6.8e-01	3.6e-01	3.0e+02
	1000	7	11	11	4	466	51	1125	217	5.2e-03	2.7e-02	1.4e-01	3.6e-03
MSQRTALS	50	3	3	3	3	2379	11	233	2446	2.1e-01	2.4e+00	1.3e+00	2.1e-01
	4900	37	12	12	23	446	55	442	373	3.3e-02	5.5e-02	9.9e-02	1.6e-02
MSQRTBLS	49	34	16	30	39	11619	81	10003	12017	2.4e+02	2.0e+02	2.0e+02	2.4e+02
	4900	26	13	13	27	454	59	576	446	1.4e-02	1.9e-02	1.1e-01	1.9e-02
NCB20B	49	32	16	35	40	11081	81	10587	10835	2.3e+02	2.0e+02	2.2e+02	2.2e+02
	1000	582	42	52	39	4143	262	2405	551	7.1e-01	1.6e-01	6.2e-01	5.8e-02
NCB20	180	1458	68	70	131	7965	425	3924	1062	7.6e+00	1.7e+00	5.0e+00	1.1e+00
	1010	4219	14	14	41	34429	187	6090	1783	1.1e+01	2.8e-01	1.5e+00	3.3e-01
	110	4199	14	15	48	35664	169	4117	2861	3.5e+01	1.3e+00	6.2e+00	2.8e+00

Continued on next page

表 3.B.2 CUTEst 数据集的完整结果，迭代和时间

name	n	k				$k^g$				t			
		A	H	Hv	N	A	H	Hv	N	A	H	Hv	N
NONCVXU2	1000	106	10	10	22	496	46	128	99	2.5e-02	1.0e-03	6.0e-02	3.5e-03
	10	3710	691	721	362	19119	3871	33069	3386	5.6e+00	1.2e+01	1.7e+01	1.1e+00
NONCVXUN	1000	67	9	9	20	293	41	103	80	2.8e-02	2.0e-03	5.8e-02	3.1e-03
	10	-	4852	2811	283	-	26262	282701	4668	-	2.0e+02	2.0e+02	1.4e+00
NONDIA	1000	11	8	8	14	35	33	61	49	2.6e-03	5.0e-03	5.0e-02	2.7e-03
	90	8	7	7	8	26	29	52	26	3.6e-03	7.5e-02	3.7e-02	4.2e-03
NONDQUAR	1000	71	52	57	22	5572	253	1508	643	7.1e-02	7.2e-02	3.7e-01	9.5e-03
	100	87	71	62	114	8516	350	1543	8063	4.4e-01	8.7e-01	4.2e-01	3.1e-01
NONMSQRT	4900	860	20000	5332	1141	18284	108076	114356	51594	8.4e-01	5.1e+00	9.8e+01	4.5e+00
	49	158	119	78	89	14331	653	14645	14859	2.0e+02	2.0e+02	2.1e+02	2.0e+02
OSCIGRAD	1000	13	332	12	19	133	2816	220	168	9.9e-03	4.4e-02	9.5e+01	3.5e-03
	15	14	-	12	19	160	-	548	183	2.4e-02	-	1.8e-01	3.2e-02
OSCIPATH	25	4	3	3	4	22	10	22	22	7.0e-04	2.0e-03	2.5e-02	8.2e-04
	500	4	3	3	4	22	10	22	22	2.5e-03	1.5e-02	2.2e-02	2.7e-03
PENALTY1	1000	65	50	47	58	196	308	479	193	9.8e-03	6.0e-02	2.3e-01	9.0e-03
	50	29	22	22	33	86	186	265	99	8.7e-03	3.3e+00	1.2e-01	1.1e-02
PENALTY2	1000	37	13	13	39	242	79	301	353	3.8e-02	1.4e-02	8.3e-02	1.4e-02
	50	0	240	20000	2529	3	1510	161371	7581	0.0e+00	3.1e+02	1.3e+02	3.0e+02
PENALTY3	50	63	25	25	26	286	129	496	144	3.5e-01	5.6e-02	3.3e-01	6.6e-02
	1000	19	13	13	19	110	64	133	110	1.2e-02	6.0e-03	7.3e-02	3.3e-03
POWELLSG	60	20	38	34	20	116	190	357	114	8.3e-03	3.5e-01	2.1e-01	1.0e-02
	1000	24	8	8	24	253	37	153	253	5.8e-03	9.0e-03	5.3e-02	5.0e-03

Continued on next page

表 3.B.2 CUTEst 数据集的完整结果，迭代和时间

name	n	k				$k^g$				t			
		A	H	Hv	N	A	H	Hv	N	A	H	Hv	N
QUARTC	50	30	14	14	30	647	71	723	653	3.4e-02	2.8e+00	1.7e-01	3.5e-02
	1000	26	10	10	28	225	59	155	219	6.3e-03	7.0e-03	5.9e-02	5.7e-03
	100	30	15	15	34	252	107	192	252	1.6e-02	1.3e-01	9.2e-02	1.6e-02
SBRYND	1000	20001	112	2005	251	372440	678	1139968	7075	1.4e+01	2.0e+02	9.5e+01	9.3e+01
	50	504	18	581	7130	984414	149	712211	810823	2.0e+02	2.0e+02	2.0e+02	2.0e+02
SCHNVET	1000	4	6	6	5	37	24	83	47	1.3e-03	1.0e-03	3.8e-02	1.4e-03
	10	5	6	6	5	77	25	116	73	3.5e-02	7.5e-02	1.1e-01	3.7e-02
SCOSINE	1000	42	-	-	4793	19355851	-	-	41338	2.0e+02	-	-	2.3e+02
	10	12824	249	56	530	2048247	3964	1662	54720	2.0e+02	2.1e+02	1.2e+02	4.6e+00
SCURLY10	1000	30	4	4	30	103	17	42	103	4.5e-03	1.0e-03	2.8e-02	3.9e-03
	10	30	9	9	31	510	53	232	553	5.2e-02	1.3e-01	7.0e-02	4.8e-02
SCURLY20	1000	30	8	8	31	378	46	180	389	5.8e-02	1.5e-01	6.6e-02	5.1e-02
	1000	-	9	9	31	-	48	199	350	-	2.1e-01	8.0e-02	6.6e-02
SENSORS	1000	40	13	13	18	217	60	174	65	9.7e-03	5.0e-03	7.3e-02	5.7e-03
	10	180	13	13	37	887	68	232	134	1.8e+02	1.0e+01	5.4e+01	2.7e+01
SINQUAD	1000	33	10	10	13	130	47	93	47	1.2e-02	3.0e-03	5.5e-02	3.2e-03
	50	49	20	20	18	195	106	197	65	5.2e-02	1.8e-01	2.6e-01	1.8e-02
SPARSINE	1000	7	9	9	32	183	39	488	433	3.4e-03	1.6e-02	7.6e-02	1.2e-02
	50	11	15	19	32	9666	71	11762	5875	1.7e+00	1.7e+01	3.9e+00	1.2e+00
SPARSQR	1000	18	5	5	18	156	22	78	156	3.8e-03	4.0e-03	3.4e-02	5.6e-03
	50	22	7	7	22	216	34	91	217	3.3e-02	1.6e-01	5.1e-02	3.7e-02
SPMSRTLS	1000	21	10	10	22	239	46	377	211	8.2e-03	1.5e-02	7.8e-02	1.1e-02

Continued on next page



表 3.B.2 CUTEst 数据集的完整结果，迭代和时间

name	n	k				$k^g$				t			
		A	H	Hv	N	A	H	Hv	N	A	H	Hv	N
SROSENBR	100	20	15	15	21	360	74	503	295	6.5e-02	2.0e-01	2.3e-01	6.4e-02
	500	10	8	8	13	33	32	57	45	1.5e-03	4.0e-03	4.7e-02	2.1e-03
	50	10	8	8	11	35	34	61	38	2.8e-03	3.2e-02	4.5e-02	3.7e-03
SSBRYBND	1000	306	120	1	117	29656	1005	2	2814	5.8e-01	2.0e+02	1.0e+02	4.5e-02
	50	153	30	203	79	292466	356	194973	746	5.8e+01	2.1e+02	2.0e+02	3.9e+02
SSCOSINE	1000	20001	20000	20000	51	174630	153859	246624	325	5.9e+00	9.7e-01	8.9e+01	6.9e-03
	10	7524	73	-	550	2025778	1016	-	20787	2.0e+02	2.0e+02	-	2.3e+00
TESTQUAD	1000	7	22	22	8	1053	109	3377	1049	5.4e-02	3.3e+01	8.0e-01	5.3e-02
TOINTGSS	1000	30	6	6	10	156	23	76	59	7.2e-03	6.0e-03	4.2e-02	3.1e-03
	50	7	5	5	9	33	22	43	40	8.0e-03	5.0e-02	3.7e-02	1.3e-02
TQUARTIC	1000	22	14	14	2	71	62	113	6	1.0e-02	5.0e-03	7.5e-02	3.0e-04
	50	12	44	42	8	42	219	369	26	5.0e-03	4.1e-01	2.3e-01	4.1e-03
TRIDIA	1000	5	10	10	5	128	46	303	116	1.2e-03	8.0e-03	7.1e-02	1.6e-03
	50	6	21	21	7	734	100	1865	738	3.5e-02	2.7e-01	3.5e-01	3.1e-02
VARDIM	200	28	5	5	28	83	23	39	83	5.8e-03	3.1e-02	3.1e-02	5.6e-03
VAREIGVL	100	13	9	12	24	116	40	508	1351	3.8e-03	1.4e-02	1.3e-01	3.5e-02
	500	22	11	12	27	2130	51	646	2536	1.9e-01	1.0e-01	2.4e-01	2.0e-01
WATSON	12	13	103	14	13	99	571	206	98	3.1e-03	2.6e-02	5.0e-01	5.1e-03
	4000	97	40	40	85	425	203	432	435	1.3e-02	6.0e-03	1.9e-01	9.4e-03
WOODS	4	99	29	29	62	459	154	317	321	1.6e-01	3.5e+00	2.8e-01	9.5e-02
	120	20001	36	36	247	53998	188	328	831	1.2e+01	4.1e-02	1.9e-01	7.4e-02
YATP1LS	2600	88	33	33	52	278	168	292	188	4.2e-01	6.4e+00	1.1e+00	2.8e-01

Continued on next page

表 3.B.2 CUTEst 数据集的完整结果，迭代和时间

name	n	k			k <sup>g</sup>			t		
		A	H	Hv	N	A	H	Hv	N	Hv
YATP2LS	8	38	8	8	56	124	40	65	505	1.1e-01 2.1e+00 1.2e-01 4.2e-01
	2600	301	7	7	386	1508	33	54	2259	1.3e-01 2.0e-03 4.4e-02 8.8e+01

### 第三章 齐次二阶下降法

表 3.B.3 CUTEst 数据集的完整结果，函数值和梯度范数

name	n	$f$				$\ g\ $			
		A	H	Hv	N	A	H	Hv	N
ARGLINA	200	+1.2e-22	+7.0e-28	+6.7e-29	+2.8e-26	2.2e-11	3.7e-13	3.1e-14	3.4e-13
ARGLINB	200	+5.0e+01	+5.0e+01	+0.0e+00	+5.0e+01	1.7e-03	1.4e+01	0.0e+00	2.0e-03
ARGLINC	200	+5.1e+01	+5.1e+01	+5.1e+01	+5.1e+01	1.2e+01	5.0e+01	1.8e-01	5.0e-04
ARGTRIGLS	200	+2.1e-19	+7.1e-23	+6.9e-17	+2.2e-19	1.4e-08	2.4e-08	3.3e-06	1.5e-08
ARWHEAD	1000	+0.0e+00	+0.0e+00	+0.0e+00	+0.0e+00	1.2e-13	5.1e-09	5.1e-09	1.2e-13
	100	+0.0e+00	+0.0e+00	+0.0e+00	+0.0e+00	1.1e-11	1.4e-07	1.4e-07	1.1e-11
BDQRTIC	1000	+3.8e+02	+3.8e+02	+3.8e+02	+3.8e+02	1.1e-08	1.2e-10	4.7e-07	1.1e-08
	100	+4.0e+03	+4.0e+03	+4.0e+03	+4.0e+03	1.4e-08	2.9e-06	3.0e-06	1.4e-08
BOXPOWER	1000	-1.7e-01	-1.7e-01	-1.7e-01	-1.7e-01	6.6e-13	4.3e-10	4.3e-10	1.5e-12
	10	-1.8e+02	-1.8e+02	-1.8e+02	-1.8e+02	9.3e-10	5.7e-06	5.7e-06	1.6e-08
BOX	1000	+8.0e-09	+5.2e-16	+5.2e-16	+8.6e-09	4.0e-07	1.5e-06	1.5e-06	4.4e-07
	10	+1.4e-08	+7.2e-10	+2.8e-09	+1.6e-08	4.8e-07	8.1e-07	2.9e-06	8.0e-07
BROWNAL	1000	+6.4e-13	+2.0e-22	+4.2e-23	+2.2e-19	1.4e-07	5.0e-07	2.8e-06	1.5e-09
	200	+2.3e-12	+3.2e-21	+2.3e-12	+2.3e-12	6.9e-07	2.6e-08	6.8e-07	5.6e-07
BROYDN3DLS	1000	+4.9e-15	+4.6e-29	+7.5e-17	+4.7e-15	6.5e-07	1.1e-10	5.8e-07	6.4e-07
	50	+6.3e-19	+7.1e-01	+7.1e-01	+3.9e-15	7.9e-09	7.0e-07	1.1e-06	6.1e-07
BROYDN7D	500	+1.7e+01	+1.7e+01	+1.7e+01	+1.7e+01	7.2e-08	4.2e-06	4.3e-06	1.4e-08
	50	+1.8e+02	+2.8e+00	+2.8e+00	+1.9e+02	8.0e-09	3.0e-06	1.5e-06	2.4e-08
BROYDNBDLS	1000	+1.5e-17	+7.7e-19	+4.6e-15	+1.2e-17	7.7e-09	1.4e-06	1.5e-06	1.5e-08
	50	+2.2e-14	+1.1e-22	+6.6e-17	+6.2e-18	8.9e-07	2.4e-10	4.9e-08	1.3e-08
BRYBND	1000	+1.5e-17	+7.7e-19	+4.6e-15	+1.2e-17	7.7e-09	1.4e-06	1.5e-06	1.5e-08
	50	+2.2e-14	+8.4e-23	+6.6e-17	+6.2e-18	8.9e-07	2.4e-10	4.9e-08	1.3e-08
CHAINWOOD	1000	+1.0e+00	+1.0e+00	+1.0e+00	+1.0e+00	1.3e-08	9.9e-06	9.9e-06	5.0e-09
	4	+8.5e+02	+4.2e+02	+3.8e+02	+7.9e+02	2.7e-08	4.4e-10	1.4e-06	1.7e-07
CHNROSNB	25	+1.0e-18	+9.1e-29	+3.2e-16	+6.5e-19	1.8e-08	6.4e-08	3.9e-07	9.5e-09
CHNRSNBM	25	+2.5e-19	+1.3e-27	+2.8e-15	+2.7e-17	1.0e-08	8.2e-07	1.1e-06	5.7e-08
COSINE	1000	-9.9e+01	-9.9e+01	-9.9e+01	-9.9e+01	1.4e-07	2.2e-09	4.4e-06	7.7e-09
	100	-1.0e+03	-1.0e+03	-1.0e+03	-1.0e+03	7.8e-07	1.9e-10	2.5e-07	4.0e-09
CRAGGLVY	1000	+1.5e+01	+1.5e+01	+1.5e+01	+1.5e+01	4.0e-08	2.5e-06	2.7e-06	4.2e-08
	50	+3.4e+02	+3.4e+02	+3.4e+02	+3.4e+02	1.4e-08	2.4e-09	9.2e-07	9.0e-07
CURLY10	1000	-1.0e+04	-1.0e+04	-1.0e+04	-1.0e+04	3.9e-07	1.7e-08	9.1e-07	1.7e-08
	100	-1.0e+05	-1.0e+05	-1.0e+05	-1.0e+05	9.4e-07	2.4e-07	9.3e-06	6.4e-04
CURLY20	1000	-1.0e+04	-1.0e+04	-1.0e+04	-1.0e+04	1.5e-08	1.6e-08	9.5e-07	1.3e-08

Continued on next page

### 第三章 齐次二阶下降法

表 3.B.3 CUTEst 数据集的完整结果，函数值和梯度范数

name	$n$	$f$				$\ g\ $			
		A	H	Hv	N	A	H	Hv	N
	100	-1.0e+05	-1.0e+05	-1.0e+05	-1.0e+05	1.4e-08	2.5e-09	4.1e-06	1.9e-05
CURLY30	1000	-1.0e+05	-1.0e+05	-1.0e+05	-1.0e+05	1.5e-08	1.6e-03	7.5e-06	1.6e-04
DIXMAANA	3000	+1.0e+00	+1.0e+00	+1.0e+00	+1.0e+00	8.7e-12	1.6e-06	1.6e-06	5.0e-17
	90	+1.0e+00	+1.0e+00	+1.0e+00	+1.0e+00	8.2e-17	3.6e-08	3.6e-08	6.3e-09
DIXMAANB	3000	+1.0e+00	+1.0e+00	+1.0e+00	+1.0e+00	8.5e-08	3.4e-06	3.4e-06	2.7e-09
	90	+1.0e+00	+1.0e+00	+1.0e+00	+1.0e+00	2.6e-09	6.9e-10	6.9e-10	6.1e-09
DIXMAANC	3000	+1.0e+00	+1.0e+00	+1.0e+00	+1.0e+00	7.7e-14	6.7e-10	6.7e-10	4.2e-17
	90	+1.0e+00	+1.0e+00	+1.0e+00	+1.0e+00	1.3e-14	3.4e-08	3.4e-08	1.2e-19
DIXMAAND	3000	+1.0e+00	+1.0e+00	+1.0e+00	+1.0e+00	5.8e-12	3.0e-08	3.0e-08	2.6e-12
	90	+1.0e+00	+1.0e+00	+1.0e+00	+1.0e+00	4.7e-08	8.3e-06	8.3e-06	1.4e-09
DIXMAANE	3000	+1.0e+00	+1.0e+00	+1.0e+00	+1.0e+00	1.3e-08	1.3e-08	9.8e-07	5.3e-07
	90	+1.0e+00	+1.0e+00	+1.0e+00	+1.0e+00	2.6e-07	3.0e-06	2.9e-06	3.5e-07
DIXMAANF	3000	+1.0e+00	+1.0e+00	+1.0e+00	+1.0e+00	1.3e-08	4.1e-08	8.1e-07	9.4e-09
	90	+1.0e+00	+1.0e+00	+1.0e+00	+1.0e+00	1.5e-08	6.1e-06	6.0e-06	1.6e-08
DIXMAANG	3000	+1.0e+00	+1.0e+00	+1.0e+00	+1.0e+00	1.1e-08	5.6e-08	9.3e-07	2.5e-07
	90	+1.0e+00	+1.0e+00	+1.0e+00	+1.0e+00	7.4e-08	5.5e-06	5.4e-06	8.0e-08
DIXMAANH	3000	+1.0e+00	+1.0e+00	+1.0e+00	+1.0e+00	3.5e-08	1.2e-07	8.7e-07	1.9e-08
	90	+1.0e+00	+1.0e+00	+1.0e+00	+1.0e+00	1.5e-07	1.3e-06	1.3e-06	2.2e-08
DIXMAANI	3000	+1.0e+00	+1.0e+00	+1.0e+00	+1.0e+00	3.6e-09	4.8e-06	4.3e-06	1.8e-09
	90	+1.0e+00	+1.0e+00	+1.0e+00	+1.0e+00	6.7e-07	7.2e-06	9.4e-06	1.3e-08
DIXMAANJ	3000	+1.0e+00	+1.0e+00	+1.0e+00	+1.0e+00	5.8e-08	1.8e-06	1.8e-06	1.3e-08
	90	+1.0e+00	+1.0e+00	+1.0e+00	+1.0e+00	8.6e-07	8.1e-06	9.4e-06	3.3e-07
DIXMAANK	3000	+1.0e+00	+1.0e+00	+1.0e+00	+1.0e+00	1.3e-07	5.7e-07	5.9e-07	1.4e-08
	90	+1.0e+00	+1.0e+00	+1.0e+00	+1.0e+00	4.2e-07	7.9e-06	9.9e-06	7.9e-08
DIXMAANL	3000	+1.0e+00	+1.0e+00	+1.0e+00	+1.0e+00	1.5e-08	7.0e-06	7.3e-06	7.8e-08
	90	+1.0e+00	+1.0e+00	+1.0e+00	+1.0e+00	5.8e-07	9.2e-06	7.8e-06	6.8e-07
DIXMAANM	3000	+1.0e+00	+1.0e+00	+1.0e+00	+1.0e+00	1.5e-08	7.0e-07	7.7e-07	1.3e-08
	90	+1.0e+00	+1.0e+00	+1.0e+00	+1.0e+00	1.6e-07	8.4e-06	9.9e-06	5.0e-08
DIXMAANN	3000	+1.0e+00	+1.0e+00	+1.0e+00	+1.0e+00	1.5e-08	4.6e-06	4.6e-06	3.7e-08
	90	+1.0e+00	+1.0e+00	+1.0e+00	+1.0e+00	7.7e-07	7.9e-06	1.0e-05	3.9e-07
DIXMAANO	3000	+1.0e+00	+1.0e+00	+1.0e+00	+1.0e+00	8.9e-08	4.4e-06	4.4e-06	2.2e-08
	90	+1.0e+00	+1.0e+00	+1.0e+00	+1.0e+00	8.7e-07	6.3e-06	8.8e-06	1.5e-08
DIXMAANP	3000	+1.0e+00	+1.0e+00	+1.0e+00	+1.0e+00	1.3e-08	1.4e-06	2.7e-06	1.9e-08

Continued on next page

### 第三章 齐次二阶下降法

表 3.B.3 CUTEst 数据集的完整结果，函数值和梯度范数

name	$n$	$f$				$\ g\ $			
		A	H	Hv	N	A	H	Hv	N
DIXON3DQ	90	+1.0e+00	+1.0e+00	+1.0e+00	+1.0e+00	3.0e-07	8.1e-06	9.5e-06	4.3e-07
	1000	+1.9e-16	+3.0e-14	+7.2e-13	+2.4e-27	1.4e-08	3.4e-06	4.3e-06	1.9e-13
	100	+3.2e-08	+5.7e-10	+6.1e-12	+7.8e-17	5.7e-07	7.6e-06	2.8e-06	8.3e-09
DQDRTIC	1000	+1.2e-23	+2.7e-29	+9.4e-18	+7.1e-30	7.0e-12	8.4e-08	8.4e-08	5.4e-15
	50	+5.2e-15	+1.8e-30	+2.6e-41	+1.7e-42	1.4e-07	3.2e-10	3.2e-10	6.6e-21
	1000	+3.4e-09	+1.8e-09	+1.8e-09	+3.8e-09	8.5e-07	3.5e-06	3.5e-06	8.3e-07
DQRTIC	50	+7.5e-07	+1.4e-10	+1.4e-10	+5.1e-07	2.3e-05	1.1e-06	1.2e-06	1.7e-05
	2000	+2.2e+02	+2.2e+02	+2.2e+02	+2.2e+02	9.0e-07	1.1e-07	5.3e-07	1.6e-08
	36	+1.2e+04	+1.2e+04	+1.2e+04	+1.2e+04	7.8e-09	1.9e-06	1.9e-06	7.8e-09
EIGENALS	2550	+1.1e-16	+7.5e-22	+7.5e-22	+5.2e-22	2.4e-08	2.4e-06	2.4e-06	1.2e-10
	6	+3.9e-14	+7.4e+01	+7.3e+01	+8.2e-11	5.1e-07	1.6e+02	9.3e+01	8.8e-07
	2550	+1.8e-01	+9.6e-24	+1.8e-01	+1.8e-01	7.2e-08	2.2e-06	8.2e-07	3.3e-07
EIGENBLS	6	+1.5e-02	+1.5e-02	+4.1e-03	+5.4e-04	1.6e-03	1.2e-01	4.1e-02	4.6e-03
	2652	+3.8e-17	+2.7e-23	+9.7e-14	+5.0e-17	1.1e-08	4.6e-06	4.8e-06	1.4e-08
	30	+1.3e+03	+2.9e+03	+8.6e+02	+4.2e-03	1.4e+00	1.6e+01	8.5e+00	1.3e-01
ENGVAL1	1000	+5.4e+01	+5.4e+01	+5.4e+01	+5.4e+01	8.9e-09	5.7e-06	5.7e-06	8.7e-09
	50	-	+1.1e+03	+1.1e+03	+1.1e+03	-	1.7e-11	3.8e-07	1.4e-08
	25	+1.8e+01	+1.8e+01	+1.8e+01	+1.8e+01	1.3e-08	7.4e-08	3.8e-06	7.2e-07
ERRINROS	25	+1.8e+01	+1.8e+01	-	+1.8e+01	7.6e-10	8.0e-06	-	5.5e-08
	1000	+3.1e-08	+7.1e-28	+1.8e-18	+3.3e-09	8.8e-07	3.6e-09	7.4e-08	1.6e-08
	100	+3.3e-08	+3.0e-09	+2.4e-09	+5.1e-07	9.7e-07	9.4e-06	1.0e-05	7.4e-04
FLETBV3M	1000	-2.2e-03	+1.2e-05	+1.2e-05	-2.2e-03	8.8e-07	7.7e-06	7.7e-06	4.9e-07
	10	-2.0e+03	-2.0e+03	-2.0e+03	-2.0e+03	3.6e-12	4.4e-11	6.6e-07	1.6e-08
	1000	-5.5e-01	-5.5e-01	-5.5e-01	-5.5e-01	1.4e-08	5.7e-07	5.7e-07	1.0e-08
FLETBVB2	10	-5.0e-01	-5.0e-01	-5.0e-01	-5.0e-01	1.5e-08	2.8e-06	3.0e-06	2.1e-09
	1000	-3.2e-02	+1.2e-05	+1.2e-05	-3.2e-02	9.9e-07	7.7e-06	7.7e-06	5.6e-09
	10	-8.0e+07	-2.9e+11	-2.5e+11	-1.0e+11	6.3e-01	7.5e-01	7.6e-01	9.4e-01
FLETCHBV	1000	-2.7e+06	-2.7e+06	-2.7e+06	-2.7e+06	2.9e-08	1.3e-07	1.3e-07	0.0e+00
	10	-5.6e+19	-3.0e+19	-2.7e+19	-9.2e+18	4.7e+07	7.3e+07	7.4e+07	8.8e+07
	1000	+4.5e-19	+3.7e-27	+7.1e-16	+1.6e-19	2.2e-08	1.2e-06	3.1e-06	1.5e-08
FLETCHCR	100	+2.8e-19	+1.4e-22	+6.5e-18	+7.6e+02	1.9e-08	5.3e-06	6.9e-07	4.6e+00
	16	+1.0e+00	+1.0e+00	+1.0e+00	+1.0e+00	1.6e-09	6.1e-06	6.1e-06	1.4e-08
	961	+1.0e+00	+1.0e+00	+1.0e+00	+1.0e+00	2.0e-08	5.5e-06	8.3e-06	1.6e-08

Continued on next page

### 第三章 齐次二阶下降法

表 3.B.3 CUTest 数据集的完整结果，函数值和梯度范数

name	$n$	$f$				$\ g\ $			
		A	H	Hv	N	A	H	Hv	N
FMINSURF	16	+1.0e+00	+1.0e+00	+1.0e+00	+1.0e+00	8.4e-09	2.8e-07	2.8e-07	2.2e-09
	961	+1.0e+00	+1.0e+00	+1.0e+00	+1.0e+00	1.1e-07	6.1e-07	9.6e-07	2.7e-08
FREUROTH	1000	+5.9e+03	+5.9e+03	+5.9e+03	+5.9e+03	1.3e-08	3.3e-07	3.3e-07	1.4e-08
	50	+1.2e+05	+1.2e+05	+1.2e+05	+1.2e+05	9.3e-07	3.2e-10	4.8e-07	2.8e-08
GENHUMPS	1000	+2.9e+04	+8.6e-32	+7.5e-20	+1.3e-17	8.5e+01	2.2e-08	7.9e-07	2.1e-09
	10	+1.0e+07	+4.2e-17	+5.7e+02	+1.2e+05	1.9e+03	7.2e-06	2.9e+01	5.0e+02
GENROSE	100	+1.0e+00	+1.0e+00	+1.0e+00	+1.0e+00	1.8e-08	3.9e-07	6.9e-07	9.3e-09
	500	+1.0e+00	+1.0e+00	+1.0e+00	+1.0e+00	1.7e-08	7.8e-06	5.5e-07	8.9e-09
HILBERTA	6	+2.3e-11	+3.1e-13	+4.6e-11	+5.0e-09	2.9e-08	2.8e-07	9.9e-06	3.6e-07
HILBERTB	5	+3.4e-19	+1.2e-31	+3.1e-23	+8.9e-20	2.8e-09	8.5e-09	3.7e-07	1.3e-09
INDEFM	1000	-1.0e+09	-2.9e+05	-9.2e+14	-2.2e+05	7.1e+00	7.1e+00	7.3e+00	1.4e+01
	50	-2.0e+10	-5.9e+03	-4.5e+03	-1.3e+06	3.2e+01	3.2e+01	3.2e+01	6.0e+01
INDEF	1000	-5.0e+03	-4.7e+03	-4.7e+03	-4.9e+03	1.3e-11	3.0e-06	6.9e-07	2.2e-08
	50	-1.0e+05	-1.0e+05	-1.0e+05	-9.5e+04	3.7e-10	1.7e-07	2.5e-10	3.1e-07
INTEQNELS	102	+3.2e-18	+7.9e-31	+1.1e-17	+1.1e-18	3.7e-09	4.9e-10	1.7e-07	2.2e-09
	502	+1.6e-17	+9.2e-30	+7.1e-14	+1.0e-14	8.3e-09	3.2e-09	5.4e-07	2.1e-07
JIMACK	1521	+8.7e-01	+8.7e-01	+8.7e-01	+9.1e-01	3.1e-06	8.7e-06	8.5e-06	3.6e-01
	81	+8.9e-01	+8.7e-01	+1.1e+00	+8.9e-01	1.6e-02	3.4e-04	9.4e-01	1.7e-02
LIARWHD	1000	+7.4e-19	+6.2e-28	+5.6e-28	+6.8e-19	8.9e-09	3.1e-06	3.1e-06	8.5e-09
	36	+1.0e-25	+4.9e-29	+0.0e+00	+7.2e-22	2.3e-11	4.7e-09	4.7e-09	1.9e-09
MANCINO	50	+1.5e-21	+6.0e-24	+5.9e-20	+1.3e-21	5.4e-08	5.6e-07	5.6e-07	5.2e-08
MODBEALE	10	+2.3e-21	+9.9e-26	+1.5e-25	+1.7e-14	1.6e-10	3.3e-06	3.3e-06	2.5e-07
	2000	+3.0e-15	+1.8e-25	+5.2e-14	+8.0e+00	1.3e-07	3.5e-08	6.7e-07	4.8e-04
MOREBV	1000	+6.7e-12	+7.8e-13	+3.1e-12	+1.8e-14	2.9e-08	5.3e-06	7.6e-06	5.7e-09
	50	+1.2e-09	+1.2e-09	+1.2e-09	+1.1e-09	1.5e-07	1.5e-06	8.9e-06	3.9e-07
MSQRTALS	4900	+1.1e-14	+9.8e-28	+7.9e-14	+2.5e-17	1.4e-07	9.0e-08	1.0e-06	7.9e-09
	49	+8.5e-04	+1.5e-01	+3.6e-04	+7.7e-05	4.9e-01	5.9e+00	1.7e-01	4.6e-03
MSQRTBLS	4900	+4.3e-17	+6.2e-24	+6.8e-14	+1.7e-14	1.0e-08	8.6e-07	1.3e-06	2.1e-07
	49	+4.5e-03	+1.6e-01	+2.2e-04	+1.3e-05	6.1e-01	6.0e+00	6.6e-02	1.8e-03
NCB20B	1000	+1.9e+02	+1.9e+02	+1.9e+02	+1.9e+02	1.3e-08	4.8e-07	9.8e-07	3.0e-08
	180	+9.2e+02	+9.2e+02	+9.2e+02	+9.1e+02	1.3e-08	2.6e-06	9.2e-08	2.7e-08
NCB20	1010	+3.5e+02	+3.5e+02	+3.5e+02	+3.5e+02	3.3e-08	1.7e-07	8.9e-07	4.1e-08
	110	+1.7e+03	+1.7e+03	+1.7e+03	+1.7e+03	3.8e-08	2.8e-06	2.9e-06	2.1e-07

Continued on next page

### 第三章 齐次二阶下降法

表 3.B.3 CUTEst 数据集的完整结果，函数值和梯度范数

name	$n$	$f$				$\ g\ $			
		A	H	Hv	N	A	H	Hv	N
NONCVXU2	1000	+2.3e+01	+2.3e+01	+2.3e+01	+2.3e+01	1.4e-11	1.3e-06	1.3e-06	7.1e-12
	10	+2.3e+03	+2.3e+03	+2.3e+03	+2.3e+03	1.5e-08	7.0e-06	1.5e-06	6.4e-07
NONCVXUN	1000	+2.3e+01	+2.3e+01	+2.3e+01	+2.3e+01	3.4e-10	5.0e-09	4.8e-09	4.0e-07
	10	-	+2.3e+03	+2.3e+03	+2.3e+03	-	3.0e-02	5.3e-02	7.0e-02
NONDIA	1000	+3.3e-23	+5.0e-28	+0.0e+00	+2.6e-26	4.1e-10	3.4e-10	3.3e-10	9.1e-12
	90	+2.0e-23	+6.4e-27	+2.0e-29	+2.1e-23	2.6e-09	1.9e-07	1.9e-07	2.6e-09
NONDQUAR	1000	+9.0e-07	+3.5e-06	+3.0e-06	+5.0e-05	9.4e-07	7.2e-06	7.5e-06	3.2e-03
	100	+1.6e-06	+3.5e-06	+5.0e-06	+5.3e-07	9.8e-07	9.1e-06	7.0e-06	9.0e-07
NONMSQRT	4900	+1.1e+00	+1.1e+00	+1.1e+00	+1.1e+00	9.5e-07	3.1e-02	8.5e-02	6.4e-07
	49	+7.2e+02	+7.5e+02	+7.9e+02	+7.3e+02	2.5e+00	4.8e+01	6.8e+01	3.7e+02
OSCIGRAD	1000	+2.8e-09	+2.8e-09	+2.4e-09	+2.8e-09	4.6e-08	9.2e-06	1.3e-03	3.6e-08
	15	+5.6e-24	-	+1.7e-21	+3.2e-23	6.1e-08	-	6.8e-07	8.9e-08
OSCIPATH	25	+1.0e+00	+1.0e+00	+1.0e+00	+1.0e+00	4.6e-09	2.6e-06	2.6e-06	4.6e-09
	500	+1.0e+00	+1.0e+00	+1.0e+00	+1.0e+00	4.6e-09	2.6e-06	2.6e-06	4.6e-09
PENALTY1	1000	+4.3e-04	+4.3e-04	+4.3e-04	+4.3e-04	2.6e-07	8.6e-06	1.5e-06	1.2e-08
	50	+9.7e-03	+9.7e-03	+9.7e-03	+9.7e-03	4.3e-03	1.8e-07	1.8e-07	1.9e-03
PENALTY2	1000	+4.3e+00	+4.3e+00	+4.3e+00	+4.3e+00	1.4e-08	1.6e-09	7.7e-06	1.9e-07
	50	+1.4e+83	+4.9e+82	+4.1e+82	+1.1e+83	4.9e+38	9.1e+69	2.5e+67	2.4e+67
PENALTY3	50	+1.0e-03	+1.0e-03	+1.0e-03	+1.0e-03	8.1e-09	2.5e-08	8.9e-07	1.2e-07
POWELLSG	1000	+5.1e-10	+1.6e-11	+1.8e-11	+5.1e-10	5.5e-07	6.9e-06	6.9e-06	5.4e-07
	60	+1.7e-09	+6.4e-13	+7.4e-12	+1.6e-09	6.6e-07	3.5e-06	2.3e-08	6.4e-07
POWER	1000	+1.1e-10	+1.2e-18	+2.4e-12	+1.1e-10	7.6e-07	3.1e-08	4.1e-08	7.6e-07
	50	+1.3e-09	+2.5e-18	+7.4e-12	+1.3e-09	1.9e-05	5.7e-08	6.7e-07	1.9e-05
QUARTC	1000	+4.6e-09	+3.1e-10	+4.5e-10	+2.8e-09	8.9e-07	4.6e-06	4.7e-06	6.0e-07
	100	+7.5e-07	+1.4e-10	+1.4e-10	+5.1e-07	2.3e-05	1.1e-06	1.2e-06	1.7e-05
SBRYBND	1000	+1.3e+02	+4.5e+02	+4.7e-05	+3.5e+02	4.2e+02	2.1e+07	3.4e+03	1.2e+05
	50	+2.9e+03	+2.1e+04	+5.0e+03	+3.6e+03	2.8e+03	1.1e+08	5.1e+06	1.8e+05
SCHMVETT	1000	-2.4e+01	-2.4e+01	-2.4e+01	-2.4e+01	3.1e-07	1.0e-10	1.0e-10	1.2e-08
	10	-3.0e+03	-3.0e+03	-3.0e+03	-3.0e+03	2.5e-08	2.1e-07	1.0e-06	5.4e-08
SCOSINE	1000	+7.1e+00	-	-	-6.5e+00	1.3e+05	-	-	5.9e+00
	10	-4.7e+02	-1.2e+02	+8.3e+02	-4.4e+02	4.5e+02	9.1e+16	1.3e+15	6.8e+04
SCURLY10	1000	+1.7e+06	+9.7e+10	+9.7e+10	+1.7e+06	5.4e+10	1.1e+20	1.1e+20	5.4e+10
	10	+1.8e+11	+3.4e+23	+3.4e+23	+1.0e+11	1.4e+14	6.6e+23	6.6e+23	9.0e+13

Continued on next page

### 第三章 齐次二阶下降法

表 3.B.3 CUTEst 数据集的完整结果，函数值和梯度范数

name	$n$	$f$				$\ g\ $			
		A	H	Hv	N	A	H	Hv	N
SCURLY20	1000	+2.1e+12	+8.6e+24	+8.6e+24	+1.2e+12	1.4e+15	1.6e+25	1.6e+25	9.1e+14
SCURLY30	1000	-	+1.5e+25	+1.5e+25	+4.1e+12	-	3.7e+25	3.7e+25	3.2e+15
SENSORS	1000	-2.0e+01	-2.1e+01	-2.1e+01	-2.0e+01	8.6e-11	2.1e-10	1.2e-07	1.9e-08
	10	-2.0e+05	-2.0e+05	-2.0e+05	-2.0e+05	1.6e-08	9.6e-06	9.6e-06	7.1e-08
SINQUAD	1000	-1.1e+03	-1.1e+03	-1.1e+03	-1.1e+03	2.5e-11	9.1e-06	9.1e-06	2.7e-08
	50	-2.9e+05	-2.9e+05	-2.9e+05	-2.9e+05	1.3e-09	2.1e-07	2.1e-07	1.2e-10
SPARSINE	1000	+3.6e-18	+1.4e-27	+1.8e-15	+6.3e-18	1.5e-08	7.5e-08	7.1e-07	1.0e-08
	50	+3.2e-18	+3.3e-20	+9.8e-13	+1.7e-11	1.9e-08	5.3e-09	9.6e-06	2.8e-04
SPARSQR	1000	+3.8e-10	+1.2e-13	+5.3e-13	+3.8e-10	4.7e-07	5.5e-09	5.5e-09	4.7e-07
	50	+2.3e-10	+1.2e-14	+5.6e-10	+2.3e-10	3.2e-07	1.6e-08	1.8e-06	3.2e-07
SPMSRTL	1000	+8.9e-17	+2.5e-17	+2.7e-13	+4.4e-16	1.1e-08	2.5e-06	2.5e-06	4.1e-08
	100	+5.4e-16	+4.3e-16	+2.1e-15	+8.5e-15	1.5e-08	3.5e-08	9.5e-07	3.6e-07
SROSENBR	500	+9.4e-17	+1.4e-28	+1.2e-28	+1.8e-18	1.7e-08	6.5e-07	6.5e-07	1.3e-09
	50	+3.4e-28	+6.5e-28	+1.3e-27	+2.8e-29	7.9e-13	4.7e-08	4.7e-08	1.1e-14
SSBRYBND	1000	+2.1e-17	+1.0e-10	+1.9e-16	+1.7e+00	2.2e-08	1.7e-01	2.3e-04	1.5e+03
	50	+1.7e-19	+1.9e+02	+1.7e-12	+1.1e+04	1.3e-08	1.9e+05	2.2e-02	9.0e+04
SSCOSINE	1000	-8.3e+00	-9.0e+00	-8.5e+00	-9.0e+00	9.7e-04	2.9e-01	6.0e+04	3.1e-09
	10	-9.9e+02	+2.5e+01	-	-1.0e+03	2.4e-02	2.4e+10	-	2.9e-03
TESTQUAD	1000	+1.2e-16	+3.0e-19	+2.4e-17	+2.8e-17	5.3e-08	1.5e-07	8.5e-08	3.3e-08
TOINTGSS	1000	+1.0e+01	+1.0e+01	+1.0e+01	+1.0e+01	4.5e-07	8.1e-09	5.4e-07	1.5e-08
	50	+1.0e+01	+1.0e+01	+1.0e+01	+1.0e+01	4.1e-07	4.8e-08	1.2e-07	5.0e-09
TQUARTIC	1000	+8.2e-22	+4.0e-30	+2.1e-30	+1.6e-28	5.5e-10	2.1e-08	2.1e-08	3.5e-13
	50	+1.5e-17	+6.0e-31	+2.9e-23	+1.2e-13	2.5e-10	6.8e-10	6.8e-10	2.2e-08
TRIDIA	1000	+1.6e-17	+4.9e-27	+1.1e-15	+6.4e-18	5.6e-08	5.9e-07	9.4e-07	4.1e-08
	50	+9.5e-19	+2.7e-25	+3.2e-16	+3.5e-19	4.0e-08	6.1e-10	9.3e-07	2.2e-08
VARDIM	200	+1.7e-06	+3.6e-02	+3.6e-02	+1.7e-06	4.3e+00	2.2e+05	2.2e+05	4.3e+00
VAREIGVL	100	+3.9e-18	+3.2e-26	+1.6e-13	+1.9e-11	7.5e-09	8.4e-07	2.9e-06	2.0e-07
	500	+8.2e-12	+3.5e-26	+6.1e-13	+7.4e-11	1.9e-07	9.4e-07	4.5e-06	8.9e-07
WATSON	12	+1.2e-07	+1.1e-08	+1.4e-08	+1.2e-07	9.3e-07	9.9e-06	4.5e-06	9.3e-07
WOODS	4000	+1.0e-19	+7.4e-24	+7.4e-24	+1.3e-20	1.3e-08	9.9e-06	9.9e-06	5.0e-09
	4	+1.8e-23	+3.0e-27	+1.4e-27	+1.6e-23	1.5e-10	1.1e-06	1.1e-06	1.1e-10
YATP1LS	120	+1.7e+00	+5.5e-26	+8.0e-27	+2.3e-17	1.6e+00	4.2e-07	4.2e-07	4.4e-10
	2600	+1.1e-21	+3.4e-24	+6.0e-25	+5.3e-23	3.4e-09	3.0e-07	3.0e-07	6.6e-10

Continued on next page



表 3.B.3 CUTEst 数据集的完整结果，函数值和梯度范数

name	$n$	$f$				$\ g\ $			
		A	H	Hv	N	A	H	Hv	N
YATP2LS	8	+1.1e+02	+3.7e-31	+6.3e-28	+1.1e+02	9.9e-07	5.2e-10	5.2e-10	1.9e-07
	2600	+2.6e-29	+3.8e-28	+8.0e-27	+1.3e+02	1.2e-13	7.5e-12	7.3e-12	4.0e-07

## 第四章 广义齐次模型及下降框架

### 第一节 简介

上一章，我们提出了一种齐次二阶下降方法 (Homogeneous Second-Order Descent Method, HSODM)。此方法中，通过求解具有聚合矩阵  $F_k$  的**普通齐次模型** (Ordinary Homogeneous Model, OHM) 来构建迭代点：

$$\min_{\|[v;t]\| \leq 1} \psi_k(v, t; F_k) := [v; t]^T F_k [v; t], F_k := \begin{bmatrix} H_k & g_k \\ g_k^T & \delta \end{bmatrix}, \quad (4-1)$$

其中变量  $v \in \mathbb{R}^n, t \in \mathbb{R}$ 。

齐次化方法**预设**了一个  $\delta \leq 0$ ，使得向量  $[v; t]$  对应于聚合矩阵  $F_k$  的最左特征向量。然后，HSODM 采用一种简单策略来找到步长  $\eta_k$ ，例如通过线搜索方法，并使用 OHM 生成的方向  $[v_k; t_k]$  更新迭代点为  $x_{k+1} = x_k + \eta_k(v_k/t_k)$ 。该方法对于非凸问题有  $O(\epsilon^{-3/2})$  的迭代复杂度。前面已经通过数值验证 HSODM 相对于标准二阶方法的优势。

在本章中，我们扩展齐次化的思想，不仅限于具有二阶 Lipschitz 连续性的非凸问题。我们引入了以下**广义齐次模型 (Generalized Homogeneous Model, GHM)**

$$\min_{\|[v;t]\| \leq 1} \psi_k(v, t; F_k) := [v; t]^T F_k [v; t] \quad \text{with} \quad F_k := \begin{bmatrix} H_k & \phi_k(x_k) \\ \phi_k(x_k)^T & \delta_k \end{bmatrix}, \quad (4-2)$$

使得  $\delta_k \in \mathbb{R}$  允许某种适应性。此外，我们引入了变换  $\phi_k : \mathbb{R}^n \mapsto \mathbb{R}^n$  来替代梯度  $g_k$ ，同时保留使用对称特征值的优势。这种灵活性便于我们设计一个通用的齐次框架，在此框架中还可以设计新的算法。

本章的贡献可以总结如下。我们在 [算法 4-1](#) 中提出了一种齐次二阶下降框架 (Homogeneous Second-Order Descent Framework, HSODF)，其中子问题中所需的线性系统替换为广义齐次模型 (4-2) (可以作为特征值问题进行求解)。我们详细对比求解线性方程组和求解 GHM 的问题条件数，并通过一些简单数值实验说明优势。

如果允许对  $\delta_k$  进行搜索，我们讨论如何利用 HSODF 生成一系列二阶优化方法。在此背景下，提出一种自适应的 HSODM ([第 5 章](#))，适用于二阶 Lipschitz 函数，这是之前基于线搜索的 HSODM 的加强版本。这种方法在非凸优化中保持了  $O(\epsilon^{-3/2})$  的迭代复杂度，同时将原始 HSODM 扩展到凸问题。接下来，我们提供了一种用来辅助定位  $\delta_k$  的二分法，其复杂度为  $O(\log(1/\epsilon))$ ；同时作为一种基础技术，可以适用于类似的、需要对  $\delta_k$  进行搜索的二阶方法。

在同一框架下，随后的章节(第6章)会讨论一种变体，该变体适用于自协 Lipschitz 而非标准的二阶 Lipschitz 函数。在这种情况下，对 GHM 种的  $\delta_k$  和  $\phi_k$  同时进行了调整。我们展示了相应的同伦 HSODM 通过一种“非内点法”的路径跟随技术表现出全局线性收敛率。注意，我们称其为同伦 (homotopy) HSODM, 因其求解路径类似于内点法等同伦算法，但该算法不能归类为内点法；故我们称之为使用了“非内点法”的路径跟随技术。对于这种变体，值得一提的是，无需任何辅助过程来定位  $\delta_k$ ；相比上述方法，每次迭代最多只需 2 个 GHM。

这里我们对本章的符号做一些调整。记  $\|\cdot\|$  为  $\mathbb{R}^n$  空间中的标准欧几里得范数。对于矩阵  $A \in \mathbb{R}^{n \times n}$ ， $\|A\|$  表示其诱导的  $\ell_2$  范数。记  $A^\star$  为矩阵  $A$  的伪逆。令  $P_{\mathcal{X}}$  为投影到空间的正交投影算子，其中  $\mathcal{X} \subseteq \mathbb{R}^n$ 。我们使用 mod 表示二进制模运算。我们说向量  $y$  与子空间  $\mathcal{S}$  正交，即  $y \perp \mathcal{S}$ ，如果对于任意非零向量  $u \in \mathcal{S}$ ，都有  $u^T y = 0$ 。

同时我们为 Hessian 阵  $H_k$  的特征值引入以下符号。在算法迭代点  $x_k$  处，我们假定  $H_k$  有  $r$  个 ( $1 \leq r \leq n$ ) 不同的特征值  $\{\lambda_1(H_k), \dots, \lambda_r(H_k)\}$ ，其中  $\lambda_1(H_k) < \dots < \lambda_r(H_k)$ ，且  $\mathcal{S}_1(H_k), \dots, \mathcal{S}_r(H_k)$  是由相应特征向量生成的子空间。我们有时将  $\lambda_{\min}$  和  $\lambda_{\max}$  作为  $\lambda_1$  和  $\lambda_r$  的同义词。我们将  $H_k$  的条件数记为  $\kappa(H_k) = \frac{\lambda_r(H_k)}{\lambda_1(H_k)}$ 。由于对特征值的讨论主要限于迭代点  $x_k$ ，我们有时会为了简便省略索引  $k$ 。

## 第二节 齐次二阶下降框架

我们首先给出 (4-2) 的最优性条件。我们省略证明，因为它遵循于标准信赖域子问题的全局最优性条件。

### 引理 4.1: 最优性条件

$[v_k; t_k]$  是子问题 (4-2) 的最优解，当且仅当存在一个对偶变量  $\theta_k \geq 0$ ，使得

$$\begin{bmatrix} H_k + \theta_k \cdot I & \phi_k \\ \phi_k^T & \delta_k + \theta_k \end{bmatrix} \geq 0, \quad (4-3)$$

$$\begin{bmatrix} H_k + \theta_k \cdot I & \phi_k \\ \phi_k^T & \delta_k + \theta_k \end{bmatrix} \begin{bmatrix} v_k \\ t_k \end{bmatrix} = 0, \quad (4-4)$$

$$\theta_k \cdot (\|[v_k; t_k]\| - 1) = 0. \quad (4-5)$$

GHM 中的最优解在若干方面与 OHM 略有不同。由于引入了  $\delta_k$  的适应性， $[v_k; t_k]$  可能不会达到单位球的边界，因此根据 (4-5)， $\theta_k = 0$ 。这在凸函数和某些  $\delta_k \gg 0$  的情况下是可能的。为了坚持使用特征值过程，我们应防止  $\delta_k$  过大，特别是在迭代点接近局部最小值时。

另一个棘手的情况是当  $t_k = 0$  时，无法规范化一个方向。这个困境对应于信赖域子问题的**难解**

**情况 (Hard Case)**<sup>[118,146]</sup>。然而，在 OHM 中，这个问题可以通过使用固定半径策略或在  $|t_k| < \nu$  时引入截断  $\nu$  来更新  $d_k$ ，从而轻松解决。因此，原始 HSODM 不需要额外的操作，而 GHM 需要一些非平凡的分析，如 第 5.2.3 节 中所示。

接下来，我们在 算法 4-1 中展示了齐次二阶下降框架 (HSODF)，该框架使用 GHM (4-2) 作为子程序。

---

**算法 4-1:** 齐次二阶下降框架 (HSODF)

---

```

1  给定初始点  $x_1$ ，控制参数  $\delta_1$ ，最大迭代次数  $K_{\max}$ ;
2  forall  $k = 1, \dots, K_{\max}$  do
3      forall  $j = 1, \dots, \mathcal{T}_k$  do
4          构建一个 GHM:  $F_{k,j} = \begin{bmatrix} H_k & \phi_{k,j}(x_{k,j}) \\ \phi_{k,j}(x_{k,j})^T & \delta_{k,j} \end{bmatrix}$ ;
5          获取  $[v_{k,j}; t_{k,j}] = \arg \min_{\|[v;t]\| \leq 1} \psi_{k,j}(v, t; F_{k,j})$  (参见 (4-2));
6          设置  $d_{k,j} := v_{k,j}/t_{k,j}$ ;
7          if  $d_{k,j}$  满足 (内部) 终止条件 then
8              设置  $d_k := d_{k,j}$ ; 跳出;
9          调整  $\delta_{k,j}$  和  $\phi_{k,j}$ 。
10     更新  $x_{k+1} = x_k + d_k$ ;
11     if  $x_{k+1}$  满足 (外部) 终止条件 then
12         Output:  $x_k$ 

```

---

请注意，HSODF 中的适应性策略排除了算法参数中 Lipschitz 常数的需求。与原始 HSODM 不同，那里  $\delta_k \equiv \delta < 0$  是固定的，且  $\phi_k = g_k$ ，每次迭代  $k$  都涉及一个由  $j$  标记的内部循环，以搜索合适的  $\delta_{k,j}$  或  $\phi_{k,j}$ ，当  $d_{k,j}$  满足某些条件时终止 (参见 行 6)。通过适当设计这些条件，在类似迭代复杂度下，可以恢复到一些现存的二阶方法。通常，对于每个  $k$ ，内部迭代的规模  $\mathcal{T}_k$  大约为  $O(\log(1/\epsilon))$ ；但在同伦 HSODM 中，对于自协 Lipschitz 函数只需要 2 个 GHM。

#### 4.2.1 HSODF 的动机

广义齐次模型及下降框架主要受一些具有强稀疏性和低秩结构的高维数据问题的启发。具体来说，我们的计算发现表明，解决 GHM 可能需要比牛顿方程更少的 Krylov 迭代次数，特别是当 Hessian  $H_k$  是退化的情况下。这些发现部分归因于特征值问题和 Krylov 子空间方法的线性方程的条件数差异。基于此，我们进一步比较了 HSODF 和牛顿型方法在每次迭代成本上的差异。

## 4.2.1.1 HSODF 与牛顿型方法的计算成本比较

首先，我们比较牛顿型方程、和相应的齐次模型利用迭代方法的计算成本。特别地，我们考虑 Hessian  $H_k$  是正定但病态的情况。假定在某次迭代  $x_k \in \mathbb{R}^n$  中，求解以下扰动牛顿型方程，

$$(H_k + \epsilon_N I) d_k = -g_k. \quad (4-6)$$

在不失一般性的情况下，我们假定  $\epsilon_N \in (0, 1)$ 。当  $n$  很大时，计算依赖于迭代方法，如共轭梯度方法 (CG), 广义最小残差方法 (GMRES), 和重启的广义最小残差方法 (rGMRES)<sup>[71]</sup>。假定在这种情况下 HSODM 使用具有  $\delta_k = -\epsilon_L$  的 GHM,

$$F_k := \begin{bmatrix} H_k & g_k \\ g_k^T & -\epsilon_L \end{bmatrix}, \quad (4-7)$$

其核心工具是对称特征值问题的方法，如 Lanczos 方法 (Lanczos)<sup>[149]</sup>。众所周知，求解线性系统的复杂度取决于扰动矩阵  $H_k + \epsilon_N I$  的条件数。共轭梯度方法所需的迭代次数通常为

$$O(\sqrt{\kappa(H_k + \epsilon_N I)} \log(1/\epsilon)) \quad \text{其中} \quad \kappa(\cdot) = \frac{\lambda_{\max}(H_k) + \epsilon_N}{\lambda_1(H_k) + \epsilon_N}, \quad (4-8)$$

其中  $\kappa(\cdot)$  通常被称为条件数。相比之下，对于寻找最小特征值及其对应的特征向量，Lanczos 方法<sup>[103]</sup> 的迭代复杂度界限为 (另见 定理 4.18)

$$O\left(\sqrt{\kappa_L(F_k)} \log(1/\epsilon)\right) \quad \text{其中} \quad \kappa_L(F_k) := \frac{\lambda_{\max}(F_k) - \lambda_1(F_k)}{\lambda_2(F_k) - \lambda_1(F_k)} \quad (4-9)$$

以上结果表明我们会以高概率接受某个最小特征值的估计  $\xi$ , 使得  $\xi - \lambda_1(F_k) \leq \epsilon$ 。以下结果表明， $\kappa_L(F_k)$  总是有界的，这与当  $\epsilon_N$  逼近 0 时， $\kappa(H_k + \epsilon_N I)$  无界的情况形成对比。此外，在退化情况下， $\kappa_L(F_k)$  可以远小于  $\kappa(H_k + \epsilon_N I)$ 。

**定理 4.2**

对于 (4-7) 中的聚合矩阵  $F_k$ ，假定  $U_H := \lambda_{\max}(H_k) \gg \epsilon_N$ ，则有

(a) 对于任意  $\epsilon_L > 0$ ， $\kappa_L$  总是有限的，具体而言，

$$\kappa_L(F_k) \leq \frac{2(\lambda_{\max}(H_k) - \epsilon_L - \lambda_1(F_k))}{-U_H + \epsilon_L + \sqrt{(U_H + \epsilon_L)^2 + \|g_k\|^2/n}} < \infty. \quad (4-10)$$

(b) 此外，假定  $\lambda_1(H_k) = 0$ ，则

$$\frac{\kappa_L(F_k)}{\kappa(H_k + \epsilon_N I)} \leq O\left(\frac{\epsilon_N}{\frac{\|g_k\|^2}{U_H + \epsilon_L} + \epsilon_L}\right). \quad (4-11)$$

为了简洁起见，我们将上述定理的证明推迟到 第 4.A.3 节。定理第二部分中的比率分别比较了解牛顿型方程和 GHM 时的两个条件数。比率的较小值意味着 GHM 的条件数相对于牛顿型方程更好。这个比率表明，由于  $\|g_k\|$  的隐式缩放， $\kappa_L$  通常比牛顿型方程的条件数更好且更稳健。例如，令

$\epsilon_N = \epsilon_L \rightarrow 0$ , 则分子趋近于 0, 而分母保持常数阶, 表明在这种情况下  $\kappa_L$  要小得多。相反的极端情况表明, 这两个条件数变得接近; 如果  $\|g_k\| \rightarrow 0$ , 情况也是如此。此外, 这一分析呼应了梯度正则化步骤, 其中扰动被设定为  $\epsilon_N = \|g_k\|^{1/2}$  [49,115]。在数值上, 以上事实在图 4.2.1 中得到了可视化, 其中估计的  $\kappa_L$  来自引理 4.19。请注意, 当  $g_k \perp \mathcal{S}_1$  时, 估计是紧的。

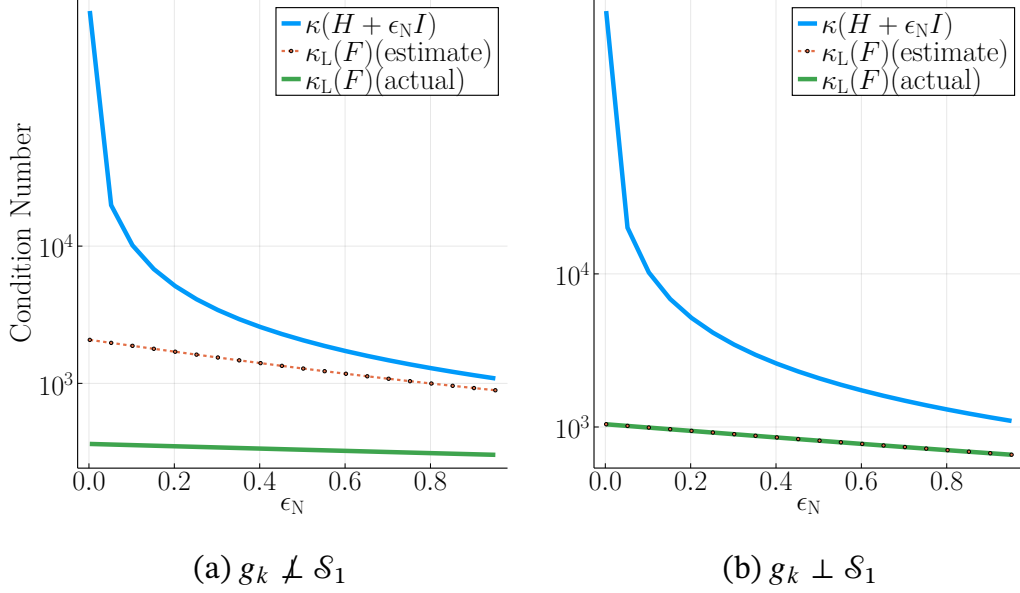


图 4.2.1  $\kappa(H_k + \epsilon_N I)$  和  $\kappa_L(F_k)$  在退化情况下的比较。

#### 4.2.1.2 每次迭代成本的进一步数值分析

受上述分析的启发, 我们进行了实验, 比较了使用迭代方法的 Krylov 迭代次数。我们使用 CG、GMRES、rGMRES 分别表示它们在 (4-6) 上的对应结果。我们在 (4-7) 中使用 Lanczos 方法。对于求解线性方程 (4-6), 我们将残差设定为  $10^{-5}$ , 并且在 (4-7) 中的 Lanczos 方法在固定容忍度  $10^{-7}$  下终止。

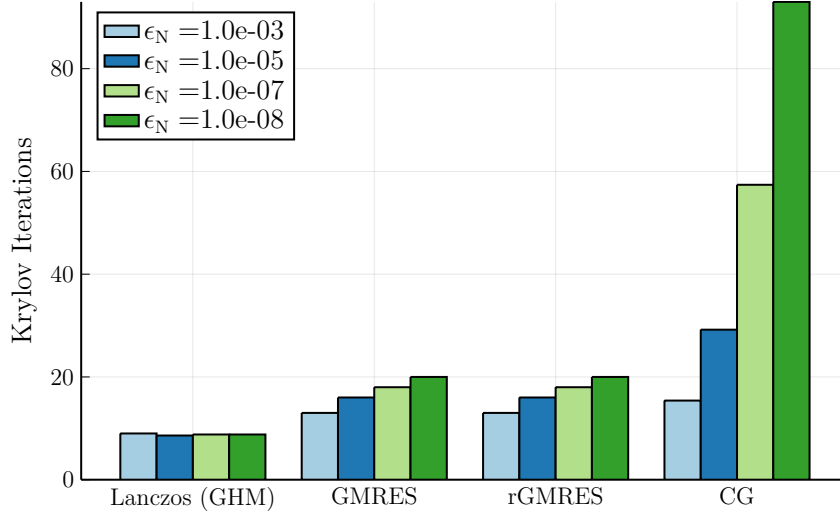
**Hilbert 矩阵生成的例子** 考虑 Hilbert 矩阵 [83] 作为已知病态的  $H$ 。维度为  $n$  的 Hilbert 矩阵具有以下解析形式:

$$H_{ij} = \frac{1}{i+j-1}, i \leq n, j \leq n.$$

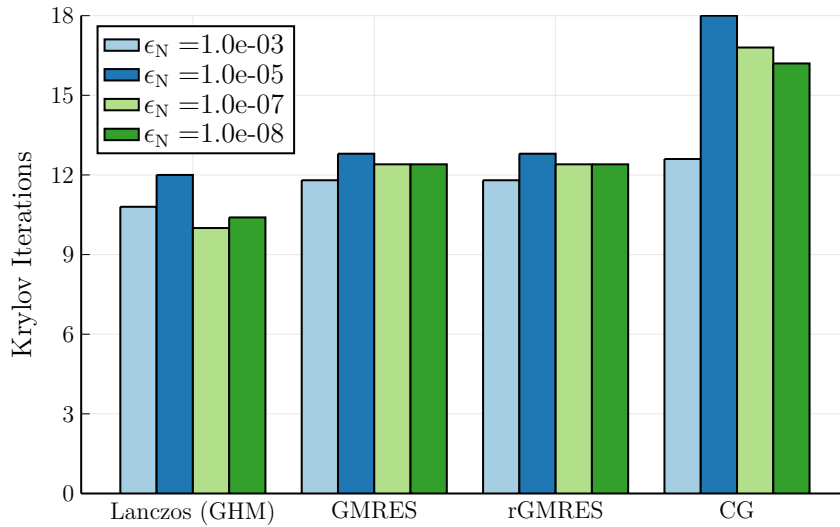
已知条件数  $\kappa(H)$  以  $O((1 + \sqrt{2})^{4n}/\sqrt{n})$  增长。在不同的  $\epsilon_N \in \{10^{-3}, 10^{-5}, 10^{-7}, 10^{-8}\}$  下, 我们比较了四种提到的算法 ( $\epsilon_L = \epsilon_N$ )。我们设定  $n = 300$ , 并随机生成具有大范数和小范数的  $g_k$ , 然后在图 4.2.2 中收集 Krylov 迭代次数的平均值。

结果基本上表明, 当  $\|g_k\|$  较大时 (见图 4.2.2a), 对于 GHM (通过 Lanczos), Krylov 迭代次数在不同的  $\epsilon_L$  下几乎保持不变, 而 CG、GMRES 和 rGMRES 的迭代次数可能随条件数的增加而增长。这

与我们的理论分析一致，即解决 GHM 的复杂度较少受 Hessian 矩阵条件数的影响。对于  $\|g_k\|$  非常小的情况 (见 图 4.2.2b)，求解牛顿方程和 GHM 之间没有太大差异。我们还注意到，对于 GHM，选择较小的  $\epsilon_L$  而不会导致条件数恶化是相当安全的。这一特性在实践中很有利，因为通常我们会自适应地为牛顿方法设置正则化参数。



(a)  $\|g_k\| = 5.17 \times 10^{-1}$



(b)  $\|g_k\| = 5.17 \times 10^{-7}$

图 4.2.2 计算扰动 Hilbert 矩阵和 GHM 的牛顿型方向的结果。蓝色到绿色的条形图表示 Krylov 迭代次数。图例中显示的数字对应于不同的正则化  $\epsilon_N$ 。

LIBSVM 实例 接下来, 考虑源自实际最小二乘问题的牛顿型系统:

$$\min_{\beta} \frac{1}{2N} \sum_{i=1}^N \|x_i^T \beta - y_i\|^2 + \frac{\gamma}{2} \|\beta\|^2. \quad (4-12)$$

其中,  $x_i \in \mathbb{R}^n, \gamma > 0, \beta \in \mathbb{R}^n$ ,  $N$  表示数据点的数量。易见, 在某个  $\beta$  处, 牛顿系统具有以下形式:

$$\left( \frac{1}{N} X^T X + \gamma \cdot I \right) \Delta \beta = -\frac{1}{N} X^T (X\beta - y). \quad (4-13)$$

在该情况下,  $\epsilon_N = \gamma$  作为扰动参数; 我们在 LIBSVM 库<sup>注1</sup>的一组问题上测试该系统, 其中  $\gamma \in \{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$ 。类似地, 我们跟踪 Krylov 迭代次数, 以找到满足相应残差的迭代点。我们从均匀分布中随机生成五个  $\beta$  样本, 并生成相应的牛顿方程和 GHM。通过计算所需 Krylov 迭代次数的平均值, Lanczos、CG、GMRES 以及 rGMRES 的结果展示在表 4.2.1 中。当存在退化现象时, GHM 表现更优, 例如在 rcv1 和 news20 实例中。在这些情况下, Lanczos 方法的表现最佳, 并且当  $\gamma$  变小时不会恶化。

## 4.2.2 原始-对偶解的刻画

在通用二阶方法框架下(算法 4-1), 将 GHM 作为子问题使用, 引发了进一步的兴趣。为了分析所提出的算法, 我们对子问题 (4-2) 在控制参数  $\delta_k$  和  $\phi_k$  方面的原始-对偶解进行全面刻画。请注意, 这些分析中的许多内容是专门针对 GHM 开发的, 而 OHM 并不需要这些分析。我们将本小节的证明推迟到第 4.B 节。

### 4.2.2.1 基本结果

现在, 我们考虑  $\theta_k = 0$  的情况, 此时与特征值问题的等价性不再成立。

#### 引理 4.3: 负曲率的存在性

在 (4-2) 中,  $\theta_k = 0$  当且仅当  $F_k \geq 0$ 。

GHM 和 HSODF 的成功依赖于  $F_k$  为不定矩阵。直观上, 控制变量  $\delta_k$  不能太大, 否则 (4-2) 的最优解将落入单位球的内部。我们在引理 4.17 中给出了全面的刻画。需要注意的是, 仅当  $H_k \geq 0$  时, 才可能出现  $F_k \geq 0$ , 因此当 Hessian 是半正定时, 我们提供了一种特殊描述。

#### 引理 4.4: 凸情况的充分性

在齐次模型 (3-8) 中, 假定  $H_k \geq 0$ , 则只要  $\delta_k < \overline{\delta_k^{\text{cvx}}} := \phi_k^T H_k^* \phi_k$ , 就有  $\lambda_1(F_k) < 0$ 。

下一个引理描述了  $\theta_k$  的上界。

注1 <https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/>



表 4.2.1 线性最小二乘问题中计算一个牛顿型方向或 GHM 的数据集详情及 Krylov 迭代次数的平均值。

name	Details		Method	$\gamma$			
	$n$	$N$		$10^{-3}$	$10^{-4}$	$10^{-5}$	$10^{-6}$
a4a	122	4781	GMRES	18.6	29.2	31.6	32.2
			rGMRES	18.6	29.2	31.6	32.2
			CG	20.8	46.8	77.2	80.6
			Lanczos	6.0	6.0	6.0	6.0
a9a	123	32561	GMRES	18.6	29.2	29.4	31.4
			rGMRES	18.6	29.2	29.4	31.4
			CG	20.6	44.6	73.2	70.2
			Lanczos	6.0	6.0	6.0	6.0
w4a	300	6760	GMRES	19.2	32.0	39.0	42.8
			rGMRES	19.2	32.0	39.0	42.8
			CG	20.6	45.0	81.0	93.6
			Lanczos	6.0	6.0	6.0	6.0
rcv1	47236	20242	GMRES	14.0	28.0	52.0	87.8
			rGMRES	14.0	28.0	52.0	103.8
			CG	13.0	33.2	84.4	199.0
			Lanczos	5.0	5.0	5.0	5.0
covtype	54	581012	GMRES	8.0	9.6	9.6	9.6
			rGMRES	8.0	9.6	9.6	9.6
			CG	7.0	11.6	14.0	14.0
			Lanczos	6.0	6.0	6.0	6.0
news20	1355191	19996	GMRES	11.0	20.0	35.0	60.0
			rGMRES	11.0	20.0	75.0	71.2
			CG	11.0	27.0	74.2	124.0
			Lanczos	5.0	5.0	5.0	5.0

#### 引理 4.5: $\theta_k$ 的上界

在 GHM 中, 有:

$$\theta_k \leq \max\{-\delta_k, -\lambda_1(H_k), 0\} + \|\phi_k(x)\|. \quad (4-14)$$

接下来, 我们讨论  $t_k = 0$  的情况。回顾以下关于  $F_k$  的谱的引理, 当  $\phi_k \perp \mathcal{S}_1(H_k)$  时成立。

**引理 4.6:** Rojas et al.<sup>[146]</sup> 的引理 3.1, 3.2

对于任意  $q \in \mathcal{S}_j(H_k), 1 \leq j \leq r$ , 定义

$$p_j = -(H_k - \lambda_j(H_k)I)^\star \phi_k, \tilde{\alpha}_j = \lambda_j(H_k) - \phi_k^T p_j,$$

则:

- (a) 当且仅当  $\phi_k \perp \mathcal{S}_j(H_k)$  时,  $(\lambda_j(H_k), [0; q])$  是  $F_k$  的一个特征对。
- (b) 当且仅当  $\phi_k \perp \mathcal{S}_j(H_k)$  且  $\delta_k = \tilde{\alpha}_j$  时,  $(\lambda_j(H_k), [1; p_j])$  是  $F_k$  的一个特征对。

上述结果基本上表明, 如果  $\phi_k$  正交于特征空间, 则  $H_k$  的特征向量可用于构造  $F_k$  的特征向量。 $F_k$  的第  $j$  个特征空间的维度取决于  $\delta_k$  是否与上述定义的临界值  $\tilde{\alpha}_j$  一致。如果  $\phi_k \perp \mathcal{S}_1(H_k)$ , 则该结果仅表明  $\lambda_1(H_k)$  是  $F_k$  的一个特征值, 但不一定是最小的特征值。仅在某些特殊的  $\delta_k$  取值下, 才有  $\lambda_1(H_k) = \lambda_1(F_k)$ 。

我们现在给出使得  $t_k$  为零的必要条件。

**推论 4.7:** 使  $t_k = 0$  的必要条件

在 (4-2) 中, 假定  $\phi_k \perp \mathcal{S}_1(H_k)$ , 如果  $t_k = 0$ , 则  $\delta_k \geq \tilde{\alpha}_1$ 。更具体地:

- (a) 如果  $\delta_k < \tilde{\alpha}_1$ , 则  $\lambda_1(F_k) < \lambda_1(H_k)$  且  $t_k \neq 0$ ;
- (b) 如果  $\delta_k = \tilde{\alpha}_1$ , 则  $\lambda_1(F_k) = \lambda_1(H_k)$ , 且  $[0; q], [1; p_1]$  是与  $\lambda_1(F_k)$  相关的特征向量, 其中  $q \in \mathcal{S}_1$ ,  $p_1$  由 引理 4.6 定义, 这进一步意味着  $t_k \in [0, 1]$ ;
- (c) 如果  $\delta_k > \tilde{\alpha}_1$ , 则  $\lambda_1(F_k) = \lambda_1(H_k)$  且  $t_k = 0$ 。

该推论的证明由其前两个引理直接推得。然而, 需要注意的是, **逆命题**并不成立。确切地说, 如果  $\phi_k \perp \mathcal{S}_1(H_k)$  且  $\delta_k = \tilde{\alpha}_1$ , 那么确实有  $\lambda_1(F_k) = \lambda_1(H_k)$ 。然而,  $t_k$  可能不为零, 因为其特征向量可以是  $[0; q]$  和  $[1; p_1]$  的线性组合, 其中  $q \in \mathcal{S}_1(H_k)$ 。充分性仅在  $\delta_k > \tilde{\alpha}_1$  时成立。实际上, 虽然  $\phi_k \perp \mathcal{S}_1(H_k)$ , 但  $t_k \neq 0$  并不是一个不希望出现的情况。在算法设计中, 每当  $t_k = 0$  时,  $\delta_k$  都不应被增加, 这一策略正是由必要条件所引导的。

作为一个附加结果, 下面的引理刻画了  $\tilde{\alpha}_1$  的量级。

**引理 4.8:**  $\tilde{\alpha}_1$  的排序

定义  $p_1, \tilde{\alpha}_1$  类似于 引理 4.6, 则对于我们关注的最小特征值, 满足  $\lambda_1(H_k) \leq \tilde{\alpha}_1$ 。

## 4.2.2.2 辅助函数的连续性

为了便于讨论，我们将对偶变量  $\theta_k$ 、步长的范数  $\|d_k\|$  以及比率  $\theta_k(\delta_k)/\|d_k\|$  视作  $\delta_k$  的函数，这些被称为辅助函数。接下来，我们讨论这些函数的连续性和可微性。

**定义 4.9:  $\delta_k$  的辅助函数**

在每次迭代  $x_k$  处，考虑 GHM 及其对应的  $\delta_k$ ，令  $v_k, t_k$  为相应解。设  $\bar{\Delta}_k < +\infty$  为步长的上界。

$$\begin{aligned}\Delta_k : \mathbb{R} &\mapsto \mathbb{R}_+, \Delta_k(\delta_k) := \begin{cases} \|v_k/t_k\|^2 & \text{if } \delta_k < \tilde{\alpha}_1 \\ \bar{\Delta}_k & \text{o.w.} \end{cases} \\ \omega_k : \mathbb{R} &\mapsto \mathbb{R}_+, \omega_k(\delta_k) := \theta_k^2 \\ h_k : \mathbb{R} &\mapsto \mathbb{R}_+, h_k(\delta_k) := \frac{\omega_k(\delta_k)}{\Delta_k(\delta_k)}\end{aligned}$$

在后续讨论中，我们将简写  $\Delta_k = \Delta_k(\delta_k), \omega_k = \omega(\delta_k), h_k = h_k(\delta_k)$ 。默认这些带有下标  $k$  的值对应当前迭代点  $x_k$ 。需要回顾的是，如果  $\delta_k = \tilde{\alpha}_1$  (参见 [推论 4.7](#))，则  $t_k \in [0, 1]$  是**集值**的。因此，此时  $v_k/t_k$  并不严格定义。然而，这种情况仅在  $\delta_k = \tilde{\alpha}_1$  时发生，因此我们无需担心这一问题，因为对  $F_k$  进行任何扰动都将消除该集值情况。

在 [图 4.2.3](#) 中，我们给出了  $g_k \perp \mathcal{S}_1(H_k)$  **近似成立**时的凸和非凸示例。根据该示例，我们有如下观察。首先，在凸和非凸情况下，当  $\delta_k \nearrow \tilde{\alpha}_1$  (如 [引理 4.6](#) 定义)， $t_k$  可以从接近 1 的值跳变为 0。此外， $\Delta_k$  在两种情况下均是递增的，并由于  $t_k$  也出现明显的跳变。相比之下， $\theta_k$  (以及  $\omega_k$ ) 在整个  $\delta_k \in \mathbb{R}$  上是连续的。实际上， $\theta_k$  的可微性与  $t_k = 0$  直接相关。我们形式化地总结如下结论。

**引理 4.10:  $\theta_k, \omega_k$  的连续性**

对于  $\phi_k \neq 0$ ，我们有： $\theta_k, \omega_k$  是递减凸并在所有  $\delta_k \in \mathbb{R}$  上连续。此外，在  $t_k \neq 0$  时， $\theta_k$  可微，满足

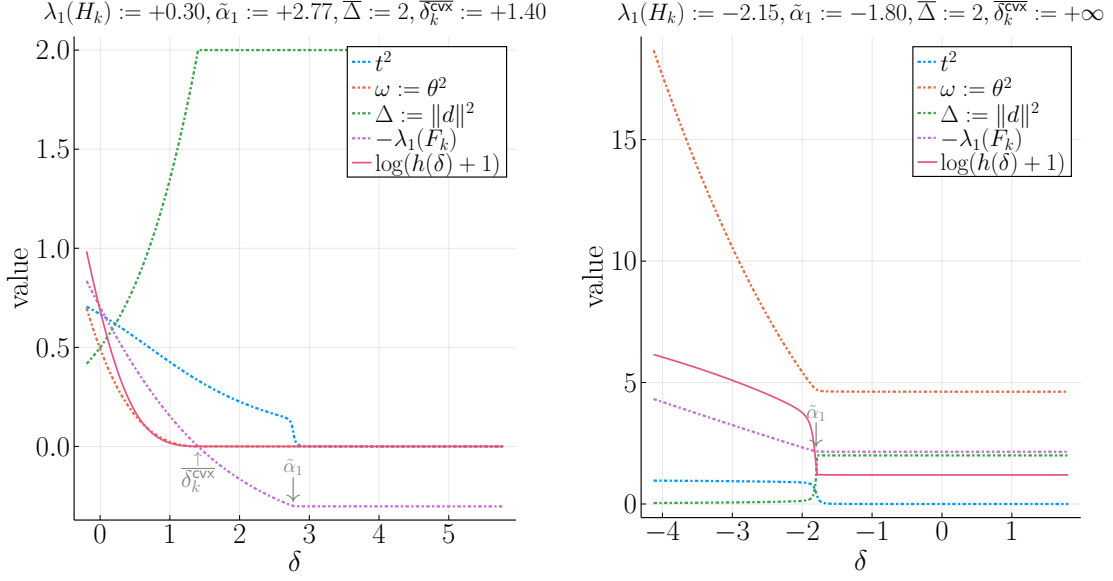
$$d\delta_k \theta_k = -\frac{1}{\Delta_k + 1} \quad (4-15)$$

现在我们转向  $\Delta_k$  的连续性。

**引理 4.11:  $\Delta_k$  的连续性**

对于每次迭代  $x_k$ ，将  $\Delta_k$  视作  $\delta_k$  的函数 (见 [定义 4.9](#))。则  $\Delta_k$  在  $\delta_k$  上连续当且仅当  $t_k \neq 0$ ，即：

(a) 若  $\phi_k \not\perp \mathcal{S}_1(H_k)$ ，则  $\Delta_k$  在所有  $\delta_k \in \mathbb{R}$  上连续。



(a) 凸情况

(b) 非凸情况

图 4.2.3 当  $g_k \perp \mathcal{S}_1(H_k)$  时的“扰动”情形示意图，其中  $\phi_k = g_k + \varepsilon \cdot u_1, u_1 \in \mathcal{S}_1(H_k)$  以提供更清晰的展示。图中标注了拐点  $\delta_k^{\text{cvx}}$  和  $\tilde{\alpha}_1$ 。

(b) 若  $\phi_k \perp \mathcal{S}_1(H_k)$ ，则  $\Delta_k$  在  $\delta_k = \tilde{\alpha}_1$  处不连续。

其证明与 引理 4.10 类似，故省略。根据  $h_k$  的定义可知，其连续性也依赖于  $\Delta_k$  的连续性。一旦  $\phi_k \not\perp \mathcal{S}_1(H_k)$ ，我们可以证明  $h_k$  可微且单调递减。

#### 引理 4.12: 可微性与单调性

若  $\phi_k \perp \mathcal{S}_1(H_k)$ ，则  $h_k(\delta)$  在  $\tilde{\alpha}_1$  处不连续；否则， $h_k(\delta)$  连续。此外， $h_k(\delta)$  在  $\delta$  上可微且单调递减。

现在我们给出**困难情况** (hard case) 的定义。

#### 定义 4.13: HSODF 的困难情况

若在迭代  $k$  处， $h_k(\delta)$  在  $\tilde{\alpha}_1$  处不连续，则称该迭代处于困难情况。

在 HSODF 中，**困难情况** 仅在  $x_k$  处  $f$  为非凸时才有影响。如图 4.2.3 所示，若  $f$  在  $x_k$  附近是凸的，则函数  $h_k(\delta)$  在  $\tilde{\alpha}_1$  处可能不连续，但  $\tilde{\alpha}_1$  大于拐点  $\delta_k^{\text{cvx}}$ ，此时  $\theta_k = 0$ 。这不会影响我们的算法：若聚合矩阵  $F_k$  具有正的最小特征对，我们可以减小  $\delta_k$ ，此时困难情况无关紧要。

相反，若  $f$  在局部是非凸的，则  $t_k = 0$  确实会阻碍计算。我们可以对  $\phi_k$  施加微小扰动以避免

困难情况。

#### 4.2.3 在二阶算法中使用 GHM

我们现在简要讨论如何使用 GHM 来恢复一些标准的二阶方法。例如，我们展示了如何通过二分法使用  $O(\log(\epsilon^{-1}))$  个 GHM 来求解信赖域子问题。此外，我们还可以提供梯度正则化牛顿步的替代方案<sup>[50,51,115]</sup>，其中内层循环中的搜索过程可以通过牛顿法（而非二分法）在  $O(\log \log(\epsilon^{-1}))$  个 GHM 内完成。

##### 4.2.3.1 利用 GHM 恢复信赖域方法

考虑用于非凸优化的信赖域方法，其涉及求解以下信赖域子问题 (TRS)，其中  $\Delta > 0$ ：

$$\min_{\|d\| \leq \Delta} m_k(d) := f(x_k) + g_k^T d + \frac{1}{2} d^T H_k d. \quad (4-16)$$

我们可以通过将  $\phi_k$  设为  $g_k$  来利用 GHM：

$$F_k = [H_k, g_k; g_k^T, \delta_k].$$

显然，对于任意  $v \in \mathbb{R}^n, t \neq 0$ ，有：

$$m_k(v/t) - f(x_k) + \frac{1}{2} \delta_k = \frac{1}{2t^2} \psi_k(v, t; F_k).$$

这导致 HSODF 采用以下更新方式：

$$x_{k+1} = x_k - d_k, \quad d_k := v_k/t_k.$$

使用 [算法 4-1](#)，我们能够恢复信赖域方法，其中 HSODF 的外层迭代  $k$  遵循相同的  $\Delta$  更新规则，并在内层迭代  $j$  中重复求解 GHM 以获得信赖域子问题 (TRS) 的相同解。基本思想是搜索合适的  $\delta_k$ ，使得在每次外层迭代  $k$  时满足  $\|d_k\| = \Delta$ 。随后，GHM 关联的对偶变量  $\theta_k$  扮演与信赖域子问题中的对偶变量相同的角色。我们给出以下结果，以描述内层迭代次数  $\mathcal{T}_k$  的上界。

#### 定理 4.14: 利用 GHM 恢复信赖域方法

在每次迭代  $k$ ，对于任意半径  $\Delta$ ，通过适当选择  $\delta_k$ ，解  $[v_k; t_k]$  也能解信赖域子问题。此外，搜索所需  $\delta_k$  的内迭代  $\mathcal{T}_k$  通过二分法最多需要  $O(\log(\epsilon^{-1}))$  步。

该定理的证明见 [第 4.C.1 节](#)。利用参数本征值问题求解 TRS 的方法已在<sup>[4,146]</sup>中探讨。然而，这些方法仅限于求解子问题 (4-16)，而未涉及在迭代算法中改进解。为了实现最先进的收敛性，该方法可以使用<sup>[41,43]</sup>中关于  $\Delta$  的更新规则。

利用 GHM 恢复梯度正则化方法 梯度正则化牛顿方法最近在<sup>[50,115]</sup>中用于凸优化，其步长定义如下：

$$x_{k+1} = x_k - (H_k + \gamma_k \|g_k\|^{1/2})^{-1} g_k, \quad (4-17)$$

其中  $\gamma_k > 0$ 。类似地，只要  $t_k \neq 0$ ，则  $d_k = v_k/t_k$  形成一个正则化牛顿方向。一个直接的方法是选择合适的  $\delta_k$  使得  $\theta_k \approx \Theta(\|g_k\|^{1/2})$ ，从而建立等价性。我们总结如下定理：

**定理 4.15: 利用 GHM 恢复正则化牛顿方法**

对于任意  $\gamma_k > 0$ ，通过适当选择  $\delta_k$ ，解  $[v_k; t_k]$  生成与 (4-17) 相同的迭代。此外，寻找所需  $\delta_k$  的内迭代  $\mathcal{T}_k$  通过二分法最多需要  $O(\log(\epsilon^{-1}))$  步，而使用牛顿法最多需要  $O(\log \log(\epsilon^{-1}))$  步。

该定理的证明见 第 4.C.2 节。我们不进一步探讨专门的变体，因为它们超出了本文的讨论范围。广义而言，任何牛顿型步长，无论是正则化还是信赖域方法，都可以通过一系列 GHM 方法恢复，前

表 4.2.2 利用齐次框架通过自适应  $\delta_k, \phi_k$  来恢复或替代其他二阶方法。内循环复杂度  $\mathcal{T}_k$  表示与外部迭代  $k$  相关的内部迭代次数上界。

具体二阶方法	内循环复杂度 $\mathcal{T}_k$	详细讨论
信赖域方法 <sup>[41,43]</sup>	$O(\log(1/\epsilon))$	见 <sup>[146]</sup> 及 定理 4.14
正则化牛顿方法 <sup>[50,115]</sup>	$O(\log \log(1/\epsilon))$	见 定理 4.15
自适应 HSODM	$O(\log(1/\epsilon))$	见 第 5 章
同伦 HSODM	$\leq 2$	见 第 6 章

提是采用合理的策略来选择  $\delta_k$  (以及可能的  $\phi_k$ )。此外，内部迭代  $\mathcal{T}_k$  (定位  $\delta_k$ ) 的复杂度依赖于特定齐次子问题的原对偶解的计算。

在此背景下，我们讨论了在二阶 Lipschitz 连续性下用于非凸优化的自适应 HSODM 方法。该方法自然地扩展到凸优化，以补充原始 HSODM，后者仅涉及带线搜索的非凸情形。在自适应 HSODM 中，对偶变量  $\theta_k$  仍可解释为正则化项 (见 第 5 章 中的局部模型 (5-3))，并且内部迭代次数  $\mathcal{T}_k$  典型地满足  $O(\log(1/\epsilon))$ 。

此外，我们在 第 6 章 中提出了一种新型同伦方法，该方法在某些结构化的非强凸优化中具有  $O(\log(\epsilon^{-1}))$  的迭代复杂度，并且每次迭代  $x_k$  的内部迭代次数满足  $\mathcal{T}_k \leq 2$ 。为便于理解，我们在 表 4.2.2 中总结了上述讨论。

## 本章附录

### 第一节 第 4.2.1 节的主要证明

#### 4.A.1 技术引理

我们引入几个技术引理。

##### 引理 4.16: 半正定矩阵的伪逆

若  $A \in \mathbb{R}^{n \times n}$ ,  $A \geq 0$ , 则其伪逆满足  $A^\star \geq 0$ 。

*Proof.* 取  $A$  的谱分解  $A = UDU^T$ , 假设其有  $d$  个严格正的特征值。则有  $A^\star = U_+ D_+^{-1} U_+^T$ , 其中  $D_+ \in \mathbb{R}^{n \times d}$ ,  $U_+ \in \mathbb{R}^{n \times d}$  分别为严格正的特征值及其对应特征向量。因此,  $A^\star \geq 0$ 。  $\square$

##### 引理 4.17: 严格负曲率

设  $A \in \mathbb{R}^{n \times n}$ ,  $b \in \mathbb{R}^n$ , 以及  $c \in \mathbb{R}$ , 定义聚合矩阵:

$$B = \begin{bmatrix} A & b \\ b^T & c \end{bmatrix}$$

若满足以下任意条件之一, 则  $\lambda_1(B) < 0$ :

(a)  $c \leq 0$  或  $\lambda_1(A) < 0$ 。

(b)  $\lambda_1(A) = 0$  且  $b \notin \mathcal{S}_1$ , 其中  $\mathcal{S}_1$  是  $\lambda_1(A)$  对应特征向量所张成的子空间。

*Proof.* 若  $c < 0$  或  $\lambda_1(A) < 0$ , 结论显然成立。

现在考虑  $c = 0$  的情况, 对于任意满足  $v \in \mathbb{R}^n, t \neq 0$  的向量  $\xi = [\eta \cdot v; t] \in \mathbb{R}^{n+1}$ , 我们有:

$$R_B([\eta \cdot v; t]) := \begin{bmatrix} \eta v \\ t \end{bmatrix}^T B \begin{bmatrix} \eta v \\ t \end{bmatrix} = \eta^2 \cdot (v^T A v) + 2\eta \cdot (b^T v \cdot t).$$

易见, 存在  $v, t$  使得  $b^T v \cdot t \neq 0$ , 因此存在  $\eta$  使得  $R_B([\eta \cdot v; t]) < 0$ , 这意味着  $\xi$  对应一个负曲率方向。

最后考虑  $\lambda_1(A) = 0$  且  $b \notin \mathcal{S}_1$  的情况。对于任意非零向量  $u \in \mathcal{S}_1$ , 有  $b^T u \neq 0$ 。类似地, 对于

向量  $[u; \eta t]$ , 其中  $t \neq 0$ , 我们有:

$$\begin{aligned} R_B([u; \eta t]) &:= \begin{bmatrix} u \\ \eta t \end{bmatrix}^T B \begin{bmatrix} u \\ \eta t \end{bmatrix} = (u^T A u) + 2\eta \cdot (b^T u \cdot t) + \eta^2 \cdot ct^2 \\ &= 2\eta \cdot (b^T u \cdot t) + \eta^2 \cdot ct^2. \end{aligned}$$

由于  $b^T u \cdot t \neq 0$ , 因此存在  $\eta$  使得  $R_B([u; \eta t]) < 0$ , 从而完成证明.  $\square$

#### 4.A.2 基本结果

##### 定理 4.18: 特征向量计算

对于聚合矩阵  $F_k$ , Lanczos 方法需要

$$\min \left\{ O \left( \sqrt{\frac{\|F_k\|}{\varepsilon}} \log \left( \frac{n+1}{\delta} \right) \right), O \left( \sqrt{\frac{\|F_k\|}{\lambda_2(F_k) - \lambda_1(F_k)}} \log \left( \frac{n+1}{\varepsilon \cdot \delta} \right) \right) \right\}$$

次迭代, 以计算满足  $|\xi - \lambda_1(F_k)| \leq \varepsilon$  的估计  $\xi$ , 其概率为  $1 - \delta$ .

*Proof.* 记  $\lambda_{r+1}$  为最大特征值 (由于  $F_k$  维数增加 1). 计算  $F_k$  的最小特征值等价于计算  $B_k \cdot I - F_k$  的最大特征值, 其中  $B_k \geq \lambda_{r+1}(F_k)$ . 根据<sup>[103]</sup> Theorem 4.2 (a), 第  $k$  次 Lanczos 迭代产生的估计  $-\xi$  满足:

$$\mathbb{P} \left\{ \frac{|\lambda_{r+1}(B_k \cdot I - F_k) - (-\xi)|}{(\lambda_{r+1}(B_k \cdot I - F_k) - \lambda_1(B_k \cdot I - F_k))} \geq \tilde{\varepsilon} \right\} \leq 1.648\sqrt{n+1} \exp\{-\sqrt{\tilde{\varepsilon}} \cdot (2k-1)\}.$$

取右侧上界为  $\delta$ , 设  $\tilde{\varepsilon} = \varepsilon / (\lambda_{r+1}(F_k) - \lambda_1(F_k))$ , 可得:

$$\lambda_{r+1}(B_k \cdot I - F_k) - \lambda_1(B_k \cdot I - F_k) = \lambda_{r+1}(F_k) - \lambda_1(F_k),$$

$$\lambda_{r+1}(B_k \cdot I - F_k) - (-\xi) = \xi - \lambda_1(F_k).$$

并且有  $\lambda_{r+1}(F_k) - \lambda_1(F_k) \leq 2\|F_k\|$ .

根据<sup>[103]</sup> Theorem 4.2 (b), 还可以得到:

$$\begin{aligned} \mathbb{P} \left\{ \frac{|\lambda_{r+1}(B_k \cdot I - F_k) - (-\xi)|}{(\lambda_{r+1}(B_k \cdot I - F_k) - \lambda_1(B_k \cdot I - F_k))} \geq \tilde{\varepsilon} \right\} \\ \leq 1.648\sqrt{n+1} \left( \frac{1 - \sqrt{(\lambda_{r+1} - \lambda_r)/(\lambda_{r+1} - \lambda_1)}}{1 + \sqrt{(\lambda_{r+1} - \lambda_r)/(\lambda_{r+1} - \lambda_1)}} \right)^{k-1}. \end{aligned}$$

由此可得:

$$k \geq 1 + \log(1.648\sqrt{n+1}/(\delta\sqrt{\tilde{\varepsilon}}))/\log \left( \frac{1 + \sqrt{(\lambda_{r+1} - \lambda_r)/(\lambda_{r+1} - \lambda_1)}}{1 - \sqrt{(\lambda_{r+1} - \lambda_r)/(\lambda_{r+1} - \lambda_1)}} \right).$$

进一步观察  $\log(1 + 1/t) \geq \frac{1}{1/2+t}$ , 以及

$$\lambda_{r+1}(B_k \cdot I - F_k) - \lambda_r(B_k \cdot I - F_k) = \lambda_2(F_k) - \lambda_1(F_k),$$



从而得出第二个估计式。 □

括号内的第一个估计通常称为“无间隙”复杂度 (gap-free complexity)，而第二个称为“依赖间隙”复杂度 (gap-dependent complexity)。在非凸优化问题中，由于 Hessian 的谱沿实数轴分布，我们只能依赖“无间隙”保证。在前一章节，已经严格证明了求解 GHM 的“近似”解的复杂度与获得近似牛顿方向的复杂度相匹配<sup>[44,147]</sup>。

**引理 4.19:  $F_k$  的估计**

考虑聚合矩阵  $F_k$ ，则有：

$$\lambda_2(F_k) - \lambda_1(F_k) \geq \lambda_1(H_k) + \max_j \left\{ \frac{-\lambda_j(H_k) - \delta_k + \sqrt{(\lambda_j(H_k) - \delta_k)^2 + 4\|P_{\mathcal{S}_j}(g_k)\|^2}}{2} \right\}$$

$$\lambda_{r+1}(F_k) \leq \lambda_r(H_k) + \delta_k - \lambda_1(F_k)$$

*Proof.* 根据 引理 4.1 中的二阶条件 (4-3)，我们有：

$$H_k + \theta_k I - \frac{g_k g_k^T}{\theta_k + \delta_k} \geq 0,$$

由 Schur 补公式可得。投影到  $\mathcal{S}_j(H_k)$ ,  $j = 1, \dots, d$  上，得到：

$$\theta_k + \lambda_j(H_k) - \frac{\|P_{\mathcal{S}_j}(g_k)\|^2}{\theta_k + \delta_k} \geq 0, \forall j$$

整理后可得：

$$\theta_k^2 + (\lambda_j(H_k) + \delta_k)\theta_k + \lambda_j(H_k)\delta_k - \|P_{\mathcal{S}_j}(g_k)\|^2 \geq 0,$$

进而得到：

$$\theta_k \geq \max_j \left\{ \frac{-(\lambda_j(H_k) + \delta_k) + \sqrt{(\lambda_j(H_k) - \delta_k)^2 + 4\|P_{\mathcal{S}_j}(g_k)\|^2}}{2} \right\}.$$

由 Cauchy 交错定理，可得：

$$\begin{aligned} \lambda_2(F_k) - \lambda_1(F_k) &= \lambda_2(F_k) + \theta_k \geq \lambda_1(H_k) + \theta_k \\ &\geq \max_j \left\{ \frac{2\lambda_1(H_k) - (\lambda_j(H_k) + \delta_k) + \sqrt{(\lambda_j(H_k) - \delta_k)^2 + 4\|P_{\mathcal{S}_j}(g_k)\|^2}}{2} \right\}. \end{aligned}$$

注意：

$$\text{tr}(F_k) = \sum_{j=1}^{r+1} \lambda_j(F_k) = \sum_{j=1}^r \lambda_j(H_k) + \delta_k.$$

由此可得：

$$\begin{aligned}
 \lambda_{\max}(F_k) &:= \lambda_{r+1}(F_k) = \sum_{j=2}^r (\lambda_j(H_k) - \lambda_j(F_k)) + \lambda_1(H_k) - \lambda_1(F_k) + \delta_k \\
 &\leq \sum_{j=2}^r (\lambda_j(H_k) - \lambda_{j-1}(H_k)) + \delta_k + \lambda_1(H_k) - \lambda_1(F_k) \\
 &= \lambda_r(H_k) - \lambda_1(H_k) + \delta_k + \lambda_1(H_k) - \lambda_1(F_k) \\
 &= \lambda_r(H_k) + \delta_k - \lambda_1(F_k) \\
 &= \lambda_r(H_k) + \delta_k + \theta_k.
 \end{aligned}$$

□

#### 4.A.3 对定理 4.2 的证明

*Proof.* 对于部分 (a)，我们参考引理 4.19 并设  $\delta_k = -\epsilon_L$ 。先注意到  $\forall \eta \leq 1/2\sqrt{n}$ ，存在  $j'$  使得  $\|P_{\mathcal{S}_{j'}}(g_k)\|^2 \geq \eta^2 \|g_k\|^2$ ，否则：

$$\sum_{j=1}^r \|P_{\mathcal{S}_j}(g_k)\|^2 < \sum_{j=1}^r \eta^2 \|g_k\|^2 \leq \frac{\|g_k\|^2}{4},$$

这与  $\bigcup_j \mathcal{S}_j = \mathbb{R}^n$  矛盾。因此，对于  $\forall \eta \in [0, 1/2\sqrt{n}]$ ，有：

$$\lambda_2(F_k) - \lambda_1(F_k) \geq \lambda_1(H_k) + \frac{-\lambda_{j'}(H_k) + \epsilon_L + \sqrt{(\lambda_{j'}(H_k) + \epsilon_L)^2 + 4\eta^2 \|g_k\|^2}}{2} \quad (4-18a)$$

$$\geq \lambda_1(H_k) + \frac{-U_H + \epsilon_L + \sqrt{(U_H + \epsilon_L)^2 + 4\eta^2 \|g_k\|^2}}{2} > 0. \quad (4-18b)$$

由于一维函数

$$\ell(x) := \frac{-x + \epsilon_L + \sqrt{(x + \epsilon_L)^2 + 4\eta^2 \|g_k\|^2}}{2}$$

在  $x > 0$  上单调递减，且  $\lambda_{j'}(H_k) \leq U_H$ ，由定义可得：

$$\kappa_L(F_k) \leq \frac{2(\lambda_r(H_k) - \epsilon_L - \lambda_1(F_k))}{-U_H + \epsilon_L + \sqrt{(U_H + \epsilon_L)^2 + \|g_k\|^2/n}} < \infty$$

其中  $\lambda_1(F_k)$  有界。

对于部分 (a)，我们观察到：

$$\begin{aligned}
 \frac{\kappa_L(F_k)}{\kappa(H_k + \epsilon_N I)} &\leq \frac{2\|F_k\|}{\lambda_2(F_k) - \lambda_1(F_k)} \cdot \frac{\lambda_1(H_k) + \epsilon_N}{\lambda_r(H_k) + \epsilon_N} \\
 &= \frac{2\|F_k\|}{\epsilon_N + U_H} \cdot (\lambda_1(H_k) + \epsilon_N) \cdot \underbrace{\frac{1}{\lambda_2(F_k) - \lambda_1(F_k)}}_{\ddagger}.
 \end{aligned}$$

对于  $\ddagger$ ，我们有：

$$\lambda_2(F_k) - \lambda_1(F_k) \geq \epsilon_L + \frac{-(U_H + \epsilon_L) + \sqrt{(U_H + \epsilon_L)^2 + 4\eta^2 \|g_k\|^2}}{2}$$

$$= \epsilon_L + \frac{U_H + \epsilon_L}{2} \cdot \left( \sqrt{1 + \left( \frac{2\eta \|g_k\|}{U_H + \epsilon_L} \right)^2} - 1 \right).$$

令  $\Gamma = \frac{2\eta \|g_k\|}{U_H + \epsilon_L}$ , 则:

$$\lambda_2(F_k) - \lambda_1(F_k) \geq \epsilon_L + \frac{U_H + \epsilon_L}{2} \cdot \left( \frac{\Gamma^2}{2} - \frac{\Gamma^4}{8} + O(\Gamma^6) \right).$$

由于  $\Gamma = \Theta(\|g_k\|)$  且  $U_H + \epsilon_L$  可视为常数, 可推出:

$$\frac{\kappa_L(F_k)}{\kappa(H_k + \epsilon_N I)} \leq O\left(\frac{\epsilon_N}{\frac{\|g_k\|^2}{U_H + \epsilon_L} + \epsilon_L}\right).$$

证毕。

□

## 第二节 第 4.2.2 节的主要证明

下面的引理是<sup>[146]</sup>的 Lemma 3.3 的一个略微重组版本。

### 引理 4.20

设  $p_j, \tilde{\alpha}_j, 1 \leq j \leq r$  的定义与 引理 4.6 相同, 并假定  $\phi_k \perp \mathcal{S}_i(H_k)$ , 其中  $i = 1, 2, \dots, \ell$ , 且  $1 \leq \ell < r$ , 则:

- (a) 若  $\delta_k = \tilde{\alpha}_j, 1 \leq j \leq \ell$ , 则  $\lambda_j(F_k) = \lambda_j(H_k), j = 1, 2, \dots, \ell$ ;
- (b) 若  $\delta_k < \tilde{\alpha}_1$ , 则  $\lambda_1(F_k) < \lambda_1(H_k)$  且  $\lambda_j(F_k) = \lambda_{j-1}(H_k), j = 2, \dots, \ell + 1$ ;
- (c) 若  $\tilde{\alpha}_{i-1} < \delta_k < \tilde{\alpha}_i, 2 \leq i \leq \ell$ , 则  $\lambda_j(F_k) = \lambda_j(H_k)$  对于  $j = 1, \dots, i-1$  成立, 并且  $\lambda_j(F_k) = \lambda_{j-1}(H_k)$  对于  $j = i+1, \dots, \ell+1$  成立;
- (d) 若  $\delta_k > \tilde{\alpha}_\ell$ , 则  $\lambda_j(F_k) = \lambda_j(H_k), j = 1, 2, \dots, \ell$ 。

### 4.B.1 对 引理 4.5 的证明

*Proof.* 回忆  $\theta_k$  在 (4-2) 中的定义, 我们有:

$$\begin{aligned} -\theta_k &= \min_{v^T v + t^2 \leq 1} v^T H_k v + 2\phi_k^T v \cdot t + \delta_k t^2 \\ &\geq \min_{v^T v + t^2 \leq 1} v^T H_k v + \delta_k t^2 + \min_{v^T v + t^2 \leq 1} 2\phi_k^T v \cdot t. \end{aligned} \quad (4-19)$$

现在分别分析两个项, 对于第一个项, 有:

$$\min_{v^T v + t^2 \leq 1} v^T H_k v + \delta_k t^2 = \min_{\| [v; t] \| \leq 1} \begin{bmatrix} v \\ t \end{bmatrix}^T \begin{bmatrix} H_k & 0 \\ 0^T & \delta_k \end{bmatrix} \begin{bmatrix} v \\ t \end{bmatrix}$$

$$= \begin{cases} 0 & H_k \geq 0, \delta_k \geq 0 \\ \lambda_1 \left( \begin{bmatrix} H_k & 0 \\ 0^T & \delta_k \end{bmatrix} \right) & \text{否则.} \end{cases}$$

因此, 我们得到:

$$\min_{v^T v + t^2 \leq 1} v^T H_k v + \delta_k t^2 = \min\{\lambda_1, \delta_k, 0\}. \quad (4-20)$$

对于第二项, 使用类似的分析方法, 可以得出:

$$\min_{v^T v + t^2 \leq 1} 2\phi_k^T v \cdot t = \lambda_1 \left( \begin{bmatrix} 0 & \phi_k \\ \phi_k^T & 0 \end{bmatrix} \right) = -\|\phi_k\|. \quad (4-21)$$

将 (4-20) 和 (4-21) 代入 (4-19), 即可完成证明。□

#### 4.B.2 对引理 4.8 的证明

*Proof.* 由于  $H_k - \lambda_1(H_k)I \geq 0$ , 根据引理 4.16, 有  $(H_k - \lambda_1(H_k)I)^\star \geq 0$ , 因此:

$$\tilde{\alpha}_1 - \lambda_1(H_k) = -\phi_k^T p_1 = \phi_k^T (H_k - \lambda_1(H_k)I)^\star \phi_k \geq 0.$$

证毕。□

#### 4.B.3 对引理 4.10 的证明

*Proof.* 我们分以下两种情况讨论。

**情况 (1)**  $\phi_k \notin \mathcal{S}_1(H_k)$ 。在这种情况下, 对于所有  $\delta_k, t_k \neq 0$ 。最优性条件 (4-4) 退化为类似 Newton 方程的形式, 因为  $d_k = v_k/t_k$  是良定义的:

$$(H_k + \theta_k I) d_k = -\phi_k, \quad -\phi_k^T d_k = \theta_k + \delta_k.$$

根据二阶条件 (4-3), 可得:

$$\begin{bmatrix} H_k + \theta_k I & \phi_k \\ \phi_k^T & \theta_k + \delta_k \end{bmatrix} \geq 0,$$

这意味着  $H_k + \theta_k I > 0$ , 否则由引理 4.17 可知, 始终存在负曲率。因此, 逆矩阵是良定义的:

$$d_k = -(H_k + \theta_k I)^{-1} \phi_k, \quad (4-22)$$

由此得到:

$$\theta_k + \delta_k = -d_k^T \phi_k = \phi_k^T (H_k + \theta_k I)^{-1} \phi_k.$$

两边求导:

$$d\theta_k(\theta_k + \delta_k) = d\theta_k \left( \phi_k^T (H_k + \theta_k I)^{-1} \phi_k \right)$$

$$\begin{aligned}\Rightarrow 1 + d\theta_k \delta_k &= -\phi_k^T (H_k + \theta_k I)^{-2} \phi_k \stackrel{(4-22)}{=} -\Delta_k \\ \Rightarrow d\delta_k \theta_k &= -\frac{1}{\Delta_k + 1}.\end{aligned}$$

因此，我们得到 (4-15)，这意味着  $\theta_k$  在此情况下是连续且递减的。

关于  $\theta_k$  的凸性，首先观察到  $\|d_k\|$  随着  $\theta_k$  的增加而递减，因为  $\|d_k\| = \|(H_k + \theta_k I)^{-1} \phi_k\|$ 。由于  $\delta_k \nearrow, \theta_k \searrow$ ，因此  $\|d_k\| \nearrow$ ，进而意味着  $d\delta_k \theta_k \nearrow$ ，由此得出  $\theta_k$  在  $t_k \neq 0$  时是凸的。根据 引理 4.6， $\phi_k \perp \mathcal{S}_1(H_k)$  时， $t_k \neq 0$  对所有  $\delta_k \in \mathbb{R}$  成立。综上， $\theta_k$  是递减的、连续的、凸的且可微的。由于  $\theta_k \geq 0$ ，同样适用于  $\omega_k$ 。

**情况 (2)**  $\phi_k \perp \mathcal{S}_1(H_k)$ 。在此情况下，我们分三种子情况讨论：

(i)  $\lambda_1(H) > 0$ 。在此情况下，根据 Schur 补公式，当  $\delta_k > \phi_k^T H_k^{-1} \phi_k$  时，可得  $\lambda_1(F_k) > 0$ ，因此  $\theta_k = 0$ 。此外，注意到：

$$\phi_k^T H_k^{-1} \phi_k < \tilde{\alpha}_1 := \phi_k^T (H_k - \lambda_1(H_k)I)^* \phi_k,$$

这意味着在  $\delta_k = \tilde{\alpha}_1$  之前， $\theta_k$  早已变为零。这表明，当  $\delta_k \leq \phi_k^T H_k^{-1} \phi_k$  时， $t_k \neq 0$ ，进而由 **情况 (1)** 可得  $\theta_k$  是凸的和连续的。当  $\delta_k > \phi_k^T H_k^{-1} \phi_k$  时， $\theta_k = 0$  为常数。

(ii)  $\lambda_1(H) = 0$ 。根据 引理 4.17，我们知道  $F_k$  始终存在严格负曲率，这意味着  $\theta_k = -\lambda_1(F_k)$  对所有  $\delta_k$  成立。首先，根据 推论 4.7，对于任何  $\delta_k \leq \tilde{\alpha}_1$ ，始终存在  $t_k \neq 0$ 。结合 **情况 (1)** 的结果，我们得出  $\theta_k$  在  $\delta_k \leq \tilde{\alpha}_1$  时是凸的和连续的。此外，对于  $\delta_k \geq \tilde{\alpha}_1$ ，有  $\theta_k = -\lambda_1(F_k) = -\lambda_1(H_k)$ 。因此， $-\theta_k$  对所有  $\delta_k$  是凸的和连续的。

(iii)  $\lambda_1(H) < 0$ 。这一情况与上一个类似，省略证明，结论是  $\theta_k = -\lambda_1(F_k)$  对所有  $\delta_k$  成立， $\theta_k$  在  $\delta_k \leq \tilde{\alpha}_1$  时是凸的和连续的，且当  $\delta_k \geq \tilde{\alpha}_1$  时， $\theta_k = -\lambda_1(F_k) = -\lambda_1(H_k)$ 。

□

#### 4.B.4 对 引理 4.12 的证明

*Proof.* 对于  $\phi_k \perp \mathcal{S}_1(H_k)$  的情况，若  $\delta_k = \tilde{\alpha}_1$ ， $\Delta_k$  不连续，因此我们可得出  $h_k$  也是不连续的。

若  $\phi_k \not\perp \mathcal{S}_1(H_k)$ ，根据最优性条件 (4-3) 和 引理 4.17，有：

$$H_k + \theta_k I > 0,$$

于是我们对  $H_k$  进行谱分解，并记作：

$$H_k = \sum_{i=1}^r \lambda_i(H_k) v_i v_i^T, \quad \beta_i = \phi_k^T v_i, \quad i = 1, \dots, d. \quad (4-23)$$

因此, 我们可以将  $h_k(\delta)$  重写为:

$$\begin{aligned} h_k(\delta) &= \frac{\theta_k^2}{\|d_k\|^2} = \frac{\theta_k^2}{\phi_k^T (H_k + \theta_k I)^{-2} \phi_k} \\ &= \frac{\theta_k^2}{\sum_{i=1}^n \frac{\beta_i^2}{(\lambda_i + \theta_k)^2}}, \end{aligned} \quad (4-24)$$

因此函数  $h_k(\delta)$  可视为  $\omega_k$  的复合函数。回忆 [引理 4.10](#),  $\omega_k(\delta)$  是可微的, 因此  $h_k(\delta)$  也是可微的。根据链式法则及 [引理 4.10](#) 的结果, 我们可以得出:

$$\begin{aligned} h'_k(\delta) &= d\theta h_k(\delta) d\delta\theta \\ &= -\frac{2\theta\|d\|^2 + \theta^2 d^T (H_k + \theta I)^{-1} d}{\|d\|^2} \frac{1}{1 + \|d\|^2} \\ &= -\frac{2\theta\|d\|^2 + \theta^2 d^T (H_k + \theta I)^{-1} d}{\|d\|^2 (1 + \|d\|^2)}, \end{aligned} \quad (4-25)$$

这表明  $h_k(\delta)$  是单调递减的。  $\square$

### 第三节 第 4.2.3 节的主要证明

#### 4.C.1 对 [定理 4.14](#) 的证明

*Proof.* 对于信赖域法, 当  $g_k \perp \mathcal{S}_1(H_k)$  时, 通常需要单独处理。为简洁起见, 我们仅考虑  $g_k \not\perp \mathcal{S}_1(H_k)$  的情况。设  $([v_k; t_k], \theta_k)$  为 GHM 子问题的原对偶解, 由 [推论 4.7](#) 可得对于所有  $\delta_k \in \mathbb{R}$ , 都有  $t_k \neq 0$ , 从而  $\Delta_k(\delta_k) = \|v_k/t_k\|^2$  是连续的。因此, 我们可以在 [第 5.3 节](#) 中采用类似的二分法找到合适的  $\delta_k$  使得

$$\Delta_k(\delta_k) = \Delta^2.$$

结合 (4-3) 中的最优性条件, 设  $d_k = v_k/t_k$ , 可得

$$\begin{aligned} H_k + \theta_k \cdot I &\geq 0, \\ (H_k + \theta_k \cdot I)d_k &= -g_k, \\ \theta_k(\|d_k\| - \Delta) &= 0, \end{aligned}$$

这表明  $(d_k, \theta_k)$  是信赖域子问题 (4-16) 的解。通过选择合适的  $\delta_k$ , 我们始终可以利用  $[v_k; t_k]$  还原信赖域子问题 (4-16) 的解。由于我们已采用自适应 HSODM, 这里仅简要讨论  $\delta_k$  的区间搜索阶段。

由于  $g_k \not\perp \mathcal{S}_1(H_k)$ , 由 [引理 4.11](#) 知  $\Delta_k$  在  $\delta_k$  上是连续的。因此, 我们的关键目标是找到一个区间  $(\delta_{low}, \delta_{up})$  使得

$$\Delta^2 \in [\Delta_k(\delta_{low}), \Delta_k(\delta_{up})].$$

我们主要分析  $\Delta < 1$  的情况, 因为  $\Delta \geq 1$  的情况类似。

首先, 由 [引理 5.20](#), 可取  $\delta_{up} = \lambda_r(H_k)$ , 则  $\Delta^2 \leq \Delta_k(\delta_{up})$ 。

接下来，我们估计  $\delta_{low}$ 。实际上，不失一般性，我们可以假设  $\lambda_1(H_k) \leq 0$  并设  $\delta_{low} = \lambda_1(H_k) - \|\phi_k\| \frac{1-\Delta^2}{\Delta^2}$ 。当  $\lambda_1(H_k) > 0$  时，仅需做轻微调整。我们将在下文证明  $\Delta^2 \geq \Delta_k(\delta_{low})$ 。

设  $\delta_k = \delta_{low}$ ，类似于引理 5.20，我们有

$$\begin{aligned} -\theta_k &= v_k^T H_k v_k + 2t_k \phi_k^T v_k + \delta_k t_k^2 \\ &\geq \lambda_1(H_k) \|v_k\|^2 + 2t_k \phi_k^T v_k + \delta_k t_k^2 \\ &= \lambda_1(H_k) \|v_k\|^2 - 2t_k^2 (\theta_k + \delta_k) + \delta_k t_k^2 \\ &= (\delta_k + \|\phi_k\| \frac{1-\Delta^2}{\Delta^2}) \|v_k\|^2 - 2t_k^2 (\theta_k + \delta_k) + \delta_k t_k^2 \\ &= \delta_k - 2t_k^2 (\theta_k + \delta_k) + \|\phi_k\| \frac{1-\Delta^2}{\Delta^2} (1-t_k^2), \end{aligned}$$

进而可得

$$(2t_k^2 - 1)(\theta_k + \delta_k) \geq \|\phi_k\| \frac{1-\Delta^2}{\Delta^2} (1-t_k^2).$$

另外，由引理 4.5 可知

$$(2t_k^2 - 1)\|\phi_k\| \geq \|\phi_k\| \frac{1-\Delta^2}{\Delta^2} (1-t_k^2),$$

重新整理可得  $\Delta_{low} \leq \Delta^2$ ，从而证明成立。

对于  $g_k \perp \mathcal{S}_1(H_k)$  的情况，我们参考<sup>[146]</sup>第3节中的不同处理方式，以应对信赖域子问题的困难情况。 □

#### 4.C.2 对定理 4.15 的证明

*Proof.* 为了通过 GHM 还原正则化牛顿法，我们首先显式设定  $\phi_k = g_k$ 。由于正则化牛顿法适用于凸目标函数，我们需要谨慎选择  $\delta_k$  以确保  $\theta_k(\delta_k)$  严格为正值，这也意味着根据引理 4.8 的排序结果， $t_k \neq 0$ 。由引理 4.3，可得  $\theta_k = 0$  当且仅当

$$H_k \geq 0, (I - H_k H_k^\star) g_k = 0, \delta_k \geq \overline{\delta_k^{cvx}} = g_k^T H_k^\star g_k \geq 0.$$

因此，在每次迭代中，我们划分以下两种情况：

- (a)  $(I - H_k H_k^\star) g_k = 0$ 。在此情况下，我们需要保持  $\delta_k < \overline{\delta_k^{cvx}}$ ，从而  $F_k$  是不定的。另外，由于  $t_k \neq 0$ ，我们有  $\theta_k > -\delta_k$ ，这说明搜索合适的  $\delta_k$  时应采用以下区间：

$$[-2\gamma_k \|g_k\|^{1/2}, \overline{\delta_k^{cvx}}), \forall \gamma_k > 0.$$

结合  $\theta_k$  的连续性(引理 4.10)，可知  $\forall \gamma_k > 0$ ， $\exists \delta_k$  使得  $\theta_k(\delta_k) = \gamma_k \|g_k\|^{1/2}$ 。计算  $\delta_k$  可采用多种方法。首先，我们可以使用类似于第 5.3 节的可证明二分法，在  $O(\log(\epsilon^{-1}))$  时间内求解。

其次, 由于  $t_k \neq 0$ ,  $\omega_k = \theta_k^2$  具有已知导数 (参见 (4-15)), 这意味着牛顿法<sup>[164]</sup> 也是一种可行选择, 时间复杂度为  $O(\log \log(\epsilon^{-1}))$ 。

采用上述任一方法, HSODM 计算出的步长与正则化牛顿法完全一致:

$$d_k = -(H_k + \theta_k \cdot I)^{-1} g_k = -\left(H_k + \gamma_k \|g_k\|^{1/2}\right)^{-1} g_k$$

其中  $\theta_k(\delta_k) = \gamma_k \|g_k\|^{1/2}$ 。

(b)  $(I - H_k H_k^\star)g_k \neq 0$ 。在此情况下,  $F_k$  始终是不定的, 因此对于所有  $\delta_k$ , 我们有  $\theta_k > 0$ , 即可能需要  $\delta_k \rightarrow \infty$ 。然而, 一个简单的修正方案即可解决问题。由于  $H_k \geq 0$ , 我们有:

$$\tilde{H}_k := H_k + \frac{1}{2}\gamma_k \|g_k\|^{1/2} > 0.$$

由此, 我们直接构造基于  $\tilde{H}_k$  的 GHM, 并找到  $\theta_k(\delta_k) = \frac{1}{2}\gamma_k \|g_k\|^{1/2}$ 。由于此时满足  $(I - \tilde{H}_k \tilde{H}_k^\star)g_k = 0$ , 则问题可回归至情况 (a) 的讨论。

结合以上两种情况, 证明完成。 □

为了实现正则化牛顿法的步长, 并不需要在每次迭代时检查  $(I - H_k H_k^\star)g_k = 0$  是否成立。完全可以直接采用情况 (b) 中的策略。然而, 这一发现更具有理论意义。

此外, 我们意识到<sup>[50,51,115]</sup> 中存在其他技术, 而不仅仅是选择合适的  $\gamma_k$ 。因此, 如何设计高效的方法来实现正则化牛顿法仍是一个值得研究的方向。



## 第五章 自适应的齐次二阶法

### 第一节 方法概述

在本节中，我们考虑 HSODF 的一种特定实现，称为“自适应 HSODM”。我们的直觉来自于<sup>[128]</sup>的既定结果和最近的专著<sup>[29]</sup>。我们目标不变，仍然是寻找近似二阶稳定点 (SOSP) (2-2)；与 HSODM 类似，我们考虑一般二阶 Lipschitz 连续的函数 (3-17)。这些定义我们不再复述。

我们进一步对 GHMs 中每个迭代点  $x_k$  处的函数  $\phi_k(x_k)$  做出以下假定。

#### 假设 5.1

假设存在一个统一常数  $\varsigma_\phi > 0$ 。对于一个迭代点  $x_k \in \mathbb{R}^n$ ，如果  $t_k \neq 0$ ，则

$$\|\phi_k(x_k) - g_k\| \leq \varsigma_\phi \|d_k\|^2 \quad (5-1)$$

其中  $d_k = v_k/t_k$ 。

这一条件的目的是确保  $\phi_k$  相对于方向  $d_k$  的范数保持与梯度  $g_k$  充分接近。值得注意的是，这一假定并不具有约束性。例如，如果我们简单地选择  $\phi_k = g_k$ ，则 (5-1) 的左侧变为零，从而自动满足条件。实际上，这种选择在大多数情况下可能已足够。现在我们准备在 算法 5-1 中介绍适用于二阶 Lipschitz 连续函数的自适应 HSODM。

自适应框架主要受到<sup>[27,28]</sup>的启发。我们在迭代  $k$  期间采用了一个众所周知的比率测试，计算比率  $\rho_k = \frac{f(x_k+d_k)-f(x_k)}{m_k(d_k)-f(x_k)}$ ：如果比率  $\rho_k \geq \eta_2$ ，我们称之为 **非常成功的迭代**；如果  $\eta_1 \leq \rho_k \leq \eta_2$ ，则称为 **成功的迭代**；否则称为 **不成功的迭代**。根据比率测试，我们自适应地构建了一个期望区间  $I_h$  用于  $\sqrt{h_k}$ ，该区间作为以下三次模型的正则化系数，

$$m_k(d) := f(x_k) + \phi_k^T d + \frac{1}{2} d^T H_k d + \frac{\sqrt{h_k(\delta_k)}}{3} \|d\|^3, \text{ 其中 } h_k(\delta_k) := \frac{\theta_k^2}{\|d_k\|^2}. \quad (5-3)$$

注意，我们使用  $\phi_k$  来替代一阶近似。与之前一样，我们利用 GHM 来求解 (5-3)。以下两种情况通过  $t_k$  的值 (见 行 5) 加以区分。

如果  $t_k \neq 0$ ，则  $h_k$  是连续且光滑的，因此 (5-2) 的解  $d_k$  最小化 (5-3)。为了将后验  $\sqrt{h_k(\delta_k)} \in I_h$ ，我们依赖于基于内部循环的二分法来解决标记为  $j$  的 GHMs。在定位  $\sqrt{h_k} \in [\sqrt{\ell}, \sqrt{\nu}]$  时允许存在一个容差  $\sigma$ ，例如  $\sqrt{h_k} \in [\sqrt{\ell}, \sqrt{\nu+\sigma}]$ 。我们将  $\sigma$  设置为自适应的： $\sigma \leq \Omega(h_k)$ ，或者作为一个满足  $\sigma < h_{\min}$  的常数。自适应 HSODM 的整体收敛性对  $\sigma$  的选择不敏感。如前所述，内部迭代次数由 第 4.2.2.2 节 中的分析保证上界为  $O(\log(\epsilon^{-1}))$ 。

如果  $t_k = 0$ ，则在 算法 5-2 中对  $\phi_k$  进行扰动。我们随后将展示这最终会产生一个成功的迭代点，使得主算法 (算法 5-1) 继续进行下一次迭代。为了使我们的表述流畅，我们暂时忽略“困难情况”和二分法，详细内容将在稍后回顾。

**算法 5-1:** 自适应 HSODM

```

1  初始点  $x_0 \in \mathbb{R}^n$ ,  $\delta_0 \in \mathbb{R}$ ,  $I_h = \mathbb{R}$ ,  $h_{\min} > 0$ , 参数  $0 < \eta_1 < \eta_2 < 1$ ,
    $\gamma_1 > 1, \gamma_3 \geq \gamma_2 > 1, 0 < \gamma_4 \leq 1, \sigma > 0$ ;
2  for  $k = 0, 1, 2, \dots$  do
3       $\phi_k = g_k, \delta_{k,0} = \delta_{k-1}$ 
4      for  $j = 0, 1, \dots, \mathcal{T}_k$  do
5          获取子问题的解对  $(\theta_{k,j}, [v_{k,j}; t_{k,j}])$ 
               
$$\min_{\| [v;t] \| \leq 1} \begin{bmatrix} v \\ t \end{bmatrix}^T \begin{bmatrix} H_k & \phi_{k,j} \\ (\phi_{k,j})^T & \delta_{k,j} \end{bmatrix} \begin{bmatrix} v \\ t \end{bmatrix} \quad (5-2)$$

6          if  $t_{k,j} = 0$  then // 检查困难情况, 见 第 5.2.3 节
7              转到 算法 5-2;
8          设置  $d_{k,j} = v_{k,j}/t_{k,j}$ ,  $h_k(\delta_{k,j}) := (\theta_{k,j}/\|d_{k,j}\|)^2$ ;
9          if  $\sqrt{h_k(\delta_k)} \in I_h$  且在容差  $\sigma$  内 then
10             设置  $d_k := d_{k,j}$ ,  $\delta_k = \delta_{k,j}$ ;
11             中断
12         else // 搜索  $\delta_{k,j}$ , 见 第 5.3 节
13             更新  $\delta_{k,j}$ 
14     计算
               
$$\rho_k := \frac{f(x_k + d_k) - f(x_k)}{m_k(d_k) - f(x_k)};$$

15     if  $\rho_k > \eta_2$  then // 非常成功的迭代
16          $I_h = \left[ \max \left\{ \sqrt{h_{\min}}, \gamma_4 \sqrt{h_k(\delta_k)} \right\}, \sqrt{h_k(\delta_k)} \right], x_{k+1} = x_k + d_k$ 
17     if  $\eta_1 \leq \rho_k \leq \eta_2$  then // 成功的迭代
18          $I_h = \left[ \sqrt{h_k(\delta_k)}/\gamma_1, \gamma_2 \sqrt{h_k(\delta_k)} \right], x_{k+1} = x_k + d_k$ 
19     else // 不成功的迭代
20          $I_h = \left[ \gamma_2 \sqrt{h_k(\delta_k)}, \gamma_3 \sqrt{h_k(\delta_k)} \right], x_{k+1} = x_k$ 

```

## 第二节 收敛性分析

### 5.2.1 全局收敛

我们首先引入一些技术性引理，以帮助我们在接受迭代点时识别充分下降量。

为此，我们将 [算法 5-1](#) 中子问题 (5-2) 的最优性条件改写为关于函数  $h_k(\delta_k)$  的形式。

#### 引理 5.2: 子问题 (5-2) 的最优性条件

假定当前迭代点  $x_k$  未落入 [定义 4.13](#) 中定义的困难情况，则子问题 (5-2) 的最优原对偶解  $([v_k; t_k], -\theta_k)$  满足以下条件：

$$\phi_k + H_k d_k + \sqrt{h_k(\delta_k)} \|d_k\| d_k = 0, \quad (5-4a)$$

$$H_k + \sqrt{h_k(\delta_k)} \|d_k\| I \geq 0, \quad (5-4b)$$

其中  $d_k = v_k/t_k$ 。也就是说，若将  $h_k(\delta_k)$  视为常数，则  $d_k$  是 (5-3) 中定义的  $m_k(d)$  的最小化解。

*Proof.* 这一结果可通过将  $\theta_k$  替换为  $\sqrt{h_k(\delta_k)} \|d_k\|$  代入最优性条件得出。  $\square$

我们接下来给出模型下降量的估计。

#### 引理 5.3: 模型下降估计

假定  $x_k$  未落入困难情况，则模型下降满足：

$$f(x_k) - m_k(d_k) \geq \frac{\sqrt{h_k(\delta_k)}}{6} \|d_k\|^3, \quad (5-5)$$

其中  $m_k(\cdot)$  由 (5-3) 定义。

*Proof.* 由以下推导可得：

$$\begin{aligned} f(x_k) - m_k(d_k) &= -\phi_k^T d_k - \frac{1}{2} d_k^T H_k d_k - \frac{\sqrt{h_k(\delta_k)}}{3} \|d_k\|^3 \\ &= \frac{1}{2} d_k^T H_k d_k + \frac{2\sqrt{h_k(\delta_k)}}{3} \|d_k\|^3 \\ &= \frac{1}{2} d_k^T (H_k + \sqrt{h_k(\delta_k)} \|d_k\| I) d_k + \frac{\sqrt{h_k(\delta_k)}}{6} \|d_k\|^3 \\ &\geq \frac{\sqrt{h_k(\delta_k)}}{6} \|d_k\|^3. \end{aligned} \quad (5-6)$$

第二个等式由 (5-4a) 得到，最后一个不等式由 (5-4b) 得到。  $\square$

此外，如果当前迭代点  $x_k$  被接受，则下一步的梯度范数可以得到上界估计。

**引理 5.4: 步长与梯度的关系**

假定  $x_k$  未落入困难情况，且 假设 5.24 成立，则如果第  $k$  次迭代是成功的，则步长  $d_k$  与新点  $x_{k+1}$  处的梯度之间满足以下关系：

$$\|g_{k+1}\| \leq \frac{2\sqrt{h_k(\delta_k)} + M + 2\varsigma_\phi}{2} \|d_k\|^2. \quad (5-7)$$

*Proof.* 对 (5-4a) 取范数可得：

$$\|\phi_k + H_k d_k\| = \sqrt{h_k(\delta_k)} \|d_k\|^2. \quad (5-8)$$

进一步观察：

$$\begin{aligned} \|g_{k+1}\| &\leq \|g_{k+1} - g_k - H_k d_k\| + \|\phi_k - g_k\| + \|\phi_k + H_k d_k\| \\ &\leq \frac{M}{2} \|d_k\|^2 + \sqrt{h_k(\delta_k)} \|d_k\|^2 + \|\phi_k - g_k\|. \end{aligned} \quad (5-9)$$

其中最后一个不等式由二阶 Lipschitz 条件得到。结合 假设 5.24，即得所需结论。  $\square$

结合 引理 5.4 和 引理 5.3，在步长被接受的情况下，我们可以通过梯度或 Hessian 估计函数值的下降量。

**引理 5.5: 函数下降估计**

假定  $x_k$  未落入困难情况，且 假设 5.24 成立，则如果第  $k$  次迭代是成功的，则函数值的下降量满足：

$$f(x_k) - f(x_{k+1}) \geq \eta_1 \frac{\sqrt{h_k(\delta_k)}}{12} \left( \frac{2}{M + 2\sqrt{h_k(\delta_k)} + 2\varsigma_\phi} \right)^{3/2} \|g_{k+1}\|^{3/2}. \quad (5-10)$$

*Proof.* 将 (5-7) 代入 (5-5) 可得：

$$f(x_k) - m_k(d_k) \geq \frac{\sqrt{h_k(\delta_k)}}{12} \left( \frac{2}{M + 2\sqrt{h_k(\delta_k)} + 2\varsigma_\phi} \right)^{3/2} \|g_{k+1}\|^{3/2} > 0.$$

由于第  $k$  次迭代是成功的，因此得证 (5-10)。  $\square$

**引理 5.6: 二阶下降估计**

假定  $x_k$  未落入困难情况，则如果第  $k$  次迭代是成功的，则函数值的下降量满足：

$$f(x_k) - f(x_k + d_k) \geq -\frac{\eta_1 (\lambda_1(H_k))^3}{6h_k(\delta_k)}. \quad (5-11)$$

*Proof.* 由于  $h_k(\delta_k) \geq h_{\min} > 0$ , 我们将不等式 (5-4b) 代入模型下降式 (5-5), 从而完成证明。  $\square$

接下来, 我们证明, 只要  $\phi_k$  足够接近  $g_k$  (如 假设 5.24 所述), 则不成功的迭代次数将被成功迭代次数的某个数量界定。我们首先给出正则化项  $h_k(\delta_k)$  的上界估计。注意, 这也意味着 (5-7) 右侧的上界是存在的。

**引理 5.7: 正则化项的上界**

在 算法 5-1 中, 若 假设 5.24 成立, 则对于所有  $k \geq 1$ ,  $h_k(\delta_k)$  具有一个一致的上界:

$$h_k(\delta_k) \leq \max \left\{ h_0(\delta_0), 9\gamma_3^2 \left( \frac{M}{2} + \varsigma_\phi \right)^2 \right\} =: \varsigma_h. \quad (5-12)$$

*Proof.* 只需证明当  $h_k(\delta_k) \geq 9\left(\frac{M}{2} + \varsigma_\phi\right)^2$  时, 迭代  $k$  必定是非常成功的, 即:

$$\begin{aligned} m_k(d_k) - f(x_k + d_k) &= (\phi_k - g_k)^T d_k + \frac{1}{2} d_k^T (H_k - \nabla^2 f(x_k + \xi d_k)) d_k + \frac{\sqrt{h_k(\delta_k)}}{3} \|d_k\|^3 \\ &\geq -\varsigma_\phi \|d_k\|^3 - \frac{M}{2} \|d_k\|^3 + \frac{\sqrt{h_k(\delta_k)}}{3} \|d_k\|^3 \\ &= \left( \frac{\sqrt{h_k(\delta_k)}}{3} - \frac{M}{2} - \varsigma_\phi \right) \|d_k\|^3. \end{aligned} \quad (5-13)$$

其中  $\xi \in [0, 1]$ 。因此, 只要  $h_k(\delta_k) \geq 9\left(\frac{M}{2} + \varsigma_\phi\right)^2$ , 则  $m_k(d_k) - f(x_k + d_k) \geq 0$  成立。

结合 引理 5.3, 比率  $\rho_k$  计算如下:

$$\begin{aligned} \rho_k &= \frac{f(x_k) - f(x_{k+1})}{f(x_k) - m_k(d_k)} = \frac{f(x_k) - m_k(d_k) + m_k(d_k) - f(x_{k+1})}{f(x_k) - m_k(d_k)} \\ &= 1 + \frac{m_k(d_k) - f(x_k + d_k)}{f(x_k) - m_k(d_k)} \geq 1. \end{aligned} \quad (5-14)$$

因此, 迭代  $k$  必定是非常成功的, 从而得到了所需的上界。  $\square$

**推论 5.8: 函数下降估计**

若 假设 5.24 成立, 则存在常数  $\varsigma_C > 0$ , 使得对于任何成功迭代  $k$ , 有:

$$f(x_k) - f(x_{k+1}) \geq \varsigma_C \|g_{k+1}\|^{3/2}, \quad (5-15)$$

其中  $\varsigma_C := \eta_1 \frac{\sqrt{h_{\min}}}{12} \left( \frac{2}{M + 2\sqrt{h_{\min}} + 2\varsigma_\phi} \right)^{3/2}$ 。

*Proof.* 由于  $h_k$  在  $[h_{\min}, \varsigma_h]$  内有界, 该表达式:

$$\eta_1 \frac{\sqrt{h_k(\delta_k)}}{12} \left( \frac{2}{M + 2\sqrt{h_k(\delta_k)} + 2\varsigma_\phi} \right)^{3/2}$$

在该区间上连续, 因此存在最小值  $\varsigma_C$ 。通过基本分析可知, 该函数是单调递增的, 因此其最小值出现在  $h_k = h_{\min}$ , 从而得到  $\varsigma_C$ 。□

**推论 5.9: 连续成功迭代下降估计**

若 **假设 5.24** 成立, 则对于任意两个连续的成功迭代 (记为  $j$  和  $j+1$ ), 必定满足以下两种情况之一:

- (a) 若  $x_{j+1}$  满足 (2-2a) 和 (2-2b), 则  $x_{j+1}$  是一个二阶稳定点
- (b) 否则,  $f(x_j) - f(x_{j+2}) \geq \Omega(\epsilon^{3/2})$ 。

*Proof.* 若第一种情况不成立, 则必有:

$$\|g_{j+1}\| \geq \Omega(\epsilon) \text{ 或 } \lambda_1(\nabla^2 f(x_{j+1})) \leq O(-\sqrt{\epsilon}).$$

结合 (5-15) 和 (5-11) 估计的函数值下降量:

$$f(x_j) - f(x_{j+2}) \geq \max \left\{ -\frac{\eta_1 \lambda_1 (H_{j+1})^3}{6\varsigma_h}, \varsigma_C \|g_{j+1}\|^{3/2} \right\} \geq \varsigma_f \epsilon^{3/2},$$

其中  $\varsigma_f = \min \left\{ \varsigma_C, \frac{\eta_1}{6\varsigma_h} \right\}$ , 因此第二种情况成立。□

现在, 我们准备对 **算法 5-1** 进行复杂度分析。我们定义以下集合, 这些集合在后续分析中经常被使用: 成功迭代索引集  $\mathcal{S}_j$  及不成功迭代索引集  $\mathcal{U}_j$ , 它们分别表示截至第  $j$  次迭代的成功和不成功迭代:

$$\mathcal{S}_j := \{k \leq j : \rho_k \geq \eta_1\} \text{ 和 } \mathcal{U}_j := \{k \leq j : \rho_k < \eta_1\}. \quad (5-16)$$

事实上, 集合  $\mathcal{U}_j$  的基数可以由集合  $\mathcal{S}_j$  的基数上界约束, 如下引理所示。

**引理 5.10: 不成功迭代次数的上界**

对于任意  $j \geq 0$ , 令  $\mathcal{S}_j$  和  $\mathcal{U}_j$  由 (5-16) 定义, 则有:

$$|\mathcal{U}_j| \leq \frac{1}{\log \gamma_2} \left[ \frac{1}{2} \log \frac{\varsigma_h}{h_0(\delta_0)} + |\mathcal{S}_j| \log \frac{1}{\gamma_4} \right]. \quad (5-17)$$

*Proof.* 由于算法的更新机制, 我们有:

$$\gamma_4 \sqrt{h_k(\delta_k)} \leq \sqrt{h_{k+1}(\delta_{k+1})}, \quad k \in \mathcal{S}_j, \quad (5-18)$$

以及

$$\gamma_2 \sqrt{h_k(\delta_k)} \leq \sqrt{h_{k+1}(\delta_{k+1})}, \quad k \in \mathcal{U}_j. \quad (5-19)$$

由引理 5.7,  $h_k$  存在上界, 通过归纳法推导可得:

$$\sqrt{h_0(\delta_0)}\gamma_4^{|\mathcal{S}_j|}\gamma_2^{|\mathcal{U}_j|} \leq \sqrt{\varsigma_h}, \quad (5-20)$$

这等价于:

$$|\mathcal{S}_j| \log \gamma_4 + |\mathcal{U}_j| \log \gamma_2 \leq \frac{1}{2} \log \frac{\varsigma_h}{h_0(\delta_0)}. \quad (5-21)$$

重新整理 (5-21) 后, 即得所需结论。□

接下来, 我们假设  $j_f$  是满足以下条件的最小索引:

$$\|g_{j_f+1}\| \leq O(\epsilon), \lambda_1(H_{j_f+1}) \geq \Omega(-\sqrt{\epsilon}),$$

即下一个迭代点已经是  $\epsilon$ -近似的二阶稳定点。我们给出集合  $|\mathcal{S}_{j_f}|$  的上界。

#### 引理 5.11: 成功迭代次数的上界

集合  $\mathcal{S}_{j_f}$  的基数满足:

$$|\mathcal{S}_{j_f}| \leq \frac{f_0 - f_{\text{low}}}{\varsigma_f} \epsilon^{-3/2} = O(\epsilon^{-3/2}). \quad (5-22)$$

*Proof.* 令  $k_j^s$  表示  $\mathcal{S}_{j_f}$  中的第  $j$  个元素, 即  $\mathcal{S}_{j_f} = \{k_1^s, k_2^s, \dots, k_{|\mathcal{S}_{j_f}|}^s\}$ 。由推论 5.9, 对于任意  $k_j^s$ , 有:

$$f(x_{k_j^s}) - f(x_{k_{j+2}^s}) \geq \varsigma_f \epsilon^{3/2}. \quad (5-23)$$

因此, 在  $j_f$  之前, 每两次连续的成功迭代都会导致  $O(\epsilon^{3/2})$  的函数值下降。假设第一步是成功迭代, 则有:

$$f(x_0) - f(x_{j_f+1}) \geq \sum_{j=|\mathcal{S}_{j_f}| \bmod 2+1}^{|\mathcal{S}_{j_f}|-1} [f(x_{k_j^s}) - f(x_{k_{j+2}^s})] \geq \varsigma_f |\mathcal{S}_{j_f}| \epsilon^{3/2}. \quad (5-24)$$

其中 (5-24) 中的 “mod” 为取模运算。由于  $\varsigma_f > 0$  依赖于问题参数 (如  $M$ ), 因此:

$$|\mathcal{S}_{j_f}| \leq \frac{f(x_0) - f_{\text{low}}}{\varsigma_f} \epsilon^{-3/2} = O(\epsilon^{-3/2}).$$

□

#### 定理 5.12: 全局收敛性

自适应 HSODM 需要  $O(\epsilon^{-3/2})$  次迭代, 以达到满足  $\|g_k\| \leq O(\epsilon)$  且  $\lambda_1(H_k) \geq \Omega(-\sqrt{\epsilon})$  的点  $x_k$ 。

*Proof.* 由于  $x_{j_f+1}$  已经是一个二阶稳定点, 总迭代次数为  $|\mathcal{S}_{j_f}|$  和  $|\mathcal{U}_{j_f}|$  之和。最终结果由 (5-22) 和 (5-17) 直接推出。□



除了算法的迭代次数外，我们还关注算法求解 GHMs 的总次数。由于算法需要找到合适的  $h_k$  使其落入所需区间，因此必须搜索合适的  $\delta_k$ 。由第 4.2.2.2 节的分析可知， $h_k$  在大多数情况下是连续的，因此可以使用二分法。特别地，确定  $h_k$  需要额外求解  $O(\log(\epsilon^{-1}))$  个 GHMs (参见第 5.3 节)。

**定理 5.13: 二分法的复杂度**

在适应性 HSODM 的某次迭代  $x_k$  处，二分法的迭代次数满足：

$$O\left(\log\left(\frac{s_h U_\phi U_H}{h_{\min} \sigma}\right)\right). \quad (5-25)$$

若我们将二分法的不精确性设定为  $\sigma = \epsilon$ ，则迭代界变为：

$$O\left(\log\left(\frac{s_h U_\phi U_H}{h_{\min}} \epsilon^{-1}\right)\right). \quad (5-26)$$

我们现在得到 算法 5-1 中 GHM 求解的总次数。

**定理 5.14: GHM 总复杂度**

令  $\mathcal{K}_\psi$  表示 算法 5-1 中求解 GHM 的总次数。若二分法的容差设定为  $\sigma = \epsilon$ ，则在达到满足 (2-2a) 和 (2-2b) 的迭代点之前，我们有以下界：

$$\mathcal{K}_\psi = O\left(\epsilon^{-3/2} \log(\epsilon^{-1})\right). \quad (5-27)$$

*Proof.* 该结论由 定理 5.12 和 定理 5.13 直接推得。 □

### 5.2.2 局部收敛性

现在我们分析自适应 HSODM 的局部收敛性，假定收敛到一个非退化的局部极小点。与假设 3.19 的定义一致。由于  $x_k \rightarrow x^*$  且  $H(x)$  是 Lipschitz 连续的，我们有：

$$H(x_k) \geq \mu I \quad \text{对于充分大的 } k. \quad (5-28)$$

接下来，我们分析自适应 HSODM 的局部收敛性质。

**引理 5.15: 步长与当前梯度的关系**

假定 假设 5.24 和 假设 3.19 成立，则对于充分大的  $k$ ，对应的迭代是非常成功的，即  $\rho_k \geq \eta_2$ ，并且：

$$\|d_k\| \leq \frac{1}{\mu} \|\phi_k\|. \quad (5-29)$$

*Proof.* 首先证明当  $k$  足够大时  $\rho_k \geq \eta_2$ ，注意在我们的算法中，一旦  $d_k \neq 0$ ，我们始终能保证

$m_k(d_k) - f(x_k) < 0$ , 因此可以定义:

$$r_k := f(x_k + d_k) - m_k(d_k) + (1 - \eta_2)(m_k(d_k) - f(x_k)).$$

其中  $\rho_k \geq \eta_2$  等价于  $r_k \leq 0$ , 接下来证明当  $k \rightarrow +\infty$  时  $r_k \leq 0$ .

对于  $f(x_k + d_k) - m_k(d_k)$ , 由 Taylor 展开可得:

$$\begin{aligned} f(x_k + d_k) - m_k(d_k) &= f(x_k) + g_k^T d_k + \frac{1}{2} d_k^T H(x_k + \xi d_k) d_k \\ &\quad - f(x_k) - \phi_k^T d_k - \frac{1}{2} d_k^T H_k d_k - \frac{\sqrt{h_k(\delta_k)}}{3} \|d_k\|^3 \\ &\leq \frac{1}{2} d_k^T (H(x_k + \xi d_k) - H_k) d_k + (g_k - \phi_k)^T d_k \\ &\leq \frac{1}{2} \|H(x_k + \xi d_k) - H_k\| \|d_k\|^2 + \varsigma_\phi \|d_k\|^3, \end{aligned} \quad (5-30)$$

其中  $\xi \in (0, 1)$ . 对于  $m_k(d_k) - f(x_k)$ , 由 (5-4a) 可得:

$$\begin{aligned} f(x_k) - m_k(d_k) &= -\phi_k^T d_k - \frac{1}{2} d_k^T H_k d_k - \frac{\sqrt{h_k(\delta_k)}}{3} \|d_k\|^3 \\ &= \frac{1}{2} d_k^T H_k d_k + \frac{2}{3} \sqrt{h_k(\delta_k)} \|d_k\|^3 \\ &\geq \frac{1}{2} \mu \|d_k\|^2. \end{aligned} \quad (5-31)$$

代入 (5-30) 和 (5-31) 得:

$$r_k \leq \frac{1}{2} \|d_k\|^2 \left\{ \|H(x_k + \xi d_k) - H_k\| + 2\varsigma_\phi \|d_k\| - (1 - \eta_2)\mu \right\}, \quad \xi \in (0, 1). \quad (5-32)$$

由于  $x_k \rightarrow x^*$ , 可知  $d_k \rightarrow 0$ , 因此当  $k \rightarrow +\infty$  时,  $r_k \leq 0$  成立, 完成第一部分的证明。

对于第二部分, 由 (5-4a) 可得:

$$-\phi_k^T d_k = d_k^T H_k d_k + \sqrt{h_k(\delta_k)} \|d_k\|^3. \quad (5-33)$$

因此, 对于充分大的  $k$ , 有:

$$\mu \|d_k\|^2 \leq d_k^T H_k d_k + \sqrt{h_k(\delta_k)} \|d_k\|^3 = -\phi_k^T d_k \leq \|\phi_k\| \|d_k\|. \quad (5-34)$$

重新整理后, 即得所需结论。  $\square$

接下来, 我们证明自适应 HSODM 具有局部二次收敛速率。

#### 定理 5.16: 局部收敛性

假定 假设 5.24 和 假设 3.19 成立, 则自适应 HSODM 具有二次收敛速率, 即:

$$\lim_{k \rightarrow +\infty} \frac{\|\phi_{k+1}\|}{\|\phi_k\|^2} \leq \frac{2\sqrt{\varsigma_h} + M + 4\varsigma_\phi}{2\mu^2}. \quad (5-35)$$

*Proof.* 由引理 5.4 和 引理 5.7 可得:

$$\|g_{k+1}\| \leq \frac{2\sqrt{\varsigma_h} + M + 2\varsigma_\phi}{2} \|d_k\|^2.$$

结合 假设 5.24, 可得:

$$\|\phi_{k+1}\| \leq \|\phi_{k+1} - g_{k+1}\| + \|g_{k+1}\| \leq \frac{2\sqrt{\varsigma_h} + M + 4\varsigma_\phi}{2} \|d_k\|^2.$$

由于 (5-7), 可得:

$$\sqrt{\frac{2}{2\sqrt{\varsigma_h} + M + 4\varsigma_\phi}} \|\phi_{k+1}\|^{\frac{1}{2}} \leq \|d_k\| \leq \frac{1}{\mu} \|\phi_k\|. \quad (5-36)$$

取极限, 即得所需结论。  $\square$

### 5.2.3 关于困难情况的讨论

在本节中, 我们补充之前的分析, 讨论 算法 5-1 中关于困难情况 (见第 5 行) 的处理方法。回顾<sup>[41]</sup> 中类似收敛性的分析, 适应性 HSODM 依赖于对  $\delta_k$  进行二分搜索, 以确保  $\sqrt{h_k(\delta_k)}$  在迭代点  $x_k$  处落入给定区间。困难情况的挑战在于,  $h_k$  在  $\tilde{\alpha}_1$  处的连续性 (参见 引理 4.12) 受损, 因此无法找到合适的  $\delta_k$ 。

幸运的是, GHM (4-2) 的灵活性使我们能够通过适当的扰动  $\phi_k$  以替代  $g_k$  来规避困难情况, 并确保成功迭代。具体而言, 这种简单的扰动可以同时满足以下两个目标:

- (a) 假设 5.24 成立;
- (b) 函数值  $h_k(\delta)$  落入给定区间。

注意, 当  $H_k \geq 0$  时,  $\tilde{\alpha}_1 > 0$ , 此时转折点  $\overline{\delta^{\text{cvx}}} \leq \tilde{\alpha}_1$ , 且  $h_k(\tilde{\alpha}_1) = 0$ , 因此  $\tilde{\alpha}_1$  总是不在目标区间内。因此, 我们仅关注 GHM 为不定情况的情形, 因为根据方法的设计, 当当前点的 Hessian 是半正定时, 不会出现困难情况。此外, 我们通过 引理 4.4 和 推论 4.7 的分析来防止  $\delta_k$  过大, 即保持  $h_k > h_{\min} > 0$ . 因此, 困难情况仅在以下条件下发生:

- (a)  $H_k < 0$ ;
- (b) 先前  $(k-1)$  次迭代是成功的。

#### 5.2.3.1 困难情况的扰动处理

我们现在描述 算法 5-2, 该算法使用扰动梯度  $\phi_k$  来处理困难情况。为了便于理解, 令  $k$  表示当前迭代, 则后续迭代用  $i$  表示, 即  $k, k+1, \dots, k+i$ 。类似于之前的讨论, 我们记  $\lambda_1 = \lambda_1(H_k)$ 。由于出现了困难情况,  $v_k$  现在是最左侧的特征向量。其基本思想如下:

在每个后续迭代  $i$  中, 我们基于前一轮的  $h_{k+i-1}$  进行  $\phi_{k+i}$  的扰动。当逐步增大  $h_{k+i}$  时, 我们使用同样的二分方法 (索引为  $j$ ) 寻找合适的  $\delta_{k+i,j}$ 。最终, 该方法能够产生成功的迭代。同时, 一旦某次迭代成功, 它必须满足 [假设 5.24](#)。

**算法 5-2: 困难情况的扰动处理**

```

1  输入: 迭代步  $k$ ,  $x_k \in \mathbb{R}^n$ ,  $g_k$ ,  $H_k$ ,  $h_{k-1}$ ,  $\delta_{k-1}$ , 其中  $g_k \perp \mathcal{S}_1$ , 容差  $\sigma > 0$ ;
2  for  $i = 0, 1, \dots$  do
3      设置
          
$$\phi_{k+i} = g_k + \frac{\varsigma_\phi}{\gamma_3^2 h_{k+i-1} + \sigma} \lambda_1^2 v_k. \quad (5-37)$$

          计算区间  $I_h := [\gamma_2 \sqrt{h_{k+i-1}}, \gamma_3 \sqrt{h_{k+i-1}}]$ ;
4      repeat // 内层迭代  $j$  通过二分法 (见 第 5.3 节)
5          求解 GHM 子问题的解  $[v_{k+i,j}; t_{k+i,j}]$ 
          
$$\min_{\|[v;t]\| \leq 1} \begin{bmatrix} v \\ t \end{bmatrix}^T \begin{bmatrix} H_k & \phi_{k+i} \\ (\phi_{k+i})^T & \delta_{k+i,j} \end{bmatrix} \begin{bmatrix} v \\ t \end{bmatrix}$$

          设置  $d_{k+i,j} = v_{k+i,j}/t_{k+i,j}$ ,  $h_{k+i} := (\theta_{k+i,j}/\|d_{k+i,j}\|)^2$ ;
6          更新  $\delta_{k+i,j}$ , 增加  $j = j + 1$ 
7      until  $\sqrt{h_{k+i}} \in I_h$ , 满足容差  $\sigma$ ;
8      令  $d_{k+i} = d_{k+i,j}$ ,  $\delta_{k+i} = \delta_{k+i,j}$ ;
9      计算
          
$$\rho_{k,i} = \frac{f(x_k + d_{k+i}) - f(x_k)}{m_k(d_{k+i}) - f(x_k)}$$

          if  $\rho_{k+i} \geq \eta_1$  then
10         跳出循环
    
```

首先, 我们证明如果  $\phi_{k+i}$  采用扰动形式 (5-37) 设定, 一旦某次迭代成功, [假设 5.24](#) 必定成立。

**引理 5.17: 扰动梯度导致成功迭代**

假设  $\phi_{k+i}$  采用以下形式的扰动:

$$\phi_{k+i} = g_k + \frac{\varsigma_\phi}{\gamma_3^2 h_{k+i-1} + \sigma} \lambda_1^2 v_k,$$

则  $\phi_{k+i}$  和  $d_{k+i}$  满足  $\|\phi_{k+i} - g_k\| \leq \varsigma_\phi \|d_{k+i}\|^2$ , 因此, 每当  $d_{k+i}$  被接受时, [假设 5.24](#) 必定成立。

*Proof.* 由于  $(k+i)$ -次迭代是成功的, 回顾最优性条件 (5-4), 有:

$$H_k + \sqrt{h_{k+i}(\delta_{k+i})} \|d_{k+i}\| I \geq 0.$$

上述不等式进一步意味着:

$$\|d_{k+i}\|^2 \geq \frac{1}{h_{k+i}(\delta_{k+i})} \lambda_1^2 \geq \frac{1}{\gamma_3^2 h_{k+i-1}(\delta_{k+i-1}) + \sigma} \lambda_1^2. \quad (5-38)$$

因此,  $\phi_{k+i}$  与  $g_k$  之间的差距可界定如下:

$$\|\phi_{k+i} - g_k\| = \frac{\varsigma_\phi}{\gamma_3^2 h_{k+i-1} + \sigma} \lambda_1^2 \leq \varsigma_\phi \|d_{k+i}\|^2,$$

这意味着  $\phi_{k+i}$  满足 假设 5.24. □

接下来, 我们证明当出现困难情况时, 算法 5-2 最终能够产生成功的迭代。

**定理 5.18: 困难情况的收敛性**

算法 5-2 在至多  $\lfloor \log_{\gamma_2} \frac{\varsigma_h}{h_{k-1}(\delta_{k-1})} \rfloor + 1$  次迭代内产生成功步长。此外, 假设 5.24 必定成立。

*Proof.* 类似于 引理 5.7, 我们有:

$$\begin{aligned} m_k(d_{k+i}) - f(x_k + d_{k+i}) &= (\phi_{k+i} - g_k)^T d_{k+i} + \frac{1}{2} d_{k+i}^T (H_k - \nabla^2 f(x_k + \xi d_{k+i})) d_{k+i} \\ &\quad + \frac{\sqrt{h_{k+i}(\delta_{k+i})}}{3} \|d_{k+i}\|^3 \end{aligned} \quad (5-39)$$

$$\geq -\|\phi_{k+i} - g_k\| \|d_{k+i}\| - \frac{M}{2} \|d_{k+i}\|^3 + \frac{\sqrt{h_{k+i}(\delta_{k+i})}}{3} \|d_{k+i}\|^3 \quad (5-40)$$

$$\geq -\kappa_\phi \|d_{k+i}\|^3 - \frac{M}{2} \|d_{k+i}\|^3 + \frac{\sqrt{h_{k+i}(\delta_{k+i})}}{3} \|d_{k+i}\|^3 \quad (5-41)$$

$$= \left( \frac{\sqrt{h_{k+i}(\delta_{k+i})}}{3} - \frac{M}{2} - \kappa_\phi \right) \|d_{k+i}\|^3, \quad (5-42)$$

其中第二个不等式由 引理 5.17 得出。因此, 当  $h_{k+i}$  超过  $\varsigma_h$  时, 步长  $d_{k+i}$  必然被接受。此外, 由 引理 5.17 可知, 假设 5.24 必定成立。 □

现在可以看到,  $\phi_k$  的选择以及  $h_k$  的调整机制在困难情况出现时起到了关键作用。这一机制可以直接嵌入 算法 5-1 中。事实上, 定理 5.39 说明, 困难情况可以通过有限次迭代的 算法 5-2 规避。

此外, 我们强调, 对于 算法 5-1, 非凸问题的分析也可推广至 Hessian 为 Lipschitz 连续的凸优化问题。注意, 引理 5.4 和 引理 5.5 所建立的条件类似于三次正则化牛顿法的性质, 参见文献<sup>[128]</sup>中的 Lemma 3, Lemma 4. 因此, 我们的方法可以合理地扩展至凸优化问题及其他具有类似复杂度保证的结构化问题。由于篇幅限制, 我们将其留作未来研究。

允许子问题的不精确求解是牛顿型方法中的常见问题<sup>[27,44]</sup>。对于将 GHM 作为子问题的方法而言，这意味着我们只能获得近似特征对，即 Ritz 对，用于更新迭代点。该问题已在原始 HSODM 中得到解决。这里未包含关于不精确 **自适应的** HSODM 的复杂度分析。简单地说，只要 Lanczos 方法的误差足够小，适应性 HSODM 仍可达到与精确版本相同的复杂度。但这些分析需要对困难情况、步长选择、以及二分法进行详细讨论。为了简洁起见，这个部分放在。

## 第三节 齐次框架中的二分法

 5.3.1 基于  $h_k$  的二分法

我们讨论在 算法 5-1 中搜索  $\delta_k$  使得  $\sqrt{h_k} \in I_h$  的复杂度。简而言之，我们利用 第 4.B.3 节 中的分析来证明  $h_k$  的连续性和单调性。这些结果使我们能够采用简单的二分法。为了便于分析二分法，我们首先对梯度和 Hessian 进行如下假定：

**假设 5.19**

假定沿着迭代序列  $x_k$ ，近似梯度  $\phi_k$  和 Hessian  $H_k$  的范数均有上界，即存在两个常数  $U_\phi > 0$  和  $U_H > 0$  使得  $\|\phi_k\| \leq U_\phi, \|H_k\| \leq U_H$ 。

接下来，我们考虑以下情形下的最优解  $[v_k; t_k]$ 。

**引理 5.20**

在 GHM (4-2) 中，若  $\phi_k \not\perp \mathcal{S}_1$ ，则有：

- (a) 若  $\delta_k \leq \lambda_1(H_k)$ ，则  $|t_k| \geq \frac{\sqrt{2}}{2}$ 。
- (b) 若  $\delta_k \geq \lambda_d(H_k)$ ，则  $|t_k| \leq \frac{\sqrt{2}}{2}$ ，其中  $\lambda_r(\cdot)$  表示最大特征值。

*Proof.* 对于第一个结论，回顾 GHM 的定义 (4-2)，我们有

$$\begin{aligned}
 -\theta_k &= v_k^T H_k v_k + 2t_k \phi_k^T v_k + \delta_k t_k^2 \\
 &\geq \lambda_1(H_k) \|v_k\|^2 + 2t_k \phi_k^T v_k + \delta_k t_k^2 \\
 &= \lambda_1(H_k) \|v_k\|^2 - 2t_k^2 (\theta_k + \delta_k) + \delta_k t_k^2 \\
 &\geq \delta_k \|v_k\|^2 - 2t_k^2 (\theta_k + \delta_k) + \delta_k t_k^2 = \delta_k - 2t_k^2 (\theta_k + \delta_k),
 \end{aligned} \tag{5-43}$$

其中，第二个等式来自最优性条件 (4-3)，后续的不等式是由于  $\delta_k \leq \lambda_1(H_k)$ ，而最后一个等式成立是因为  $[v_k; t_k]$  是单位向量。重排可得： $(2t_k^2 - 1)(\theta_k + \delta_k) \geq 0$ ，进而可得  $|t_k| \geq \frac{\sqrt{2}}{2}$ 。类似地，可证明第二个结论。  $\square$

注意，在搜索  $\delta$  的过程中，我们始终可以假定  $\phi_k \not\perp \mathcal{S}_1$ ；否则，可以通过对  $\phi_k$  进行扰动来避免  $\phi_k \perp \mathcal{S}_1$  的情况 (见 算法 5-2)。基于前述分析，我们知道当  $\phi_k \not\perp \mathcal{S}_1$  时，函数  $h_k(\delta)$  关于  $\delta$  是连续且单调递减的 (见 引理 4.12)。因此，在 算法 5-1 的每次迭代 (第 4 行)，我们可以使用二分法搜索合适的  $\delta_k$ 。

不失一般性，假设  $I_h = [\sqrt{\ell}, \sqrt{v}]$ 。为了应用二分法，我们首先需要进行有效的区间搜索，即找

到一个区间  $I_k := [\delta_{low}, \delta_{up}]$  使其包含所需的  $\delta_k$ 。基于  $h_k(\delta)$  的连续性 (见 引理 4.12)，我们可以得到以下结果。

**引理 5.21**

设  $\phi_k \notin \mathcal{S}_1$ ，且对于任意区间  $I_h := [\sqrt{\ell}, \sqrt{\nu}]$ ，使得函数值满足  $h_k(\cdot) \in [\ell, \nu]$ ，若我们设定

$$\delta_{low} = \min \{ \lambda_1(H_k), -\sqrt{\nu} \}, \quad (5-44a)$$

$$\delta_{up} = \max \left\{ \frac{(1 + |\lambda_1(H_k)|)^2 (1 + \lambda_d(H_k) + |\lambda_1(H_k)|)}{\ell}, \|\phi_k\|^2 \right\}, \quad (5-44b)$$

则有

$$[\ell, \nu] \subseteq [h_k(\delta_{up}), h_k(\delta_{low})].$$

*Proof.* 要证明  $[\ell, \nu] \subseteq [h_k(\delta_{up}), h_k(\delta_{low})]$ ，只需验证  $h_k(\delta_{low}) \geq \nu$  且  $h_k(\delta_{up}) \leq \ell$ 。由 引理 5.20，当  $\delta_{low} = \min \{ \lambda_1(H_k), -\sqrt{\nu} \}$  时， $|t_k| \geq \frac{\sqrt{\nu}}{2}$ ，因此  $\frac{t_k^2}{1-t_k^2} \geq 1$ 。另一方面，取  $\delta_k = \delta_{low}$ ，由最优性条件 (4-3)，可得  $\theta_k + \delta_{low} \geq 0$ ，从而  $\theta_k \geq \sqrt{\nu}$ 。因此，

$$h_k(\delta_{low}) = \frac{t_k^2}{1-t_k^2} \theta_k^2 \geq \nu. \quad (5-45)$$

接下来，证明  $h_k(\delta_{up}) \leq \ell$ 。类似地，我们需要分别界定  $\theta_k^2$  和  $\frac{t_k^2}{1-t_k^2}$ 。关于  $\theta_k$ ，当  $\delta_k \geq \|\phi_k\|^2$  时，我们有  $-\lambda_1(H_k) \leq \theta_k \leq 1 + |\lambda_1(H_k)|$ 。回顾  $\phi_k \notin \mathcal{S}_1$  的情况，显然  $\theta_k > -\lambda_1(H_k)$ 。

此外， $\theta_k$  满足以下方程 (见 [146]，定理 3.1)：

$$\delta + \theta = \sum_{i=1}^r \frac{\beta_i^2}{\lambda_i(H_k) + \theta}, \quad (5-46)$$

其中  $\beta_i, i = 1, \dots, r$  由 第 4.B.4 节 证明中的 (4-23) 方式定义。若设  $\theta_k = |\lambda_1(H_k)| + 1$ ，根据 (5-46) 可得

$$\begin{aligned} \delta_k &= -\theta_k + \sum_{i=1}^r \frac{\beta_i^2}{\lambda_i(H_k) + \theta_k} \\ &\leq -1 - |\lambda_1(H_k)| + \frac{\|\phi_k\|^2}{\lambda_1(H_k) + |\lambda_1(H_k)| + 1} \\ &\leq \frac{\|\phi_k\|^2}{\lambda_1(H_k) + |\lambda_1(H_k)| + 1} \leq \|\phi_k\|^2. \end{aligned} \quad (5-47)$$

由此可得，当  $\delta \geq \|\phi_k\|^2$  时， $\theta_k \leq 1 + |\lambda_1(H_k)|$ ，因为  $\theta_k$  随  $\delta_k$  单调递减。

另一方面，若取  $\delta_k = \lambda_d(H_k) + \alpha$ ，其中  $\alpha > 0$ ，我们估计  $\alpha$  如下：

$$\begin{aligned} -\theta_k &= v_k^T H_k v_k + 2t_k \phi_k^T v_k + \delta_k t_k^2 \\ &\leq (\delta_k - \alpha) \|v_k\|^2 - 2t_k^2 (\delta_k + \theta_k) + \delta_k t_k^2 \\ &= \delta_k - \alpha \|v_k\|^2 - 2t_k^2 (\delta_k + \theta_k), \end{aligned} \quad (5-48)$$



变形得

$$\begin{aligned}
 \frac{t_k^2}{1-t_k^2} &\leq 1 - \frac{\alpha}{\theta_k + \delta_k} = 1 - \frac{\alpha}{\theta_k + \lambda_d(H_k) + \alpha} \\
 &\leq 1 - \frac{\alpha}{1 + |\lambda_1(H_k)| + \lambda_d(H_k) + \alpha} \\
 &= \frac{1 + |\lambda_1(H_k)| + \lambda_d(H_k)}{1 + |\lambda_1(H_k)| + \lambda_d(H_k) + \alpha}.
 \end{aligned} \tag{5-49}$$

由于 (5-49) 左侧关于  $\alpha$  单调递减, 因此当

$$\alpha \geq \alpha_u := -1 - |\lambda_1(H_k)| - \lambda_d(H_k) + \frac{(1 + |\lambda_1(H_k)|)^2(1 + \lambda_d(H_k) + |\lambda_1(H_k)|)}{\ell}$$

时, 有

$$\frac{t_k^2}{1-t_k^2} \leq \frac{\ell}{(1 + |\lambda_1(H_k)|)^2}. \tag{5-50}$$

结合 (5-44b) 可知  $\delta_{up} \geq \lambda_d(H_k) + \alpha_u$ , 并结合 (5-47), 可得

$$\begin{aligned}
 h_k(\delta_{up}) &\leq \theta_k^2 \cdot \frac{\ell}{(1 + |\lambda_1(H_k)|)^2} \\
 &\leq (|\lambda_1(H_k)| + 1)^2 \frac{\ell}{(1 + |\lambda_1(H_k)|)^2} \leq \ell.
 \end{aligned}$$

证明完成。  $\square$

在上述引理中, 我们通过显式构造  $I_k$  证明了  $\delta_k$  的区间存在性, 使得  $h_k(\delta_k) \in [\ell, \nu]$  对于任意  $[\ell, \nu]$  成立。为了分析二分法的复杂度, 我们定义如下针对  $\delta_k$  的目标区间, 并允许一个容差  $\sigma > 0$ :

$$I_k^\sigma = [\underline{\delta}, \bar{\delta}], \quad h_k(\underline{\delta}) = \nu + \sigma, \quad h_k(\bar{\delta}) = \ell. \tag{5-51}$$

注意, 如果  $\sigma > 0$  充分小, 该不精确性不会影响 算法 5-1 框架的收敛率。接下来, 我们关注  $I_k^\sigma$  的长度与  $I_k$  长度的比值, 这是分析二分法复杂度的关键步骤。接下来, 我们给出区间  $I_k^\sigma$  长度的下界。

#### 引理 5.22

在第  $k$  次迭代中, 若  $d_k$  是满足  $\delta_k \in I_k^\sigma$  的 (4-2) 解, 则区间  $I_k^\sigma$  的长度至少为

$$\frac{\sigma \|d_k\|}{2\sqrt{\nu + \sigma} + (\nu + \sigma)\|\phi_k\|}. \tag{5-52}$$

*Proof.* 由 (5-51) 定义可知,  $\ell \leq h_k(\delta) = \frac{\theta^2}{\|d\|^2} \leq \nu + \sigma$ , 于是

$$\ell \|d\|^2 \leq \theta^2 \leq (\nu + \sigma) \|d\|^2.$$

将上述结果代入 (4-25), 可得

$$|h'_k(\delta)| \leq \frac{2\sqrt{\nu + \sigma} \|d\|^3 + (\nu + \sigma) \|d\|^2 d^T (H_k + \theta I)^{-1} d}{\|d\|^4}. \tag{5-53}$$

另一方面，由最优性条件 (4-3) 及其 Schur 补可得

$$H_k + \theta I - \frac{\phi_k \phi_k^T}{\delta + \theta} \geq 0,$$

因此，

$$H_k + \theta I \geq \frac{\phi_k \phi_k^T}{\delta + \theta} = -\frac{\phi_k \phi_k^T}{\phi_k^T d} \geq \frac{\phi_k \phi_k^T}{\|\phi_k\| \|d\|},$$

进而得到

$$d^T (H_k + \theta I)^{-1} d \leq \frac{(\phi_k^T d)^2}{\|\phi_k\| \|d\|}. \quad (5-54)$$

将 (5-54) 代入 (5-53)，可得

$$|h'_k(\delta)| \leq \frac{2\sqrt{\nu + \sigma} + (\nu + \sigma)\|\phi_k\|}{\|d\|}. \quad (5-55)$$

由均值定理，

$$h_k(\underline{\delta}) - h_k(\bar{\delta}) = h'_k(\xi)(\bar{\delta} - \underline{\delta}), \quad \xi \in [\underline{\delta}, \bar{\delta}].$$

结合 (5-51) 和 (5-55)，可得

$$\bar{\delta} - \underline{\delta} \geq \frac{\sigma \|d_k\|}{2\sqrt{\nu + \sigma} + (\nu + \sigma)\|\phi_k\|},$$

证毕。 □

我们注意到，在 算法 5-1 过程中，可以假设  $\|d_k\| \geq \sqrt{\epsilon}$ 。否则，由 (5-7) 和 (5-4b) 可得

$$\|g_{k+1}\| \leq O(\epsilon), \quad \lambda_{\min}(H_k) \geq \Omega(-\sqrt{\epsilon}).$$

由于  $\nabla^2 f(x)$  具有 Lipschitz 连续性，我们知道  $\lambda_{\min}(H_{k+1}) \geq \Omega(-\sqrt{\epsilon})$ ，因此  $x_{k+1} := x_k + d_k$  满足 (2-2a) 和 (2-2b)，可以在迭代  $x_{k+1}$  处终止算法。

为了分析  $|I_k|$  和  $|I_k^\sigma|$  之间的比值，根据 假设 5.19，我们对区间  $I_k^\sigma$  的长度进行粗略估计：

$$\bar{\delta} - \underline{\delta} = \Omega\left(\frac{\sigma\sqrt{\epsilon}}{c_h U_\phi}\right), \quad (5-56)$$

该结果由 引理 5.7 以及  $\nu$  由函数值  $h_{k-1}(\delta_{k-1})$  确定的事实推出。

现在，我们可以对二分法的复杂度进行最终分析。

### 5.3.1.1 证明 定理 5.13

*Proof.* 由于我们使用二分法在初始区间  $I_k$  内搜索  $I_k^\sigma$ ，根据二分法的机制，其复杂度为

$$O\left(\log \frac{|I_k|}{|I_k^\sigma|}\right) = O\left(\log \frac{\delta_{up} - \delta_{low}}{\bar{\delta} - \underline{\delta}}\right). \quad (5-57)$$

由不等式 (5-56), 我们已证明  $\bar{\delta} - \underline{\delta}$  的下界。接下来, 我们估计  $\delta_{up} - \delta_{low}$ 。回顾 引理 5.21 的设定, 可得

$$\begin{aligned}\delta_{up} - \delta_{low} &\leq \max \left\{ (U_\phi)^2, \frac{(1 + U_H)^2(1 + 2U_H)}{h_{\min}} \right\} + \max\{U_H, \sqrt{\varsigma_h}\} \\ &\leq O \left( \frac{(U_\phi)^2(U_H)^3\sqrt{\varsigma_h}}{h_{\min}} \right),\end{aligned}\tag{5-58}$$

将 (5-58) 和 (5-56) 代入 (5-57), 即可得到所需结论。□

#### 第四节 数值实验

与之前类似, 我们对自适应的 HSODM 进行一些数值实验。我们在 Mac OS 桌面端实现了我们的算法, 设备配备 3.2 GHz 6 核 Intel Core i7 处理器和 32 GB DDR4 内存。大多数子程序均可在标准的 Julia 包中找到。例如, 我们使用 Optim.jl 包中的线搜索算法。CUTEst 基准测试的访问由 CUTEst.jl 提供支持。KrylovKit.jl 用于共轭梯度法和 Lanczos 方法。实现的详细信息仍可在 [github.com/bzhangcw/DRSOM.jl](https://github.com/bzhangcw/DRSOM.jl) 找到。简而言之, 我们在本论文中实现了 Adaptive-HSODM (算法 5-1)。在求解 GHM 时, 我们在 Lanczos 方法中设置收敛容差为  $\min\{10^{-5}, 10^{-2}\|g_k\|\}$ 。

##### 5.4.1 CUTEst 基准测试

我们在 CUTEst 问题的一个子集上进行基准测试, 以评估非凸优化的性能。我们选择决策变量维度满足  $500 \leq n \leq 5000$  的问题, 共计 81 个实例。我们将其与三次正则化牛顿法 (ARC) 和牛顿信赖域法 (Newton-TR-STCG) 进行比较。其中, Newton-TR-STCG 在求解信赖域子问题时使用 Steihaug-Toint 共轭梯度方法。我们采用 [53] 中最新的 Julia 实现, 并使用其中的默认设置。

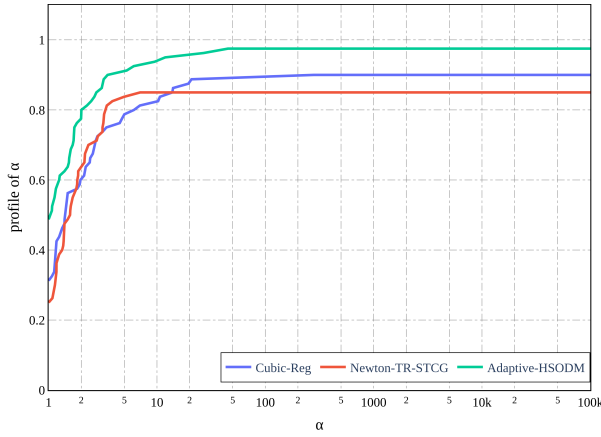
我们设定终止准则, 以找到一个迭代点  $x_k$  使得  $\|g_k\| \leq 10^{-5}$ 。每个实例的时间限制为 200 秒。如果某个实例失败, 则将其迭代次数和求解时间设为 20,000。表 5.4.1 统计了三种算法的结果。

在该表格中,  $\mathcal{R}$  表示成功求解的实例数量;  $\bar{t}_G, \bar{k}_G$  分别表示时间和迭代次数的缩放几何均值 (SGM, 分别以 1 秒和 50 次迭代进行缩放);  $\bar{k}_G^f, \bar{k}_G^g$  分别是目标函数和梯度评估的 SGM, 其中每次 Hessian-向量乘积计作两次梯度评估。图 5.4.1 进一步展示了性能概况。从结果来看, 我们的 Adaptive-HSODM 具有较强的鲁棒性: 它在成功求解的实例数量、迭代次数和运行时间方面均表现最佳。三种方法在目标函数评估次数上相近, 而 Newton-TR-STCG 在梯度评估方面表现最佳。Adaptive-HSODM 在每次迭代中的梯度评估次数 ( $\bar{k}_G^g/\bar{k}_G$ ) 略高, 这可能是由于其对问题维度的依赖性较高。

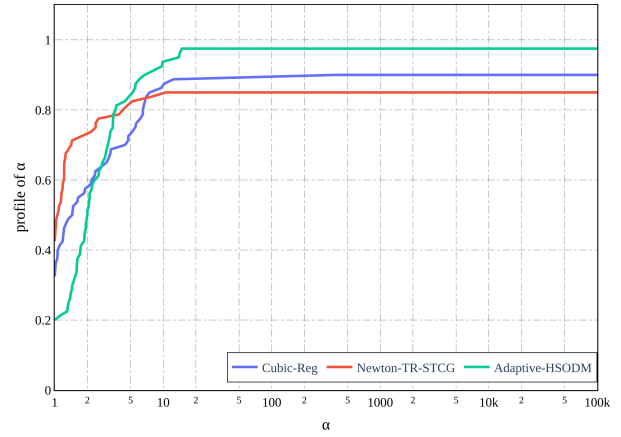
与凸 (接近退化) 函数下的情况不同, Lanczos 方法可能需要调整以适应更好的不精确策略, 因为在一般情况下缺少基于间隙的条件数。我们将这一改进留作未来研究。总的来说, 实验结果表明, 我们对 Adaptive-HSODM 的初步实现在 CUTEst 问题上与当前最先进的软件包 [53] 具有竞争力。

表 5.4.1 CUTEst 基准测试结果汇总 (81 个实例)。

方法	$\mathcal{H}$	$\bar{t}_G$	$\bar{k}_G$	$\bar{k}_G^f$	$\bar{k}_G^g$
ARC	72.00	14.20	304.94	304.94	1810.82
Adaptive-HSODM	78.00	10.49	189.70	323.47	1642.13
Newton-TR-STCG	68.00	21.80	353.21	353.21	1327.13



(a) 迭代次数的性能概况



(b) 梯度评估的性能概况

图 5.4.1 CUTEst 问题的性能概况。

## 本章附录

### 第一节 基于非精确特征值的自适应齐次二阶法

当使用不精确特征值时,所有提及的量

$$(\theta_k, v_k, t_k, h_k)$$

在一定程度上均为近似形式. 在 Lanczos 方法中, 我们记  $\gamma_k, [\hat{v}_k; \hat{t}_k]$  为近似最左特征对的 Ritz 对. 当误差容限  $e_k := \theta_k - \gamma_k$  足够小时, 例如  $e_k \leq O(\sqrt{\epsilon})$ , Lanczos 方法终止. 我们回顾如下的近似最优条件.

#### 引理 5.23: 近似最优条件

使用 Lanczos 方法求解 (5-2) 使得  $e_k := \theta_k - \gamma_k$ , 则有

$$H_k \hat{v}_k + \phi_k \hat{t}_k + \gamma_k \hat{v}_k = r_k, \quad (5-59a)$$

$$\phi_k^T \hat{v}_k + \delta_k \hat{t}_k + \gamma_k \hat{t}_k = \sigma_k, \quad (5-59b)$$

$$F_k + (\gamma_k + e_k)I \geq 0, \quad (5-59c)$$

$$[\hat{v}_k; \hat{t}_k] \perp [r_k; \sigma_k], \quad (5-59d)$$

其中  $[r_k; \sigma_k]$  为 Ritz 误差.

我们首先讨论  $\hat{t}_k \neq 0$  的情况, 因为当  $\hat{t}_k = 0$  时可以识别出类似于困难情况的情形. 当子问题被不精确求解时, 精确对  $(\theta_k, [v_k; t_k])$  的信息不存在. 因此, 辅助函数  $h_k(\delta_k)$  与模型函数  $m_k(d)$  都只能利用  $\gamma_k, [\hat{v}_k; \hat{t}_k]$  进行近似. 它们可以定义为

$$\hat{h}_k(\delta_k) = \left( \frac{\gamma_k}{\|\hat{d}_k\|} \right)^2, \quad \text{其中 } \hat{d}_k = \hat{v}_k / \hat{t}_k, \quad (5-60)$$

以及三次正则

$$\hat{m}_k(\hat{d}_k) = f(x_k) + \phi_k^T \hat{d}_k + \frac{1}{2}(\hat{d}_k)^T H_k \hat{d}_k + \frac{\sqrt{\hat{h}_k(\delta_k)}}{3} \|\hat{d}_k\|^3. \quad (5-61)$$

现在我们给出适用于二阶 Lipschitz 连续函数的自适应齐次二阶下降法, 如 算法 5-1 所示.

我们进一步对每个迭代点  $x_k$  在 GHM 中的函数  $\phi_k(x_k)$  做如下假设.

#### 假设 5.24

假设存在一个统一常数  $\varsigma_\phi > 0$ . 给定迭代点  $x_k \in \mathbb{R}^n$ , 若  $t_k \neq 0$ , 则有

$$\|\phi_k(x_k) - g_k\| \leq \varsigma_\phi \|\hat{d}_k\|^2, \quad (5-63)$$

**算法 5-3:** 自适应齐次二阶下降法

```

1  初始点  $x_0 \in \mathbb{R}^n$ ,  $\delta_0 \in \mathbb{R}$ ,  $I_h = \mathbb{R}$ ,  $h_{\min} > 0$ , 参数  $0 < \iota_1 < \eta_2 < 1$ ,  $\iota_1 > 1$ ,  $\iota_3 \geq \iota_2 > 1$ ,  $0 < \iota_4 \leq 1$ ,
    $\sigma > 0$ ;
2  for  $k = 0, 1, 2, \dots$  do
3      令  $\phi_k = g_k$ ,  $\delta_{k,0} = \delta_{k-1}$ ;
4      for  $j = 0, 1, \dots, \mathcal{T}_k$  do
5          求解子问题 (5-2) 得到解对  $(\gamma_{k,j}, [\hat{v}_{k,j}; \hat{t}_{k,j}])$ :
              
$$\min_{\| \begin{bmatrix} v \\ t \end{bmatrix} \| \leq 1} \begin{bmatrix} v \\ t \end{bmatrix}^T \begin{bmatrix} \hat{h}_k & \phi_{k,j} \\ (\phi_{k,j})^T & \delta_{k,j} \end{bmatrix} \begin{bmatrix} v \\ t \end{bmatrix} \quad (5-62)$$

              if  $\hat{t}_{k,j} = 0$  then // 检查困难情况, 见 第 5.A.3 节
                  转至 算法 5-2
                  跳出
              令  $\hat{d}_{k,j} = \hat{v}_{k,j} / \hat{t}_{k,j}$ ,  $\hat{h}_k(\delta_{k,j}) := (\gamma_{k,j} / \|\hat{d}_{k,j}\|)^2$ ;
              if  $\sqrt{\hat{h}_k(\delta_k)} \in I_h$  在容差  $\sigma$  内 then
                  令  $\hat{d}_k := \hat{d}_{k,j}$ ,  $\delta_k = \delta_{k,j}$ 
                  跳出
              更新  $\delta_{k,j}$ ;
12  计算
              
$$\rho_k := \frac{f(x_k + \hat{d}_k) - f(x_k)}{m_k(\hat{d}_k) - f(x_k)};$$

              if  $\rho_k > \eta_2$  then // 非常成功的迭代
                  令  $I_h = \left[ \max \left\{ \sqrt{h_{\min}}, \iota_4 \sqrt{\hat{h}_k(\delta_k)} \right\}, \sqrt{\hat{h}_k(\delta_k)} \right]$ ,  $x_{k+1} = x_k + \hat{d}_k$ 
14  if  $\iota_1 \leq \rho_k \leq \eta_2$  then // 成功的迭代
                  令  $I_h = \left[ \sqrt{\hat{h}_k(\delta_k)} / \iota_1, \iota_2 \sqrt{\hat{h}_k(\delta_k)} \right]$ ,  $x_{k+1} = x_k + \hat{d}_k$ 
16  else // 不成功的迭代
                  令  $I_h = \left[ \iota_2 \sqrt{\hat{h}_k(\delta_k)}, \iota_3 \sqrt{\hat{h}_k(\delta_k)} \right]$ ,  $x_{k+1} = x_k$ 
18

```

其中  $\hat{d}_k = \hat{v}_k / \hat{t}_k$ .

我们提出如下停止准则以终止子问题求解器.

**条件 5.25: 子问题的不精确性**

假设通过 Lanczos 方法以预定误差容限  $e_k \leq O(\epsilon^{1/2})$  求解 (5-2). 当满足以下条件时, 终止 Lanczos 方法:

$$\|\phi_k + (H_k + \sqrt{\hat{h}_k(\delta_k)}\|\hat{d}_k\|I)\hat{d}_k\| \leq \varsigma_r \|\hat{d}_k\|^2, \quad (5-64a)$$

$$\gamma_k + \delta_k \geq 0, \quad (5-64b)$$

$$\sqrt{h_{\min}}\|\hat{d}_k\| \geq e_k. \quad (5-64c)$$

注意, 上述不精确条件会随着 Krylov 子空间的演化逐渐满足. 这里  $\varsigma_r$  可取任一正常数, 它以类似于 inexact Newton 方法<sup>[40]</sup> 中的方式衡量一阶最优性条件的相对误差. 条件 (5-64b) 要求 Ritz 值至少与  $\delta_k$  一样好. 这种要求在 第 3 章 中通过在 Lanczos 方法中使用偏随机初始化进行了讨论. 而我们提出的最后一个条件 (5-64c) 仅用于确保收敛到二阶平稳性, 这一条件在算法初期可能不满足.

建立不精确算法的收敛性引出了一组关键问题:

- (1) 不精确性如何传递到下降引理中;
- (2) 如何使用  $\hat{h}_k(\delta)$  代替  $h_k$  进行有效的二分, 且仍具有  $O(\log(1/\sigma))$  的收敛速度;
- (3) 在不精确求解下, 扰动思想如何发挥作用.

这些问题将在以下各小节中依次解答.

### 5.A.1 收敛性分析

目前, 我们假设困难情况  $\hat{t}_k = 0$  不会发生, 这些问题将在后续讨论中解决. 下面我们证明, 在足够大的  $\hat{h}_k$  及辅助函数  $\hat{h}_k$  的上界下, 最终步将成为成功步. 显然, 我们给出以下关于下降步的论证.

**引理 5.26**

假设  $\hat{d}_k$  满足 (5-64b), 则存在如下模型下降量

$$f(x_k) - \hat{m}_k(\hat{d}_k) \geq \frac{\sqrt{\hat{h}_k(\delta_k)}}{6} \|\hat{d}_k\|^3 \geq \Omega(\|\hat{d}_k\|^3). \quad (5-65)$$

*Proof.* 注意, 根据 (5-59) 与 (5-64b), 有

$$(\hat{d}_k)^T H_k \hat{d}_k + \gamma_k \|\hat{d}_k\|^2 + 2\phi_k^T \hat{d}_k = -\delta_k - \gamma_k \leq 0.$$

则有

$$\begin{aligned}
 m_k(\hat{d}_k) - f(x_k) &= \phi_k^T \hat{d}_k + \frac{1}{2}(\hat{d}_k)^T H_k \hat{d}_k + \frac{\sqrt{\hat{h}_k(\delta_k)}}{2} \|\hat{d}_k\|^3 - \frac{\sqrt{\hat{h}_k(\delta_k)}}{6} \|\hat{d}_k\|^3 \\
 &\leq \phi_k^T \hat{d}_k + \frac{1}{2}(\hat{d}_k)^T H_k \hat{d}_k + \frac{\gamma_k}{2} \|\hat{d}_k\|^2 - \frac{\sqrt{\hat{h}_k(\delta_k)}}{6} \|\hat{d}_k\|^3 \\
 &= -\frac{1}{2}(\delta_k + \gamma_k) - \frac{\sqrt{\hat{h}_k(\delta_k)}}{6} \|\hat{d}_k\|^3 \leq -\frac{\sqrt{\hat{h}_k(\delta_k)}}{6} \|\hat{d}_k\|^3.
 \end{aligned} \tag{5-66}$$

又因为  $\hat{h}_k(\delta_k) \geq h_{\min}$ , 证毕.  $\square$

#### 引理 5.27

假设不精确解满足 (5-64c), 则有

$$\|\hat{d}_k\| \geq -\frac{1}{2\sqrt{\epsilon_h}} \lambda_1(H_k). \tag{5-67}$$

*Proof.* 注意, 根据 (5-59c), 有

$$F_k + \gamma_k I + e_k I \geq 0.$$

因此,

$$\begin{aligned}
 \gamma_k &= \sqrt{\hat{h}_k(\delta_k)} \|\hat{d}_k\| \\
 &\geq -\lambda_{\min}(F_k) - e_k \\
 &\geq -\lambda_{\min}(H_k) - e_k.
 \end{aligned}$$

结合 (5-64c), 得到

$$2\sqrt{\hat{h}_k(\delta_k)} \|\hat{d}_k\| \geq -\lambda_1(H_k).$$

$\square$

以上引理表明, 满足 条件 5.25 的不精确步保证了与精确情况相同的下降性质, 回答了问题 (a). 作为对应, 可以将上述引理与本章中的精确情况进行比较. 其余分析与精确情况相同.

因此, 在 Ritz 对的质量假设 (条件 5.25) 下, 可得不精确自适应 HSODM 与精确版本具有相同的  $O(\epsilon^{-3/2})$  迭代复杂度, 即在  $O(\epsilon^{-3/2})$  次迭代内可获得满足  $\|g_k\| \leq O(\epsilon)$  且  $\lambda_1(H_k) \geq \Omega(-\sqrt{\epsilon})$  的点  $x_k$ . 我们省略证明, 因为证明过程仅需将精确量替换为上述不精确量.

#### 定理 5.28

假设子问题 (5-2) 通过 Lanczos 方法以误差容限  $e_k \leq O(\epsilon^{1/2})$  求解, 且近似解满足 条件 5.25. 则自适应 HSODM 在  $O(\epsilon^{-3/2})$  次迭代内可得到满足  $\|g_k\| \leq O(\epsilon)$  以及  $\lambda_1(H_k) \geq \Omega(-\sqrt{\epsilon})$  的点  $x_k$ .



## 5.A.2 二分法的复杂度

若没有二分法来确定  $\delta_k$ , 则无法建立复杂度率。在本小节中, 我们简单验证使用  $\hat{h}$  替代  $h$  时, 二分法依然有效。我们展示  $h$  与  $\hat{h}$  之间的差异可以通过不精确性的质量来度量。

由于我们在搜索过程中允许容差  $\sigma$ , 因此可以将搜索区间  $\hat{I}_h$  与  $h_k$  的真实区间  $I_h$  关联起来。我们记  $\hat{I}_h: [\ell, \nu]$  为  $\hat{h}_k$  的目标区间。

在开始之前, 我们作出如下假设。

**假设 5.29:** 对

齐次系统 (5-2), 假设存在  $\sigma_F > 0$ , 使得下式成立

$$v^T F_k v - \lambda_1(F_k) \geq \sigma_F, \quad (5-68)$$

对于某个满足  $\|v\| = 1$  的  $v$ 。

注意, 若  $\|F_k\| \gg e_k$ , 则 (5-68) 无失一般性; 否则, Lanczos 方法几乎立即终止。以下结果摘自第 3 章。为完整起见, 我们也给出证明。

**引理 5.30**

假设  $\sigma_F := \lambda_2(F_k) - \lambda_1(F_k) > 0$ , 则对于 (5-59) 中的  $[r_k; \sigma_k]$ , 有

$$\|[r_k; \sigma_k]\| \leq O\left(\sqrt{\frac{e_k}{\sigma_F}}\right). \quad (5-69)$$

并且如果  $v_1 = [v_1; t_1] \in \mathcal{S}_1(F_k)$ , 则有

$$\|[\hat{v}_k; \hat{t}_k] - [v_1; t_1]\| \leq O\left(\sqrt{\frac{e_k}{\sigma_F}}\right). \quad (5-70)$$

*Proof.* 根据 (5-59), 我们有

$$\hat{d}_k^T F_k \hat{d}_k + \gamma_k \|\hat{d}_k\|^2 = 0. \quad (5-71)$$

我们可以将  $\hat{d}_k$  写成  $\hat{d}_k = \alpha v_1 + s$ , 其中  $s \perp v_1$ 。又因为  $\hat{d}_k$  为单位向量, 则有

$$\alpha^2 + \|s\|^2 = 1. \quad (5-72)$$

则由 (5-71) 可得

$$\begin{aligned} -\theta_k + e_k &\geq -\gamma_k = \hat{d}_k^T F_k \hat{d}_k \\ &= -\theta_k \alpha^2 + s^T F_k s \\ &\geq -\theta_k \alpha^2 + (-\theta_k + \sigma_F) \|s\|^2. \end{aligned} \quad (5-73)$$

上式的第二个等式是通过展开  $\hat{d}_k$  且利用  $s \perp v_1$  得到的. 这意味着

$$\|s\|^2 \leq \frac{e_k}{\sigma_F}. \quad (5-74)$$

因此, 有

$$\begin{aligned} r_k &= F_k \hat{d}_k + \gamma_k \hat{d}_k \\ &= (F_k + \gamma_k I)(\alpha v_1 + s) \\ &= \alpha(\gamma_k - \theta_k)v_1 + (F_k + \gamma_k I)s. \end{aligned} \quad (5-75)$$

由 (5-74), 可得残差范数的界:

$$\begin{aligned} \|r_k\| &\leq \alpha(\theta_k - \gamma_k) + \|(F_k + \gamma_k I)s\| \\ &\leq \alpha e_k + \|F_k + \gamma_k I\| \sqrt{\frac{e_k}{\sigma_F}} \\ &\leq \alpha e_k + 2U_F \sqrt{\frac{e_k}{\sigma_F}}, \end{aligned} \quad (5-76)$$

其中  $U_F$  表示  $\|F_k\|$  的上界.

对于第二部分, 注意

$$\begin{aligned} \|\hat{d}_k - v_1\| &= \sqrt{\|\hat{d}_k - v_1\|^2} \\ &= \sqrt{2\|s\|^2} \leq O\left(\sqrt{\frac{e_k}{\sigma_F}}\right). \end{aligned} \quad (5-77)$$

□

我们现在将辅助函数用  $t_k$  表示, 即

$$\hat{h}_k(\delta_k) = \gamma_k^2 \cdot g(\hat{t}_k), \quad h_k(\delta_k) = \theta_k^2 \cdot g(t_k), \quad (5-78)$$

其中定义  $g(t) = \frac{t^2}{1-t^2}$ ; 注意, 我们已经得到  $\hat{t}_k$  的上界. 现在我们将  $\hat{h}_k$  用  $h_k$  的盒状区间界定如下.

#### 引理 5.31: 辅

函数  $\hat{h}_k$  可由函数  $h_k$  同时给出上界与下界, 具体为:

$$h_k(\delta_k) - \hat{h}_k(\delta_k) \leq 2\frac{\varsigma_h}{\gamma_k} e_k + 2\sqrt{\varsigma_h} \frac{(\gamma_k^2 + \varsigma_h)^{3/2}}{\gamma_k^2} |t_k - \hat{t}_k| + o(|t_k - \hat{t}_k|), \quad (5-79)$$

以及

$$\hat{h}_k(\delta_k) - h_k(\delta_k) \leq 2\sqrt{\varsigma_h} \frac{(\gamma_k^2 + \varsigma_h)^{3/2}}{\gamma_k^2} |t_k - \hat{t}_k| + o(|t_k - \hat{t}_k|). \quad (5-80)$$

*Proof.* 我们有

$$\begin{aligned}
 h_k(\delta_k) - \hat{h}_k(\delta_k) &= \theta_k^2 g(t_k) - \gamma_k^2 g(\hat{t}_k) \\
 &= \theta_k^2 g(t_k) - \gamma_k^2 g(t_k) + \gamma_k^2 g(t_k) - \gamma_k^2 g(\hat{t}_k) \\
 &= g(t_k)(\theta_k + \gamma_k)(\theta_k - \gamma_k) + \gamma_k^2 g'(\hat{t}_k)(t_k - \hat{t}_k) + o(|t_k - \hat{t}_k|) \\
 &\leq 2g(t_k)\theta_k e_k + \gamma_k^2 |g'(\hat{t}_k)| |t_k - \hat{t}_k| + o(|t_k - \hat{t}_k|) \\
 &\leq 2g(t_k)\theta_k e_k + \gamma_k^2 |g'(\hat{t}_k)| |t_k - \hat{t}_k| + o(|t_k - \hat{t}_k|).
 \end{aligned} \tag{5-81}$$

关于  $g'(\cdot)$  的部分, 由于  $g'(\cdot)$  是单调递增的, 且根据前述分析  $\hat{t}_k$  有上界, 故有

$$g'(\hat{t}_k) = \frac{2\hat{t}_k}{(1 - (\hat{t}_k)^2)^2} \leq 2\sqrt{s_h} \frac{(\gamma_k^2 + s_h)^{3/2}}{\gamma_k^4}.$$

因此, 由 (5-81) 得到

$$h_k(\delta_k) - \hat{h}_k(\delta_k) \leq 2\frac{s_h}{\gamma_k} e_k + 2\sqrt{s_h} \frac{(\gamma_k^2 + s_h)^{3/2}}{\gamma_k^2} |t_k - \hat{t}_k| + o(|t_k - \hat{t}_k|).$$

类似地, 有

$$\begin{aligned}
 \hat{h}_k(\delta_k) - h_k(\delta_k) &= \gamma_k^2 g(\hat{t}_k) - \theta_k^2 g(t_k) \\
 &= \gamma_k^2 g(\hat{t}_k) - \gamma_k^2 g(t_k) + \gamma_k^2 g(t_k) - \theta_k^2 g(t_k) \\
 &\leq \gamma_k^2 |g'(\hat{t}_k)| |t_k - \hat{t}_k| + o(|t_k - \hat{t}_k|) \\
 &\leq 2\sqrt{s_h} \frac{(\gamma_k^2 + s_h)^{3/2}}{\gamma_k^2} |t_k - \hat{t}_k| + o(|t_k - \hat{t}_k|).
 \end{aligned}$$

□

由此可见, 当 Lanczos 方法求解的足够精确时, 搜索过程仍然可以正常进行.

如果搜索过程基于  $\hat{h}_k$ , 令目标区间为  $\hat{I}_h = [\ell, \nu]$ , 则有如下推论.

#### 推论 5.32

假设存在区间  $\hat{I}_h = [\ell, \nu]$  使得  $\hat{h}_k \in \hat{I}_h$ , 那么满足  $h_k \in I_h$  的区间

$$I_h := \left[ \ell + 2\frac{s_h}{\gamma_k} e_k + 2\sqrt{s_h} \frac{(\gamma_k^2 + s_h)^{3/2}}{\gamma_k^2} |t_k - \hat{t}_k|, \nu - 2\sqrt{s_h} \frac{(\gamma_k^2 + s_h)^{3/2}}{\gamma_k^2} |t_k - \hat{t}_k| \right]$$

满足。

忽略高阶项, 上述结果的一个直接后果是: 如果  $\hat{h}_k(\delta_k) \notin \hat{I}_h$ , 则必有  $h_k(\delta_k) \notin I_h$ 。只要区间  $\hat{I}_h$  的长度足够大, 精确区间  $I_h$  就不会为空, 从而二分法是良定义的。具体来说, 通过允许  $\sigma$  作为  $\hat{I}_h$  的长度, 就能保证存在非平凡的  $I_h$ 。

**引理 5.33**

假设 (5-64) 成立, 且区间  $\hat{I}_h$  的长度为  $\sigma$ , 则区间  $I_h$  的长度至少为  $\frac{\sigma}{2}$ 。

*Proof.* 注意, 当 (5-64) 以及 (5-85) 成立时, 我们有

$$2\sqrt{s_h} \frac{(\gamma_k^2 + s_h)^{3/2}}{\gamma_k^2} |t_k - \hat{t}_k| \leq 2\sqrt{s_h} \frac{(\gamma_k^2 + s_h)^{3/2}}{\gamma_k^2} \sqrt{\frac{e_k}{\sigma_F}} \leq \frac{1}{8}\sigma, \quad 2\frac{s_h}{\gamma_k} e_k \leq \frac{1}{8}\sigma.$$

将上述不等式与  $\hat{I}_h$  的长度相结合, 即可证明结论。  $\square$

综上所述, 我们有以下定理。

**定理 5.34**

在某一迭代点  $x_k$ , 假设子问题 (5-2) 被求解以满足 条件 5.25 以及 条件 5.35, 则二分法在至多

$$O\left(\log\left(\frac{s_h U_\phi U_H}{h_{\min} \sigma}\right)\right) \quad (5-82)$$

次迭代内输出某个  $\delta_k$  使得  $\hat{h}_k(\delta_k) \in \hat{I}_h$  在容差  $\sigma$  内成立。

*Proof.* 综合前述所有结果, 我们注意到将  $h_k \in I_h$  的条件可以隐式地通过尝试  $\hat{h}_k(\delta_k) \in \hat{I}_h$  来实现。我们让二分法在  $\hat{h}_k(\delta_k) \notin \hat{I}_h$  时继续进行, 此时必有  $h_k(\delta_k) \notin I_h$ 。由于从  $\hat{I}_h$  长度为  $\sigma$  可保证  $|I_h| > \frac{\sigma}{2}$ , 因此二分法所需的算术运算次数与精确情况相同。  $\square$

### 5.A.3 处理困难情况

不同于使用精确特征值的方法, 不精确二分法过程与困难情况交织在一起。当  $\hat{t}_k = 0$  时, 根据 (5-59a)–(5-59d) 得到的不精确困难情况为

$$(H_k + \gamma_k I) \hat{v}_k = r_k, \quad (5-83a)$$

$$\hat{v}_k \perp r_k, \quad \phi_k^T \hat{v}_k = \sigma_k, \quad (5-83b)$$

$$F_k + \gamma_k I + e_k I \geq 0, \quad (5-83c)$$

其中  $\hat{t}_k = 0$  不再意味着我们获得了  $\lambda_1(H_k)$  的精确值。我们采用与<sup>[81]</sup> Algorithm 3 相同的扰动处理方法, 同时使用 Ritz 对。为完整起见, 我们在 算法 5-2 中给出该方法。记当前迭代为  $k$ , 随后迭代记为  $i$ :  $k, k+1, \dots, k+i, \dots$

**算法 5-4:** 困难情况的扰动处理

```

1 输入: 当前迭代  $k$ ,  $x_k \in \mathbb{R}^n$ ,  $g_k$ ,  $H_k$ ,  $\hat{h}_{k-1}, \delta_{k-1}$ , 其中满足  $g_k \perp \mathcal{S}_1$ ;
2 for  $i = 0, 1, \dots$  do
3     令
        
$$\phi_{k+1} = \phi_k + \left( \text{sign}(\sigma_k) \frac{\varsigma_\phi \gamma_k^2}{4\nu} \right) \cdot \hat{v}_k. \quad (5-84)$$

        计算区间
        
$$I_h := [\iota_2 \sqrt{\hat{h}_{k+i-1}}, \iota_3 \sqrt{\hat{h}_{k+i-1}}].$$

        repeat // 通过二分法进行内层迭代  $j$ , 参见 ??
4         求解 GHM 子问题
            
$$\min_{\| [v; t] \| \leq 1} \begin{bmatrix} v \\ t \end{bmatrix}^T \begin{bmatrix} \hat{h}_k & \phi_{k+i} \\ (\phi_{k+i})^T & \delta_{k+i,j} \end{bmatrix} \begin{bmatrix} v \\ t \end{bmatrix}$$

            得到解  $[\hat{v}_{k+i,j}; \hat{t}_{k+i,j}]$ ; 令  $\hat{d}_{k+i,j} = \hat{v}_{k+i,j} / \hat{t}_{k+i,j}$ , 并定义  $\hat{h}_{k+i} := (\theta_{k+i,j} / \|\hat{d}_{k+i,j}\|)^2$ ;
5         更新  $\delta_{k+i,j}$ , 令  $j = j + 1$ .
6     until  $\sqrt{\hat{h}_{k+i}} \in I_h$  在容差  $\sigma$  内成立;
7     令  $\hat{d}_{k+i} = \hat{d}_{k+i,j}$ ,  $\delta_{k+i} = \delta_{k+i,j}$ ;
8     计算
        
$$\rho_{k,i} = \frac{f(x_k + \hat{d}_{k+i}) - f(x_k)}{m_k(\hat{d}_{k+i}) - f(x_k)}.$$

        if  $\rho_{k+i} \geq \iota_1$  then
9         跳出
    
```

**条件 5.35**

 假设 Lanczos 方法一直执行, 直到不精确性  $e_k$  满足

$$e_k \leq \min \left\{ \frac{\gamma_k \sigma}{16\varsigma_h}, \frac{\sigma^2 \sigma_F}{256} \frac{\gamma_k^4}{(\gamma_k^2 + \varsigma_h)^3} \right\}. \quad (5-85)$$

虽然  $\hat{t}_k = 0$  不再直接产生  $H_k$  的特征对, 但若残差  $r_k, \sigma_k$  可容忍,  $\gamma_k, \hat{v}_k$  仍然可作为  $H_k$  左侧特征对的近似。

**引理 5.36**

假设发生  $\hat{t}_k = 0$ , 则有

$$\gamma_k \leq -\lambda_1(H_k). \quad (5-86)$$

此外, 假设  $\hat{v}_k = \alpha v_1 + s$ , 其中  $v_1$  为对应于  $\lambda_1(H_k)$  的特征向量, 且  $s \perp v_1$ , 则有

$$\alpha \geq \sqrt{1 - \frac{e_k}{\sigma_H}}, \quad (5-87)$$

其中  $\sigma_H := \lambda_2(H_k) - \lambda_1(H_k)$ .

*Proof.* 将 (5-83a) 两边左乘  $\hat{v}_k$ , 并注意到  $\hat{v}_k \perp r_k$ , 得到

$$\hat{v}_k^T H_k \hat{v}_k = -\gamma_k \|\hat{v}_k\|^2.$$

因此可得 (5-86). 为验证 (5-87), 注意前一引理表明

$$\|s\| \leq \sqrt{\frac{e_k}{\sigma_H}},$$

再结合  $\alpha^2 + \|s\|^2 = 1$ , 即可得到 (5-87).  $\square$

在精确情形下, 对于以下的每个迭代  $i$ , 我们基于先前的  $h_{k+i-1}$  对  $\phi_{k+i}$  进行扰动。当逐步增大  $h_{k+i}$  时, 我们使用以  $j$  为指标的二分法来寻找  $\delta_{k+i,j}$ 。我们将证明最终能产生一次**成功的迭代**; 一旦成功, 该迭代必满足 **假设 5.24**。我们接下来证明这些目标可以通过 Ritz 向量来实现。

**引理 5.37**

假设在第  $k$  次迭代时发生  $\hat{t}_k = 0$  且  $\hat{h}_{k+1}$  的搜索区间为  $[\ell, \nu]$ 。若按照 (5-84) 设定  $\phi_k$ , 并且不精确性  $e_{k+1}$  满足

$$e_{k+1} \leq e_k \leq \frac{-2\|\phi_k\| + \sqrt{4\|\phi_k\|^2 + \frac{\varsigma_\phi \gamma_k^2}{\nu} \left( \|\phi_k\| + \frac{\varsigma_\phi \gamma_k^2}{2\nu} \right)}}{2\|\phi_k\| + \frac{\varsigma_\phi \gamma_k^2}{\nu}}, \quad (5-88)$$

则后续的  $t_{k+1} \neq 0$ , 即困难情况被消除。

*Proof.* 我们采用反证法证明。假设  $t_{k+1} = 0$ , 记  $\hat{v}_{k+1} = \alpha_1 v_1 + s_1$ 。根据 **引理 5.36**, 有  $\alpha_1 \geq \sqrt{1 - \frac{e_{k+1}}{\sigma_H}}$ 。

于是有

$$\begin{aligned}
 \hat{v}_{k+1}^T \phi_{k+1} &= \left( \phi_k + \text{sign}(\sigma_k) \frac{\varsigma_\phi \gamma_k^2}{4\nu} \hat{v}_k \right)^T \hat{v}_{k+1} \\
 &= \phi_k^T \hat{v}_{k+1} + \text{sign}(\sigma_k) \frac{\varsigma_\phi \gamma_k^2}{4\nu} \hat{v}_k^T \hat{v}_{k+1} \\
 &= \phi_k^T \hat{v}_k + \phi_k^T (\hat{v}_{k+1} - \hat{v}_k) + \text{sign}(\sigma_k) \frac{\varsigma_\phi \gamma_k^2}{4\nu} \hat{v}_k^T \hat{v}_{k+1} \\
 &= \sigma_k + \phi_k^T (\hat{v}_{k+1} - \hat{v}_k) + \text{sign}(\sigma_k) \frac{\varsigma_\phi \gamma_k^2}{4\nu} \hat{v}_k^T \hat{v}_{k+1} \\
 &= \sigma_k + \phi_k^T (\alpha_1 v_1 + s_1 - \alpha v_1 - s) + \text{sign}(\sigma_k) \frac{\varsigma_\phi \gamma_k^2}{4\nu} (\alpha \alpha_1 + s^T s_1).
 \end{aligned}$$

在最后一步中，我们使用了  $\hat{v}_{k+1} = \alpha_1 v_1 + s_1$  以及  $\hat{v}_k = \alpha v_1 + s$ 。利用  $e_{k+1} \leq e_k$ ，可得

$$\begin{aligned}
 |\hat{v}_{k+1}^T \phi_{k+1}| &\geq \sigma_k + \frac{\varsigma_\phi \gamma_k^2}{4\nu} \left( \sqrt{1 - \frac{e_{k+1}}{\sigma_H}} \sqrt{1 - \frac{e_k}{\sigma_H}} - \|s\| \|s_1\| \right) - \|\phi_k\| \|(\alpha_1 - \alpha)v_1 + s_1 - s\| \\
 &\geq \sigma_k + \frac{\varsigma_\phi \gamma_k^2}{4\nu} \left( \sqrt{1 - \frac{e_{k+1}}{\sigma_H}} \sqrt{1 - \frac{e_k}{\sigma_H}} - \sqrt{\frac{e_k}{\sigma_H}} \sqrt{\frac{e_{k+1}}{\sigma_H}} \right) - \|\phi_k\| (|\alpha_1 - \alpha| + \|s_1\| + \|s\|) \\
 &\geq \sigma_k + \frac{\varsigma_\phi \gamma_k^2}{4\nu} \left( 1 - 2\frac{e_k}{\sigma_H} \right) - \|\phi_k\| \left( 1 - \sqrt{1 - \frac{e_k}{\sigma_H}} + 2\sqrt{\frac{e_k}{\sigma_H}} \right) \\
 &> \sigma_k + \frac{\varsigma_\phi \gamma_k^2}{4\nu} \left( 1 - 2\frac{e_k}{\sigma_H} \right) - \|\phi_k\| \left( \frac{e_k}{\sigma_H} + 2\sqrt{\frac{e_k}{\sigma_H}} \right).
 \end{aligned}$$

最后一不等式利用了当  $0 < x < 1$  时有  $1 - \sqrt{1 - x^2} \leq x^2$ 。由此，根据 (5-88) 可知

$$|\hat{v}_{k+1}^T \phi_{k+1}| > \sigma_k,$$

这与  $t_{k+1} = 0$  矛盾，从而证明了命题。 □

我们现在证明 算法 5-2 会逐步产生满足 假设 5.24 的  $\phi_{k+1}$ 。

#### 引理 5.38

假设在第  $k$  次迭代时发生困难情况，我们按照 (5-84) 对梯度进行扰动，并求得  $\delta_{k+1}$  使得  $\hat{h}_{k+1}(\delta_{k+1}) \in [\ell, \nu]$  且满足  $e_{k+1} \leq \sqrt{h_{\min}} \|\hat{d}_{k+1}\|$ ，那么在第  $(k+1)$  次迭代时，有

$$\|\nabla f(x_{k+1}) - \phi_{k+1}\| \leq \kappa_\phi \|\hat{d}_{k+1}\|^2.$$

*Proof.* 由 引理 5.37 已证明  $t_{k+1} \neq 0$ ，因此  $d_{k+1}$  是良定义的。从 (5-83c) 可知，

$$\sqrt{\hat{h}_{k+1}(\delta_{k+1})} \|\hat{d}_{k+1}\| \geq -\lambda_1(F_{k+1}) - e_{k+1} \geq -\lambda_1(H_k) - e_{k+1}.$$

整理得

$$\|\hat{d}_{k+1}\| \geq -\frac{\lambda_1(H_k)}{2\nu} \geq -\frac{\gamma_k}{2\nu},$$

由于在发生困难情况时不更新，则有

$$\|\nabla f(x_{k+1}) - \phi_{k+1}\| = \frac{s_\phi \gamma_k^2}{4\nu} \|\hat{d}_{k+1}\|^2.$$

令  $\kappa_\phi = \frac{s_\phi \gamma_k^2}{4\nu}$ ，便得证。 □

剩下的工作是证明在发生困难情况时，[算法 5-2](#) 最终会产生一次成功的迭代。该证明与原论文<sup>[81]</sup> Theorem 3.5 中的证明完全相同，这里省略。

**定理 5.39**

[算法 5-2](#) 最多需要  $\lfloor \log_{l_2} \frac{s_h}{\bar{h}_{k-1}(\delta_{k-1})} \rfloor + 1$  次迭代即可获得一次成功步。此外，[假设 5.24](#) 必然成立。



## 第六章 齐次同伦法

再此之前，我们都是对一般(二阶) Lipschitz 连续的函数进行分析。有时全局 Lipschitz 条件是不满足的，反而具有一些局部性质。在本节中，我们假设目标函数满足自协 Lipschitz 条件<sup>[111]</sup>，并提出**同伦 HSODM**，这是一种 HSODF 对此类问题的具体实现。

## 第一节 自协 Lipschitz 函数与同伦模型

我们首先给出定义如下。

**定义 6.1: 自协 (二阶) Lipschitz 条件**

若函数  $f$  满足以下条件，则称其为自协 Lipschitz: 存在常数  $\beta > 0$  使得对于任意  $x \in \text{dom}(f)$ ，有：

$$\|\nabla f(x+d) - \nabla f(x) - \nabla^2 f(x)d\| \leq \beta \cdot d^T \nabla^2 f(x)d, \quad (6-1)$$

其中  $d$  满足  $\|d\| \leq C$ ，且  $x+d \in \text{dom}(f)$ ，其中  $C > 0$ 。

我们称满足该条件的函数为**自协 (二阶) Lipschitz 函数**，或  $\beta$ -**自协 Lipschitz 函数**。实际上，许多函数(参见第 6.1.1 节)都满足上述自协 Lipschitz 条件。我们可以将自协 Lipschitz 条件看作是凸优化中一类“缩放 Lipschitz 条件 (scaled Lipschitz condition, SLC)”的简化版本<sup>[46,102]</sup>。事实上，该函数类在机器学习问题中广泛出现，尤其是在 Hessian 矩阵因高度稀疏的数据结构而退化的情况下。

## 6.1.1 自协 Lipschitz 函数的基本性质

我们首先提供一些基本性质。

**引理 6.2**

如果函数  $f$  满足自协 Lipschitz 条件，则  $f$  是凸的。

*Proof.* 不等式 (6-1) 表明对于所有  $x \in \text{dom}(f)$ ， $\nabla^2 f(x) \geq 0$ 。因此， $f$  是凸的。  $\square$

**Remark 6.3**

上述推论并不表明自协 Lipschitz 函数是**严格凸或强凸的**，这意味着我们允许 Hessian 的退化。

现在我们展示自协 Lipschitz 函数一些有用性质，使我们能够推导出更多“复杂”的例子。值得注

意的是，我们展示了自协 Lipschitz 函数空间在正标量乘法和求和下是封闭的。这一性质在变量的仿射变换中也得以保留。

**引理 6.4: 自协 Lipschitz 函数的和**

设  $f_i$  是满足 (6-1) 的  $\beta_i$ -自协 Lipschitz 函数，其中  $\beta_i \geq 0$ ，对于  $i = 1, \dots, m$ 。则  $\sum_{i=1}^m f_i$  是一个自协 Lipschitz 函数。

*Proof.* 根据 定义 6.1，对于任何点  $x \in \cap_{i=1}^m \text{dom}(f_i)$ ，有

$$\begin{aligned} & \left\| \sum_{i=1}^m \nabla f_i(x+d) - \sum_{i=1}^m \nabla f_i(x) - \left( \sum_{i=1}^m \nabla^2 f_i(x) \right) d \right\| \\ & \leq \sum_{i=1}^m \left\| \nabla f_i(x+d) - \nabla f_i(x) - \nabla^2 f_i(x) d \right\| \\ & \leq \max_{1 \leq i \leq m} \{\beta_i\} \cdot d^T \sum_{i=1}^m \nabla^2 f_i(x) d, \end{aligned}$$

当我们选择  $\|d\| \leq \min_{1 \leq i \leq m} C_i$  时。 □

**引理 6.5: 常数系数的自协 Lipschitz**

假设函数  $f$  满足  $\beta$ -自协 Lipschitz 条件，则对于任何系数  $c > 0$ ，函数  $c \cdot f$  是  $c\beta$ -自协 Lipschitz。

*Proof.* 自协 Lipschitz 函数的定义直接得出结果。 □

**引理 6.6: 复合函数**

复合函数  $f(x) = \phi(Ax - b)$  是自协 Lipschitz，如果  $\phi(\cdot)$  是  $\beta$ -自协 Lipschitz。

*Proof.* 注意到  $\nabla f(x) = A^T \nabla \phi(Ax - b)$  和  $\nabla^2 f(x) = A^T \nabla^2 \phi(Ax - b) A$ ，那么对于任何使得  $Ax - b \in \text{dom}(\phi)$  的点  $x$ ，我们有

$$\begin{aligned} & \left\| \nabla f(x+d) - \nabla f(x) - \nabla^2 f(x) d \right\| \\ & = \left\| A^T (\nabla \phi(Ax - b + Ad) - \nabla \phi(Ax - b) - \nabla^2 \phi(Ax - b) Ad) \right\| \\ & \leq \|A^T\| \cdot \left\| \nabla \phi(Ax - b + Ad) - \nabla \phi(Ax - b) - \nabla^2 \phi(Ax - b) Ad \right\| \\ & \leq \|A^T\| \cdot \beta \cdot (Ad)^T \nabla^2 \phi(Ax - b) (Ad) \\ & = \|A^T\| \beta \cdot d^T \nabla^2 f(x) d. \end{aligned}$$

□

此外，我们提供了确保自协 Lipschitz 的充分条件。

**引理 6.7: 充分条件**

设函数  $\phi(y)$ ,  $y \in \mathbb{R}^m$  是标准  $M$  二阶利普希茨连续且  $\mu$  强凸的，则函数  $f(x) = \phi(Ax - b)$  对于所有满足  $y = Ax - b$  在  $\phi$  的定义域内的  $x \in \mathbb{R}^n$ ，是自协 Lipschitz，其中  $A \in \mathbb{R}^{m \times n}$ ,  $m \leq n$  是一个具有秩  $m$  的常数系数矩阵。

*Proof.* 根据引理 6.6 的类似论证，得到

$$\begin{aligned} & \|\nabla f(x+d) - \nabla f(x) - \nabla^2 f(x)d\| \\ &= \|A^T(\nabla \phi(Ax - b + Ad) - \nabla \phi(Ax - b) - \nabla^2 \phi(Ax - b)Ad)\| \\ &\leq \|A^T\| \cdot \|\nabla \phi(Ax - b + Ad) - \nabla \phi(Ax - b) - \nabla^2 \phi(Ax - b)Ad\| \\ &\leq \|A^T\| \cdot \frac{M}{2} \|Ad\|^2 \leq \|A^T\| \cdot \frac{M}{2\mu} \cdot d^T(A^T \nabla^2 \phi(Ax - b)A)d = \frac{\|A^T\|M}{2\mu} \cdot d^T \nabla^2 f(x)d. \end{aligned}$$

设  $\beta = \frac{\|A^T\|M}{2\mu}$ ，则证明完成。  $\square$

**推论 6.8: 充分条件**

如果函数  $f$  是标准二阶利普希茨连续且  $\mu$  强凸的，则它是自协 Lipschitz。

我们可以利用上述结果验证逻辑回归问题：例 1.1。为了方便，我们重新写出该问题：

$$f(x) = \frac{1}{m} \sum_{i=1}^m \log(1 + e^{-b_i \cdot a_i^T x}) + \frac{\gamma}{2} \|x\|^2, \quad (6-2)$$

其中  $\gamma > \frac{2}{m} \sum_{i=1}^m \|a_i\|^2$ ,  $a_i \in \mathbb{R}^n$ ,  $b_i \in \{-1, 1\}$ ,  $i = 1, 2, \dots, m$ 。则函数  $f(x)$  满足自协 Lipschitz 条件。

*Proof.* 我们首先设

$$\nu = \lambda_{\max} \left( \frac{1}{m} \sum_{i=1}^m b_i^2 \cdot a_i a_i^T \right)$$

为最大特征值。然后定义单变量函数  $\phi(y) = \log(1 + e^{-y}) + \frac{\gamma}{2\nu} \cdot y^2$ ,  $y \in \mathbb{R}$ ，则  $\phi(y)$  是标准二阶利普希茨连续且  $\frac{\gamma}{\nu}$  强凸的。通过引理 6.7，这意味着

$$g_i(x) = \phi_i(b_i \cdot a_i^T x) = \log(1 + e^{-b_i \cdot a_i^T x}) + \gamma \cdot \frac{(b_i \cdot a_i^T x)^2}{2\nu},$$

对于所有  $i = 1, \dots, m$ ，都是自协 Lipschitz。接下来需要看的是

$$h(x) = x^T \left( \frac{\gamma}{2} I - \frac{\gamma}{2\nu} \frac{1}{m} \sum_{i=1}^m b_i^2 \cdot a_i a_i^T \right) x$$

是一个凸二次函数，因此也是自协 Lipschitz。使用加法规则，我们看到函数  $f$  是  $g_i(x)$  和  $h(x)$  的求和，即

$$f(x) = \frac{1}{m} \sum_{i=1}^m g_i(x) + h(x).$$

结合 引理 6.4 和 引理 6.5，我们得出  $f(x)$  满足自协 Lipschitz 条件。  $\square$

### 6.1.1.1 同伦模型

我们考虑满足 (6-1) 性质的同伦模型：

$$\min_{x \in \mathbb{R}^n} f(x) + \frac{\mu}{2} \|x\|^2. \quad (6-3)$$

该模型在<sup>[172]</sup>提出的路径跟踪法中被用于求解凸优化问题，通过构造一系列良性、严格凸的子问题来处理退化情况。该同伦模型具有以下性质。

#### 引理 6.9: <sup>[172]</sup>

设  $f$  为二阶连续可微的凸函数，且其值下界有界， $x^*$  为  $f(x)$  的最小  $\ell_2$  范数解。令  $x_\mu = \arg \min \{f(x) + \frac{\mu}{2} \|x\|^2\}$ ，则：

- (a) 对于任意  $\mu > 0$ ， $x_\mu$  是唯一的，并且当  $\mu$  变化时， $x_\mu$  形成一条连续路径；
- (b)  $f(x_\mu)$  是  $\mu$  的单调递增函数，而  $\|x_\mu\|$  是  $\mu$  的单调递减函数；
- (c) 当  $\mu \rightarrow 0^+$  时， $x_\mu$  收敛至  $x^*$ ；
- (d) 当  $\mu \rightarrow \infty$  时， $x_\mu \rightarrow 0$ 。

由于原文并未提供完整的分析，我们在给出简要证明。

*Proof.* 由于正则化目标函数  $f(x) + \frac{\mu}{2} \|x\|^2$  对于任意给定的  $\mu > 0$  都是强凸的，因此其最小化解  $x_\mu$  是唯一的。同时，由于  $x_\mu = -\nabla f(x_\mu)/\mu$ ，结合  $\nabla f(x_\mu)$  的连续性和  $\frac{1}{\mu}$  的连续性，可知  $x_\mu$  是连续函数，从而第一条性质成立。

对于第二条性质，取任意  $0 < \mu' < \mu$ ，有：

$$f(x_{\mu'}) + \frac{\mu'}{2} \|x_{\mu'}\|^2 < f(x_\mu) + \frac{\mu'}{2} \|x_\mu\|^2, \quad (6-4)$$

以及

$$f(x_\mu) + \frac{\mu}{2} \|x_\mu\|^2 < f(x_{\mu'}) + \frac{\mu}{2} \|x_{\mu'}\|^2. \quad (6-5)$$

将 (6-4) 和 (6-5) 两式相加并整理，得到：

$$\frac{\mu - \mu'}{2} \|x_{\mu'}\|^2 > \frac{\mu - \mu'}{2} \|x_\mu\|^2.$$

由于  $\mu - \mu' > 0$ , 可得  $\|x_{\mu'}\| > \|x_\mu\|$ , 即  $\|x_\mu\|$  是  $\mu$  的严格单调递减函数。将  $\|x_{\mu'}\| > \|x_\mu\|$  代入不等式 (6-4), 进一步得到:  $f(x_{\mu'}) < f(x_\mu)$ , 从而证明了第二条性质。

现在我们证明第三条性质。由定义,  $x^*$  是  $f(x)$  的最小  $\ell_2$  范数解, 因此  $\nabla f(x^*) = 0$ 。结合  $x_\mu = \arg \min \{f(x) + \frac{\mu}{2}\|x\|^2\}$ , 有:

$$\nabla f(x_\mu) - \nabla f(x^*) + \mu x_\mu = 0.$$

将两边同时乘以  $x_\mu - x^*$ , 利用  $f$  的凸性, 得到:

$$-\mu(x_\mu - x^*)^T x_\mu = (x_\mu - x^*)^T (\nabla f(x_\mu) - \nabla f(x^*)) \geq 0.$$

进一步推出:  $\|x_\mu\|^2 \leq x_\mu^T x^* \leq \|x^*\| \|x_\mu\|$ , 即对于任意  $\mu > 0$ , 有  $\|x_\mu\| \leq \|x^*\|$ 。由  $x^*$  的唯一性, 可得  $\lim_{\mu \rightarrow 0^+} x_\mu = x^*$ 。

对于第四条性质, 假设  $x_\mu \rightarrow z \in \mathbb{R}^n \neq 0$  当  $\mu \rightarrow \infty$ , 则可以选取  $\mu > \frac{2(f(0) - f(z))}{\|x^*\|^2 - \|z\|^2}$ , 从而得到:

$$f(z) + \frac{\mu}{2}\|z\|^2 > f(0).$$

该不等式与  $x_\mu = \arg \min \{f(x) + \frac{\mu}{2}\|x\|^2\}$  矛盾, 因此第四条性质成立。□

上述引理表明, 当  $\mu$  逐渐趋于 0 时, 轨迹  $\{x_\mu\}_{\mu \rightarrow 0}$  收敛至最优解  $x^*$ 。此外, 基于这些性质, 我们得到以下结果。

#### 推论 6.10: 同伦解的有界性

给定一个单调递减序列  $\{\mu_k\}_{k=0}^\infty$ , 使得  $\mu_k \rightarrow 0$ , 则对于任意  $x_{\mu_k} = \arg \min \{f(x) + \frac{\mu_k}{2}\|x\|^2\}$ , 有:  $\|x_{\mu_k}\| \leq \|x^*\|$ 。

该结果由 引理 6.9 的第二、三条直接推出。

类似于内点法, 我们可以通过最小化一系列  $\{\mu_k\}_{k=0}^\infty$  所定义的同伦模型, 并在每次迭代  $k$  使用牛顿法求解最优方程:

$$\nabla f(x) + \mu_k \cdot x = 0.$$

在算法框架中, 若惩罚参数  $\mu_k$  以线性速率递减, 例如:

$$\mu_{k+1} = \rho_k \cdot \mu_k, \quad 0 < \rho_k < 1,$$

则可以预期该方法具有全局线性收敛速率。由于该方法和内点法一般依赖于求解线性系统, 因此自然的想法是设计一种同伦 HSODM。我们观察到, 对于某个  $\mu$ , 牛顿法求解 (6-3) 的二阶模型  $m^H$ :

$$\begin{aligned} m^H(x, d) - f(x) &= \nabla f(x)^T d + \frac{1}{2} d^T \nabla^2 f(x) d + \frac{\mu}{2} \|x + d\|^2 \\ &= (\nabla f(x) + \mu \cdot x)^T d + \frac{1}{2} d^T (\nabla^2 f(x) + \mu I) d + \frac{\mu}{2} \|x\|^2. \end{aligned} \quad (6-6)$$

相比于 (3-8), (6-6) 额外包含当前点  $x$  相关的信息。因此, OHM 无法直接应用于 (6-6)。得益于 GHM 中  $\delta, \phi$  的灵活性, 我们可以对 (6-6) 进行齐次化:

$$F^H(x) := \begin{bmatrix} \nabla^2 f(x) & \nabla f(x) + \mu \cdot x \\ \nabla f(x)^T + \mu \cdot x^T & -\mu \end{bmatrix}. \quad (6-7)$$

构造  $\psi^H(v, t; F^H) = [v; t]^T F^H [v; t]$ , 则在适当缩放  $\mu$  并设定  $d := v/t$  且  $t \neq 0$  时, 该表达式等价于 (6-6):

$$m^H(x, d) - f(x) = \frac{1}{2t^2} \psi^H(v, t; F^H) + \frac{\mu}{2} (\|x\|^2 + \|[v; t]\|^2).$$

类似地, 我们可以求解带有单位球约束的 GHM:

$$\min_{\|[v; t]\| \leq 1} \psi^H(v, t; F^H). \quad (6-8)$$

本质上, 这是一个对称特征值问题, 因为 (6-8) 的最优解总是达到单位球面, 参见引理 4.17.

### 6.1.2 同伦 HSODM

利用 (6-7) 中定义的特定 GHM, 我们可以构造**同伦 HSODM** (算法 6-1), 其中  $\{\mu_k\}_{k=0}^\infty$  以线性速率递减。在外层迭代  $k$  处, 我们更新  $\mu_k$ 。然后, 对于每个中间目标, 我们应用一系列 GHMs (算法 6-2) 来计算一个近似的**中心**。我们用  $x_{k,j}$  表示 算法 6-1 中的迭代点, 其中  $k$  是外层迭代次数,  $j$  计算内层 GHM 次数。

---

#### 算法 6-1: 同伦 HSODM

---

- 1 **初始化:** 初始点  $x_{0,0} = 0$ , 迭代计数  $k = 0$ , 参数  $\mu_0 = 2(\beta + 1)(1 + \|g_{0,0}\|^2)$ ;
  - 2 **for**  $k = 0, 1, \dots, K$  **do**
  - 3     计算  $(x_{k,j}, \rho_k) = \text{iACGHM}(x_{k,0}, \mu_k)$ ;
  - 4     更新  $\mu_{k+1} = \rho_k \cdot \mu_k$ ;
  - 5     设定  $x_{k+1,0} := x_{k,j}$ ;
  - 6 **输出**  $x_{K+1,0}$ .
- 

值得注意的是, 对于每个  $\mu_k$ , 算法 6-2 在迭代点满足**近似居中条件** (由  $\mu_k$  和自协 Lipschitz 常数  $\beta$  控制, 见行 3) 时终止。我们证明 算法 6-2 具有二次收敛速率, 并且对于每个  $\mu_k$  至多需要 2 次 GHM 计算。此外, 类似于内点法, 每个  $\mu_k$  关联的最后一内层迭代点  $x_{k,j}$  也位于  $x_{\mu_k}$  的邻域内。进一步地, 随着  $\mu_k \rightarrow 0$ , 该邻域的宽度逐渐缩小, 并且邻域内的每个点相对于  $x_{\mu_k}$  都具有**固定的**偏差界限。

**算法 6-2:** 通过 GHM 计算的不精确近似中心 (iACGHM)

---

```

1 输入:  $x_{k,0}, \mu_k$ ;
2 for  $j = 0, \dots, \mathcal{T}_k$  do
3   if  $\|g_{k,j} + \mu_k \cdot x_{k,j}\| \leq \frac{\mu_k}{1+3(\beta+1)}$  then
4     计算  $\rho_k = \frac{3(\beta+1)(1+\|x_{k,j}\|)}{1+3(\beta+1)(1+\|x_{k,j}\|)}$ ;
5     返回  $(x_{k,j}, \rho_k)$ ;
6   else
7     求解 GHM 子问题的解  $[v_{k,j}; t_{k,j}]$ 

```

$$\min_{\|[v;t]\| \leq 1} \begin{bmatrix} v \\ t \end{bmatrix}^T \begin{bmatrix} H_{k,j} & g_{k,j} + \mu_k \cdot x_{k,j} \\ (g_{k,j} + \mu_k \cdot x_{k,j})^T & -\mu_k \end{bmatrix} \begin{bmatrix} v \\ t \end{bmatrix}; \quad (6-9)$$

```

    设定  $d_{k,j} = v_{k,j}/t_{k,j}$  并更新  $x_{T,j+1} = x_{k,j} + d_{k,j}$ ;

```

---

## 第二节 收敛性分析

在本小节中, 我们分析 [算法 6-1](#) 的收敛性质。我们证明, 如果  $\mu_k$  以几何速率递减, 则算法具有线性收敛速率。此外, 对于每个  $\mu_k$ , 相应的同伦模型可以通过有限次 GHM (6-7) 进行求解。与用于非凸和二阶 Lipschitz 连续凸函数的 [算法 5-1](#) 不同, [算法 6-1](#) 不需要 [假设 5.24](#)。

## 6.2.1 近似居中条件的性质

我们首先给出关于近似居中条件的结果。

**引理 6.11:** 中心路径的宽度

设  $f$  为凸函数, 若迭代点  $x_{k,j}$  满足近似居中条件:

$$\|g_{k,j} + \mu_k \cdot x_{k,j}\| \leq \frac{\mu_k}{1 + 3(\beta + 1)},$$

则有

$$\|x_{k,j} - x_{\mu_k}\| \leq \frac{1}{1 + 3(\beta + 1)},$$

其中  $x_{\mu_k} = \arg \min \{f(x) + \frac{\mu_k}{2} \|x\|^2\}$ 。

*Proof.* 由于目标函数  $f(x) + \frac{\mu_k}{2} \|x\|^2$  是  $\mu_k$ -强凸的, 根据<sup>[125]</sup> Theorem 2.1.10, 可得:

$$\mu_k \|x_{k,j} - x_{\mu_k}\| \leq \|g_{k,j} + \mu_k \cdot x_{k,j}\|.$$

结合近似居中条件，进一步得到：

$$\|x_{k,j} - x_{\mu_k}\| \leq \frac{1}{1 + 3(\beta + 1)}.$$

□

### 6.2.2 GHM 的基本特性

接下来，我们讨论特定 GHM (6-9) 的基本性质。特别地，下面引理中的 (a) 说明了困难情况不会发生。

#### 引理 6.12: 同伦 GHM 的最优性条件

设  $f$  满足自协 Lipschitz 条件。对于  $\mu_k > 0$ ，令  $([v_{k,j}; t_{k,j}], -\theta_{k,j})$  为 GHM (6-9) 的最优原对偶解，则：

(a)  $t_{k,j} \neq 0, \theta_{k,j} > 0$ ;

(b)  $\theta_{k,j} - \mu_k \leq \|g_{k,j} + \mu_k \cdot x_{k,j}\|$ ;

(c) 令  $d_{k,j} = v_{k,j}/t_{k,j}$ ，则  $d_k$  满足：

$$\|d_{k,j}\| \leq \frac{\|g_{k,j} + \mu_k \cdot x_{k,j}\|}{\mu_k}, \quad d_{k,j}^T H_{k,j} d_{k,j} \leq \frac{\|g_{k,j} + \mu_k \cdot x_{k,j}\|^2}{\mu_k}.$$

*Proof.* 对于 (a)，在同伦 HSODM 的 GHM (6-8) 中，对角线元素设为  $\delta_{k,j} = -\mu_k < 0$ 。由于  $f(\cdot)$  是凸函数，并结合引理 4.8 的结果，有  $\delta_{k,j} < \tilde{\alpha}_1$ ，从而保证  $t_{k,j} \neq 0$ 。由引理 4.17，单位球约束始终处于活动状态，意味着  $\lambda_1(F_{k,j}) < 0$ ，因此  $\theta_{k,j} > 0$ 。

(b) 由引理 4.5 直接推出。

(c) 由 (3-10)， $d_k$  满足：

$$H_{k,j} d_{k,j} + \theta_{k,j} d_{k,j} = -g_{k,j} - \mu_k \cdot x_{k,j},$$

进而可得：

$$\|H_{k,j} d_{k,j}\|^2 + \theta_{k,j}^2 \|d_{k,j}\|^2 + 2\theta_{k,j} d_{k,j}^T H_{k,j} d_{k,j} = \|g_{k,j} + \mu_k \cdot x_{k,j}\|^2.$$

由于交叉项  $d_{k,j}^T H_{k,j} d_{k,j} \geq 0$  且  $\theta_{k,j} \geq -\delta_{k,j} = \mu_k > 0$ ，可得：

$$\|H_{k,j} d_{k,j}\| \leq \|g_{k,j} + \mu_k \cdot x_{k,j}\|, \quad \theta_{k,j} \|d_{k,j}\| \leq \|g_{k,j} + \mu_k \cdot x_{k,j}\|.$$

于是  $d_{k,j}$  的范数满足：

$$\|d_{k,j}\| \leq \frac{\|g_{k,j} + \mu_k \cdot x_{k,j}\|}{\theta_{k,j}} \leq \frac{\|g_{k,j} + \mu_k \cdot x_{k,j}\|}{\mu_k},$$



进一步推出交叉项满足：

$$d_{k,j}^T H_{k,j} d_{k,j} \leq \|d_{k,j}\| \cdot \|H_{k,j} d_{k,j}\| \leq \frac{\|g_{k,j} + \mu_k \cdot x_{k,j}\|^2}{\mu_k}.$$

证毕。  $\square$

在接下来的两个引理中，我们证明对于每个固定的  $\mu_k > 0$ ，对应的内层问题具有二次收敛性。这需要分别讨论初始  $\mu_0$  和后续的  $\mu_k, k \geq 1$ 。注意，在第一次迭代中，我们初始化  $x_{0,0} = 0$ 。相比之下，其余迭代将**热启动**于上一次的最终点，即当  $x_{k,j}$  满足 [算法 6-2](#) 中的近似居中条件时，设定  $x_{k+1,0} := x_{k,j}$ 。在这两种情况下，二次收敛速率进一步导致内层问题在有限步内收敛。

### 6.2.2.1 初始 $\mu_0$ 的选择

在初始迭代中，我们需要选择合适的  $\mu_0$  以确保二次收敛性。由 [引理 6.9](#) 的 (c) 部分可知，当  $\mu_k \rightarrow \infty$  时， $x^*$  充分接近 0。因此，考虑到我们选择  $x_{0,0} = 0$ ，直觉上  $\mu_0$  应该足够大。以下引理给出了如何选择合适的初始  $\mu_0$ 。

#### 引理 6.13: 初始 $\mu_0$ 下的二次收敛

令  $x_{0,0} = 0$ ，并令序列  $\{x_{0,j}\}$  通过以下方式更新：

$$x_{0,j+1} = x_{0,j} + d_{0,j}, \quad d_{0,j} = v_{0,j}/t_{0,j},$$

其中  $[v_{0,j}; t_{0,j}]$  为 [\(6-9\)](#) 在迭代点  $x_{0,j}$  处的解。若  $\mu_0 \geq 2(\beta + 1) \cdot \max\{1, \|g_{0,0}\|^2\}$ ，则残差误差  $e_{0,j} = \|g_{0,j} + \mu_0 \cdot x_{0,j}\|$  具有二次收敛性，即：

$$e_{0,1} \leq \frac{1}{2}, \text{ 且 } e_{0,j+1} \leq e_{0,j}^2, \quad \forall k \geq 1.$$

*Proof.* 首先证明  $e_{0,1} < 1$ 。注意到  $x_{0,1} = x_{0,0} + d_{0,0} = d_{0,0}$ ，结合 [引理 6.12](#) 可得：

$$e_{0,0} = \|g_{0,0}\|, \quad 0 < \theta_{0,0} - \mu_0 \leq e_{0,0}, \quad \|d_{0,0}\| \leq e_{0,0}/\mu_0, \quad \text{且 } d_{0,0}^T H_{0,0} d_{0,0} \leq e_{0,0}^2/\mu_0. \quad (6-10)$$

由  $e_{0,j}$  的定义，得到：

$$\begin{aligned} e_{0,1} &= \|g_{0,1} + \mu_0 \cdot x_{0,1}\| = \|g_{0,1} + \mu_0 \cdot d_{0,0}\| \\ &= \|g_{0,1} - g_{0,0} - H_{0,0} d_{0,0} + g_{0,0} + H_{0,0} d_{0,0} + \mu_0 \cdot d_{0,0}\| \\ &\leq \|g_{0,1} - g_{0,0} - H_{0,0} d_{0,0}\| + \|g_{0,0} + H_{0,0} d_{0,0} + \mu_0 \cdot d_{0,0}\| \\ &\leq \beta \cdot d_{0,0}^T H_{0,0} d_{0,0} + |\theta_{0,0} - \mu_0| \cdot \|d_{0,0}\| \\ &\stackrel{(6-10)}{\leq} (\beta + 1) \cdot \frac{e_{0,0}^2}{\mu_0} \leq \frac{1}{2} < 1, \end{aligned} \quad (6-11)$$

其中最后一步利用了  $\mu_0 \geq 2(\beta + 1) \cdot \max \{1, \|g_{0,0}\|^2\}$ 。对于该引理的第二部分，我们可使用类似的推导：

$$e_{0,j+1} = \|g_{0,j+1} + \mu_0 \cdot x_{0,j+1}\| \leq (\beta + 1) \cdot \frac{e_{0,j}^2}{\mu_0} \leq e_{0,j}^2, \quad (6-12)$$

因此，二次收敛速率得证。  $\square$

#### 推论 6.14: 初始阶段的迭代次数界

在初始迭代  $k = 0$  时，iACGHM 内的迭代次数  $j$  上界为：

$$\mathcal{T}_0 = \left\lceil \log_2 \left( \frac{\max \{\log(1 + 3(1 + \beta)) - \log \mu_0, \log 2\}}{\log 2} \right) \right\rceil + 1 \leq 2.$$

*Proof.* 由于  $e_{0,1} \leq \frac{1}{2}$  且  $e_{0,j+1} \leq e_{0,j}^2$ ，设  $e_{0,j} \leq \frac{\mu_0}{1+3(1+\beta)}$ ，可得：

$$j \geq \log_2 \left( \frac{\max \{\log(1 + 3(1 + \beta)) - \log \mu_0, \log 2\}}{\log 2} \right) + 1,$$

进而推出：

$$\mathcal{T}_0 = \left\lceil \log_2 \left( \frac{\max \{\log(1 + 3(1 + \beta)) - \log \mu_0, \log 2\}}{\log 2} \right) \right\rceil + 1.$$

由于：

$$\mu_0 \geq 2(\beta + 1) \cdot \max \{1, \|g_{0,0}\|^2\} \geq 2(\beta + 1),$$

可得：

$$\mathcal{T}_0 \leq \left\lceil \log_2 \left( \frac{\log 2.5}{\log 2} \right) \right\rceil + 1 = 2.$$

证毕。  $\square$

给定足够大的  $\mu_0$ ，推论 6.14 说明我们可以自然地得到一个满足近似居中条件的迭代点  $x_{0,j}$ 。对于后续迭代，我们证明当罚因子  $\mu_k$  线性递减时，二次收敛性依然成立。

#### 引理 6.15: 后续迭代的二次收敛性

对于  $k \geq 1$ ，序列  $\{x_{k,j}\}$  通过以下方式更新：

$$x_{k,j+1} = x_{k,j} + d_{k,j}, \quad d_{k,j} = v_{k,j}/t_{k,j},$$

其中  $[v_{k,j}; t_{k,j}]$  为 (6-9) 在迭代点  $x_{k,j}$  处的解。定义  $e_{k,j} = \|g_{k,j} + \mu_k \cdot x_{k,j}\|$ ，则  $\frac{\beta+1}{\mu_k} e_{k,j}$  具有二次收敛性，即：

$$\frac{\beta+1}{\mu_k} e_{k,0} \leq \frac{2}{3}, \quad \frac{\beta+1}{\mu_k} e_{k,j+1} \leq \left( \frac{\beta+1}{\mu_k} e_{k,j} \right)^2, \quad \forall k \geq 1.$$

*Proof.* 由算法机制, 对于  $k \geq 1$  的初始点  $x_{k,0}$ , 必然有:

$$\begin{aligned} \|g_{k,0} + \mu_{k-1} \cdot x_{k,0}\| &\leq \frac{\mu_{k-1}}{1 + 3(\beta + 1)}, \\ \rho_{k-1} &= \frac{3(\beta + 1)(1 + \|x_{k,0}\|)}{1 + 3(\beta + 1)(1 + \|x_{k,0}\|)}, \\ \mu_k &= \rho_{k-1} \cdot \mu_{k-1}. \end{aligned} \quad (6-13)$$

注意:

$$\frac{1 - \rho_{k-1}}{\rho_{k-1}} = \frac{1}{3(\beta + 1)(1 + \|x_{k,0}\|)} \quad (6-14)$$

于是,

$$\begin{aligned} \frac{\beta + 1}{\mu_k} e_0 &= \frac{\beta + 1}{\mu_k} \cdot \|g_{k,0} + \mu_k \cdot x_{k,0}\| \\ &= \frac{\beta + 1}{\mu_k} \cdot \|g_{k,0} + \mu_{k-1} \cdot x_{k,0} - (1 - \rho_{k-1})\mu_{k-1} \cdot x_{k,0}\| \\ &\leq \frac{\beta + 1}{\mu_k} \cdot \|g_{k,0} + \mu_{k-1} \cdot x_{k,0}\| + \frac{(\beta + 1)(1 - \rho_{k-1})\mu_{k-1}}{\mu_k} \cdot \|x_{k,0}\| \\ &\leq \frac{\beta + 1}{\mu_k} \cdot \frac{\mu_{k-1}}{1 + 3(\beta + 1)} + \frac{(1 - \rho_{k-1})\mu_{k-1}}{\mu_k} \cdot (\beta + 1)\|x_{k,0}\| \end{aligned} \quad (6-15a)$$

$$\leq \frac{1}{3} + \frac{\|x_{k,0}\|}{3(1 + \|x_{k,0}\|)} \leq \frac{2}{3}, \quad (6-15b)$$

其中 (6-15a) 由 (6-14) 得出, 因此第一部分证明完成。

对于第二部分, 由 引理 6.12 中 GHM 的性质, 可得:

$$\|d_{k,j}\| \leq e_{k,j}/\mu_k, \quad d_{k,j}^T H_{k,j} d_{k,j} \leq e_{k,j}^2/\mu_k, \quad |\theta_{k,j} - \mu_k| \leq e_{k,j}. \quad (6-16)$$

将上述不等式代入  $\frac{\beta+1}{\mu_k} e_{k,j+1}$ , 可得:

$$\begin{aligned} \frac{\beta + 1}{\mu_k} e_{k,j+1} &= \frac{\beta + 1}{\mu_k} \cdot \|g_{k,j+1} + \mu_k \cdot x_{k,j+1}\| \\ &\leq \frac{\beta + 1}{\mu_k} \cdot \left( \beta \cdot d_{k,j}^T H_{k,j} d_{k,j} + |\theta_{k,j} - \mu_k| \cdot \|d_{k,j}\| \right) \leq \left( \frac{\beta + 1}{\mu_k} e_{k,j} \right)^2. \end{aligned} \quad (6-17)$$

其中第一步使用了与 (6-11) 相同的技术, 第二步由 (6-16) 得出。  $\square$

#### 推论 6.16: 后续迭代的内层迭代次数上界

在  $k \geq 1$  的迭代中, 内层迭代次数  $j$  的上界为:

$$\mathcal{T} := \mathcal{T}_k = \left\lceil \log_2 \left( \frac{\log(1 + 3(\beta + 1)) - \log(\beta + 1)}{\log 3 - \log 2} \right) \right\rceil \leq 2.$$

*Proof.* 轻度滥用符号, 设  $E_{k,j} = \frac{\beta+1}{\mu_k} e_{k,j}$ , 因此:

$$E_{k,0} \leq \frac{2}{3}, \quad E_{k,j+1} \leq E_{k,j}^2.$$

由近似居中条件, 可知:

$$E_{k,j} \leq \frac{\beta+1}{1+3(\beta+1)},$$

这意味着:

$$e_{k,j} = \frac{\mu_k}{\beta+1} E_{k,j} \leq \frac{\mu_k}{1+3(\beta+1)}.$$

因此,

$$j \geq \log_2 \left( \frac{\log(1+3(\beta+1)) - \log(\beta+1)}{\log 3 - \log 2} \right).$$

类似地,

$$\begin{aligned} \mathcal{T}_k &= \left\lceil \log_2 \left( \frac{\log(1+3(\beta+1)) - \log(\beta+1)}{\log 3 - \log 2} \right) \right\rceil \\ &\leq \left\lceil \log_2 \left( \frac{\log(4(\beta+1)) - \log(\beta+1)}{\log 3 - \log 2} \right) \right\rceil \leq \left\lceil \log_2 \left( \frac{\log 4}{\log 3 - \log 2} \right) \right\rceil = \lceil 1.77 \rceil = 2. \end{aligned}$$

证毕.  $\square$

在两种情况下 ( $k=0, k \geq 1$ ), 均表明内层循环是有限收敛的。注意, 对于  $k \geq 1$ , 我们的估计是统一的, 因此可直接令  $\mathcal{T}$  为 GHM 的查询次数。

在证明迭代次数的上界之前, 我们首先给出  $\rho_k$  的统一上界, 它同伦 HSODM 的收敛率分析中起着重要作用。

#### 引理 6.17: $\rho_k$ 的上界

存在常数  $\tau \in (0, 1)$  使得对于所有  $k \geq 0$ , 有  $\rho_k \leq \tau$ .

*Proof.* 由  $\rho_k = \frac{3(\beta+1)(1+\|x_{k,j}\|)}{1+3(\beta+1)(1+\|x_{k,j}\|)}$  的定义可知, 我们只需找到  $x_{k,j}$  的统一上界。由 [引理 6.11](#) 和 [推论 6.10](#) 可得:

$$\|x_{k,j}\| \leq \|x_{k,j} - x_{\mu_k}\| + \|x_{\mu_k}\| \leq \frac{1}{1+3(\beta+1)} + \|x^*\|.$$

因此, 设

$$\tau = \frac{3(\beta+1) \left( 1 + \frac{1}{1+3(\beta+1)} + \|x^*\| \right)}{1+3(\beta+1) \left( 1 + \frac{1}{1+3(\beta+1)} + \|x^*\| \right)}$$

即可完成证明.  $\square$

#### 引理 6.18: 外层迭代次数上界

假设  $f$  满足自协 Lipschitz 条件。对于任意给定的  $\epsilon > 0$ , 至多经过

$$K = \left\lceil \log_\tau \left( \frac{(1+3(\beta+1))\epsilon}{2(\beta+1)(1+\|\nabla f(0)\|^2)((3\beta+4)\|x^*\|+2)} \right) \right\rceil$$

次迭代, 其中  $\tau$  由 [引理 6.17](#) 定义, 最终的输出迭代点  $x_{K+1,0}$  满足  $\|\nabla f(x_{K+1,0})\| \leq \epsilon$ 。

*Proof.* 由 [算法 6-1](#) 可知,  $x_{K+1,0} = x_{K,j}$ , 且满足近似居中条件。结合 [推论 6.10](#), 可得:

$$\begin{aligned} \|\nabla f(x_{K+1,0})\| &= \|g_{k,j}\| \leq \|g_{k,j} + \mu_K \cdot x_{K,j}\| + \mu_K \|x_{K,j}\| \\ &\leq \frac{\mu_K}{1+3(\beta+1)} + \mu_K \|x_{K,j}\| \\ &\leq \left( \frac{2}{1+3(\beta+1)} + \|x^*\| \right) \cdot \mu_K. \end{aligned} \quad (6-18)$$

结合 [引理 6.17](#),  $\mu_K$  有上界:

$$\mu_K = \rho_{K-1} \cdot \mu_{K-1} \leq \tau \cdot \mu_{K-1} \leq \tau^K \cdot \mu_0. \quad (6-19)$$

将上式代入 (6-18), 可得:

$$\|\nabla f(x_{K+1,0})\| \leq \left( \frac{2}{1+3(\beta+1)} + \|x^*\| \right) \cdot \tau^K \cdot \mu_0 \leq \epsilon,$$

其中最后一个不等式由  $K \geq \log_\tau \left( \frac{(1+3(\beta+1))\epsilon}{2(\beta+1)(1+\|\nabla f(0)\|^2)((3\beta+4)\|x^*\|+2)} \right)$  以及  $\mu_0 = 2(\beta+1)(1+\|\nabla f(0)\|^2)$  推得, 证毕。  $\square$

与之前相同, 我们建立同伦 HSODM 所求解的 GHM 总数。

#### 定理 6.19: 同伦 HSODM 的复杂度

假设  $f$  满足自协 Lipschitz 条件。对于任意给定的  $\epsilon > 0$ , 记  $\mathcal{K}_\psi$  为 [算法 6-1](#) 返回一个近似全局最优解所需的 GHM 总数。则:

$$\mathcal{K}_\psi = \left\lceil 2 \log_\tau \left( \frac{(1+3(\beta+1))\epsilon}{2(\beta+1)(1+\|\nabla f(0)\|^2)((3\beta+4)\|x^*\|+2)} \right) \right\rceil.$$

*Proof.* 该结果直接由 [引理 6.18](#)、[推论 6.14](#) 和 [推论 6.16](#) 推得。  $\square$

通常情况下, 自协 Lipschitz 参数  $\beta$  的具体数值在实践中是未知的。然而, 由于 [算法 6-1](#) 的机制, 我们可以用较大的  $\beta$  代替其真实值, 同时仍然保持线性收敛 (详见 [引理 6.17](#))。因此, 一个直接的启发式方法是在给定点  $x$  处利用小扰动  $d$  计算近似估计值:

$$\frac{\|\nabla f(x+d) - \nabla f(x) - \nabla^2 f(x)d\|}{d^T \nabla^2 f(x)d},$$

并基于此构造一个上界。我们用以下备注来总结本节讨论的内容。

**Remark 6.20:** 需

强调的是, 这里我们并不假设水平集是有界的, 这在凸优化复杂度分析中被广泛应用<sup>[121,125]</sup>。此外, 算法 6-1 通过找到一个稳定点进一步隐含了函数值的收敛。具体而言, 对于凸函数, 我们知道:

$$f(x) - f^* \leq \|\nabla f(x)\| \cdot \|x - x^*\|.$$

但反之, 即已知  $f(x) - f^* \leq \epsilon$ , 如何保证梯度范数足够小并不是一个平凡问题, 详见<sup>[62,124]</sup>。最后值得注意的是, 自协 Lipschitz 函数并不一定是强凸的 (甚至不是严格凸的), 然而, 通过同伦 HSODM 仍然可以建立全局线性收敛性。

## 第三节 数值实验

我们保持与第 5 章, 为了比较, 我们自己实现了同伦 HSODM (Homotopy-HSODM), 以及一些非精确牛顿法 (iNewton-Grad).

6.3.1  $\ell_2$  正则化逻辑回归

这里, 我们回到开头提到的例子例 1.1, 我们提供了求解  $\ell_2$  正则化逻辑回归 (6-2) 的初步数值结果。我们选取了 LIBSVM 库中的两个退化高维数据集: rcv1 和 news20。问题的初始点随机选取自正态分布:

$$x_0 \sim \mathbf{N}(0, 100 \cdot I_n)$$

以确保初始点远离局部最优解。算法在迭代点  $x_k$  满足  $\|g_k\| \leq 10^{-8}$  时终止。

由于该问题是凸的且高度退化, 我们将其与不精确正则化牛顿法进行比较。我们使用梯度范数作为正则化项, 即 iNewton-Grad 在每次迭代中计算:

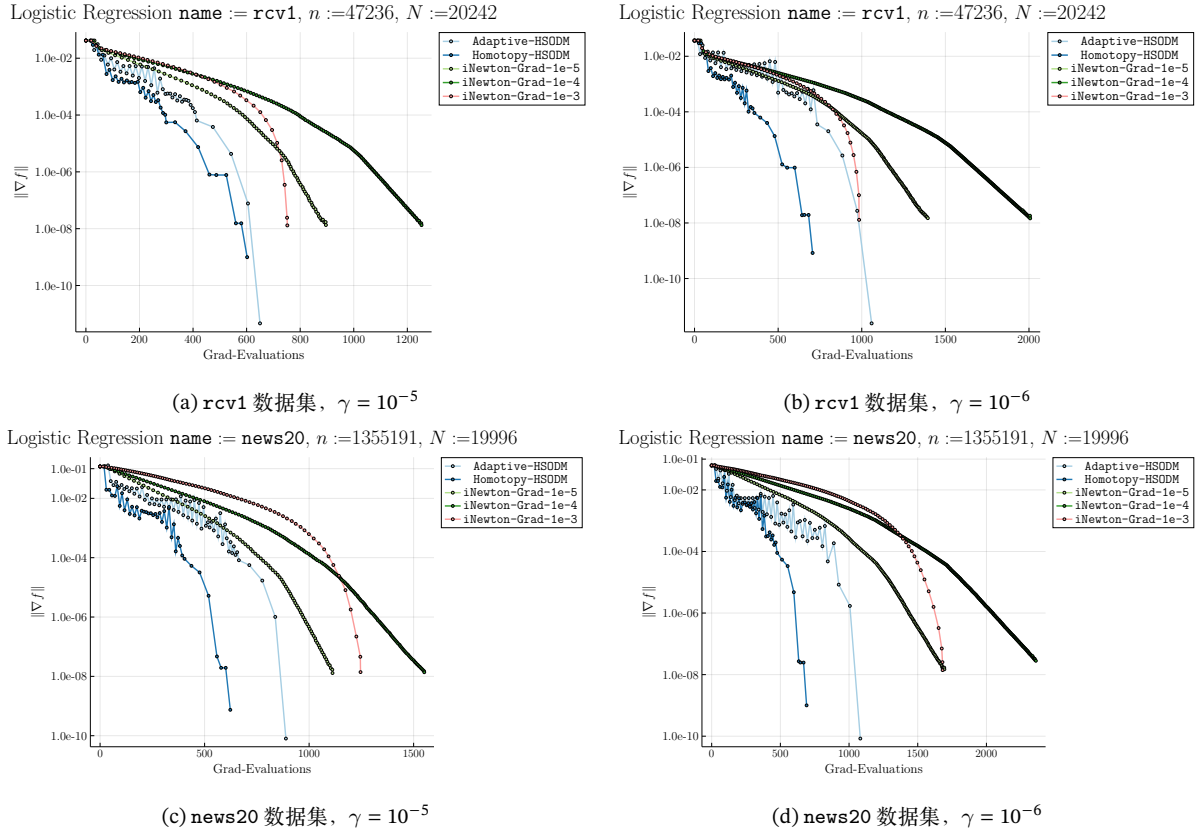
$$\left( H_k + \sigma \|g_k\|^{1/2} I \right) d_k = -g_k. \quad (6-20)$$

为简单起见, 我们尝试固定正则化参数:  $\sigma \in \{10^{-5}, 10^{-4}, 10^{-3}\}$ 。由于  $n$  较大, 对 Hessian 矩阵的操作可能成本较高, 因此在 iNewton-Grad 中使用共轭梯度法 (CG) 求解线性系统, 并在迭代满足以下条件时停止:

$$\left\| \left( H_k + \sigma \|g_k\|^{1/2} I \right) d_k + g_k \right\| \leq \min\{10^{-4}, \zeta \|g_k\|\}, \zeta \approx \Theta(10^3),$$

该容差选择受到<sup>[44]</sup>的启发。相对于  $\|g_k\|$  的容差在达到高精度时生效, 此时我们稍微收紧容差以防止算法陷入停滞, 否则则使用较宽松的精度要求。在 GHM 求解的特征值问题中, 我们在 Lanczos 方法中采用相同的容差策略。所有方法均使用回溯线搜索算法。

在图 6.3.1 中, 我们展示了梯度范数随梯度计算次数的变化轨迹。同样地, 我们将每次 Hessian-向

图 6.3.1  $\ell_2$  正则化逻辑回归问题上不同 SOM 方法的性能表现。

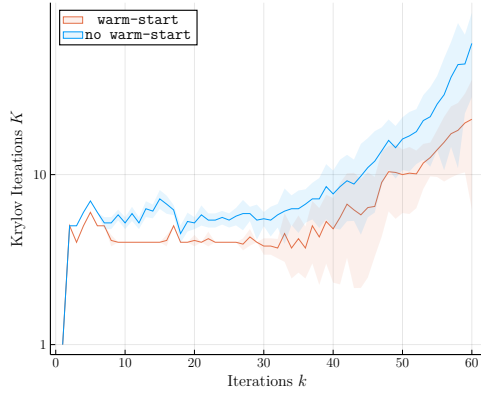
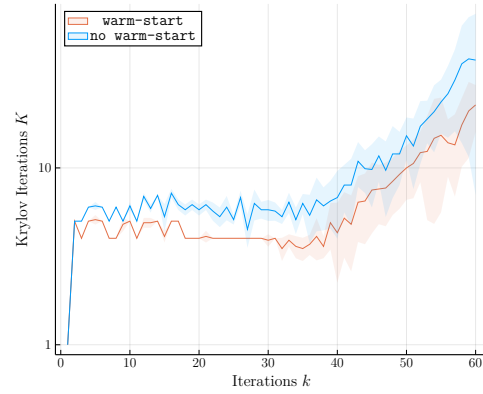
量乘法计算视为两次梯度计算。结果表明，当正则化参数  $\gamma$  足够大且 (6-20) 中的  $\sigma$  选择得当时，Homotopy-HSODM、Adaptive-HSODM 和 iNewton-Grad 的表现相近。若问题变得更加退化 ( $\gamma$  变小)，所有方法的收敛速度均有所下降，而 Homotopy-HSODM 似乎具有最强的稳健性。这一发现的可能原因是 Homotopy-HSODM 依赖于一致性条件 (6-1)，而不是通常的 Lipschitz 条件。

### 6.3.1.1 热启动的优势

在 Homotopy-HSODM 中，我们表明迭代点在某种意义上是连续的。由于 GHM 求解使用了 Lanczos 方法 (Lanczos)，自然地，我们可以利用上一次迭代的解  $[v_{k-1}; t_{k-1}]$  作为当前特征值问题的**热启动**，即让 Lanczos 方法以  $[v_{k-1}; t_{k-1}]$  作为初始点。

我们针对这一策略进行初步实验。在相同的数据集上，我们比较 Homotopy-HSODM 在不使用和使用上次迭代特征向量热启动的情况下 Krylov 迭代次数  $K$ ，以达到所需的  $10^{-8}$  精度。图 6.3.2 展示了数据集 rcv1 和 news20 在主迭代  $x_k$  过程中的 Krylov 迭代次数变化情况。

我们的初步结果表明，在 GHM 求解过程中，Homotopy-HSODM 通过利用先前特征向量显著减少

Warm-start for Homotopy HSODM on `name := rcv1`Warm-start for Homotopy HSODM on `name := news20`图 6.3.2  $\ell_2$  正则化逻辑回归问题上 Homotopy-HSODM 采用热启动策略的性能表现。

了 Krylov 迭代次数。这一策略的进一步优化留待未来研究。



## 第七章 齐次框架在交换市场中的应用

前面的章节中，我们讨论了基于齐次模型的二阶方法。这些算法已经足够解决绪论中提到的一些问题，但如果将这些二阶方法应用在例 1.3 上，无法得到满意的求解速度，主要原因是 Fisher 模型的对偶问题含有非负约束，同时 Fisher 模型的间接效用函数不具有一般欧式空间的 Lipschitz 连续性（一阶、二阶均不满足）。

本章中，我们设计一种基于齐次模型的内点法，该内点法可以求解更一般的、允许仿射约束的问题。我们将该内点法定制到求解 Fisher 模型，我们发现，由于 Hessian 具有可分的结构，天然可以写成若干个秩一矩阵的和。我们建立了一个简单的秩一近似，使得近似误差可以解释为某个矩阵的估计的方差，利用集中不等式，我们证明该近似误差对于玩家数量较多时非常高效。由于近似矩阵是秩一的，故可以利用 Sherman-Morrison 公式，直接写出 Hessian 的逆。

### 第一节 介绍

考虑一组可分割商品 (Divisible Goods)  $j \in \mathcal{J} = \{1, \dots, n\}$  和玩家  $i \in \mathcal{I} = \{1, \dots, m\}$  的交换市场。每个玩家  $i$  拥有初始现金 (Endowment)  $w_i \in \mathbb{R}_+$ ，并依据市场参与者设定的当前价格购买商品组合  $\mathbf{x}_i$  以最大化其效用函数  $u_i$ 。每种商品  $j \in \mathcal{J}$  的总量为 1。

引入一对互为对偶的线性空间  $(\mathbb{E}, \mathbb{E}^*)$ ，其中  $\mathbb{E}^*$  为  $\mathbb{E}$  的对偶空间，我们设定价格向量  $\mathbf{p} \in \mathcal{P} \subseteq \mathbb{E}$ ，其中  $\mathcal{P}$  是某个可行价格集。对于每个玩家  $\forall i \in \mathcal{I}$ ，分配向量  $\mathbf{x}_i \in \mathcal{X}_i(\mathbf{p}) \subseteq \mathbb{E}^*$  表示在价格  $\mathbf{p}$  下可行的商品组合。在 Fisher 模型中 (参见 [20])，我们取  $\mathcal{P} = \mathbb{R}_+^n$ ，且  $\mathcal{X}_i(\mathbf{p}) = \{x_i \in \mathbb{R}_+^n : \langle \mathbf{p}, \mathbf{x}_i \rangle \leq w_i\}$ 。不失一般性地，假定  $\sum_{i \in \mathcal{I}} w_i = 1$ 。市场均衡定义为这样的  $(\mathbf{x}_1, \dots, \mathbf{x}_m, \mathbf{p})$ ，其中每位玩家  $i$  采用效用最大化 (Utility Maximization, UM) 策略，并满足市场出清条件：

$$\begin{aligned} \text{存在 } \mathbf{p} \in \mathcal{P}, \text{ 使得 } \mathbf{x}_i \in \arg \max_{x_i \in \mathcal{X}_i} u_i(x_i), \\ \text{s.t. } \sum_{i \in \mathcal{I}} \mathbf{x}_i = \mathbf{1}. \end{aligned} \quad (7-1)$$

即使在中心化的设定下 (市场参与者有足够权力同时决定  $\{\mathbf{x}_i\}_{i \in \mathcal{I}}$  和  $\mathbf{p}$ ) 下，计算市场均衡仍然存在一定的挑战。当玩家的效用函数是齐次的时，一种计算均衡和价格的方法是借助 Eisenberg-Gale (EG) 凸优化问题 [36, 57]：

$$\max \sum_i w_i \log(u_i(\mathbf{x}_i)) \quad (7-2a)$$

$$\text{s.t. } \sum_{i \in \mathcal{I}} \mathbf{x}_i \leq \mathbf{1}, \quad (7-2b)$$

$$\mathbf{x}_i \in \mathcal{X}_i, \forall i \in \mathcal{I}. \quad (7-2c)$$

已知 (7-2b) 的对偶解对应于清算市场所需的价格  $\mathbf{p}$ 。该问题 (7-2) 可以利用几何规划 (geometric programming) 中的标准技术<sup>[76]</sup> 进行求解, 例如使用椭圆法<sup>[76,100]</sup>。然而, 即使是线性 Fisher 均衡的多项式时间算法也是近年才受到关注的<sup>[48]</sup>。Ye<sup>[170]</sup> 提出了适用于一类的市场问题 (包括 Arrow-Debreu 模型) 的内点法, 也可以用来求解 Fisher 模型。

### 7.1.1 研究动机

这些中心化的方法无法解释价格如何随玩家响应而演化。此外, 即使目的是计算均衡价格, 也不一定需要构造一个将分配和价格耦合起来的问题。另一种方法是设计 (分布式的) 拍卖算法。例如一个最简单的 tâtonnement 拍卖可以简单表述为如下流程:

- 令  $\mathbf{x}_i(\mathbf{p}) = \arg \max_{\mathbf{x}_i \in \mathcal{X}_i(\mathbf{p})} u_i(\mathbf{x}_i)$  为**最佳响应 (Best-Response BR) 映射** (或直接简称为**需求**), **间接效用函数 (Indirect Utility)** 由该点对应的最优目标值给出:

$$\nu_i(\mathbf{p}, w_i) = \max_{\mathbf{x}_i \in \mathcal{X}_i(\mathbf{p})} u_i(\mathbf{x}_i); \quad (7-3)$$

- 计算**市场超额需求函数**:

$$\mathbf{z}(\mathbf{p}) = \sum_{i \in \mathcal{I}} \mathbf{x}_i(\mathbf{p}) - \mathbf{1};$$

- 根据某个拍卖规则  $G$  更新  $\mathbf{p}$ :  $\mathbf{p}_+ \leftarrow \mathbf{p} + G \circ \mathbf{z}(\mathbf{p})$ .

对于 Fisher 模型, 取  $f_i(\mathbf{p}) = \log(\nu_i(\mathbf{p}, w_i)) - \log(w_i)$ , 则 Eisenberg-Gale 问题 (7-2) 的对偶问题可写为<sup>[68,113]</sup>:

$$\min_{\mathbf{p} \geq 0} f(\mathbf{p}) := \langle \mathbf{p}, \mathbf{1} \rangle + \sum_{i \in \mathcal{I}} w_i \log(w_i) + \sum_{i \in \mathcal{I}} w_i f_i(\mathbf{p}). \quad (7-4)$$

上述凸问题具有显式的可分结构, 可以很简单地想到利用一阶方法<sup>[16,34,63,69]</sup> 来构造拍卖算法。通过下面的表达式可以建立一阶算法与经典 tâtonnement 拍卖的联系:

$$w_i \nabla f_i(\mathbf{p}) = -\mathbf{x}_i(\mathbf{p}) \quad \text{且} \quad \nabla f(\mathbf{p}) = \mathbf{1} - \sum_{i \in \mathcal{I}} \mathbf{x}_i(\mathbf{p}). \quad (7-5)$$

对于某些连续可微的效用函数, 高阶导数 (如  $\nabla^p f_i, p \geq 2$ ) 也是可用的。由于大多数拍卖过程是一阶的, 我们探究二阶及更高阶 ( $p \geq 2$ ) 方法是否具有优势。具体来说, 我们研究以下问题:

“我们能否设计一个具有更好收敛性的二阶拍卖过程?”

针对这个问题, 我们考虑一种“非集中式”的内点方法。据我们所知, 除了早期针对一般均衡模型的研究<sup>[58]</sup>, 本章内容是第一个具有复杂性保证的尝试。

### 7.1.2 相关工作

众所周知, Fisher 模型是 Arrow-Debreu 模型<sup>[11,158]</sup>的特例, 其中货币预算  $w_i(\mathbf{p})$  依赖于价格。在 Arrow-Debreu 设置下, EG 问题并不可用, 而且通常无法构造一个凸的势函数<sup>注1</sup>。对于线性效用函数, 该问题等价于线性互补问题 (Linear Complementary Problem, LCP)<sup>[56,60]</sup>, 可通过 Lemke 方法实现有限步 (但不一定是多项式) 收敛。Curtis Eaves<sup>[45]</sup> 证明了 Cobb-Douglas 效用市场可在强多项式时间内求解。其他一般设定可借助一些可以求解广义 Nash 均衡问题的方法<sup>[59,95]</sup>。若效用函数是线性的, 另一种策略是使用<sup>[84]</sup>提出的凸可行系统 (最早见于<sup>[120]</sup>)。对于如何拓展到齐次准凹效用函数, 参见文献<sup>[85]</sup>。对于某些 CES 经济问题, 文献<sup>[37]</sup>提出了一种非线性变换, 使得问题转化为一个凸的可行性系统。但一般而言, 在 Arrow-Debreu 市场中, 可能存在多个不连通的均衡<sup>[35,67]</sup>。

## 第二节 方法概述

对于实数  $a \in \mathbb{R}$ , 我们定义  $[a]_+ = \max\{a, 0\}$  为  $a$  的非负部分。我们用  $\|\cdot\|, \|\cdot\|^*$  分别表示向量空间  $\mathbb{E}, \mathbb{E}^*$  上的范数及其对偶范数。算子范数表示为  $\|\mathbf{A}\| := \sup_{\|\mathbf{p}\| \leq 1} \|\mathbf{A}\mathbf{p}\|^*$ 。对于自伴算子,  $\lambda_1(\mathbf{A})$  代表其最小特征值,  $\lambda_n(\mathbf{A})$  代表最大特征值。我们用  $\ker(\mathbf{A})$  表示线性算子的零空间, 即  $\{\mathbf{p} \in \mathbb{E} | \mathbf{A}\mathbf{p} = 0\}$ 。我们用  $\mathcal{H}^d(\mathbb{E})$  表示所有  $d$  次齐次连续映射的集合, 若上下文清楚, 则省略  $\mathbb{E}$ 。即, 若  $f \in \mathcal{H}^d(\mathbb{E})$ , 则对于任意  $\mathbf{p} \in \mathbb{E}$ , 有  $f(\lambda\mathbf{p}) = \lambda^d f(\mathbf{p})$ 。特别地, 记  $\mathcal{H}^1(\mathbb{E}) \equiv \mathcal{H}(\mathbb{E})$ 。类似地, 我们用  $\mathcal{C}^p$  表示所有  $p$  阶可微的连续映射。 $D^p[\mathbf{h}_1, \dots, \mathbf{h}_p]$  代表沿方向  $(\mathbf{h}_1, \dots, \mathbf{h}_p)$  的  $p$  阶方向导数。对于梯度与 Hessian, 我们直接记作  $\nabla f$  和  $\nabla^2 f$ 。最后, 我们用大写字母表示对角矩阵, 其对角元为对应的小写字母。例如,  $\mathbf{P} = \text{diag}(\mathbf{p})$ ,  $\mathbf{P}^{-1} = \text{diag}(\mathbf{p}^{-1})$ 。

### 7.2.1 消费者理论的预备知识

我们考虑, 每个消费者  $i \in \mathcal{I}$  的效用函数属于常替代弹性 (Constant Elasticity of Substitution, CES) 效用族的情况:

$$u_i(\mathbf{x}_i) = \left[ \sum_j c_{ij} x_{ij}^{\rho_i} \right]^{\frac{1}{\rho_i}} = \langle \mathbf{c}_i, \mathbf{x}_i^{\rho_i} \rangle^{\frac{1}{\rho_i}}, \rho_i \in (-\infty, 1) \quad (7-6)$$

其中省略了线性效用和 Leontief 效用。令  $\sigma_i = \frac{\rho_i}{1-\rho_i} \in (-1, \infty)$ , 其关于  $\rho_i$  是递增的。通常,  $\delta_i = \frac{1}{1-\rho_i} = 1 + \sigma_i \in (0, \infty)$  被称为**替代弹性**。假设  $\sum_{i \in \mathcal{I}} w_i = 1$ , 且  $\mathbf{c}_i \in \mathbb{R}_+^n, \forall i \in \mathcal{I}$ 。由于  $\rho \in (-\infty, 1)$ , 已知  $\mathbf{x}_i(\mathbf{p})$  是单值的, 并且可微<sup>[113]</sup>。对 CES 市场, 有如下著名引理。

<sup>注1</sup>这本质上是因为超额需求函数  $\mathbf{z}(\mathbf{p})$  的 Jacobian 矩阵不对称, 故无法解释为梯度映射。

引理 7.1: CES 间接效用函数<sup>[113]</sup>

若  $u_i(\mathbf{x}_i) = \langle \mathbf{c}_i, \mathbf{x}_i^{\rho_i} \rangle^{\frac{1}{\rho_i}}$ ,  $\rho_i \in (-\infty, 1)$ , 则

$$\nu_i(\mathbf{p}, w_i) = w_i \left[ \langle \mathbf{c}_i^{1+\sigma_i}, \mathbf{p}^{-\sigma_i} \rangle \right]^{\frac{1}{\sigma_i}} \quad (7-7a)$$

$$\mathbf{x}_i(\mathbf{p}, w_i) = w_i \frac{[\mathbf{p}^{-1} \mathbf{c}_i]^{1+\sigma_i}}{\langle \mathbf{c}_i^{1+\sigma_i}, \mathbf{p}^{-\sigma_i} \rangle} \quad (7-7b)$$

对于 Fisher 市场,  $w_i$  为常数, 则对于任意  $\alpha \in \mathbb{R}_+$ ,

$$u_i(\alpha \mathbf{x}_i) = \alpha u_i(\mathbf{x}_i);$$

$$\mathbf{x}_i(\alpha \mathbf{p}, w_i) = w_i \frac{\alpha^{-(\sigma_i+1)} [\mathbf{p}^{-1} \mathbf{c}_i]^{1+\sigma_i}}{\alpha^{-\sigma_i} \langle \mathbf{c}_i^{1+\sigma_i}, \mathbf{p}^{-\sigma_i} \rangle} = \alpha^{-1} \mathbf{x}_i(\mathbf{p}, w_i); \quad (7-8)$$

即,  $u_i(\cdot) \in \mathcal{H}^1$  且  $\mathbf{x}_i(\cdot) \in \mathcal{H}^{-1}$ 。此外, 易验证

$$\nabla \log(\nu_i(\mathbf{p}, w_i)) = -\frac{\mathbf{x}_i(\mathbf{p})}{w_i}.$$

注意到

$$\begin{aligned} f(\mathbf{p}) &= \langle \mathbf{p}, \mathbf{1} \rangle + \sum_{i \in \mathcal{I}} w_i \log(\nu_i(\mathbf{p}, w_i)) \\ &= \langle \mathbf{p}, \mathbf{1} \rangle + \sum_{i \in \mathcal{I}} w_i \log w_i + \sum_{i \in \mathcal{I}} w_i \frac{1}{\sigma_i} \log \left( \langle \mathbf{c}_i^{1+\sigma_i}, \mathbf{p}^{-\sigma_i} \rangle \right) \\ &= \langle \mathbf{p}, \mathbf{1} \rangle + \sum_{i \in \mathcal{I}} w_i \log w_i + \sum_{i \in \mathcal{I}} w_i f_i(\mathbf{p}) \end{aligned} \quad (7-9)$$

其中取  $f_i(\mathbf{p}) = \frac{1}{\sigma_i} \log \left( \langle \mathbf{c}_i^{1+\sigma_i}, \mathbf{p}^{-\sigma_i} \rangle \right)$ , 显然,  $f(\mathbf{p})$  具有可分结构。

### 7.2.2 基于齐次模型的 Barrier 算法

为了设计“二阶”拍卖方法, 我们考虑一个稍微更一般的问题, 该问题允许仿射约束:

$$\begin{aligned} \min \quad & f(\mathbf{p}) \\ \text{s.t.} \quad & \mathbf{A}\mathbf{p} = \mathbf{a} \\ & \mathbf{p} \in \mathbb{R}_+^n \end{aligned} \quad (7-10)$$

换言之, 价格空间是第一象限与线性子空间的交集:

$$\mathcal{P} = \{ \mathbf{p} \in \mathbb{R}_+^n : \mathbf{A}\mathbf{p} = \mathbf{a} \}. \quad (7-11)$$

$\mathcal{P}$  的相对内点记作  $\mathcal{P}^\circ$ . 我们考虑对数 Barrier 模型, 取  $\mu > 0$ ,

$$\begin{aligned} \min_{\mathbf{p} \in \mathbb{R}_+^n} \quad & f_\mu(\mathbf{p}) := f(\mathbf{p}) - \mu h(\mathbf{p}) \\ \text{s.t.} \quad & \mathbf{A}\mathbf{p} = \mathbf{a}. \end{aligned} \quad (7-12)$$

其中，目标函数  $f: \mathbb{E} \mapsto \mathbb{R}$ ，类似于 (7-9)，是一个允许我们计算  $\nabla f$  和  $\nabla^2 f$  的凸函数。函数  $h(\mathbf{p}) = \sum_i^n \log(p_i)$  对于  $\mathbb{R}_+^n$  的标准对数 Barrier 函数。根据 Barrier 函数诱导的半范数定义如下： $\|\mathbf{h}\|_{\mathbf{p}} = \|\mathbf{P}^{-1}\mathbf{h}\|$ ，其对偶范数为  $\|\mathbf{h}\|_{\mathbf{p}}^* = \|\mathbf{P}\mathbf{h}\|$ 。由于  $\mathbf{A}$  结构较为简单，例如仅仅是一个单纯形，我们不妨假设存在初始点  $\mathbf{p}_0$  使得  $\mathbf{A}\mathbf{p}_0 = \mathbf{a}$ 。为了方便起见，我们将拉格朗日函数及其梯度记作：

$$\mathcal{L}(\mathbf{p}, \mathbf{y}) = f(\mathbf{p}) + \langle \mathbf{y}, \mathbf{A}\mathbf{p} - \mathbf{a} \rangle \quad \text{以及} \quad \ell(\mathbf{p}) := \nabla \mathcal{L}(\mathbf{p}, \mathbf{y}) = \nabla f(\mathbf{p}) + \mathbf{A}^T \mathbf{y}. \quad (7-13)$$

我们在 算法 7-1 中提出了一个齐次 Barrier 算法，其中下降方向是在  $\mathbf{A}$  的零空间中计算的。我们利用齐次模型进行迭代步的计算。

---

**算法 7-1: 齐次 Barrier 算法**


---

**Input:**  $k = T = 0, \mu_0 > 0, \mathbf{p} := \mathbf{p}_0, \eta < 1$

---

```

1 while  $\mu > \epsilon$  do
2     由 (7-22) 解出  $[\mathbf{v}, t]$ ;                                // 求解一个特征值问题
3     设  $\mathbf{d} = \frac{\mathbf{v}}{t}$ ;                                            // 将结果缩放回原始空间
4     按照 (7-42) 选择步长  $\alpha$ ;
5     令  $\mathbf{p}_+ = \mathbf{p} + \mathbf{h}, \mathbf{h} = \alpha \mathbf{P}\mathbf{d}$ ;
6     if  $\mathbf{p} \in \mathcal{Q}_{\mathbf{p}}(\eta, \mu)$  then
7         选择某个  $\sigma < 1$ ;
8         令  $\mu_+ = \sigma \mu$ ;
9     置  $\mathbf{p} \leftarrow \mathbf{p}_+, \mu \leftarrow \mu_+$ ;
```

---

该方法之所以称为“齐次”，是因为其升维后的  $(n+1)$  维变量  $[\mathbf{v}; t]$  是由一个齐次的二次规划规划求解的，该问题可以表述为一个广义特征值问题。求解后，我们将其归一化回原始空间 (行 3)。步长的选择保证下次更新仍然满足  $\mathbf{p}_+ \in \mathcal{P}^\circ$ 。类似于典型的内点法，我们定义中心路径的以下邻域，对于任意  $\eta \in (0, 1]$ ，

$$\mathcal{Q}_{\mathbf{p}}(\eta, \mu) = \left\{ \mathbf{p} \in \mathbb{R}_+^n : \frac{\|\mathbf{P}\ell(\mathbf{p}) - \mu \mathbf{1}\|}{\mu} \leq \frac{\eta}{Q_f} \right\},$$

$$Q_f = \frac{(\sqrt{n} + \beta_f)(1 + 2[\sigma]_{\max})C_f}{1 - C_f\beta_f} + \frac{\sqrt{n} + \beta_f}{\beta_f} + 2\beta_f + 1. \quad (7-14)$$

$C_f, \beta_f$  及其他参数的定义及其直观意义将在后续进行讨论。当价格迭代步进入该邻域时，我们按一定比例减少参数  $\mu$  (行 8)。我们证明该算法能够找到满足以下  $\epsilon$ -近似 KKT 条件的价格：

**定义 7.2: KKT 条件**

若一个元组  $(\mathbf{p}, \mathbf{y})$  满足以下条件，则称其为 (7-10) 的  $\epsilon$ -近似 KKT 点：

$$\mathbf{p} \in \mathbb{R}_+^n, \mathbf{A}\mathbf{p} = \mathbf{a}, \quad (7-15a)$$

$$\|\mathbf{P}(\nabla f(\mathbf{p}) + \mathbf{A}^T \mathbf{y})\|_\infty \leq O(\epsilon), \quad (7-15b)$$

$$\nabla f(\mathbf{p}) + \mathbf{A}^T \mathbf{y} \in \mathbb{R}_+^n. \quad (7-15c)$$

为简洁起见，我们称满足上述条件的点为  $\epsilon$ -KKT 点。在本文的其余部分，我们做如下假定。

### 假设 7.3

在 算法 7-1 中，我们保证迭代变量  $\mathbf{p}$  满足

$$f_i(\mathbf{p}) \leq \bar{f}_i \quad \text{且} \quad \|\mathbf{p}\|_\infty \leq D_{\mathcal{P}}. \quad (7-16)$$

注意，上述假定并不具有局限性。例如，在经典 Fisher 模型中，若采用 CES 效用函数，我们知道  $u_i(\cdot) \in \mathcal{H}^1$ 。由强对偶性，(7-2) 与 (7-9) 等价，我们可以写出 (7-2) 的最优解，其中  $\mathbf{p}$  是出清约束的对偶变量：

$$\begin{aligned} -w_i \frac{\nabla u_i(\mathbf{x}_i)}{u_i(\mathbf{x}_i)} + \mathbf{p} + \mathbf{s}_i &= 0, \quad \forall i \in \mathcal{I}, \\ \sum_{i \in \mathcal{I}} \mathbf{x}_i &= \mathbf{1}, \\ \mathbb{R}_+^n \ni \mathbf{x}_i \perp \mathbf{s}_i &\in \mathbb{R}_+^n, \quad \forall i \in \mathcal{I}. \end{aligned} \quad (7-17)$$

将第一组方程两侧乘以  $\mathbf{x}_i$  并对  $i \in \mathcal{I}$  求和，由  $u_i \in \mathcal{H}^1$  可得：

$$\langle \mathbf{p}, \sum_{i \in \mathcal{I}} \mathbf{x}_i \rangle = \sum_{i \in \mathcal{I}} \langle w_i \frac{1}{u_i(\mathbf{x}_i)} \nabla u_i(\mathbf{x}_i) - \mathbf{s}_i, \mathbf{x}_i \rangle = \sum_{i \in \mathcal{I}} w_i. \quad (7-18)$$

换言之， $\mathbf{p} \in \Delta_n$ ，且  $D_{\mathcal{P}} = 1$ 。

### 7.2.3 通过特征值问题更新迭代

记  $\mathbf{g}(\mathbf{p}) = \nabla f(\mathbf{p})$ ,  $\mathbf{H}(\mathbf{p}) = \nabla^2 f(\mathbf{p})$ 。在每次迭代中，考虑如下子问题：

$$\begin{aligned} \min_{\mathbf{h}} \quad & \frac{1}{2} \langle \mathbf{h}, (\mathbf{H} + \mu \mathbf{P}^{-2}) \mathbf{h} \rangle + \langle \mathbf{h}, \mathbf{g}(\mathbf{p}) - \mu \mathbf{P}^{-1} \mathbf{1} \rangle \\ \text{s.t.} \quad & \mathbf{h} \in \ker(\mathbf{A}). \end{aligned} \quad (7-19)$$

该问题适用于任何原始可行算法 (Primal Feasible Method<sup>[111]</sup>)，不难看出，算法 7-1 也属于该范畴。可以通过“仿射缩放 (Affine Scaling)”处理上述子问题，令  $\mathbf{d} = \mathbf{P}^{-1} \mathbf{h}$ ，则有

$$\begin{aligned} \min_{\mathbf{h} \in \ker(\mathbf{A})} \quad & \frac{1}{2} \langle \mathbf{h}, (\mathbf{H} + \mu \mathbf{P}^{-2}) \mathbf{h} \rangle + \langle \mathbf{h}, \mathbf{g}(\mathbf{p}) - \mu \mathbf{P}^{-1} \mathbf{1} \rangle \\ = \min_{\mathbf{d} \in \ker(\mathbf{AP})} \quad & \frac{1}{2} \langle \mathbf{d}, (\mathbf{PHP} + \mu \mathbf{I}) \mathbf{d} \rangle + \langle \mathbf{d}, \mathbf{Pg}(\mathbf{p}) - \mu \mathbf{1} \rangle. \end{aligned} \quad (7-20)$$

设  $\mathbf{A_p} = \mathbf{AP}$ ，我们将 (7-20) 提升 (lift) 至  $(n+1)$  维问题：

$$\begin{aligned} \min_{\|\mathbf{v}; t\| \leq 1} \quad & \psi(\mathbf{v}, t) = \frac{1}{2} \langle \mathbf{v}, \mathbf{PHPv} \rangle + \langle \mathbf{v}, \mathbf{Pg}(\mathbf{p}) - \mu \mathbf{1} \rangle \cdot t + \frac{1}{2} \mu \|\mathbf{v}\|^2 \\ \text{s.t.} \quad & \mathbf{A_p v} = 0. \end{aligned} \quad (7-21)$$

注意, (7-21) 现在是齐次的。零空间  $\ker(\mathbf{A}_p)$  的投影算子为  $\Pi_p = \mathbf{I} - \mathbf{A}_p^T(\mathbf{A}_p\mathbf{A}_p^T)^{-1}\mathbf{A}_p$ 。在  $\mathbf{p} \in \Delta_n$  的情况下, 易得  $\Pi_p = \mathbf{I} - \frac{\mathbf{p}\mathbf{p}^T}{\|\mathbf{p}\|^2}$ 。为简洁起见, 设  $\mathbf{g}_\mu = \Pi_p(\mathbf{P}\mathbf{g}(\mathbf{p}) - \mu\mathbf{1})$ 。注意到:

$$\Pi_p(\mathbf{P}\mathbf{H}\mathbf{P} + \mu\mathbf{I})\Pi_p = \Pi_p\mathbf{P}\mathbf{H}\mathbf{P}\Pi_p + \mu\Pi_p.$$

设  $\mathbf{Q}(\mathbf{p}) = \Pi_p\mathbf{P}\mathbf{H}\mathbf{P}\Pi_p$ 。遵循齐次模型的常规处理方式, 我们需计算以下特征值问题, 其中我们已按  $-\mu$  进行了缩放:

$$\min_{\|[\mathbf{v}; t]\| = 1} \frac{1}{2} \langle [\mathbf{v}; t], F(\mathbf{p}, \mu)[\mathbf{v}; t] \rangle, \quad F(\mathbf{p}, \mu) = \begin{bmatrix} \mathbf{Q}(\mathbf{p}) & \mathbf{g}_\mu(\mathbf{p}) \\ \mathbf{g}_\mu(\mathbf{p}) & -\mu \end{bmatrix}. \quad (7-22)$$

下文将建立这些问题的等价性 (仅相差某种缩放)。(本节的证明见 第 7.B 节)

#### 定理 7.4: 问题的等价性

以下问题在某种缩放意义下等价:

$$(7-21) \iff (7-22). \quad (7-23)$$

需要注意两点:

- 通过上述等价性, 我们提供了一种的有效计算子问题的方法: 首先求解广义特征值问题 (7-22), 那么原始-对偶解  $([\mathbf{v}; t], \theta)$  对应于  $F(\mathbf{p}, \mu)$  的最小特征向量和最小特征值。由于  $\Pi_p$  易于计算, 问题 (7-22) 可通过标准的特征值求解器 (如 ARPACK<sup>[106]</sup>) 高效求解。如果  $[\mathbf{v}; 0]$  是 (7-21) 的解, 则意味着  $\mathbf{v} \in \mathcal{S}_1(\mathbf{Q}(\mathbf{p}))$  且  $\mathbf{g}_\mu(\mathbf{p}) \perp \mathcal{S}_1(\mathbf{Q}(\mathbf{p}))$ <sup>[81,146]</sup>。在这种情况下, 我们可适当扰动  $\mathbf{Q}(\mathbf{p})$  处理 (类似于对信赖域子问题的做法<sup>[157]</sup>), 使得  $t \neq 0$  这种情况被消除。
- 其次, 如下所示, 我们得到了带有齐次仿射约束齐次二次问题的**全局最优性条件**。接下来, 我们在 推论 7.5 中引入一个等价的最优性条件, 该条件避免了投影算子  $\Pi_p$ 。

#### 推论 7.5: (7-21) 的最优性条件

存在  $\mathbf{d} \in \mathbb{R}^n, \theta > 0, \mathbf{y} \in \mathbb{R}^m$  使得 (7-21) 的最优性条件为  $\mathbf{d} = \frac{\mathbf{v}}{t}$ , 即:

$$(\mathbf{P}\mathbf{H}\mathbf{P} + \theta\mathbf{I})\mathbf{d} + (\mathbf{P}\mathbf{g} - \mu\mathbf{1}) + \mathbf{A}_p^T\mathbf{y} = 0, \quad (7-24a)$$

$$\theta - \mu + \langle \mathbf{d}, \mathbf{P}\mathbf{g} - \mu\mathbf{1} \rangle = 0, \quad (7-24b)$$

$$\mathbf{A}_p\mathbf{d} = 0, \quad (7-24c)$$

$$\theta - \mu > 0, \quad (7-24d)$$

$$(\mathbf{P}\mathbf{H}\mathbf{P} + \theta\mathbf{I}) - \frac{1}{\theta - \mu}(\mathbf{P}\mathbf{g}(\mathbf{p}) - \mu\mathbf{1})(\mathbf{P}\mathbf{g}(\mathbf{p}) - \mu\mathbf{1})^T \geq 0 \quad \text{on } \ker(\mathbf{A}_p). \quad (7-24e)$$

需要注意的是, 上述条件不仅仅是关于稳定点的标准二阶条件, 这些标准条件仅要求 Schur 补在线性约束  $\ker(\mathbf{A}_p)$  和球约束的切空间交集上为半正定<sup>[111,130]</sup>。相比之下, (7-24e) 更强, 因为它保



证了在整个  $\ker(\mathbf{A}_p)$  上的半正定性。严格来说, 该全局最优性条件是广义信赖域子问题的全局最优性条件的类比<sup>[29]</sup>, 而不仅仅是局部非全局的二阶条件<sup>[112,159]</sup>。由于  $\mathbf{y}$  仅用于分析, 我们无需显式计算  $\mathbf{y}$ 。基于上述最优性条件, 我们提出一个比较有用的引理。

**引理 7.6**

若  $(\mathbf{d}, \theta, \mathbf{y})$  是 (7-21) 的原始-对偶解, 则:

$$\|\mathbf{P}\mathbf{H}\mathbf{P}\mathbf{d}\| \leq \|\mathbf{P}\ell(\mathbf{p}) - \mu\mathbf{1}\|, \quad (7-25a)$$

$$\|\mathbf{d}\| \leq \frac{\|\mathbf{P}\ell(\mathbf{p}) - \mu\mathbf{1}\|}{\theta} \leq \frac{\|\mathbf{P}\ell(\mathbf{p}) - \mu\mathbf{1}\|}{\mu}, \quad (7-25b)$$

$$\mathbf{d}^T \mathbf{P}\mathbf{H}\mathbf{P}\mathbf{d} \leq \frac{\|\mathbf{P}\ell(\mathbf{p}) - \mu\mathbf{1}\|^2}{\mu}, \quad (7-25c)$$

$$0 \leq \theta - \mu \leq \frac{2\|\mathbf{P}\ell(\mathbf{p}) - \mu\mathbf{1}\|^2}{\mu}. \quad (7-25d)$$

### 第三节 对 CES 效用函数的新分析

对于 CES 效用函数, 回顾

$$\begin{aligned} f(\mathbf{p}) &= \langle \mathbf{p}, \mathbf{1} \rangle + \sum_{i \in \mathcal{I}} w_i \log(\nu_i(\mathbf{p}, w_i)) \\ &= \langle \mathbf{p}, \mathbf{1} \rangle + \sum_{i \in \mathcal{I}} w_i \left[ \log w_i + \underbrace{\frac{1}{\sigma_i} \log(\langle \mathbf{c}_i^{1+\sigma_i}, \mathbf{p}^{-\sigma_i} \rangle)}_{f_i(\mathbf{p})} \right]. \end{aligned} \quad (7-26)$$

同时, 我们记  $f_i(\mathbf{p}) = \frac{1}{\sigma_i} \log(r_i(\mathbf{p}))$ , 其中

$$r_i(\mathbf{p}) = \langle \mathbf{c}_i^{1+\sigma_i}, \mathbf{p}^{-\sigma_i} \rangle. \quad (7-27)$$

引入的函数  $r_i(\mathbf{p})$  具有可进一步探索的结构。取

$$\mathbf{v}_i = \mathbf{C}_i^{1+\sigma_i} \mathbf{p}^{-\sigma_i}, \quad \mathbf{V}_i = \mathbf{C}_i^{1+\sigma_i} \mathbf{P}^{-\sigma_i}, \quad (7-28)$$

则  $\mathbf{V}_i \geq \mathbf{0}$ , 其内积  $\langle \cdot, \cdot \rangle_{\mathbf{V}_i}$  及由  $\mathbf{V}_i$  诱导的半范数  $\|\cdot\|_{\mathbf{V}_i}$  均是良定义的。特别地,  $r_i$  可视为在该特殊度量下单位向量的长度:

$$r_i(\mathbf{p}) = \langle \mathbf{1}, \mathbf{v}_i \rangle = \langle \mathbf{1}, \mathbf{V}_i \mathbf{1} \rangle = \langle \mathbf{1}, \mathbf{1} \rangle_{\mathbf{V}_i} = \|\mathbf{1}\|_{\mathbf{V}_i}^2.$$

由此, 我们得出以下引理。

**引理 7.7: 严格但非强凸性**

对于所有  $\mathbf{p} \in \mathbb{R}_+^n$  和  $\mathbf{h} \in \mathbb{R}^n$ , 有

$$D^2 f_i(\mathbf{p})[\mathbf{h}, \mathbf{h}] \geq \min\{1, \sigma_i + 1\} \frac{\|\mathbf{P}^{-1} \mathbf{h}\|_{\mathbf{V}_i}^2}{\|\mathbf{1}\|_{\mathbf{V}_i}^2}. \quad (7-29)$$



另一种表述 (7-29) 的方式是 Hessian 矩阵关于一个“仿射缩放”的秩一矩阵是正定的：

$$\begin{aligned}\nabla^2 f_i(\mathbf{p}) &\geq \min\{1, \sigma_i + 1\} \left( \frac{\mathbf{P}^{-1}\mathbf{v}_i}{\langle \mathbf{1}, \mathbf{v}_i \rangle} \right) \left( \frac{\mathbf{P}^{-1}\mathbf{v}_i}{\langle \mathbf{1}, \mathbf{v}_i \rangle} \right)^T \\ &= \frac{\min\{1, \sigma_i + 1\}}{\exp(\sigma_i f_i(\mathbf{p}))^2} \left( \left( \frac{\mathbf{c}_i}{\mathbf{p}} \right)^{\sigma_i+1} \right) \left( \left( \frac{\mathbf{c}_i}{\mathbf{p}} \right)^{\sigma_i+1} \right)^T.\end{aligned}\quad (7-30)$$

注意  $\frac{\mathbf{c}_i}{\mathbf{p}} = \mathbf{P}^{-1}\mathbf{c}_i$ , 这在内点法中可以看成是利用对数 barrier 函数“仿射缩放”后的向量。这表明, 每个  $f_i(\mathbf{p})$  (因此对偶函数) 都是严格凸的, 但**非强凸**。同时, 它也不是 Lipschitz 光滑的, 因为  $\lim_{\mathbf{p} \rightarrow 0_+} \|\nabla^2 f_i(\mathbf{p})\| = +\infty$ 。幸运的是, 经过适当缩放后, 我们可以证明 Hessian 矩阵的算子范数是有界的。

#### 引理 7.8: 有界的算子范数

对于所有  $\mathbf{p} \in \mathbb{R}_+^n$ , 有

$$\begin{aligned}\|\mathbf{P}\nabla^2 f_i(\mathbf{p})\mathbf{P}\| &\leq 2[\sigma_i]_+ + 1, \\ \|\mathbf{P}\nabla^2 f(\mathbf{p})\mathbf{P}\| &\leq 2[\sigma]_{\max} + 1.\end{aligned}\quad (7-31)$$

其中  $[\sigma]_{\max} = \max_{i \in \mathcal{I}} [\sigma_i]_+$ 。

严格来说, 对偶函数满足自协性 (参见 <sup>[127]</sup> Definition 2.1.1)。

#### 引理 7.9: 自协性

对于所有  $\mathbf{h} \in \mathbb{R}^n$  和  $\mathbf{p} \in \mathbb{R}_+^n$ , 有

$$|D^3 f_i(\mathbf{p})[\mathbf{h}, \mathbf{h}, \mathbf{h}]| \leq 2C_i(\sigma_i) [D^2 f_i(\mathbf{p})[\mathbf{h}, \mathbf{h}]]^{\frac{3}{2}}. \quad (7-32)$$

其中

$$C_i(\sigma_i) = \begin{cases} \frac{(\sigma_i+1)(\sigma_i+2) \exp(\sigma_i \bar{f}_i) + 5\sigma_i^2 + 3\sigma_i}{2}, & \text{若 } \sigma_i > 0, \\ \frac{(\sigma_i+1)(\sigma_i+2) D_{\mathcal{P}} \|\mathbf{c}_i^{1+\sigma_i}\|_1 + 5\sigma_i^2 - 3\sigma_i}{2}, & \text{o.w.} \end{cases} \quad (7-33)$$

因此, 我们有如下推论。

#### 推论 7.10

对偶函数  $f(\mathbf{p})$  具有自协性质, 且常数为

$$C_f = \max_{i \in \mathcal{I}} \left\{ \frac{1}{\sqrt{w_i}} C_i(\sigma_i) \right\}. \quad (7-34)$$

我们称  $C_f$  为自协系数。该结论直接来自协函数对求和的稳定性; 参见 <sup>[127]</sup> 的 Proposition 2.1.1 和 <sup>[125]</sup> 的 Theorem 5.1.1。上述结果表明, 最困难的情况出现在  $\rho_i$  接近 1 时。我们给出以下条件作为分析的直接推论。

**引理 7.11: 局部稳定性条件**

若函数  $f_i(\mathbf{p})$  是  $C_i$ -自协函数, 则对于任意  $\|\mathbf{h}\|_{\mathbf{p}} \leq \beta_f < \frac{1}{C_f}$ , 有

$$\|\nabla f(\mathbf{p} + \mathbf{h}) - \nabla f(\mathbf{p}) - \nabla^2 f(\mathbf{p})\mathbf{h}\|_{\mathbf{p}}^* \leq (1 + 2[\sigma]_{\max}) \frac{C_f \|\mathbf{h}\|_{\mathbf{p}}^2}{1 - C_f \|\mathbf{h}\|_{\mathbf{p}}}, \quad (7-35)$$

其中  $[\sigma]_{\max}$  的定义见 引理 7.8。

此外, 可以观察到自协系数满足

$$C_f \approx \Theta(\max_{i \in \mathcal{I}} \{\sigma_i^2\}) \quad \text{以及} \quad \|\mathbf{P} \nabla^2 f(\mathbf{p}) \mathbf{P}\| \leq \Theta([\sigma]_{\max} + 1).$$

对于处于互补范围（即在该效用函数下, 商品是互补的）内的  $\rho_i$  (即  $\rho_i \leq 0$ ), 这意味着 Hessian 不会随着  $\rho_i \searrow -\infty$  而增长。一种策略是对  $f_i(\mathbf{p})$  进行适当缩放, 例如乘以  $\max_{i \in \mathcal{I}} \{\sigma_i^2\}$  的某个倍数, 使得  $C_f$  接近 1. 这将增大局部稳定性的有效半径 (7-35), 并允许更大的步长。为保持完整性, 我们未进行此操作, 但根据<sup>[127]</sup> 定理 2.1.1, 这种操作是可行的。

**7.3.1 对角线加秩一 (Diagonal plus rank-one, DR1) 近似**

在本节中, 我们介绍一种对偶函数 Hessian 矩阵的简单而有效的近似, 我们称之为 **对角线加秩 1 (DR1) 近似**. 考虑

$$\mathbf{P} \nabla^2 f(\mathbf{p}) \mathbf{P} = \sum_{i \in \mathcal{I}} w_i [(\sigma_i + 1) \mathbf{U}_i - \sigma_i \mathbf{u}_i \mathbf{u}_i^T] = \underbrace{\sum_{i \in \mathcal{I}} w_i (\sigma_i + 1) \mathbf{U}_i}_{\mathbf{D}} - \underbrace{\sum_{i \in \mathcal{I}} \sigma_i w_i \mathbf{u}_i \mathbf{u}_i^T}_{(\sum_{i \in \mathcal{I}} \sigma_i w_i) \Xi} \quad (7-36)$$

其中我们取

$$\Xi := \frac{\sum_{i \in \mathcal{I}} \sigma_i w_i \mathbf{u}_i \mathbf{u}_i^T}{\sum_{i \in \mathcal{I}} \sigma_i w_i} = \sum_{i \in \mathcal{I}} \gamma_i \mathbf{u}_i \mathbf{u}_i^T, \quad \sum_{i \in \mathcal{I}} \gamma_i = 1. \quad (7-37)$$

对  $\mathbf{H}(\mathbf{p}) = \nabla^2 f(\mathbf{p})$  的一个自然近似为  $\tilde{\mathbf{H}}(\mathbf{p}) = \mathbf{D} + (\sum_{i \in \mathcal{I}} \sigma_i w_i) \tilde{\Xi}$ , 其中取  $\mathbf{D} = \sum_{i \in \mathcal{I}} w_i (\sigma_i + 1) \mathbf{U}_i$ , 并通过下面的秩 1 矩阵用  $\tilde{\Xi}$  来近似  $\Xi$ :

$$\tilde{\Xi} := \tilde{\mathbf{u}} \tilde{\mathbf{u}}^T, \quad \tilde{\mathbf{u}} := \sum_{i \in \mathcal{I}} \gamma_i \mathbf{u}_i. \quad (7-38)$$

我们有以下定理.

**定理 7.12: DR1 近似的误差**

若对于所有  $i \in \mathcal{I}$ ,  $\mathbf{c}_i \sim \tau$  且相互独立, 如果玩家数量满足

$$|\mathcal{I}| \geq \frac{2 \left(4 + \frac{2\epsilon}{3}\right) \ln \left(\frac{n}{\delta}\right)}{\epsilon^2}.$$

则至少以下不等式以至少为  $1 - \delta$  的概率成立:

$$\|\mathbf{P}(\nabla^2 f(\mathbf{p}) - (\mathbf{D} + (\sum_{i \in \mathcal{I}} \sigma_i w_i) \tilde{\Xi})) \mathbf{P}\| \leq \epsilon. \quad (7-39)$$

以上定理说明, 只要市场中玩家数量大于  $\Omega(\epsilon^{-2})$ , 则 DR1 近似与真实 Hessian 矩阵的误差小于  $\epsilon$  的概率至少为  $1 - \delta$ . 使用 DR1 近似的好处是, 利用 Sherman-Morrison 公式, 我们可以直接写出所需矩阵的逆,

$$\begin{aligned} [\nabla^2 f(\mathbf{p})]^{-1} &\approx [\tilde{\mathbf{H}}]^{-1} = \left[ \mathbf{D} + \left( \sum_{i \in \mathcal{J}} \sigma_i w_i \right) \bar{\mathbf{u}} \bar{\mathbf{u}}^T \right]^{-1} \\ &= \mathbf{D}^{-1} - \frac{(\sum_{i \in \mathcal{J}} \sigma_i w_i) \mathbf{D}^{-1} \bar{\mathbf{u}} \bar{\mathbf{u}}^T \mathbf{D}^{-1}}{1 + (\sum_{i \in \mathcal{J}} \sigma_i w_i) \langle \bar{\mathbf{u}}, \mathbf{D}^{-1} \bar{\mathbf{u}} \rangle}. \end{aligned} \quad (7-40)$$

所以即使我们考虑的是二阶算法, 我们也不需要求解线性方程组。

我们随机生成一些向量  $\{\mathbf{c}_i\}_{i \in \mathcal{J}}$ , 并不断扩大  $\mathcal{J}$ , 下图展示了估计  $\mathbf{P} \nabla^2 f(\mathbf{p}) \mathbf{P} - \|\mathbf{P} \tilde{\mathbf{H}} \mathbf{P}\|$  的误差随玩家数目  $|\mathcal{J}|$  的变化。我们可以看到, 正如理论所预测的, 随着玩家数目的增加, 估计的误差迅速

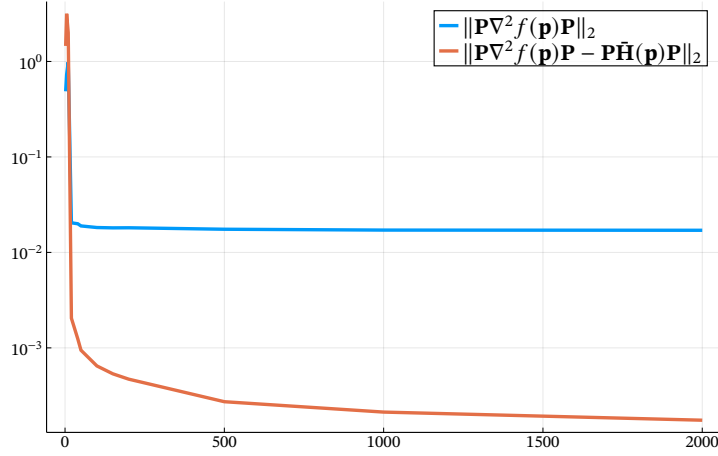


图 7.3.1 估计  $\mathbf{P} \nabla^2 f(\mathbf{p}) \mathbf{P}$  的误差随玩家数目  $|\mathcal{J}|$  的变化

减小。对于二阶算法而言, 取得误差小于  $10^{-2}$  左右的误差已经足够小。在高概率下  $(1 - \delta)$ , 只要迭代的不动点误差  $\|\mathbf{h}\|$  足够大, 我们不会停止迭代, 此时我们可得

$$\|[\nabla^2 f(\mathbf{p}) - \tilde{\mathbf{H}}(\mathbf{p})] \mathbf{h}\| \leq \epsilon \|\mathbf{h}\| \implies \|[\nabla^2 f(\mathbf{p}) - \tilde{\mathbf{H}}(\mathbf{p})] \mathbf{h}\| \leq O(\|\mathbf{h}\|^2). \quad (7-41)$$

这个条件强于拟牛顿法的 Dennis-Moré 条件, 与不精确二阶算法建立全局复杂度的条件类似<sup>[28,162]</sup>.

## 第四节 价格更新算法的收敛性分析

为了保持迭代点  $\mathbf{p} \in \mathcal{D}^\circ$  并维持自协性质的有效性，我们需要以下步长选择规则。

**引理 7.13**

设步长按以下规则选取：

$$\alpha = \min \left\{ 1, \frac{\beta_f \mu}{\|\mathbf{P}\ell(\mathbf{p}) - \mu \mathbf{1}\|} \right\}, \quad (7-42)$$

则有

$$\alpha \|\mathbf{d}\| \leq \beta_f, \quad (7-43a)$$

$$1 - \alpha \leq \frac{1}{\beta_f} \frac{\|\mathbf{P}\ell(\mathbf{p}) - \mu \mathbf{1}\|}{\mu}. \quad (7-43b)$$

*Proof.* 我们分两种情况讨论：

(1) 若  $\frac{\|\mathbf{P}\ell(\mathbf{p}) - \mu \mathbf{1}\|}{\mu} \leq \beta_f$ ，则  $\frac{\beta_f \mu}{\|\mathbf{P}\ell(\mathbf{p}) - \mu \mathbf{1}\|} \geq 1$ ，从而  $\alpha = 1$ ：

$$\|\alpha \mathbf{d}\| = \alpha \|\mathbf{d}\| \leq \alpha \frac{\|\mathbf{P}\ell(\mathbf{p}) - \mu \mathbf{1}\|}{\mu} \leq \beta_f.$$

(2) 否则， $\frac{\|\mathbf{P}\ell(\mathbf{p}) - \mu \mathbf{1}\|}{\mu} > \beta_f$ ，即  $\frac{\beta_f \mu}{\|\mathbf{P}\ell(\mathbf{p}) - \mu \mathbf{1}\|} \leq 1$ ，于是：

$$\|\alpha \mathbf{d}\| \leq \frac{\beta_f \mu}{\|\mathbf{P}\ell(\mathbf{p}) - \mu \mathbf{1}\|} \frac{\|\mathbf{P}\ell(\mathbf{p}) - \mu \mathbf{1}\|}{\mu} \leq \beta_f.$$

进一步有：

$$1 - \alpha \leq 1 \leq \frac{\|\mathbf{P}\ell(\mathbf{p}) - \mu \mathbf{1}\|}{\beta_f \mu}.$$

证毕。  $\square$

在继续讨论之前，我们回顾以下定义（参见 (7-14)）：

$$\begin{aligned} \mathcal{Q}_{\mathbf{p}}(\eta, \mu) &= \left\{ \mathbf{p} \in \mathbb{R}_+^n : \frac{\|\mathbf{P}\ell(\mathbf{p}) - \mu \mathbf{1}\|}{\mu} \leq \frac{\eta}{Q_f} \right\}, \\ Q_f &= \frac{(\sqrt{n} + \beta_f)(1 + 2[\sigma]_{\max})C_f}{1 - C_f\beta_f} + \frac{\sqrt{n}}{\beta_f} + 2\beta_f + 2. \end{aligned} \quad (7-44)$$

证明关键是，在固定  $\mu$  的情况下，迭代点将以二次速度收敛到中心路径。以下结果表明二次收敛区域大致为  $\mathcal{Q}_{\mathbf{p}}(1, \mu)$ 。

**定理 7.14: 子问题的二次收敛性**

对于任意迭代点  $\mathbf{p}$  和固定  $\mu > 0$ , 有:

$$\frac{\|\mathbf{P}_+\ell(\mathbf{p}_+) - \mu\mathbf{1}\|}{\mu} \leq Q_f \left( \frac{\|\mathbf{P}\ell(\mathbf{p}) - \mu\mathbf{1}\|}{\mu} \right)^2. \quad (7-45)$$

*Proof.* 设  $\mathbf{p}$  为可行点, 则

$$\mathbf{P}_+ = \mathbf{P} + \alpha\mathbf{P}\mathbf{D} = (\mathbf{I} + \alpha\mathbf{D})\mathbf{P}. \quad (7-46)$$

于是

$$\begin{aligned} \mathbf{P}_+\ell(\mathbf{p}_+) - \mu\mathbf{1} &= \mathbf{P}_+(\ell_+ - \mu\mathbf{1}) = \mathbf{P}_+(\mathbf{g}(\mathbf{p}_+) + \mathbf{A}^T\mathbf{y}) - \mu\mathbf{1}, \\ &= \mathbf{P}_+(\mathbf{g}(\mathbf{p}_+) - \mathbf{g}(\mathbf{p}) - \alpha\mathbf{H}(\mathbf{p})\mathbf{P}\mathbf{d}) + \mathbf{P}_+(\mathbf{g} + \alpha\mathbf{H}\mathbf{P}\mathbf{d} + \mathbf{A}^T\mathbf{y} - \mu\mathbf{P}^{-1}\mathbf{1}) + \mu\mathbf{P}_+\mathbf{P}^{-1}\mathbf{1} - \mu\mathbf{1}, \\ &= \mathbf{P}_+(\mathbf{g}(\mathbf{p}_+) - \mathbf{g}(\mathbf{p}) - \alpha\mathbf{H}(\mathbf{p})\mathbf{P}\mathbf{d}) + \alpha\mathbf{P}_+(\mathbf{g} + \mathbf{H}\mathbf{P}\mathbf{d} + \mathbf{A}^T\mathbf{y} - \mu\mathbf{P}^{-1}\mathbf{1}) \\ &\quad + (1 - \alpha)\mathbf{P}_+(\mathbf{g} + \mathbf{A}^T\mathbf{y} - \mu\mathbf{P}^{-1}\mathbf{1}) - \mu\mathbf{1} + \mu(1 + \alpha\mathbf{d}), \\ &= (\mathbf{I} + \alpha\mathbf{D})\mathbf{P}(\mathbf{g}(\mathbf{p}_+) - \mathbf{g}(\mathbf{p}) - \alpha\mathbf{H}(\mathbf{p})\mathbf{P}\mathbf{d}) + \alpha(\mathbf{I} + \alpha\mathbf{D})(\mathbf{P}\mathbf{g} + \mathbf{P}\mathbf{H}\mathbf{P}\mathbf{d} + \mathbf{A}_p^T\mathbf{y} - \mu\mathbf{1}) \\ &\quad + (1 - \alpha)(\mathbf{I} + \alpha\mathbf{D})(\mathbf{P}\mathbf{g} + \mathbf{A}_p^T\mathbf{y} - \mu\mathbf{1}) + \alpha\mu\mathbf{d}. \end{aligned} \quad (7-47)$$

由 (7-24), 可得

$$\begin{aligned} \mathbf{P}_+\ell(\mathbf{p}_+) - \mu\mathbf{1} &= (\mathbf{I} + \alpha\mathbf{D})\mathbf{P}(\mathbf{g}(\mathbf{p}_+) - \mathbf{g}(\mathbf{p}) - \alpha\mathbf{H}(\mathbf{p})\mathbf{P}\mathbf{d}) + \alpha(\mathbf{I} + \alpha\mathbf{D})(-\theta\mathbf{d}) \\ &\quad + (1 - \alpha)(\mathbf{I} + \alpha\mathbf{D})(\mathbf{P}\ell(\mathbf{p}) - \mu\mathbf{1}) + \alpha\mu\mathbf{d}, \\ &= (\mathbf{I} + \alpha\mathbf{D})\mathbf{P}(\mathbf{g}(\mathbf{p}_+) - \mathbf{g}(\mathbf{p}) - \alpha\mathbf{H}(\mathbf{p})\mathbf{P}\mathbf{d}) \\ &\quad + (1 - \alpha)(\mathbf{I} + \alpha\mathbf{D})(\mathbf{P}\ell(\mathbf{p}) - \mu\mathbf{1}) + (\alpha(\mu - \theta)\mathbf{d} - \alpha^2\theta\mathbf{d}^2). \end{aligned} \quad (7-48)$$

由 (7-43), 得  $\|(\mathbf{I} + \alpha\mathbf{D})\| \leq \sqrt{n} + \beta_f$ , 进一步有:

$$\begin{aligned} \|(\mathbf{I} + \alpha\mathbf{D})\mathbf{P}(\mathbf{g}(\mathbf{p}_+) - \mathbf{g}(\mathbf{p}) - \alpha\mathbf{H}(\mathbf{p})\mathbf{P}\mathbf{d})\| &\stackrel{(7-35)}{\leq} (\sqrt{n} + \beta_f)(1 + 2[\sigma]_{\max}) \frac{\alpha^2 C_f \|\mathbf{d}\|^2}{1 - \alpha C_f \|\mathbf{d}\|}, \\ &\stackrel{(7-25b), (7-43)}{\leq} (\sqrt{n} + \beta_f)(1 + 2[\sigma]_{\max}) \frac{\alpha^2 C_f \|\mathbf{d}\|^2}{1 - C_f \beta_f}, \\ &\stackrel{(7-25c)}{\leq} \alpha^2 \frac{(\sqrt{n} + \beta_f)(1 + 2[\sigma]_{\max}) C_f \|\mathbf{P}\ell(\mathbf{p}) - \mu\mathbf{1}\|^2}{1 - C_f \beta_f \mu}. \end{aligned} \quad (7-49)$$

并且

$$\begin{aligned} \|(1 - \alpha)(\mathbf{I} + \alpha\mathbf{D})(\mathbf{P}\ell(\mathbf{p}) - \mu\mathbf{1})\| &\leq (1 - \alpha)(\sqrt{n} + \beta_f) \|\mathbf{P}\ell(\mathbf{p}) - \mu\mathbf{1}\| \\ &\stackrel{(7-43b)}{\leq} \frac{(\sqrt{n} + \beta_f)}{\beta_f} \frac{\|\mathbf{P}\ell(\mathbf{p}) - \mu\mathbf{1}\|^2}{\mu}. \end{aligned} \quad (7-50)$$

以及

$$\begin{aligned} \|\alpha(\mu - \theta)\mathbf{d} - \alpha^2\theta\mathbf{d}^2\| &\stackrel{(7-25), (7-43)}{\leq} 2\beta_f \frac{\|\mathbf{P}\ell(\mathbf{p}) - \mu\mathbf{1}\|^2}{\mu} + \alpha^2 \frac{\|\mathbf{P}\ell(\mathbf{p}) - \mu\mathbf{1}\|^2}{\mu}, \\ &\leq (\alpha^2 + 2\beta_f) \frac{\|\mathbf{P}\ell(\mathbf{p}) - \mu\mathbf{1}\|^2}{\mu}. \end{aligned} \quad (7-51)$$

综上所述,

$$\frac{\|\mathbf{P}_+\ell(\mathbf{p}_+) - \mu\mathbf{1}\|}{\mu} \leq \left( \frac{(\sqrt{n} + \beta_f)(1 + 2[\sigma]_{\max})C_f}{1 - C_f\beta_f} + \frac{\sqrt{n} + \beta_f}{\beta_f} + 2\beta_f + 1 \right) \left( \frac{\|\mathbf{P}\ell(\mathbf{p}) - \mu\mathbf{1}\|}{\mu} \right)^2. \quad (7-52)$$

证毕。  $\square$

注意, 在 [算法 7-1](#) 的 [行 8](#) 处, 我们将  $\mu$  乘以  $\sigma < 1$  进行更新。类似于标准短步策略, 只要  $\sigma$  选取得当, 迭代点将始终保持在有效区域  $\mathbb{Q}_{\mathbf{p}}$  内, 如 [\(7-14\)](#) 所示。

#### 引理 7.15: 短步策略

当减少  $\mu$  时, 若我们选择  $\sigma$  使得

$$\mu_+ = \sigma\mu, \quad \frac{\sqrt{n} + \eta Q_f^{-1}}{\sqrt{n} + Q_f^{-1}} \leq \sigma < 1, \quad (7-53)$$

则当  $\frac{\|\mathbf{P}\ell - \mu\mathbf{1}\|}{\mu} \leq \frac{\eta}{Q_f}$  时, 有  $\frac{1}{\mu_+} \|\mathbf{P}\ell - \mu_+\mathbf{1}\| \leq \frac{1}{Q_f}$ 。

*Proof.*

$$\begin{aligned} \frac{1}{\mu_+} \|\mathbf{P}\ell - \mu_+\mathbf{1}\| &= \frac{1}{\sigma\mu} \|\mathbf{P}\ell - \mu\mathbf{1} + (1 - \sigma)\mu\mathbf{1}\| \\ &\leq \frac{1}{\sigma\mu} \|\mathbf{P}\ell - \mu\mathbf{1}\| + \frac{1}{\sigma\mu} (1 - \sigma) \|\mu\mathbf{1}\| \\ &\leq \frac{1}{\sigma} \frac{\eta}{Q_f} + \frac{\sqrt{n}(1 - \sigma)}{\sigma}. \end{aligned}$$

令右侧小于等于  $\frac{1}{Q_f}$ , 即得结论。  $\square$

在迭代开始时, 我们简单选择  $\mu_0$  使得我们在下述定理中总结算法的整体迭代复杂度。

#### 定理 7.16: [算法 7-1](#) 的复杂度

若在 [算法 7-1](#) 中  $\sigma$  依照 [\(7-53\)](#) 选取, 且  $1 - \eta = \eta_0 Q_f$ , 则算法在求解  $O(\sqrt{n} \log(\frac{1}{\epsilon}) \log \log(\frac{Q_f}{\eta}))$  个特征值问题后, 能找到满足 [\(7-15\)](#) 的  $\epsilon$ -KKT 点。

*Proof.* 由 [\(7-53\)](#), 可得

$$\mu_+ = \mu \cdot \left( 1 - \frac{(1-\eta)Q_f^{-1}}{\sqrt{n} + Q_f^{-1}} \right) = \mu \cdot \left( 1 - \frac{(1-\eta)}{Q_f\sqrt{n} + 1} \right) \leq \mu \cdot \left( 1 - \frac{1-\eta}{Q_f\sqrt{n}} \right). \quad (7-54)$$

设  $\mu_K$  为经过  $K$  次衰减后的 barrier 参数, 则

$$\mu_K \leq \mu_0 \left(1 - \frac{1-\eta}{Q_f \sqrt{n}}\right)^K = \left(1 - \frac{\eta_0}{\sqrt{n}}\right)^K. \quad (7-55)$$

令  $\mu_K \leq \epsilon$ , 选取合适的  $K$ , 有

$$K \geq \frac{\sqrt{n}}{\eta_0} \log\left(\frac{1}{\epsilon}\right) \geq \frac{\log\left(\frac{1}{\epsilon}\right)}{-\log\left(1 - \frac{\eta_0}{\sqrt{n}}\right)}. \quad (7-56)$$

在每个  $\mu_k, k \leq K$  迭代过程中,  $\mathbf{p} \in \mathbb{Q}_{\mathbf{p}}(\eta, \mu_k)$ ; 对于相同的  $\mathbf{p}$ , 有  $\mathbf{p} \in \mathbb{Q}_{\mathbf{p}}(1, \mu_{k+1})$ , 即  $\frac{Q_f \|\mathbf{P}\ell(\mathbf{p}) - \mu_{k+1} \mathbf{1}\|}{\mu_{k+1}} \leq 1$ . 特别地, 从  $\mu_{k+1}$  开始, 经过  $\log \log\left(\frac{Q_f}{\eta}\right)$  个子问题后, 我们再次得到  $\mathbf{p} \in \mathbb{Q}_{\mathbf{p}}(\eta, \mu_{k+1})$ . 由定义, 在  $\mu_K$  处, 有

$$\begin{aligned} \|\mathbf{P}\ell(\mathbf{p}) - \mu_K \mathbf{1}\| &\leq \frac{\eta}{Q_f} \mu_K \leq \frac{\eta}{Q_f} \epsilon, \\ \implies \|\mathbf{P}(\nabla f(\mathbf{p}) + \mathbf{A}^T \mathbf{y})\|_{\infty} &\leq \left(\frac{\eta}{Q_f} + 1\right) \epsilon. \end{aligned} \quad (7-57)$$

这表明 (7-15b) 成立。由于迭代点始终保持在可行集内部, (7-15a) 亦成立。证毕。  $\square$

## 第五节 数值实验

我们对算法 7-1 进行一些数值实验。我们在 Mac OS 桌面端实现了我们的算法, 设备配备 14-核 Apple M4 Pro 处理器和 48 GB LPDDR5 内存。大多数子程序均可在标准的 Julia 包中找到。

在求解 GHM 时, 我们在 Lanczos 方法中设置收敛容差为  $\min\{10^{-5}, 10^{-2} \|g_k\|\}$ 。图 7.5.1 展示了不同弹性系数下, 算法 7-1 的收敛速度。

我们进行一些简单的实验。对于货物数  $n$ , 玩家数目  $m$ , 我们随机生成系数  $\mathbf{c}_1, \dots, \mathbf{c}_m$ , 其中  $c_{ij}, \forall i \in \mathcal{I}, j \in \mathcal{J}$  服从 0-1 均匀分布。对每一个玩家  $i$ , 我们设定  $c_i$  的稀疏度为 0.3, 同时保证至少有一个货物是被需要的。算法 7-1 展示了  $m = 200, n = 100$  时, 算法的收敛情况: 一般情况下, 只需要十步以内的迭代, 一个例外是接近于线性效用函数时, 问题的难度较大。据我们所知, 是否可以基于拍卖规则设计线性收敛的算法仍然是一个开放问题。

为了进一步说明算法的效率, 我们与具有收敛性保证的一阶 tâtonnement 算法进行比较, 算法可以看作某种镜像梯度法 (Mirror Descent Method), 具体实现参照文献<sup>[33]</sup>。另外一种算法需要将问题重新表述为 Shmyrev 均衡问题<sup>[151]</sup>, 具体实现参照文献<sup>[34]</sup>。计算结果可见图 7.5.2 和 7.5.3。可以看到, 不论是在迭代数还是时间上, 我们的算法都具有显著优势。

由于本章的分析, 我们首次给出了一个可以分布式运行的二阶算法, 其迭代代价类似于一个一阶算法, 我们对如下的大问题进行简单的验证。我们令  $\rho = 0.6$ , 按相同规则生成  $\mathbf{c}_1, \dots, \mathbf{c}_m$ , 我们设置 100.0 秒的时间限制, 停机准则设置成梯度容差  $10^{-7}$ , 同时, 我们可以将一个完整问题建模成标准的线性锥规划<sup>注 2</sup>。从表 7.5.1 可以看到, 在问题规模较大时, 我们的算法具有显著优势。

<sup>注 2</sup>即线性目标, 线性约束, 变量满足凸锥。<sup>[127]</sup>

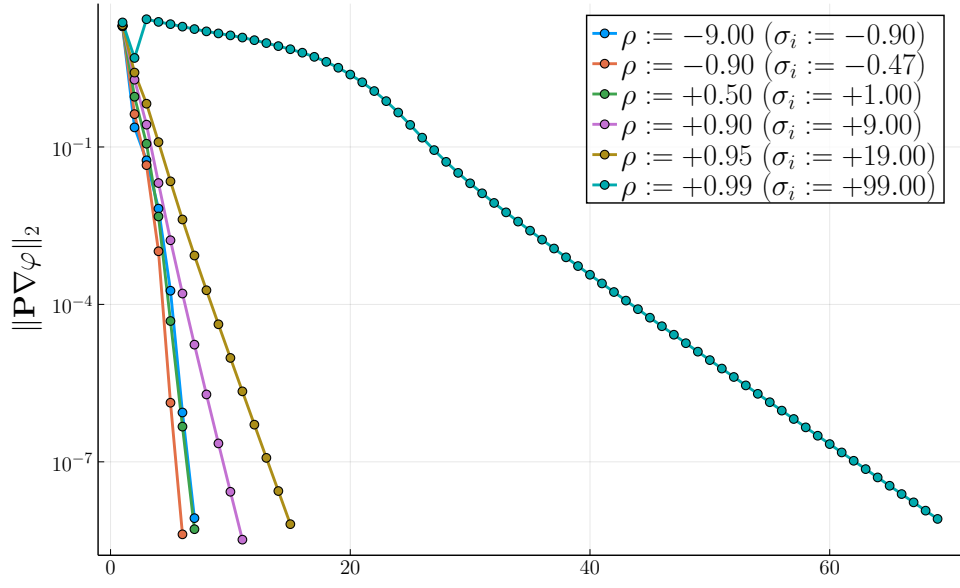


图 7.5.1 算法 7-1 在不同弹性系数下的比较，其中

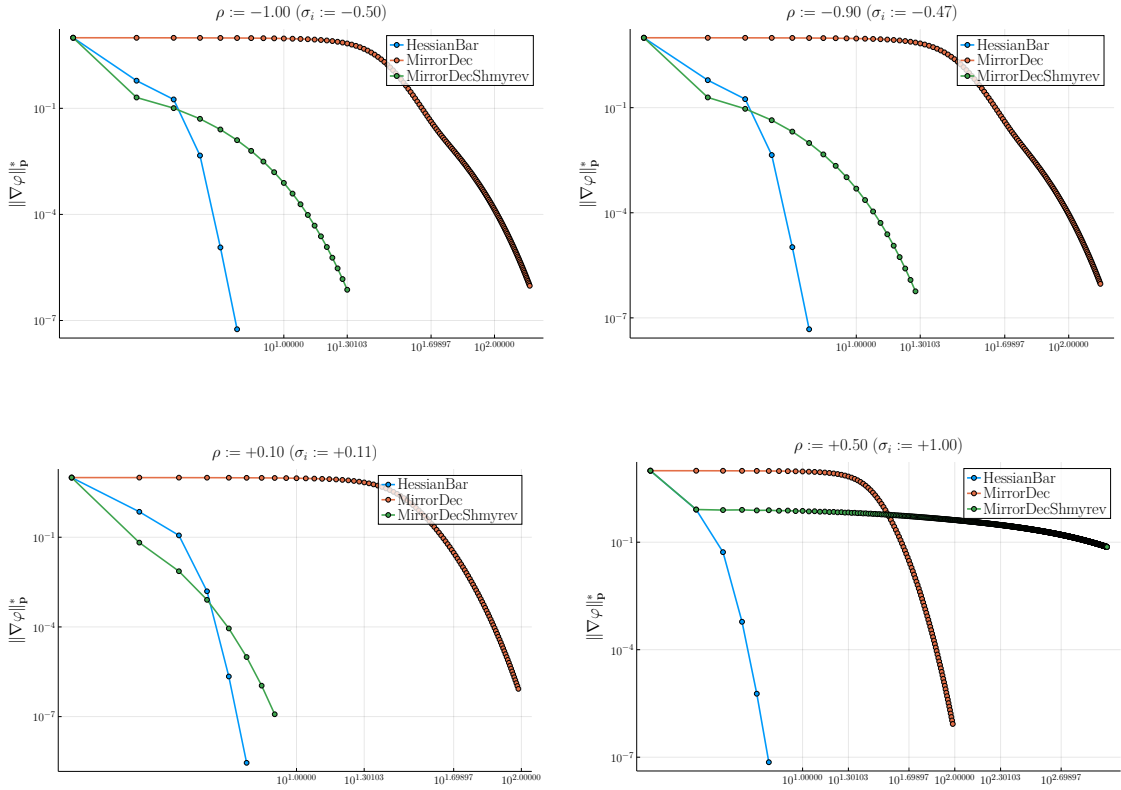


图 7.5.2 不同算法在  $m = 200, n = 100$  时的收敛情况 (迭代数)



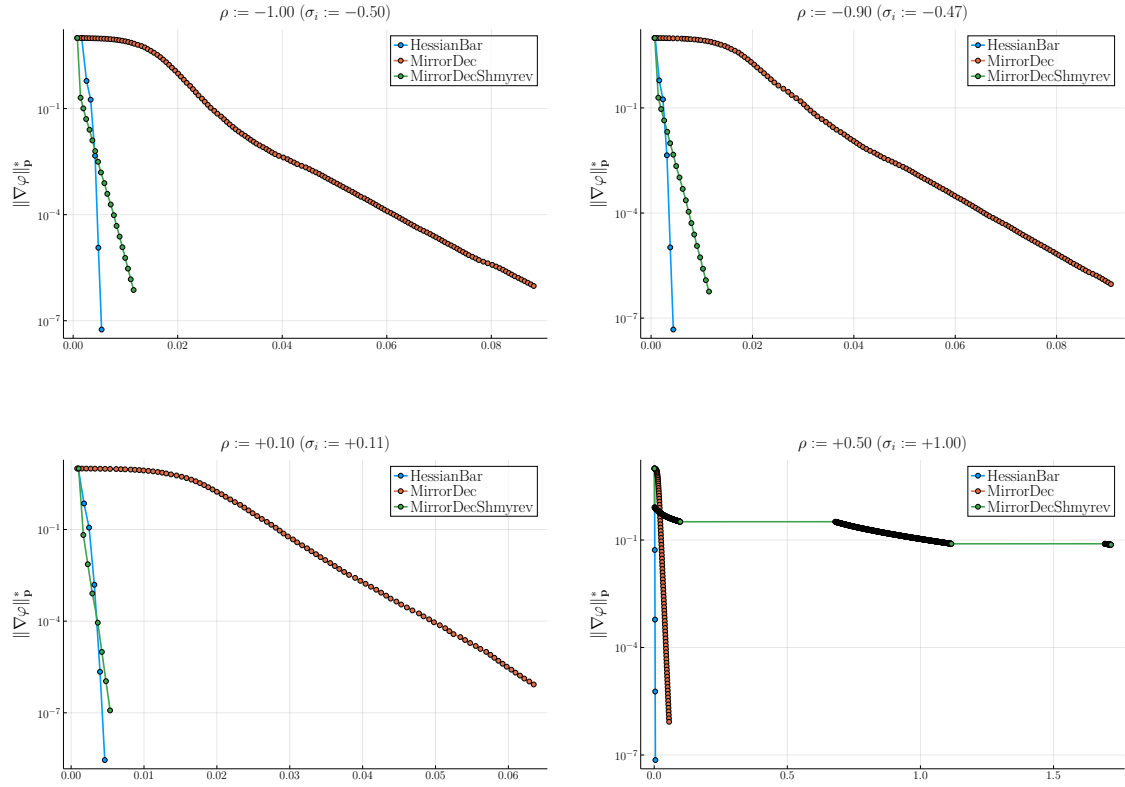


图 7.5.3 不同算法在  $m = 200, n = 100$  时的收敛情况 (时间)

表 7.5.1 不同问题规模下, 算法 7-1 的运行时间。\* 表示在 100.0 秒内未收敛到容差, 括号内表示停机时的梯度范数大小。

$n$	$m$	Time (算法 7-1)	Time (Tâtonnement <sup>[33]</sup> )	Time (Mosek)
100	1,000	0.11	0.20	2.97
100	10,000	2.93	2.34	57.00
100	100,000	8.91	21.51	-
1,000	10,000	3.02	16.16	-
2,000	10,000	4.95	34.91	-
2,000	100,000	67.27	100.0* ( $2.4 \times 10^{-4}$ )	-
5,000	10,000	11.97	-	-
5,000	50,000	100.66	-	-

## 本章附录

## 第一节 CES 效用函数下自协性质的证明

若取  $\mathbf{u}_i = \frac{\mathbf{v}_i}{r_i(\mathbf{p})}$ , 则满足  $\langle \mathbf{u}_i, \mathbf{1} \rangle = 1, \mathbf{u}_i \geq 0$ , 从而

$$\|\mathbf{U}_i\| = \|\mathbf{u}_i\|_\infty \leq 1, \|\mathbf{u}_i \mathbf{u}_i^T\| \leq 1. \quad (7-58)$$

根据 CES 效用函数的构造, 计算  $r_i(\mathbf{p})$  的导数:

$$\nabla r_i(\mathbf{p}) = -\sigma_i \mathbf{C}_i^{1+\sigma_i} \mathbf{P}^{-\sigma_i-1} \mathbf{1} = -\sigma_i \mathbf{P}^{-1} \mathbf{v}_i, \quad (7-59a)$$

$$\nabla^2 r_i(\mathbf{p}) = \sigma_i(\sigma_i + 1) \mathbf{C}_i^{1+\sigma_i} \mathbf{P}^{-\sigma_i-2} = \sigma_i(\sigma_i + 1) \mathbf{P}^{-1} \mathbf{v}_i \mathbf{P}^{-1}, \quad (7-59b)$$

$$D^3 r_i(\mathbf{p}) = -\sigma_i(\sigma_i + 1)(\sigma_i + 2) \left( \mathbf{C}_i^{1+\sigma_i} \mathbf{P}^{-\sigma_i-3} \right). \quad (7-59c)$$

因此,

$$\nabla f_i(\mathbf{p}) = \frac{1}{\sigma_i r_i(\mathbf{p})} D r_i(\mathbf{p}), \quad (7-60a)$$

$$\nabla^2 f_i(\mathbf{p}) = \frac{1}{\sigma_i r_i(\mathbf{p})} D^2 r_i(\mathbf{p}) - \frac{1}{\sigma_i (r_i(\mathbf{p}))^2} D r_i(\mathbf{p}) D r_i(\mathbf{p})^T \quad (7-60b)$$

$$= \frac{1}{\sigma_i r_i(\mathbf{p})^2} [r_i(\mathbf{p}) D^2 r_i(\mathbf{p}) - D r_i(\mathbf{p}) D r_i(\mathbf{p})^T]. \quad (7-60c)$$

等价地,

$$\nabla^2 f_i(\mathbf{p}) = \frac{1}{r_i(\mathbf{p})} (\sigma_i + 1) \mathbf{P}^{-1} \mathbf{v}_i \mathbf{P}^{-1} - \frac{\sigma_i \mathbf{P}^{-1} \mathbf{v}_i \mathbf{v}_i^T \mathbf{P}^{-1}}{(r_i(\mathbf{p}))^2}. \quad (7-61)$$

三阶导数为

$$\begin{aligned} D^3 f_i(\mathbf{p}) &= \frac{d}{d\mathbf{p}} \left[ \frac{1}{\sigma_i r_i(\mathbf{p})} D^2 r_i(\mathbf{p}) - \frac{1}{\sigma_i (r_i(\mathbf{p}))^2} D r_i(\mathbf{p}) D r_i(\mathbf{p})^T \right] \\ &= \frac{D^3 r_i(\mathbf{p})}{\sigma_i r_i(\mathbf{p})} - \frac{3 D^2 r_i(\mathbf{p}) \otimes D r_i(\mathbf{p})}{\sigma_i r_i(\mathbf{p})^2} + \frac{2 D r_i(\mathbf{p}) \otimes D r_i(\mathbf{p}) \otimes D r_i(\mathbf{p})}{\sigma_i r_i(\mathbf{p})^3}. \end{aligned} \quad (7-62)$$

对于所有  $\mathbf{h} \in \mathbb{R}^n$ , 设  $\tilde{\mathbf{h}} = \mathbf{P}^{-1} \mathbf{h}$ , 则

$$D r_i(\mathbf{p})[\mathbf{h}] = -\sigma_i \langle \mathbf{v}_i, \mathbf{P}^{-1} \mathbf{h} \rangle = -\sigma_i \langle \mathbf{v}_i, \tilde{\mathbf{h}} \rangle, \quad (7-63a)$$

$$D^2 r_i(\mathbf{p})[\mathbf{h}, \mathbf{h}] = \sigma_i(\sigma_i + 1) \langle \mathbf{v}_i, \tilde{\mathbf{h}}^2 \rangle, \quad (7-63b)$$

$$D^3 r_i(\mathbf{p})[\mathbf{h}, \mathbf{h}, \mathbf{h}] = -\sigma_i(\sigma_i + 1)(\sigma_i + 2) \langle \mathbf{v}_i, \tilde{\mathbf{h}}^3 \rangle. \quad (7-63c)$$

于是

$$D f_i(\mathbf{p})[\mathbf{h}] = -\frac{\langle \mathbf{v}_i, \tilde{\mathbf{h}} \rangle}{r_i(\mathbf{p})}. \quad (7-64)$$

二阶导数满足

$$D^2 f_i(\mathbf{p})[\mathbf{h}, \mathbf{h}] = \frac{\sigma_i + 1}{\|\mathbf{1}\|_{\mathbf{v}_i}^2} \|\tilde{\mathbf{h}}\|_{\mathbf{v}_i}^2 - \frac{\sigma_i}{\|\mathbf{1}\|_{\mathbf{v}_i}^4} \langle \mathbf{1}, \tilde{\mathbf{h}} \rangle_{\mathbf{v}_i}^2. \quad (7-65)$$

三阶导数为

$$D^3 f_i(\mathbf{p})[\mathbf{h}, \mathbf{h}, \mathbf{h}] = -\frac{(\sigma_i + 1)(\sigma_i + 2)\langle \mathbf{1}, \tilde{\mathbf{h}}^3 \rangle_{\mathbf{v}_i}}{r_i(\mathbf{p})} + \frac{3\sigma_i(\sigma_i + 1)\|\tilde{\mathbf{h}}\|_{\mathbf{v}_i}^2 \langle \mathbf{1}, \tilde{\mathbf{h}} \rangle_{\mathbf{v}_i}}{r_i(\mathbf{p})^2} + \frac{2\sigma_i^2 \langle \mathbf{1}, \tilde{\mathbf{h}} \rangle_{\mathbf{v}_i}^3}{r_i(\mathbf{p})^3}. \quad (7-66a)$$

### 7.A.1 证明 引理 7.7

*Proof.* 由于  $\|\mathbf{1}\|_{\mathbf{v}_i} = \langle \mathbf{c}_i^{1+\sigma_i}, \mathbf{p}^{-\sigma_i} \rangle$ , 我们分两种情况分析:

(1) 若  $\sigma_i = \frac{\rho_i}{1-\rho_i} > 0$ , 则

$$\begin{aligned} D^2 f_i(\mathbf{p})[\mathbf{h}, \mathbf{h}] &= \frac{\sigma_i + 1}{\|\mathbf{1}\|_{\mathbf{v}_i}^2} \|\tilde{\mathbf{h}}\|_{\mathbf{v}_i}^2 - \frac{\sigma_i}{\|\mathbf{1}\|_{\mathbf{v}_i}^4} \langle \mathbf{1}, \tilde{\mathbf{h}} \rangle_{\mathbf{v}_i}^2 \\ &\geq \frac{\sigma_i + 1}{\|\mathbf{1}\|_{\mathbf{v}_i}^4} \|\tilde{\mathbf{h}}\|_{\mathbf{v}_i}^2 \|\mathbf{1}\|_{\mathbf{v}_i}^2 - \frac{\sigma_i}{\|\mathbf{1}\|_{\mathbf{v}_i}^4} \|\tilde{\mathbf{h}}\|_{\mathbf{v}_i}^2 \|\mathbf{1}\|_{\mathbf{v}_i}^2 \geq \frac{\|\tilde{\mathbf{h}}\|_{\mathbf{v}_i}^2}{\|\mathbf{1}\|_{\mathbf{v}_i}^2}. \end{aligned} \quad (7-67)$$

(2) 若  $\sigma_i \in (-1, 0]$ , 则

$$D^2 f_i(\mathbf{p})[\mathbf{h}, \mathbf{h}] = \frac{\sigma_i + 1}{\|\mathbf{1}\|_{\mathbf{v}_i}^2} \|\tilde{\mathbf{h}}\|_{\mathbf{v}_i}^2 - \frac{\sigma_i}{\|\mathbf{1}\|_{\mathbf{v}_i}^4} \langle \mathbf{1}, \tilde{\mathbf{h}} \rangle_{\mathbf{v}_i}^2 \geq (\sigma_i + 1) \frac{\|\tilde{\mathbf{h}}\|_{\mathbf{v}_i}^2}{\|\mathbf{1}\|_{\mathbf{v}_i}^2}. \quad (7-68)$$

证毕。  $\square$

### 7.A.2 证明 引理 7.8

回顾 Hessian  $\nabla^2 f_i(\mathbf{p})$  的表达式 (7-61):

$$\nabla^2 f_i(\mathbf{p}) = \frac{1}{r_i(\mathbf{p})} (\sigma_i + 1) \mathbf{P}^{-1} \mathbf{V}_i \mathbf{P}^{-1} - \frac{\sigma_i \mathbf{P}^{-1} \mathbf{v}_i \mathbf{v}_i^T \mathbf{P}^{-1}}{(r_i(\mathbf{p}))^2}. \quad (7-69)$$

利用 (7-58), 即  $\mathbf{u}_i = \frac{\mathbf{v}_i}{r_i(\mathbf{p})}$ , 我们有

$$\mathbf{P} \nabla^2 f_i(\mathbf{p}) \mathbf{P} = (\sigma_i + 1) \mathbf{U}_i - \sigma_i \mathbf{u}_i \mathbf{u}_i^T. \quad (7-70)$$

观察可得:

$$\|\mathbf{P} \nabla^2 f_i(\mathbf{p}) \mathbf{P}\| \leq (\sigma_i + 1) \|\mathbf{U}_i\| + \sigma_i \|\mathbf{u}_i \mathbf{u}_i^T\| \leq 2[\sigma_i]_+ + 1. \quad (7-71)$$

类似地, 特征值问题 (7-22) 中的矩阵满足:

$$\begin{aligned} \|\Pi_{\mathbf{p}} \mathbf{P} \mathbf{H}(\mathbf{p}) \mathbf{P} \Pi_{\mathbf{p}}\| &\leq \|\mathbf{P} (\sum_{i \in \mathcal{F}} w_i \nabla^2 f_i(\mathbf{p})) \mathbf{P}\| \\ &\leq \sum_{i \in \mathcal{F}} w_i (2[\sigma_i]_+ + 1) \leq 2[\sigma]_{\max} + 1. \end{aligned} \quad (7-72)$$

证毕。

### 7.A.3 证明 引理 7.9

我们首先引入以下引理。

**引理 7.17**

若 [假设 7.3](#) 成立, 则

$$r_i(\mathbf{p}) := \|\mathbf{1}\|_{\mathbf{v}_i}^2 \leq \begin{cases} \exp(\sigma_i \bar{f}_i) & \text{if } \sigma_i > 0, \\ D_{\mathcal{F}} \|\mathbf{c}_i^{1+\sigma_i}\|_1 & \text{if } \sigma_i \leq 0. \end{cases} \quad (7-73)$$

进一步, 若  $\sigma_i > 0$ , 则

$$\mathbf{v}_i \geq \mathbf{c}_i^{1+\sigma_i} D_{\mathcal{F}}^{-\sigma_i}. \quad (7-74)$$

*Proof.* 若  $\sigma_i > 0$ , 则

$$\begin{aligned} r_i(\mathbf{p}) &= \langle \mathbf{c}_i^{1+\sigma_i}, \mathbf{p}^{-\sigma_i} \rangle = \exp(\sigma_i f_i(\mathbf{p})) \leq \exp(\sigma_i \bar{f}_i), \\ \mathbf{v}_i &= \mathbf{C}_i^{1+\sigma_i} \mathbf{p}^{-\sigma_i} \geq \mathbf{c}_i^{1+\sigma_i} D_{\mathcal{F}}^{-\sigma_i}. \end{aligned} \quad (7-75)$$

否则,

$$r_i(\mathbf{p}) = \langle \mathbf{c}_i^{1+\sigma_i}, \mathbf{p}^{-\sigma_i} \rangle \leq D_{\mathcal{F}} \langle \mathbf{1}, \mathbf{c}_i^{1+\sigma_i} \rangle. \quad (7-76)$$

证毕。  $\square$

现在我们证明 [引理 7.9](#)。

*Proof.* 根据定义,

$$\begin{aligned} D^3 f_i(\mathbf{p})[\mathbf{h}, \mathbf{h}, \mathbf{h}] &= -\frac{(\sigma_i + 1)(\sigma_i + 2) \langle \mathbf{1}, \tilde{\mathbf{h}}^3 \rangle_{\mathbf{v}_i}}{r_i(\mathbf{p})} + \frac{3\sigma_i(\sigma_i + 1) \|\tilde{\mathbf{h}}\|_{\mathbf{v}_i}^2 \langle \mathbf{1}, \tilde{\mathbf{h}} \rangle_{\mathbf{v}_i}}{r_i(\mathbf{p})^2} + \frac{2\sigma_i^2 \langle \mathbf{1}, \tilde{\mathbf{h}} \rangle_{\mathbf{v}_i}^3}{r_i(\mathbf{p})^3}. \end{aligned} \quad (7-77)$$

由此可得

$$\begin{aligned} |D^3 f_i(\mathbf{p})[\mathbf{h}, \mathbf{h}, \mathbf{h}]| &\leq \frac{(\sigma_i + 1)(\sigma_i + 2) \|\mathbf{1}\|_{\mathbf{v}_i} \|\tilde{\mathbf{h}}\|_{\mathbf{v}_i}^3}{\sigma_i \|\mathbf{1}\|_{\mathbf{v}_i}^2} + \frac{3|\sigma_i|(\sigma_i + 1) \|\mathbf{1}\|_{\mathbf{v}_i} \|\tilde{\mathbf{h}}\|_{\mathbf{v}_i}^3}{\|\mathbf{1}\|_{\mathbf{v}_i}^4} + \frac{2\sigma_i^2 \|\mathbf{1}\|_{\mathbf{v}_i}^3 \|\tilde{\mathbf{h}}\|_{\mathbf{v}_i}^3}{\|\mathbf{1}\|_{\mathbf{v}_i}^6} \\ &= \frac{(\sigma_i + 1)(\sigma_i + 2) \|\tilde{\mathbf{h}}\|_{\mathbf{v}_i}^3}{\|\mathbf{1}\|_{\mathbf{v}_i}} + \frac{3|\sigma_i|(\sigma_i + 1) \|\tilde{\mathbf{h}}\|_{\mathbf{v}_i}^3}{\|\mathbf{1}\|_{\mathbf{v}_i}^3} + \frac{2\sigma_i^2 \|\tilde{\mathbf{h}}\|_{\mathbf{v}_i}^3}{\|\mathbf{1}\|_{\mathbf{v}_i}^3}. \end{aligned} \quad (7-78)$$

由于  $\|\mathbf{1}\|_{\mathbf{v}_i} = \langle \mathbf{c}_i^{1+\sigma_i}, \mathbf{p}^{-\sigma_i} \rangle$ , 我们按照 [引理 7.7](#) 分两种情况分析:

(1) 若  $\sigma_i = \frac{\rho_i}{1-\rho_i} > 0$ , 则

$$|D^3 f_i(\mathbf{p})[\mathbf{h}, \mathbf{h}, \mathbf{h}]| \quad (7-79a)$$

$$\leq \left( (\sigma_i + 1)(\sigma_i + 2) \exp(\sigma_i \bar{f}_i) + 5\sigma_i^2 + 3\sigma_i \right) \left[ D^2 f_i(\mathbf{p})[\mathbf{h}, \mathbf{h}] \right]^{\frac{3}{2}}. \quad (7-79b)$$

(2) 若  $\sigma_i \in (-1, 0]$ , 则

$$|D^3 f_i(\mathbf{p})[\mathbf{h}, \mathbf{h}, \mathbf{h}]| \leq \left( (\sigma_i + 1)(\sigma_i + 2) D_{\mathcal{P}} \langle \mathbf{1}, \mathbf{c}_i^{1+\sigma_i} \rangle + 5\sigma_i^2 - 3\sigma_i \right) [D^2 f_i(\mathbf{p})[\mathbf{h}, \mathbf{h}]]^{\frac{3}{2}}. \quad (7-80a)$$

综上所述, 我们得到

$$|D^3 f_i(\mathbf{p})[\mathbf{h}, \mathbf{h}, \mathbf{h}]| \leq 2C_i(\sigma_i) [D^2 f_i(\mathbf{p})[\mathbf{h}, \mathbf{h}]]^{\frac{3}{2}}. \quad (7-81a)$$

其中

$$C_i(\sigma_i) = \begin{cases} \frac{(\sigma_i+1)(\sigma_i+2) \exp(\sigma_i \bar{f}_i) + 5\sigma_i^2 + 3\sigma_i}{2}, & \text{if } \sigma_i > 0, \\ \frac{(\sigma_i+1)(\sigma_i+2) D_{\mathcal{P}} \|\mathbf{c}_i^{1+\sigma_i}\|_1 + 5\sigma_i^2 - 3\sigma_i}{2}, & \text{otherwise.} \end{cases} \quad (7-82)$$

□

#### 7.A.4 证明 引理 7.11

*Proof.* 若  $f(\mathbf{p})$  是  $C_f$ -自协的, 则由<sup>[125]</sup> Theorem 5.1.7 可得:

$$\begin{aligned} (1 - tC_f r)^2 \nabla^2 f(\mathbf{p}) &\leq \nabla^2 f(\mathbf{p} + t\mathbf{h}) \leq \frac{1}{(1 - tC_f r)^2} \nabla^2 f(\mathbf{p}), \quad \text{对所有 } \|\mathbf{h}\|_{\mathbf{p}} = r \leq \frac{1}{C_f}, \\ \implies ((1 - tC_f r)^2 - 1) \nabla^2 f(\mathbf{p}) &\leq \underbrace{\nabla^2 f(\mathbf{p} + t\mathbf{h}) - \nabla^2 f(\mathbf{p})}_{\mathbf{G}(t)} \leq \left( \frac{1}{(1 - tC_f r)^2} - 1 \right) \nabla^2 f(\mathbf{p}). \end{aligned} \quad (7-83)$$

其中左侧的不等式是一个负定矩阵, 因此有:

$$\begin{aligned} \|\mathbf{G}(t)\|_{\mathbf{p}}^* &\leq \max \left\{ 1 - (1 - tC_f r)^2, \frac{1}{(1 - tC_f r)^2} - 1 \right\} \|\nabla^2 f(\mathbf{p})\|_{\mathbf{p}}^*, \\ &\leq \left( \frac{1}{(1 - tC_f r)^2} - 1 \right) \|\nabla^2 f(\mathbf{p})\|_{\mathbf{p}}^*, \end{aligned} \quad (7-84)$$

由于  $1 - (1 - tC_f r)^2 \leq \frac{1}{(1 - tC_f r)^2} - 1$ . 对于 CES 经济, 结合 (7-61),

$$\begin{aligned} \nabla^2 f_i(\mathbf{p}) &= \frac{(\sigma_i + 1) \mathbf{P}^{-1} \mathbf{V}_i \mathbf{P}^{-1}}{r_i(\mathbf{p})} - \frac{\sigma_i \mathbf{P}^{-1} \mathbf{v}_i \mathbf{v}_i^T \mathbf{P}^{-1}}{(r_i(\mathbf{p}))^2}, \\ \nabla^2 f(\mathbf{p}) &= \sum_{i \in \mathcal{I}} w_i \nabla^2 f_i(\mathbf{p}). \end{aligned} \quad (7-85)$$

于是

$$\begin{aligned} \|\nabla^2 f(\mathbf{p})\|_{\mathbf{p}}^* &= \sup_{\|\xi\|_{\mathbf{p}}=1} \|\mathbf{P} \nabla^2 f(\mathbf{p}) \xi\| = \sup_{\|\xi\|_{\mathbf{p}}=1} \left\| \sum_{i \in \mathcal{I}} w_i \left( \frac{(\sigma_i + 1) \mathbf{V}_i \mathbf{P}^{-1} \xi}{r_i(\mathbf{p})} - \frac{\sigma_i \mathbf{v}_i \mathbf{v}_i^T \mathbf{P}^{-1} \xi}{(r_i(\mathbf{p}))^2} \right) \right\|, \\ &\leq \sup_{\|\xi\|_{\mathbf{p}}=1} \left\| \sum_{i \in \mathcal{I}} w_i \left( \frac{(\sigma_i + 1) \mathbf{V}_i}{r_i(\mathbf{p})} - \frac{\sigma_i \mathbf{v}_i \mathbf{v}_i^T}{(r_i(\mathbf{p}))^2} \right) \right\| \|\mathbf{P}^{-1} \xi\|. \end{aligned} \quad (7-86)$$

利用 (7-58) 及  $\|\mathbf{P}^{-1}\xi\| = \|\xi\|_{\mathbf{p}} = 1$ , 可得

$$\begin{aligned}\|\nabla^2 f(\mathbf{p})\|_{\mathbf{p}}^* &\leq \sum_{i \in \mathcal{F}} w_i \|(\sigma_i + 1)\mathbf{U}_i - \sigma_i \mathbf{u}_i \mathbf{u}_i^T\|, \\ &\leq \sum_{i \in \mathcal{F}} w_i \left( (\sigma_i + 1)\|\mathbf{U}_i\| + \|\sigma_i \mathbf{u}_i \mathbf{u}_i^T\| \right), \\ &\leq \sum_{i \in \mathcal{F}} w_i (1 + 2[\sigma]_+) \leq 1 + 2[\sigma]_{\max}.\end{aligned}\quad (7-87)$$

由于  $\int_0^1 \left( \frac{1}{(1-tC_f r)^2} - 1 \right) dt = \frac{C_f r}{1-C_f r}$ , 我们有

$$\begin{aligned}\|\nabla f(\mathbf{p} + \mathbf{h}) - \nabla f(\mathbf{p}) - \nabla^2 f(\mathbf{p})\mathbf{h}\|_{\mathbf{p}}^* &= \left\| \int_0^1 (\nabla^2 f(\mathbf{p} + t \cdot \mathbf{h}) - \nabla^2 f(\mathbf{p}))\mathbf{h} dt \right\|_{\mathbf{p}}^* \leq \int_0^1 \|\mathbf{G}(t)\|_{\mathbf{p}}^* \|\mathbf{h}\|_{\mathbf{p}} dt, \\ &\leq \int_0^1 (1 + 2[\sigma]_{\max}) \left( \frac{1}{(1-tC_f r)^2} - 1 \right) r dt, \\ &\leq (1 + 2[\sigma]_{\max}) \frac{C_f r^2}{1 - C_f r}.\end{aligned}\quad (7-88)$$

于是可得

$$\|\nabla f(\mathbf{p} + \mathbf{h}) - \nabla f(\mathbf{p}) - \nabla^2 f(\mathbf{p})\mathbf{h}\|_{\mathbf{p}}^* \leq (1 + 2[\sigma]_{\max}) \frac{C_f \|\mathbf{h}\|_{\mathbf{p}}^2}{1 - C_f \|\mathbf{h}\|_{\mathbf{p}}}.\quad (7-89)$$

证毕。  $\square$

#### 7.A.5 证明 定理 7.12

我们引入矩阵 Bernstein 不等式如下:

##### 引理 7.18: Matrix Bernstein inequality (Theorem 6.1, [156])

考虑一系列维数为  $d$  的独立随机自伴矩阵  $\{\mathbf{X}_k\}$ , 并且假定

$$\mathbb{E}\mathbf{X}_k = \mathbf{0} \quad \text{且} \quad \lambda_{\max}(\mathbf{X}_k) \leq R \quad \text{几乎必然成立.}$$

计算总方差的范数,

$$\sigma^2 := \left\| \sum_k \mathbb{E}(\mathbf{X}_k^2) \right\|.$$

则对于所有  $t \geq 0$ , 有以下一系列不等式成立:

$$\begin{aligned}\mathbb{P} \left\{ \lambda_{\max} \left( \sum_k \mathbf{X}_k \right) \geq t \right\} &\leq d \cdot \exp \left( -\frac{\sigma^2}{R^2} \cdot h \left( \frac{Rt}{\sigma^2} \right) \right) \leq d \cdot \exp \left( \frac{-t^2/2}{\sigma^2 + Rt/3} \right) \\ &\leq \begin{cases} d \cdot \exp \left( -\frac{3t^2}{8\sigma^2} \right) & \text{当 } t \leq \sigma^2/R, \\ d \cdot \exp \left( -\frac{3t}{8R} \right) & \text{当 } t \geq \sigma^2/R. \end{cases}\end{aligned}\quad (7-90)$$

其中函数  $h(u) := (1+u) \log(1+u) - u, u \geq 0$ .

为了证明这一点, 我们再次参考公式 (7-61)

$$\nabla^2 f_i(\mathbf{p}) = \frac{1}{r_i(\mathbf{p})}(\sigma_i + 1)\mathbf{P}^{-1}\mathbf{V}_i\mathbf{P}^{-1} - \frac{\sigma_i\mathbf{P}^{-1}\mathbf{v}_i\mathbf{v}_i^T\mathbf{P}^{-1}}{(r_i(\mathbf{p}))^2}. \quad (7-91)$$

再次利用  $\mathbf{u}_i$  的归一化 (其中  $\mathbf{u}_i \in \Delta_n$ ), 我们有

$$\begin{aligned} \nabla^2 f(\mathbf{p}) &= \sum_{i \in \mathcal{J}} w_i [(\sigma_i + 1)\mathbf{U}_i - \sigma_i \mathbf{u}_i \mathbf{u}_i^T] \\ &= \sum_{i \in \mathcal{J}} w_i (\sigma_i + 1) \mathbf{U}_i - \sum_{i \in \mathcal{J}} \sigma_i w_i \mathbf{u}_i \mathbf{u}_i^T. \end{aligned} \quad (7-92)$$

我们取

$$\Xi := \frac{\sum_{i \in \mathcal{J}} \sigma_i w_i \mathbf{u}_i \mathbf{u}_i^T}{\sum_{i \in \mathcal{J}} \sigma_i w_i} = \sum_{i \in \mathcal{J}} \gamma_i \mathbf{u}_i \mathbf{u}_i^T, \quad \sum_{i \in \mathcal{J}} \gamma_i = 1. \quad (7-93)$$

其思想是通过下面的秩 1 矩阵来近似  $\Xi$ :

$$\tilde{\Xi} := \bar{\mathbf{u}} \bar{\mathbf{u}}^T, \quad \bar{\mathbf{u}} := \sum_{i \in \mathcal{J}} \gamma_i \mathbf{u}_i. \quad (7-94)$$

不失一般性, 我们假定  $\sigma_i > 0$ , 否则, 我们可以估计

$$\sum_{i \in \mathcal{J}} (\sigma_i + 1) w_i \mathbf{u}_i \mathbf{u}_i^T \quad \text{和} \quad \sum_{i \in \mathcal{J}} w_i \mathbf{u}_i \mathbf{u}_i^T.$$

取

$$\mathbf{E} = \Xi - \bar{\mathbf{u}} \bar{\mathbf{u}}^T = \sum_{i \in \mathcal{J}} \gamma_i (\mathbf{u}_i - \bar{\mathbf{u}})(\mathbf{u}_i - \bar{\mathbf{u}})^T. \quad (7-95)$$

为了确保

$$\left\| \sum_{i \in \mathcal{J}} \sigma_i w_i \mathbf{u}_i \mathbf{u}_i^T - \left( \sum_{i \in \mathcal{J}} \sigma_i w_i \right) \bar{\mathbf{u}} \bar{\mathbf{u}}^T \right\| = \left| \sum_{i \in \mathcal{J}} \sigma_i w_i \right| \|\mathbf{E}\| = \left| \sum_{i \in \mathcal{J}} \sigma_i w_i \right| \|\Xi - \bar{\mathbf{u}} \bar{\mathbf{u}}^T\|_2 \leq \varepsilon, \quad (7-96)$$

只需令  $W = \max_{i \in \mathcal{J}} |2\sigma_i w_i|$ , 则有

$$\left\| \sum_{i \in \mathcal{J}} \gamma_i (\mathbf{u}_i - \bar{\mathbf{u}})(\mathbf{u}_i - \bar{\mathbf{u}})^T \right\|_2 \leq \varepsilon := \frac{\epsilon}{W \|\mathcal{J}\|}. \quad (7-97)$$

其失败概率至多为  $\delta$ . 定义中心化矩阵

$$\Sigma_i = \gamma_i (\mathbf{u}_i - \bar{\mathbf{u}})(\mathbf{u}_i - \bar{\mathbf{u}})^T,$$

则有  $\mathbb{E}[\Sigma_i] = \mathbf{0}$ . 因为

$$\|\mathbf{u}_i \mathbf{u}_i^T\|_2 \leq \text{tr}(\mathbf{u}_i \mathbf{u}_i^T) = \|\mathbf{u}_i\|_2^2 \leq \|\mathbf{u}_i\|_1 = 1,$$

所以我们有

$$\|\Sigma_i\|_2 \leq |2\gamma_i| \leq 2. \quad (7-98)$$

另外, 注意到

$$\|\Sigma_i^2\|_2 \leq \|\Sigma_i\|_2^2 \leq |2\gamma_i|^2 \leq 4.$$

因此, 方差参数满足

$$\sigma^2 = \left\| \sum_{i \in \mathcal{J}} \mathbb{E}[\Sigma_i^2] \right\|_2 \leq 4 \sum_{i \in \mathcal{J}} \gamma_i^2 \leq 4.$$

根据引理 7.18, 对于任意  $t > 0$ , 有

$$\mathbb{P} \left( \left\| \sum_{i \in \mathcal{J}} \Sigma_i \right\|_2 \geq t \right) \leq n \exp \left( -\frac{t^2/2}{4 + 4t/3} \right).$$

取  $t = \frac{\epsilon}{W\|\mathcal{J}\|}$ , 则要求

$$n \exp \left( -\frac{\epsilon^2/2}{4|\mathcal{J}|W^2 + 2Wt/3} \right) \leq \delta. \quad (7-99)$$

在统一权重的特例中,

$$w_i = \frac{1}{\|\mathcal{J}\|},$$

此时  $W = \frac{1}{\|\mathcal{J}\|}$ . 则不等式变为

$$\frac{\epsilon^2}{2 \left( \frac{4}{\|\mathcal{J}\|} + \frac{2\epsilon}{3\|\mathcal{J}\|} \right)} \geq \ln \left( \frac{n}{\delta} \right) \iff \frac{\|\mathcal{J}\|\epsilon^2}{2 \left( 4 + \frac{2\epsilon}{3} \right)} \geq \ln \left( \frac{n}{\delta} \right).$$

解得

$$\|\mathcal{J}\| \geq \frac{2 \left( 4 + \frac{2\epsilon}{3} \right) \ln \left( \frac{n}{\delta} \right)}{\epsilon^2}.$$

## 第二节 齐次模型的相关证明

### 7.B.1 证明 定理 7.4

*Proof.* 我们首先证明 (7-21) 等价于一个带投影算子  $\Pi_{\mathbf{p}}$  的广义信赖域问题。为了简化表示, 令  $\mathbf{v} \leftarrow \Pi_{\mathbf{p}}\mathbf{v}$ , 则仿射约束可以省略:

$$\begin{aligned} \min_{[\mathbf{v}; t] \in \mathbb{R}^{n+1}} \quad & \frac{1}{2} \langle \mathbf{v}, \Pi_{\mathbf{p}} \mathbf{P} \mathbf{H} \mathbf{P} \Pi_{\mathbf{p}} \mathbf{v} \rangle + \langle \mathbf{v}, \Pi_{\mathbf{p}} (\mathbf{P} \mathbf{g}(\mathbf{p}) - \mu \mathbf{1}) \rangle \cdot t + \frac{1}{2} \mu \|\Pi_{\mathbf{p}} \mathbf{v}\|^2 \\ \text{s.t.} \quad & \|[\Pi_{\mathbf{p}} \mathbf{v}; t]\| \leq 1. \end{aligned} \quad (7-100)$$

该问题是一个广义信赖域子问题, 目标函数是齐次的。根据<sup>[29]</sup> Theorem 8.2.7, (7-100) 的全局最优性条件如下:

$$\Pi_{\mathbf{p}} (\mathbf{P} \mathbf{H} \mathbf{P} + \mu \mathbf{I}) \Pi_{\mathbf{p}} \mathbf{v} + \theta \Pi_{\mathbf{p}} \mathbf{v} + t \cdot \Pi_{\mathbf{p}} (\mathbf{P} \mathbf{g}(\mathbf{p}) - \mu \mathbf{1}) = 0, \quad (7-101a)$$

$$\theta t + \langle \mathbf{v}, \Pi_{\mathbf{p}} (\mathbf{P} \mathbf{g}(\mathbf{p}) - \mu \mathbf{1}) \rangle = 0, \quad (7-101b)$$

$$\begin{bmatrix} \Pi_{\mathbf{p}} (\mathbf{P} \mathbf{H} \mathbf{P} + \mu \mathbf{I}) \Pi_{\mathbf{p}} & \Pi_{\mathbf{p}} (\mathbf{P} \mathbf{g}(\mathbf{p}) - \mu \mathbf{1}) \\ (\mathbf{P} \mathbf{g}(\mathbf{p}) - \mu \mathbf{1})^T \Pi_{\mathbf{p}} & 0 \end{bmatrix} + \theta \begin{bmatrix} \Pi_{\mathbf{p}} & 0 \\ 0 & 1 \end{bmatrix} \geq 0, \quad (7-101c)$$

$$0 \leq \theta \perp (1 - \|[\mathbf{v}; t]\|) \leq 0. \quad (7-101d)$$

由 (7-101c) 可得:

$$\begin{aligned} (7-101c) \implies & \begin{bmatrix} \Pi_{\mathbf{p}} (\mathbf{P} \mathbf{H} \mathbf{P} + (\theta + \mu) \mathbf{I}) \Pi_{\mathbf{p}} & \Pi_{\mathbf{p}} (\mathbf{P} \mathbf{g}(\mathbf{p}) - \mu \mathbf{1}) \\ (\mathbf{P} \mathbf{g}(\mathbf{p}) - \mu \mathbf{1})^T \Pi_{\mathbf{p}} & \theta \end{bmatrix} \geq 0, \\ \implies & \theta - \langle \Pi_{\mathbf{p}} (\mathbf{P} \mathbf{g}(\mathbf{p}) - \mu \mathbf{1}), (\Pi_{\mathbf{p}} (\mathbf{P} \mathbf{H} \mathbf{P} + (\theta + \mu) \mathbf{I}) \Pi_{\mathbf{p}})^{-1} \Pi_{\mathbf{p}} (\mathbf{P} \mathbf{g}(\mathbf{p}) - \mu \mathbf{1}) \rangle \geq 0. \end{aligned} \quad (7-102)$$



其中最后一个不等式由 Schur 补得出, 意味着  $\theta > 0$ , 因此  $\|[\Pi_{\mathbf{p}}\mathbf{v}; t]\| = 1$ 。进一步计算:

$$\frac{1}{2}\langle \mathbf{v}, \Pi_{\mathbf{p}}(\mathbf{PHP} + \mu\mathbf{I})\Pi_{\mathbf{p}}\mathbf{v} \rangle + \langle \mathbf{v}, \Pi_{\mathbf{p}}(\mathbf{Pg}(\mathbf{p}) - \mu\mathbf{1}) \rangle \cdot t - \frac{1}{2}\mu(\|\mathbf{v}\|^2 + t^2) = \frac{1}{2}\langle [\mathbf{v}; t], F(\mathbf{p}, \mu)[\mathbf{v}; t] \rangle.$$

证毕。  $\square$

### 7.B.2 证明 推论 7.5

*Proof.* 由于已知  $\|[\mathbf{v}; t]\| = 1$ , 我们仅需证明最后一个不等式。

$$\begin{aligned} \psi(\mathbf{v}, t) &= \frac{1}{2}\langle \mathbf{v}, \mathbf{PHPv} \rangle + \langle \mathbf{v}, \mathbf{Pg}(\mathbf{p}) - \mu\mathbf{1} \rangle \cdot t + \frac{1}{2}\mu\|\mathbf{v}\|^2 \\ &= \frac{1}{2}\langle \mathbf{v}, \mathbf{PHPv} \rangle + \langle \mathbf{v}, \mathbf{Pg}(\mathbf{p}) - \mu\mathbf{1} \rangle \cdot t + \frac{1}{2}\mu - \frac{1}{2}\mu t^2. \end{aligned} \quad (7-103)$$

由二阶条件 (7-101c) 再次进行推导 (此处已按  $-\mu$  进行缩放):

$$\begin{bmatrix} \Pi_{\mathbf{p}}(\mathbf{PHP} + \theta\mathbf{I})\Pi_{\mathbf{p}} & \Pi_{\mathbf{p}}(\mathbf{Pg}(\mathbf{p}) - \mu\mathbf{1}) \\ (\mathbf{Pg}(\mathbf{p}) - \mu\mathbf{1})^T\Pi_{\mathbf{p}} & \theta - \mu \end{bmatrix} \geq 0.$$

由 Schur 补得:

$$\Pi_{\mathbf{p}}(\mathbf{PHP} + \theta\mathbf{I})\Pi_{\mathbf{p}} - \frac{1}{\theta - \mu}\Pi_{\mathbf{p}}(\mathbf{Pg}(\mathbf{p}) - \mu\mathbf{1})(\Pi_{\mathbf{p}}(\mathbf{Pg}(\mathbf{p}) - \mu\mathbf{1}))^T \geq 0. \quad (7-104)$$

这意味着  $(\mathbf{PHP} + \theta\mathbf{I}) - \frac{1}{\theta - \mu}(\mathbf{Pg}(\mathbf{p}) - \mu\mathbf{1})(\mathbf{Pg}(\mathbf{p}) - \mu\mathbf{1})^T \geq 0$  在  $\ker(\mathbf{A}_p)$  上成立。  $\square$

### 第八章 总结与展望

随着时代的发展，优化方法的研究范畴不再局限于凸优化问题，非凸优化问题越来越受到人们的关注。我在博士一年级的接触到这些问题，最早的兴趣是在二次约束二次规划问题上，其中由何斯迈、江波、葛冬冬老师主讲的《锥优化》课程，以及一些与半正定规划相关的文献对我产生了较大的影响，如<sup>[90,137,152,167]</sup>，这些文献最终指向了齐次化方法的诞生。

2023 年，我受邀去复旦大学大数据学院报告。当时我将一些早期的研究成果与酆旭东，江如俊教授进行了交流，他们建议我应该仔细地思考齐次化方法的主要优势，这些讨论最终形成了对使用齐次化方法时，条件数估计的分析。2024 年七月，我受香港中文大学（深圳）Andre Milzarek 教授邀请，在国际优化大会（International Symposium on Mathematical programming）上做了关于齐次化方法的报告，期间 Kate Zhu (Oxford), Kim-Chuan Toh (NUS) 教授对论文的内容提出了不少建设性的意见。同年 11 月，东京大学的 Akiko Takeda 教授及其合作者将随机线性代数应用到齐次化方法的特征值求解上，取得了一些降维的效果。这个方法思路本质上可以认为是某种 Johnson-Lindenstrauss 定理的结果，这个结果最终作为 Spotlight 论文发表在 2025 年的 ICRL 上。

齐次化方法的研究不仅限于此，我们提出几个目前正在进行和发展的研究方向。对于内点法而言，已经可以证明，齐次化方法可以作用于 Primal, Dual, 以及 Primal-Dual 内点法上，尤其是对于 Potential Reduction 内点法，齐次化方法具有一定的计算优势；这些方法的分析与传统的牛顿法差异不大。由于内点法的矩阵分解模块已经发展的较为完善，基于迭代法的齐次化方法目前还很难与这些方法竞争。另一个可能的方法是高阶优化方法，利用齐次化框架对高阶多项式进行估计，并产生迭代步。利用齐次化方法后，高阶迭代步可以看成是张量特征值问题<sup>[87]</sup>。

参考文献

- [1] Martín Abadi. TensorFlow: Learning functions at scale. In **Proceedings of the 21st ACM SIGPLAN International Conference on Functional Programming**, ICFP 2016, page 1, New York, NY, USA, 2016. Association for Computing Machinery.
- [2] Tobias Achterberg. What’s new in gurobi 9.0, 2019. URL <https://www.gurobi.com/wp-content/uploads/2019/12/Gurobi-90-Overview-Webinar-Slides-1.pdf>.
- [3] Satoru Adachi and Yuji Nakatsukasa. Eigenvalue-based algorithm and analysis for nonconvex QCQP with one constraint. **Mathematical Programming**, 173(1-2):79–116, 2019.
- [4] Satoru Adachi, Satoru Iwata, Yuji Nakatsukasa, and Akiko Takeda. Solving the Trust-Region Subproblem By a Generalized Eigenvalue Problem. **SIAM Journal on Optimization**, 27(1):269–291, 2017.
- [5] Naman Agarwal, Zeyuan Allen-Zhu, Brian Bullins, Elad Hazan, and Tengyu Ma. Finding approximate local minima faster than gradient descent. In **Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing**, pages 1195–1199, 2017.
- [6] Ravindra K. Ahuja, Thomas L. Magnanti, and James B. Orlin. **Network Flows: Theory, Algorithms, and Applications**. Prentice Hall, Upper Saddle River, NJ, 1993.
- [7] Edward Anderson, Zhaojun Bai, Christian Bischof, L. Susan Blackford, James Demmel, Jack Dongarra, Jeremy Du Croz, Anne Greenbaum, Sven Hammarling, and Alan McKenney. **LAPACK Users’ Guide**. SIAM, 1999.
- [8] David Applegate, Mateo Díaz, Oliver Hinder, Haihao Lu, Miles Lubin, Brendan O’Donoghue, and Warren Schudy. Practical Large-Scale Linear Programming using Primal-Dual Hybrid Gradient, 2022.

- 
- [9] David Applegate, Oliver Hinder, Haihao Lu, and Miles Lubin. Faster first-order primal-dual methods for linear programming using restarts and sharpness. **Mathematical Programming**, 201(1-2):133–184, 2023.
- [10] Ronald D. Armstrong and Zhiying Jin. A new strongly polynomial dual network simplex algorithm. **Mathematical Programming**, 78(2):131–148, 1997.
- [11] Kenneth J. Arrow and Gerard Debreu. Existence of an Equilibrium for a Competitive Economy. **Econometrica**, 22(3):265–290, 1954.
- [12] Achraf Bahamou and Donald Goldfarb. Layer-wise Adaptive Step-Sizes for Stochastic First-Order Methods for Deep Learning, 2023.
- [13] R. Baldick. The generalized unit commitment problem. **IEEE Transactions on Power Systems**, 10(1):465–475, 1995.
- [14] Aharon Ben-Tal and Arkadi Nemirovski. **Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications**. Society for Industrial and Applied Mathematics, 2001.
- [15] Jeff Bezanson, Alan Edelman, Stefan Karpinski, and Viral B. Shah. Julia: A fresh approach to numerical computing. **SIAM review**, 59(1):65–98, 2017.
- [16] Benjamin Birnbaum, Nikhil R. Devanur, and Lin Xiao. Distributed algorithms via gradient descent for fisher markets. In **Proceedings of the 12th ACM Conference on Electronic Commerce**, EC ’11, pages 127–136, New York, NY, USA, 2011. Association for Computing Machinery.
- [17] L. Susan Blackford, Antoine Petitet, Roldan Pozo, Karin Remington, R. Clint Whaley, James Demmel, Jack Dongarra, Iain Duff, Sven Hammarling, and Greg Henry. An updated set of basic linear algebra subprograms (BLAS). **ACM Transactions on Mathematical Software**, 28(2):135–151, 2002.
- [18] Lenore Blum, Mike Shub, and Steve Smale. On a theory of computation and complexity over the real numbers: - completeness, recursive functions and universal machines. **Bulletin of the American Mathematical Society**, 21(1):1–46, 1989.

- 
- [19] Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization Methods for Large-Scale Machine Learning. **SIAM Review**, 60(2):223–311, 2018.
- [20] William C. Brainard and Herbert E. Scarf. How to Compute Equilibrium Prices in 1891. **The American Journal of Economics and Sociology**, 64(1):57–83, 2005.
- [21] R. H. Byrd, S. L. Hansen, Jorge Nocedal, and Y. Singer. A Stochastic Quasi-Newton Method for Large-Scale Optimization. **SIAM Journal on Optimization**, 26(2):1008–1031, 2016.
- [22] Richard H. Byrd, Jorge Nocedal, and Richard A. Waltz. Knitro: An Integrated Package for Nonlinear Optimization. In G. Di Pillo and M. Roma, editors, **Large-Scale Nonlinear Optimization**, pages 35–59. Springer US, Boston, MA, 2006.
- [23] Yair Carmon, John C. Duchi, Oliver Hinder, and Aaron Sidford. Accelerated Methods for NonConvex Optimization. **SIAM Journal on Optimization**, 28(2):1751–1772, 2018.
- [24] Yair Carmon, John C. Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points I. **Mathematical Programming**, 184(1):71–120, 2020.
- [25] C. Cartis, N. I. M. Gould, and Ph. L. Toint. On the Complexity of Steepest Descent, Newton’s and Regularized Newton’s Methods for Nonconvex Unconstrained Optimization Problems. **SIAM Journal on Optimization**, 20(6):2833–2852, 2010.
- [26] C. Cartis, N. I. M. Gould, and Ph. L. Toint. Complexity bounds for second-order optimality in unconstrained optimization. **Journal of Complexity**, 28(1):93–108, 2012.
- [27] Coralia Cartis, Nicholas I. M. Gould, and Philippe L. Toint. Adaptive cubic regularisation methods for unconstrained optimization. Part II: Worst-case function- and derivative-evaluation complexity. **Mathematical Programming**, 130(2):295–319, 2011.

- 
- [28] Coralia Cartis, Nicholas I. M. Gould, and Philippe L. Toint. Adaptive cubic regularisation methods for unconstrained optimization. Part I: Motivation, convergence and numerical results. **Mathematical Programming**, 127(2):245–295, 2011.
- [29] Coralia Cartis, Nicholas I. M. Gould, and Philippe L. Toint. **Evaluation Complexity of Algorithms for Nonconvex Optimization: Theory, Computation and Perspectives**. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2022.
- [30] Augustin Cauchy. Méthode générale pour la résolution des systemes d’équations simultanées. **Comp. Rend. Sci. Paris**, 25(1847):536–538, 1847.
- [31] Xiaojun Chen, Lingfeng Niu, and Yaxiang Yuan. Optimality Conditions and a Smoothing Trust Region Newton Method for NonLipschitz Optimization. **SIAM Journal on Optimization**, 23(3):1528–1552, 2013.
- [32] Xiaojun Chen, Dongdong Ge, Zizhuo Wang, and Yinyu Ye. Complexity of unconstrained L2-Lp minimization. **Mathematical Programming**, 143(1-2):371–383, 2014.
- [33] Yun Kuen Cheung, Richard Cole, and Nikhil Devanur. Tatonnement beyond gross substitutes? gradient descent to the rescue. In **Proceedings of the Forty-Fifth Annual ACM Symposium on Theory of Computing**, STOC ’13, pages 191–200, New York, NY, USA, 2013. Association for Computing Machinery.
- [34] Yun Kuen Cheung, Richard Cole, and Yixin Tao. Dynamics of Distributed Updating in Fisher Markets. In **Proceedings of the 2018 ACM Conference on Economics and Computation**, EC ’18, pages 351–368, New York, NY, USA, 2018. Association for Computing Machinery.
- [35] John S. Chipman. Multiple equilibrium under CES preferences. **Economic Theory**, 45(1):129–145, 2010.
- [36] Bruno Codenotti and Kasturi Varadarajan. Computation of Market Equilibria by Convex Programming. In Eva Tardos, Noam Nisan, Tim Roughgarden, and Vijay V. Vazirani, editors, **Algorithmic Game Theory**, pages 135–158. Cambridge University Press, Cambridge, 2007.

- 
- [37] Bruno Codenotti, Benton McCune, Sriram Penumatcha, and Kasturi Varadarajan. Market Equilibrium for CES Exchange Economies: Existence, Multiplicity, and Computation. In Sundar Sarukkai and Sandeep Sen, editors, **FSTTCS 2005: Foundations of Software Technology and Theoretical Computer Science**, pages 505–516, Berlin, Heidelberg, 2005. Springer.
- [38] Andrew R. Conn, Nicholas I. M. Gould, and Philippe L. Toint. **Trust-Region Methods**. MPS-SIAM Series on Optimization. Society for Industrial and Applied Mathematics, Philadelphia, Pa, 2000.
- [39] Frank E. Curtis and Daniel P. Robinson. Exploiting negative curvature in deterministic and stochastic optimization. **Mathematical Programming**, 176(1):69–94, 2019.
- [40] Frank E. Curtis and Qi Wang. Worst-Case Complexity of TRACE with Inexact Subproblem Solutions for Nonconvex Smooth Optimization. **SIAM Journal on Optimization**, 33(3):2191–2221, 2023.
- [41] Frank E. Curtis, Daniel P. Robinson, and Mohammadreza Samadi. A trust region algorithm with a worst-case iteration complexity of  $\mathcal{O}(\epsilon^{-3/2})$  for nonconvex optimization. **Mathematical Programming**, 162(1):1–32, 2017.
- [42] Frank E. Curtis, Zachary Lubberts, and Daniel P. Robinson. Concise complexity analyses for trust region methods. **Optimization Letters**, 12(8):1713–1724, 2018.
- [43] Frank E. Curtis, Daniel P. Robinson, and Mohammadreza Samadi. An inexact regularized Newton framework with a worst-case iteration complexity of for nonconvex optimization. **IMA Journal of Numerical Analysis**, 39(3):1296–1327, 2019.
- [44] Frank E. Curtis, Daniel P. Robinson, Clément W. Royer, and Stephen J. Wright. Trust-Region Newton-CG with Strong Second-Order Complexity Guarantees for Nonconvex Optimization. **SIAM Journal on Optimization**, 31(1):518–544, 2021.
- [45] B. Curtis Eaves. Finite solution of pure trade markets with Cobb-Douglas utilities. In Alan S. Manne, editor, **Economic Equilibrium: Model Formulation and Solution**, pages 226–239. Springer, Berlin, Heidelberg, 1985.

- 
- [46] D. den Hertog, F. Jarre, C. Roos, and T. Terlaky. A sufficient condition for self-concordance, with application to some classes of structured convex programming problems. **Mathematical Programming**, 69(1):75–88, 1995.
- [47] Qi Deng, Qing Feng, Wenzhi Gao, Dongdong Ge, Bo Jiang, Yuntian Jiang, Jingsong Liu, Tianhao Liu, Chenyu Xue, Yinyu Ye, and Chuwen Zhang. An Enhanced Alternating Direction Method of Multipliers-Based Interior Point Method for Linear and Conic Optimization. **INFORMS Journal on Computing**, 2024.
- [48] N.R. Devanur, C.H. Papadimitriou, A. Saberi, and V.V. Vazirani. Market equilibrium via a primal-dual-type algorithm. In **The 43rd Annual IEEE Symposium on Foundations of Computer Science, 2002. Proceedings.**, pages 389–395, 2002.
- [49] Nikita Doikov and Yurii Nesterov. Minimizing Uniformly Convex Functions by Cubic Regularization of Newton Method. **Journal of Optimization Theory and Applications**, 189(1):317–339, 2021.
- [50] Nikita Doikov and Yurii Nesterov. Gradient regularization of Newton method with Bregman distances. **Mathematical Programming**, 204(1):1–25, 2024.
- [51] Nikita Doikov, Konstantin Mishchenko, and Yurii Nesterov. Super-Universal Regularized Newton Method. **SIAM Journal on Optimization**, 34(1):27–56, 2024.
- [52] Elizabeth D. Dolan and Jorge J. Moré. Benchmarking optimization software with performance profiles. **Mathematical Programming**, 91(2):201–213, 2002.
- [53] Jean-Pierre Dussault, Tangi Migot, and Dominique Orban. Scalable adaptive cubic regularization methods. **Mathematical Programming**, 207(1):191–225, 2024.
- [54] Pavel Dvurechensky and Mathias Staudigl. Hessian barrier algorithms for non-convex conic optimization. **Mathematical Programming**, 2024.
- [55] Pavel Dvurechensky, Kamil Safin, Shimrit Shtern, and Mathias Staudigl. Generalized self-concordant analysis of Frank–Wolfe algorithms. **Mathematical Programming**, pages 1–69, 2022.



- 
- [56] B. Curtis Eaves. A finite algorithm for the linear exchange model. **Journal of Mathematical Economics**, 3(2):197–203, 1976.
- [57] Edmund Eisenberg and David Gale. Consensus of Subjective Probabilities: The Pari-Mutuel Method. **The Annals of Mathematical Statistics**, 30(1):165–168, 1959.
- [58] Mercedes Esteban-Bravo. Computing Equilibria in General Equilibrium Models via Interior-point Methods. **Computational Economics**, 23(2):147–171, 2004.
- [59] Francisco Facchinei and Christian Kanzow. Generalized Nash Equilibrium Problems. **Annals of Operations Research**, 175(1):177–211, 2010.
- [60] Francisco Facchinei and Jong-Shi Pang, editors. **Finite-Dimensional Variational Inequalities and Complementarity Problems**. Springer Series in Operations Research and Financial Engineering. Springer New York, New York, NY, 2004.
- [61] Ilyas Fatkhullin, Anas Barakat, Anastasia Kireeva, and Niao He. Stochastic Policy Gradient Methods: Improved Sample Complexity for Fisher-non-degenerate Policies, 2023.
- [62] Dylan J. Foster, Ayush Sekhari, and Karthik Sridharan. Uniform convergence of gradients for non-convex learning and optimization. **Advances in Neural Information Processing Systems**, 31, 2018.
- [63] Yuan Gao and Christian Kroer. First-Order Methods for Large-Scale Market Equilibrium Computation. In **Advances in Neural Information Processing Systems**, volume 33, pages 21738–21750. Curran Associates, Inc., 2020.
- [64] Dongdong Ge, Xiaoye Jiang, and Yinyu Ye. A note on the complexity of  $L_p$  minimization. **Mathematical Programming**, 129(2):285–299, 2011.
- [65] Dongdong Ge, Rongchuan He, and Simai He. An improved algorithm for the  $L_2$ - $L_p$  minimization problem. **Mathematical Programming**, 166(1):131–158, 2017.
- [66] Dongdong Ge, Qi Huangfu, Zizhuo Wang, Jian Wu, and Yinyu Ye. Cardinal Optimizer (COPT) User Guide, 2022.

- 
- [67] Steven Gjerstad. Multiple equilibria in exchange economies with homothetic, nearly identical preferences. Discussion Paper No. 288, Center for Economic Research, 1996.
- [68] Denizalp Goktas, Enrique Areyan Viqueira, and Amy Greenwald. A Consumer-Theoretic Characterization of Fisher Market Equilibria. In Michal Feldman, Hu Fu, and Inbal Talgam-Cohen, editors, **Web and Internet Economics**, pages 334–351, Cham, 2022. Springer International Publishing.
- [69] Denizalp Goktas, Jiayi Zhao, and Amy Greenwald. Tâtonnement in Homothetic Fisher Markets. In **Proceedings of the 24th ACM Conference on Economics and Computation**, EC '23, pages 760–781, New York, NY, USA, 2023. Association for Computing Machinery.
- [70] Donald Goldfarb and Jianxiu Hao. A primal simplex algorithm that solves the maximum flow problem in at most  $n^2$  pivots and  $O(n^2m)$  time. **Mathematical Programming**, 47(1):353–365, 1990.
- [71] Gene H. Golub and Charles F. Van Loan. **Matrix Computations**. Johns Hopkins Studies in the Mathematical Sciences. The Johns Hopkins University Press, Baltimore, fourth edition edition, 2013.
- [72] Nicholas I. M. Gould, Yueling Loh, and Daniel P. Robinson. A Nonmonotone Filter SQP Method: Local Convergence and Numerical Results. **SIAM Journal on Optimization**, 25(3):1885–1911, 2015.
- [73] Nicholas IM Gould and Valeria Simoncini. Error estimates for iterative algorithms for minimizing regularized quadratic subproblems. **Optimization Methods and Software**, 35(2):304–328, 2020.
- [74] Geovani N. Grapiglia, Jinyun Yuan, and Ya-xiang Yuan. On the convergence and worst-case complexity of trust-region and regularization methods for unconstrained optimization. **Mathematical Programming**, 152(1-2):491–520, 2015.

- [75] A Griewank. The modification of Newton’s method for unconstrained optimization by bounding cubic terms. Technical report, University of Cambridge, Department of Applied Mathematics and Theoretical Physics, University of Cambridge.
- [76] Martin Grötschel, László Lovász, and Alexander Schrijver. **Geometric Algorithms and Combinatorial Optimization**, volume 2 of **Algorithms and Combinatorics**. Springer Berlin Heidelberg, Berlin, Heidelberg, 1993.
- [77] Jutho Haegeman. KrylovKit. Zenodo, 2024.
- [78] Gabriel Haeser, Hongcheng Liu, and Yinyu Ye. Optimality condition and complexity analysis for linearly-constrained optimization without differentiability on the boundary. **Mathematical Programming**, 178(1):263–299, 2019.
- [79] William W. Hager and Hongchao Zhang. Algorithm 851: CG\_DESCENT, a conjugate gradient method with guaranteed descent. **ACM Transactions on Mathematical Software (TOMS)**, 32(1):113–137, 2006.
- [80] Elad Hazan and Tomer Koren. A linear-time algorithm for trust region problems. **Mathematical Programming**, 158(1):363–381, 2016.
- [81] Chang He, Yuntian Jiang, Chuwen Zhang, Dongdong Ge, Bo Jiang, and Yinyu Ye. Homogeneous Second-Order Descent Framework: A Fast Alternative to Newton-Type Methods, 2023.
- [82] Chuan He and Zhaosong Lu. A Newton-CG Based Barrier Method for Finding a Second-Order Stationary Point of Nonconvex Conic Optimization with Complexity Guarantees. **SIAM Journal on Optimization**, 33(2):1191–1222, 2023.
- [83] David Hilbert. Ein Beitrag zur Theorie des Legendre’schen Polynoms. **Acta Mathematica**, 18(none):155–159, 1900.
- [84] Kamal Jain. A Polynomial Time Algorithm for Computing an Arrow–Debreu Market Equilibrium for Linear Utilities. **SIAM Journal on Computing**, 37(1):303–318, 2007.

- 
- [85] Kamal Jain, Vijay V. Vazirani, and Yinyu Ye. Market equilibria for homothetic, quasi-concave utilities and economies of scale in production. In **Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms**, SODA '05, pages 63–71, USA, 2005. Society for Industrial and Applied Mathematics.
- [86] Xiaojing Jia, Xin Liang, Chungeng Shen, and Lei-Hong Zhang. Solving the Cubic Regularization Model by a Nested Restarting Lanczos Method. **SIAM Journal on Matrix Analysis and Applications**, 43(2):812–839, 2022.
- [87] Bo Jiang, Shiqian Ma, and Shuzhong Zhang. Tensor principal component analysis via convex optimization. **Mathematical Programming**, 150(2):423–457, May 2015. ISSN 1436-4646. doi: 10.1007/s10107-014-0774-0.
- [88] Bo Jiang, Tianyi Lin, and Shuzhong Zhang. A Unified Adaptive Tensor Approximation Scheme to Accelerate Composite Convex Optimization. **SIAM Journal on Optimization**, 30(4):2897–2926, 2020.
- [89] Bo Jiang, Haoyue Wang, and Shuzhong Zhang. An optimal high-order tensor method for convex optimization. **Mathematics of Operations Research**, 46(4):1390–1412, 2021.
- [90] Rujun Jiang and Duan Li. Simultaneous diagonalization of matrices and its applications in quadratically constrained quadratic programming. **SIAM Journal on Optimization**, 26(3):1649–1668, 2016.
- [91] Rujun Jiang and Duan Li. A Linear-Time Algorithm for Generalized Trust Region Subproblems. **SIAM Journal on Optimization**, 30(1):915–932, 2020.
- [92] Yuntian Jiang, Chang He, Chuwen Zhang, Dongdong Ge, Bo Jiang, and Yinyu Ye. Beyond Nonconvexity: A Universal Trust-Region Method with New Analyses, 2024.
- [93] Chi Jin. Lecture 14: Lanczos Algorithm (March 22, 2021), 2021.
- [94] Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M. Kakade, and Michael I. Jordan. How to escape saddle points efficiently. In **International Conference on Machine Learning**, pages 1724–1732. PMLR, 2017.

- 
- [95] Alejandro Jofré, R. Terry Rockafellar, and Roger J-B. Wets. Variational Inequalities and Economic Equilibrium. **Mathematics of Operations Research**, 32(1):32–50, 2007.
- [96] Norman H. Josephy. Newton’s Method for Generalized Equations. Technical report, Wisconsin Univ-Madison Mathematics Research Center, 1979.
- [97] Patrick K Mogensen and Asbjørn N Riseth. Optim: A mathematical optimization package for Julia. **Journal of Open Source Software**, 3(24):615, 2018.
- [98] L. V. Kantorovich and Gleb Pavlovich Akilov. **Functional Analysis**. Pergamon Press, Oxford ; New York, 2d ed edition, 1982.
- [99] Narendra Karmarkar. A new polynomial-time algorithm for linear programming. In **Proceedings of the Sixteenth Annual ACM Symposium on Theory of Computing**, pages 302–311, 1984.
- [100] Leonid Genrikhovich Khachiyan. A polynomial algorithm in linear programming. In **Doklady Akademii Nauk**, volume 244, pages 1093–1096. Russian Academy of Sciences, 1979.
- [101] Victor Klee and George J. Minty. How good is the simplex algorithm. **Inequalities**, 3(3):159–175, 1972.
- [102] K. O. Kortanek and Jishan Zhu. A Polynomial Barrier Algorithm for Linearly Constrained Convex Programming Problems. **Mathematics of Operations Research**, 18(1):116–127, 1993.
- [103] J. Kuczyski and H. Woniakowski. Estimating the Largest Eigenvalue by the Power and Lanczos Algorithms with a Random Start. **SIAM Journal on Matrix Analysis and Applications**, 13(4):1094–1122, 1992.
- [104] Guanghui Lan. Complexity of stochastic dual dynamic programming. **Mathematical Programming**, 2020.

- 
- [105] Guanghui Lan. **First-Order and Stochastic Optimization Methods for Machine Learning**. Springer Series in the Data Sciences Ser. Springer International Publishing AG, Cham, 2020.
- [106] R. B. Lehoucq, D. C. Sorensen, and C. Yang. **ARPACK Users' Guide: Solution of Large-Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods**. Software, Environments, Tools. SIAM, Philadelphia, 1998.
- [107] Claude Lemaréchal. Cauchy and the gradient method. **Doc Math Extra**, 251(254):10, 2012.
- [108] Haihao Lu and Jinwen Yang. cuPDLP.jl: A GPU Implementation of Restarted Primal-Dual Hybrid Gradient for Linear Programming in Julia, 2023.
- [109] Haihao Lu, Jinwen Yang, Haodong Hu, Qi Huangfu, Jinsong Liu, Tianhao Liu, Yinyu Ye, Chuwen Zhang, and Dongdong Ge. cuPDLP-C: A Strengthened Implementation of cuPDLP for Linear Programming by C language, 2024.
- [110] David G. Luenberger. The gradient projection method along geodesics. **Management Science**, 18(11):620–631, 1972.
- [111] David G. Luenberger and Yinyu Ye. **Linear and Nonlinear Programming**, volume 228 of **International Series in Operations Research & Management Science**. Springer International Publishing, Cham, 2021.
- [112] José Mario Martínez. Local Minimizers of Quadratic Functions on Euclidean Balls and Spheres. **SIAM Journal on Optimization**, 4(1):159–176, 1994.
- [113] Andreu Mas-Colell, Michael D. Whinston, and Jerry R. Green. **Microeconomic Theory**. Oxford Univ. Press, New York, NY, 1995.
- [114] Saeed Masiha, Saber Salehkaleybar, Niao He, Negar Kiyavash, and Patrick Thiran. Stochastic Second-Order Methods Provably Beat SGD For Gradient-Dominated Functions. **arXiv preprint arXiv:2205.12856**, 2022.

- 
- [115] Konstantin Mishchenko. Regularized Newton Method with Global  $\mathcal{O}(1/k^2)$  Convergence. **SIAM Journal on Optimization**, 33(3):1440–1462, 2023.
  - [116] Shinji Mizuno, Michael J. Todd, and Yinyu Ye. On adaptive-step primal-dual interior-point algorithms for linear programming. **Mathematics of Operations research**, 18(4):964–981, 1993.
  - [117] Renato D. C. Monteiro and B. F. Svaiter. An Accelerated Hybrid Proximal Extragradient Method for Convex Optimization and Its Implications to Second-Order Methods. **SIAM Journal on Optimization**, 23(2):1092–1125, 2013.
  - [118] Jorge J. Moré and D. C. Sorensen. Computing a Trust Region Step. **SIAM Journal on Scientific and Statistical Computing**, 4(3):553–572, 1983.
  - [119] A. S. Nemirovskii and D. B. Yudin. **Problem complexity and method efficiency in optimization**. Wiley-Interscience series in discrete mathematics. Wiley, Chichester ; New York, 1983.
  - [120] EI Nenakov and ME Primak. One algorithm for finding solutions of the arrow-debreu model. **Kibernetika**, 3:127–128, 1983.
  - [121] Yu Nesterov. Accelerating the cubic regularization of Newton’s method on convex problems. **Mathematical Programming**, 112(1):159–181, 2008.
  - [122] Yurii Nesterov. **Introductory Lectures on Convex Optimization**, volume 87 of **Applied Optimization**. Springer US, Boston, MA, 2004.
  - [123] Yurii Nesterov. Unconstrained Convex Minimization in Relative Scale. **Mathematics of Operations Research**, 34(1):180–193, 2009.
  - [124] Yurii Nesterov. How To Make the Gradients Small. **Optima. Mathematical Optimization Society Newsletter**, (88), 2012.
  - [125] Yurii Nesterov. **Lectures on Convex Optimization**, volume 137 of **Springer Optimization and Its Applications**. Springer International Publishing, Cham, 2018.

- [126] Yurii Nesterov. Implementable tensor methods in unconstrained convex optimization. **Mathematical Programming**, 186(1):157–183, 2021.
- [127] Yurii Nesterov and Arkadii Nemirovskii. **Interior-Point Polynomial Algorithms in Convex Programming**. Society for Industrial and Applied Mathematics, 1994.
- [128] Yurii Nesterov and B.T. Polyak. Cubic regularization of Newton method and its global performance. **Mathematical Programming**, 108(1):177–205, 2006.
- [129] Isaac Newton. **De Analysi per Aequationes Numero Terminorum Infinitas**. 1711.
- [130] Jorge Nocedal and Stephen J. Wright. **Numerical Optimization**. Springer Series in Operations Research and Financial Engineering. Springer, New York, NY, second edition edition, 2006.
- [131] Brendan O’Donoghue. Operator Splitting for a Homogeneous Embedding of the Linear Complementarity Problem. **SIAM Journal on Optimization**, 31(3):1999–2023, 2021.
- [132] Dominique Orban and Abel Siqueira. JuliaSmoothOptimizers. Zenodo, 2019.
- [133] Brendan O’donoghue, Eric Chu, Neal Parikh, and Stephen Boyd. Conic optimization via operator splitting and homogeneous self-dual embedding. **Journal of Optimization Theory and Applications**, 169(3):1042–1068, 2016.
- [134] François Pacaud and Sungho Shin. GPU-accelerated dynamic nonlinear optimization with ExaModels and MadNLP, 2024.
- [135] François Pacaud, Michel Schanen, Sungho Shin, Daniel Adrian Maldonado, and Mihai Anitescu. Parallel Interior-Point Solver for Block-Structured Nonlinear Programs on SIMD/GPU Architectures, 2023.
- [136] Christos H. Papadimitriou and Kenneth Steiglitz. **Combinatorial Optimization: Algorithms and Complexity**. Dover Publ, Mineola, NY, corr., unabridged republ. of the work orig. publ. in 1982 by prentice-hall edition, 1998.



- [137] Panos M. Pardalos and Stephen A. Vavasis. Quadratic programming with one negative eigenvalue is NP-hard. **Journal of Global Optimization**, 1(1):15–22, 1991.
- [138] Panos M. Pardalos and Stephen A. Vavasis. Open questions in complexity theory for numerical optimization. **Mathematical Programming**, 57(1):337–339, 1992.
- [139] Beresford N. Parlett. **The Symmetric Eigenvalue Problem**. Number 20 in Classics in Applied Mathematics. Society for Industrial and Applied Mathematics, Philadelphia, 1998.
- [140] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In **Advances in Neural Information Processing Systems**, volume 32. Curran Associates, Inc., 2019.
- [141] Yves Pochet and Laurence A. Wolsey. **Production Planning by Mixed Integer Programming**. Springer Series in Operations Research and Financial Engineering. Springer, New York ; Berlin, 2006.
- [142] Florian Potra and Yinyu Ye. A Quadratically Convergent Polynomial Algorithm for Solving Entropy Optimization Problems. **SIAM Journal on Optimization**, 3(4): 843–860, 1993.
- [143] Joseph Raphson. **Analysis aequationum universalis seu ad aequationes algebraicas resolvendas methodus generalis, & expedita, ex nova infinitarum serierum methodo, deducta ac demonstrata; cui annexum est de spatio reali, seu ente infinito conamen mathematico-metaphysicum**. Th. Braddyll, 1697.
- [144] R. T. Rockafellar. Augmented Lagrangians and Applications of the Proximal Point Algorithm in Convex Programming. **Mathematics of Operations Research**, 1(2): 97–116, 1976.

- 
- [145] R Tyrrell Rockafellar and Roger J-B Wets. **Variational Analysis**, volume 317. Springer Science & Business Media, 2009.
- [146] Marielba Rojas, Sandra A. Santos, and Danny C. Sorensen. A New Matrix-Free Algorithm for the Large-Scale Trust-Region Subproblem. **SIAM Journal on Optimization**, 11(3):611–646, 2001.
- [147] Clément W. Royer and Stephen J. Wright. Complexity analysis of second-order line-search algorithms for smooth nonconvex optimization. **SIAM Journal on Optimization**, 28(2):1448–1477, 2018.
- [148] Clément W. Royer, Michael O’Neill, and Stephen J. Wright. A Newton-CG algorithm with complexity guarantees for smooth unconstrained optimization. **Mathematical Programming**, 180(1):451–488, 2020.
- [149] Y. Saad. **Numerical Methods for Large Eigenvalue Problems**. Number 66 in Classics in Applied Mathematics. Society for Industrial and Applied Mathematics, Philadelphia, rev. ed edition, 2011.
- [150] Sungho Shin, François Pacaud, and Mihai Anitescu. Accelerating Optimal Power Flow with GPUs: SIMD Abstraction of Nonlinear Programs and Condensed-Space Interior-Point Methods, 2024.
- [151] V. I. Shmyrev. An algorithm for finding equilibrium in the linear exchange model with fixed budgets. **Journal of Applied and Industrial Mathematics**, 3(4):505–518, 2009.
- [152] Jos F. Sturm and Shuzhong Zhang. On Cones of Nonnegative Quadratic Functions. **Mathematics of Operations Research**, 28(2):246–267, 2003.
- [153] Tianxiao Sun and Quoc Tran-Dinh. Generalized self-concordant functions: A recipe for Newton-type methods. **Mathematical Programming**, 178(1):145–213, 2019.
- [154] Quoc Tran-Dinh, Tianxiao Sun, and Shu Lu. Self-concordant inclusions: A unified framework for path-following generalized Newton-type algorithms. **Mathematical Programming**, 177(1):173–223, 2019.

- 
- [155] J. F. Traub and H. Woniakowski. Complexity of linear programming. **Operations Research Letters**, 1(2):59–62, 1982.
- [156] Joel A. Tropp. User-Friendly Tail Bounds for Sums of Random Matrices. **Foundations of Computational Mathematics**, 12(4):389–434, 2012.
- [157] Stephen A. Vavasis and Richard Zippel. Proving polynomial-time for sphere-constrained quadratic programming. Technical report, Cornell University, 1990.
- [158] Léon Walras. **Éléments d'économie politique pure ou Théorie de la richesse sociale**. Corbaz & Cie, Lausanne/Paris, 1874.
- [159] Jiulin Wang and Yong Xia. Closing the Gap between Necessary and Sufficient Conditions for Local Nonglobal Minimizer of Trust Region Subproblem. **SIAM Journal on Optimization**, 30(3):1980–1995, 2020.
- [160] Andreas Wächter and Lorenz T. Biegler. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. **Mathematical Programming**, 106(1):25–57, 2006.
- [161] Yong Xia. A Survey of Hidden Convex Optimization. **Journal of the Operations Research Society of China**, 8(1):1–28, 2020.
- [162] Peng Xu, Jiyan Yang, Fred Roosta, Christopher Ré, and Michael W. Mahoney. Subsampled Newton methods with non-uniform sampling. **Advances in Neural Information Processing Systems**, 29, 2016.
- [163] Yinyu Ye. A New Complexity Result on Minimization of a Quadratic Function with a Sphere Constraint. In **Recent Advances in Global Optimization**, volume 176, pages 19–31. Princeton University Press, 1991.
- [164] Yinyu Ye. Combining Binary Search and Newtons Method to Compute Real Roots for a Class of Real Functions. **Journal of Complexity**, 10(3):271–280, 1994.
- [165] Yinyu Ye. **Interior Point Algorithms: Theory and Analysis**. Wiley, 1 edition, 1997.

- [166] Yinyu Ye. On the complexity of approximating a KKT point of quadratic programming. **Mathematical Programming**, 80(2):195–211, 1998.
- [167] Yinyu Ye. Approximating quadratic programming with bound and quadratic constraints. **Mathematical Programming**, 84(2):219–226, 1999.
- [168] Yinyu Ye. Second Order Optimization Algorithms I, 2005.
- [169] Yinyu Ye. Second Order Optimization Algorithms II: Interior-Point Algorithms, 2005.
- [170] Yinyu Ye. A path to the Arrow–Debreu competitive market equilibrium. **Mathematical Programming**, 111(1):315–348, 2008.
- [171] Yinyu Ye. The Simplex and Policy-Iteration Methods Are Strongly Polynomial for the Markov Decision Problem with a Fixed Discount Rate. **Mathematics of Operations Research**, 36(4):593–603, 2011.
- [172] Yinyu Ye. A Second-Order Path-Following Algorithm for Unconstrained Convex Optimization. 2017.
- [173] Yinyu Ye and Shuzhong Zhang. New results on quadratic minimization. **SIAM Journal on Optimization**, 14(1):245–267, 2003.
- [174] Yinyu Ye, Osman Güler, Richard A. Tapia, and Yin Zhang. A quadratically convergent  $O(L)$ -iteration algorithm for linear programming. **Mathematical programming**, 59(1-3):151–162, 1993.
- [175] Man-Chung Yue, Zirui Zhou, and Anthony Man-Cho So. A family of inexact SQA methods for non-smooth convex minimization with provable convergence guarantees based on the Luo–Tseng error bound property. **Mathematical Programming**, 174(1-2):327–358, 2019.
- [176] Chuwen Zhang, Dongdong Ge, Chang He, Bo Jiang, Yuntian Jiang, Chenyu Xue, and Yinyu Ye. A Homogeneous Second-Order Descent Method for Nonconvex Optimization, 2022.

- [177] Chuwen Zhang, Dongdong Ge, Chang He, Bo Jiang, Yuntian Jiang, and Yinyu Ye. DRSOM: A Dimension Reduced Second-Order Method, 2022.
- [178] Lei-Hong Zhang, Chungen Shen, and Ren-Cang Li. On the generalized Lanczos trust-region method. **SIAM Journal on Optimization**, 27(3):2110–2142, 2017.
- [179] Leihong Zhang, Weihong Yang, Chungen Shen, and Jiang Feng. Error bounds of Lanczos approach for trust-region subproblem. **Frontiers of Mathematics in China**, 13:459–481, 2018.