



中国科学院大学
University of Chinese Academy of Sciences

博士学位论文

计算材料科学中的优化问题、理论与算法

作者姓名: 胡雨宽

指导教师: 刘歆 研究员

中国科学院数学与系统科学研究院

学位类别: 理学博士

学科专业: 计算数学

培养单位: 中国科学院数学与系统科学研究院

2024 年 6 月

Optimization Models, Theories, and Methods
in Computational Materials Science

A dissertation submitted to
University of Chinese Academy of Sciences
in partial fulfillment of the requirement
for the degree of
Doctor of Philosophy
in Computational Mathematics
By
HU Yukuan
Supervisor: Professor LIU Xin

Academy of Mathematics and Systems Science
Chinese Academy of Sciences

June, 2024

中国科学院大学

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。承诺除文中已经注明引用的内容外，本论文不包含任何其他个人或集体享有著作权的研究成果，未在以往任何学位申请中全部或部分提交。对本论文所涉及的研究工作做出贡献的其他个人或集体，均已在文中以明确方式标明或致谢。本人完全意识到本声明的法律结果由本人承担。

作者签名：

日 期：

中国科学院大学

学位论文授权使用声明

本人完全了解并同意遵守中国科学院大学有关收集、保存和使用学位论文的规定，即中国科学院大学有权按照学术研究公开原则和保护知识产权的原则，保留并向国家指定或中国科学院指定机构送交学位论文的电子版和印刷版文件，且电子版与印刷版内容应完全相同，允许该论文被检索、查阅和借阅，公布本学位论文的全部或部分内容，可以采用扫描、影印、缩印等复制手段以及其他法律许可的方式保存、汇编本学位论文。

涉密及延迟公开的学位论文在解密或延迟期后适用本声明。

作者签名：

导师签名：

日 期：

日 期：

摘要

计算材料科学通过理论模型和数值模拟,帮助科学家和工程师们理解和预测材料的性能和行为,进而促进材料研发和应用。尽管计算材料科学以计算机科学与材料科学为基础,但其中有许多关键的问题可归结为某种能量的极小化。这一联系为优化这门学科提供了新机遇、新挑战,同时也有望为计算材料科学提供强有力的新方法。立足于优化与计算材料科学的交叉点,本文研究具有特殊结构的优化问题及其理论与算法,并将它们用于两类重要问题——强关联电子体系计算和第一性原理晶体结构弛豫。

我们首先研究一类带广义互补约束优化问题的 ℓ_1 罚函数精确性。此类问题可见于强关联电子体系的计算。由于广义互补约束的存在,此类问题可能不满足约束规范条件,从而 Karush-Kuhn-Tucker 条件可能不再是其局部最优解满足的必要条件。为此,一种常用的解决方案是将广义互补约束以 ℓ_1 形式惩罚到目标函数上,再证明 ℓ_1 罚函数的精确性。我们先给出了此类问题的一个实例,其 ℓ_1 罚函数的精确性无法被已有理论结果覆盖。接着,我们充分借助所考虑问题类的代数与几何结构,证明了 ℓ_1 罚问题的精确性。这些结果为之后优化算法的应用提供了理论基础。

随后,我们考虑具有块状结构的优化问题,并研究不可行非精确的邻近交替线性化极小化 (PALM) 算法的理论性质。前述源于强关联电子体系计算的 ℓ_1 罚问题是此类问题的特例。对于 PALM 算法,已有工作的理论分析需建立在目标函数值序列的某种单调性之上。然而,在许多情形下,人们会倾向或不得不使用不可行方法求解 PALM 算法中的子问题。此时,目标函数值序列的非单调性无法避免,已有理论结果不再适用。为此,我们先为 PALM 算法子问题的求解设计了一个可实现的终止准则。之后,基于子问题满足的误差界与所设计的终止准则,我们巧妙地构造了一个单调下降的代理序列。以此,我们首次在可实现的条件下,证明了不可行非精确 PALM 算法的收敛性质与渐进收敛速度。通过测试问题上的数值实验,我们展示了不可行非精确 PALM 算法的效率优势。此外,我们还将其嵌入了一个瀑布型多重网格优化 (CMGOPT) 框架,通过求解前述 ℓ_1 罚问题,模拟了一维、二维强关联电子体系。我们取得了符合理论预测与物理直观的数值结果,并首次可视化了二维情形下电子位置之间的映射。

接着,我们考虑运输多胞体上的分块矩阵优化问题,并为其设计了完全无需全矩阵的块坐标下降型算法。前述源于强关联电子体系计算的 ℓ_1 罚问题是此类问题的特例。为求解之,已有块坐标下降型算法需要显式存储或计算全矩阵。对于大规模问题的求解,这些算法具有较高的空间与计算复杂度。为此,我们结合最优运输工具与矩阵逐元素随机近似设计了全新的块坐标下降型算法。其中,使用矩阵逐元素随机近似等价于在子问题中增加随机置零约束。这完全免去了全矩阵的存储与计算,使新算法与已有随机块坐标下降型算法有根本的不同。我们为

新算法建立了概率意义下的收敛性质,首次为矩阵逐元素随机近似在分块非凸问题上的应用提供了理论保证.在数值实验中,新算法展现出了更好的计算标度.我们还将新算法嵌入了 CMGOPT 框架,通过求解前述 ℓ_1 罚问题,成功模拟了二维、三维强关联电子体系.得益于新算法的低标度,我们首次可视化了三维情形电子位置之间的映射.

最后,我们考虑行列式约束优化问题,并为其设计了基于矩阵缩放的投影梯度下降 (PGD) 算法.此类问题与第一性原理固定晶格体积晶体结构弛豫紧密相关.后者在材料结构物态方程计算中具有重要应用.相较于正交投影,矩阵缩放可以充分利用行列式的性质,将迭代点显式地拉回至可行域.然而,目前暂无已有工作在使用矩阵缩放保持迭代点可行性时,分析算法的理论性质.为此,我们以负梯度在可行域切锥上的正交投影为搜索方向,结合矩阵缩放,设计了 PGD 算法.借助可行域切锥的代数结构与非单调线搜索,我们证明了 PGD 算法的收敛性质.在数值实验中,我们将 PGD 算法推广至求解第一性原理固定晶格体积晶体结构弛豫问题.在含有 223 个来自不同类别结构的基准算例集上,相较于常用材料模拟软件中实现的非线性共轭梯度算法与拟牛顿算法,PGD 算法展现出了显著且普遍的效率与鲁棒性优势.我们还将 PGD 算法用于计算高熵合金 AlCoCrFeNi 的物态方程,通过与已有实验测定数据对比验证了计算结果的正确性.

关键词: 强关联电子体系计算, 第一性原理晶体结构弛豫, 罚函数, 块坐标下降型算法, 投影梯度下降算法

Abstract

Through theoretical models and numerical simulations, computational materials science aids scientists and engineers in understanding and predicting the properties and behaviors of materials, thereby facilitating materials design and applications. Despite its root in computer science and materials science, numerous critical issues in this field can be attributed to the minimization of some energy. This connection not only presents new opportunities and challenges for the optimization community but also holds promise for novel methods in the realm of computational materials science. Centered at the intersection of optimization and computational materials science, this dissertation endeavors to investigate optimization problems with special structures as well as their theories and numerical methods, and apply them to two important topics—calculations of strongly correlated electron systems and *ab initio* crystal structure relaxation.

Firstly, we examine the exactness of ℓ_1 penalty functions for a class of mathematical programs with generalized complementarity constraints. This class of problems finds applications in the calculations of strongly correlated electron systems. Due to the presence of generalized complementarity constraints, problems in this class may not satisfy constraint qualifications, rendering Karush-Kuhn-Tucker conditions no longer necessary for their local minimizers. A conventional treatment for this deficiency is to penalize the generalized complementarity constraints in ℓ_1 form and then prove the exactness of ℓ_1 penalty functions. In our work, we first identify an instance from the problem class, where the exactness of ℓ_1 penalty function is not covered by existing theoretical results. By fully exploiting the inherent algebraic and geometric structures, we then establish the exactness for the problem class. Our findings lay the theoretical foundation for the subsequent algorithm applications.

Secondly, we delve into block-structured optimization problems and study the theoretical properties of infeasible inexact proximal alternating linearized minimization (PALM) method. The problem class under consideration encompasses the aforementioned ℓ_1 penalty problem arising from the calculations of strongly correlated electron systems as a special case. For the PALM method, existing theoretical analyses rest on some monotonicity of objective value sequence. However, in many cases, practitioners tend or have to employ infeasible approaches for the subproblems within the PALM method. In such scenarios, the nonmonotonicity of objective value sequence becomes inevitable, rendering the existing theoretical results inapplicable. For this purpose, we first devise an implementable stopping criterion for solving the subproblems within the PALM method. Leveraging the error bounds of subproblems as well as the devised stopping criterion, we then construct a monotonically decreasing surrogate sequence in

a nuanced manner. This enables the first proof of convergence properties and asymptotic convergence rates for the infeasible inexact PALM method under computationally implementable conditions. We demonstrate the efficiency advantage of the infeasible inexact PALM method through numerical experiments on test problems. Moreover, we integrate it into a cascadic multigrid optimization (CMGOPT) framework and simulate one/two-dimensional strongly correlated electron systems by solving the aforementioned ℓ_1 penalty problem. We obtain numerical results that align with both theoretical predictions and physical intuitions and, remarkably, provide the first visualization of mappings between electron positions in two-dimensional contexts.

Thirdly, we explore multi-block matrix optimization problems over transport polytopes and develop block coordinate descent-type methods that dispense the use of full matrices. The problem class under consideration also encompasses the aforementioned ℓ_1 penalty problem arising from the calculations of strongly correlated electron systems as a special case. Existing block coordinate descent-type methods for solving problems in this class require explicit storage or computations of the full matrices. Consequently, they exhibit formidable storage and computational burdens in large-scale contexts. To this end, we develop novel block coordinate descent-type methods by combining tools from the optimal transport field and matrix entrywise random sparsification. Notably, adopting the matrix entrywise random sparsification amounts to introducing random zeroing constraints to subproblems. This feature completely waives the need for storing and computing full matrices, fundamentally distinguishing the new methods from existing ones. We rigorously establish the convergence properties of the new methods in a probabilistic sense, providing the first theoretical guarantees for the application of matrix entrywise random sparsification to multi-block nonconvex problems. The new methods display improved computational scaling in numerical experiments. Furthermore, we embed the new methods into the CMGOPT framework and successfully simulate two/three-dimensional strongly correlated electron systems by solving the aforementioned ℓ_1 penalty problem. The improved scaling of the new methods enables the first visualization of mappings between electron positions in three-dimensional contexts.

Lastly, we address determinant-constrained optimization problem and design a projected gradient descent (PGD) method based on matrix scaling. The problem class under consideration is closely related to *ab initio* crystal structure relaxation under a fixed unit cell volume, which holds particular significance in computing the equations of state for given material structures. Compared with the orthogonal projection, the matrix scaling fully exploits the determinant property and pulls iterates back to feasible region via a closed-form formula. Nevertheless, to date, there have been no works dedicated to analyzing the theoretical properties of algorithms employing the matrix scaling to maintain the feasibility of iterates. In this study, we develop the PGD method, which takes the

orthogonal projections of negative gradients onto the tangent cones of feasible region as search directions and adopts the matrix scaling. By leveraging the algebraic structure of tangent cone and incorporating nonmonotone line search, we establish the convergence properties of the PGD method. In numerical experiments, we extend the PGD method to tackle the *ab initio* crystal structure relaxation under a fixed unit cell volume. Across a benchmark test set comprising 223 structures from different categories, the PGD method demonstrates prominent and universal advantages in terms of both efficiency and robustness over the nonlinear conjugate gradient method and quasi-Newton method implemented in popular materials simulation software. In addition, we apply the PGD method to compute the equations of state for the high-entropy alloy AlCoCrFeNi. Existing experimental data corroborate the validity of our numerical results.

Key Words: Calculations of strongly correlated electron systems, *ab initio* crystal structure relaxation, penalty function, block coordinate descent-type methods, projected gradient descent method

目 录

第 1 章 引言	1
1.1 研究背景	1
1.2 数学优化的基本概念	2
1.3 电子结构计算	7
1.3.1 电子总能极小与密度泛函理论	7
1.3.2 Kohn-Sham 密度泛函理论与算法	9
1.3.3 严格关联电子密度泛函理论与算法	11
1.4 第一性原理晶体结构弛豫	17
1.4.1 晶体结构的数学描述	17
1.4.2 第一性原理晶体结构弛豫问题与算法	18
1.5 本文主要内容	21
第 2 章 一类带广义互补约束优化问题的 ℓ_1 罚函数精确性	23
2.1 问题描述与研究现状	23
2.1.1 问题描述	23
2.1.2 研究现状	26
2.1.3 本章主要内容	26
2.2 已有理论结果不适用之例	26
2.3 ℓ_1 罚函数精确性的证明	29
2.4 本章小结	32
第 3 章 求解具有块状结构优化问题的不可行非精确邻近交替线性化极小化算法	35
3.1 问题描述与研究现状	35
3.1.1 问题描述	35
3.1.2 邻近交替线性化极小化算法	35
3.1.3 研究现状	35
3.1.4 本章主要内容	38
3.2 算法描述	38
3.3 收敛性分析	41
3.3.1 全局依子列收敛性	42
3.3.2 全局依点列收敛性	47
3.4 渐进收敛速度分析	50

3.4.1 误差控制序列指数下降情形	52
3.4.2 误差控制序列次线性下降情形	55
3.5 数值实验	57
3.5.1 算法比较	58
3.5.2 强关联电子体系计算	64
3.6 本章小结	76
第 4 章 求解运输多胞体上分块矩阵优化问题的块坐标下降型算 法	79
4.1 问题描述与研究现状	79
4.1.1 问题描述	79
4.1.2 研究现状	79
4.1.3 本章主要内容	80
4.2 算法设计	80
4.2.1 熵正则交替线性化极小化算法	81
4.2.2 熵正则交替线性化极小化算法的采样版本	82
4.2.3 基于 Kullback-Leibler 散度的交替线性化极小化算法	84
4.2.4 基于 Kullback-Leibler 散度的交替线性化极小化算法的采样版本	86
4.2.5 单步计算代价比较	87
4.3 收敛性分析	87
4.4 数值实验	97
4.4.1 待模拟强关联电子体系	98
4.4.2 实验设置	99
4.4.3 算法比较	100
4.4.4 二维、三维强关联电子体系计算	105
4.4.5 算法标度测试	107
4.5 本章小结	109
第 5 章 求解行列式约束优化问题的投影梯度下降算法	113
5.1 问题描述与研究现状	113
5.1.1 本章主要内容	113
5.2 算法设计	114
5.3 收敛性分析	115
5.4 数值实验	119
5.4.1 求解固定晶格体积晶体结构弛豫问题的投影梯度下降算法	120
5.4.2 实验设置	122
5.4.3 基准算例集上的测试	123

5.4.4 高熵合金物态方程计算 ······	124
5.5 本章小结 ······	129
第 6 章 总结与展望 ······	131
参考文献 ······	135
附录一 带 Coulomb 费用多边际最优运输问题的离散化 ······	155
致谢 ······	157
作者简历及攻读学位期间发表的学术论文与其他相关学术成果 ·	159

图目录

图 1.1 晶体结构示意图, 其中红色(大)球与蓝色(小)球代表不同类型原子. 左图: 单胞. 右图: 晶格	18
图 1.2 二维情形下行列式约束 $\det(A) = 1$ 对应的可行域、正交投影与缩放算子示意图. 其中, 横纵坐标为 A 的两个奇异值, 蓝色实线代表可行域, 红色圆点代表待投影点, 黑色方形代表其到可行域上的正交投影, 红色五角星代表缩放算子作用在该点上的结果	19
图 2.1 一般 LPCC、MPCC、MPGCC 与问题 (2.5) 之间的关系. 带实线边界的蓝色椭圆表示 MPGCC, 带虚线边界的蓝色圆盘表示 MPCC, 带点线边界的蓝色圆盘表示 LPCC, 带点划线边界的红色椭圆表示问题 (2.5)	25
图 2.2 当 $K = 6$ 时, 问题 (2.11) 的一个最优解 (Y_1^*, Y_2^*) . 深紫色方块代表值为 1 的元素, 蓝色方块代表值为 0 的元素	27
图 2.3 当 $K = 6$ 时, 构造的 ℓ_1 罚问题可行点 $(Y_1(\varepsilon), Y_2(\varepsilon))$. 深紫色方块代表值为 $1 - \varepsilon$ 的元素, 浅紫色方块代表值为 ε 的元素, 蓝色方块代表值为 0 的元素	28
图 3.1 当误差控制序列取做 $\varepsilon_k = \bar{\varepsilon}/(k+1)^l$ ($k \geq 0, l > 1$) 时, PALM-I 算法的渐进收敛速度负指数. 其中, x 轴为 F 在收敛点处的 Łojasiewicz 指数, y 轴为 l 的取值, z 轴为 PALM-I 算法的渐进收敛速度负指数, 虚线表示 θ 与 l 的临界关系 $\theta = l/(2l-1)$	59
图 3.2 从 100 个随机初始点出发, PALM-E 算法与 PALM-I 算法在求解问题 (2.8) 时的平均相对 KKT 违反度与平均相对目标函数值误差随迭代过程的变化. 其中蓝色实线与红色点划线分别表示 PALM-E 算法与 PALM-I 算法的结果	60
图 3.3 从最优解附近的 100 个随机初始点出发, PALM-E 算法与 PALM-I 算法在求解问题 (2.8) 时的平均相对 KKT 违反度与平均相对目标函数值误差随迭代过程的变化. 其中蓝色实线与红色点划线分别表示 PALM-E 算法与 PALM-I 算法的结果	61
图 3.4 从 100 个随机初始点出发, PALM-E 算法、PALM-F 算法与 PALM-I 算法在求解问题 (3.30) 时的平均相对 KKT 违反度与平均相对目标函数值误差随迭代过程的变化, 其中蓝色实线表示 PALM-E 算法的结果, 红色点划线表示 PALM-I 算法的结果, 绿色划线表示 PALM-F 算法的结果	63
图 3.5 表 3.2 所列单电子密度的可视化	65
图 3.6 一维一致加密情形插值算子示意图. 红色方块代表在网格 $\mathcal{T}^{(\ell)}$ 上, 解的元素 $y_{i,34}^{(\ell,\star)}$ 为正. 蓝色方块代表在插值算子 $\mathcal{I}_\ell^{\ell+1}$ 的作用下, 在网格 $\mathcal{T}^{(\ell+1)}$ 上, 初始点的元素 $y_{i,57}^{(\ell+1,0)}, y_{i,67}^{(\ell+1,0)}, y_{i,58}^{(\ell+1,0)}, y_{i,68}^{(\ell+1,0)}$ 为正 ..	67

图 3.7 二维一致加密情形插值算子示意图, 其中网格 $\mathcal{T}^{(\ell)}$ 与 $\mathcal{T}^{(\ell+1)}$ 的离散单元数分别为 7×7 与 14×14 . 离散单元在网格中的二维坐标与 Y_i 的行序号或列序号一一对应. 例如, $\mathcal{T}^{(\ell)}$ 中坐标为 $(2, 4)$ 的单元对应 Y_i 的第 11 行或列. 红色方框代表在网格 $\mathcal{T}^{(\ell)}$ 上, 解的元素 $y_{i,5,11}^{(\ell,\star)}$ 为正. 蓝色方块代表在插值算子 $\mathcal{I}_\ell^{\ell+1}$ 的作用下, 在网格 $\mathcal{T}^{(\ell+1)}$ 上, 初始点的元素 $y_{i,j,j'}$ 为正, 其中 $j \in \{9, 10, 23, 24\}$, $j' \in \{35, 36, 49, 50\}$	68
图 3.8 在一维三电子体系上, CMGOPT 框架在第 $\ell = 0 \sim 3$ 层网格上输出的 $\hat{\lambda}$ 以及精确 SCE 势. 从左至右分别为体系一、二、三的 $\hat{\lambda}$ 与 u_{SCE} . 其中蓝色点线、点划线、虚线、实线分别表示 CMGOPT 框架在第 $\ell = 0 \sim 3$ 层网格上输出的 $\hat{\lambda}$, 红色点线表示精确 SCE 势	71
图 3.9 在一维三电子体系上, CMGOPT 框架在第 $\ell = 0, 2, 4, 6$ 层网格上输出的近似映射 $\{\hat{\mathcal{T}}_i\}_{i=2}^N$. 从左至右分别为体系一、二、三的近似映射. 其中蓝点与红点分别表示单元重心在 $\hat{\mathcal{T}}_2$ 与 $\hat{\mathcal{T}}_3$ 下的像	72
图 3.10 在一维七电子体系上, CMGOPT 框架在第 $\ell = 0 \sim 3$ 层网格上输出的 $\hat{\lambda}$ 以及精确 SCE 势. 从左至右分别为体系四、五、六的 $\hat{\lambda}$ 与 u_{SCE} . 其中蓝色点线、点划线、虚线、实线分别表示 CMGOPT 框架在第 $\ell = 0 \sim 3$ 层网格上输出的 $\hat{\lambda}$, 红色点线表示精确 SCE 势	73
图 3.11 在一维七电子体系上, CMGOPT 框架在第 $\ell = 0, 2, 4, 6$ 层网格上输出的近似映射 $\{\hat{\mathcal{T}}_i\}_{i=2}^N$. 从左至右分别为体系四、五、六的近似映射. 其中不同颜色的点分别表示单元重心在 $\{\hat{\mathcal{T}}_i\}_{i=2}^N$ 下的像	74
图 3.12 在二维三电子体系上, CMGOPT 框架在第 $\ell = 0 \sim 3$ 层网格上输出的 $\hat{\lambda}$. (a) 体系七. (b) 体系八	76
图 3.13 在二维三电子体系上, 初始离散网格 (第一行) 与 CMGOPT 框架在第 $\ell = 3$ 层网格上输出的近似映射 $\{\hat{\mathcal{T}}_i\}_{i=2}^N$ (余下三行) 切片. 从左至右分别为体系七与八的初始离散网格与近似映射. 其中灰色、绿色与蓝色部分分别表示原像 $\tilde{\Omega} \subseteq \Omega$ 及其在 $\hat{\mathcal{T}}_2$ 、 $\hat{\mathcal{T}}_3$ 下的像	77
图 3.14 在二维三电子体系上, 解 $\{Y_i\}_{i=2}^N$ 的 $\{\text{nnz}_{ij}\}_{ij}$ 与 $\{d_{ij}\}_{ij}$ 联合频数百分比分布图. (a) 体系七. (b) 体系八	78
图 4.1 表 4.2 所列单电子密度的可视化. 对于三维体系, 我们仅展示单电子密度大于 0.01 的区域	99
图 4.2 在不同 (K, \hat{t}) 下, S-KLALM 算法模拟体系一 (等质量剖分) 时的平均 err_obj、err_sce 和 T. 其中, 带有右向三角形标记的蓝色、绿色和紫色虚线分别表示 S-KLALM 算法在 $\hat{t} = 0, 5, 10$ 时所取得的结果. 左图: err_obj. 中图: err_sce. 右图: T	101

图 4.3 不同 σ 下, ERALM 算法、KLALM 算法、S-ERALM 算法与 S-KLALM 算法模拟体系一(等质量剖分, $K = 90$)时平均 err_obj、err_sce 和 T. 其中, 带有三角型标记的红色实线和虚线分别表示 ERALM 算法和 S-ERALM 算法的结果, 带有右向三角形标记的蓝色实线和虚线分别表示 KLALM 算法和 S-KLALM 算法的结果. 左图: err_obj. 中图: err_sce. 右图: T	102
图 4.4 在不同 K 下, S-ERALM 算法与 S-KLALM 算法 ($\sigma = 1$) 模拟体系一(等质量剖分)时的平均 err_obj、err_sce 和 T. 其中, 带有三角形标记的红色虚线表示 S-ERALM 算法的结果, 带有右向三角形标记的蓝色虚线表示 S-KLALM 算法的结果. 左图: err_obj. 中图: err_sce. 右图: T	103
图 4.5 在不同 K 下, KLALM 算法与 S-KLALM 算法在模拟体系一(等质量剖分)时的平均 err_obj 随 CPU 时间的收敛曲线. 其中, 蓝色实线和虚线分别表示 KLALM 算法与 S-KLALM 算法的结果, T_{inter} 代表两个算法的曲线最后一次相交时的 CPU 时间	104
图 4.6 在体系一(等质量剖分)上 T_{inter} 和 K 的三次多项式拟合结果. 拟合得到的多项式为 $9.09 \times 10^{-7}K^3 - 1.86 \times 10^{-3}K^2 + 1.00K - 106.38$. 其中, 蓝色五角星为实验中得到的 T_{inter} , 蓝色点划线表示拟合的三次关系	105
图 4.7 在不同 K 下, PALM-I 算法、KLALM 算法与 S-KLALM 算法在模拟一维三电子体系(等质量剖分)时的平均 err_obj、err_sce 和 T. 其中, 带有方形标记的橙色实线代表 PALM-I 算法的结果, 带有右向三角形标记的蓝色实线和虚线分别代表 KLALM 算法和 S-KLALM 算法的结果. 由左至右: err_obj、err_sce 和 T. (a) 体系一. (b) 体系二	106
图 4.8 在不同 K 下, PALM-I 算法、KLALM 算法与 S-KLALM 算法在模拟一维七电子体系(等质量剖分)时的平均 err_obj、err_sce 和 T. 其中, 带有方形标记的橙色实线代表 PALM-I 算法的结果, 带有右向三角形标记的蓝色实线和虚线分别代表 KLALM 算法和 S-KLALM 算法的结果. 由左至右: err_obj、err_sce 和 T. (a) 体系三. (b) 体系四	107
图 4.9 在二维、三维体系上, S-KLALM-CMG 算法在每层网格上输出的 $\hat{\lambda}$. (a) 体系五. (b) 体系六. (c) 体系七. (d) 体系八	109
图 4.10 在二维、三维体系上, S-KLALM-CMG 算法在最后一层网格上输出的近似映射 $\{\hat{\mathcal{T}}_i\}_{i=2}^N$ 切片. 其中, 红色部分表示原像 $\tilde{\Omega} \subseteq \Omega$, 其他颜色表示原像在 $\{\hat{\mathcal{T}}_i\}_{i=2}^N$ 下的像. (a) 体系一. (b) 体系二. (c) 体系三. (d) 体系四	110

图 4.11 在不同 K 和 N 下, KLALM 算法与 S-KLALM 算法在模拟一维余弦型体系(等质量剖分)时的平均 err_obj、err_sce 和 T. 其中, 蓝色实线与虚线分别表示 KLALM 算法与 S-KLALM 算法的结果, 粉色实线与虚线分别代表所拟合的 KLALM 算法与 S-KLALM 算法的 T 与 K 或 T 与 N 之间的关系. 左图: err_obj. 中图: err_sce. 右图: T. (a) 对于 K 的标度测试. 对于 KLALM 算法, $T \sim K^{2.6}$. 对于 S-KLALM 算法, $T \sim K^{2.2}$. (b) 对于 N 的标度测试. 对于 KLALM 算法, $T \sim N^{1.3}$. 对于 S-KLALM 算法, $T \sim N^{1.1}$	111
图 5.1 CG 算法、QN 算法与 PGD 算法在基准算例集上的性能剖面. 为比较公平性, 我们 (1) 剔除三个算法收敛到的最大与最小能量之差超过每原子 3 meV 的结构; (2) 保留至少有一个算法无法在 #KS 超过 1000 之前终止的结构, 并将其 #KS 与 CPU 置为无穷大. 最终, 用于绘制性能剖面的结构数为 204. 红色点线、蓝色虚线与黑色实线分别表示 CG 算法、QN 算法与 PGD 算法的结果. (a) #KS 的性能剖面. (b) CPU 的性能剖面	124
图 5.2 PGD 算法对 CG 算法分结构类别平均加速比. 为比较公平性, 我们仅保留二者均正常收敛且收敛到的能量差绝对值不超过每原子 3 meV 的结构. 此外, 我们将结构数少于 5 的类别统一并入“其他”中. 左图: 分结构类别平均 #KS 加速比. 右图: 分结构类别平均 CPU 加速比	125
图 5.3 由 SAE 方法与整数规划生成的初始原子与磁构型. (a) 由 SAE 方法生成的初始原子构型. (b) 由整数规划选取的初始磁构型中的 64 个上旋位点(红色箭头). (c) 由整数规划选取的初始磁构型中的 64 个下旋位点(蓝色箭头). (d) 初始磁构型中磁元素的径向分布函数	127
图 5.4 AlCoCrFeNi 的 FM 态与 PM 态的能量-体积关系.“计算”表示第一性原理计算结果, “拟合”表示拟合的静态 BM3 物态方程, 红色圆点与实线表示 FM 态的结果, 黑色方形与点线表示 PM 态的结果	128
图 5.5 AlCoCrFeNi 在 300 K 等温的体积-压强关系. 其中, “实验测定”表示使用 X 射线衍射与金刚石压砧实验测定的结果, “PGD + MMFP”表示由修正平均场势方法基于 PGD 算法的弛豫结果给出的估计	129
图 5.6 AlCoCrFeNi 的 FM 态弛豫后的轴长比-体积关系. 其中, “ c/a ”与“ b/a ”分别表示 c 轴和 a 轴长度之比与 b 轴与 a 轴长度之比. 红色实线与点线表示 CG 算法的结果, 黑色实线与点线表示 PGD 算法的结果. 对于 CG 算法无法正常收敛的体积点, 我们使用其最后迭代构型的轴长比	130
图 6.1 本文应用领域、研究内容、主要贡献及它们的关系	131

表目录

表 1.1 MMOT (1.18) 的已有低维转化模型与数值方法	17
表 3.1 PALM-I 算法在不同情形下的收敛速度	58

表 3.2 待模拟的一维和二维强关联电子体系. 第二列为未归一化的单电子密度 ρ , 第三列为截断区域 Ω , 第四列为体系所含电子数 N	64
表 3.3 不同问题规模 K 对应的 β	69
表 3.4 在一维三电子体系上, CMGOPT 框架在每层网格上输出的目标函数值与映射误差. 其中“-”表示初始网格上没有初始点, “err_map _s ”与“err_map _e ”分别表示初始点与 PALM-I 算法输出解的映射误差	70
表 3.5 在一维七电子体系上, CMGOPT 框架在每层网格上输出的目标函数值与映射误差. 其中“-”表示初始网格没有初始点, “err_map _s ”与“err_map _e ”分别表示初始点与 PALM-I 算法输出解的映射误差	73
表 3.6 在一维三电子体系 ($K = 768$) 上, 从随机初始点与由插值算子构造的初始点出发, PALM-I 算法收敛到的目标函数值与所需运行时间. 随机初始化的结果取 10 次模拟的平均值	75
表 3.7 在二维三电子体系上, CMGOPT 框架在每层网格上输出的目标函数值	75
表 4.1 四个新算法单步计算代价的比较	89
表 4.2 待模拟的一维、二维和三维强关联电子体系. 第二列为未归一化的单电子密度 ρ , 第三列为截断区域 Ω , 第四列为体系所含电子数 N ..	98
表 4.3 在不同采样概率下 S-ERALM 算法模拟体系一 ($K = 90$, 等质量剖分) 时的平均 err_obj、err_sce 和 T. 其中, 在“采样概率”列, “随机”表示随机生成采样概率, 其余表示使用不同 γ 的重要性采样概率	101
表 4.4 在二维、三维体系上, S-KLALM-CMG 算法在每层网格上输出的目标函数值	108
表 5.1 第一性原理固定晶格体积晶体结构弛豫算法基准测试算例集信息	123
表 5.2 在 Wigner-Seitz 半径范围 [2.45 Å, 2.70 Å] 内按等间距选取的 18 个晶格体积点	127
表 5.3 拟合的静态 BM3 物态方程中的参数. 其中“FM/PM (PGD)”表示使用 PGD 算法弛豫的结果, “FM/PM (CG)”表示使用 CG 算法弛豫的结果, 其余为已有工作的结果	128

符号列表

集合

如不特别说明, 我们用花体大写字母表示集合. 以下是一些具有特殊含义的集合及其符号:

\mathbb{C}	复数域
\mathbb{N}	自然数集合
\mathbb{R}	实数域
\mathbb{R}_+	非负实数集合
\mathbb{R}_{++}	正实数集合
\mathbb{R}^n	全体 n 维实向量构成的欧氏空间
$\mathbb{R}^{m \times n}$	全体 $m \times n$ 维实矩阵构成的欧氏空间
\mathbb{Z}	整数环
\mathcal{A}^c	集合 \mathcal{A} 的补集
$B_r(\mathbf{x})$	以 \mathbf{x} 为中心、 r 为半径的球
\mathcal{F}	优化问题的可行域
$\mathcal{H}^1(\mathcal{A}; \mathcal{B})$	全体由 \mathcal{A} 到 \mathcal{B} 本身与广义导数 L^2 可积的函数构成的 Sobolev 空间
$\mathcal{P}(\mathcal{A})$	集合 \mathcal{A} 上全体概率测度构成的空间
$\mathcal{P}_{\text{sym}}(\mathcal{A})$	集合 \mathcal{A} 上全体对称概率测度构成的空间
\mathcal{P}_N	集合 $\{1, \dots, N\}$ 上的置换群
$\text{rbd}(\mathcal{A})$	集合 \mathcal{A} 的相对边界

标量、向量与矩阵

如不特别说明, 我们用常规小写字母表示标量, 粗体小写字母表示向量, 用常规大写字母表示矩阵. 以下是一些具有特殊含义的标量、向量与矩阵及其符号:

$\delta_{\mathbf{a}, \mathbf{b}}$	Kronecker 记号
x^*	标量 x 的复共轭
$\mathbf{1}_n$	元素全为 1 的 n 维向量
\mathbf{x}^\top	向量 \mathbf{x} 的转置
x_i	向量 \mathbf{x} 的第 i 个分量
\mathbf{x}_i	向量 \mathbf{x} 的第 i 个子块
$\mathbf{x}^{(k)}$	向量 \mathbf{x} 的第 k 次迭代
$\mathbf{x}^{(k,j)}$	向量 \mathbf{x} 在第 $(k+1)$ 次迭代中的第 j 次子迭代

I_n	$n \times n$ 维单位矩阵
X^\top	矩阵 X 的转置
X^{-1}	矩阵 X 的逆
X^\dagger	矩阵 X 的 Moore-Penrose 伪逆
$X_{\mathcal{I}}$	矩阵 X 中指标在集合 \mathcal{I} 中的元素
x_{ij}	矩阵 X 在第 i 行第 j 列的元素
$X_{\cdot,j}$	矩阵 X 的第 j 列
X_i	矩阵 X 的第 i 个子块
$x_{i,jk}$	矩阵 X 第 i 个子块在第 j 行第 k 列的元素
$X^{(k)}$	矩阵 X 的第 k 次迭代
$X^{(k,j)}$	矩阵 X 在第 $(k+1)$ 次迭代中的第 j 次子迭代

函数与算子

如不特别说明, 我们用手写体大写字母表示算子. 以下是一些具有特殊含义的函数与算子及其符号:

\times, \times	集合的 Cartesian 乘积
$\langle \cdot, \cdot \rangle$	标准向量或矩阵内积
\odot	向量或矩阵逐元素相乘
\oslash	向量或矩阵逐元素相除
$\lfloor \cdot \rfloor$	向下取整函数
$ \cdot $	有限集合元素个数、集合(外)测度或标量的模
$\ \cdot\ $	欧式向量范数或矩阵范数
$\ \cdot\ _2$	矩阵 ℓ_2 -范数
$\ \cdot\ _\infty$	矩阵逐元素 ℓ_∞ -范数
$\ \cdot\ _{2,\infty}$	矩阵 $\ell_{2,\infty}$ -范数
$\kappa(\cdot)$	矩阵谱条件数
$\delta(\cdot)$	Dirac delta 函数
$\delta_{\mathcal{A}}(\cdot)$	集合 \mathcal{A} 的指示函数
$\delta_{\mathbf{a}}(\cdot)$	在 \mathbf{a} 处有单位质量的 Dirac 测度
$\mathcal{P}_{\mathcal{A}}(\cdot)$	到集合 \mathcal{A} 上的正交投影算子
$\mathcal{P}_{\mathcal{V}}(\cdot)$	矩阵缩放算子
$\text{Prob}(\cdot)$	随机事件发生的概率
$E(\cdot)$	随机变量的期望
$\text{Var}(\cdot)$	随机变量的方差

$\text{Diag}(\mathbf{x})$	以向量 \mathbf{x} 中的元素为对角元构造对角矩阵
$\text{diag}(X)$	以矩阵 X 的对角元构造列向量
$\det(X)$	矩阵 X 的行列式
$\text{supp}(X)$	矩阵 X 的支撑指标集
$\text{Tr}(X)$	矩阵 X 的迹
$\nabla f(\mathbf{x})$	函数 f 在 \mathbf{x} 处的梯度
$\nabla_{\mathbf{x}_i} f(\mathbf{x})$	函数 f 在 \mathbf{x} 处对第 i 个子块的梯度
$\partial f(\mathbf{x})$	函数 f 在 \mathbf{x} 处的次微分
$\Delta f(\mathbf{x})$	函数 f 在 \mathbf{x} 处的 Laplacian
$\text{dist}(\mathbf{x}, \mathcal{A})$	向量 \mathbf{x} 到集合 \mathcal{A} 的欧式距离

与电子结构计算相关的符号

E_0	基态能量
E_{KS}	Kohn-Sham 密度泛函理论基态能量
E_{SCE}	严格关联电子密度泛函理论基态能量
E_{KSSCE}	Kohn-Sham–严格关联电子密度泛函理论基态能量
K	离散规模
N	电子个数
\mathbf{r}_i	第 i 个电子的位置
c_{ee}	Coulomb 势
v_{ext}	外势
Ψ	波函数
ρ	单电子密度
E_{kd}	动能去相关能量泛函
E_{xc}	交换–关联能泛函
F_{LL}	Levy-Lieb 泛函
J	静电自能泛函
T_{e}	动能泛函
T_s	Kohn-Sham 动能泛函
V_{ee}	电子–电子相互作用能泛函
V_{ne}	电子–原子核相互作用能泛函
$V_{\text{ee,SCE}}$	严格关联电子能量泛函
\mathcal{H}	Hamiltonian 算子
\mathcal{H}_{KS}	Kohn-Sham Hamiltonian 算子

与晶体结构弛豫相关的符号

M	原子个数
\mathbf{a}_i	第 i 个晶格基矢
R_i	第 i 个原子的位置
A	晶格基矢矩阵
R	原子位置矩阵
F_{atom}	原子受力
F_{latt}	晶格受力
Σ	晶格应力
Σ_{dev}	晶格偏应力
E	能量泛函

其他

\propto	正比例于
s. t.	“subject to”的缩写
\AA	长度单位“埃”; $1 \text{\AA} = 10^{-10}$ 米
eV	能量单位“电子伏特”; $1 \text{ eV} \approx 1.6022 \times 10^{-19}$ 焦尔

缩写

ADMM	交替方向乘子法
BCG	分块条件梯度
CG	非线性共轭梯度
CMGOPT	瀑布型多重网格优化
DFT	密度泛函理论
HP	混合投影
KKT	Karush-Kuhn-Tucker
KS	Kohn-Sham
KSDFT	Kohn-Sham 密度泛函理论
LL	Levy-Lieb
LPCC	互补约束线性规划
MMOT	带库伦费用的多边际最优运输问题
MPCC	互补约束优化问题
MPGCC	带广义互补约束的优化问题
PALM	邻近交替线性化极小化

PALM-E	精确邻近交替线性化极小化
PALM-F	可行非精确邻近交替线性化极小化
PALM-I	不可行非精确邻近交替线性化极小化
PGD	投影梯度下降
QE	Quantum ESPRESSO
QN	拟牛顿
SCE	严格关联电子
SCEDFT	严格关联电子密度泛函理论
VASP	Vienna <i>Ab initio</i> Simulation Package

第1章 引言

1.1 研究背景

优化 (optimization) 是计算数学和运筹学的交叉学科^[1], 主要研究如何在无任何限制条件或一定限制条件下, 选取适当的参数或决策方案, 使得目标在某种意义上达到最优. 它与工程、管理、计算机科学等学科紧密相关, 在现代社会中扮演着重要的角色, 已被广泛应用于科学与工程计算、数据科学、机器学习、人工智能、图像和信号处理、金融和经济、管理科学等领域. 求解优化问题一般需要经历三个步骤: 为优化问题建立恰当的数学优化模型, 使用或设计优化算法数值求解之, 验证或评估解的质量或合理性. 数学上, 我们可以研究数学优化模型的理论性质, 分析优化算法的收敛性、收敛速度、复杂度等.

计算材料科学 (computational materials science) 是一门综合性学科, 旨在通过理论模型和计算机模拟, 理解和预测材料的性能和行为^[2]. 它将计算机科学、材料科学、物理学、数学等学科融合在一起, 帮助科学家和工程师们探索材料的原子结构、能带结构、力学性质、热力学性质等方面的信息, 进而促进材料研发和应用的过程. 计算材料科学关注的对象具有多尺度的特点, 从微观纳米尺度到宏观尺度. 不同尺度下的研究对象需使用不同的数学或物理模型刻画. 这些模型往往呈现非线性、高维数、不确定性等特征, 为其数值求解带来了巨大的困难.

实际上, 计算材料科学中的许多问题可归结为某种能量的极小化. 这一联系为优化提供了新机遇、新挑战, 同时也有望为计算材料科学提供强有力的新方法. 立足于优化与计算材料科学的交叉点, 本文研究具有特殊结构的优化问题及其理论与算法, 并将它们用于两类重要问题——**强关联电子体系 (strongly correlated electron systems)** 计算和**第一性原理晶体结构弛豫 (ab initio crystal structure relaxation)**.

电子结构计算通过求解量子力学 Schrödinger 方程^[3] 的近似数学模型, 帮助人们理解物质机理、理解与预测材料性质. 强关联电子体系在高温超导^[4,5]、磁性材料^[6,7]、热电材料^[8] 等领域具有重要的应用. 针对这类体系, 广泛使用的数学模型 (如 Kohn-Sham 密度泛函理论^[9]) 在刻画电子之间的相互作用时采用了较为粗糙的近似, 因此有着较大的系统误差, 无法较好地描述一些物理现象^[10,11]. 而显式地刻画电子之间相互作用则需直面模型的维数灾难 (curse of dimensionality), 即搜索空间的维数随电子个数增加指数上升. 因此, **为强关联电子体系的模拟提供高效、可扩展且有理论保证的数值算法是目前电子结构计算领域的前沿热点.**

第一性原理晶体结构弛豫以电子结构计算为基础, 通过优化原子层面的自由度搜寻结构的平衡构型. 它是结构物态方程计算^[12,13]、晶体结构预测^[14–16]、材料高通量设计^[17,18] 等应用的重要组成部分, 但同时也是它们的效率瓶颈. 一方面, 已有弛豫算法需要经历漫长的迭代过程, 而在每一次更新构型后均需通过昂贵的电子结构计算获取能量等信息. 另一方面, 已有弛豫算法缺乏收敛性保证, 在

实际使用中频繁地非正常终止, 进而造成大量人力与计算资源的浪费. 因此, **为第一性原理晶体结构弛豫设计高效且具有理论保证的数值算法对实际应用意义重大.**

在本章接下来的内容中, 我们首先在第 1.2 节简要介绍数学优化的基本概念, 随后在第 1.3 与 1.4 节分别介绍电子结构计算与第一性原理晶体结构弛豫的基本模型和计算方法, 最后在第 1.5 节概括本文主要内容.

1.2 数学优化的基本概念

本节介绍数学优化的基本概念, 包括优化问题的基本形式、优化问题最优解与稳定点的定义、优化算法的终止准则与理论性质.

我们考虑的数学优化问题具有如下基本形式:

$$\min_{\mathbf{x}} f(\mathbf{x}), \text{ s. t. } \mathbf{x} \in \mathcal{F}, \quad (1.1)$$

其中 $\mathbf{x} \in \mathbb{R}^n$ 是变量 (variable), $n \in \mathbb{N}$ 为变量维数, $f : \mathbb{R}^n \rightarrow \mathbb{R}$ 是目标函数 (objective function), $\mathcal{F} \subseteq \mathbb{R}^n$ 为可行域 (feasible region). 我们将可行域中包含的点称为可行解或可行点 (feasible point). 若要在可行域上极大化目标函数 f , 则需将 “min” 替换为 “max”. 由于极大化问题总可以通过给目标函数添加负号变成极小化问题, 因此在后文中我们只针对极小化问题叙述.

若 $\mathcal{F} = \mathbb{R}^n$, 我们称问题 (1.1) 是无约束优化问题 (unconstrained optimization problem), 否则称之为约束优化问题 (constrained optimization problem). 我们所关心的约束优化问题通常具有如下形式:

$$\min_{\mathbf{x}} f(\mathbf{x}), \text{ s. t. } \mathbf{g}(\mathbf{x}) \geq 0, \mathbf{h}(\mathbf{x}) = 0, \quad (1.2)$$

其中 $\mathbf{g} = [g_1, \dots, g_p]^\top : \mathbb{R}^n \rightarrow \mathbb{R}^p$ 为不等式约束函数 ($p \in \mathbb{N}$), $\mathbf{h} = [h_1, \dots, h_q]^\top : \mathbb{R}^n \rightarrow \mathbb{R}^q$ 为等式约束函数 ($q \in \mathbb{N}$). 此时, 可行域具有代数表征 $\mathcal{F} = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{g}(\mathbf{x}) \geq 0, \mathbf{h}(\mathbf{x}) = 0\}$. 若目标函数 f 是凸函数且 \mathcal{F} 是凸集, 我们称问题 (1.1) 是凸优化问题 (convex optimization problem), 否则称之为非凸优化问题 (nonconvex optimization problem). 我们还可以将变量 \mathbf{x} 拆分成若干个变量块 $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_s)$, 其中 $s \in \mathbb{N}$ 为变量块数, $\mathbf{x}_i \in \mathbb{R}^{m_i}$ 代表第 i 个变量块, $m_i \in \mathbb{N}$ 为其维数 ($i = 1, \dots, s$), 满足 $\sum_{i=1}^s m_i = n$. 此时, $f(\mathbf{x}) = f(\mathbf{x}_1, \dots, \mathbf{x}_s)$. 在一些问题上, 若固定其中若干变量块, 目标函数 f (以及可行域 \mathcal{F}) 的结构会得到极大的简化. 此时, 我们称问题 (1.1) 是具有块状结构的优化问题 (block-structured optimization problem).

下面, 我们给出问题 (1.1) 全局最优解 (global minimizer) 和局部最优解 (local minimizer) 的定义.

定义 1.1 (全局最优解与局部最优解).

(1) 我们称 \mathbf{x}^* 是问题 (1.1) 的 (全局) 最优解, 若 $\mathbf{x}^* \in \mathcal{F}$ 且 $f(\mathbf{x}^*) \leq f(\mathbf{x})$ 对所有 $\mathbf{x} \in \mathcal{F}$ 均成立. 此时, 称 $f(\mathbf{x}^*)$ 是问题 (1.1) 的 (全局) 最优值.

(2) 我们称 \mathbf{x}^* 是问题 (1.1) 的局部最优解, 若 $\mathbf{x}^* \in \mathcal{F}$ 且存在 \mathbf{x}^* 的一个邻域 $\mathcal{N}(\mathbf{x}^*)$, 使得 $f(\mathbf{x}^*) \leq f(\mathbf{x})$ 对所有 $\mathbf{x} \in \mathcal{N}(\mathbf{x}^*) \cap \mathcal{F}$ 均成立.

(3) 我们称 \mathbf{x}^* 是问题 (1.1) 的严格局部最优解, 若 $\mathbf{x}^* \in \mathcal{F}$ 且存在 \mathbf{x}^* 的一个邻域 $\mathcal{N}(\mathbf{x}^*)$, 使得 $f(\mathbf{x}^*) < f(\mathbf{x})$ 对所有 $\mathbf{x} \in \mathcal{N}(\mathbf{x}^*) \cap \mathcal{F} \setminus \{\mathbf{x}^*\}$ 均成立.

根据定义, 全局最优解一定是局部最优解, 反之则不一定成立. 对于凸优化问题, 局部最优解也是全局最优解. 除了全局与局部最优解, 我们再引入一阶稳定点 (first-order stationary point) 的概念.

首先考虑光滑情形. 下面是光滑无约束优化问题 (1.1) 局部最优解满足的一阶必要条件.

定理 1.1 (光滑无约束优化问题局部最优解的一阶必要条件^[1,19,20]). 若 \mathbf{x}^* 是光滑无约束优化问题 (1.1) 的局部最优解, 则 $\nabla f(\mathbf{x}^*) = 0$. 我们称满足 $\nabla f(\mathbf{x}) = 0$ 的 \mathbf{x} 为光滑无约束优化问题 (1.1) 的一阶稳定点.

类似地, 我们可以给出光滑约束优化问题 (1.2) 局部最优解满足的一阶必要条件. 在此之前, 我们先定义可行域在一点处的切锥 (tangent cone) 与线性化可行方向锥 (linearized feasible direction cone).

定义 1.2 (切锥^[1,19,20]). 设 $\mathbf{x} \in \mathcal{F}$. 我们称

$$\mathcal{T}_{\mathcal{F}}(\mathbf{x}) := \left\{ \mathbf{d} : \exists \mathcal{F} \supseteq \{\mathbf{x}^{(k)}\} \rightarrow \mathbf{x}, \{t_k\} \rightarrow 0^+, \lim_{k \rightarrow \infty} \frac{\mathbf{x}^{(k)} - \mathbf{x}}{t_k} = \mathbf{d} \right\}$$

为 \mathcal{F} 在 \mathbf{x} 处的切锥.

定义 1.3 (线性化可行方向锥^[1,19,20]). 设 $\mathbf{x} \in \mathcal{F}$. 我们称

$$\mathcal{D}_{\mathcal{F}}(\mathbf{x}) := \left\{ \mathbf{d} : \nabla \mathbf{g}(\mathbf{x})^\top \mathbf{d} \geq 0, \nabla \mathbf{h}(\mathbf{x})^\top \mathbf{d} = 0 \right\}$$

为 \mathcal{F} 在 \mathbf{x} 处的线性化可行方向锥, 其中 $\nabla \mathbf{g}(\mathbf{x}) \in \mathbb{R}^{n \times p}$, $\nabla \mathbf{h}(\mathbf{x}) \in \mathbb{R}^{n \times q}$ 分别是 \mathbf{g} , \mathbf{h} 在 \mathbf{x} 处的 Jacobian.

注. 根据定义, 切锥由可行域与可行点唯一确定, 而线性化可行方向锥则与可行域的代数表征相关. 一般地, 对任一 $\mathbf{x} \in \mathcal{F}$, 我们有 $\mathcal{T}_{\mathcal{F}}(\mathbf{x}) \subseteq \mathcal{D}_{\mathcal{F}}(\mathbf{x})$.

定理 1.2 (光滑约束优化问题局部最优解的一阶必要条件^[1,19,20]). 若 \mathbf{x}^* 是光滑约束优化问题 (1.2) 的局部最优解, 满足 $\mathcal{T}_{\mathcal{F}}(\mathbf{x}^*) = \mathcal{D}_{\mathcal{F}}(\mathbf{x}^*)$, 则存在 $\lambda_g^* \in \mathbb{R}^p$ 与 $\lambda_h^* \in \mathbb{R}^q$, 使得如下条件成立:

$$\begin{cases} \nabla f(\mathbf{x}^*) - \nabla \mathbf{g}(\mathbf{x}^*) \lambda_g^* - \nabla \mathbf{h}(\mathbf{x}^*) \lambda_h^* = 0; \\ \mathbf{g}(\mathbf{x}^*) \geq 0, \mathbf{h}(\mathbf{x}^*) = 0, \lambda_g^* \geq 0; \\ \lambda_g^{*\top} \mathbf{g}(\mathbf{x}^*) = 0. \end{cases} \quad (1.3)$$

上述条件 (1.3) 也被称作 *Karush-Kuhn-Tucker (KKT)* 条件, 其中 λ_g^* 与 λ_h^* 分别称为在 \mathbf{x}^* 处对应于不等式约束与等式约束的对偶变量 (*dual variable*) 或 *Lagrange* 乘子 (*Lagrange multiplier*). 我们称满足 KKT 条件的 \mathbf{x} 为光滑约束优化问题 (1.2) 的一阶稳定点或 KKT 点.

注. 使定理 1.2 中的假设 “ $\mathcal{T}_F(\mathbf{x}^*) = \mathcal{D}_F(\mathbf{x}^*)$ ” 成立的条件被称为约束规范条件 (constraint qualification). 常用的约束规范条件有线性函数约束规范条件、线性独立约束规范条件、Mangasarian-Fromovitz 约束规范条件等^[1,19,20]. 对一些特殊的优化问题, 例如互补约束优化问题 (mathematical program with complementarity constraints, MPCC), 假设 “ $\mathcal{T}_F(\mathbf{x}^*) = \mathcal{D}_F(\mathbf{x}^*)$ ” 可能在全局最优解处都不成立. 例子可见第 2 章问题 (2.2). 对此类优化问题, 我们需拓展上述一阶稳定点的概念.

注. 关于光滑无约束与约束优化问题局部最优解的二阶充分或必要条件以及二阶稳定点的概念, 我们推荐读者参阅专著^[1,19,20].

再考虑非光滑情形. 我们先将函数的值域拓宽为拓展实数轴 $(-\infty, \infty] := \mathbb{R} \cup \{\infty\}$ ¹, 并称此类函数为拓展实值 (extended valued) 函数. 由此, 我们仅需考虑无约束优化问题. 这是因为任何约束优化问题 (1.1) 都可以写成

$$\min_{\mathbf{x}} f(\mathbf{x}) + \delta_F(\mathbf{x}),$$

其中 δ_F 是可行域 F 的指示函数 (indicator function): 若 $\mathbf{x} \in F$, 则 $\delta_F(\mathbf{x})$ 取 0; 否则, 取无穷大. 对任一拓展实值函数 f , 记其定义域 (domain) 为 $\text{dom}(f) := \{\mathbf{x} : f(\mathbf{x}) < \infty\}$. 我们称其是适定的 (proper), 若 $\text{dom}(f) \neq \emptyset$; 称其在 \mathbf{x} 处是下半连续的 (lower semi-continuous), 若 $\liminf_{\mathbf{y} \rightarrow \mathbf{x}} f(\mathbf{y}) \geq f(\mathbf{x})$. 我们再引入函数 Fréchet 次微分的概念.

定义 1.4 (Fréchet 次微分^[22]). 设 $f : \mathbb{R}^n \rightarrow (-\infty, \infty]$ 为一适定下半连续函数. 我们称

$$\partial f(\mathbf{x}) := \left\{ \mathbf{d} : \liminf_{\mathbf{y} \neq \mathbf{x}, \mathbf{y} \rightarrow \mathbf{x}} \frac{f(\mathbf{y}) - f(\mathbf{x}) - \langle \mathbf{d}, \mathbf{y} - \mathbf{x} \rangle}{\|\mathbf{y} - \mathbf{x}\|} \geq 0 \right\}$$

为 f 在 \mathbf{x} 处的 Fréchet 次微分. 若 $\mathbf{x} \notin \text{dom}(f)$, 则规定 $\partial f(\mathbf{x}) = \emptyset$.

定理 1.3 (凸函数的 Fréchet 次微分^[22]). 设 $f : \mathbb{R}^n \rightarrow (-\infty, \infty]$ 为一适定下半连续凸函数, 则

$$\partial f(\mathbf{x}) = \{ \mathbf{d} : f(\mathbf{y}) - f(\mathbf{x}) \geq \langle \mathbf{d}, \mathbf{y} - \mathbf{x} \rangle, \forall \mathbf{y} \}, \quad \forall \mathbf{x} \in \text{dom}(f).$$

定理 1.4 (非光滑优化问题局部最优解的一阶必要条件^[22]). 设 $f : \mathbb{R}^n \rightarrow (-\infty, \infty]$ 为一适定下半连续函数. 若 $\mathbf{x}^* \in \text{dom}(f)$ 是非光滑优化问题

$$\min_{\mathbf{x}} f(\mathbf{x})$$

的局部最优解, 则 $0 \in \partial f(\mathbf{x}^*)$. 我们称满足 $0 \in \partial f(\mathbf{x})$ 的 \mathbf{x} 为该问题的一阶稳定点.

注. 非光滑函数具有多种次微分的定义, 对应不同的一阶必要条件. 感兴趣的读者可阅读专著^[22] 及其中的参考文献.

¹ 在原始的定义^[21] 中, 拓展实值函数可以取负无穷大. 由于在讨论极小化问题时通常不会遇到函数值取负无穷大的情形, 所以本文将其忽略.

由于凸优化问题的一阶稳定点一定是其局部与全局最优解,因此优化算法可以通过搜寻一阶稳定点求得凸优化问题的最优解.而对于一般的非凸优化问题,根据定理1.1、1.2与1.4,在一定条件下,局部最优解一定是一阶稳定点,但反之不一定成立.因此优化算法可以搜寻一阶稳定点作为局部最优解的备选.

由于从实际应用导出的优化问题往往缺乏显式解,因此人们通常采用迭代型的优化算法求解之.为了使算法能在有限步内终止,我们一般会设置一些终止准则(stopping criterion).下面,我们只考虑最优解与最优值未知的情形.对于光滑无约束优化问题,根据定理1.1,我们可在 $\mathbf{x}^{(k)}$ 满足

$$\frac{\|\nabla f(\mathbf{x}^{(k)})\|}{\max \{\|\nabla f(\mathbf{x}^{(0)})\|, 1\}} \leq \epsilon_1 \quad (1.4)$$

时终止算法,其中 $0 < \epsilon_1 \ll 1$.对于光滑约束优化问题,根据定理1.2,我们可在 $\mathbf{x}^{(k)}$ 满足

$$\frac{\text{KKTViol}_k}{\max \{\text{KKTViol}_0, 1\}} \leq \epsilon_2 \quad (1.5)$$

时终止算法,其中

$$\begin{aligned} \text{KKTViol}_k := \max \Big\{ & \left\| \nabla f(\mathbf{x}^{(k)}) - \nabla \mathbf{g}(\mathbf{x}^{(k)}) \lambda_{\mathbf{g}}^{(k)} - \nabla \mathbf{h}(\mathbf{x}^{(k)}) \lambda_{\mathbf{h}}^{(k)} \right\|, \\ & \left\| \max \{-\mathbf{g}(\mathbf{x}^{(k)}), 0\} \right\|, \left\| \mathbf{h}(\mathbf{x}^{(k)}) \right\|, \left\| \max \left\{ -\lambda_{\mathbf{g}}^{(k)}, 0 \right\} \right\|, \left| \lambda_{\mathbf{g}}^{(k)\top} \mathbf{g}(\mathbf{x}^{(k)}) \right| \Big\}, \end{aligned}$$

$0 < \epsilon_2 \ll 1$, $\lambda_{\mathbf{g}}^{(k)} \in \mathbb{R}^p$ 和 $\lambda_{\mathbf{h}}^{(k)} \in \mathbb{R}^q$ 为迭代过程中产生的对偶变量.我们也称 KKTViol_k 为在 $\mathbf{x}^{(k)}$ 处的 KKT 违反度(KKT violation).此外,不论是求解何种优化问题,我们总可以通过监测相邻迭代点的距离或相邻迭代步目标函数值的差终止算法:

$$\frac{\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|}{\max \{\|\mathbf{x}^{(k)}\|, 1\}} \leq \epsilon_3, \quad \frac{|f(\mathbf{x}^{(k)}) - f(\mathbf{x}^{(k-1)})|}{\max \{|f(\mathbf{x}^{(k)})|, 1\}} \leq \epsilon_4, \quad (1.6)$$

其中 $0 < \epsilon_3, \epsilon_4 \ll 1$.需要指明的是,若优化问题本身不满足一定的误差界(error bound)条件,上述终止准则(1.4)、(1.5)或(1.6)通常并不能直接用来评判优化算法求得解的质量.当优化问题来源于实际应用,我们还需要从应用的角度设计终止准则或解质量的评判标准.

在分析优化算法的理论性质时,我们通常研究它的收敛性(convergence)与收敛速度(convergence rate).

定义1.5(优化算法的收敛性^[1,19,20]).设 $\mathbf{x}^\star \in \mathcal{F}$.我们

(1) 称一个优化算法依子列收敛到 \mathbf{x}^\star ,若其产生的迭代点序列 $\{\mathbf{x}^{(k)}\}$ 至少存在一个子列 $\{\mathbf{x}^{(k_j)}\}_j$,满足

$$\lim_{j \rightarrow \infty} \|\mathbf{x}^{(k_j)} - \mathbf{x}^\star\| = 0.$$

(2) 称一个优化算法全局依子列收敛到 \mathbf{x}^* , 若从任意初始点 $\mathbf{x}^{(0)}$ 出发, 其产生的迭代点序列 $\{\mathbf{x}^{(k)}\}$ 都至少存在一个子列 $\{\mathbf{x}^{(k_j)}\}_j$, 满足

$$\lim_{j \rightarrow \infty} \|\mathbf{x}^{(k_j)} - \mathbf{x}^*\| = 0.$$

(3) 称一个优化算法依点列收敛到 \mathbf{x}^* , 若其产生的迭代点序列 $\{\mathbf{x}^{(k)}\}$ 满足

$$\lim_{k \rightarrow \infty} \|\mathbf{x}^{(k)} - \mathbf{x}^*\| = 0.$$

(4) 称一个优化算法全局依点列收敛到 \mathbf{x}^* , 若从任意初始点 $\mathbf{x}^{(0)}$ 出发, 其产生的迭代点序列 $\{\mathbf{x}^{(k)}\}$ 都满足

$$\lim_{k \rightarrow \infty} \|\mathbf{x}^{(k)} - \mathbf{x}^*\| = 0.$$

类似地, 我们也可以定义函数值序列 $\{f(\mathbf{x}^{(k)})\}$ 的收敛性.

定义 1.6 (优化算法的收敛速度^[1,19,20]). 设 $\mathbf{x}^* \in \mathcal{F}$ 且有一优化算法依点列收敛于 \mathbf{x}^* . 我们

(1) 称该算法 Q-次线性收敛于 \mathbf{x}^* , 若其产生的迭代点序列 $\{\mathbf{x}^{(k)}\}$ 满足

$$\lim_{k \rightarrow \infty} \frac{\|\mathbf{x}^{(k+1)} - \mathbf{x}^*\|}{\|\mathbf{x}^{(k)} - \mathbf{x}^*\|} = 1.$$

称其 Q-线性收敛于 \mathbf{x}^* , 若其产生的迭代点序列 $\{\mathbf{x}^{(k)}\}$ 满足

$$0 < \liminf_{k \rightarrow \infty} \frac{\|\mathbf{x}^{(k+1)} - \mathbf{x}^*\|}{\|\mathbf{x}^{(k)} - \mathbf{x}^*\|} \leq \limsup_{k \rightarrow \infty} \frac{\|\mathbf{x}^{(k+1)} - \mathbf{x}^*\|}{\|\mathbf{x}^{(k)} - \mathbf{x}^*\|} < 1.$$

称其 Q-超线性收敛于 \mathbf{x}^* , 若其产生的迭代点序列 $\{\mathbf{x}^{(k)}\}$ 满足

$$\lim_{k \rightarrow \infty} \frac{\|\mathbf{x}^{(k+1)} - \mathbf{x}^*\|}{\|\mathbf{x}^{(k)} - \mathbf{x}^*\|} = 0.$$

称其 Q-二次收敛于 \mathbf{x}^* , 若其产生的迭代点序列 $\{\mathbf{x}^{(k)}\}$ 满足

$$0 < \liminf_{k \rightarrow \infty} \frac{\|\mathbf{x}^{(k+1)} - \mathbf{x}^*\|}{\|\mathbf{x}^{(k)} - \mathbf{x}^*\|^2} \leq \limsup_{k \rightarrow \infty} \frac{\|\mathbf{x}^{(k+1)} - \mathbf{x}^*\|}{\|\mathbf{x}^{(k)} - \mathbf{x}^*\|^2} < \infty;$$

(2) 称该算法 R-次线性 (线性、超线性或二次) 收敛于 \mathbf{x}^* , 若存在 Q-次线性 (线性、超线性或二次) 收敛于 0 的非负序列 $\{t_k\}$ 满足 $\|\mathbf{x}^{(k)} - \mathbf{x}^*\| \leq t_k$.

注. 为使理论结果有意义, 定义 1.5 和 1.6 中的 \mathbf{x}^* 往往取做优化问题的全局最优解、局部最优解或 (一阶) 稳定点. 优化算法的收敛速度与其复杂度理论紧密相关. 我们推荐感兴趣的读者阅读综述^[23] 及其中的参考文献.

1.3 电子结构计算

当前, 电子结构计算已成为探索物质机理、理解与预测材料性质的重要手段和工具. 本节简要介绍电子结构计算的基本模型与计算方法. 对本节内容感兴趣的读者可参阅专著^[24,25]、综述^[26–30] 及其中的参考文献.

为叙述方便, 下文仅考虑孤立的、非相对论的、定态的、无自旋的多电子体系, 并采用原子单位 (atomic units), 即 $m_e = e = \hbar = 1/(4\pi\epsilon_0) = k_B = 1$, 其中 m_e 是单电子质量, e 是单位电荷量, \hbar 是约化 Planck 常数, ϵ_0 是 Coulomb 介电常数, k_B 是 Boltzmann 常数.

1.3.1 电子总能极小与密度泛函理论

电子结构计算最基本的数学模型是量子力学 Schrödinger 方程^[3]. 基于 Born-Oppenheimer 近似^[31], 求解多电子体系的基态 (ground state) 即是要求最小的 E_0 , 使得存在 $\Psi \in \mathcal{W}_N$, 满足

$$\mathcal{H}\Psi = E_0\Psi. \quad (1.7)$$

这里 E_0 称为多电子体系的基态能量², Ψ 是描述电子的波函数 (wave function), $N \in \mathbb{N}$ 和 $d \in \{1, 2, 3\}$ 分别是体系中电子的个数和体系维数. 我们用 $\mathbf{r}_1, \dots, \mathbf{r}_N \in \mathbb{R}^d$ 表示 N 个电子的位置. \mathcal{H} 是作用在波函数上的 Hamiltonian 算子³, 定义为

$$\mathcal{H} := \sum_{i=1}^N -\frac{1}{2}\Delta_{\mathbf{r}_i} + c_{ee}(\mathbf{r}_1, \dots, \mathbf{r}_N) + \sum_{i=1}^N v_{\text{ext}}(\mathbf{r}_i).$$

其中 $c_{ee} : (\mathbb{R}^d)^N \rightarrow \mathbb{R}$ 是 N 体 Coulomb 势, $v_{\text{ext}} : \mathbb{R}^d \rightarrow \mathbb{R}$ 为 $M \in \mathbb{N}$ 个原子核诱导的外势, 分别定义为

$$\begin{aligned} c_{ee}(\mathbf{r}_1, \dots, \mathbf{r}_N) &:= \sum_{1 \leq i < j}^N \frac{1}{\|\mathbf{r}_i - \mathbf{r}_j\|}, & \forall \{\mathbf{r}_i\}_{i=1}^N \subseteq \mathbb{R}^d, \\ v_{\text{ext}}(\mathbf{r}) &:= -\sum_{j=1}^M \frac{z_j}{\|\mathbf{r} - \mathbf{R}_j\|}, & \forall \mathbf{r} \in \mathbb{R}^d, \end{aligned}$$

$z_j \in \mathbb{N}$ 与 $\mathbf{R}_j \in \mathbb{R}^d$ 分别为第 j 个原子核的电荷数与坐标 ($j = 1, \dots, M$). 集合

$$\mathcal{W}_N := \left\{ \Psi \in \mathcal{H}^1((\mathbb{R}^d)^N; \mathbb{C}) : \Psi \text{ 反对称且 } \int_{(\mathbb{R}^d)^N} \Psi^* \Psi = 1 \right\},$$

其中 “ Ψ 反对称” 来源于 Pauli 不相容原理 (Pauli exclusion principle)^[33], 指对任何 $i, j \in \{1, \dots, N\} : i \neq j$,

$$\Psi(\mathbf{r}_1, \dots, \mathbf{r}_i, \dots, \mathbf{r}_j, \dots, \mathbf{r}_N) = -\Psi(\mathbf{r}_1, \dots, \mathbf{r}_j, \dots, \mathbf{r}_i, \dots, \mathbf{r}_N), \quad \forall \{\mathbf{r}_i\}_{i=1}^N \subseteq \mathbb{R}^d. \quad (1.8)$$

²对于只有一个原子核的阳离子或原子, 基态能量 E_0 的存在性具有理论保证^[32].

³当 Hamiltonian 算子 \mathcal{H} 作用在波函数上时, c_{ee} 与 v_{ext} 的作用应理解为函数相乘. 后文提及“算子作用”时做类似理解.

由变分原理, 求解方程 (1.7) 等价于求解下述电子总能极小问题:

$$E_0 := \inf_{\Psi} \left\{ \int_{(\mathbb{R}^d)^N} \Psi^* \mathcal{H} \Psi : \Psi \in \mathcal{W}_N \right\}. \quad (1.9)$$

根据 Hamilton 算子的定义, 问题 (1.9) 的目标函数可以分解成如下动能 $T_e[\Psi]$ 、电子-电子相互作用能 $V_{ee}[\Psi]$ 与电子-原子核相互作用能 $V_{ne}[\Psi]$ 之和:

$$\begin{aligned} T_e[\Psi] &:= \frac{1}{2} \sum_{i=1}^N \int_{(\mathbb{R}^d)^N} \left\| \nabla_{\mathbf{r}_i} \Psi(\mathbf{r}_1, \dots, \mathbf{r}_N) \right\|^2 d\mathbf{r}_1 \cdots d\mathbf{r}_N, \\ V_{ee}[\Psi] &:= \sum_{1 \leq i < j}^N \int_{(\mathbb{R}^d)^N} \frac{|\Psi(\mathbf{r}_1, \dots, \mathbf{r}_N)|^2}{\|\mathbf{r}_i - \mathbf{r}_j\|} d\mathbf{r}_1 \cdots d\mathbf{r}_N, \\ V_{ne}[\Psi] &:= \sum_{i=1}^N \int_{(\mathbb{R}^d)^N} v_{\text{ext}}(\mathbf{r}_i) |\Psi(\mathbf{r}_1, \dots, \mathbf{r}_N)|^2 d\mathbf{r}_1 \cdots d\mathbf{r}_N. \end{aligned}$$

由于在 Coulomb 势 c_{ee} 中所有电子坐标耦合在一起, 波函数 Ψ 无法解耦, 问题 (1.7) 和 (1.9) 中存在维数灾难, 即搜索空间的维数随电子个数的增加而指数增长. 当电子个数较多时, 直接求解问题 (1.7) 和 (1.9) 几乎不可能. 我们必须寻求问题 (1.7) 和 (1.9) 的近似或转化. 其中, Hohenberg 和 Kohn 于 1964 年提出的密度泛函理论 (density functional theory, DFT)^[34] 因其理论上的精确性及近似计算的操作性而被广泛采用.

为介绍 DFT, 我们先给出单电子密度 (single-particle density) 的定义. 给定一个波函数 $\Psi \in \mathcal{W}_N$, 对应的单电子密度 $\rho_\Psi : \mathbb{R}^d \rightarrow \mathbb{R}$ 定义为

$$\rho_\Psi(\mathbf{r}) := N \int_{(\mathbb{R}^d)^{N-1}} |\Psi(\mathbf{r}, \mathbf{r}_2, \dots, \mathbf{r}_N)|^2 d\mathbf{r}_2 \cdots d\mathbf{r}_N, \quad \forall \mathbf{r} \in \mathbb{R}^d. \quad (1.10)$$

由于 Ψ 具有反对称性 (1.8), 此处选取哪 $N - 1$ 个电子位置积分不会改变 ρ_Ψ 的定义. 按照习惯, 我们用 “ $\Psi \mapsto \rho$ ” 表示一个波函数 $\Psi \in \mathcal{W}_N$ 和一个密度 ρ 满足关系式 (1.10). 此时, 我们也称 ρ 是 N 可表示 (N -representable) 密度.

DFT 表明多电子体系的基态性质可完全由基态单电子密度决定, 而复杂的波函数是冗余的^[34]. 具体地, Hohenberg-Kohn 第一定理指出除一可加常数外, 体系所处外势 v_{ext} 由体系基态单电子密度唯一确定, 其严格数学证明及推广可见文献^[35,36]. Hohenberg-Kohn 第二定理 (也称 Hohenberg-Kohn 变分定理)^[34,37,38] 则将电子总能极小问题 (1.9) 转化为

$$E_0 = \inf_{\rho} \left\{ F_{\text{LL}}[\rho] + \int_{\mathbb{R}^d} v_{\text{ext}}(\mathbf{r}) \rho(\mathbf{r}) d\mathbf{r} : \rho \in \mathcal{D}_N \right\}, \quad (1.11)$$

其中

$$F_{\text{LL}}[\rho] := \min_{\Psi} \left\{ T_e[\Psi] + V_{ee}[\Psi] : \Psi \in \mathcal{W}_N, \Psi \mapsto \rho \right\}$$

通常被称作 Levy-Lieb 能量, F_{LL} 则被称作 Levy-Lieb 泛函. 集合

$$\mathcal{D}_N := \left\{ \rho \geq 0 : \sqrt{\rho} \in \mathcal{H}^1(\mathbb{R}^d; \mathbb{R}), \int_{\mathbb{R}^d} \rho(\mathbf{r}) d\mathbf{r} = N \right\}$$

包含了所有 N 可表示密度.

由于问题 (1.11) 以密度 ρ 为变量, 搜索空间维数与电子个数 N 无关, DFT 在理论上克服了问题 (1.7) 和 (1.9) 中的维数灾难. 然而时至今日, 泛函 F_{LL} 依然没有显式表达式. 为 DFT 在实际计算中得以使用, 对 F_{LL} 的近似不可或缺. 幸运的是, 在 DFT 诞生后的第二年, Kohn 和 Sham 就提出了一个可计算的近似模型 Kohn-Sham DFT^[9]. 目前, 它已经成为电子结构计算中最为广泛使用的数学模型. Kohn 也因在材料电子性质理解上的突出贡献, 获得了 1998 年诺贝尔化学奖.

1.3.2 Kohn-Sham 密度泛函理论与算法

Kohn-Sham DFT (KSDFT) 从一个假想的 (fictitious)、单电子密度是 ρ 的无相互作用 (non-interacting) 体系出发, 将重点放在了 $F_{\text{LL}}[\rho]$ 中动能部分的近似上. 具体地, 首先考虑 $F_{\text{LL}}[\rho]$ 的无相互作用极限

$$\min_{\Psi} T_e[\Psi], \text{ s. t. } \Psi \in \mathcal{W}_N, \Psi \mapsto \rho.$$

将上述问题的变量 Ψ 限制为由标准正交轨道 $\{\phi_i\}_{i=1}^N \subseteq \mathcal{H}^1(\mathbb{R}^d; \mathbb{C})$ 构成的 Slater 行列式^[39]

$$\Psi(\mathbf{r}_1, \dots, \mathbf{r}_N) = \det \begin{bmatrix} \phi_1(\mathbf{r}_1) & \phi_1(\mathbf{r}_2) & \cdots & \phi_1(\mathbf{r}_N) \\ \phi_2(\mathbf{r}_1) & \phi_2(\mathbf{r}_2) & \cdots & \phi_2(\mathbf{r}_N) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_N(\mathbf{r}_1) & \phi_N(\mathbf{r}_2) & \cdots & \phi_N(\mathbf{r}_N) \end{bmatrix}, \quad \forall \{\mathbf{r}_i\}_{i=1}^N \subseteq \mathbb{R}^d$$

时, 我们可以用如下 KS 动能近似 $F_{\text{LL}}[\rho]$ 的动能部分:

$$T_s[\rho] := \inf_{\{\phi_i\}_{i=1}^N} \left\{ \frac{1}{2} \sum_{i=1}^N \int_{\mathbb{R}^d} \|\nabla \phi_i\|^2 : \{\phi_i\}_{i=1}^N \subseteq \mathcal{H}^1(\mathbb{R}^d; \mathbb{C}), \right. \\ \left. \sum_{i=1}^N |\phi_i|^2 = \rho, \int_{(\mathbb{R}^d)^2} \phi_i^* \phi_j = \delta_{ij}, i, j = 1, \dots, N \right\},$$

其中 $\delta_{\mathbf{a}, \mathbf{b}}$ 为 Kronecker 记号: 若 $\mathbf{a} = \mathbf{b}$, $\delta_{\mathbf{a}, \mathbf{b}}$ 取 1; 否则, $\delta_{\mathbf{a}, \mathbf{b}}$ 取 0. 接着, 基于平均场来近似 $F_{\text{LL}}[\rho]$ 中的相互作用能部分:

$$J[\rho] := \frac{1}{2} \int_{(\mathbb{R}^d)^2} \frac{\rho(\mathbf{r}_1)\rho(\mathbf{r}_2)}{\|\mathbf{r}_1 - \mathbf{r}_2\|} d\mathbf{r}_1 d\mathbf{r}_2.$$

我们有时也将 $J[\rho]$ 称为静电自能 (electrostatic self energy). 最后, 将由两部分近似导致的误差归结为交换-关联能 (exchange-correlation energy) $E_{\text{xc}}[\rho]$. 结合 $T_s[\rho]$ 、 $J[\rho]$ 与 $E_{\text{xc}}[\rho]$, 我们就得到了 KSDFT 的数学模型:

$$E_{\text{KS}} := \inf_{\rho} \left\{ T_s[\rho] + J[\rho] + E_{\text{xc}}[\rho] + \int_{\mathbb{R}^d} v_{\text{ext}}(\mathbf{r})\rho(\mathbf{r}) d\mathbf{r} : \rho \in \mathcal{D}_N \right\}. \quad (1.12)$$

交换–关联能泛函 E_{xc} 无显式表达式。自 1965 年起, 对 E_{xc} 的近似模型便不断涌现。现有近似可分为半局部型 (semi-local) 与非局部型 (non-local) 两种。其中半局部型近似是密度泛函, 包括局域密度近似 (local density approximation) 及其变体、广义梯度近似 (generalized gradient approximation) 及其变体等; 非局部型近似不再是密度泛函, 包括杂化泛函 (hybrid functional)、轨道泛函 (orbital functional) 等。我们推荐感兴趣的读者阅读综述^[40] 及其中的参考文献。

选定交换–关联能泛函近似后, 我们可从 KSDFT (1.12) 对应的 Euler-Lagrange 方程出发求解之。以选取半局部型 (semi-local) 交换–关联能泛函近似为例, KS-DFT 对应的 Euler-Lagrange 方程为

$$\begin{cases} (\mathcal{H}_{KS}[\rho])\phi_i = \varepsilon_i \phi_i, \quad i = 1, \dots, N; \\ \sum_{i=1}^N |\phi_i|^2 = \rho; \quad \int_{(\mathbb{R}^d)^2} \phi_i^* \phi_j = \delta_{ij}, \quad i, j = 1, \dots, N. \end{cases} \quad (1.13)$$

这里, $\{\varepsilon_i\}_{i=1}^N \subseteq \mathbb{R}$, \mathcal{H}_{KS} 是 KS Hamiltonian 算子, 定义为

$$\mathcal{H}_{KS}[\rho] := -\frac{1}{2} \Delta_{\mathbf{r}} + \int_{\mathbb{R}^d} \frac{\rho(\mathbf{r}')}{\|\mathbf{r} - \mathbf{r}'\|} d\mathbf{r}' + \frac{\delta E_{xc}[\rho]}{\delta \rho} + v_{\text{ext}},$$

其中 $\delta E_{xc}/\delta \rho$ 是 E_{xc} 的泛函导数。若构造原理 (aufbau principle) 成立, 则求解 KS-DFT (1.12) 就等价于求满足方程 (1.13) 的最小的 $\{\varepsilon_i\}_{i=1}^N$ 和对应的 $\{\phi_i\}_{i=1}^N$ 。这可视为求解一个非线性特征值问题 (nonlinear eigenvalue problem)。

目前, 求解问题 (1.13) 最常用的计算方法是自洽场 (self-consistent field) 迭代^[41] 及其变体^[42–46]。数学上, 自洽场迭代可看作对密度 ρ 的不动点迭代^[47]。此类算法的收敛性已被学者广泛研究。已有结果需要对离散 KS Hamiltonian 算子的特征值间隙 (eigengap) 及选取的交换–关联能泛函近似做假设^[48–51]。自洽场迭代最大的计算开销在于每次迭代需求解一个线性特征值问题, 可调用已有基于正交化的算法^[52–54] 或无需正交化的算法^[55–58]。

除了 Euler-Lagrange 方程, 我们还可以直接求解离散的 KSDFT (1.12)⁴。这对应一个正交约束优化问题:

$$\min_X \hat{T}_s(X) + \hat{J}(X) + \hat{E}_{xc}(X) + \hat{E}_{ne}(X), \quad \text{s. t. } X^* X = I_N. \quad (1.14)$$

这里, \hat{T}_s 、 \hat{J} 、 \hat{E}_{xc} 和 \hat{E}_{ne} 分别是对 KSDFT (1.12) 中 KS 动能、静电自能、交换–关联能和电子–原子核相互作用能泛函的离散。由于正交矩阵全体构成 Stiefel 流形, 因此我们可结合黎曼流形的几何工具^[59] 设计优化算法。这些算法一般需在每步进行显式或隐式的正交化。相关工作可见文献^[60–67]。另外, 还有一些无需正交化的算法。它们一般不保证迭代过程中离散轨道的标准正交性, 但具有更好的并行可扩展性。相关工作可见文献^[55,68–76]。

⁴事实上, 若构造原理不成立, 非线性特征值问题 (1.13) 与 KSDFT (1.12) 并不等价。反例可见文献^[50]。相比之下, 正交约束优化问题 (1.14) 直接通过离散 KSDFT (1.12) 得到, 并不依赖于构造原理。

如前所述, KSDFT 将重点放在近似 $F_{\text{LL}}[\rho]$ 中的动能部分, 仅用静电自能近似相互作用能部分. 所有近似误差的刻画则依赖人为构造的交换–关联能近似. 基于现有的交换–关联能近似, KSDFT 能够较好地描述弱关联体系 (weakly correlated system), 但在刻画强关联体系时有着较大的系统误差^[11]. 然而, 强关联电子体系在实际生活中具有关键的应用, 例如前文所提到的高温超导^[4,5]、磁性材料^[6,7]、热电材料^[8]等. 这激励我们研究适用于强关联电子体系的数学模型与计算方法.

1.3.3 严格关联电子密度泛函理论与算法

强关联电子体系数学模型与计算方法的研究一直是电子结构计算的前沿热点. 现有模型与方法包括变分 Monte Carlo (variational Monte Carlo)^[77]、耦合簇方法 (coupled cluster method)^[78]、密度矩阵重整化群 (density matrix renormalization group)^[79]、动力学平均场理论 (dynamical mean-field theory)^[80]、严格关联电子 (strictly-correlated-electron, SCE) DFT^[81–84]、密度矩阵嵌入理论 (density matrix embedding theory)^[85] 等. 其中, SCEDFT 具有完备的数学理论^[30], 但仍缺乏高效的计算方法. 这是本文的重点内容之一.

SCEDFT 由 KSDFT (1.12) 的对立面导出, 借助一个假想的、具有完全关联作用 (complete correlation) 的体系, 将近似的重点放在了 $F_{\text{LL}}[\rho]$ 中的相互作用能部分. 具体而言, 首先考虑 $F_{\text{LL}}[\rho]$ 的强相互作用极限:

$$\begin{aligned} V_{\text{ee}, \text{SCE}}[\rho] &:= \inf_{\Psi} \left\{ V_{\text{ee}}[\Psi] : \Psi \in \mathcal{W}_N, \Psi \mapsto \rho \right\} \\ &= \inf_{\Psi} \left\{ \int_{(\mathbb{R}^d)^N} c_{\text{ee}}(\mathbf{r}_1, \dots, \mathbf{r}_N) |\Psi(\mathbf{r}_1, \dots, \mathbf{r}_N)|^2 d\mathbf{r}_1 \cdots d\mathbf{r}_N : \Psi \in \mathcal{W}_N, \right. \\ &\quad \left. \int_{(\mathbb{R}^d)^{N-1}} |\Psi(\mathbf{r}, \mathbf{r}_2, \dots, \mathbf{r}_N)|^2 d\mathbf{r}_2 \cdots d\mathbf{r}_N = \frac{\rho(\mathbf{r})}{N}, \forall \mathbf{r} \in \mathbb{R}^d \right\} \\ &= \inf_{\pi} \left\{ \int_{(\mathbb{R}^d)^N} c_{\text{ee}}(\mathbf{r}_1, \dots, \mathbf{r}_N) \pi(\mathbf{r}_1, \dots, \mathbf{r}_N) d\mathbf{r}_1 \cdots d\mathbf{r}_N : i = 1, \dots, N, \forall \mathbf{r} \in \mathbb{R}^d, \right. \\ &\quad \left. \int_{(\mathbb{R}^d)^{N-1}} \pi(\mathbf{r}_1, \dots, \mathbf{r}_{i-1}, \mathbf{r}, \mathbf{r}_{i+1}, \dots, \mathbf{r}_N) d\mathbf{r}_1 \cdots d\mathbf{r}_{i-1} d\mathbf{r}_{i+1} \cdots d\mathbf{r}_N = \frac{\rho(\mathbf{r})}{N} \right\}. \end{aligned} \quad (1.15)$$

我们称 $V_{\text{ee}, \text{SCE}}[\rho]$ 为 SCE 能量, 将其作为 $F_{\text{LL}}[\rho]$ 中相互作用能部分的近似. 接着, 将这一近似的误差与动能部分归为动能去相关能量 (kinetic decorrelation energy) $E_{\text{kd}}[\rho]$ ^[86–88]. 结合 $V_{\text{ee}, \text{SCE}}[\rho]$ 与 $E_{\text{kd}}[\rho]$, 我们就得到了 SCEDFT 的数学模型:

$$E_{\text{SCE}} := \inf_{\rho} \left\{ V_{\text{ee}, \text{SCE}}[\rho] + E_{\text{kd}}[\rho] + \int_{\mathbb{R}^d} v_{\text{ext}}(\mathbf{r}) \rho(\mathbf{r}) d\mathbf{r} : \rho \in \mathcal{D}_N \right\}. \quad (1.16)$$

我们还可以将 KSDFT 与 SCEDFT 结合起来: 使用 KS 动能和 SCE 能量分别近似 $F_{\text{LL}}[\rho]$ 中动能与相互作用能部分, 从而自适应地处理不同关联程度下的电子体系. 由此得到的 KSSCE 数学模型^[89–91] 为

$$E_{\text{KSSCE}} := \inf_{\rho} \left\{ T_s[\rho] + V_{\text{ee}, \text{SCE}}[\rho] + \int_{\mathbb{R}^d} v_{\text{ext}}(\mathbf{r}) \rho(\mathbf{r}) d\mathbf{r} : \rho \in \mathcal{D}_N \right\}. \quad (1.17)$$

显然, SCEDFT (1.16) 和 KSSCE (1.17) 的求解都依赖 $V_{ee,SCE}$ 及其泛函导数 (也被称为 SCE 势)

$$u_{SCE} := \frac{\delta V_{ee,SCE}}{\delta \rho}$$

的计算. 这等同于要求问题 (1.15) 的最优值及其最优对偶势^[91,92].

值得说明的是, (1.15) 式定义的极小化问题未必有最优解^[93]. 为此, 我们只需将问题 (1.15) 的可行域“扩大”⁵为 $(\mathbb{R}^d)^N$ 上所有以 ρ/N 为边际分布的 N 点概率测度构成的集合^[94,95]:

$$\begin{aligned} V_{ee,SCE}[\rho] = \min_{\Pi} & \left\{ \int_{(\mathbb{R}^d)^N} c_{ee}(\mathbf{r}_1, \dots, \mathbf{r}_N) d\Pi(\mathbf{r}_1, \dots, \mathbf{r}_N) : \Pi \in \mathcal{P}((\mathbb{R}^d)^N), \right. \\ & \left. \int_{(\mathbb{R}^d)^{N-1}} d\Pi(\mathbf{r}_1, \dots, \mathbf{r}_{i-1}, \mathbf{r}, \mathbf{r}_{i+1}, \dots, \mathbf{r}_N) = \frac{\rho(\mathbf{r})}{N}, \forall \mathbf{r} \in \mathbb{R}^d, i = 1, \dots, N \right\}. \end{aligned} \quad (1.18)$$

这里, $\mathcal{P}((\mathbb{R}^d)^N)$ 为 $(\mathbb{R}^d)^N$ 上的全体 N 点概率测度构成的空间, 等式约束也称为边际约束 (marginal constraints). 我们称满足该边际约束的 Π 以 ρ/N 为边际分布. 在电子结构计算的背景下, 我们可将 Π 理解为所有电子位置的联合概率测度:

$$\int_{\times_{i=1}^N \mathcal{A}_i} d\Pi(\mathbf{r}_1, \dots, \mathbf{r}_N) = \text{Prob} \left\{ \text{第 } i \text{ 个电子位于集合 } \mathcal{A}_i \subseteq \mathbb{R}^d \text{ 中, } i = 1, \dots, N \right\},$$

其中“ \times ”表示 Cartesian 直积.

注. 如果将问题 (1.18) 的可行域限制为 $(\mathbb{R}^d)^N$ 上所有以 ρ/N 为边际分布的对称 N 点概率测度构成的集合, 由 c_{ee} 的对称性, 问题 (1.18) 的最优值不变且存在对应的最优解. 这里, 对任一 N 点概率测度 $\Pi \in \mathcal{P}((\mathbb{R}^d)^N)$, 我们称其是对称的, 若

$$\int_{\mathcal{A}_1 \times \dots \times \mathcal{A}_N} d\Pi(\mathbf{r}_1, \dots, \mathbf{r}_N) = \int_{\mathcal{A}_{\sigma(1)} \times \dots \times \mathcal{A}_{\sigma(N)}} d\Pi(\mathbf{r}_1, \dots, \mathbf{r}_N)$$

对任何开集 $\mathcal{A}_1, \dots, \mathcal{A}_N \subseteq \mathbb{R}^d$ 以及任何 $\{1, \dots, N\}$ 上的置换 σ (记作 $\sigma \in \mathcal{P}_N$) 都成立. 我们将 $(\mathbb{R}^d)^N$ 上的全体对称 N 点概率测度构成的空间记为 $\mathcal{P}_{\text{sym}}((\mathbb{R}^d)^N)$.

问题 (1.18) 可被视作带 Coulomb 费用的多边际最优运输问题 (multimarginal optimal transport problems with Coulomb cost, MMOT). 关于最优运输, 感兴趣的读者可参阅综述^[96] 和专著^[97]. 在传统最优运输问题中, 运输费用随距离单调递增. 由于 Coulomb 费用具有完全相反的特性, 许多最优运输领域的工具^[97]都无法使用. 当 $N > 2$ (即考虑多边际) 时, 离散的问题 (1.18) 不再等价于某个组合优化问题, 从而无法用现有组合优化算法求解^[98]. 最后, 同电子总能极小问题 (1.9) 类似, 由于显式地纳入了 Coulomb 势 c_{ee} , MMOT (1.18) 中存在维数灾难, 无法直接数值求解. 目前, 一个主流的研究方向是提出替代的低维转化模型并相应地设计计算方法.

⁵尽管问题 (1.15) 在形式上以 N 点概率密度 π 为变量, 我们仍可视其为以 N 点概率测度为变量的极小化问题, 只是要求概率测度有对应的概率密度. 所谓的“扩大”即允许概率测度没有对应的概率密度.

Monge型拟设 (Monge-type ansatz) 是被研究最久的低维假设. 它假定概率测度 Π 可分解为两两电子耦合函数的乘积. 借助此类拟设, 我们可将随电子个数指数增长的自由度降为线性增长的自由度. 现有 Monge型拟设有 Monge 拟设 (Monge ansatz)^[81–83]、类 Monge 拟设 (Monge-like ansatz)^[91] 以及拟 Monge 拟设 (quasi-Monge ansatz)^[99].

Monge 拟设^[81–83] 假定

$$d\Pi(\mathbf{r}_1, \dots, \mathbf{r}_N) = \frac{\rho(\mathbf{r}_1)}{N} \delta(\mathbf{r}_2 - \mathcal{T}_2(\mathbf{r}_1)) \cdots \delta(\mathbf{r}_N - \mathcal{T}_N(\mathbf{r}_1)) d\mathbf{r}_1 \cdots d\mathbf{r}_N,$$

其中 δ 是 Dirac delta 函数, $\mathcal{T}_i : \mathbb{R}^d \rightarrow \mathbb{R}^d$ 是第 1 个和第 i 个电子位置之间的映射⁶, 满足保质量条件

$$\rho(\mathcal{T}_i(\mathbf{r})) d\mathcal{T}_i(\mathbf{r}) = \rho(\mathbf{r}) d\mathbf{r}, \quad i = 2, \dots, N. \quad (1.19)$$

不难看出, Monge 拟设描述了多电子体系的一种特殊状态: 第 1 个电子可以通过 $\{\mathcal{T}_i\}_{i=2}^N$ 决定其他 $N - 1$ 个电子的位置. 这能反映出体系所蕴含的丰富的物理信息. Monge 拟设的正确性已在两电子体系^[95]、一维体系^[100]与一些球对称体系^[101–105]上得到了严格证明. 对这些情形, 我们可以通过求解偏微分方程组从 $\{\mathcal{T}_i\}_{i=1}^N$ 直接得到 SCE 势^[101]. 在 Monge 拟设下, MMOT (1.18) 可被转化为一个带偏微分方程约束的优化问题:

$$\inf_{\{\mathcal{T}_i\}_{i=2}^N} \frac{1}{N} \int_{\mathbb{R}^d} c_{ee}(\mathbf{r}, \mathcal{T}_2(\mathbf{r}), \dots, \mathcal{T}_N(\mathbf{r})) \rho(\mathbf{r}) d\mathbf{r}, \quad \text{s. t. } \{\mathcal{T}_i\}_{i=2}^N \text{ 满足 (1.19) 式.} \quad (1.20)$$

目前, 暂无探讨问题 (1.20) 数值求解的工作.

类 Monge 拟设^[91] 假定

$$d\Pi(\mathbf{r}_1, \dots, \mathbf{r}_N) = \frac{\rho(\mathbf{r}_1)}{N} \gamma_2(\mathbf{r}_1, \mathbf{r}_2) \cdots \gamma_N(\mathbf{r}_1, \mathbf{r}_N) d\mathbf{r}_1 \cdots d\mathbf{r}_N, \quad (1.21)$$

其中 $\gamma_i : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ 是第 1 个和第 i 个电子位置的耦合函数⁷, 满足

$$\int_{\mathbb{R}^d} \gamma_i(\cdot, \mathbf{r}_i) d\mathbf{r}_i = 1, \quad \int_{\mathbb{R}^d} \gamma_i(\mathbf{r}_1, \cdot) \rho(\mathbf{r}_1) d\mathbf{r}_1 = \rho, \quad \gamma_i \geq 0, \quad i = 2, \dots, N. \quad (1.22)$$

类 Monge 拟设可视为 Monge 拟设的推广, 即将 $\delta(\mathbf{r}_i - \mathcal{T}_i(\mathbf{r}_1))$ 替换为 $\gamma_i(\mathbf{r}_1, \mathbf{r}_i)$. 反过来, 我们可使用 γ_i 近似 \mathcal{T}_i :

$$\mathcal{T}_i(\mathbf{r}) \approx \int_{\mathbb{R}^d} \mathbf{r}' \gamma_i(\mathbf{r}, \mathbf{r}') d\mathbf{r}', \quad \forall \mathbf{r} \in \mathbb{R}^d. \quad (1.23)$$

$\rho \gamma_i$ 可以理解成第 1 个和第 i 个电子位置的联合概率密度:

$$\int_{\mathcal{A}_1 \times \mathcal{A}_i} \rho(\mathbf{r}_1) \gamma_i(\mathbf{r}_1, \mathbf{r}_i) d\mathbf{r}_1 d\mathbf{r}_i = \text{Prob}\left\{\text{第 1 个与第 } i \text{ 个电子分别位于 } \mathcal{A}_1 \text{ 与 } \mathcal{A}_i\right\}. \quad (1.24)$$

⁶按惯例, 取 \mathcal{T}_1 为恒等映射.

⁷类似于 Monge 拟设, 取 $\gamma_1 \equiv 1$.

在类 Monge 拟设下, MMOT (1.18) 可被转化为⁸

$$\inf_{\{\gamma_i\}_{i=2}^N} \sum_{i=2}^N \int_{(\mathbb{R}^d)^2} \frac{\rho(\mathbf{r})\gamma_i(\mathbf{r}, \mathbf{r}')}{\|\mathbf{r} - \mathbf{r}'\|} d\mathbf{r} d\mathbf{r}' + \sum_{2 \leq i < j}^N \int_{(\mathbb{R}^d)^3} \frac{\rho(\mathbf{r})\gamma_i(\mathbf{r}, \mathbf{r}')\gamma_j(\mathbf{r}, \mathbf{r}'')}{\|\mathbf{r}' - \mathbf{r}''\|} d\mathbf{r} d\mathbf{r}' d\mathbf{r}'',$$

s. t. $\gamma_2, \dots, \gamma_N$ 满足 (1.22) 式. (1.25)

在引入 K 个单元的有限单元离散后 (离散化细节可见附录一), 我们可从问题 (1.25) 得到一个带广义互补约束的优化问题 (mathematical program with generalized complementarity constraints, MPGCC):

$$\begin{aligned} \min_{\{Y_i\}_{i=2}^N} \quad & \sum_{i=2}^N \langle Y_i, \text{Diag}(\boldsymbol{\rho})C \rangle + \sum_{2 \leq i < j}^N \langle Y_i, \text{Diag}(\boldsymbol{\rho})Y_j C \rangle, \\ \text{s. t.} \quad & Y_i \mathbf{1}_K = \mathbf{1}_K, \quad Y_i^\top \boldsymbol{\rho} = \boldsymbol{\rho}, \quad \text{Tr}(Y_i) = 0, \quad Y_i \geq 0, \quad i = 2, \dots, N, \\ & \langle Y_i, Y_j \rangle = 0, \quad \forall i, j \in \{2, \dots, N\} : i \neq j, \end{aligned} \quad (1.26)$$

其中 $C := (c_{ij}) \in \mathbb{R}^{K \times K}$ 、 $Y_i := (y_{i,jk}) \in \mathbb{R}^{K \times K}$ ($i = 2, \dots, N$) 分别是对两体 Coulomb 势 $1/\|\mathbf{r} - \mathbf{r}'\|$ 、 γ_i ($i = 2, \dots, N$) 的离散, $\boldsymbol{\rho} := [\varrho_1, \dots, \varrho_K]^\top \in \mathbb{R}_+^K$ 是对 ρ 的离散, $\mathbf{1}_K$ 是 \mathbb{R}^K 中的全一向量. 这里非负约束与零内积约束构成了广义互补约束. 由于 MPGCC (1.26) 可行域非凸且不满足常用的约束规范条件^[106], 已有数值求解的工作仅考虑了两电子情形^[91]. 此时, MPGCC (1.26) 退化成一个线性规划 (linear programming).

拟 Monge 拟设^[99] 针对离散情形, 即单电子密度由有限个 Dirac 测度构成: $\rho = \sum_{k=1}^K \varrho_k \delta_{\mathbf{s}_k}$, 其中 $S_K := \{\mathbf{s}_k\}_{k=1}^K \subseteq \mathbb{R}^d$ 是离散位点. 拟 Monge 拟设的数学形式如下:

$$d\Pi(\mathbf{r}_1, \dots, \mathbf{r}_N) = \sum_{k=1}^K \alpha_k \left(\frac{1}{N!} \sum_{\sigma \in \mathcal{P}_N} \prod_{i=1}^N \delta \left(\mathbf{r}_i - \hat{\mathcal{T}}_{\sigma(i)}(\mathbf{s}_k) \right) \right) d\mathbf{r}_1 \cdots d\mathbf{r}_N, \quad (1.27)$$

其中 $\boldsymbol{\alpha} := [\alpha_1, \dots, \alpha_K]^\top \in \mathbb{R}_+^K$ 是 S_K 上的未知概率向量, $\{\hat{\mathcal{T}}_i\}_{i=1}^N$ 是从 S_K 到 S_K 上的未知映射. 它们满足

$$\sum_{k'=1}^K \alpha_{k'} \left(\frac{1}{N} \sum_{i=1}^N \delta_{\mathbf{s}_k, \hat{\mathcal{T}}_i(\mathbf{s}_{k'})} \right) = \frac{\varrho_k}{N}, \quad k = 1, \dots, K. \quad (1.28)$$

当将 $\boldsymbol{\alpha}$ 和 $\hat{\mathcal{T}}_1$ 分别固定为 $\boldsymbol{\rho}/N$ 和恒等映射时, 拟 Monge 拟设就退化成离散情形下的对称 Monge 拟设. 几何上, (1.27) 式右端括号中的部分实际上是 $\mathcal{P}_{\text{sym}}(S_K)$ 的极点, 其总数随电子数 N 指数增长^[99]. 因此, 拟 Monge 拟设本质上假定离散 MMOT (1.18) 存在可仅用 K 个 $\mathcal{P}_{\text{sym}}(S_K)$ 的极点的凸组合表示的最优解, 即存在稀疏极

⁸为节省空间, 我们略去了问题 (1.25) 目标函数的系数 $1/N$ (参见问题 (1.20)).

点表示的最优解, 其正确性已得到了严格证明^[99]. 在拟 Monge 拟设下, MMOT (1.18) 可被转化为一个混合整数规划 (mixed integer programming):

$$\min_{\alpha, \{\hat{\mathcal{T}}_i\}_{i=1}^N} \sum_{k=1}^K c_{ee}(\hat{\mathcal{T}}_1(\mathbf{s}_k), \dots, \hat{\mathcal{T}}_N(\mathbf{s}_k)) \alpha_k, \text{ s. t. } \alpha \geq 0, \{\hat{\mathcal{T}}_i\}_{i=1}^N \text{ 满足 (1.28) 式. (1.29)}$$

目前, 暂无探讨问题 (1.29) 数值求解的工作. 值得一提的是, 借助稀疏极点表示, 有学者进一步研究了离散 MMOT (1.18) 极点表示形式的求解^[107,108].

除了基于 Monge 型拟设的低维转化, 还有一些学者研究 MMOT (1.18) 的 N 可表示形式^[109–111] 与矩约束松弛 (moment constrained relaxation) 形式^[112,113].

MMOT (1.18) 的 N 可表示形式^[109] 为

$$V_{ee,SCE}[\rho] = \min_{\Gamma} \left\{ \frac{N(N-1)}{2} \int_{(\mathbb{R}^d)^2} \frac{1}{\|\mathbf{r} - \mathbf{r}'\|} d\Gamma(\mathbf{r}, \mathbf{r}') : \right. \\ \left. \Gamma \text{ 是 } N \text{ 可表示的, } \int_{\mathbb{R}^d} d\Gamma(\mathbf{r}, \mathbf{r}') d\mathbf{r}' = \frac{\rho(\mathbf{r})}{N}, \forall \mathbf{r} \in \mathbb{R}^d \right\}, \quad (1.30)$$

其以 N 可表示两点概率测度为变量. 这里, 类似于 N 可表示密度的定义 (1.10), 我们称两点概率测度 Γ 是 N 可表示的, 若它是某个 $\Pi \in \mathcal{P}_{sym}((\mathbb{R}^d)^N)$ 的 2-边际测度 (2-marginal):

$$\int_{\mathcal{A}_1 \times \mathcal{A}_2} d\Gamma(\mathbf{r}, \mathbf{r}') = \int_{\mathcal{A}_1 \times \mathcal{A}_2 \times (\mathbb{R}^d)^{N-2}} d\Pi(\mathbf{r}, \mathbf{r}', \mathbf{r}_3, \dots, \mathbf{r}_N), \quad \forall \text{ 开集 } \mathcal{A}_1, \mathcal{A}_2 \subseteq \mathbb{R}^d.$$

然而, 与 N 可表示密度不同的是, N 可表示两点测度并没有一个显式的刻画. 为此, 有学者通过研究两点测度 N 可表示的必要条件, 提出了离散问题 (1.30) 的半定规划松弛 (semidefinite programming relaxation)^[110,111]:

$$\begin{aligned} & \min_{P, Y} \frac{N(N-1)}{2} \langle C, P \rangle, \\ \text{s. t. } & P = \frac{N}{N-1} Y - \frac{1}{N-1} \text{Diag}(Y \mathbf{1}_K), \\ & Y \mathbf{1}_K = \frac{\rho}{N}, \quad \mathbf{1}_K^\top Y \mathbf{1}_K = 1, \quad \text{diag}(Y) = \frac{1}{N} Y \mathbf{1}_K, \\ & Y \geq 0, \quad Y \succeq 0, \end{aligned} \quad (1.31)$$

其中 $C \in \mathbb{R}^{K \times K}$ 、 $\rho \in \mathbb{R}^K$ 同问题 (1.26) 中一样, $P \in \mathbb{R}^{K \times K}$ 是对 Γ 的离散, $Y \in \mathbb{R}^{K \times K}$, 约束 “ $Y \succeq 0$ ” 要求 Y 是一个半正定矩阵. 问题 (1.31) 的最优值可为离散的问题 (1.30) 提供一个下界. 问题 (1.31) 的求解可调用现有的半定规划求解器, 如 SDPNAL+^[114]. 然而, 这并不适合离散规模较大的情形 ($K = 10^4 \sim 10^6$), 这是因为现有求解半定规划的算法在每次迭代需要对 $K \times K$ 矩阵做奇异值分解.

MMOT (1.18) 的矩约束松弛形式^[112,113] 仅对边际约束做矩离散:

$$\int_{(\mathbb{R}^d)^N} \varphi_s(\mathbf{r}_i) d\Pi(\mathbf{r}_1, \dots, \mathbf{r}_N) = \int_{\mathbb{R}^d} \varphi_s(\mathbf{r}) \frac{\rho(\mathbf{r})}{N} d\mathbf{r}, \quad i = 1, \dots, N, \quad s = 1, \dots, S, \quad (1.32)$$

其中 $\{\varphi_s\}_{s=1}^S$ 为人为选定的 $S \in \mathbb{N}$ 个测试函数. 由于边际约束被松弛, 为防止在无穷远处出现不合理的正概率, 我们还需添加额外的约束:

$$\sum_{i=1}^N \int_{(\mathbb{R}^d)^N} \theta(|\mathbf{r}_i|) d\Pi(\mathbf{r}_1, \dots, \mathbf{r}_N) \leq A. \quad (1.33)$$

这里, $\theta : [0, \infty) \rightarrow [0, \infty)$ 是人为选定的函数, 满足 $\lim_{r \rightarrow \infty} \theta(r) = \infty$, $A > 0$. 如此一来, MMOT (1.18) 就被松弛为

$$\begin{aligned} \min_{\Pi} \quad & \int_{(\mathbb{R}^d)^N} c_{ee}(\mathbf{r}_1, \dots, \mathbf{r}_N) d\Pi(\mathbf{r}_1, \dots, \mathbf{r}_N), \\ \text{s. t.} \quad & \Pi \in \mathcal{P}((\mathbb{R}^d)^N) \text{ 满足 (1.32)、(1.33) 式.} \end{aligned} \quad (1.34)$$

在一定条件下, 可证明随着 M 趋于无穷大, 问题 (1.34) 的最优值趋于 MMOT (1.18) 的最优值^[112]. 此外, 问题 (1.34) 存在与拟 Monge 拟设 (1.27) 形式类似的最优解^[112]:

$$d\Pi(\mathbf{r}_1, \dots, \mathbf{r}_N) = \sum_{k=1}^{\tilde{K}} \tilde{\alpha}_k \left(\frac{1}{N!} \sum_{\sigma \in \mathcal{P}_N} \prod_{i=1}^N \delta \left(\mathbf{r}_i - \mathbf{a}_{\sigma(i)}^{(k)} \right) \right) d\mathbf{r}_1 \cdots d\mathbf{r}_N,$$

其中 $\tilde{K} \in \mathbb{N}$ 满足 $\tilde{K} \leq S + 2$, $\tilde{\boldsymbol{\alpha}} := [\tilde{\alpha}_1, \dots, \tilde{\alpha}_{\tilde{K}}]^\top \in \mathbb{R}_{+}^{\tilde{K}}$ 与 $\mathbf{a}_i^{(k)} \in \mathbb{R}^d$ ($i = 1, \dots, N$, $k = 1, \dots, \tilde{K}$) 是未知向量, $\mathbf{1}_{\tilde{K}}^\top \tilde{\boldsymbol{\alpha}} = 1$. 因此, 在选定 $\{\varphi_s\}_{s=1}^S$ 、 θ 与 \tilde{K} 后, 我们只需考虑如下低维问题:

$$\begin{aligned} \min_{\tilde{\boldsymbol{\alpha}}, \{\mathbf{a}_i^{(k)}\}_{i,k}} \quad & \sum_{k=1}^{\tilde{K}} \tilde{\alpha}_k c_{ee} \left(\mathbf{a}_1^{(k)}, \dots, \mathbf{a}_N^{(k)} \right), \\ \text{s. t.} \quad & \frac{1}{N} \sum_{k=1}^{\tilde{K}} \sum_{i=1}^N \tilde{\alpha}_k \varphi_s(\mathbf{a}_i^{(k)}) = \int_{\mathbb{R}^d} \varphi_s(\mathbf{r}) \frac{\rho(\mathbf{r})}{N} d\mathbf{r}, \quad s = 1, \dots, S, \\ & \sum_{k=1}^{\tilde{K}} \sum_{i=1}^N \tilde{\alpha}_k \theta \left(\left| \mathbf{a}_i^{(k)} \right| \right) \leq A, \quad \mathbf{1}_{\tilde{K}}^\top \tilde{\boldsymbol{\alpha}} = 1, \quad \tilde{\boldsymbol{\alpha}} \geq 0. \end{aligned} \quad (1.35)$$

在文献^[113]中, 作者为求解问题 (1.35) 设计了约束过阻尼 Langevin 动力学 (constrained overdamped Langevin dynamics), 其本质是投影梯度下降算法, 投影由 Newton 算法近似计算. 该算法的数值表现严重依赖测试函数与参数的选取.

除了上面介绍的 MMOT (1.18) 的低维转化, 还有学者研究其 Kantorovich 对偶 (Kantorovich duality) 形式^[115] 以及熵正则化 (entropy regularization) 形式^[116]. 它们因保留了随电子数指数增长的自由度或复杂度, 均不适用于处理电子数较多或离散规模较大的情形, 在此不作介绍.

我们将 MMOT (1.18) 的已有低维转化模型与计算方法总结在表 1.1 中. 不难看出, 现有低维转化模型都十分复杂、难以数值求解; 绝大多数现有数值方法可扩展性不高或对实验设置敏感, 还难以用来模拟实际的强关联电子体系. 不过, 相

表 1.1 MMOT (1.18) 的已有低维转化模型与数值方法

Table 1.1 The existing reformulations and numerical approaches for MMOT (1.18)

假设或转化	优化模型	现有数值方法	现有工作主要缺陷	参考文献
Monge 拟设	带偏微分方程约束的优化问题 (1.20)	暂无	暂无	[81–83]
类 Monge 拟设	带广义互补约束的优化问题 (1.26)	线性规划算法	只能处理两电子情形	[91]
拟 Monge 拟设	混合整数规划 (1.29)	暂无	暂无	[99]
N 可表示	半定规划 (1.31)	半定规划算法	无法处理较大规模离散	[109–111]
矩约束松弛	约束优化问题 (1.35)	投影梯度下降算法	数值表现对测试函数与参数选取敏感	[112,113]

比于 N 可表示形式半定规划松弛 (1.31) 与矩约束松弛形式 (1.35), Monge 型拟设下的优化问题蕴含了丰富的两电子关联信息. 我们可基于此评估算法所得解的质量, 可见后文第 3 与第 4 章的内容. 相比于其他 Monge 型拟设, 类 Monge 拟设导出的 MPGCC (1.26) 结构更加简单, 适合作为模型分析与算法设计的突破口.

围绕 SCEDFT (1.16) 中 MMOT (1.18) 的求解, 本文聚焦类 Monge 拟设下的优化模型 MPGCC (1.26), 将分析其模型性质, 为其设计可扩展算法.

1.4 第一性原理晶体结构弛豫

晶体是由大量微观物质单位(原子、分子等)在三维空间按周期性排列形成的结构^[117]. 晶体的分布非常广泛. 自然界的固体物质绝大多数是晶体, 而气体、液体和非晶物质在一定条件下也可以转变成晶体. 晶体结构由原子的排布方式决定. 原子的排布可直接影响材料的物理、化学、热力学等性质. 第一性原理晶体结构弛豫通过改变原子的位置搜索势能面(potential energy surface)上的平衡态(equilibrium state), 其中势能面信息(如能量、受力等)通过电子结构计算(如求解KSDFT (1.12))获取. 本节简要介绍第一性原理晶体结构弛豫的基本模型与计算方法.

1.4.1 晶体结构的数学描述

由于晶体内部存在三维周期性结构, 因此我们只需要一个可重复的结构单元以及三维空间中的无限整数点阵就可描述晶体的结构. 这里, 可重复的结构单元称为单胞(unit cell), 无限整数点阵称为 Bravais 晶格^[13]. 数学上, 我们可以用三个晶格基矢(lattice vectors) $A := [\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3] \in \mathbb{R}^{3 \times 3}$ 与单胞中原子的位置(简称原子位置) $\mathbf{R} := [R_1, \dots, R_M] \in \mathbb{R}^{3 \times M}$ 描述晶体结构⁹:

$$\mathbf{R}_m + \mathbb{Z}\mathbf{a}_1 + \mathbb{Z}\mathbf{a}_2 + \mathbb{Z}\mathbf{a}_3, \quad m = 1, \dots, M.$$

在后文中, 我们将 (\mathbf{R}, A) 称为构型(configuration), 分别称 \mathbf{a}_1 、 \mathbf{a}_2 与 \mathbf{a}_3 为单胞的 a 轴、b 轴与 c 轴. 图 1.1 展示了一个立方晶格¹⁰. 由于三维周期性, 单胞中只有两个

⁹晶体结构的描述不是唯一的. 这取决于晶体结构的对称性以及所选取的晶格基矢.

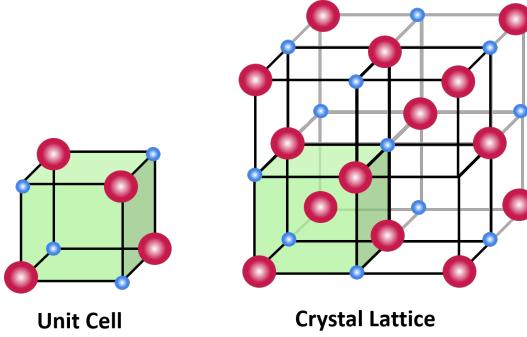


图 1.1 晶体结构示意图, 其中红色(大)球与蓝色(小)球代表不同类型原子. 左图: 单胞. 右图: 晶格

Figure 1.1 An illustration of crystal structure, where the (big) red and (small) blue balls represent different kinds of atoms. Left: Unit cell. Right: Crystal lattice

原子. 若以单胞左下前方格点为原点, 则原子位置为

$$R = \begin{pmatrix} 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

1.4.2 第一性原理晶体结构弛豫问题与算法

在定义好晶格基矢 A 与原子位置 R 后, 一般的第一性原理晶体结构弛豫就是要基于电子结构计算, 寻找下面优化问题的一阶稳定点:

$$\min_{R, A} E(R, A), \text{ s. t. } (R, A) \in \mathcal{F} := \mathcal{F}_{\text{atom}} \times \mathcal{F}_{\text{latt}}, \quad (1.36)$$

其中 E 是能量泛函, $\mathcal{F}_{\text{atom}} \subseteq \mathbb{R}^{3 \times M}$ 、 $\mathcal{F}_{\text{latt}} \subseteq \mathbb{R}^{3 \times 3}$ 分别是原子位置、晶格基矢的可行域. 所谓的势能面即为 $\{(R, A), E(R, A) : (R, A) \in \mathcal{F}\}$. 电子结构计算(如求解 KSDFT (1.12)) 可为弛豫算法提供势能面的信息: 对给定构型 (R, A) , 结构的能量 $E(R, A)$ 可通过电子结构计算直接得到¹⁰, 原子受力 $F_{\text{atom}}(R, A) := -\nabla_R E(R, A)$ 与晶格应力(stress) $\Sigma(R, A)$ 则可通过 Hellmann-Feynman 定理^[118,119] 计算. 进一步地, 晶格受力 $F_{\text{latt}}(R, A) := -\nabla_A E(R, A)$ 与其他量有如下关系^[120,121]:

$$F_{\text{latt}}(R, A) = \det(A)\Sigma(R, A)A^{-\top} - F_{\text{atom}}(R, A)(A^{-1}R)^{\top}.$$

可行域 $\mathcal{F}_{\text{atom}} \times \mathcal{F}_{\text{latt}}$ 刻画了实际应用对构型的限制条件, 较为简单的取法有

- (1) $\mathcal{F}_{\text{atom}} = \mathbb{R}^{3 \times M}$ 、 $\mathcal{F}_{\text{latt}} = \mathbb{R}^{3 \times 3}$: 无约束情形;
- (2) $\mathcal{F}_{\text{atom}} = \{R : R_I = \bar{R}_I \in \mathbb{R}^{3 \times |\mathcal{I}|}\}$: 固定序号在集合 $\mathcal{I} \subseteq \{1, \dots, M\}$ 中的原子的位置, 其中 \bar{R}_I 是给定原子位置;

¹⁰ 图片来源: <https://www.expii.com/t/crystal-lattice-structure-formation-7999>.

¹¹ 为完整描述无限周期性结构, 我们需要使用无穷个离散基函数. 实际计算往往对离散基组做截断.

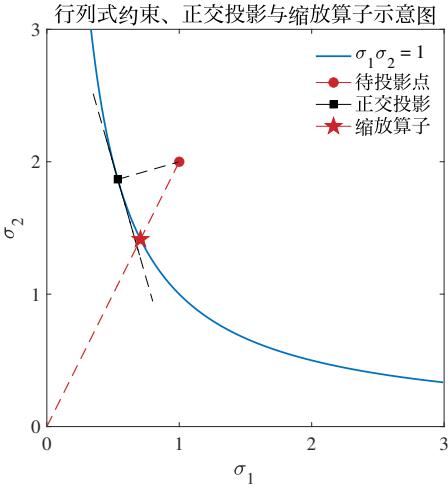


图 1.2 二维情形下行列式约束 $\det(A) = 1$ 对应的可行域、正交投影与缩放算子示意图. 其中, 横纵坐标为 A 的两个奇异值, 蓝色实线代表可行域, 红色圆点代表待投影点, 黑色方形代表其到可行域上的正交投影, 红色五角星代表缩放算子作用在该点上的结果

Figure 1.2 An illustration of the feasible region corresponding to the determinant constraint $\det(A) = 1$, orthogonal projection, and scaling operator. The horizontal and vertical axes denote two singular values of A , the blue solid line stands for the feasible region, the red circle is for the point to be projected, the black square is for its orthogonal projection onto the feasible region, the red pentagon is for the action of scaling operator on the point

- (3) $\mathcal{F}_{\text{latt}} = \{A : A = \bar{A} \in \mathbb{R}^{3 \times 3}\}$: 固定晶格基矢, 其中 \bar{A} 是给定晶格基矢;
- (4) $\mathcal{F}_{\text{latt}} = \{A : A = \gamma \bar{A}, \gamma \geq 0, \bar{A} \in \mathbb{R}^{3 \times 3}\}$: 固定晶格形状, 其中 \bar{A} 是给定晶格基矢;
- (5) $\mathcal{F}_{\text{latt}} = \{A : \det(A) = V > 0\}$: 固定晶格体积, 其中 V 是给定晶格体积.

上述取法中, 较为特殊的是取法 (5), 其对应一个非凸行列式约束优化问题 (determinant-constrained optimization):

$$\min_{R, A} E(R, A), \text{ s. t. } \det(A) = V. \quad (1.37)$$

问题 (1.37) 常见于材料物态方程计算, 对材料力学、热力学等性质的计算颇为重要^[12,13]. 常用材料模拟软件 (如 Vienna *Ab initio* Simulation Package (VASP)^[122,123]、Quantum ESPRESSO (QE)^[124]) 均提供固定晶格体积晶体结构弛豫的功能. 不过, 由于可行域的非凸性 (可见图 1.2), 传统优化中的许多工具均无法直接使用. 例如, 任取 $A \in \mathbb{R}^{3 \times 3}$, 其到问题 (1.37) 可行域的正交投影没有显式表达式且可能不唯一. 这给算法的设计与收敛性分析带来了巨大挑战. 目前, 尚无已有工作为行列式优化问题设计具有收敛性保证的算法.

尽管随着实际应用场景的改变, 问题 (1.36) 的具体形式千差万别, 且往往呈现高度非线性、非凸等特征, 但人们通常不会从数学上严格地处理 (甚至不处理) 约束, 使用的弛豫算法也大抵相同. 最为广泛使用的弛豫算法有二: 非线性共轭梯度 (nonlinear conjugate gradient, CG) 算法^[125–130] 与拟 Newton (quasi-Newton, QN) 算法^[42,43,131–135]. 其中, CG 算法相较 QN 算法更加稳定, 而 QN 算法在局部最优

附近收敛更快。下面，我们简要介绍这两个算法在 VASP 中的实现。

VASP 中实现的 CG 算法与传统教科书^[1,20]中的 CG 算法（使用 Polak-Ribière-Polyak 共轭参数）在更新原子位置时基本一致，而在更新晶格基矢时有较大差异。以无约束情形为例，VASP 中 CG 算法的更新格式为

$$\begin{aligned}\beta_k &:= \frac{\left\langle F_{\text{atom}}^{(k)}, F_{\text{atom}}^{(k)} - F_{\text{atom}}^{(k-1)} \right\rangle + \left\langle \Sigma^{(k)}, \Sigma^{(k)} - \Sigma^{(k-1)} \right\rangle / M}{\left\| F_{\text{atom}}^{(k-1)} \right\|^2 + \left\| \Sigma^{(k-1)} \right\|^2 / M}, \\ D_{\text{atom}}^{(k)} &:= F_{\text{atom}}^{(k)} + \beta_k D_{\text{atom}}^{(k-1)}, \quad R^{(k+1)} := R^{(k+1)} + \eta_k D_{\text{atom}}^{(k)}, \\ D_{\text{latt}}^{(k)} &:= \frac{\Sigma^{(k)}}{M} + \beta_k D_{\text{latt}}^{(k-1)}, \quad A^{(k+1)} := \left(I_3 + \eta_k D_{\text{latt}}^{(k)} \right) A^{(k)},\end{aligned}$$

其中 $F_{\text{atom}}^{(k)} := F_{\text{atom}}(R^{(k)}, A^{(k)}) \in \mathbb{R}^{3 \times M}$, $\Sigma^{(k)} := \Sigma(R^{(k)}, A^{(k)}) \in \mathbb{R}^{3 \times 3}$, $\eta_k > 0$ 是通过线搜索确定的步长。主要区别有二：(1) 对与晶格应力相关的量添加了额外的缩放因子。形式上，这与预条件 CG 算法相似，但不完全相同；(2) 使用晶格应力更新搜索方向，使用类似于应变 (strain) 定义^[12,13]的方式更新晶格基矢。如此一来，搜索方向是否是下降方向便不那么明确。不过，在实际数值模拟时，VASP 的 CG 算法相比传统 CG 算法效率普遍更高，且具有相似的收敛性。

在无约束情形下，VASP 中 QN 算法的更新格式为

$$\begin{aligned}\left(D_{\text{atom}}^{(k)}, D_{\text{latt}}^{(k)} \right) &:= S^{(k)} Y^{(k)\dagger} \left(F_{\text{atom}}^{(k)}, \Sigma^{(k)} \right), \\ R^{(k+1)} &:= R^{(k)} + D_{\text{atom}}^{(k)}, \quad A^{(k+1)} := \left(I_3 + D_{\text{latt}}^{(k)} \right) A^{(k)}.\end{aligned}$$

这里，

$$\begin{aligned}S^{(k)} &:= \left((R^{(k-j)} - R^{(k-j-1)}, A^{(k-j)} - A^{(k-j-1)}) \right)_{j=0}^{\ell-1} \in \mathbb{R}^{3 \times (M+3)\ell}, \\ Y^{(k)} &:= \left(\left(F_{\text{atom}}^{(k-j-1)} - F_{\text{atom}}^{(k-j)}, \Sigma^{(k-j-1)} - \Sigma^{(k-j)} \right) \right)_{j=0}^{\ell-1} \in \mathbb{R}^{3 \times (M+3)\ell},\end{aligned}$$

$\ell \in \mathbb{N}$ 为存储构型的个数。该算法可以看作是特殊的 Broyden 方法^[42,43]，其通过如下方式更新对 Hessian 逆的近似：

$$\min_{\mathbf{H}} \left\| \mathbf{H} - \mathbf{H}^{(k-1)} \right\|^2, \quad \text{s. t. } \mathbf{H}[Y^{(k)}] = S^{(k)},$$

并始终令 $\mathbf{H}^{(k-1)} = 0$ 。

在有约束情形时，VASP 中 CG 算法与 QN 算法的更新方式与上述稍有不同。例如，对固定晶格体积的晶体结构弛豫，需将上面的 Σ 改为晶格偏应力 (deviatoric stress)

$$\Sigma_{\text{dev}} := \Sigma - \frac{\text{Tr}(\Sigma)}{3} \cdot I_3 \in \mathbb{R}^{3 \times 3}.$$

此外，为保证每次迭代后晶格体积满足约束，还需增加如下步骤：

$$A^{(k+1)} := \mathcal{P}_V(A^{(k+1)}) := \sqrt[3]{\frac{V}{\det(A^{(k+1)})}} A^{(k+1)}. \quad (1.38)$$

我们将 \mathcal{P}_V 称为缩放算子 (scaling operator). (1.38) 式充分地利用了行列式的性质, 可将任一晶格基矢矩阵显式拉回至可行域中. 与正交投影算子不同, \mathcal{P}_V 不保证对晶格基矢矩阵做最小改变, 但是单值的且有解析的表达式. 缩放算子与正交投影的对比可见图 1.2.

不可否认的是, VASP 中实现的 CG 算法与 QN 算法均为工程师们仔细打磨的产物. 然而, 它们在效率与收敛性上仍有不足, 尤其是在处理实际应用场景时, 如固定晶格体积情形. 另外, 它们在每次迭代均需要通过昂贵的电子结构计算获取势能面的信息. 这些特点使第一性原理晶体结构弛豫成为下游应用的计算瓶颈. 本文将为一般的行列式优化问题设计高效且具有理论保证的算法, 并将其用于第一性原理固定晶格体积晶体结构弛豫问题 (1.37).

1.5 本文主要内容

立足于优化与计算材料科学的交叉点, 本文研究具有特殊结构的优化问题及其理论与算法, 并将它们应用于强关联电子体系计算与第一性原理晶体结构弛豫. 以下是各章内容简介.

在第 2 章中, 我们研究了一类特殊的 MPGCC, 并证明了其 ℓ_1 罚函数精确性 (exactness). 我们所考虑的 MPGCC 具有目标函数多仿射、可行域为可分多面体的特征. MPGCC (1.26) 是此类问题的特例. 由于广义互补约束的存在, MPGCC 可能不满足约束规范条件, 从而 KKT 条件可能不再是其局部最优解的必要条件. 为此, 一种常用的解决方案是将广义互补约束以 ℓ_1 形式惩罚到目标函数上, 以此消除约束规范条件不成立带来的困难. 如此一来, 证明原问题与罚问题的等价性 (即罚函数的精确性) 就尤为重要. 在本章中, 我们首先给出了 MPGCC (1.26) 的一个实例, 其 ℓ_1 罚函数的精确性无法被已有理论结果覆盖. 接着, 我们充分借助所考虑问题类的代数与几何结构, 证明了 ℓ_1 罚函数的精确性. 这些结果为后面章节中优化算法的应用提供了理论基础.

在第 3 章中, 我们考虑了具有块状结构的优化问题, 并研究了不可行非精确邻近交替线性化极小化 (proximal alternating linearized minimization, PALM) 算法的理论性质. MPGCC (1.26) 的 ℓ_1 罚问题是此类问题的特例. 对于 PALM 算法, 已有工作的理论分析需建立在目标函数值序列的某种单调性之上. 然而, 在许多情形下, 人们会倾向或不得不使用不可行算法求解 PALM 算法中的子问题. 例如, 在使用 PALM 算法求解 MPGCC (1.26) 的 ℓ_1 罚问题时, 子问题变量个数随离散规模平方增长, 而等式约束个数线性增长. 当离散规模较大时, 从对偶的角度求解子问题显然更加高效. 在这种情形下, PALM 算法的迭代点是不可行的, 其目标函数值序列的非单调性无法避免. 我们称此时的 PALM 算法为不可行非精确的 PALM (infeasible inexact PALM, PALM-I) 算法. 目前, 唯一一个考虑 PALM-I 算法收敛性的工作需要对子问题的求解设置无法验证的终止准则, 所得理论结果不具有实际意义. 在本章中, 我们首先为 PALM-I 算法子问题的求解设计了一个可实现的终止准则. 随后, 基于子问题满足的误差界与所设计的终止准则, 我们巧妙地构造了

一个单调下降的代理序列. 以此, 我们首次在可实现的条件下, 证明了 PALM-I 算法的全局依子(点)列收敛性与渐进收敛速度. 在数值实验中, 我们 (1) 在测试问题上对比了可行的 PALM 算法与 PALM-I 算法的效率; (2) 将 PALM-I 算法嵌入了一个瀑布型多重网格优化 (cascadic multigrid optimization, CMGOPT) 框架, 通过求解 MPGCC (1.26) 的 ℓ_1 罚问题, 模拟了一维、二维强关联电子体系. 值得一提的是, 我们首次可视化了二维情形下电子位置之间的映射.

在第 4 章中, 我们考虑了运输多胞体 (transport polytopes) 上的分块矩阵优化问题, 并为其设计了完全无需全矩阵的块坐标下降型算法. MPGCC (1.26) 的 ℓ_1 罚问题同样是此类问题的特例. 为求解此类问题, 已有块坐标下降型算法需要显式存储或操作矩阵变量. 但来源于实际应用的问题往往维数较高. 例如, MPGCC (1.26) 的 ℓ_1 罚问题所涉及的矩阵变量维数由离散规模决定, 可达数十万甚至上百万. 在这种情形下, 已有块坐标下降型算法具有较高的空间与计算复杂度. 在本章中, 我们结合最优运输工具与矩阵逐元素随机近似设计了全新的块坐标下降型算法. 其中, 使用矩阵逐元素随机近似等价于在子问题中增加随机置零约束. 这完全免去了全矩阵的存储与计算, 使新算法与已有随机块坐标下降型算法有根本的不同. 我们为新算法建立了概率意义下的收敛性质, 首次为矩阵逐元素随机近似在分块非凸问题上的应用提供了理论保证. 在数值实验中, 我们 (1) 通过一维强关联电子体系的模拟, 对比了 PALM-I 算法与新算法的效率; (2) 将新算法嵌入了 CMGOPT 框架, 通过求解 MPGCC (1.26) 的 ℓ_1 罚问题, 模拟了二维、三维强关联电子体系. 特别地, 得益于新算法的低标度, 我们首次可视化了三维情形下电子位置之间的映射.

在第 5 章中, 我们考虑了行列式约束优化问题, 并为其设计了基于缩放算子 (1.38) 的投影梯度下降 (projected gradient descent, PGD) 算法. 此类问题与第一性原理固定晶格体积晶体结构弛豫问题 (1.37) 紧密相关. 目前, 暂无已有工作在使用缩放算子 (1.38) 保持迭代点可行性时, 分析算法的理论性质. 为此, 我们以负梯度在可行域切锥上的正交投影为搜索方向, 结合缩放算子 (1.38), 设计了 PGD 算法. 借助可行域切锥的代数结构与非单调线搜索, 我们证明了 PGD 算法的全局依子列收敛性. 在数值实验中, 我们 (1) 将 PGD 算法推广至求解第一性原理固定晶格体积晶体结构弛豫问题 (1.37); (2) 构建了含有 223 个来自不同类别结构的基准算例集, 并在其上对比了 CG 算法、QN 算法与 PGD 算法的效率与鲁棒性; (3) 将 PGD 算法用于计算难弛豫的高熵合金 AlCoCrFeNi 的物态方程. 我们得到的计算结果与已有实验测定数据高度吻合.

最后, 在第 6 章中, 我们总结全文的主要内容, 并探讨后续工作中值得继续研究的问题.

第2章 一类带广义互补约束优化问题的 ℓ_1 罚函数精确性

在本章中, 我们研究一类 MPGCC 的 ℓ_1 罚函数, 并证明其精确性. 第1章中的 MPGCC (1.26) 是此类问题的特例. 我们先给出了 MPGCC (1.26) 的一个实例, 其 ℓ_1 罚函数的精确性无法被已有理论结果覆盖. 接着, 借助所考虑问题类的代数与几何结构, 我们证明了 ℓ_1 罚函数的精确性. 本章内容为后文优化算法的设计与应用提供了优化模型的理论基础.

2.1 问题描述与研究现状

2.1.1 问题描述

MPGCC 一般具有如下形式:

$$\begin{aligned} & \min_{\mathbf{x}_1, \dots, \mathbf{x}_s} f(\mathbf{x}_1, \dots, \mathbf{x}_s), \\ \text{s. t. } & \mathbf{g}(\mathbf{x}_1, \dots, \mathbf{x}_s) \geq 0, \mathbf{x}_i \geq 0, i = 1, \dots, s, \\ & \langle \mathbf{x}_i, \mathbf{x}_j \rangle = 0, \forall i, j \in \{1, \dots, s\} : i \neq j, \end{aligned} \quad (2.1)$$

其中 $\mathbf{x}_i \in \mathbb{R}^m (i = 1, \dots, s)$, $f : (\mathbb{R}^m)^s \rightarrow \mathbb{R}$, $\mathbf{g} : (\mathbb{R}^m)^s \rightarrow \mathbb{R}^p$, $m, s \in \mathbb{N}$. 非负约束与零内积约束构成了广义互补约束: 对任意两个变量块 \mathbf{x}_i 与 $\mathbf{x}_j (i \neq j)$, 若 \mathbf{x}_i 中某个分量大于 0, 则 \mathbf{x}_j 在对应位置必然是 0. 当 s 等于 2 时, MPGCC (2.1) 退化成 MPCC, 其在经济、工程等领域具有诸多应用. 感兴趣的读者可参阅专著^[136] 及其中的参考文献. 进一步地, 若 f 与 \mathbf{g} 还是仿射函数, 则 MPCC 将退化为互补约束线性规划 (linear program with complementarity constraints, LPCC)^[137,138].

由于互补型约束的存在, 不论是 MPCC 还是 MPGCC, 它们在任意可行点处都不满足常用的约束规范条件 (例如 Mangasarian-Fromovitz 约束规范条件^[106]). 因此, 它们的局部最优解 (从而全局最优解) 未必满足 KKT 条件. 一个简单的例子如下所示:

$$\begin{aligned} & \min_{x_1, x_2, x_3} x_1^2 + x_2^2 - x_3, \\ \text{s. t. } & -4x_1 + x_3 \leq 0, -4x_2 + x_3 \leq 0, x_1, x_2 \geq 0, x_1 x_2 = 0. \end{aligned} \quad (2.2)$$

可以验证, 对上面这个 MPCC, KKT 条件在其全局最优解 $[0, 0, 0]^\top$ 处都不成立. 此外, 还可以证明, 求解一个一般的 MPCC 或 MPGCC 是 NP-完全的^[139–141]. 这些困难给算法的设计与分析带来了巨大困难. 为此, 许多学者针对 MPCC 提出了更弱的约束规范条件以及相应的稳定性概念^[106,142–147]. 这些稳定性概念的代数表征大多较为复杂.

除了直接处理原本的 MPCC 或 MPGCC, 我们还可以将互补型约束以 ℓ_1 形式惩罚到目标函数上¹. 之后, 只需考虑 ℓ_1 罚问题的求解. 以 MPGCC (2.1) 为例,

¹ 我们只惩罚零内积约束, 保留非负约束. 这样做的好处是, 在非负约束下, 惩罚项是光滑的 (可见问题 (2.4)).

其对应的 ℓ_1 罚问题为

$$\begin{aligned} \min_{\mathbf{x}_1, \dots, \mathbf{x}_s} \quad & f(\mathbf{x}_1, \dots, \mathbf{x}_s) + \beta p(\mathbf{x}_1, \dots, \mathbf{x}_s), \\ \text{s. t.} \quad & \mathbf{g}(\mathbf{x}_1, \dots, \mathbf{x}_s) \geq 0, \mathbf{x}_i \geq 0, i = 1, \dots, s. \end{aligned} \quad (2.3)$$

这里, $\beta > 0$ 为罚参数, ℓ_1 罚项

$$p(\mathbf{x}_1, \dots, \mathbf{x}_s) := \sum_{1 \leq i < j}^s |\langle \mathbf{x}_i, \mathbf{x}_j \rangle|, \quad \forall \mathbf{x}_1, \dots, \mathbf{x}_s \in \mathbb{R}^m.$$

由于变量均非负, 问题 (2.3) 进一步等价于

$$\begin{aligned} \min_{\mathbf{x}_1, \dots, \mathbf{x}_s} \quad & f_\beta(\mathbf{x}_1, \dots, \mathbf{x}_s) := f(\mathbf{x}_1, \dots, \mathbf{x}_s) + \beta \sum_{1 \leq i < j}^s \langle \mathbf{x}_i, \mathbf{x}_j \rangle, \\ \text{s. t.} \quad & \mathbf{g}(\mathbf{x}_1, \dots, \mathbf{x}_s) \geq 0, \mathbf{x}_i \geq 0, i = 1, \dots, s. \end{aligned} \quad (2.4)$$

由于罚问题 (2.4) 中没有互补型约束, 因此在其局部最优点处常用的约束规范条件可能成立. 我们只需调用优化算法求解罚问题 (2.4). 需要指出的是, 按这条路线求解原本的 MPGCC (2.1), 其合理性需建立在 ℓ_1 罚函数的精确性上, 即当 β 充分大时, MPGCC (2.1) 与其罚问题 (2.4) 的最优解集相同.

注. 我们还可以使用其他形式的罚函数^[148,149], 例如

$$p_1(\mathbf{x}_1, \dots, \mathbf{x}_s) := \sum_{1 \leq i < j}^s \min\{\mathbf{x}_i, \mathbf{x}_j\}, \quad p_2(\mathbf{x}_1, \dots, \mathbf{x}_s) := \sum_{1 \leq i < j}^s \sqrt{\langle \mathbf{x}_i, \mathbf{x}_j \rangle}.$$

在一定条件下, 这些罚函数是精确的. 但它们或是非光滑的, 或是在原点处导数无界. 因此从计算角度而言, 这些罚函数不如 ℓ_1 罚函数性质好. 我们在下文仅讨论 ℓ_1 罚函数及其精确性.

本章考虑一类特殊的 MPGCC, 其数学形式为

$$\begin{aligned} \min_{\mathbf{x}_1, \dots, \mathbf{x}_s} \quad & f(\mathbf{x}_1, \dots, \mathbf{x}_s), \\ \text{s. t.} \quad & \mathbf{x}_i \in \mathcal{F}_i, \mathbf{x}_i \geq 0, i = 1, \dots, s, \\ & \langle \mathbf{x}_i, \mathbf{x}_j \rangle = 0, \forall i, j \in \{1, \dots, s\} : i \neq j. \end{aligned} \quad (2.5)$$

此处, $f : (\mathbb{R}^m)^s \rightarrow \mathbb{R}$ 是多仿射的 (multi-affine), 即对任一 $i \in \{1, \dots, s\}$, 在固定其他 $s - 1$ 个变量块后, f 对 \mathbf{x}_i 是仿射的. 变量块的可行域 $\{\mathcal{F}_i\}_{i=1}^s$ 是 \mathbb{R}^m 中的多面体 (polyhedron). 在后文中, 我们记 $\mathbf{x} := (\mathbf{x}_1, \dots, \mathbf{x}_s) \in (\mathbb{R}^m)^s$, $\mathcal{F}_i^+ := \mathcal{F}_i \cap \mathbb{R}_+^m$ ($i = 1, \dots, s$), $\mathcal{F}^+ := \bigtimes_{i=1}^s \mathcal{F}_i^+$. 为方便读者理解, 我们用图 2.1 表示一般 LPCC、MPCC、MPGCC 与问题 (2.5) 之间的关系.

问题 (2.5) 在多个领域中存在应用. 例如, 第 1 章中用于刻画强关联电子体系的 MMOT (1.18), 在类 Monge 拟设 (1.21) 下的离散问题 (1.26) 具有问题 (2.5) 的形式. 其中广义互补约束的存在排除了电子碰撞的可能. 当 s 等于 2 时, 问题

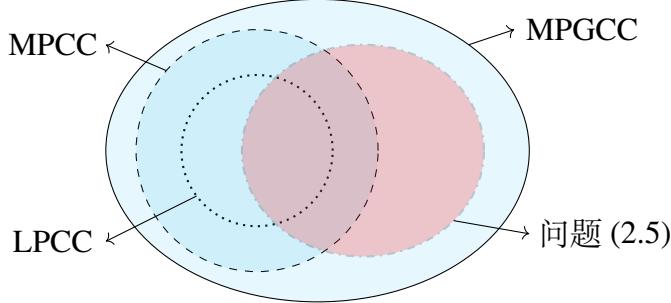


图 2.1 一般 LPCC、MPCC、MPGCC 与问题 (2.5) 之间的关系. 带实线边界的蓝色椭圆表示 MPGCC, 带虚线边界的蓝色圆盘表示 MPCC, 带点线边界的蓝色圆盘表示 LPCC, 带点划线边界的红色椭圆表示问题 (2.5)

Figure 2.1 The relations among the scopes of general LPCC, MPCC, MPGCC as well as problem (1.26). The blue ellipsoid with solid boundary stands for the scope of MPGCC, the larger blue disk with dashed boundary stands for the scope of MPCC, and the smaller blue disk with dotted boundary stands for the scope of LPCC. The red ellipsoid with dashdotted boundary refers to the scope of problem (2.5)

(2.5) 还可用于建模运输征税^[150–153]、生物燃料生产^[154]、航空工业^[155]与通信服务^[156–158]中的序列决策过程.

类似于问题 (2.4), 问题 (2.5) 的 ℓ_1 罚问题是

$$\min_{\mathbf{x}} f(\mathbf{x}) + \beta p(\mathbf{x}), \quad \text{s. t. } \mathbf{x} \in \mathcal{F}^+, \quad (2.6)$$

其等价于

$$\min_{\mathbf{x}} f_\beta(\mathbf{x}) := f(\mathbf{x}) + \beta \sum_{1 \leq i < j}^s \langle \mathbf{x}_i, \mathbf{x}_j \rangle, \quad \text{s. t. } \mathbf{x} \in \mathcal{F}^+, \quad i = 1, \dots, s. \quad (2.7)$$

之后, 我们仅研究问题 (2.7), 得到的理论结果对问题 (2.6) 也同样成立. 由 ℓ_1 罚函数的性质, f_β 仍然是多仿射的.

例 2.1. 对于强关联电子体系计算应用, MPGCC (1.26) 的 ℓ_1 罚问题是

$$\begin{aligned} \min_{\{Y_i\}_{i=2}^N} \quad & \sum_{i=2}^N \langle Y_i, \text{Diag}(\boldsymbol{\rho})C \rangle + \sum_{2 \leq i < j}^N \langle Y_i, \text{Diag}(\boldsymbol{\rho})Y_j C + \beta Y_j \rangle, \\ \text{s. t.} \quad & Y_i \mathbf{1}_K = \mathbf{1}_K, \quad Y_i^\top \boldsymbol{\rho} = \boldsymbol{\rho}, \quad \text{Tr}(Y_i) = 0, \quad Y_i \geq 0, \quad i = 2, \dots, N. \end{aligned}$$

由于线性系统满足 Hoffman 误差界^[159], 因此, 我们还可以将零迹约束惩罚至目标函数^[160], 得到如下罚问题:

$$\begin{aligned} \min_{\{Y_i\}_{i=2}^N} \quad & \sum_{i=2}^N \langle Y_i, \text{Diag}(\boldsymbol{\rho})C + \beta I_K \rangle + \sum_{2 \leq i < j}^N \langle Y_i, \text{Diag}(\boldsymbol{\rho})Y_j C + \beta Y_j \rangle, \\ \text{s. t.} \quad & Y_i \mathbf{1}_K = \mathbf{1}_K, \quad Y_i^\top \boldsymbol{\rho} = \boldsymbol{\rho}, \quad Y_i \geq 0, \quad i = 2, \dots, N. \end{aligned} \quad (2.8)$$

2.1.2 研究现状

我们介绍研究 MPGCC ℓ_1 罚函数精确性的已有工作。现有结果大体分为两类：一类针对一般的 MPCC，另一类针对问题 (2.5) 的特殊情形。

对于一般的 MPCC，已有学者在额外的正则性 (regularity) 假设下证明了 ℓ_1 罚函数的精确性。在文献^[136,148] 中，作者基于如下严格互补条件与误差界条件证明了精确性：

$$\text{严格互补: } \mathbf{x}_1 + \mathbf{x}_2 > 0, \quad \forall (\mathbf{x}_1, \mathbf{x}_2) \in \tilde{\mathcal{F}}, \quad (2.9)$$

$$\text{误差界: } \exists \tau > 0, \quad \text{s. t. } \text{dist}((\mathbf{x}_1, \mathbf{x}_2), \tilde{\mathcal{F}}) \leq \tau \langle \mathbf{x}_1, \mathbf{x}_2 \rangle, \quad \forall (\mathbf{x}_1, \mathbf{x}_2) \in \tilde{\mathcal{F}}^+, \quad (2.10)$$

其中 $\tilde{\mathcal{F}} \subseteq (\mathbb{R}^m)^2$ 与 $\tilde{\mathcal{F}}^+ \subseteq (\mathbb{R}^m)^2$ 分别代表 MPCC 与其 ℓ_1 罚问题的可行域。对任意向量 \mathbf{x} 与集合 \mathcal{A} ，距离 $\text{dist}(\mathbf{x}, \mathcal{A}) := \inf_{\mathbf{y} \in \mathcal{A}} \|\mathbf{x} - \mathbf{y}\|$ 。之后在文献^[161] 中，作者在假设所谓的正乘子非退化 (positive-multiplier nondegeneracy) 条件与罚问题的 Mangasarian-Fromovitz 约束规范条件成立时，证明了 ℓ_1 罚函数的精确性。然而，对于实际应用中的 MPCC，上述严格互补条件 (2.9) 可能过于严苛，而正乘子非退化条件是否成立往往难以验证。对于一般的 MPGCC，目前暂无工作研究其 ℓ_1 罚函数的理论性质。

还有一些工作研究了问题 (2.5) 的特殊情形：目标函数仿射且 s 等于 2。利用可行域极点集合的有限性，早期的一些工作^[150,152,162,163] 证明了当 β 充分大时，问题 (2.5) 的最优解一定是罚问题 (2.7) 的最优解，但没有考虑反之是否成立。而在文献^[164,165] 中，作者给出了 ℓ_1 罚函数精确性的完整证明。这些工作都假设问题 (2.5) 的目标函数是仿射的且只有两块变量。目前，暂无工作讨论目标函数非线性或变量块数大于 2 的情形。

2.1.3 本章主要内容

我们首先给出问题 (2.5) 的一个实例，其 ℓ_1 罚函数的精确性无法由已有结果推得。随后，我们充分利用问题 (2.5) 的内在代数与几何结构（如目标函数的多仿射性、多面体可行域的可分性等），在要求 ℓ_1 罚问题可行域非空紧的条件下，证明了 ℓ_1 罚函数的精确性。我们的结果可覆盖已有工作^[150,152,162–165]，且适用于目标函数非线性、变量块数大于 2 的情形。

2.2 已有理论结果不适用之例

本节给出问题 (2.5) 的一个实例，其来源于强关联电子体系的计算（可见问题 (1.26))。由于目标函数具有非线性性，该问题不是文献^[150,152,162–165] 中分析的 LPCC。此外，严格互补条件 (2.9) 与误差界条件 (2.10) 对该问题不成立，因此已有工作^[136,148] 中的理论结果并不适用。

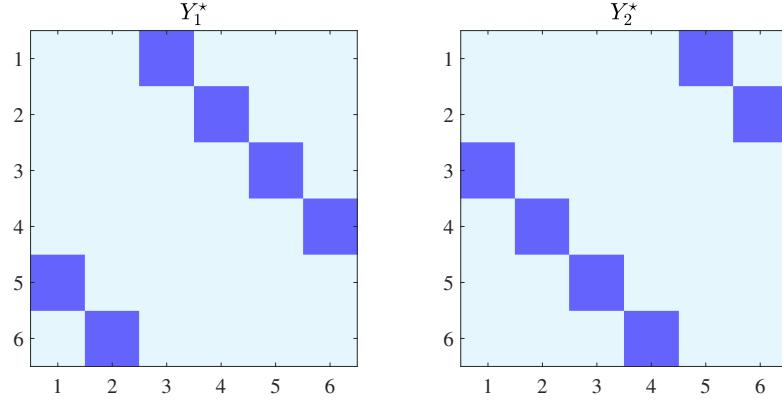


图 2.2 当 $K = 6$ 时, 问题 (2.11) 的一个最优解 (Y_1^*, Y_2^*) . 深紫色方块代表值为 1 的元素, 蓝色方块代表值为 0 的元素

Figure 2.2 The illustration of an optimal solution (Y_1^*, Y_2^*) to problem (2.11) when $K = 6$. Each deep purple block stands for the value of 1, while each blue block for the value of 0

我们考虑如下具有两块矩阵变量的 MPCC:

$$\begin{aligned} & \min_{Y_1, Y_2} \langle Y_1 + Y_2, C \rangle + \langle Y_1, Y_2 C \rangle, \\ \text{s. t. } & Y_i \mathbf{1}_K = \mathbf{1}_K, Y_i^\top \mathbf{1}_K = \mathbf{1}_K, \text{Tr}(Y_i) = 0, Y_i \geq 0, i = 1, 2, \\ & \langle Y_1, Y_2 \rangle = 0, \end{aligned} \quad (2.11)$$

其中 $K \in \mathbb{N}$, $Y_1, Y_2 \in \mathbb{R}^{K \times K}$, $C = (c_{ij}) \in \mathbb{R}^{K \times K}$ 定义为

$$c_{ij} = \begin{cases} K/|i-j|, & \text{若 } i \neq j; \\ 0, & \text{否则.} \end{cases}$$

它是问题 (1.25) 在 $[0, 1]$ 上的均匀离散, 对应 $[0, 1]$ 上的单电子密度为常数的三电子体系. 易见问题 (2.11) 具有问题 (2.5) 的形式. 根据文献^[100] 定理 1.1, 若 K 可被 3 整除, 我们可直接写出问题 (2.11) 的一个最优解 (Y_1^*, Y_2^*) : 对 $i, j = 1, \dots, K$,

$$\begin{aligned} y_{1,ij}^* &:= \begin{cases} 1, & \text{若 } i-j = 2K/3 \text{ 或 } j-i = K/3; \\ 0, & \text{否则,} \end{cases} \\ y_{2,ij}^* &:= \begin{cases} 1, & \text{若 } i-j = K/3 \text{ 或 } j-i = 2K/3; \\ 0, & \text{否则.} \end{cases} \end{aligned} \quad (2.12)$$

我们在图 2.2 中示出该最优解.

由于目标函数非线性, 问题 (2.11) 不是文献^[150,152,162–165] 中分析的 LPCC. 此外, 由于零迹约束的存在, 严格互补条件 (2.9) 在问题 (2.11) 的每个可行点处都不成立. 下面, 我们证明只要 $K > 3$ 可被 3 整除, 则误差界 (2.10) 在问题 (2.11) 上也不成立.

定理 2.1 (问题 (2.11) 不满足误差界 (2.10)). 若问题 (2.11) 满足 $K > 3$ 且 $\text{mod}(K, 3) = 0$, 则不存在 $\tau > 0$ 使得

$$\text{dist}((Y_1, Y_2), \mathcal{F}) \leq \tau \langle Y_1, Y_2 \rangle, \quad \forall (Y_1, Y_2) \in \mathcal{F}^+. \quad (2.13)$$

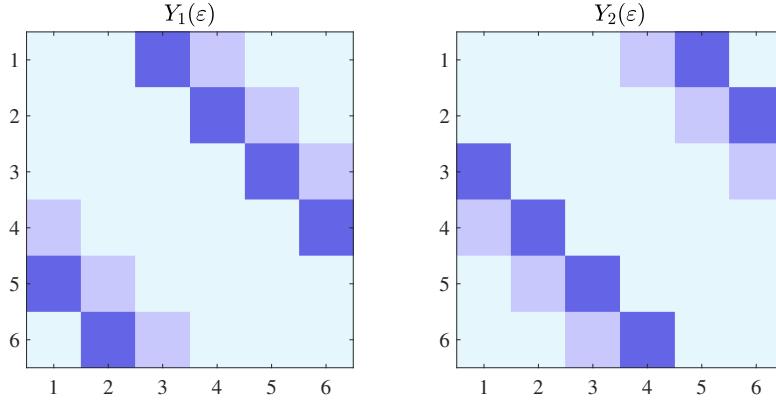


图 2.3 当 $K = 6$ 时, 构造的 ℓ_1 罚问题可行点 $(Y_1(\varepsilon), Y_2(\varepsilon))$. 深紫色方块代表值为 $1 - \varepsilon$ 的元素, 浅紫色方块代表值为 ε 的元素, 蓝色方块代表值为 0 的元素

Figure 2.3 The illustration of the constructed feasible point $(Y_1(\varepsilon), Y_2(\varepsilon))$ for the ℓ_1 penalty problem when $K = 6$. Each deep purple block stands for the value of $1 - \varepsilon$, each shallow purple block for the value of ε , while each blue block for the value of 0

证明. 根据最优解 (2.12), 对任一 $\varepsilon \in (0, 1)$, 我们构造 $(Y_1(\varepsilon), Y_2(\varepsilon))$, 其定义为

$$(Y_1(\varepsilon))_{ij} := \begin{cases} 1 - \varepsilon, & \text{若 } i - j = 2K/3 \text{ 或 } j - i = K/3; \\ \varepsilon, & \text{若 } i - j = 2K/3 - 1 \text{ 或 } j - i = K/3 + 1; \\ 0, & \text{否则,} \end{cases}$$

$$(Y_2(\varepsilon))_{ij} := \begin{cases} 1 - \varepsilon, & \text{若 } i - j = K/3 \text{ 或 } j - i = 2K/3; \\ \varepsilon, & \text{若 } i - j = 2K/3 - 1 \text{ 或 } j - i = K/3 + 1; \\ 0, & \text{否则.} \end{cases}$$

我们在图 2.3 中示出 $(Y_1(\varepsilon), Y_2(\varepsilon))$. 容易验证 $(Y_1(\varepsilon), Y_2(\varepsilon)) \in \mathcal{F}^+$, 但对任给 $\varepsilon \in (0, 1)$,

$$\begin{aligned} \langle Y_1(\varepsilon), Y_2(\varepsilon) \rangle &= \sum_{i,j} (Y_1(\varepsilon))_{ij} (Y_2(\varepsilon))_{ij} \\ &= \left(\sum_{i-j=2K/3-1} + \sum_{j-i=K/3+1} \right) (Y_1(\varepsilon))_{ij} (Y_2(\varepsilon))_{ij} = K\varepsilon^2, \end{aligned} \tag{2.14}$$

且随着 ε 趋于 0, $(Y_1(\varepsilon), Y_2(\varepsilon))$ 趋于 (Y_1^*, Y_2^*) .

若恰好有 $(Y_1^*, Y_2^*) \in \arg \min_{(Y_1, Y_2) \in \mathcal{F}} \|(Y_1, Y_2) - (Y_1(\varepsilon), Y_2(\varepsilon))\|$, 则

$$\text{dist}((Y_1(\varepsilon), Y_2(\varepsilon)), \mathcal{F}) = \|(Y_1^*, Y_2^*) - (Y_1(\varepsilon), Y_2(\varepsilon))\| = 2\sqrt{K}\varepsilon = \frac{2}{\sqrt{K\varepsilon}} \langle Y_1(\varepsilon), Y_2(\varepsilon) \rangle.$$

令 ε 趋于 0 即可知使 (2.13) 式成立的 τ 是不存在的.

若 $(Y_1^*, Y_2^*) \notin \arg \min_{(Y_1, Y_2) \in \mathcal{F}} \|(Y_1, Y_2) - (Y_1(\varepsilon), Y_2(\varepsilon))\|$, 我们下面证明

$$\limsup_{\varepsilon \rightarrow 0^+} \frac{\text{dist}((Y_1(\varepsilon), Y_2(\varepsilon)), \mathcal{F})}{\varepsilon} > 0, \tag{2.15}$$

从而与(2.14)式结合仍可导出使(2.13)式成立的 τ 是不存在的. 用反证法证明, 假设

$$\text{dist}((Y_1(\varepsilon), Y_2(\varepsilon)), \mathcal{F}) = \|(\tilde{Y}_1(\varepsilon), \tilde{Y}_2(\varepsilon)) - (Y_1(\varepsilon), Y_2(\varepsilon))\| = o(\varepsilon), \quad \varepsilon \rightarrow 0^+, \quad (2.16)$$

其中 $(\tilde{Y}_1(\varepsilon), \tilde{Y}_2(\varepsilon)) \in \mathcal{F}$. 由于 $Y_1(\varepsilon)$ 与 $Y_2(\varepsilon)$ 中的非零元素不是 $1 - \varepsilon$ 就是 ε , 且由(2.16)式可知

$$|(\tilde{Y}_i(\varepsilon))_{jk} - (Y_i(\varepsilon))_{jk}| = o(\varepsilon), \quad j, k = 1, \dots, K, \quad i = 1, 2,$$

于是 $\text{supp}(\tilde{Y}_i(\varepsilon)) \supseteq \text{supp}(Y_i(\varepsilon))$ ($i = 1, 2$). 注意到 $\tilde{Y}_1(\varepsilon), \tilde{Y}_2(\varepsilon) \geq 0$. 因此我们可从 $\langle Y_1(\varepsilon), Y_2(\varepsilon) \rangle > 0$ 得知 $\langle \tilde{Y}_1(\varepsilon), \tilde{Y}_2(\varepsilon) \rangle > 0$. 这与 $(\tilde{Y}_1(\varepsilon), \tilde{Y}_2(\varepsilon)) \in \mathcal{F}$ 矛盾. 从而(2.15)式成立.

综合上述两款, 我们完成了证明. \square

注. 利用文献^[100]定理1.1, 我们可推广定理2.1的结论至 s 不小于3的情形.

根据定理2.1, 现有文献^[136, 148, 150, 152, 162–165]中的结果不再适用于问题(2.11). 下面, 我们利用问题(2.5)的特殊结构, 证明其 ℓ_1 罚函数的精确性.

2.3 ℓ_1 罚函数精确性的证明

我们的证明依赖于下面的条件.

条件2.2. 集合 \mathcal{F}^+ 非空且紧.

相比于已有工作^[136, 148, 161]中的正则性假设, 条件2.2在许多实际问题中是自动满足的.

我们规定如下记号:

$$\begin{aligned} \text{ext}(\mathcal{F}_i^+) : & \mathcal{F}_i^+ \text{的极点集合}, & \text{ext}(\mathcal{F}^+) : & \mathcal{F}^+ \text{的极点集合}, \\ \mathcal{S}^{\text{opt}} : & \text{问题(2.5)的最优解集}, & \mathcal{S}_\beta^{\text{opt}} : & \text{问题(2.7)的最优解集}, \\ \tilde{\mathcal{S}}^{\text{opt}} : & \text{问题(2.5)的极点最优解集}, & \tilde{\mathcal{S}}_\beta^{\text{opt}} : & \text{问题(2.7)的极点最优解集}. \end{aligned}$$

我们首先证明罚问题(2.7)必定有极点最优解.

引理2.3 (罚问题可行域极点集合非空有限). 对 $i = 1, \dots, s$, 集合 \mathcal{F}_i^+ 满足 $0 < |\text{ext}(\mathcal{F}_i^+)| < \infty$. 进一步地, 罚问题(2.7)可行域 \mathcal{F}^+ 满足 $0 < |\text{ext}(\mathcal{F}^+)| < \infty$.

证明. 对 $i = 1, \dots, s$, 根据文献^[166]定理2.6可证明 \mathcal{F}_i^+ 极点集合非空, 根据文献^[21]推论19.1.1可证明 \mathcal{F}_i^+ 极点集合有限. 由于 \mathcal{F}^+ 可分, 因此其极点集合也非空有限. 证毕. \square

引理2.4 (罚问题存在极点最优解). 假设条件2.2成立. 则对任意 $\beta \in \mathbb{R}$, $\tilde{\mathcal{S}}_\beta^{\text{opt}} \neq \emptyset$.

证明. 由条件 2.2, 对任意 $\beta \in \mathbb{R}$, $S_\beta^{\text{opt}} \neq \emptyset$. 任取 $\hat{\mathbf{x}} := (\hat{\mathbf{x}}_1, \dots, \hat{\mathbf{x}}_s) \in S_\beta^{\text{opt}}$. 若恰巧 $\hat{\mathbf{x}} \in \text{ext}(\mathcal{F}^+)$, 则得证. 否则, 考虑线性规划

$$\min_{\mathbf{x}_1} f_\beta(\mathbf{x}_1, \hat{\mathbf{x}}_2, \dots, \hat{\mathbf{x}}_s), \quad \text{s. t. } \mathbf{x}_1 \in \mathcal{F}_1^+, \quad (2.17)$$

并令 $\mathbf{x}_1^\# \in \text{ext}(\mathcal{F}_1^+)$ 为其极点最优解. $\mathbf{x}_1^\#$ 的存在性可由引理 2.3 得到. 由 $\mathbf{x}_1^\#$ 的最优性以及 $\hat{\mathbf{x}}_1$ 的可行性, 可知 $f_\beta(\mathbf{x}_1^\#, \hat{\mathbf{x}}_2, \dots, \hat{\mathbf{x}}_s) \leq f_\beta(\hat{\mathbf{x}}_1, \dots, \hat{\mathbf{x}}_s)$. 随后, 对 $i = 2, \dots, s$, 依次考虑如下线性规划:

$$\min_{\mathbf{x}_i} f_\beta(\mathbf{x}_1^\#, \dots, \mathbf{x}_{i-1}^\#, \mathbf{x}_i, \hat{\mathbf{x}}_{i+1}, \dots, \hat{\mathbf{x}}_s), \quad \text{s. t. } \mathbf{x}_i \in \mathcal{F}_i^+,$$

并令 $\mathbf{x}_i^\# \in \text{ext}(\mathcal{F}_i^+)$ 为其极点最优解. 类似地, 对 $i = 2, \dots, s$, 有

$$f_\beta(\mathbf{x}_1^\#, \dots, \mathbf{x}_{i-1}^\#, \mathbf{x}_i^\#, \hat{\mathbf{x}}_{i+1}, \dots, \hat{\mathbf{x}}_s) \leq f_\beta(\mathbf{x}_1^\#, \dots, \mathbf{x}_{i-1}^\#, \hat{\mathbf{x}}_i, \hat{\mathbf{x}}_{i+1}, \dots, \hat{\mathbf{x}}_s).$$

最后, 我们有不等式

$$f_\beta(\mathbf{x}_1^\#, \dots, \mathbf{x}_s^\#) \leq f_\beta(\mathbf{x}_1^\#, \dots, \mathbf{x}_{s-1}^\#, \hat{\mathbf{x}}_s) \leq \dots \leq f_\beta(\mathbf{x}_1^\#, \hat{\mathbf{x}}_2, \dots, \hat{\mathbf{x}}_s) \leq f_\beta(\hat{\mathbf{x}}_1, \dots, \hat{\mathbf{x}}_s).$$

由 \mathcal{F}^+ 的可分性可知 $\text{ext}(\mathcal{F}^+) = \bigtimes_{i=1}^s \text{ext}(\mathcal{F}_i^+)$ 且 $\mathbf{x}^\# := (\mathbf{x}_1^\#, \dots, \mathbf{x}_s^\#) \in \text{ext}(\mathcal{F}^+)$. 另外, 因为 $\hat{\mathbf{x}}$ 是罚问题 (2.7) 的一个最优解, 所以由上面的不等式可得 $\mathbf{x}^\# \in S_\beta^{\text{opt}}$. 证毕. \square

对于问题 (2.5) 与其罚问题 (2.7), 容易证明如下引理.

引理 2.5 (不可行罚问题最优解对应罚项非零). 若对某个 $\beta \in \mathbb{R}$, $\mathbf{x} \in S_\beta^{\text{opt}} \setminus S^{\text{opt}}$, 则 $p(\mathbf{x}) > 0$.

证明. 如若不然, 假设 $p(\mathbf{x}) = 0$, 则 $\mathbf{x} \in \mathcal{F}$. 我们证明 $\mathbf{x} \in S^{\text{opt}}$, 从而导出矛盾: 对任意 $\tilde{\mathbf{x}} \in \mathcal{F}$,

$$f(\tilde{\mathbf{x}}) = f_\beta(\tilde{\mathbf{x}}) \geq f_\beta(\mathbf{x}) = f(\mathbf{x}) + \beta p(\mathbf{x}) = f(\mathbf{x}).$$

其中不等号是因为 $\mathbf{x} \in S_\beta^{\text{opt}}$. 证毕. \square

利用引理 2.3、2.4 与 2.5, 我们可证明 ℓ_1 罚函数的部分精确性.

命题 2.6 (罚函数的部分精确性). 假设条件 2.2 成立. 则存在 $\bar{\beta} > 0$, 使得对任意 $\beta \geq \bar{\beta}$, 有包含关系 $\tilde{S}_\beta^{\text{opt}} \subseteq S^{\text{opt}} \subseteq S_\beta^{\text{opt}}$.

证明. 我们先用反证法证明当 β 充分大时, $\tilde{S}_\beta^{\text{opt}} \subseteq S^{\text{opt}}$. 假设不然, 存在 $\{\beta_k\}$ 趋于无穷大以及对应的 $\{\mathbf{x}^{(k)} \in \tilde{S}_{\beta_k}^{\text{opt}} \setminus S^{\text{opt}}\}$. 根据引理 2.5, 对任意 $k \in \mathbb{N}$, 有 $p(\mathbf{x}^{(k)}) > 0$. 从引理 2.3 可知, \mathcal{F} 的极点集合非空有限, 于是

$$\inf_k f(\mathbf{x}^{(k)}) > -\infty, \quad \inf_k p(\mathbf{x}^{(k)}) > 0. \quad (2.18)$$

任取问题(2.5)的可行点 \mathbf{x} , 对任意 $k \in \mathbb{N}$,

$$\begin{aligned}\infty > f(\mathbf{x}) &= f(\mathbf{x}) + \beta_k p(\mathbf{x}) \geq f_{\beta_k}(\mathbf{x}^{(k)}) \\ &= f(\mathbf{x}^{(k)}) + \beta_k p(\mathbf{x}^{(k)}) \\ &\geq \inf_l f(\mathbf{x}^{(l)}) + \beta_k \inf_l p(\mathbf{x}^{(l)}).\end{aligned}$$

基于(2.18)式, 若令上述不等式右端 k 趋于无穷大, 会导出矛盾. 因此, 存在 $\bar{\beta} > 0$, 使得对任意 $\beta \geq \bar{\beta}$, 有包含关系 $\tilde{S}_\beta^{\text{opt}} \subseteq S^{\text{opt}}$.

我们接着证明当 $\beta \geq \bar{\beta}$ 时, $S^{\text{opt}} \subseteq S_\beta^{\text{opt}}$, 也即对任意 $\mathbf{x}^* \in S^{\text{opt}}$ 与 $\mathbf{x} \in \mathcal{F}^+$, $f_\beta(\mathbf{x}^*) \leq f_\beta(\mathbf{x})$. 由引理2.4, $\tilde{S}_\beta^{\text{opt}} \neq \emptyset$. 任取 $\hat{\mathbf{x}} \in \tilde{S}_\beta^{\text{opt}}$. 由前一段的证明可知 $\hat{\mathbf{x}} \in S^{\text{opt}}$. 因此

$$f_\beta(\mathbf{x}^*) = f(\mathbf{x}^*) = f(\hat{\mathbf{x}}) = f_\beta(\hat{\mathbf{x}}).$$

由于 $\hat{\mathbf{x}} \in S_\beta^{\text{opt}}$, 因此 $f_\beta(\hat{\mathbf{x}}) \leq f_\beta(\mathbf{x})$. 结合上面的等式即得证. \square

为证明罚函数的精确性, 我们只需证明反方向: $S_\beta^{\text{opt}} \subseteq S^{\text{opt}}$. 我们先证明如下引理, 其表明若罚问题最优解只有一个变量块不是对应可行域的极点, 则它必定是原问题的最优解.

引理2.7 (单变量块非极点不影响最优化). 假设条件2.2成立. 对任意 $\beta \geq \bar{\beta}$, 令 $\hat{\mathbf{x}} := (\hat{\mathbf{x}}_1, \dots, \hat{\mathbf{x}}_s) \in S_\beta^{\text{opt}}$. 若仅存在一个 $i \in \{1, \dots, s\}$ 使得 $\hat{\mathbf{x}}_i \notin \text{ext}(\mathcal{F}_i^+)$, 则 $\hat{\mathbf{x}} \in S^{\text{opt}}$.

证明. 不失一般性, 假定 $i = 1$. 我们用反证法证明. 假设 $\hat{\mathbf{x}} \notin S^{\text{opt}}$. 则由引理2.5可知 $p(\hat{\mathbf{x}}) > 0$. 由于 $\hat{\mathbf{x}} \in S_\beta^{\text{opt}}$, 因此 $\hat{\mathbf{x}}_1$ 是线性规划(2.17)的最优解, 且 $\hat{\mathbf{x}}_1 \in \text{rbd}(\mathcal{F}_1^+)$. 因为 $\hat{\mathbf{x}}_1 \notin \text{ext}(\mathcal{F}_1^+)$, 所以线性规划(2.17)在 \mathcal{F}_1^+ 的某个面上是退化的. 记这个面为 Ω_1 , 其上的每个点都是线性规划(2.17)的最优解. 特别地, 根据命题2.6, Ω_1 的每个极点与 $\{\hat{\mathbf{x}}_j\}_{j=2}^s$ 合并后都是问题(2.5)的最优解. 记 Ω_1 的所有极点为 $\{\bar{\mathbf{x}}_1^{(l)}\}_{l \in I_1}$, 其中 $I_1 \subseteq \mathbb{N}$ 是一个有限指标集. 因为 $\hat{\mathbf{x}}_1 \notin \text{ext}(\mathcal{F}_1^+)$, $|I_1| \geq 2$. 由条件2.2, Ω_1 是紧凸集. 根据Minkowski定理^[21], Ω_1 是 $\{\bar{\mathbf{x}}_1^{(l)}\}_{l \in I_1}$ 的凸包.

任取 I_1 中的两个元素(例如 $l = 1, 2$). 则

$$(\bar{\mathbf{x}}_1^{(l)}, \hat{\mathbf{x}}_2, \dots, \hat{\mathbf{x}}_s) \in S^{\text{opt}}, \quad p(\bar{\mathbf{x}}_1^{(l)}, \hat{\mathbf{x}}_2, \dots, \hat{\mathbf{x}}_s) = 0, \quad l = 1, 2.$$

因为 p 对第一个变量块是线性的, 因此对 $\bar{\mathbf{x}}_1^{(1)}$ 与 $\bar{\mathbf{x}}_1^{(2)}$ 连线上的任意 \mathbf{x}_1 , 都有 $p(\mathbf{x}_1, \hat{\mathbf{x}}_2, \dots, \hat{\mathbf{x}}_s) = 0$.

重复前一段的推导过程, 可知在 $\text{rbd}(\Omega_1) \times \{\hat{\mathbf{x}}_2\} \times \dots \times \{\hat{\mathbf{x}}_s\}$ 上 $p \equiv 0$. 若 $\hat{\mathbf{x}}_1 \in \text{rbd}(\Omega_1)$ 即可导出矛盾. 否则, 因为 Ω_1 紧凸, $\hat{\mathbf{x}}_1$ 一定位于 $\text{rbd}(\Omega_1)$ 中某两个点的连线上. 此时, $\hat{\mathbf{x}}$ 为第一个变量块在 $\text{rbd}(\Omega_1)$ 中的向量的线性组合. 再次地, 因为 p 对第一个变量块是线性的, 可知 $p(\hat{\mathbf{x}}) = 0$, 导出矛盾. 证毕. \square

下面, 我们用数学归纳法证明 ℓ_1 罚函数的精确性.

定理 2.8 (罚函数的精确性). 假设条件 2.2 成立. 则对任意 $\beta \geq \bar{\beta}$, $S_\beta^{\text{opt}} = S^{\text{opt}}$.

证明. 包含关系 $S^{\text{opt}} \subseteq S_\beta^{\text{opt}}$ 已在命题 2.6 中得到证明. 下面, 我们用数学归纳法证明如下论断: 对 $r \in \{0, 1, \dots, s\}$, 若 $\hat{\mathbf{x}} \in S_\beta^{\text{opt}}$ 只有 r 个指标 $\{i_1, \dots, i_r\} \subseteq \{1, \dots, s\}$ 使得 $\hat{\mathbf{x}}_{i_j} \notin \text{ext}(\mathcal{F}_{i_j}^+)$ ($j = 1, \dots, r$), 则 $\hat{\mathbf{x}} \in S^{\text{opt}}$. 易见, 若此论断成立, 则相反包含关系得证. 命题 2.6 与引理 2.7 分别已经证明了 $r = 0$ 与 $r = 1$ 的情形. 下面, 假设论断对 $r = t$ 成立 ($t \in \{1, \dots, s-1\}$), 我们用反证法证明 $r = t+1$ 的情形.

不失一般性, 考虑 $i_j = j$ ($j = 1, \dots, t+1$). 假设 $\hat{\mathbf{x}} \notin S^{\text{opt}}$. 则由引理 2.5 可知 $p(\hat{\mathbf{x}}) > 0$. 由于 $\hat{\mathbf{x}} \in S_\beta^{\text{opt}}$, $\hat{\mathbf{x}}_{t+1} \in \text{rbd}(\mathcal{F}_{t+1}^+)$ 是如下线性规划的最优解:

$$\min_{\mathbf{x}_{t+1}} f_\beta(\hat{\mathbf{x}}_1, \dots, \hat{\mathbf{x}}_t, \mathbf{x}_{t+1}, \hat{\mathbf{x}}_{t+2}, \dots, \hat{\mathbf{x}}_s), \quad \text{s. t. } \mathbf{x}_{t+1} \in \mathcal{F}_{t+1}^+. \quad (2.19)$$

因为 $\hat{\mathbf{x}}_{t+1} \notin \text{ext}(\mathcal{F}_{t+1}^+)$, 线性规划 (2.19) 在 \mathcal{F}_{t+1}^+ 的某个面上是退化的. 记这个面为 Ω_{t+1} , 其上的每个点都是线性规划 (2.19) 的最优解. 由目标函数值的最优化, Ω_{t+1} 的每个极点与 $\{\hat{\mathbf{x}}_j\}_{j \neq t+1}$ 合并后都是罚问题 (2.7) 的最优解. 记 Ω_{t+1} 的所有极点为 $\{\bar{\mathbf{x}}_{t+1}^{(l)}\}_{l \in \mathcal{I}_{t+1}}$, 其中 $\mathcal{I}_{t+1} \subseteq \mathbb{N}$ 是一个有限指标集, 并记

$$\bar{\mathbf{x}}^{(l)} := (\hat{\mathbf{x}}_1, \dots, \hat{\mathbf{x}}_t, \bar{\mathbf{x}}_{t+1}^{(l)}, \hat{\mathbf{x}}_{t+2}, \dots, \hat{\mathbf{x}}_s), \quad l \in \mathcal{I}_{t+1}.$$

则 $\{\bar{\mathbf{x}}^{(l)}\}_{l \in \mathcal{I}_{t+1}} \subseteq S_\beta^{\text{opt}}$. 对每个 $l \in \mathcal{I}_{t+1}$, 注意到在 $\bar{\mathbf{x}}^{(l)}$ 中只有 t 个变量块不是对应可行域的极点. 由归纳假设, 就有 $\{\bar{\mathbf{x}}^{(l)}\}_{l \in \mathcal{I}_{t+1}} \subseteq S^{\text{opt}}$, 从而 $p(\bar{\mathbf{x}}^{(l)}) = 0$ ($l \in \mathcal{I}_{t+1}$). 经过类似于引理 2.7 证明中的推导, 即可导出矛盾. 因此, 论断在 $r = t+1$ 时也成立. 定理得证. \square

2.4 本章小结

在本章中, 我们证明了 MPGCC (2.5) ℓ_1 罚函数的精确性. 强关联电子体系计算中的 MPGCC (1.26) 是 MPGCC (2.5) 的特例. 无需已有工作中的严格互补或正乘子非退化等假设条件, 我们充分利用了问题的特殊代数与几何结构, 包括目标函数的多仿射性与多面体可行域的可分性, 证明了在 β 充分大时, $S^{\text{opt}} = S_\beta^{\text{opt}}$. 我们取得的结果覆盖已有关于 LPCC 的工作, 且适用于目标函数非线性、变量块数大于 2 的情形. 我们还给出了 MPGCC (2.5) 的一个实例, 对其已有理论结果不适用, 而我们的结果则可保证其 ℓ_1 罚函数的精确性. 本章内容可为后文优化算法的设计与应用提供优化模型的理论基础.

此外, 本章中的结果可推广至一类多仿射优化问题:

$$\min_{\mathbf{x}} \tilde{f}(\mathbf{x}_1, \dots, \mathbf{x}_s), \quad \text{s. t. } \mathbf{h}(\mathbf{x}_1, \dots, \mathbf{x}_s) = 0, \quad \mathbf{x}_i \in \tilde{\mathcal{F}}_i, \quad i = 1, \dots, s,$$

其中 $\mathbf{x} := (\mathbf{x}_1, \dots, \mathbf{x}_s)$, $\tilde{f} : (\mathbb{R}^m)^s \rightarrow \mathbb{R}$ 与 $\mathbf{h} : (\mathbb{R}^m)^s \rightarrow \mathbb{R}^q$ 是多仿射函数 ($q \in \mathbb{N}$), $\{\tilde{\mathcal{F}}_i\}_{i=1}^s$ 是多面体. 在 $\bigtimes_{i=1}^s \tilde{\mathcal{F}}_i$ 上, \mathbf{h} 非负. 上述多仿射优化问题对应的 ℓ_1 罚问题为

$$\min_{\mathbf{x}} \tilde{f}_\beta(\mathbf{x}) := \tilde{f}(\mathbf{x}) + \beta \mathbf{h}(\mathbf{x})^\top \mathbf{1}_q, \quad \text{s. t. } \mathbf{x} \in \bigtimes_{i=1}^s \tilde{\mathcal{F}}_i.$$

关于一般多仿射优化问题的理论与算法,感兴趣的读者可参阅文献^[167,168].

第3章 求解具有块状结构优化问题的 不可行非精确邻近交替线性化极小化算法

在本章中, 我们考虑具有块状结构的优化问题. 强关联电子计算中的问题 (2.8) 是此类问题的特例. 我们为其设计了可实现的不可行非精确邻近交替线性化极小化 (PALM-I) 算法, 并研究其理论性质, 包括全局依子(点)列收敛性、渐进收敛速度等. 我们还在数值实验中 (1) 对比了可行的 PALM 算法与 PALM-I 算法的效率; (2) 将 PALM-I 算法嵌入了一个瀑布型多重网格优化 (CMGOPT) 框架, 通过求解问题 (2.8), 模拟了一维、二维强关联电子体系.

3.1 问题描述与研究现状

3.1.1 问题描述

我们考虑如下具有块状结构的优化问题:

$$\min_{\mathbf{x}} f(\mathbf{x}_1, \dots, \mathbf{x}_s), \quad \text{s. t. } \mathbf{x}_i \in \mathcal{F}_i := \{\mathbf{y}_i \in \mathbb{R}^{m_i} : \mathbf{g}_i(\mathbf{y}_i) \geq 0\}, \quad i = 1, \dots, s, \quad (3.1)$$

其中 $\mathbf{x} := (\mathbf{x}_1, \dots, \mathbf{x}_s)$ ($s \in \mathbb{N}$), $\mathbf{x}_i \in \mathbb{R}^{m_i}$ ($m_i \in \mathbb{N}$), \mathcal{F}_i 为 \mathbf{x}_i 的可行域 ($i = 1, \dots, s$), $f : \times_{i=1}^s \mathbb{R}^{m_i} \rightarrow \mathbb{R}$ 可微未必凸, $\mathbf{g}_i : \mathbb{R}^{m_i} \rightarrow \mathbb{R}^{p_i}$ 凹可微 ($p_i \in \mathbb{N}$), $i = 1, \dots, s$. 我们还可以将问题 (3.1) 写做如下拓展实值形式:

$$\min_{\mathbf{x}} F(\mathbf{x}_1, \dots, \mathbf{x}_s) := f(\mathbf{x}_1, \dots, \mathbf{x}_s) + \sum_{i=1}^s \delta_{\mathcal{F}_i}(\mathbf{x}_i).$$

问题 (3.1) 在弹性力学^[169]、信号处理^[170]、图像去噪^[171]、拓扑优化^[172] 等领域具有广泛应用. 特别地, 用于刻画强关联电子体系的问题 (2.8) 具有问题 (3.1) 的形式.

3.1.2 邻近交替线性化极小化算法

考虑到问题 (3.1) 具有块状结构, 我们可使用 PALM 算法^[173] 求解之. PALM 算法的一般框架可见算法 3.1, 其中

$$\mathbf{x}_{<i} := (\mathbf{x}_1, \dots, \mathbf{x}_i), \quad \mathbf{x}_{\geq i} := (\mathbf{x}_i, \dots, \mathbf{x}_s), \quad (3.2)$$

子问题的求解需满足特定的条件.

3.1.3 研究现状

当子问题 (3.3) 被精确求解时, 我们称此时的算法为精确 PALM (exact PALM, PALM-E) 算法. 在选取合适的邻近参数后, 我们可以导出目标函数值序列的充分下降性以及稳定性违反度的上界, 进而证明算法的全局依子列收敛性. 这一套

算法 3.1: 求解问题 (3.1) 的 PALM 算法.

输入: 初始点 $\mathbf{x}^{(0)} := (\mathbf{x}_1^{(0)}, \dots, \mathbf{x}_s^{(0)}) \in \bigtimes_{i=1}^s \mathbb{R}^{m_i}$.

1 置 $k := 0$.

2 **while** 终止准则未满足 **do**

3 **for** $i = 1, \dots, s$ **do**

4 选取邻近参数 $\sigma_{i,k} > 0$.

5 求解第 i 个邻近线性化子问题

$$\min_{\mathbf{x}_i \in \mathcal{F}_i} \left\langle \nabla_{\mathbf{x}_i} f \left(\mathbf{x}_{<i}^{(k+1)}, \mathbf{x}_{\geq i}^{(k)} \right), \mathbf{x}_i - \mathbf{x}_i^{(k)} \right\rangle + \frac{\sigma_{i,k}}{2} \left\| \mathbf{x}_i - \mathbf{x}_i^{(k)} \right\|^2 \quad (3.3)$$

直至满足特定条件, 得到 $\mathbf{x}_i^{(k+1)} \in \mathbb{R}^{m_i}$.

6 **end**

7 置 $k := k + 1$.

8 **end**

输出: $\mathbf{x}^{(k)} := (\mathbf{x}_1^{(k)}, \dots, \mathbf{x}_s^{(k)}) \in \bigtimes_{i=1}^s \mathbb{R}^{m_i}$.

分析框架还可用于更加一般的算法, 例如分块连续极小化 (block successive minimization) 算法^[174]、基于 Bregman 距离的块坐标邻近梯度法 (Bregman distance-based block coordinate proximal gradient methods)^[175] 等. 此外, 若问题中的函数还满足 Łojasiewicz 性质, 我们还可以证得算法的全局依点列收敛性^[173,176].

不难看出, 求解算法 3.1 中的子问题 (3.3) 等价于计算

$$\tilde{\mathbf{x}}_i^{(k)} := \mathbf{x}_i^{(k)} - \frac{1}{\sigma_{i,k}} \nabla_{\mathbf{x}_i} f \left(\mathbf{x}_{<i}^{(k+1)}, \mathbf{x}_{\geq i}^{(k)} \right) \quad (3.4)$$

到可行域 \mathcal{F}_i 上的正交投影. 然而, 在大多数情形下, 该投影不具有显式表达式. 此时, 人们只能非精确地求解子问题 (3.3).

当子问题 (3.3) 被非精确求解且迭代点可行 (即 $\mathbf{x}_i^{(k)} \in \mathcal{F}_i$) 时, 我们称此时的算法为可行非精确 PALM (feasible inexact PALM, PALM-F) 算法. 目前, 绝大多数相关工作在分析其理论性质时仍需要目标函数值序列的单调性. 一些工作通过非精确求解子问题 (3.3) 获得下降方向, 进而做线搜索^[171,177]. 在文献^[175]中, 作者将子问题非精确求解的误差视为诱导 Bregman 距离核函数中的额外项, 从而借助 PALM-E 算法的理论设置子问题的非精确求解准则. 在文献^[178,179]中, 作者将证明 PALM-E 算法收敛性的工具直接作为子问题非精确求解准则, 只保留目标函数值序列的充分下降性, 而允许在稳定性违反度上界中存在误差.

当子问题 (3.3) 被非精确求解且迭代点不可行 (即 $\mathbf{x}_i^{(k)} \notin \mathcal{F}_i$) 时, 我们称此时的算法为不可行非精确 PALM 算法, 即 PALM-I 算法. 相较于可行情形, 研究 PALM-I 算法理论性质的工作十分稀少. 然而, 迭代点不可行的情况在实际应用中是十分常见的. 实际上, 只要子问题过于欠定 (under-determined), 我们就会使用

不可行方法(例如对偶方法、罚方法等^[20])求解之. 下面, 我们举两个例子进行说明.

例 3.1(线性约束). 可行域 \mathcal{F}_i 是 Birkhoff 多胞体 (polytope):

$$\mathcal{F}_i := \left\{ W \in \mathbb{R}^{m_i \times m_i} : W\mathbf{1}_{m_i} = \mathbf{1}_{m_i}, W^\top \mathbf{1}_{m_i} = \mathbf{1}_{m_i}, W \geq 0 \right\}.$$

此类可行域可见于最优运输问题^[96]、强关联电子体系计算^[91]等(参见问题(2.8)). 由于当 m_i 较大时, 描述 \mathcal{F}_i 的等式约束数量远少于变量的个数, 因此从对偶的角度求解子问题 (3.3) 更加合理与高效. 我们可以使用文献^[180] 中提出的半光滑 Newton 法求解对偶问题. 然而, 由此获得的迭代点是不可行的.

例 3.2(非线性约束). 可行域 \mathcal{F}_i 是 \mathbb{R}^{m_i} 中的椭球 (ellipsoid):

$$\mathcal{F}_i := \left\{ \mathbf{w} \in \mathbb{R}^{m_i} : \frac{1}{2}\mathbf{w}^\top A_i \mathbf{w} + \mathbf{b}_i^\top \mathbf{w} \leq \alpha_i \right\},$$

其中 $I_{m_i} \neq A_i \in \mathbb{R}^{m_i \times m_i}$ 是对称正定矩阵, $\mathbf{b}_i \in \mathbb{R}^{m_i}$, $\alpha_i > 0$. 计算到椭球上的投影是凸分析中的基本问题, 可应用于弹性力学^[169]、信号处理^[170]、拓扑优化^[172]等领域. 当 $\mathbf{b}_i = 0$ 时, 它还与非线性优化中的信赖域子问题相关^[181]. 最近, 有学者为其设计了交替方向乘子法(alternating direction method of multipliers, ADMM)^[182], 得到的原始解未必可行. 然而, 该不可行方法的效率却显著优于已有的可行混合投影(hybrid projection, HP) 方法^[183].

由于迭代点不可行, 我们无法保证目标函数值序列的单调性. 这一性质是绝大多数已有工作分析的关键. 目前, 仅有项工作考虑了 PALM-I 算法的理论性质^[178]. 作者对子问题的非精确求解提出了如下条件: 存在 $\beta_1, \beta_2 > 0$, 使得对 $i = 1, \dots, s$ 和 $k \geq 0$,

$$\begin{cases} \sum_{j=1}^{i-1} \|\mathbf{x}_j^{(k+1)} - \bar{\mathbf{x}}_j^{(k+1)}\| + \sum_{j=i}^s \|\mathbf{x}_j^{(k)} - \bar{\mathbf{x}}_j^{(k)}\| \leq \beta_1 \|\bar{\mathbf{x}}_i^{(k+1)} - \bar{\mathbf{x}}_i^{(k)}\|; \\ \left\langle \mathbf{x}_i^{(k)} - \bar{\mathbf{x}}_i^{(k)}, \bar{\mathbf{x}}_i^{(k+1)} - \bar{\mathbf{x}}_i^{(k)} \right\rangle \leq \beta_2 \left\| \bar{\mathbf{x}}_i^{(k+1)} - \bar{\mathbf{x}}_i^{(k)} \right\|^2, \end{cases} \quad (3.5)$$

其中 $\bar{\mathbf{x}}_i^{(k+1)}$ 是子问题 (3.3) 的唯一解:

$$\bar{\mathbf{x}}_i^{(k+1)} := \arg \min_{\mathbf{x}_i \in \mathcal{F}_i} \left\langle \nabla_{\mathbf{x}_i} f \left(\mathbf{x}_{<i}^{(k+1)}, \mathbf{x}_{\geq i}^{(k)} \right), \mathbf{x}_i - \mathbf{x}_i^{(k)} \right\rangle + \frac{\sigma_{i,k}}{2} \left\| \mathbf{x}_i - \mathbf{x}_i^{(k)} \right\|^2. \quad (3.6)$$

基于条件 (3.5), 作者证明了目标函数值序列 $\{f(\bar{\mathbf{x}}^{(k)})\}$ 上的充分下降性, 其中 $\bar{\mathbf{x}}^{(k)} := (\bar{\mathbf{x}}_1^{(k)}, \dots, \bar{\mathbf{x}}_s^{(k)})$. 然而, 条件 (3.5) 在实际计算中无法验证. 这是因为 $\bar{\mathbf{x}}_i^{(k)}$ 、 $\bar{\mathbf{x}}_i^{(k+1)}$ 以及 $\|\bar{\mathbf{x}}_i^{(k+1)} - \bar{\mathbf{x}}_i^{(k)}\|$ 都无法计算. 在文献^[178] 中, 作者并未探讨条件 (3.5) 在实际计算中如何满足. 因此, PALM-I 算法在可实现条件下的理论性质依然成谜.

算法 3.2: 求解问题 (3.1) 的 PALM-I 算法.

输入: 初始点 $\mathbf{x}^{(0)} := (\mathbf{x}_1^{(0)}, \dots, \mathbf{x}_s^{(0)}) \in \bigtimes_{i=1}^s \mathbb{R}^{m_i}$, 误差控制序列 $\{0 \leq \varepsilon_k \leq \bar{\varepsilon}\}$
 $(\bar{\varepsilon} \in (0, 1))$, 邻近参数上下界 $\sigma_{\max} \geq \sigma_{\min} > 0$.

```

1 置  $k := 0$ .
2 while 终止准则未满足 do
3   for  $i = 1, \dots, s$  do
4     选取邻近参数  $\sigma_{i,k} \in [\sigma_{\min}, \sigma_{\max}]$ .
5     求解第  $i$  个邻近线性化子问题
          
$$\min_{\mathbf{x}_i \in \mathcal{F}_i} \left\langle \nabla_{\mathbf{x}_i} f \left( \mathbf{x}_{<i}^{(k+1)}, \mathbf{x}_{\geq i}^{(k)} \right), \mathbf{x}_i - \mathbf{x}_i^{(k)} \right\rangle + \frac{\sigma_{i,k}}{2} \left\| \mathbf{x}_i - \mathbf{x}_i^{(k)} \right\|^2, \quad (3.7)$$

          得到  $\mathbf{x}_i^{(k+1)} \in \mathbb{R}^{m_i}$ , 使得存在  $\lambda_i^{(k+1)} \in \mathbb{R}_+^{p_i}$  满足
          
$$\sqrt{r_i \left( \mathbf{x}_i^{(k+1)}, \lambda_i^{(k+1)}, \tilde{\mathbf{x}}_i^{(k)} \right)} \leq \varepsilon_k, \quad (3.8)$$

          其中  $\tilde{\mathbf{x}}_i^{(k)}$  由 (3.4) 式定义.
6   end
7 end
输出:  $\mathbf{x}^{(k)} := (\mathbf{x}_1^{(k)}, \dots, \mathbf{x}_s^{(0)}) \in \bigtimes_{i=1}^s \mathbb{R}^{m_i}$ .
```

3.1.4 本章主要内容

在本章中, 我们首次在可实现的条件下, 建立了 PALM-I 算法的理论性质, 包括全局依子(点)列收敛性、渐进收敛速度等. 具体地, 我们首先为子问题 (3.3) 的非精确求解设计了可实现的终止准则, 随后基于子问题的误差界与所设计的终止准则构造了单调下降的代理序列, 从而允许非单调目标函数值序列的存在. 在数值实验中, 相比于 PALM-E 算法与 PALM-F 算法, PALM-I 算法的效率显著更高. 我们还将 PALM-I 算法嵌入了一个瀑布型多重网格优化 (CMGOPT) 框架, 通过求解问题 (2.8), 模拟了一维、二维强关联电子体系. 我们取得了符合理论预测与物理直观的数值结果, 并首次可视化了二维情形下电子位置之间的映射.

3.2 算法描述

我们所设计的 PALM-I 算法如算法 3.2 所示. 相比于算法 3.1, 我们在算法 3.2 中明确了子问题的非精确求解准则 (3.8). 对 $i = 1, \dots, s$, 残差函数 $r_i : \mathbb{R}^{m_i} \times \mathbb{R}_+^{p_i} \times \mathbb{R}^{m_i} \rightarrow \mathbb{R}_+$ 定义为

$$r_i(\mathbf{x}_i, \lambda_i, \tilde{\mathbf{x}}_i) := \max \left\{ \langle \mathbf{x}_i, \mathbf{x}_i - \tilde{\mathbf{x}}_i - \nabla \mathbf{g}_i(\mathbf{x}_i) \lambda_i \rangle, 0 \right\} + \left\| \mathbf{x}_i - \tilde{\mathbf{x}}_i - \nabla \mathbf{g}_i(\mathbf{x}_i) \lambda_i \right\|_\infty \quad (3.9)$$

$$+ \left\| \max \{-\mathbf{g}_i(\mathbf{x}_i), 0\} \right\|_\infty + \max \left\{ \langle \lambda_i, \mathbf{g}_i(\mathbf{x}_i) \rangle, 0 \right\}.$$

不难看出, r_i 可度量子问题 (3.7) 的 KKT 违反度. 因此, 当子问题被求解得足够精确时, 非精确终止准则 (3.8) 总能成立. 此外, 残差函数 r_i 总是可以计算的. 若使用(原始-)对偶算法求解子问题 (3.7), 则可直接将对偶变量 $\lambda_i^{(k+1)}$ 代入计算. 若不然, 我们也可以在 $\left\| \max \left\{ -\mathbf{g}_i(\mathbf{x}_i^{(k+1)}), 0 \right\} \right\|_\infty \leq \varepsilon_k^2/4$ 成立时求解如下线性规划:

$$\max_{\lambda_i} 0, \quad \text{s. t.} \quad \begin{cases} \left\langle \mathbf{x}_i^{(k+1)}, \mathbf{x}_i^{(k+1)} - \tilde{\mathbf{x}}_i^{(k)} - \nabla \mathbf{g}_i(\mathbf{x}_i^{(k+1)}) \lambda_i \right\rangle \leq \frac{\varepsilon_k^2}{4}, \\ -\frac{\varepsilon_k^2}{4} \mathbf{1}_{m_i} \leq \mathbf{x}_i^{(k+1)} - \tilde{\mathbf{x}}_i^{(k)} - \nabla \mathbf{g}_i(\mathbf{x}_i^{(k+1)}) \lambda_i \leq \frac{\varepsilon_k^2}{4} \mathbf{1}_{m_i}, \\ \left\langle \lambda_i, \mathbf{g}_i(\mathbf{x}_i^{(k+1)}) \right\rangle \leq \frac{\varepsilon_k^2}{4}, \quad \lambda_i \geq 0. \end{cases}$$

下面, 我们用例 3.1 说明残差函数 r_i 如何计算.

例 3.3 (残差函数的计算). 假设可行域 \mathcal{F}_i 是例 3.1 中定义的 Birkhoff 多胞体. 我们使用文献^[180] 中提出的半光滑 Newton 法求解子问题 (3.7) 的对偶问题. 在每次迭代中, 半光滑 Newton 法主要进行如下两步:

步 1: 非精确求解 Newton 方程更新对偶变量 $\lambda_{i,1}^{(k+1)}, \lambda_{i,2}^{(k+1)} \in \mathbb{R}^{m_i}$. 它们分别对应于约束 $X_i \mathbf{1}_{m_i} = \mathbf{1}_{m_i}, X_i^\top \mathbf{1}_{m_i} = \mathbf{1}_{m_i}$.

步 2: 计算原始变量

$$X_i^{(k+1)} := \max \left\{ \tilde{X}_i^{(k+1)} + \lambda_{i,1}^{(k+1)} \mathbf{1}_{m_i}^\top + \mathbf{1}_{m_i} \lambda_{i,2}^{(k+1)\top} \right\} \in \mathbb{R}_+^{m_i \times m_i},$$

并检查非精确求解准则 (3.8) 是否成立.

下面, 我们写出在此情形下残差函数的计算公式. 令

$$\Phi_i^{(k+1)} := X_i^{(k+1)} - \tilde{X}_i^{(k+1)} - \lambda_{i,1}^{(k+1)} \mathbf{1}_{m_i}^\top - \mathbf{1}_{m_i} \lambda_{i,2}^{(k+1)\top} \in \mathbb{R}_+^{m_i \times m_i}.$$

则 $\Phi_i^{(k+1)}$ 可当作非负约束对应的对偶变量. 由其定义可知

$$X_i^{(k+1)} - \tilde{X}_i^{(k+1)} - \lambda_{i,1}^{(k+1)} \mathbf{1}_{m_i}^\top - \mathbf{1}_{m_i} \lambda_{i,2}^{(k+1)\top} - \Phi_i^{(k+1)} = 0.$$

于是残差函数 (3.9) 中的前两项等于 0. 因此, 我们有简化的残差函数

$$\begin{aligned} r_i(X_i^{(k+1)}, \{\lambda_{i,1}^{(k+1)}, \lambda_{i,2}^{(k+1)}, \Phi_i^{(k+1)}\}, \tilde{X}_i^{(k)}) &= \left\| X_i^{(k+1)} \mathbf{1}_{m_i} - \mathbf{1}_{m_i} \right\|_\infty + \left\| X_i^{(k+1)\top} \mathbf{1}_{m_i} - \mathbf{1}_{m_i} \right\|_\infty \\ &+ \max \left\{ \left\langle \lambda_{i,1}^{(k+1)}, X_i^{(k+1)} \mathbf{1}_{m_i} - \mathbf{1}_{m_i} \right\rangle, 0 \right\} + \max \left\{ \left\langle \lambda_{i,2}^{(k+1)}, X_i^{(k+1)\top} \mathbf{1}_{m_i} - \mathbf{1}_{m_i} \right\rangle, 0 \right\}. \end{aligned}$$

在一定条件下, 我们可以借助非精确求解准则 (3.8) 导出子问题 (3.7) 的一个误差界.

引理 3.1 (子问题的误差界). 假设在 $\bigtimes_{i=1}^s \bar{\mathcal{F}}_i$ 上, f 对每一个变量块都连续可微, 其中

$$\bar{\mathcal{F}}_i := \left\{ \mathbf{w}_i \in \mathbb{R}^{m_i} : \text{dist}(\mathbf{w}_i, \mathcal{F}_i) \leq \bar{\varepsilon} \right\}, \quad i = 1, \dots, s.$$

对 $i = 1, \dots, s$, 假设 \mathcal{F}_i 是 \mathbb{R}^{m_i} 中的紧凸集, \mathbf{g}_i 满足如下两个条件中的一个:

(1) \mathbf{g}_i 是线性函数.

(2) \mathbf{g}_i 满足 *Slater* 约束规范条件, 即存在 $\hat{\mathbf{x}}_i \in \mathbb{R}^{m_i}$, 使得 $\mathbf{g}_i(\hat{\mathbf{x}}_i) > 0$, 以及如下 *Hoffman* 型不等式成立:

$$\text{dist}(\mathbf{x}_i, \mathcal{F}_i) \leq \tilde{c}_i \|\max\{-\mathbf{g}_i(\mathbf{x}_i), 0\}\|, \quad \forall \mathbf{x}_i \in \tilde{\mathcal{F}}_i, \quad (3.10)$$

其中 $\tilde{c}_i \geq 0$ 是常数,

$$\tilde{\mathcal{F}}_i := \left\{ \mathbf{w}_i \in \mathbb{R}^{m_i} : \text{dist}(\mathbf{w}_i, \bar{\mathcal{F}}_i) \leq \frac{\bar{M}_i}{\sigma_{\min}} \right\}, \quad \bar{M}_i := \sup_{\mathbf{x} \in \bigtimes_{i=1}^s \bar{\mathcal{F}}_i} \|\nabla_{\mathbf{x}_i} f(\mathbf{x})\|.$$

令 $\{\mathbf{x}^{(k)}\}$ 为 *PALM-I* 算法产生的迭代点序列. 则存在常数 $\omega \geq 0$, 使得

$$\|\mathbf{x}^{(k+1)} - \bar{\mathbf{x}}^{(k+1)}\| \leq \omega \varepsilon_k, \quad \forall k \geq 0.$$

证明. 根据文献^[184] 定理 2.2 及对 $\{\mathbf{g}_i\}_{i=1}^s$ 的假设条件, 对 $i = 1, \dots, s$ 与 $k \geq 0$, 有

$$\begin{aligned} \|\mathbf{x}_i^{(k+1)} - \bar{\mathbf{x}}_i^{(k+1)}\|^2 &\leq \omega_{i,1} \left\| \mathbf{x}_i^{(k+1)} - \bar{\mathbf{x}}_i^{(k)} - \nabla \mathbf{g}_i(\mathbf{x}_i^{(k+1)}) \boldsymbol{\lambda}_i^{(k+1)} \right\|^2 \\ &\quad + \max \left\{ \left\langle \mathbf{x}_i^{(k+1)}, \mathbf{x}_i^{(k+1)} - \bar{\mathbf{x}}_i^{(k)} - \nabla \mathbf{g}_i(\mathbf{x}_i^{(k+1)}) \boldsymbol{\lambda}_i^{(k+1)} \right\rangle, 0 \right\} \\ &\quad + \omega_{i,2,k} \left\| \max \left\{ -\mathbf{g}_i(\mathbf{x}_i^{(k+1)}), 0 \right\} \right\|_\infty + \max \left\{ \left\langle \boldsymbol{\lambda}_i^{(k+1)}, \mathbf{g}(\mathbf{x}_i^{(k+1)}) \right\rangle, 0 \right\}, \end{aligned}$$

其中

$$\omega_{i,1} := \max_{\mathbf{x}_i \in \mathcal{F}_i} \|\mathbf{x}_i\|_1, \quad \omega_{i,2,k} := \min_{\boldsymbol{\lambda}_i \in \mathcal{W}_i^{(k)}} \|\boldsymbol{\lambda}_i\|_1,$$

$\mathcal{W}_i^{(k)} \subseteq \mathbb{R}_+^{p_i}$ 是子问题 (3.7) 的所有最优对偶变量构成的集合. 由 *Hoffman* 型不等式 (3.10), 我们可从文献^[185] 命题 3 得知 $\sup_k \omega_{i,2,k} < \infty$. 对 $i = 1, \dots, s$, 令 $\omega_i := \max \{1, \omega_{i,1}, \sup_k \omega_{i,2,k}\}$, $\omega := \max_i \sqrt{\omega_i}$, 再注意到残差函数的定义 (3.9) 即可得证. \square

注. 相较于文献^[178] 中的非精确求解准则 (3.5), *PALM-I* 算法中的准则 (3.8) 更具有实际应用价值. 由文献^[185] 我们知道当 \mathbf{g}_i 是线性映射 (例如例 3.1) 或满足如下加强的 *Slater* 约束规范条件 (例如例 3.2)

$$\begin{cases} \exists \hat{\mathbf{x}}_i \in \mathbb{R}^{m_i}, \text{ s. t. } \mathbf{g}_i(\hat{\mathbf{x}}_i) > 0; \\ \exists \zeta \geq 0, \text{ s. t. } \frac{\|\mathbf{y}_i - \hat{\mathbf{x}}_i\| - \text{dist}(\mathbf{y}_i, \mathcal{F}_i)}{\min_{j=1, \dots, p_i} \{g_{i,j}(\hat{\mathbf{x}}_i)\}} \leq \zeta, \quad \forall \mathbf{y}_i \in \tilde{\mathcal{F}}_i \end{cases}$$

时, *Hoffman* 型不等式 (3.10) 成立. 在这些情况下, 我们无需计算 $\{\bar{\mathbf{x}}^{(k)}\}$ 即可控制子问题非精确求解的误差.

注. 由于 *PALM-I* 算法中的子问题非精确求解准则 (3.8) 没有限制迭代点是否可行, 因此我们之后取得的理论结果同样适用于 *PALM-F* 算法.

3.3 收敛性分析

在本节中, 我们研究 PALM-I 算法的收敛性, 包括全局依子列收敛性与全局依点列收敛性. 我们首先陈述对 f 、 $\{\mathcal{F}_i\}_{i=1}^s$ 、 $\{\mathbf{g}_i\}_{i=1}^s$ 与 $\{\varepsilon_k\}$ 的假设条件.

条件 3.2. 问题 (3.1) 的目标函数 f 在 $\bigtimes_{i=1}^s \bar{\mathcal{F}}_i$ 上对每个变量块都 Lipschitz 连续可微, 即对 $i = 1, \dots, s$, 存在常数 $L_i > 0$ 使得

$$\left\| \nabla_{\mathbf{x}_i} f(\mathbf{x}) - \nabla_{\mathbf{x}_i} f(\mathbf{x}') \right\| \leq L_i \|\mathbf{x} - \mathbf{x}'\|, \quad \forall \mathbf{x}, \mathbf{x}' \in \bigtimes_{i=1}^s \bar{\mathcal{F}}_i$$

成立, 其中 $\{\bar{\mathcal{F}}_i\}_{i=1}^s$ 定义在引理 3.1 中.

条件 3.3. 对 $i = 1, \dots, s$, 可行域 \mathcal{F}_i 是 \mathbb{R}^{m_i} 中的紧凸集, 且 \mathbf{g}_i 满足如下两个条件中的一个:

- (1) \mathbf{g}_i 是线性函数.
- (2) \mathbf{g}_i 满足 Slater 约束规范条件与 Hoffman 型不等式 (3.10).

注. 条件 3.3 保证了 \mathbf{x} 是 F 的一阶稳定点当且仅当它是问题 (3.1) 的 KKT 点.

条件 3.4.

- (1) 序列 $\{\varepsilon_k\}$ 非负且平方可和.
- (2) 序列 $\{\varepsilon_k\}$ 非负、可和, 且存在 $\bar{\theta} \in (0, 1)$ 使得 $\{e_k^{\bar{\theta}}\}$ 可和, 其中

$$e_k := \sum_{t=k}^{\infty} \varepsilon_t^2, \quad \forall k \geq 0.$$

注. 条件 3.4 (2) 看起来十分的严苛. 但事实上, 给定 $l > 1$, 次线性下降序列 $\{\bar{\varepsilon}/(k+1)^l\}$ 即可满足需求. 这是因为此时 $e_k = \mathcal{O}(k^{-(2l-1)})$. 为保证 $\{e_k^{\bar{\theta}}\}$ 可和, 只需选取 $(0, 1) \ni \bar{\theta} > 1/(2l-1)$. 需要强调的是, 条件 3.4 (2) 只需要 $\bar{\theta}$ 的存在性, 不需要其确切的数值.

注. 使条件 3.4 (2) 成立的一个直观的充分条件是 $\sum_{k=1}^{\infty} k \varepsilon_k^{2\bar{\theta}} < \infty$. 然而, 为了保有序列 $\{\varepsilon_k\}$ 选取的灵活性, 我们仍然使用条件 3.4 (2).

在开始证明 PALM-I 算法的收敛性之前, 我们先定义一些记号. 记 $L := \max_i L_i > 0$, 并令 $\sigma_{\min} := \gamma L$ ($\gamma > 1$), 而 σ_{\max} 是不小于 σ_{\min} 的任一标量. 记 $\underline{\sigma}_k := \min_i \sigma_{i,k}$, $\bar{\sigma}_k := \max_i \sigma_{i,k}$, $v := 12/(\gamma-1)$, $\bar{M} := \sqrt{3}(\sigma_{\max} + L\sqrt{s})$,

$$C_{0,k} := \frac{\sigma_k - L(1 + 6/v)}{2},$$

$$C_{1,k} := \frac{\bar{\sigma}_k + L[v s^2/2 + (2 + 2/v - v/2)s + 2v + 4/v + 3]}{2},$$

以及 $\bar{C}_1 := 2\omega^2 \max_k C_{1,k}$, 其中 ω 是引理 3.1 中定义的常数. 我们用“D”表示“最优”迭代点与真实迭代点之间的差, 例如

$$\mathbf{D}\mathbf{x}_j^{(k+1)} := \bar{\mathbf{x}}_j^{(k+1)} - \mathbf{x}_j^{(k+1)}, \quad \mathbf{D}\mathbf{x}^{(k+1)} := \bar{\mathbf{x}}^{(k+1)} - \mathbf{x}^{(k+1)}.$$

3.3.1 全局依子列收敛性

我们首先证明目标函数值序列上的逐块近似充分下降性.

引理 3.5 (目标函数值序列的逐块近似充分下降). 假设条件 3.2 成立. 令 $\{\mathbf{x}^{(k)}\}$ 为 PALM-I 算法产生的迭代点序列. 则对 $i = 1, \dots, s$ 与任意 $k \geq 0$,

$$\begin{aligned} & f(\bar{\mathbf{x}}_{<i}^{(k+1)}, \bar{\mathbf{x}}_i^{(k)}, \bar{\mathbf{x}}_{>i}^{(k)}) - f(\bar{\mathbf{x}}_{<i}^{(k+1)}, \bar{\mathbf{x}}_i^{(k+1)}, \bar{\mathbf{x}}_{>i}^{(k)}) \\ & \geq \frac{\sigma_{i,k} - L(1 + 6/\nu)}{2} \left\| \bar{\mathbf{x}}_i^{(k+1)} - \mathbf{x}_i^{(k)} \right\|^2 - \frac{\sigma_{i,k} + L(2\nu + 3 + 4/\nu)}{2} \left\| D\mathbf{x}_i^{(k)} \right\|^2 \\ & \quad - \frac{L[\nu(i-1) + 2 + 2/\nu]}{2} \left\| D\mathbf{x}^{(k+1)} \right\|^2 - \frac{L[\nu(s-i) + 2 + 2/\nu]}{2} \left\| D\mathbf{x}^{(k)} \right\|^2. \end{aligned} \quad (3.11)$$

证明. 我们的证明依赖于条件 3.2 与 $\bar{\mathbf{x}}_i^{(k+1)}$ 的最优性 (参见 (3.6) 式). 首先注意到 (3.11) 的左端可分解成下面的五部分之和:

- (1) $\sum_{j=1}^{i-1} \left[f(\mathbf{x}_{<j}^{(k+1)}, \bar{\mathbf{x}}_{[j,i]}^{(k+1)}, \bar{\mathbf{x}}_{\geq i}^{(k)}) - f(\mathbf{x}_{\leq j}^{(k+1)}, \bar{\mathbf{x}}_{(j,i]}^{(k+1)}, \bar{\mathbf{x}}_{\geq i}^{(k)}) \right];$
- (2) $\sum_{j=i+1}^s \left[f(\mathbf{x}_{<i}^{(k+1)}, \bar{\mathbf{x}}_i^{(k)}, \mathbf{x}_{(i,j]}^{(k)}, \bar{\mathbf{x}}_{\geq j}^{(k)}) - f(\mathbf{x}_{<i}^{(k+1)}, \bar{\mathbf{x}}_i^{(k)}, \mathbf{x}_{(i,j]}, \bar{\mathbf{x}}_{>j}^{(k)}) \right];$
- (3) $f(\bar{\mathbf{x}}_{<i}^{(k+1)}, \bar{\mathbf{x}}_i^{(k)}, \mathbf{x}_{>i}^{(k)}) - f(\bar{\mathbf{x}}_{<i}^{(k+1)}, \bar{\mathbf{x}}_i^{(k+1)}, \mathbf{x}_{>i}^{(k)});$
- (4) $\sum_{j=1}^{i-1} \left[f(\bar{\mathbf{x}}_{<j}^{(k+1)}, \mathbf{x}_{[j,i]}^{(k+1)}, \bar{\mathbf{x}}_i^{(k+1)}, \mathbf{x}_{>i}^{(k)}) - f(\bar{\mathbf{x}}_{\leq j}^{(k+1)}, \mathbf{x}_{(j,i]}^{(k+1)}, \bar{\mathbf{x}}_i^{(k+1)}, \mathbf{x}_{>i}^{(k)}) \right];$
- (5) $\sum_{j=i+1}^s \left[f(\bar{\mathbf{x}}_{\leq i}^{(k+1)}, \bar{\mathbf{x}}_{(i,j]}^{(k)}, \mathbf{x}_{\geq j}^{(k)}) - f(\bar{\mathbf{x}}_{\leq i}^{(k+1)}, \bar{\mathbf{x}}_{(i,j]}, \mathbf{x}_{>j}^{(k)}) \right].$

这里, 对 $i, j \in \{1, \dots, s\}$,

$$\mathbf{x}_{(i,j)} := (\mathbf{x}_{i+1}, \dots, \mathbf{x}_{j-1}).$$

类似可定义 $\mathbf{x}_{[i,j]}$ 、 $\mathbf{x}_{(i,j]}$ 与 $\mathbf{x}_{[i,j]}$. 当下标集合为空集时, 这些向量为空.

根据条件 3.2 与 $\bar{\mathbf{x}}_i^{(k+1)}$ 的最优性, 第 (3) 部分有如下下界:

$$\begin{aligned} (3) & \geq - \left\langle \nabla_{\mathbf{x}_i} f(\bar{\mathbf{x}}_{<i}^{(k+1)}, \bar{\mathbf{x}}_i^{(k)}, \mathbf{x}_{>i}^{(k)}), \bar{\mathbf{x}}_i^{(k+1)} - \bar{\mathbf{x}}_i^{(k)} \right\rangle - \frac{L_i}{2} \left\| \bar{\mathbf{x}}_i^{(k+1)} - \bar{\mathbf{x}}_i^{(k)} \right\|^2 \\ & = - \left\langle \nabla_{\mathbf{x}_i} f(\bar{\mathbf{x}}_{<i}^{(k+1)}, \mathbf{x}_{\geq i}^{(k)}), \bar{\mathbf{x}}_i^{(k+1)} - \bar{\mathbf{x}}_i^{(k)} \right\rangle - \frac{L_i}{2} \left\| \bar{\mathbf{x}}_i^{(k+1)} - \bar{\mathbf{x}}_i^{(k)} \right\|^2 \\ & \quad + \left\langle \nabla_{\mathbf{x}_i} f(\bar{\mathbf{x}}_{<i}^{(k+1)}, \mathbf{x}_{\geq i}^{(k)}) - \nabla_{\mathbf{x}_i} f(\bar{\mathbf{x}}_{<i}^{(k+1)}, \bar{\mathbf{x}}_i^{(k)}, \mathbf{x}_{>i}^{(k)}), \bar{\mathbf{x}}_i^{(k+1)} - \bar{\mathbf{x}}_i^{(k)} \right\rangle \\ & \geq \frac{\sigma_{i,k}}{2} \left[\left\| \bar{\mathbf{x}}_i^{(k+1)} - \mathbf{x}_i^{(k)} \right\|^2 - \left\| D\mathbf{x}_i^{(k)} \right\|^2 \right] - \frac{L}{2} \left(1 + \frac{1}{\nu} \right) \left\| \bar{\mathbf{x}}_i^{(k+1)} - \mathbf{x}_i^{(k)} \right\|^2 \\ & \quad - \frac{L(1+\nu)}{2} \left\| D\mathbf{x}_i^{(k)} \right\|^2 - \frac{L}{2} \left[(\nu+2) \left\| D\mathbf{x}_i^{(k)} \right\|^2 + \frac{1}{\nu} \left\| \bar{\mathbf{x}}_i^{(k+1)} - \mathbf{x}_i^{(k)} \right\|^2 \right] \\ & = \frac{\sigma_{i,k} - L(1 + 2/\nu)}{2} \left\| \bar{\mathbf{x}}_i^{(k+1)} - \mathbf{x}_i^{(k)} \right\|^2 - \frac{\sigma_{i,k} + L(2\nu + 3)}{2} \left\| D\mathbf{x}_i^{(k)} \right\|^2, \end{aligned} \quad (3.12)$$

其中第二个不等式还使用了均值不等式与 L 的定义.

由于第 (1)、(4) 部分与第 (2)、(5) 部分结构类似, 因此我们只详细陈述对第 (1)、(4) 部分的分析. 根据条件 3.2,

$$(1) \geq \sum_{j=1}^{i-1} \left[\left\langle \nabla_{\mathbf{x}_j} f(\bar{\mathbf{x}}_{<j}^{(k+1)}, \bar{\mathbf{x}}_{[j,i]}^{(k+1)}, \bar{\mathbf{x}}_{\geq i}^{(k)}), -D\mathbf{x}_j^{(k+1)} \right\rangle - \frac{L_j}{2} \left\| D\mathbf{x}_j^{(k+1)} \right\|^2 \right],$$

$$(4) \geq \sum_{j=1}^{i-1} \left[\left\langle \nabla_{\mathbf{x}_j} f(\bar{\mathbf{x}}_{<j}^{(k+1)}, \mathbf{x}_{[j,i]}^{(k+1)}, \bar{\mathbf{x}}_i^{(k+1)}, \mathbf{x}_{>i}^{(k)}), D\mathbf{x}_j^{(k+1)} \right\rangle - \frac{L_j}{2} \|D\mathbf{x}_j^{(k+1)}\|^2 \right].$$

结合这两个不等式, 我们有

$$\begin{aligned} (1) + (4) &\geq -L \sum_{j=1}^{i-1} \|D\mathbf{x}_j^{(k+1)}\|^2 - L \sum_{j=1}^{i-1} \left\| \begin{pmatrix} \mathbf{x}_{<j}^{(k+1)} - \bar{\mathbf{x}}_{<j}^{(k+1)} \\ \bar{\mathbf{x}}_{[j,i]}^{(k+1)} - \mathbf{x}_{[j,i]}^{(k+1)} \\ \bar{\mathbf{x}}_i^{(k)} - \bar{\mathbf{x}}_i^{(k+1)} \\ \bar{\mathbf{x}}_{>i}^{(k)} - \mathbf{x}_{>i}^{(k)} \end{pmatrix} \right\| \|D\mathbf{x}_j^{(k+1)}\| \\ &= -L \sum_{j=1}^{i-1} \|D\mathbf{x}_j^{(k+1)}\|^2 - L \left\| \begin{pmatrix} \mathbf{x}_{<i}^{(k+1)} - \bar{\mathbf{x}}_{<i}^{(k+1)} \\ \bar{\mathbf{x}}_i^{(k)} - \bar{\mathbf{x}}_i^{(k+1)} \\ \bar{\mathbf{x}}_{>i}^{(k)} - \mathbf{x}_{>i}^{(k)} \end{pmatrix} \right\| \sum_{j=1}^{i-1} \|D\mathbf{x}_j^{(k+1)}\| \\ &\geq -L \sum_{j=1}^{i-1} \|D\mathbf{x}_j^{(k+1)}\|^2 - L \left\| \begin{pmatrix} \mathbf{x}_{<i}^{(k+1)} - \bar{\mathbf{x}}_{<i}^{(k+1)} \\ \bar{\mathbf{x}}_i^{(k)} - \bar{\mathbf{x}}_i^{(k+1)} \\ \bar{\mathbf{x}}_{>i}^{(k)} - \mathbf{x}_{>i}^{(k)} \end{pmatrix} \right\| \sqrt{(i-1) \sum_{j=1}^{i-1} \|D\mathbf{x}_j^{(k+1)}\|^2} \\ &\geq -\frac{L[v(i-1)+2]}{2} \sum_{j=1}^{i-1} \|D\mathbf{x}_j^{(k+1)}\|^2 - \frac{L}{v} \left[\|\bar{\mathbf{x}}_i^{(k+1)} - \mathbf{x}_i^{(k)}\|^2 + \|D\mathbf{x}_i^{(k)}\|^2 \right] \\ &\quad - \frac{L}{2v} \left[\sum_{l < i} \|D\mathbf{x}_l^{(k+1)}\|^2 + \sum_{l > i} \|D\mathbf{x}_l^{(k)}\|^2 \right], \end{aligned}$$

其中第一个不等式使用了条件 3.2 以及 L 的定义, 第二个与最后一个不等式使用了均值不等式. 类似地, 我们有

$$(2) + (5)$$

$$\begin{aligned} &\geq -\frac{L[v(s-i)+2]}{2} \sum_{j=i+1}^s \|D\mathbf{x}_j^{(k)}\|^2 - \frac{L}{v} \left[\|\bar{\mathbf{x}}_i^{(k+1)} - \mathbf{x}_i^{(k)}\|^2 + \|D\mathbf{x}_i^{(k)}\|^2 \right] \\ &\quad - \frac{L}{2v} \left[\sum_{l < i} \|D\mathbf{x}_l^{(k+1)}\|^2 + \sum_{l > i} \|D\mathbf{x}_l^{(k)}\|^2 \right]. \end{aligned}$$

结合 (3.12) 式与前两个不等式, 我们有

$$\begin{aligned} &f(\bar{\mathbf{x}}_{<i}^{(k+1)}, \bar{\mathbf{x}}_i^{(k)}, \bar{\mathbf{x}}_{>i}^{(k)}) - f(\bar{\mathbf{x}}_{<i}^{(k+1)}, \bar{\mathbf{x}}_i^{(k+1)}, \bar{\mathbf{x}}_{>i}^{(k)}) \\ &\geq \frac{\sigma_{i,k} - L(1+6/v)}{2} \|\bar{\mathbf{x}}_i^{(k+1)} - \mathbf{x}_i^{(k)}\|^2 - \frac{\sigma_{i,k} + L(2v+3+4/v)}{2} \|D\mathbf{x}_i^{(k)}\|^2 \\ &\quad - \frac{L[v(i-1)+2+2/v]}{2} \sum_{l < i} \|D\mathbf{x}_l^{(k+1)}\|^2 - \frac{L[v(s-i)+2+2/v]}{2} \sum_{l > i} \|D\mathbf{x}_l^{(k)}\|^2. \end{aligned}$$

再根据 $\mathbf{x}^{(k+1)}$ 与 $\mathbf{x}^{(k)}$ 的定义即可得证. \square

由上面的引理, 不难得到下面的命题.

命题 3.6 (目标函数值序列的近似充分下降). 假设条件 3.2 成立. 令 $\{\mathbf{x}^{(k)}\}$ 为 *PALM-I* 算法产生的迭代点序列. 则对任意 $k \geq 0$,

$$f(\bar{\mathbf{x}}^{(k)}) - f(\bar{\mathbf{x}}^{(k+1)}) \geq C_{0,k} \|\bar{\mathbf{x}}^{(k+1)} - \mathbf{x}^{(k)}\|^2 - C_{1,k} \|\mathbf{D}\mathbf{x}^{(k)}\|^2 - C_{1,k+1} \|\mathbf{D}\mathbf{x}^{(k+1)}\|^2.$$

证明. 由引理 3.5,

$$\begin{aligned} & f(\bar{\mathbf{x}}^{(k)}) - f(\bar{\mathbf{x}}^{(k+1)}) \\ & \geq \sum_{i=1}^s \left[f(\bar{\mathbf{x}}_{<i}^{(k+1)}, \bar{\mathbf{x}}_i^{(k)}, \bar{\mathbf{x}}_{>i}^{(k)}) - f(\bar{\mathbf{x}}_{<i}^{(k+1)}, \bar{\mathbf{x}}_i^{(k+1)}, \bar{\mathbf{x}}_{>i}^{(k)}) \right] \\ & \geq \sum_{i=1}^s \left[\frac{\sigma_{i,k} - L(1 + 6/\nu)}{2} \|\bar{\mathbf{x}}_i^{(k+1)} - \mathbf{x}_i^{(k)}\|^2 - \frac{\sigma_{i,k} + L(2\nu + 3 + 4/\nu)}{2} \|\mathbf{D}\mathbf{x}_i^{(k)}\|^2 \right. \\ & \quad \left. - \frac{L[\nu(i-1) + 2 + 2\nu]}{2} \|\mathbf{D}\mathbf{x}^{(k+1)}\|^2 - \frac{L[\nu(s-i) + 2 + 2/\nu]}{2} \|\mathbf{D}\mathbf{x}^{(k)}\|^2 \right] \\ & \geq \frac{\min_i \sigma_{i,k} - L(1 + 6/\nu)}{2} \|\bar{\mathbf{x}}^{(k+1)} - \mathbf{x}^{(k)}\|^2 - \frac{L[\nu s^2/2 + (2 + 2/\nu - \nu/2)s]}{2} \|\mathbf{D}\mathbf{x}^{(k+1)}\|^2 \\ & \quad - \frac{\max_i \sigma_{i,k} + L[\nu s^2/2 + (2 + 2/\nu - \nu/2)s + 2\nu + 4/\nu + 3]}{2} \|\mathbf{D}\mathbf{x}^{(k)}\|^2, \end{aligned}$$

再根据 $\underline{\sigma}_k$ 、 $\bar{\sigma}_k$ 、 $C_{0,k}$ 与 $C_{1,k}$ 的定义即可得证. \square

从命题 3.6 可知, 迭代点的不可行性会破坏目标函数值的单调性, 并且无法避免. 受文献^[186–188] 启发, 我们构造一个单调下降的代理序列, 其显式地包含子问题非精确求解的误差.

命题 3.7 (代理序列的充分下降). 假设条件 3.2 与 3.3 成立. 令 $\{\mathbf{x}^{(k)}\}$ 为 *PALM-I* 算法产生的迭代点序列, 其中 $\{\varepsilon_k\}$ 满足条件 3.4 (I). 则有如下结论成立:

(1) 序列 $\{v_k := F(\bar{\mathbf{x}}^{(k)}) + u_k + u_{k+1}\}$ 是良定义的, 其中

$$u_k := \sum_{t=k}^{\infty} C_{1,t} \|\mathbf{D}\mathbf{x}^{(t)}\|^2;$$

(2) 对任意 $k \geq 0$, $v_k - v_{k+1} \geq C_{0,k} \|\bar{\mathbf{x}}^{(k+1)} - \mathbf{x}^{(k)}\|^2 \geq 0$;

(3) 序列 $\{v_k\}$ 单调收敛于某个 $\bar{F} \in \mathbb{R}$, 其可在 $\bigtimes_{i=1}^s \mathcal{F}_i$ 上被 F 取到. 特别地, 随着 k 趋于无穷大, $F(\bar{\mathbf{x}}^{(k)})$ 收敛于 \bar{F} ;

(4) 若存在 $\tilde{k} \in \mathbb{N}$ 使得 $v_{\tilde{k}} = \bar{F}$, 则对任意 $k \geq \tilde{k}$ 恒有 $v_k = \bar{F}$ 与 $\mathbf{x}^{(k)} = \bar{\mathbf{x}}^{(k+1)}$. 进一步地, 若还存在 $\hat{k} \geq \tilde{k}$ 使得 $\mathbf{x}^{(\hat{k})} = \bar{\mathbf{x}}^{(\hat{k})}$, 则对任意 $k \geq \hat{k}$ 恒有 $\mathbf{x}^{(k)} = \bar{\mathbf{x}}^{(k+1)} = \mathbf{x}^{(k+1)}$.

证明. (1) 由条件 3.4 (1), 序列 $\{\varepsilon_k\}$ 平方可和, 因此对任意 $k \in \mathbb{N}$,

$$u_k = \sum_{t=k}^{\infty} C_{1,t} \|\mathbf{D}\mathbf{x}^{(t)}\|^2 \leq \omega^2 \sum_{t=k}^{\infty} C_{1,t} \varepsilon_t^2 \leq \frac{\bar{C}_1}{2} \sum_{t=k}^{\infty} \varepsilon_t^2 < \infty,$$

其中第一个不等式使用了引理3.1. 因此序列 $\{v_k\}$ 良定.

(2) 将 $\{v_k\}$ 的定义直接代入命题3.6即可得第一个不等式. 第二个不等式则需用到 $\sigma_{i,k} \geq \gamma L$ (可见算法3.2第5步), 由其可推得 $C_{0,k} \geq (\gamma - 1)L/4 > 0$.

(3) 由于 $\{\varepsilon_k\}$ 平方可和, 于是 $\{u_k\}$ 收敛到 0. 根据(2)中 $\{v_k\}$ 的充分下降性、 $C_{0,k} \geq (\gamma - 1)L/4$ (任意 $k \geq 0$) 以及 F 在 $\bigtimes_{i=1}^s \mathcal{F}_i$ 上的连续性即可得证.

(4) 前半部分可直接从(2)、(3)以及 $C_{0,k} \geq (\gamma - 1)L/4$ (任意 $k \geq 0$) 得到. 为证明后半部分, 我们首先从 v_k 的定义以及 $v_{\hat{k}} = v_{\hat{k}+1}$ 得知

$$F(\bar{\mathbf{x}}^{(\hat{k})}) + u_{\hat{k}} + u_{\hat{k}+1} = v_{\hat{k}} = v_{\hat{k}+1} = F(\bar{\mathbf{x}}^{(\hat{k}+1)}) + u_{\hat{k}+1} + u_{\hat{k}+2},$$

再结合 $\bar{\mathbf{x}}^{(\hat{k}+1)} = \mathbf{x}^{(\hat{k})} = \bar{\mathbf{x}}^{(\hat{k})}$, 即有

$$0 = u_{\hat{k}} - u_{\hat{k}+2} = C_{0,\hat{k}} \|\mathbf{D}\mathbf{x}^{(\hat{k})}\| + C_{0,\hat{k}+1} \|\mathbf{D}\mathbf{x}^{(\hat{k}+1)}\| = C_{0,\hat{k}+1} \|\mathbf{D}\mathbf{x}^{(\hat{k}+1)}\|.$$

因为 $C_{0,\hat{k}+1} \geq (\gamma - 1)L/4 > 0$, 所以 $\bar{\mathbf{x}}^{(\hat{k}+1)} = \mathbf{x}^{(\hat{k}+1)}$. 由归纳假设, 即可知结论对任意 $k \geq \hat{k}$ 都成立. \square

下面, 我们证明迭代点的稳定性违反度上界.

命题3.8 (迭代点稳定性违反度上界). 假设条件3.2成立. 令 $\{\mathbf{x}^{(k)}\}$ 为PALM-I算法产生的迭代点序列. 则对任意 $k \geq 0$, 存在 $\mathbf{w}^{(k+1)} \in \partial F(\bar{\mathbf{x}}^{(k+1)})$ 使得

$$\|\mathbf{w}^{(k+1)}\| \leq \bar{M} [\|\bar{\mathbf{x}}^{(k+1)} - \mathbf{x}^{(k)}\| + \|\mathbf{D}\mathbf{x}^{(k+1)}\|].$$

证明. 对 $i = 1, \dots, s$, 根据 $\bar{\mathbf{x}}_i^{(k+1)}$ 的最优化(见(3.6)式)以及Fréchet次微分的计算公式^[22], 可知存在 $\mathbf{a}_i^{(k+1)} \in \partial \delta_{\mathcal{F}_i}(\bar{\mathbf{x}}_i^{(k+1)})$ 使得

$$0 = \nabla_{\mathbf{x}_i} f(\mathbf{x}_{<i}^{(k+1)}, \bar{\mathbf{x}}_i^{(k+1)}, \mathbf{x}_{>i}^{(k)}) + \sigma_{i,k}(\bar{\mathbf{x}}_i^{(k+1)} - \mathbf{x}_i^{(k)}) + \mathbf{a}_i^{(k+1)}.$$

进而

$$\begin{aligned} \partial F(\bar{\mathbf{x}}^{(k+1)}) &\ni \nabla f(\bar{\mathbf{x}}^{(k+1)}) + (\mathbf{a}_i^{(k+1)})_{i=1}^s \\ &= \left(\nabla_{\mathbf{x}_i} f(\bar{\mathbf{x}}^{(k+1)}) - \nabla_{\mathbf{x}_i} f(\mathbf{x}_{<i}^{(k+1)}, \bar{\mathbf{x}}_i^{(k+1)}, \mathbf{x}_{>i}^{(k)}) - \sigma_{i,k}(\bar{\mathbf{x}}_i^{(k+1)} - \mathbf{x}_i^{(k)}) \right)_{i=1}^s. \end{aligned} \quad (3.13)$$

记 $\mathbf{w}^{(k+1)} := \nabla f(\bar{\mathbf{x}}^{(k+1)}) + (\mathbf{a}_i^{(k+1)})_{i=1}^s$. 对 $i = 1, \dots, s$,

$$\begin{aligned} &\left\| \nabla_{\mathbf{x}_i} f(\bar{\mathbf{x}}^{(k+1)}) - \nabla_{\mathbf{x}_i} f(\mathbf{x}_{<i}^{(k+1)}, \bar{\mathbf{x}}_i^{(k+1)}, \mathbf{x}_{>i}^{(k)}) - \sigma_{i,k}(\bar{\mathbf{x}}_i^{(k+1)} - \mathbf{x}_i^{(k)}) \right\| \\ &\leq L \left\| \begin{pmatrix} \bar{\mathbf{x}}_{<i}^{(k+1)} - \mathbf{x}_{<i}^{(k+1)} \\ \bar{\mathbf{x}}_{>i}^{(k+1)} - \mathbf{x}_{>i}^{(k)} \end{pmatrix} \right\| + \sigma_{i,k} \|\bar{\mathbf{x}}_i^{(k+1)} - \mathbf{x}_i^{(k)}\| \\ &\leq L [\|\mathbf{D}\mathbf{x}^{(k+1)}\| + \|\bar{\mathbf{x}}^{(k+1)} - \mathbf{x}^{(k)}\|] + \sigma_{i,k} \|\bar{\mathbf{x}}_i^{(k+1)} - \mathbf{x}_i^{(k)}\|, \end{aligned}$$

其中第一个不等式使用了条件 3.2 以及 L 的定义. 结合 (3.13) 式与均值不等式, 就有

$$\begin{aligned}\|\mathbf{w}^{(k+1)}\|^2 &= \sum_{i=1}^s \left[L \|\mathbf{D}\mathbf{x}^{(k+1)}\| + L \|\bar{\mathbf{x}}^{(k+1)} - \mathbf{x}^{(k)}\| + \sigma_{i,k} \left\| \bar{\mathbf{x}}_i^{(k+1)} - \mathbf{x}_i^{(k)} \right\| \right]^2 \\ &\leq 3sL^2 \left[\|\mathbf{D}\mathbf{x}^{(k+1)}\|^2 + \|\bar{\mathbf{x}}^{(k+1)} - \mathbf{x}^{(k)}\|^2 \right] + 3\bar{\sigma}_k^2 \|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\|^2,\end{aligned}$$

再根据 \bar{M} 的定义即可得证. \square

下面, 我们证明当序列 $\{\varepsilon_k\}$ 平方可和时, PALM-I 算法全局依子列收敛于 F 的一阶稳定点. 我们定义迭代点序列的聚点集合

$$\mathcal{S}(\mathbf{x}^{(0)}) := \left\{ \mathbf{x} = (\mathbf{x}_i)_{i=1}^s \in \bigtimes_{i=1}^s \mathbb{R}^{m_i} : \exists \mathcal{K} \subseteq \mathbb{N}, \text{ s. t. } \lim_{\mathcal{K} \ni k \rightarrow \infty} \mathbf{x}^{(k)} \rightarrow \mathbf{x} \right\}.$$

由于 $\{\varepsilon_k\}$ 趋于 0, 因此从引理 3.1 可知 $\{\mathbf{D}\mathbf{x}^{(k)}\}$ 也趋于 0, 从而

$$\mathcal{S}(\mathbf{x}^{(0)}) = \left\{ \mathbf{x} = (\mathbf{x}_i)_{i=1}^s \in \bigtimes_{i=1}^s \mathbb{R}^{m_i} : \exists \mathcal{K} \subseteq \mathbb{N}, \text{ s. t. } \lim_{\mathcal{K} \ni k \rightarrow \infty} \bar{\mathbf{x}}^{(k)} \rightarrow \mathbf{x} \right\}.$$

定理 3.9 (PALM-I 算法的全局依子列收敛性). 假设条件 3.2 与 3.3 成立. 令 $\{\mathbf{x}^{(k)}\}$ 为 PALM-I 算法产生的迭代点序列, 其中 $\{\varepsilon_k\}$ 满足条件 3.4 (1). 则有如下结论成立:

- (1) $\mathcal{S}(\mathbf{x}^{(0)}) \subseteq \bigtimes_{i=1}^s \mathcal{F}_i$ 非空且其中每个元素都是 F 的一阶稳定点;
- (2) $\mathcal{S}(\mathbf{x}^{(0)})$ 紧连通且

$$\lim_{k \rightarrow \infty} \text{dist}(\bar{\mathbf{x}}^{(k)}, \mathcal{S}(\mathbf{x}^{(0)})) = 0;$$

- (3) 函数 F 在 $\mathcal{S}(\mathbf{x}^{(0)})$ 上是有限常数.

证明. (1) 因为 $\{\bar{\mathbf{x}}^{(k)}\} \subseteq \bigtimes_{i=1}^s \mathcal{F}_i$ 且 \mathcal{F}_i 有界 (条件 3.3), 所以存在子列 $\{\bar{\mathbf{x}}^{(k)}\}_{k \in \mathcal{K}}$ 以及 $\bar{\mathbf{x}}$ 使得 $\{\bar{\mathbf{x}}^{(k)}\}_{k \in \mathcal{K}}$ 收敛到 $\bar{\mathbf{x}}$. 因此 $\mathcal{S}(\mathbf{x}^{(0)}) \neq \emptyset$. 同时, 注意到 $\bigtimes_{i=1}^s \mathcal{F}_i$ 是闭集 (条件 3.3), 因此 $\bar{\mathbf{x}} \in \bigtimes_{i=1}^s \mathcal{F}_i$. 由 $\bar{\mathbf{x}}$ 的任意性可知 $\mathcal{S}(\mathbf{x}^{(0)}) \subseteq \bigtimes_{i=1}^s \mathcal{F}_i$.

根据命题 3.7 (2) 以及 $C_{0,k} \geq (\gamma - 1)L/4$ (任意 $k \geq 0$), 我们知道

$$\frac{\gamma - 1}{4} L \|\bar{\mathbf{x}}^{(k+1)} - \mathbf{x}^{(k)}\|^2 \leq C_{0,k} \|\bar{\mathbf{x}}^{(k+1)} - \mathbf{x}^{(k)}\|^2 \leq v_k - v_{k+1}, \quad \forall k \geq 0.$$

将上面的不等式按指标 k 从 r_1 累加到 r_2 ($r_2 \geq r_1 \geq 0$), 可得

$$\begin{aligned}\frac{\gamma - 1}{4} L \sum_{k=r_1}^{r_2} \|\bar{\mathbf{x}}^{(k+1)} - \mathbf{x}^{(k)}\|^2 &\leq \sum_{k=r_1}^{r_2} (v_k - v_{k+1}) = v_{r_1} - v_{r_2+1} \leq v_{r_1} - \bar{F} \\ &\leq (F(\bar{\mathbf{x}}^{(r_1)}) - \bar{F}) + 2 \sum_{t=r_1}^{\infty} C_{1,t} \|\mathbf{D}\mathbf{x}^{(t)}\|^2 \leq (F(\bar{\mathbf{x}}^{(r_1)}) - \bar{F}) + \bar{C}_1 \sum_{t=r_1}^{\infty} \varepsilon_t^2,\end{aligned}$$

其中第二个不等式使用了命题3.7(3), 而最后一个使用了引理3.1和 \bar{C}_1 的定义. 由于 $\{\varepsilon_k\}$ 平方可和, 因此上面不等式的右端项对任何 r_2 都是有限的. 因此

$$\lim_{k \rightarrow \infty} \|\bar{\mathbf{x}}^{(k+1)} - \mathbf{x}^{(k)}\| = 0. \quad (3.14)$$

我们进一步可从引理3.1、命题3.8以及 $\{\varepsilon_k\}$ 的平方可和性推出

$$\lim_{k \rightarrow \infty} \mathbf{w}^{(k+1)} = 0.$$

为证明 $S(\mathbf{x}^{(0)})$ 中的元素都是 F 的一阶稳定点, 现任取 $\bar{\mathbf{x}} \in S(\mathbf{x}^{(0)})$, 其对应收敛子列 $\{\bar{\mathbf{x}}^{(k+1)}\}_{k \in \mathcal{K}}$. 因为子列 $\{(\bar{\mathbf{x}}^{(k+1)}, \mathbf{w}^{(k+1)})\}_{k \in \mathcal{K}}$ 收敛到 $(\bar{\mathbf{x}}, 0)$, 且根据命题3.7(3)以及 F 在 $\bigtimes_{i=1}^s \mathcal{F}_i$ 上的连续性, $\{F(\bar{\mathbf{x}}^{(k+1)})\}$ 收敛到 $\bar{F} = F(\bar{\mathbf{x}})$, 我们就可从Fréchet次微分的计算公式^[22]及定理1.3推出 $0 \in \partial F(\bar{\mathbf{x}})$. 这表明 $\bar{\mathbf{x}}$ 是 F 的稳定点. 因为 $\bar{\mathbf{x}}$ 是任取的, 所以结论得证.

(2)与(3)可直接由文献^[173]引理5的证明得到. \square

3.3.2 全局依点列收敛性

下面, 我们证明当函数 F 在一阶稳定点处满足Łojasiewicz性质且序列 $\{\varepsilon_k\}$ 满足条件3.4(2)时, PALM-I算法全局依点列收敛于 F 的一阶稳定点.

函数在一点处的Łojasiewicz性质定义如下.

定义3.1 (函数的单点Łojasiewicz性质^[189]). 设 $G : \mathbb{R}^n \rightarrow (-\infty, \infty]$ 为一适定下半连续函数. 我们称其在稳定点 $\bar{\mathbf{x}}$ 处满足Łojasiewicz性质, 若存在 $c > 0$ 、 $\theta \in [0, 1)$ 以及 $\eta > 0$, 使得

$$|G(\mathbf{x}) - G(\bar{\mathbf{x}})|^\theta \leq c \cdot \text{dist}(0, \partial G(\mathbf{x})), \quad \forall \mathbf{x} \in \mathcal{B}_\eta(\bar{\mathbf{x}}) := \{\mathbf{y} : \|\mathbf{y} - \bar{\mathbf{x}}\| \leq \eta\}.$$

这里, 我们规定 $0^0 = 0$. 因此, 若 $|G(\mathbf{x}) - G(\bar{\mathbf{x}})|^0 = 0$, 就有 $G(\mathbf{x}) = G(\bar{\mathbf{x}})$. 我们称 θ 为 G 在 $\bar{\mathbf{x}}$ 处的Łojasiewicz指数.

注. Łojasiewicz性质最早是针对解析(analytic)函数定义的^[190], 随后被推广到可定义(definable)函数类^[191]以及非光滑次解析(subanalytic)函数类^[173, 192, 193]. 已有工作表明, Łojasiewicz性质对许多优化问题中常见的函数都成立, 例如实解析函数^[194]、满足一定增长条件的凸函数^[192]、半代数(semialgebraic)函数^[193]等. 感兴趣的读者可参阅文献^[193]及其中的参考文献.

在文献^[189]中, 作者证明了如下一致型(uniformized)Łojasiewicz性质.

引理3.10 (函数的一致型Łojasiewicz性质^[189]). 设 $G : \mathbb{R}^n \rightarrow (-\infty, \infty]$ 为一适定下半连续函数, $S \subseteq \mathbb{R}^n$ 为 G 的一阶稳定点构成的连通紧集. 假设 G 在任一阶稳定点处都满足Łojasiewicz性质. 则 G 在 S 上是有限常数, 且存在 $c > 0$ 、 $\theta \in [0, 1)$ 以及 $\eta > 0$ 使得

$$|G(\mathbf{x}) - G(\bar{\mathbf{x}})|^\theta \leq c \cdot \text{dist}(0, \partial G(\mathbf{x})), \quad \forall \bar{\mathbf{x}} \in S, \mathbf{x} \in \{\mathbf{y} : \text{dist}(\mathbf{y}, S) \leq \eta\}.$$

我们称 θ 为 G 关于 S 的Łojasiewicz指数.

注. 引理 3.10 中的 θ 与 c 可用于 S 中任一单点处的 Łojasiewicz 性质. 事实上, 单点处的 Łojasiewicz 指数可以小于 θ , 代表在该点附近函数曲面更加尖锐 (sharp).

当 \mathbf{x} 同时满足 $\text{dist}(\mathbf{x}, S) < \eta$ 与 $|G(\mathbf{x}) - G(\bar{\mathbf{x}})| < 1$ 时, 我们可以增大 Łojasiewicz 指数.

推论 3.11 (抬升 Łojasiewicz 指数^[187,195]). 设 $G : \mathbb{R}^n \rightarrow (-\infty, \infty]$ 为一适当下半连续函数, $S \subseteq \mathbb{R}^n$ 为 G 的一阶稳定点构成的连通紧集. 假设 G 在任一一阶稳定点处都满足 Łojasiewicz 性质. 令 $c > 0$ 、 $\theta \in [0, 1)$ 以及 $\eta > 0$ 为引理 3.10 中与 G 和 S 相关的常数. 则对任意 $\bar{\theta} \in [\theta, 1)$, 都有

$$|G(\mathbf{x}) - G(\bar{\mathbf{x}})|^{\bar{\theta}} \leq c \cdot \text{dist}(0, \partial G(\mathbf{x})),$$

$$\forall \bar{\mathbf{x}} \in S, \mathbf{x} \in \{\mathbf{y} : \text{dist}(\mathbf{y}, S) < \eta\} \cap \{\mathbf{y} : |G(\mathbf{y}) - G(\bar{\mathbf{x}})| < 1\}.$$

我们称 $\bar{\theta}$ 为 G 关于 S 的抬升 (*lifted*) Łojasiewicz 指数.

我们对函数 F 的假设条件陈述如下:

条件 3.12. 函数 F 在任一一阶稳定点处均满足 Łojasiewicz 性质.

当 F 满足条件 3.12 时, 根据定理 3.9 与引理 3.10, 我们知道若 $\{\varepsilon_k\}$ 平方可和, 则 F 在集合 $S(\mathbf{x}^{(0)})$ 上满足一致型 Łojasiewicz 性质. 记 $c > 0$ 、 $\theta \in [0, 1)$ 以及 $\eta > 0$ 为其中与 $S = S(\mathbf{x}^{(0)})$ 和 $G = F$ 相关的常数.

定理 3.13 (PALM-I 算法的全局依点列收敛性). 假设条件 3.2、3.3 以及 3.12 成立. 令 $\{\mathbf{x}^{(k)}\}$ 为 PALM-I 算法产生的迭代点序列, 其中 $\{\varepsilon_k\}$ 满足条件 3.4 (2). 则 $\{\mathbf{x}^{(k)}\}$ 收敛于 F 的一个一阶稳定点.

证明. 不失一般性, 我们假设条件 3.4 (2) 中的 $\bar{\theta} \geq \theta$, 也即 $\bar{\theta}$ 是 F 关于 $S(\mathbf{x}^{(0)})$ 的抬升 Łojasiewicz 指数. 往证序列 $\{\|\mathbf{x}^{(k)} - \mathbf{x}^{(k+1)}\|\}$ 具有有限长度. 注意到对任意 $r_1, r_2 \in \mathbb{N}$ ($r_2 \geq r_1$),

$$\sum_{k=r_1}^{r_2} \|\mathbf{x}^{(k)} - \mathbf{x}^{(k+1)}\| \leq \sum_{k=r_1}^{r_2} \|\bar{\mathbf{x}}^{(k+1)} - \mathbf{x}^{(k)}\| + \sum_{k=r_1}^{r_2} \|\mathbf{D}\mathbf{x}^{(k+1)}\|. \quad (3.15)$$

而条件 3.4 (2) 中假设 $\{\varepsilon_k\}$ 可和, 因此 (3.15) 式右端第二项对 r_2 有一致上界. 于是, 我们只需要证明序列 $\{\|\bar{\mathbf{x}}^{(k+1)} - \mathbf{x}^{(k)}\|\}$ 是可和的. 下面分两种情况讨论.

情形一: 存在 $\tilde{k} \in \mathbb{N}$ 使得 $v_{\tilde{k}} - \bar{F} = 0$.

根据命题 3.7 (4), 我们可知对任意 $k \geq \tilde{k}$, 恒有 $v_k = \bar{F}$ 与 $\mathbf{x}^{(k)} = \bar{\mathbf{x}}^{(k+1)}$ 成立. 这便可直接导出 $\{\|\bar{\mathbf{x}}^{(k+1)} - \mathbf{x}^{(k)}\|\}$ 是可和的. 基于 (3.15) 式与定理 3.9 (1), $\{\mathbf{x}^{(k)}\}$ 必收敛到 F 的一个一阶稳定点.

事实上, 若还存在 $\hat{k} \in \mathbb{N}$ ($\hat{k} \geq \tilde{k}$) 使得 $\mathbf{x}^{(\hat{k})} = \bar{\mathbf{x}}^{(\hat{k})}$, 我们还可以得到更强的结果. 由命题 3.7 (4), 我们可知此时对任意 $k \geq \hat{k}$, 恒有 $\mathbf{x}^{(k)} = \bar{\mathbf{x}}^{(k+1)} = \mathbf{x}^{(k+1)}$ 成立. 于

是, 根据命题 3.8, 对任意 $k \geq \hat{k}$, 都有 $\mathbf{w}^{(k+1)} = 0$. 这就表明, PALM-I 算法在有限步内终止于 F 的一个一阶稳定点.

情形二: 对任意 $k \geq 0, v_k - \bar{F} > 0$.

在此情形下, 我们证明 $\{\|\bar{\mathbf{x}}^{(k+1)} - \mathbf{x}^{(k)}\|\}$ 满足一个递归关系. 令 $\bar{\varphi}(a) := c \cdot a^{1-\bar{\theta}}/(1-\bar{\theta})$ ($a > 0$). 因为 $v_k - \bar{F}$ 是正数, $\bar{\varphi}'(v_k - \bar{F})$ 良定. 由 $\bar{\varphi}$ 的凹性, 我们可得

$$\begin{aligned} D_{\bar{\varphi}}^{(k,k+1)} &:= \bar{\varphi}(v_k - \bar{F}) - \bar{\varphi}(v_{k+1} - \bar{F}) \geq \bar{\varphi}'(v_k - \bar{F})(v_k - v_{k+1}) \\ &\geq c \frac{C_{0,k} \|\bar{\mathbf{x}}^{(k+1)} - \mathbf{x}^{(k)}\|^2}{[F(\bar{\mathbf{x}}^{(k)}) - \bar{F} + u_k + u_{k+1}]^{\bar{\theta}}} \geq \frac{c(\gamma-1)L}{4} \frac{\|\bar{\mathbf{x}}^{(k+1)} - \mathbf{x}^{(k)}\|^2}{[F(\bar{\mathbf{x}}^{(k)}) - \bar{F} + u_k + u_{k+1}]^{\bar{\theta}}}, \end{aligned}$$

其中第二个不等式使用了命题 3.7 (2) 以及 v_k 的定义, 最后一个不等式使用了 $C_{0,k} \geq (\gamma-1)L/4$ (任意 $k \geq 0$). 从上述不等式可进一步得到

$$\begin{aligned} &\|\bar{\mathbf{x}}^{(k+1)} - \mathbf{x}^{(k)}\| \\ &\leq 2 \sqrt{\frac{1}{c(\gamma-1)L}} \sqrt{[F(\bar{\mathbf{x}}^{(k)}) - \bar{F} + u_k + u_{k+1}]^{\bar{\theta}} \cdot D_{\bar{\varphi}}^{(k,k+1)}} \\ &\leq \sqrt{\frac{1}{c(\gamma-1)L}} \left[\frac{1}{p} [F(\bar{\mathbf{x}}^{(k)}) - \bar{F} + u_k + u_{k+1}]^{\bar{\theta}} + p D_{\bar{\varphi}}^{(k,k+1)} \right], \end{aligned} \quad (3.16)$$

其中第二个不等式使用了均值不等式, 常数 p 满足

$$p > \sqrt{\frac{c}{(\gamma-1)L}} \bar{M}. \quad (3.17)$$

根据函数 $(\cdot)^{\bar{\theta}}$ 的性质、 u_k 的定义以及引理 3.1,

$$\begin{aligned} &[F(\bar{\mathbf{x}}^{(k)}) - \bar{F} + u_k + u_{k+1}]^{\bar{\theta}} \\ &\leq |F(\bar{\mathbf{x}}^{(k)}) - \bar{F}|^{\bar{\theta}} + (2u_k)^{\bar{\theta}} = |F(\bar{\mathbf{x}}^{(k)}) - \bar{F}|^{\bar{\theta}} + \left(2 \sum_{t=k}^{\infty} C_1^{(t)} \|D\mathbf{x}^{(t)}\|^2\right)^{\bar{\theta}} \\ &\leq |F(\bar{\mathbf{x}}^{(k)}) - \bar{F}|^{\bar{\theta}} + \bar{C}_1^{\bar{\theta}} \left(\sum_{t=k}^{\infty} \varepsilon_t^2 \right)^{\bar{\theta}} = |F(\bar{\mathbf{x}}^{(k)}) - \bar{F}|^{\bar{\theta}} + \bar{C}_1^{\bar{\theta}} e_k^{\bar{\theta}}. \end{aligned} \quad (3.18)$$

因为 $\{\varepsilon_k\}$ 可和 (条件 3.4 (2)), 再结合命题 3.7 (3) 与定理 3.9 (2), 可知存在 $k_1 \in \mathbb{N}$ 使得对任意 $k \geq k_1$,

$$\bar{\mathbf{x}}^{(k)} \in \{\mathbf{x} : \text{dist}(\mathbf{x}, S(\mathbf{x}^{(0)})) < \eta\} \cap \{\mathbf{x} : |F(\mathbf{x}) - \bar{F}| < 1\}.$$

于是, 可从命题 3.8 与推论 3.11 共同推知, 对任意 $k \geq k_1$,

$$\begin{aligned} &c \bar{M} [\|\bar{\mathbf{x}}^{(k)} - \mathbf{x}^{(k-1)}\| + \|D\mathbf{x}^{(k)}\|] \geq c \|\mathbf{w}^{(k)}\| \\ &\geq c \cdot \text{dist}(0, \partial F(\bar{\mathbf{x}}^{(k)})) \geq |F(\bar{\mathbf{x}}^{(k)}) - \bar{F}|^{\bar{\theta}}. \end{aligned} \quad (3.19)$$

将(3.19)式代入(3.18)式,即有对任意 $k \geq k_1$,

$$[F(\bar{\mathbf{x}}^{(k)}) - \bar{F} + u_k + u_{k+1}]^{\bar{\theta}} \leq c\bar{M} [\|\bar{\mathbf{x}}^{(k)} - \mathbf{x}^{(k-1)}\| + \|\mathbf{D}\mathbf{x}^{(k)}\|] + \bar{C}_1^{\bar{\theta}} e_k^{\bar{\theta}}. \quad (3.20)$$

再将(3.20)式代入(3.16)式,可知对任意 $k \geq k_1$,

$$\sqrt{c(\gamma-1)L} \|\bar{\mathbf{x}}^{(k+1)} - \mathbf{x}^{(k)}\| \leq \frac{c\bar{M}}{p} [\|\bar{\mathbf{x}}^{(k)} - \mathbf{x}^{(k-1)}\| + \|\mathbf{D}\mathbf{x}^{(k)}\|] + \frac{\bar{C}_1^{\bar{\theta}}}{p} e_k^{\bar{\theta}} + pD_{\bar{\varphi}}^{(k,k+1)}.$$

将上面的不等式按指标 k 从 r_1 累加至 r_2 ($r_2 \geq r_1 \geq k_1$),我们有

$$\begin{aligned} & \sqrt{c(\gamma-1)L} \sum_{k=r_1}^{r_2} \|\bar{\mathbf{x}}^{(k+1)} - \mathbf{x}^{(k)}\| \\ & \leq \frac{c\bar{M}}{p} \left[\sum_{k=r_1-1}^{r_2} \|\bar{\mathbf{x}}^{(k+1)} - \mathbf{x}^{(k)}\| + \sum_{k=r_1}^{r_2} \|\mathbf{D}\mathbf{x}^{(k)}\| \right] + \frac{\bar{C}_1^{\bar{\theta}}}{p} \sum_{k=r_1}^{r_2} e_k^{\bar{\theta}} + pD_{\bar{\varphi}}^{(r_1,r_2+1)} \\ & \leq \frac{c\bar{M}}{p} \left[\sum_{k=r_1-1}^{r_2} \|\bar{\mathbf{x}}^{(k+1)} - \mathbf{x}^{(k)}\| + \omega \sum_{k=1}^{\infty} \varepsilon_k \right] + \frac{\bar{C}_1^{\bar{\theta}}}{p} \sum_{k=1}^{\infty} e_k^{\bar{\theta}} + p\bar{\varphi}(v_{r_1} - \bar{F}), \end{aligned}$$

其中第二个不等式使用了引理3.1、条件3.4(2)以及 $\bar{\varphi}$ 在 $(0, \infty)$ 上恒正. 我们可以进一步推得

$$\begin{aligned} & \left[\sqrt{c(\gamma-1)L} - \frac{c\bar{M}}{p} \right] \sum_{k=r_1}^{r_2} \|\bar{\mathbf{x}}^{(k+1)} - \mathbf{x}^{(k)}\| \\ & \leq \frac{c\bar{M}}{p} \left[\|\bar{\mathbf{x}}^{(r_1)} - \mathbf{x}^{(r_1-1)}\| + \omega \sum_{k=1}^{\infty} \varepsilon_k \right] + \frac{\bar{C}_1^{\bar{\theta}}}{p} \sum_{k=1}^{\infty} e_k^{\bar{\theta}} + p\bar{\varphi}(v_{r_1} - \bar{F}). \end{aligned} \quad (3.21)$$

根据 p 的取值(见(3.17)式),(3.21)式左端的系数为一正常数. 因此,在条件3.4(2)下,(3.21)式表明 $\{\|\bar{\mathbf{x}}^{(k+1)} - \mathbf{x}^{(k)}\|\}$ 可和. 基于(3.15)式与定理3.9(1), $\{\mathbf{x}^{(k)}\}$ 必收敛到 F 的一个一阶稳定点.

综合上述两款,定理得证. \square

注. 当与 F 和集合 $S(\mathbf{x}^{(0)})$ 相关的Łojasiewicz指数 θ 为0时,将其抬升为正的 $\bar{\theta}$ 对于依点列收敛性的证明十分重要. 事实上,在(3.16)式中,若依然使用 $\bar{\theta} = \theta = 0$,我们只会得到 $\{\|\bar{\mathbf{x}}^{(k+1)} - \mathbf{x}^{(k)}\|\}$ 的平方可和性. 这与PALM-E算法的证明完全不同. 对PALM-E算法, u_k 恒为0. 因此 θ 为0可以导出PALM-E算法的有限终止性^[173,176].

3.4 漸进收敛速度分析

在本节中,我们基于定理3.13研究PALM-I算法的漸进收敛速度. 以下, θ 代表 F 在 $\{\mathbf{x}^{(k)}\}$ 极限 $\bar{\mathbf{x}}$ 处的Łojasiewicz指数. 记

$$S_t := \sum_{k=t}^{\infty} \|\bar{\mathbf{x}}^{(k+1)} - \mathbf{x}^{(k)}\|, \quad \forall t \geq 0.$$

在定理 3.13 的假设条件下, $S_0 < \infty$, $\sum_{k=1}^{\infty} \|D\mathbf{x}^{(k)}\| < \infty$. 易知

$$\|\mathbf{x}^{(t)} - \bar{\mathbf{x}}\| \leq \sum_{k=t}^{\infty} \{\|\bar{\mathbf{x}}^{(k+1)} - \mathbf{x}^{(k)}\| + \|\bar{\mathbf{x}}^{(k+1)} - \mathbf{x}^{(k+1)}\|\} \leq S_t + \omega \sum_{k=t}^{\infty} \varepsilon_{k+1}, \quad (3.22)$$

其中第二个不等式使用了引理 3.1.

为得到确切的收敛速度, 我们考虑指数下降型与次线性下降型的误差控制序列 $\{\varepsilon_k\}$. 我们首先研究特殊情形: 存在 $K \in \mathbb{N}$ 使得 $v_K = \bar{F}$. 根据命题 3.7 (4), 此时 $S_t \equiv 0$ (对所有充分大的 t), 于是从 (3.22) 式可知算法的渐进收敛速度只取决于 $\{\varepsilon_k\}$ 的选取. 下面的定理较为显然. 我们略去其证明.

定理 3.14 ($\{v_k\}$ 有限终止时 PALM-I 算法的渐进收敛速度). 假设定理 3.13 中的条件均成立. 令 $\bar{\mathbf{x}}$ 为 PALM-I 算法产生的迭代点序列 $\{\mathbf{x}^{(k)}\}$ 的唯一极限. 假设存在 $K \in \mathbb{N}$ 使得 $v_K = \bar{F}$. 则有如下结论成立:

(1) 若 $\varepsilon_k = \bar{\varepsilon}\tilde{\rho}^k$ (任意 $k \geq 0$), 其中 $\tilde{\rho} \in (0, 1)$, 则

$$\|\mathbf{x}^{(k)} - \bar{\mathbf{x}}\| \leq \mathcal{O}(\tilde{\rho}^k), \quad \forall k \geq K;$$

(2) 若 $\varepsilon_k = \bar{\varepsilon}/(k+1)^l$ (任意 $k \geq 0$), 其中 $l > 1$, 则

$$\|\mathbf{x}^{(k)} - \bar{\mathbf{x}}\| \leq \mathcal{O}\left(\frac{1}{k^{l-1}}\right), \quad \forall k \geq K.$$

由于子问题的求解非精确, 在实际计算中 $v_k = \bar{F}$ 几乎不会发生. 在下面两小节中, 我们研究 $v_k - \bar{F} > 0$ (任意 $k \geq 0$) 的情形. 在此之前, 我们先证明 $\{S_t\}$ 满足的一个递归关系, 并介绍一个关于数列收敛速度的引理.

引理 3.15 ($\{S_t\}$ 上的递归关系). 假设定理 3.13 中的条件均成立且 $\bar{\theta} \in [\theta, 1)$. 令 $\bar{\mathbf{x}}$ 为 PALM-I 算法产生的迭代点序列 $\{\mathbf{x}^{(k)}\}$ 的唯一极限. 假设对任意 $k \geq 0$, $v_k - \bar{F} > 0$. 则

$$S_t \leq [S_{t-1} - S_t] + C_2 [S_{t-1} - S_t + \omega\varepsilon_t]^{\frac{1-\bar{\theta}}{\bar{\theta}}} + C_3 E_{t,\bar{\theta}}, \quad \forall t \geq 1,$$

其中

$$\begin{aligned} E_{t,\bar{\theta}} &:= \sum_{k=t}^{\infty} \varepsilon_k + \sum_{k=t}^{\infty} e_k^{\bar{\theta}} + e_t^{1-\bar{\theta}}, \\ p &:= 2\sqrt{\frac{c}{(\gamma-1)L}}\bar{M}, \quad q := \frac{\sqrt{c(\gamma-1)L}}{2}, \\ C_2 &:= \frac{cp(c\bar{M})^{\frac{1-\bar{\theta}}{\bar{\theta}}}}{q(1-\bar{\theta})}, \quad C_3 := \max \left\{ \omega, \frac{\bar{C}_1^{\bar{\theta}}}{pq}, \frac{cp\bar{C}_1^{1-\bar{\theta}}}{q(1-\bar{\theta})} \right\}. \end{aligned}$$

证明. 由定理 3.13 证明中的 (3.21) 式, 我们可以推出如下 S_t 的上界:

$$S_t \leq \|\bar{\mathbf{x}}^{(t)} - \mathbf{x}^{(t-1)}\| + \omega \sum_{k=t}^{\infty} \varepsilon_k + \frac{\bar{C}_1^{\bar{\theta}}}{pq} \sum_{k=t}^{\infty} e_k^{\bar{\theta}} \quad (3.23)$$

$$\begin{aligned}
& + \frac{cp}{q(1-\bar{\theta})} \left[F(\bar{\mathbf{x}}^{(t)}) - \bar{F} + 2 \sum_{k=t}^{\infty} C_{1,t} \|\mathbf{D}\mathbf{x}^{(k)}\|^2 \right]^{1-\bar{\theta}} \\
& \leq [S_{t-1} - S_t] + \omega \sum_{k=t}^{\infty} \varepsilon_k + \frac{\bar{C}_1^{\bar{\theta}}}{pq} \sum_{k=t}^{\infty} e_k^{\bar{\theta}} + \frac{cp}{q(1-\bar{\theta})} \left[|F(\bar{\mathbf{x}}^{(t)}) - \bar{F}|^{1-\bar{\theta}} + \bar{C}_1^{1-\bar{\theta}} e_t^{1-\bar{\theta}} \right],
\end{aligned}$$

其中第二个不等式使用了函数 $(\cdot)^{1-\bar{\theta}}$ 的性质、引理 3.1 以及条件 3.4 (2). 由命题 3.8 和推论 3.11, 可知

$$\begin{aligned}
|F(\bar{\mathbf{x}}^{(t)}) - \bar{F}|^{\bar{\theta}} & \leq c \cdot \text{dist}(0, \partial F(\bar{\mathbf{x}}^{(t)})) \leq c \|\mathbf{w}^{(t)}\| \\
& \leq c \bar{M} [\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}^{(t-1)}\| + \|\mathbf{D}\mathbf{x}^{(t)}\|] \leq c \bar{M} [S_{t-1} - S_t + \omega \varepsilon_t],
\end{aligned}$$

其中最后一个不等式使用了引理 3.1. 因为 $(1-\bar{\theta})/\bar{\theta} > 0$,

$$|F(\bar{\mathbf{x}}^{(t)}) - \bar{F}|^{1-\bar{\theta}} \leq (c \bar{M})^{\frac{1-\bar{\theta}}{\bar{\theta}}} [S_{t-1} - S_t + \omega \varepsilon_t]^{\frac{1-\bar{\theta}}{\bar{\theta}}}. \quad (3.24)$$

将 (3.24) 式代入 (3.23) 式后即有

$$\begin{aligned}
S_t & \leq [S_{t-1} - S_t] + \frac{cp(c\bar{M})^{\frac{1-\bar{\theta}}{\bar{\theta}}}}{q(1-\bar{\theta})} [S_{t-1} - S_t + \omega \varepsilon_t]^{\frac{1-\bar{\theta}}{\bar{\theta}}} \\
& + \omega \sum_{k=t}^{\infty} \varepsilon_k + \frac{\bar{C}_1^{\bar{\theta}}}{pq} \sum_{k=t}^{\infty} e_k^{\bar{\theta}} + \frac{cp\bar{C}_1^{1-\bar{\theta}}}{q(1-\bar{\theta})} e_t^{1-\bar{\theta}},
\end{aligned}$$

再结合 C_2 、 C_3 以及 $E_{t,\bar{\theta}}$ 的定义即可得证. \square

引理 3.16 (非负数列的收敛速度^[196]). 设 $\{a_k\}$ 为一非负数列. 若

$$a_{k+1} \leq \left(1 - \frac{b}{k}\right) a_k + \frac{d}{k^{r+1}},$$

其中 b 、 d 和 r 是正常数, 且 $b > r$, 则

$$a_k \leq \frac{d}{b-r} \frac{1}{k^r} + o\left(\frac{1}{k^r}\right).$$

3.4.1 误差控制序列指数下降情形

定理 3.17 (误差控制序列指数下降时 PALM-I 算法的渐进收敛速度). 假设定理 3.13 中的条件均成立且 $\varepsilon_k = \bar{\varepsilon} \tilde{\rho}^k$ (任意 $k \geq 0$), 其中 $\tilde{\rho} \in (0, 1)$. 令 $\bar{\mathbf{x}}$ 为 PALM-I 算法产生的迭代点序列 $\{\mathbf{x}^{(k)}\}$ 的唯一极限, $\theta \in [0, 1)$ 为 F 在 $\bar{\mathbf{x}}$ 处的 Łojasiewicz 指数. 假设对任意 $k \geq 0$, $v_k - \bar{F} > 0$. 则有如下结论成立:

(1) 若 $\theta = 0$, 则存在 $\rho_1 \in (0, 1)$ 使得

$$\|\mathbf{x}^{(k)} - \bar{\mathbf{x}}\| \leq \mathcal{O}(\rho_1^k), \quad \forall k \text{ 充分大};$$

(2) 若 $\theta \in (0, 1/2]$, 则存在 $\rho_2 \in (0, 1)$ 使得

$$\|\mathbf{x}^{(k)} - \bar{\mathbf{x}}\| \leq \mathcal{O}(\rho_2^k), \quad \forall k \text{ 充分大};$$

(3) 若 $\theta \in (1/2, 1)$, 则

$$\|\mathbf{x}^{(k)} - \bar{\mathbf{x}}\| \leq \mathcal{O}\left(k^{-\frac{1-\theta}{2\theta-1}}\right), \quad \forall k \text{ 充分大.}$$

(1) 与 (2) 的证明. 根据 $\{\varepsilon_k\}$ 的选取可知, 任意使得 $\theta \leq \bar{\theta} \in (0, 1/2]$ 成立的 $\bar{\theta}$ 都满足条件 3.4 (2). 因此引理 3.15 对于这样的 $\bar{\theta}$ 都适用. 又由 (3.14) 式和 $\{\varepsilon_k\}$ 的选取, 存在 $k_2 \in \mathbb{N} (k_2 \geq k_1)$ 使得

$$S_{t-1} - S_t + \omega \varepsilon_t \in [0, 1], \quad \forall t \geq k_2.$$

因为对 $\bar{\theta} \in (0, 1/2]$, 有 $(1 - \bar{\theta})/\bar{\theta} \geq 1$, 所以

$$[S_{t-1} - S_t + \omega \varepsilon_t]^{\frac{1-\bar{\theta}}{\bar{\theta}}} \leq S_{t-1} - S_t + \omega \varepsilon_t, \quad \forall t \geq k_2.$$

将上述不等式代入引理 3.15 的结论, 可得

$$S_t \leq \bar{\rho} S_{t-1} + \bar{C}_2 \varepsilon_t + \bar{C}_3 E_{t,\bar{\theta}}, \quad \forall t \geq k_2,$$

其中

$$\bar{\rho} := \frac{1+C_2}{2+C_2} \in (0, 1), \quad \bar{C}_2 := \frac{\omega C_2}{2+C_2}, \quad \bar{C}_3 := \frac{C_3}{2+C_2}.$$

反复调用上述递归关系并结合 $\{\varepsilon_k\}$ 的选取, 即有对任意 $t \geq k_2$,

$$\begin{aligned} S_t &\leq \bar{\rho}^{t-k_2+1} S_{k_2-1} + \bar{C}_2 \sum_{k=k_2}^t \bar{\rho}^{t-k} \tilde{\rho}^k + \bar{C}_3 \sum_{k=k_2}^t \bar{\rho}^{t-k} E_{k,\bar{\theta}} \\ &\leq \bar{\rho}^{t-k_2+1} S_{k_2-1} + \bar{C}_2 t \max\{\bar{\rho}, \tilde{\rho}\}^t + \bar{C}_3 \sum_{k=k_2}^t \bar{\rho}^{t-k} E_{k,\bar{\theta}}. \end{aligned} \quad (3.25)$$

我们再计算 $E_{t,\bar{\theta}}$: 对任意 $t \geq 1$,

$$E_{t,\bar{\theta}} = \sum_{k=t}^{\infty} \varepsilon_k + \sum_{k=t}^{\infty} e_k^{\bar{\theta}} + e_t^{1-\bar{\theta}} = \frac{\bar{\varepsilon} \tilde{\rho}^t}{1-\tilde{\rho}} + \frac{\bar{\varepsilon}^{2\bar{\theta}} \tilde{\rho}^{2\bar{\theta}t}}{(1-\tilde{\rho}^2)^{\bar{\theta}}(1-\tilde{\rho}^{2\bar{\theta}})} + \frac{\bar{\varepsilon}^{2(1-\bar{\theta})} \tilde{\rho}^{2(1-\bar{\theta})t}}{(1-\tilde{\rho}^2)^{1-\bar{\theta}}}.$$

因为 $\bar{\theta} \in (0, 1/2]$, 所以 $2\bar{\theta} \leq \min\{2(1-\bar{\theta}), 1\}$. 由上述不等式就可推知, 存在正常数 \bar{M}_1 使得

$$E_{t,\bar{\theta}} \leq \bar{M}_1 \tilde{\rho}^{2\bar{\theta}t}, \quad \forall t \geq 1.$$

将此代入 (3.25) 式即有

$$S_t \leq \bar{\rho}^{t-k_2+1} S_{k_2-1} + \bar{C}_2 t \max\{\bar{\rho}, \tilde{\rho}\}^t + \bar{C}_3 \bar{M}_1 t \max\{\bar{\rho}, \tilde{\rho}^{2\bar{\theta}}\}^t, \quad \forall t \geq k_2.$$

因为 $\max\{\bar{\rho}, \tilde{\rho}^{2\bar{\theta}}\} \in (0, 1)$, 所以存在 $k_3 \in \mathbb{N} (k_3 \geq k_2)$ 使得

$$S_t \leq \mathcal{O}\left(\max\{\bar{\rho}, \tilde{\rho}^{2\bar{\theta}}\}^{\frac{t}{2}}\right), \quad \forall t \geq k_3.$$

最后结合 (3.22) 式即得证. \square

(3) 的证明. 根据 $\{\varepsilon_k\}$ 的选取可知, $\bar{\theta} = \theta$ 即可满足条件 3.4 (2). 因此引理 3.15 对 θ 适用. 又由 (3.14) 式和 $\{\varepsilon_k\}$ 的选取, 存在 $k_4 \in \mathbb{N}$ ($k_4 \geq k_1$) 使得

$$S_{t-1} - S_t + \omega\varepsilon_t \in [0, 1], \quad \forall t \geq k_4.$$

因为 $\theta \in (1/2, 1)$, 所以 $(1-\theta)/\theta < 1$, 进而根据函数 $(\cdot)^{(1-\theta)/\theta}$ 的性质, 对任意 $t \geq k_4$,

$$\begin{aligned} [S_{t-1} - S_t + \omega\varepsilon_t]^{\frac{1-\theta}{\theta}} &\leq [S_{t-1} - S_t]^{\frac{1-\theta}{\theta}} + \omega\varepsilon_t^{\frac{1-\theta}{\theta}}, \\ S_{t-1} - S_t &\leq [S_{t-1} - S_t]^{\frac{1-\theta}{\theta}}. \end{aligned}$$

将上述两个不等式代入引理 3.15, 即知对任意 $t \geq k_4$,

$$S_{t-1} = [S_{t-1} - S_t] + S_t \leq (2 + C_2) [S_{t-1} - S_t]^{\frac{1-\theta}{\theta}} + \omega C_2 \varepsilon_t^{\frac{1-\theta}{\theta}} + C_3 E_{t,\theta}. \quad (3.26)$$

对 (3.26) 式右端后两项, 结合 $\{\varepsilon_k\}$ 的选取简单计算可得

$$\omega C_2 \varepsilon_t^{\frac{1-\theta}{\theta}} + C_3 E_{t,\theta} = \omega C_2 (\bar{\varepsilon} \tilde{\rho}^t)^{\frac{1-\theta}{\theta}} + C_3 \left[\frac{\bar{\varepsilon} \tilde{\rho}^t}{1-\tilde{\rho}} + \frac{\bar{\varepsilon}^{2\theta} \tilde{\rho}^{2\theta t}}{(1-\tilde{\rho}^{2\theta})(1-\tilde{\rho}^2)^\theta} + \frac{\bar{\varepsilon}^{2(1-\theta)} \tilde{\rho}^{2(1-\theta)t}}{(1-\tilde{\rho}^2)^\theta} \right].$$

将此代入 (3.26) 式, 并由 $(1-\theta)/\theta \leq \min\{1, 2\theta, 2(1-\theta)\}$, 可知存在 $\bar{M}_2 > 0$, 使得

$$S_{t-1} \leq (2 + C_2) [S_{t-1} - S_t]^{\frac{1-\theta}{\theta}} + \bar{M}_2 \tilde{\rho}^{\frac{1-\theta}{\theta}t}, \quad \forall t \geq k_4.$$

因为 $\theta/(1-\theta) > 1$, 对上式使用 Minkowski 不等式, 可进一步得到

$$S_{t-1}^{\frac{\theta}{1-\theta}} \leq 2^{\frac{2\theta-1}{1-\theta}} \left[(2 + C_2)^{\frac{\theta}{1-\theta}} (S_{t-1} - S_t) + \bar{M}_2^{\frac{\theta}{1-\theta}} \tilde{\rho}^t \right], \quad \forall t \geq k_4,$$

从而

$$S_t \leq S_{t-1} - C_4 S_{t-1}^{\frac{\theta}{1-\theta}} + C_5 \tilde{\rho}^t, \quad \forall t \geq k_4, \quad (3.27)$$

其中

$$C_4 := 2^{-\frac{2\theta-1}{1-\theta}} (2 + C_2)^{-\frac{\theta}{1-\theta}}, \quad C_5 := \bar{M}_2^{\frac{\theta}{1-\theta}} (2 + C_2)^{-\frac{\theta}{1-\theta}}.$$

考虑 $h_\theta : \mathbb{R}_+ \rightarrow \mathbb{R}$ 定义为 $h_\theta(a) := a^{\theta/(1-\theta)}$. 因为 $\theta/(1-\theta) > 1$, 所以 h_θ 在 \mathbb{R}_+ 上是凸函数. 因此对任意 $\xi \geq 0$,

$$\begin{aligned} S_{t-1}^{\frac{\theta}{1-\theta}} - (\xi t^{-\frac{1-\theta}{2\theta-1}})^{\frac{\theta}{1-\theta}} &= h_\theta(S_{t-1}) - h_\theta(\xi t^{-\frac{1-\theta}{2\theta-1}}) \\ &\geq h'_\theta(\xi t^{-\frac{1-\theta}{2\theta-1}}) \left[S_{t-1} - \xi t^{-\frac{1-\theta}{2\theta-1}} \right] = \frac{\theta \xi^{\frac{2\theta-1}{1-\theta}}}{(1-\theta)t} \left[S_{t-1} - \xi t^{-\frac{1-\theta}{2\theta-1}} \right]. \end{aligned}$$

将上式代入 (3.27) 式, 可知对任意 $t \geq \bar{k}_4 := \max\{e, k_4, \theta/[(2\theta-1)\ln(1/\tilde{\rho})]\}$,

$$S_t \leq S_{t-1} - C_4 \left[S_{t-1}^{\frac{\theta}{1-\theta}} - (\xi t^{-\frac{1-\theta}{2\theta-1}})^{\frac{\theta}{1-\theta}} \right] - C_4 (\xi t^{-\frac{1-\theta}{2\theta-1}})^{\frac{\theta}{1-\theta}} + C_5 \tilde{\rho}^t$$

$$\begin{aligned}
 &\leq S_{t-1} - \frac{C_4 \theta \xi^{\frac{2\theta-1}{1-\theta}}}{(1-\theta)t} \left[S_{t-1} - \xi t^{-\frac{1-\theta}{2\theta-1}} \right] - C_4 (\xi t^{-\frac{1-\theta}{2\theta-1}})^{\frac{\theta}{1-\theta}} + C_5 \tilde{\rho}^t \\
 &= \left[1 - \frac{C_4 \theta \xi^{\frac{2\theta-1}{1-\theta}}}{(1-\theta)t} \right] S_{t-1} + \frac{C_4 \frac{2\theta-1}{1-\theta} \xi^{\frac{\theta}{1-\theta}}}{t^{\frac{\theta}{2\theta-1}}} + C_5 \tilde{\rho}^t \\
 &\leq \left[1 - \frac{C_4 \theta \xi^{\frac{2\theta-1}{1-\theta}}}{(1-\theta)t} \right] S_{t-1} + \frac{C_4 \frac{2\theta-1}{1-\theta} \xi^{\frac{\theta}{1-\theta}} + C_5}{t^{\frac{\theta}{2\theta-1}}}.
 \end{aligned}$$

因此, 只需选取 ξ 满足

$$C_4 \theta \xi^{\frac{2\theta-1}{1-\theta}} > \frac{1-\theta}{2\theta-1},$$

我们即可使用引理 3.16 得到

$$S_t \leq \mathcal{O}\left(t^{-\frac{1-\theta}{2\theta-1}}\right), \quad \forall t \geq \bar{k}_4.$$

最后结合 (3.22) 式即得证. \square

注. 由于子问题非精确求解存在误差, 当 θ 为 0 时, 原本 PALM-E 算法的有限终止性^[173,176] 无法复刻给 PALM-I 算法. 此外, 我们在定理 3.17 (1) 和 (2) 中分别指定 ρ_1 和 ρ_2 以表示两种情形下的渐进收敛速度可能并不相同.

3.4.2 误差控制序列次线性下降情形

当 $\{\varepsilon_k\}$ 次线性下降时, 不论 θ 取值如何, 我们只能证明 PALM-I 算法的 R-次线性渐进收敛速度. 从 (3.22) 式看这是十分合理的. 在此之前, 我们给出如下引理, 其证明省略.

引理 3.18. 设 $\theta \in (1/2, 1)$ 和 $l > (\theta + 1)/2\theta$. 考虑

$$\tau(\theta, l) := \min \left\{ \frac{1-\theta}{\theta} l, l-1, (2l-1)\theta - 1, (2l-1)(1-\theta) \right\}.$$

则 $\tau(\theta, l)$ 有如下显式表达式:

$$\tau(\theta, l) = \begin{cases} \frac{1-\theta}{\theta} l, & \text{若 } l \in \left[\frac{\theta}{2\theta-1}, \infty \right); \\ (2l-1)\theta - 1, & \text{若 } l \in \left(\frac{\theta+1}{2\theta}, \frac{\theta}{2\theta-1} \right]. \end{cases}$$

定理 3.19 (误差控制序列次线性下降时 PALM-I 算法的渐进收敛速度). 假设定理 3.13 中的条件均成立且 $\varepsilon_k = \bar{\varepsilon}/(k+1)^l$ (任意 $k \geq 0$), 其中 $l > 1$. 令 \bar{x} 为 PALM-I 算法产生的迭代点序列 $\{\mathbf{x}^{(k)}\}$ 的唯一极限, $\theta \in [0, 1)$ 为 F 在 \bar{x} 处的 Lojasiewicz 指数. 假设对任意 $k \geq 0$, $v_k - \bar{F} > 0$. 则对所有充分大的 k ,

$$\|\mathbf{x}^{(k)} - \bar{x}\| \leq \begin{cases} \mathcal{O}\left(k^{-\frac{1-\theta}{2\theta-1}}\right), & \text{若 } l \in \left[\frac{\theta}{2\theta-1}, \infty \right) \text{ 且 } \theta \in \left(\frac{1}{2}, 1 \right); \\ \mathcal{O}(k^{-(l-1)}), & \text{否则.} \end{cases} \quad (3.28)$$

证明. 此定理的证明类似于定理 3.17 (3). 根据条件 3.4 下方之注, 任意使

$$\max \left\{ \sqrt{\frac{1}{2l-1}}, \frac{1}{2} \right\} < \bar{\theta} \in [\theta, 1)$$

成立的 $\bar{\theta}$ 都满足条件 3.4 (2). 因此引理 3.15 对于这样的 $\bar{\theta}$ 都适用. 又由 (3.14) 式和 $\{\varepsilon_k\}$ 的选取, 存在 $k_5 \in \mathbb{N}$ ($k_5 \geq k_1$) 使得

$$S_{t-1} - S_t + \omega \varepsilon_t \in [0, 1], \quad \forall t \geq k_5.$$

对 (3.26) 式右端后两项, 结合 $\{\varepsilon_k\}$ 的选取计算可得

$$\begin{aligned} & \omega C_2 \varepsilon_t^{\frac{1-\bar{\theta}}{\bar{\theta}}} + C_3 E_{t,\bar{\theta}} \\ &= \omega C_2 \varepsilon_t^{\frac{1-\bar{\theta}}{\bar{\theta}}} + C_3 \left[\sum_{k=t}^{\infty} \varepsilon_k + \sum_{k=t}^{\infty} e_k^{\bar{\theta}} + e_t^{1-\bar{\theta}} \right] \\ &\leq \frac{\omega C_2}{(t+1)^{\frac{1-\bar{\theta}}{\bar{\theta}}l}} + C_3 \left[\sum_{k=t}^{\infty} \frac{\bar{\varepsilon}}{(k+1)^l} + \sum_{k=t}^{\infty} \frac{\bar{\varepsilon}^{2\bar{\theta}}}{(2l-1)^{\bar{\theta}} k^{(2l-1)\bar{\theta}}} + \left(\sum_{k=t}^{\infty} \frac{\bar{\varepsilon}^2}{(k+1)^{2l}} \right)^{1-\bar{\theta}} \right] \\ &\leq \frac{\omega C_2}{t^{\frac{1-\bar{\theta}}{\bar{\theta}}l}} + C_3 \left[\frac{\bar{\varepsilon}}{(l-1)t^{l-1}} + \frac{\bar{\varepsilon}^{2\bar{\theta}}[(2l-1)^{\bar{\theta}}-1]}{(t-1)^{(2l-1)\bar{\theta}-1}} + \frac{\bar{\varepsilon}^{2(1-\bar{\theta})}}{(2l-1)^{1-\bar{\theta}} t^{(2l-1)(1-\bar{\theta})}} \right]. \end{aligned}$$

根据引理 3.18 中函数 τ 的定义, 将上述不等式代入 (3.26) 式可知存在 $\bar{M}_3 > 0$ 使得

$$S_{t-1} \leq (2 + C_2) [S_{t-1} - S_t]^{\frac{1-\bar{\theta}}{\bar{\theta}}} + \frac{\bar{M}_3}{t^{\tau(\bar{\theta}, l)}}, \quad \forall t \geq k_5.$$

经过与定理 3.17 (3) 证明类似的推导, 可知对任意 $\xi \geq 0$,

$$S_t \leq \left[1 - \frac{C_4 \bar{\theta} \xi^{\frac{2\bar{\theta}-1}{1-\bar{\theta}}}}{(1-\bar{\theta})t} \right] S_{t-1} + \frac{C_4 \frac{2\bar{\theta}-1}{1-\bar{\theta}} \xi^{\frac{\bar{\theta}}{1-\bar{\theta}}} + C_6}{t^{\min\{\frac{\bar{\theta}}{2\bar{\theta}-1}, \tau(\bar{\theta}, l)\frac{\bar{\theta}}{1-\bar{\theta}}\}}}, \quad \forall t \geq k_5,$$

其中

$$C_6 := \bar{M}_3^{\frac{\bar{\theta}}{1-\bar{\theta}}} (2 + C_2)^{-\frac{\bar{\theta}}{1-\bar{\theta}}}.$$

注意到由于 $\bar{\theta} > \sqrt{1/(2l-1)}$, 因此

$$\min \left\{ \frac{\bar{\theta}}{2\bar{\theta}-1}, \tau(\bar{\theta}, l) \frac{\bar{\theta}}{1-\bar{\theta}} \right\} > 1.$$

所以, 只需选取 ξ 满足

$$C_4 \bar{\theta} \xi^{\frac{2\bar{\theta}-1}{1-\bar{\theta}}} > \min \left\{ \frac{\bar{\theta}}{2\bar{\theta}-1}, \tau(\bar{\theta}, l) \frac{\bar{\theta}}{1-\bar{\theta}} \right\} - 1,$$

我们就可使用引理 3.16 得到

$$S_t \leq \mathcal{O} \left(t^{-[\min\{\frac{\bar{\theta}}{2\bar{\theta}-1}, \tau(\bar{\theta}, l)\frac{\bar{\theta}}{1-\bar{\theta}}\}-1]} \right), \quad \forall t \geq k_5.$$

根据引理 3.18 并结合 (3.22) 式以及 $\{\varepsilon_k\}$ 的选取, 不难推得

$$\|\mathbf{x}^{(t)} - \bar{\mathbf{x}}\| \leq \begin{cases} \mathcal{O}\left(t^{-\frac{1-\bar{\theta}}{2\bar{\theta}-1}}\right), & \text{若 } l \in \left[\frac{\bar{\theta}}{2\bar{\theta}-1}, \infty\right); \\ \mathcal{O}\left(t^{-\frac{2\bar{\theta}^2l-\bar{\theta}^2-1}{1-\bar{\theta}}}\right), & \text{若 } l \in \left(\frac{\bar{\theta}^2+1}{2\bar{\theta}^2}, \frac{\bar{\theta}}{2\bar{\theta}-1}\right), \end{cases} \quad \forall t \geq k_5. \quad (3.29)$$

因为 (3.29) 式对任意满足

$$\max \left\{ \sqrt{\frac{1}{2l-1}}, \frac{1}{2} \right\} < \bar{\theta} \in [\theta, 1)$$

的 $\bar{\theta}$ 都成立, 且 k_5 与 $\bar{\theta}$ 的选取无关, 所以最好的收敛速度负指数必定是如下二者中的一个:

$$(A) \max_{\bar{\theta} \in \left[\frac{l}{2l-1}, 1\right], \bar{\theta} \geq \theta} \frac{1-\bar{\theta}}{2\bar{\theta}-1}; \quad (B) \max_{\bar{\theta} \in \left(\sqrt{\frac{1}{2l-1}}, \frac{l}{2l-1}\right], \bar{\theta} \geq \theta} \frac{2\bar{\theta}^2l-\bar{\theta}^2-1}{1-\bar{\theta}}.$$

注意到 $\bar{\theta} \geq l/(2l-1)$ 对任意 $\bar{\theta} \geq \theta$ 成立当且仅当 $l \geq \theta/(2\theta-1)$ 与 $\theta \in (1/2, 1)$. 若 $l \geq \theta/(2\theta-1)$ 且 $\theta \in (1/2, 1)$, 则最好的收敛速度负指数必定由 (A) 在 $\bar{\theta} = \theta$ 取到. 这就证明了 (3.28) 式的第一行. 若不然, 则易知 (A) 和 (B) 的最优值 $l-1$ 都在 $\bar{\theta} = l/(2l-1)$ 取到, 从而证明了 (3.28) 式的第二行. 定理得证. \square

注. (3.28) 式第一行的渐进收敛速度和 PALM-E 算法一致^[173,176]. 结合定理 3.19 与定理 3.17 (3), 我们可以作出如下概括: 当 $\{\varepsilon_k\}$ 下降得足够快时, 子问题非精确求解的误差并不会影响 PALM 算法自身的渐进收敛速度.

我们将 PALM-I 算法在不同情形下的渐进收敛速度总结在表 3.1 中¹, 将误差控制序列 $\{\varepsilon_k\}$ 次线性下降时的渐进收敛速度负指数绘制在图 3.1 中. 从表 3.1 与图 3.1 我们可看出, 当 $\theta \in (1/2, 1)$ 时, l 的取值一旦越过临界值 $\theta/(2\theta-1)$, PALM-I 算法的渐进收敛速度负指数将不再改进. 而当 $\theta \in [0, 1/2]$ 时, PALM-I 算法的渐进收敛速度负指数将随 l 线性增长.

3.5 数值实验

本章的数值实验分为两个部分. 在第一部分中, 我们在两个测试问题上比较 PALM-E 算法、PALM-F 算法与 PALM-I 算法的性能. 在第二部分中, 我们将 PALM-I 算法嵌入一个瀑布型多重网格优化(CMGOPT)框架, 通过求解问题 (2.8), 模拟一维、二维强关联电子体系.

所有的数值实验均在包含两颗 Intel Xeon Gold 6242R CPU (3.10 GHz \times 20 \times 2) 的工作站上实现. 工作站的运行内存为 510 GB, 操作系统为 Ubuntu 20.04.5. 我们使用 MATLAB R2018b 实现所有的算法.

¹我们只在表 3.1 中列出了当 $v_k - \bar{F} > 0$ 时的结论. 这是因为该情形在实际计算中更加常见.

表 3.1 PALM-I 算法在不同情形下的收敛速度

Table 3.1 The asymptotic convergence rates of the PALM-I method under different settings

θ	ε_k	假设	渐进收敛速度	参考文献
0	0	-	有限终止	[173,176]
0	$\tilde{\rho}^k$	$\tilde{\rho} \in (0, 1)$	$\mathcal{O}(\rho_1^k)$, 其中 $\rho_1 \in (0, 1)$	本文 (定理 3.17)
	$\frac{1}{(k+1)^l}$	$l \in (1, \infty)$	$\mathcal{O}(k^{-(l-1)})$	本文 (定理 3.19)
$(0, \frac{1}{2}]$	0	-	$\mathcal{O}(\rho_3^k)$, 其中 $\rho_3 \in (0, 1)$	[173,176]
	$\tilde{\rho}^k$	$\tilde{\rho} \in (0, 1)$	$\mathcal{O}(\rho_2^k)$, 其中 $\rho_2 \in (0, 1)$	本文 (定理 3.17)
$(\frac{1}{2}, 1)$	$\frac{1}{(k+1)^l}$	$l \in (1, \infty)$	$\mathcal{O}(k^{-(l-1)})$	本文 (定理 3.19)
	0	-	$\mathcal{O}\left(k^{-\frac{1-\theta}{2\theta-1}}\right)$	[173,176]
	$\tilde{\rho}^k$	$\tilde{\rho} \in (0, 1)$	$\mathcal{O}\left(k^{-\frac{1-\theta}{2\theta-1}}\right)$	本文 (定理 3.17)
$(\frac{1}{2}, 1)$	$\frac{1}{(k+1)^l}$	$l \in (1, \infty)$	$\mathcal{O}\left(k^{-\frac{1-\theta}{2\theta-1}}\right)$, 若 $l \geq \frac{\theta}{2\theta-1}$ $\mathcal{O}(k^{-(l-1)})$, 若 $l < \frac{\theta}{2\theta-1}$	本文 (定理 3.19)

3.5.1 算法比较

我们在线性与非线性约束优化问题上比较 PALM-E 算法、PALM-F 算法与 PALM-I 算法的性能. 其中, 线性约束优化问题为强关联电子体系计算中的问题 (2.8), 其可行域形如例 3.1 中的 Birkhoff 多胞体. 非线性约束优化问题的目标函数为非凸二次函数, 其可行域为例 3.2 中的椭球.

3.5.1.1 线性约束优化问题上的测试

我们考虑强关联电子体系计算中的问题 (2.8). 容易验证, 问题 (2.8) 具有本章考虑的问题 (3.1) 的形式. 我们考虑区间 $\Omega = [-1, 1]$ 上的一维三电子体系, 其单电子密度是 Gauss 型函数²:

$$\rho(x) \propto \exp(-x^2/\sqrt{\pi}), \quad \forall x \in [-1, 1].$$

我们采用等质量网格剖分, 即 \mathbf{p} 中的所有元素都相等, 取 $K = 36$.

由于暂无可行方法求解子问题, 我们在本节测试中仅比较 PALM-E 算法与 PALM-I 算法. 我们固定两个算法的邻近参数为 $\sigma_{i,k} \equiv \sigma = 10^{-2}$ ($i = 2, \dots, N$, $k \geq 0$), 使用半光滑 Newton 法^[180] 非精确求解其中的子问题. 子问题非精确求解准则为 (3.8) 式. 根据例 3.1, 此时算法迭代点的不可行性无法避免. 对 PALM-E 算法, 我们设置充分小的 $\varepsilon_k \equiv 10^{-7}$ ($k \geq 0$) 从而让所有子问题都被求解得足够精确.

² 缩放系数需满足 $\int_{\Omega} \rho = N$. 后文类似不再赘述.

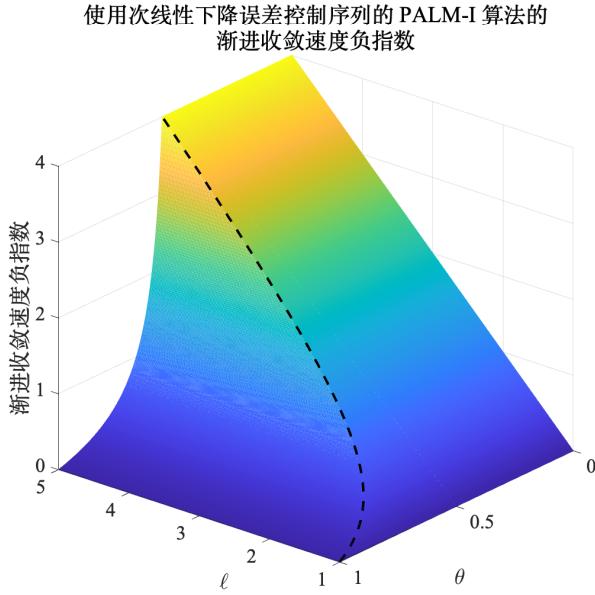


图 3.1 当误差控制序列取做 $\varepsilon_k = \bar{\varepsilon}/(k+1)^l$ ($k \geq 0, l > 1$) 时, PALM-I 算法的渐进收敛速度负指数. 其中, x 轴为 F 在收敛点处的 Łojasiewicz 指数, y 轴为 l 的取值, z 轴为 PALM-I 算法的渐进收敛速度负指数, 虚线表示 θ 与 l 的临界关系 $\theta = l/(2l - 1)$

Figure 3.1 The negative asymptotic convergence rate exponents of the PALM-I method when the error control sequence $\varepsilon_k = \bar{\varepsilon}/(k+1)^l$ ($k \geq 0, l > 1$). The x -axis refers to the Łojasiewicz exponent of F at the limit point, y -axis refers to the value of l , z -axis refers to the corresponding negative asymptotic convergence rate exponent of the PALM-I method, and the dashed line stands for the critical relation between θ and l : $\theta = l/(2l - 1)$

对 PALM-I 算法, 我们取

$$\varepsilon_k = \max \left\{ \frac{10^{-1}}{(k+1)^{0.75}}, 10^{-7} \right\}, \quad \forall k \geq 0.$$

根据条件 3.3 下方之注, 我们在问题 (2.8) 相对 KKT 违反度 (计算公式参见 (1.5) 式) 小于 10^{-6} 时终止两个算法.

我们首先比较从随机生成初始点出发的 PALM-E 算法与 PALM-I 算法在求解问题 (2.8) 时的数值表现. 我们使用 MATLAB 自带函数 “rand” 随机生成 100 个初始点. 在图 3.2 的左半部分, 我们展示 PALM-E 算法与 PALM-I 算法 100 次运行的平均相对 KKT 违反度随迭代过程的变化. 尽管两个算法需要差不多的迭代次数, 但 PALM-I 算法所需的平均运行时间仅约 0.46 秒, 与 PALM-E 算法所需约 15.97 秒形成强烈对比. PALM-I 算法的巨大效率优势是通过非精确求解子问题获得的. 因为问题 (2.8) 非凸, 所以我们还可以比较两个算法收敛到的目标函数值. 在图 3.2 的右半部分, 我们画出了 PALM-E 算法与 PALM-I 算法的 100 次平均相对目标函数值误差随迭代过程的变化. 这里的相对目标函数值误差通过下式计算:

$$\frac{|f(Y_2^{(k)}, \dots, Y_N^{(k)}) - f^*|}{\max \{f^*, 1\}},$$

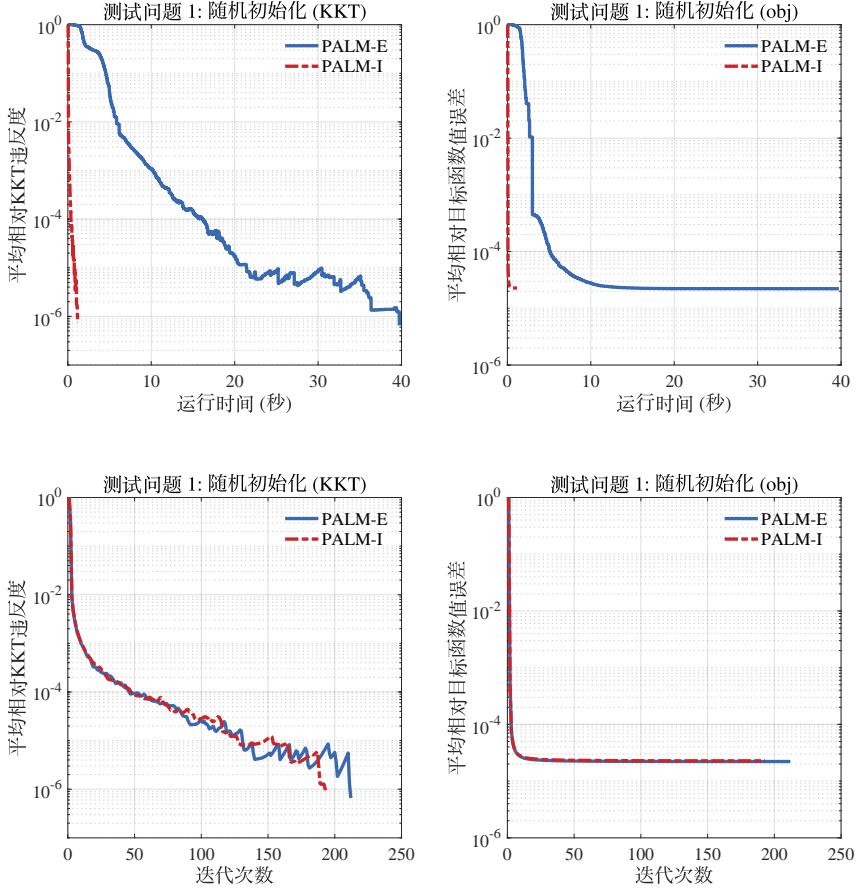


图 3.2 从 100 个随机初始点出发, PALM-E 算法与 PALM-I 算法在求解问题 (2.8) 时的平均相对 KKT 违反度与平均相对目标函数值误差随迭代过程的变化. 其中蓝色实线与红色点划线分别表示 PALM-E 算法与 PALM-I 算法的结果

Figure 3.2 With 100 random initialization, the averaged history of relative KKT violation and relative objective difference from the optimal value of the PALM-E and PALM-I methods when solving problem (2.8). The blue solid and red dashdotted lines show respectively the results of the PALM-E and PALM-I methods

其中 $f^* \in \mathbb{R}$ 为问题 (2.8) 的最优值. 对于一维体系 f^* 是可以计算的^[100]. 从平均意义上, 两个算法收敛到的目标函数值几乎一致, 得到的解质量相似.

因为在当前实验设置下问题 (2.8) 的最优值可以计算, 所以我们还可以考察 PALM-E 算法与 PALM-I 算法从最优解附近出发时的数值表现. 我们再次使用“rand”随机生成最优解附近的 100 个初始点, 其中矩阵每个元素的绝对偏差不超过 10^{-3} . 类似地, 我们在图 3.3 的左半部分与右半部分分别展示两个算法的平均相对 KKT 违反度与平均相对目标函数值误差随迭代过程的变化. PALM-E 算法所需的平均运行时间为 0.85 秒, 而 PALM-I 算法平均仅需 0.03 秒. 这表明从最优解附近出发, PALM-I 算法同样具有显著的效率优势. 此外, PALM-I 算法的非精确性并没有明显降低其输出解的质量. 在 100 次模拟中, PALM-I 算法的最大绝

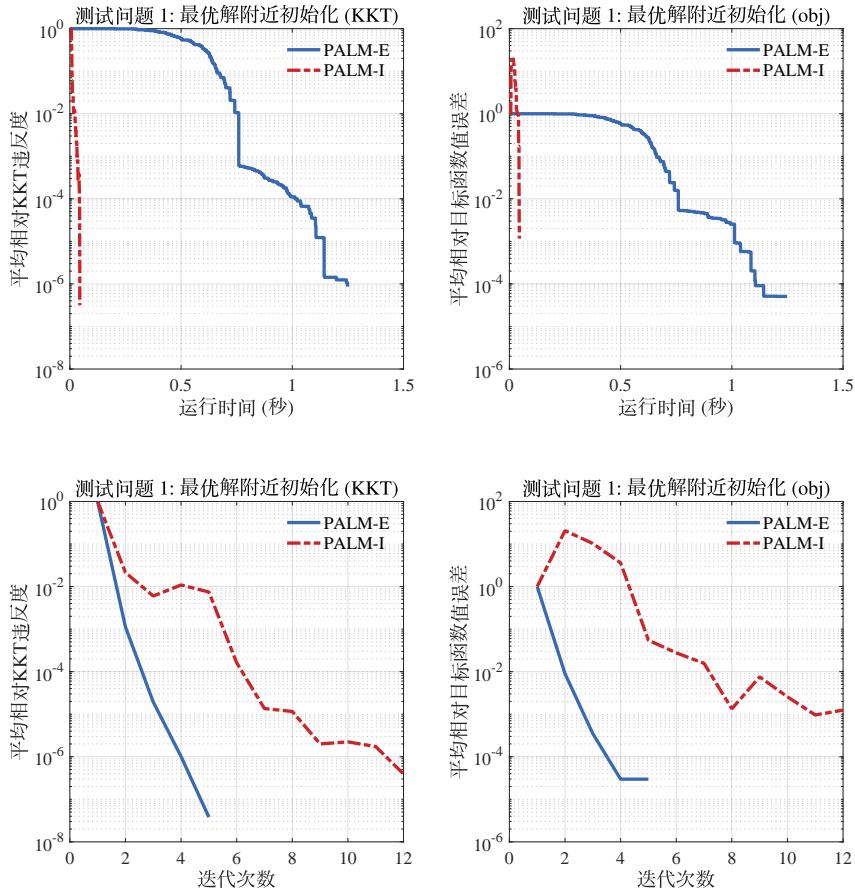


图 3.3 从最优解附近的 100 个随机初始点出发, PALM-E 算法与 PALM-I 算法在求解问题 (2.8) 时的平均相对 KKT 违反度与平均相对目标函数值误差随迭代过程的变化. 其中蓝色实线与红色点划线分别表示 PALM-E 算法与 PALM-I 算法的结果

Figure 3.3 With 100 random initialization near an optimal solution, the averaged history of relative KKT violation and relative objective difference from the optimal value for the PALM-E and PALM-I methods when solving the problem (2.8). The blue solid and red dashdotted lines show respectively the results of the PALM-E and PALM-I methods

对目标函数值误差仅为 2.39×10^{-4} ; 在超过 75% 的模拟中, PALM-I 算法的绝对目标函数值误差小于 10^{-5} .

总体而言, 求解问题 (2.8) 的数值结果充分展现了 PALM-I 算法的有效性与相对 PALM-E 算法的效率优势.

3.5.1.2 非线性约束优化问题上的测试

我们考虑的测试问题为

$$\begin{aligned} \min_{\mathbf{x}} \quad & \frac{1}{2} \langle \mathbf{x}, A\mathbf{x} \rangle + \langle \mathbf{b}, \mathbf{x} \rangle, \\ \text{s. t.} \quad & \frac{1}{2} \langle \mathbf{x}_i, B_i \mathbf{x}_i \rangle + \langle \mathbf{c}_i, \mathbf{x}_i \rangle \leq 1, \quad i = 1, \dots, s. \end{aligned} \tag{3.30}$$

问题 (3.30) 中的椭球约束可用于建模拓扑优化中的关联不确定性^[172], 也出现在弹性力学问题的对偶变分形式中^[169]. 因此, 研究问题 (3.30) 的求解对于实际应用是十分有意义的. 这里, $m_i = m$, $\mathbf{x}_i \in \mathbb{R}^m$ ($i = 1, \dots, s$), $A \in \mathbb{R}^{ms \times ms}$ 是对称矩阵, $\{\mathbf{B}_i\}_{i=1}^s \subseteq \mathbb{R}^{m \times m}$ 是对称正定矩阵, $\mathbf{b} \in \mathbb{R}^{ms}$, $\{\mathbf{c}_i\}_{i=1}^s \subseteq \mathbb{R}^m$.

在本节数值实验中, 我们用 MATLAB 的自带函数“randn”随机生成 A 和 \mathbf{b} . 我们仿照文献^[182,183] 按如下方式定义 $\{\mathbf{B}_i = (b_{i,jk})_{jk}\}_{i=1}^s$:

$$b_{i,jk} := \begin{cases} 10^{\frac{j-1}{m-1} ncond_i}, & \text{若 } j = k; \\ 0, & \text{否则,} \end{cases} \quad j, k = 1, \dots, m, i = 1, \dots, s.$$

$\{ncond_i\}_{i=1}^s \subseteq \mathbb{R}_{++}$ 是预设常数. 易见 $\{ncond_i\}_{i=1}^s$ 控制了矩阵 $\{\mathbf{B}_i\}_{i=1}^s$ 的条件数, 且 \mathbf{B}_i 的特征值就分布在区间 $[1, 10^{ncond_i}]$ 中 ($i = 1, \dots, s$). 我们取 $s = 5$, $m = 500$,

$$\{ncond_i\}_{i=1}^s = \{3.00, 3.25, 3.50, 3.75, 4.00\}.$$

向量 $\{\mathbf{c}_i\}_{i=1}^s$ 均置为 0.

我们在问题 (3.30) 上比较 PALM-E 算法、PALM-F 算法与 PALM-I 算法. 我们固定三个算法的邻近参数为 $\sigma_{i,k} \equiv \hat{\sigma} = 1$ ($i = 1, \dots, s$, $k \geq 0$). PALM-E 算法与 PALM-I 算法使用文献^[182] 中的交替方向乘子法 (ADMM) 求解子问题, 子问题非精确求解准则为 (3.8) 式. PALM-F 算法使用文献^[183] 中的混合投影法 (HP) 求解子问题. 对 $i = 1, \dots, s$ 与 $k \geq 0$, PALM-F 算法中的子问题非精确求解准则设为

$$\left\| \mathbf{x}_i^{(k+1)} - \tilde{\mathbf{x}}_i^{(k)} + \lambda_{i,\text{HP}}^{(k+1)} (\mathbf{B}_i \mathbf{x}_i^{(k+1)} + \mathbf{c}_i) \right\| \leq \frac{0.99\hat{\sigma}}{2} \left\| \mathbf{x}_i^{(k+1)} - \mathbf{x}_i^{(k)} \right\|_\infty,$$

其中 $\lambda_{i,\text{HP}}^{(k+1)} \geq 0$ 是对对应于椭球约束的 Lagrange 乘子的估计. 这一非精确求解准则可保证迭代点满足文献^[175] 中的假设条件, 进而保证 PALM-F 算法的收敛性. 由于使用 ADMM 算法求解子问题, PALM-E 算法与 PALM-I 算法的迭代点未必可行. 对于当前问题, 残差函数 r_i (见 (3.9) 式) 为

$$\begin{aligned} & \max \left\{ \left\langle \mathbf{x}_i^{(k+1)}, \mathbf{x}_i^{(k+1)} - \tilde{\mathbf{x}}_i^{(k)} + \lambda_{i,\text{ADMM}}^{(k+1)} (\mathbf{B}_i \mathbf{x}_i^{(k+1)} + \mathbf{c}_i) \right\rangle, 0 \right\} \\ & + \left\| \mathbf{x}_i^{(k+1)} - \tilde{\mathbf{x}}_i^{(k)} + \lambda_{i,\text{ADMM}}^{(k+1)} (\mathbf{B}_i \mathbf{x}_i^{(k+1)} + \mathbf{c}_i) \right\|_\infty \\ & + \lambda_{i,\text{ADMM}}^{(k+1)} \max \left\{ - \left[\frac{1}{2} \left\langle \mathbf{x}_i^{(k+1)}, \mathbf{B}_i \mathbf{x}_i^{(k+1)} \right\rangle + \left\langle \mathbf{c}_i, \mathbf{x}_i^{(k+1)} \right\rangle - 1 \right], 0 \right\} \\ & + \max \left\{ \frac{1}{2} \left\langle \mathbf{x}_i^{(k+1)}, \mathbf{B}_i \mathbf{x}_i^{(k+1)} \right\rangle + \left\langle \mathbf{c}_i, \mathbf{x}_i^{(k+1)} \right\rangle - 1, 0 \right\}, \end{aligned}$$

其中 $\lambda_{i,\text{ADMM}}^{(k+1)} \geq 0$ 是由 ADMM 算法给出的子问题近似对偶解. 对 PALM-E 算法, 我们设置充分小的 $\varepsilon_k \equiv 10^{-6}$ ($k \geq 0$) 从而让所有子问题都被求解得足够精确. 对 PALM-I 算法, 我们取

$$\varepsilon_k = \max \left\{ \frac{10^{-1}}{(k+1)^{0.75}}, 10^{-6} \right\}, \quad \forall k \geq 0.$$

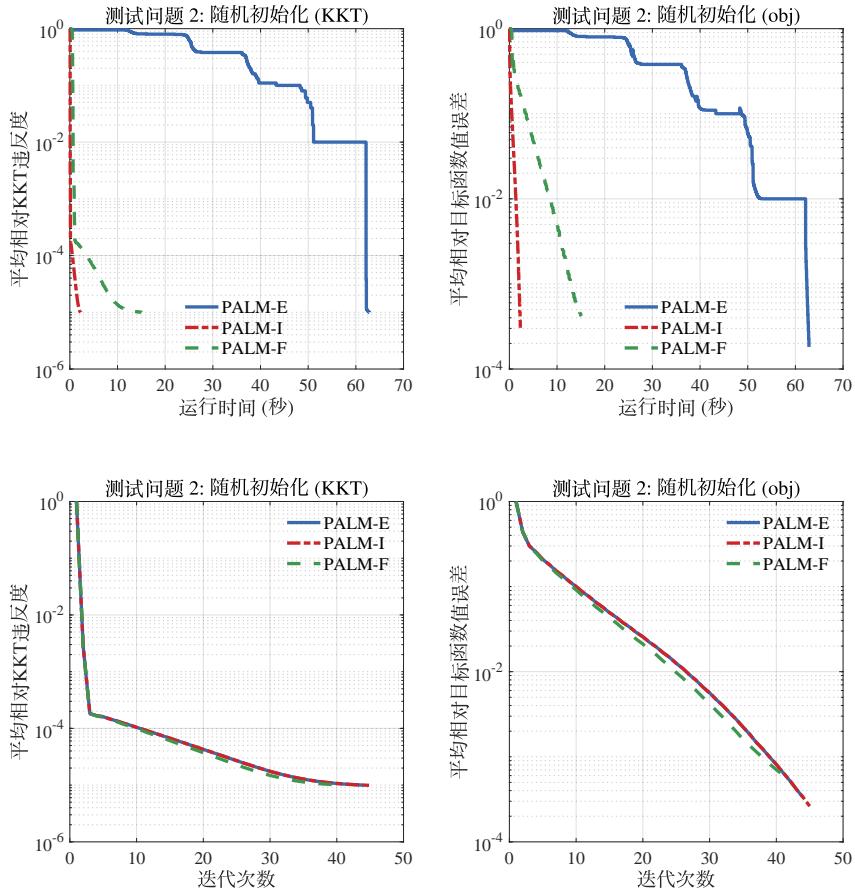


图 3.4 从 100 个随机初始点出发, PALM-E 算法、PALM-F 算法与 PALM-I 算法在求解问题 (3.30) 时的平均相对 KKT 违反度与平均相对目标函数值误差随迭代过程的变化, 其中蓝色实线表示 PALM-E 算法的结果, 红色点划线表示 PALM-I 算法的结果, 绿色划线表示 PALM-F 算法的结果

Figure 3.4 With 100 random initialization, the averaged history of relative KKT violation and relative objective difference from the minimum achieved value of the PALM-E, PALM-F, and PALM-I methods when solving problem (3.30). The blue solid line refers to the results of the PALM-E method, the red dashdotted line refers to the results of the PALM-I method, and the green dashed line refers to the results of the PALM-F method

与前一个测试问题类似, 我们在问题 (3.30) 相对 KKT 违反度 (计算公式参见 (1.5) 式) 小于 10^{-5} 时终止三个算法.

我们比较从随机初始点出发三个算法在求解问题 (3.30) 时的数值表现. 我们使用 MATLAB 自带函数 “randn” 随机生成 100 个初始点. 在图 3.4 中, 我们展示三个算法 100 次运行的平均相对 KKT 违反度与平均相对目标函数值误差随迭代过程的变化. 这里, 由于问题 (3.30) 的最优值无法获取, 我们按如下方式计算第 t

表 3.2 待模拟的一维和二维强关联电子体系. 第二列为未归一化的单电子密度 ρ , 第三列为截断区域 Ω , 第四列为体系所含电子数 N

Table 3.2 One/two-dimensional systems used for simulations. The second column lists the unnormalized single-particle densities ρ , the third gives the truncated domains Ω , and the last indicates the numbers of electrons N in systems

体系编号	单电子密度 ρ	截断区域 Ω	电子数 N
一维体系			
1	$\cos(\pi r) + 1$	$[-1, 1]$	3
2	$2\rho_6(r; -0.5) + 1.5\rho_4(r; 0.5)$	$[-1.5, 1.5]$	3
3	$\exp(- r)$	$[-5, 5]$	3
4	$\rho_{1/\sqrt{\pi}}(r; 0)$	$[-3, 3]$	7
5	$\rho_1(r; -2) + 5\rho_2(r; 0) + \rho_1(r; 2)$	$[-4, 4]$	7
6	$\rho_4(r; -2) + \rho_4(r; -1.5) + \rho_4(r; -1) + \rho_4(r; -0.5) + \rho_4(r; 2/3) + \rho_4(r; 4/3) + \rho_4(r; 2)$	$[-3, 3]$	7
二维体系			
7	$\rho_{2,5}(\mathbf{r}; [-1.5, 0]^\top) + 0.5\rho_{2,5}(\mathbf{r}; [1.5, 0]^\top)$	$[-3, 3] \times [-2, 2]$	3
8	$\rho_{2,5}(\mathbf{r}; [-1.032, -0.84]^\top) + \rho_{2,5}(\mathbf{r}; [0, 0.96]^\top) + \rho_{2,5}(\mathbf{r}; [1.032, -0.84]^\top)$	$[-2.5, 2.5]^2$	3

次模拟中的相对目标函数值误差:

$$\frac{|f(\mathbf{x}^{(k)}) - f_t^*|}{\max \{f_t^*, 1\}},$$

其中 $f_t^* \in \mathbb{R}$ 是在第 t 次模拟中三个算法收敛到的最小目标函数值 ($t = 1, \dots, 100$). PALM-E 算法、PALM-F 算法与 PALM-I 算法分别所需的平均运行时间约为 30.15 秒、9.26 秒与 1.74 秒. 由此可见, PALM-I 算法充分地发挥了不可行的 ADMM 算法的效率优势. 此外, 我们还观察到, PALM-I 算法收敛到的目标函数值较 PALM-F 算法更加接近 PALM-E 算法.

3.5.2 强关联电子体系计算

我们调用 PALM-I 算法求解问题 (2.8) 来模拟一维、二维强关联电子体系. 为在离散规模较大的情形下加速 PALM-I 算法的收敛, 我们引入一个瀑布型多重网格优化 (CMGOPT) 框架, 并将 PALM-I 算法作为其中的局部优化算法.

3.5.2.1 待模拟强关联电子体系

我们考虑八个一维和二维强关联电子体系, 在表 3.2 中列出它们的单电子密度、截断区域和电子数. 其中函数 $\rho_\alpha(\cdot; \mathbf{c})$ ($\alpha > 0, \mathbf{c} \in \mathbb{R}^d$) 定义为

$$\rho_\alpha(\mathbf{r}; \mathbf{c}) := \exp(-\alpha \|\mathbf{r} - \mathbf{c}\|^2), \quad \forall \mathbf{r} \in \mathbb{R}^d. \quad (3.31)$$

我们还在图 3.5 中示出表 3.2 里列出的单电子密度.

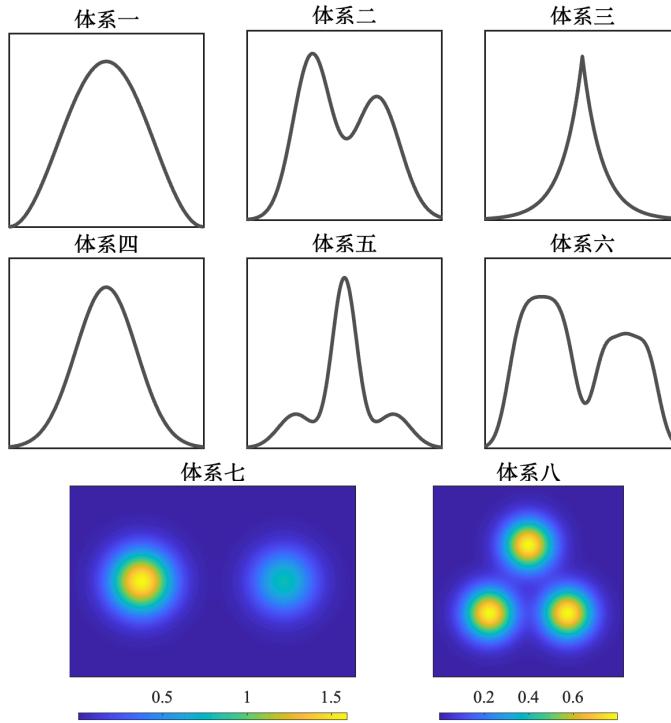


图 3.5 表 3.2 所列单电子密度的可视化

Figure 3.5 The illustrations of the single-particle densities listed in Table 3.2

3.5.2.2 瀑布型多重网格优化框架

由于问题(2.8)以矩阵为变量,其维数为离散的规模,动辄成千上万.因此,从零开始调用 PALM-I 算法直接求解实际离散规模的问题(2.8)并不是一个明智的选择.

在数值分析中,多重网格法是一种求解大规模(尤其来源于偏微分方程离散的)线性代数方程组的高效算法^[197–200].它结合粗网格与细网格的信息,通过逐层的迭代逼近,从而达到更快速、更有效地收敛到解的效果.多重网格法可依据网格之间信息传递的方式进行分类,例如 V-循环(V-cycle) 多重网格、W-循环(W-cycle) 多重网格、瀑布型多重网格、全多重网格(full multigrid) 等.我们推荐感兴趣的读者参阅专著^[197–199]、综述^[200] 及其中的参考文献.

近年来,多重网格的思想已被应用于无限维优化问题的求解,例如最优控制问题、带偏微分方程约束优化问题^[201] 等.此时,多重网格法可用于直接求解优化问题的稳定性系统^[202–204] 或与优化算法结合.对于后者,多重网格法可用于加速优化算法中子问题的求解^[205–207],也可通过提供外层框架提升优化算法的效率^[208–214].其中,与其他多重网格优化框架相比,瀑布型多重网格优化(CMGOPT) 框架^[215–217] 不需要校正(correction) 过程,只需要进行从粗到细网格的迭代,计算量小,流程统一,已被广泛应用于各个领域,如量子物理^[218–221]、分子力学^[222]、拓扑优化^[223]、最优运输^[224] 等.

我们所设计的 CMGOPT 框架如框架 3.3 所示.下面,我们对其中的细节做一些说明.

框架 3.3: 求解问题 (2.8) 的瀑布型多重网格优化框架 CMGOPT.

输入: 初始网格 $\mathcal{T}^{(0)}$ (离散规模为 $K_0 \in \mathbb{N}$), 网格加密次数 $\ell_{\max} \in \mathbb{N}$.

- 1 置 $\ell := 0$.
 - 2 选取罚参数 $\beta_0 > 0$.
 - 3 调用高精度优化算法求解网格 $\mathcal{T}^{(0)}$ 上的问题 (2.8), 得解 $Y_i^{(0,\star)} \in \mathbb{R}^{K_0 \times K_0}$ ($i = 2, \dots, N$).
 - 4 **while** $\ell < \ell_{\max}$ **do**
 - 5 加密网格 $\mathcal{T}^{(\ell)}$ 得到 $\mathcal{T}^{(\ell+1)}$ (离散规模为 $K_\ell \in \mathbb{N} : K_{\ell+1} \geq K_\ell$), 使得 $\mathcal{T}^{(\ell)}$ 嵌入 $\mathcal{T}^{(\ell+1)}$.
 - 6 构造插值算子 $\mathcal{I}_\ell^{\ell+1} : \mathbb{R}^{K_\ell \times K_\ell} \rightarrow \mathbb{R}^{K_{\ell+1} \times K_{\ell+1}}$ 并将其作用于 $Y_i^{(\ell,\star)}$ 上得到 $Y_i^{(\ell+1,0)} := \mathcal{I}_\ell^{\ell+1}(Y_i^{(\ell,\star)}) \in \mathbb{R}^{K_{\ell+1} \times K_{\ell+1}}$ ($i = 2, \dots, N$).
 - 7 选取罚参数 $\beta_{\ell+1} > 0$.
 - 8 调用局部优化算法从初始点 $Y_i^{(\ell+1,0)}$ ($i = 2, \dots, N$) 出发求解网格 $\mathcal{T}^{(\ell+1)}$ 上的问题 (2.8), 得解 $Y_i^{(\ell+1,\star)} \in \mathbb{R}^{K_{\ell+1} \times K_{\ell+1}}$ ($i = 2, \dots, N$).
 - 9 置 $\ell := \ell + 1$.
 - 10 **end**
- 输出:** $Y_i^{(\ell_{\max},\star)} \in \mathbb{R}^{K_{\ell_{\max}} \times K_{\ell_{\max}}}$ ($i = 2, \dots, N$).

在求解初始粗网格上的问题 (2.8) 时, CMGOPT 框架需调用高精度优化算法, 甚至全局优化算法. 这是精度与效率的权衡: 一方面, 我们希望求得大规模问题 (2.8) 的高质量解, 因此只要计算资源允许, 应尽可能地调用高精度优化算法; 另一方面, 使用高精度优化算法求解较细网格上的问题效率低下, 而基于插值算子 (interpolation operator) 构造的初始点可让局部优化算法热启动 (warm start), 从而快速收敛.

CMGOPT 框架中插值算子的构造基于 (1.24) 式, 即 $\rho\gamma_i$ 可看作第一个电子与第 i 个电子位置的联合概率密度 ($i = 2, \dots, N$). 首先以一维情形为例, 假设在网格 $\mathcal{T}^{(\ell)} := \{e_k^{(\ell)}\}_{k=1}^{K_\ell}$ 上已求得问题 (2.8) 的一个解 $\{Y_i^{(\ell,\star)}\}_{i=2}^N$. 对 $i = 2, \dots, N$, $y_{i,jj'}^{(\ell,\star)} > 0$ 就表明

$$\text{Prob} \left\{ \text{第一与第 } i \text{ 个电子分别位于 } e_j^{(\ell)} \text{ 与 } e_{j'}^{(\ell)} \right\} > 0.$$

当我们把网格一致加密一倍得到 $\mathcal{T}^{(\ell+1)} := \{e_k^{(\ell+1)}\}_{k=1}^{K_{\ell+1}}$ 后, $\mathcal{T}^{(\ell)}$ 中的 $e_j^{(\ell)}$ 与 $e_{j'}^{(\ell)}$ 分别对应 $\mathcal{T}^{(\ell+1)}$ 中的 $e_{2j-1}^{(\ell+1)}$ 、 $e_{2j}^{(\ell+1)}$ 与 $e_{2j'-1}^{(\ell+1)}$ 、 $e_{2j'}^{(\ell+1)}$. 因此, 我们可以推测

$$\text{Prob} \left\{ \text{第一与第 } i \text{ 个电子分别位于 } e_{j_u}^{(\ell+1)} \text{ 与 } e_{j_v}^{(\ell+1)} \right\} > 0, \quad u, v \in \{1, 2\},$$

其中 $j_1 := 2j - 1$, $j_2 := 2j$, $j'_1 := 2j' - 1$, $j'_2 := 2j'$. 对应地, 应当有 $y_{i,j_u j'_v}^{(\ell+1,0)} > 0$

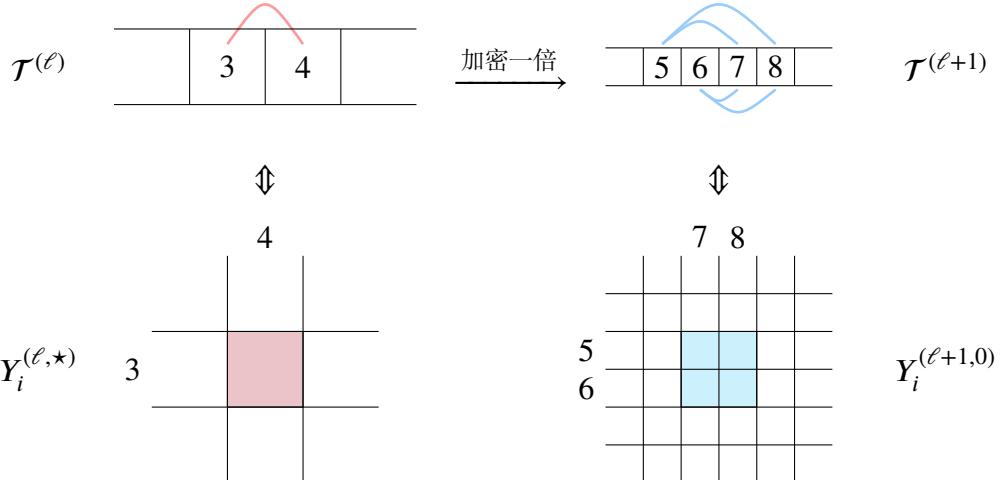


图 3.6 一维一致加密情形插值算子示意图. 红色方块代表在网格 $\mathcal{T}^{(\ell)}$ 上, 解的元素 $y_{i,34}^{(\ell,*)}$ 为正. 蓝色方块代表在插值算子 $\mathcal{J}_{\ell}^{\ell+1}$ 的作用下, 在网格 $\mathcal{T}^{(\ell+1)}$ 上, 初始点的元素 $y_{i,57}^{(\ell+1,0)}$ 、 $y_{i,67}^{(\ell+1,0)}$ 、 $y_{i,58}^{(\ell+1,0)}$ 、 $y_{i,68}^{(\ell+1,0)}$ 为正

Figure 3.6 An illustration of the interpolation operator in one-dimensional context with uniform mesh refinements. The red block represents a positive entry $y_{i,34}^{(\ell,*)}$ in the solution over the mesh $\mathcal{T}^{(\ell)}$. The blue blocks represent the positive entries $y_{i,57}^{(\ell+1,0)}$, $y_{i,67}^{(\ell+1,0)}$, $y_{i,58}^{(\ell+1,0)}$, $y_{i,68}^{(\ell+1,0)}$ in the initial point over the mesh $\mathcal{T}^{(\ell+1)}$ under the action of the interpolation operator $\mathcal{J}_{\ell}^{\ell+1}$

($u, v \in \{1, 2\}$). 基于 (1.24) 式, 插值算子 $\mathcal{J}_{\ell}^{\ell+1}$ 可定义为

$$\mathcal{J}_{\ell}^{\ell+1}(Y)_{i'j'} := \frac{\rho_i^{(\ell)}}{\rho_{i'}^{(\ell+1)}} y_{ij}, \quad i : e_{i'}^{(\ell+1)} \subseteq e_i^{(\ell)}, j : e_{j'}^{(\ell+1)} \subseteq e_j^{(\ell)}.$$

其中 $\rho^{(\ell)} := [\rho_1^{(\ell)}, \dots, \rho_{K_{\ell}}^{(\ell)}] \in \mathbb{R}^{K_{\ell}}$ 是 $\mathcal{T}^{(\ell)}$ 上单电子密度 ρ 的离散. 注意, 由于我们使用嵌入型 (embedding) 网格加密 (见框架 3.3 的步 4), 因此指标 i 和 j 总是唯一确定的. 我们将上面定义的插值算子的作用展现在图 3.6 中.

上述插值算子的构造可推广至任意 $d \in \mathbb{N}$ 维情形及任何嵌入型网格加密. 我们只需明确网格加密前后离散单元之间的对应关系. 我们将二维情形使用一致网格加密时插值算子的作用展现在图 3.7 中.

一般情形下, 插值算子的作用可通过算法 3.4 中的步骤计算.

注. 使用插值算子构造局部优化算法的初始点与求解最优运输问题的屏蔽邻域 (shielding neighborhood) 技术 [116, 225, 226] 完全不同. 屏蔽邻域技术借助线性规划的强对偶性自适应地调整变量的支撑指标集, 并将优化的变量限制在所选取的支撑上. 该技术可在不增加过多计算与空间复杂度的前提下, 保证解的最优性. 然而, 与最优运输问题相比, 问题 (2.8) 是一个非凸二次规划, 不具有强对偶性, 因此寻找与调整支撑指标集是不切实际的. 对应地, 我们在优化时保留所有的变量, 通过使用插值算子为局部优化算法提供高质量的初始点, 提升整体求解的效率.

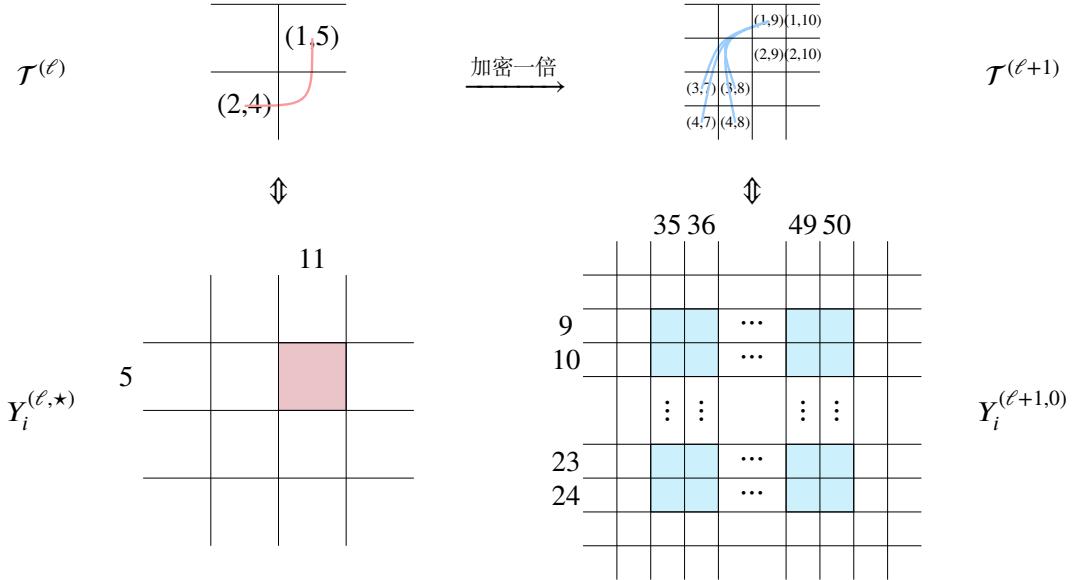


图 3.7 二维一致加密情形插值算子示意图, 其中网格 $\mathcal{T}^{(\ell)}$ 与 $\mathcal{T}^{(\ell+1)}$ 的离散单元数分别为 7×7 与 14×14 . 离散单元在网格中的二维坐标与 Y_i 的行序号或列序号一一对应. 例如, $\mathcal{T}^{(\ell)}$ 中坐标为 $(2, 4)$ 的单元对应 Y_i 的第 11 行或列. 红色方框代表在网格 $\mathcal{T}^{(\ell)}$ 上, 解的元素 $y_{i,5,11}^{(\ell,*)}$ 为正. 蓝色方块代表在插值算子 $\mathcal{J}_\ell^{\ell+1}$ 的作用下, 在网格 $\mathcal{T}^{(\ell+1)}$ 上, 初始点的元素 $y_{i,jj'}$ 为正, 其中 $j \in \{9, 10, 23, 24\}$, $j' \in \{35, 36, 49, 50\}$

Figure 3.7 An illustration of the interpolation operator in two-dimensional context with uniform mesh refinements, where the numbers of finite elements are 7×7 and 14×14 in the meshes $\mathcal{T}^{(\ell)}$ and $\mathcal{T}^{(\ell+1)}$, respectively. There is a one-by-one mapping between the two-dimensional coordinates of the finite elements in the mesh and the row or column indices in the matrix Y_i . For example, the element $(2, 4)$ in $\mathcal{T}^{(\ell)}$ corresponds to the 11th row or column in Y_i . The red block represents a positive entry $y_{i,5,11}^{(\ell,*)}$ in the solution over the mesh $\mathcal{T}^{(\ell)}$. The blue blocks represent the positive entries $y_{i,jj'}^{(\ell+1,0)}$ in the initial point over the mesh $\mathcal{T}^{(\ell+1)}$ under the action of the interpolation operator $\mathcal{J}_\ell^{\ell+1}$, where $j \in \{9, 10, 23, 24\}$, $j' \in \{35, 36, 49, 50\}$

3.5.2.3 实验设置

我们使用 CMGOPT 框架求解问题 (2.8). 对于初始粗网格上的离散问题, 我们直接使用现成全局优化算法或软件求解. 经测试, 我们发现对于规模满足 $(N-1)K^2 \leq 500$ 的问题, 随机多初始 (random multistart) 方法^[227] 只需要随机生成数十个初始点运行局部优化算法即可找到问题 (2.8) 的全局最优解. 因此, 对于一维体系, 我们采用随机多初始方法作为求解初始粗网格上离散问题的全局优化算法. 对于二维体系, 我们调用 AMPL^[228] 中的 BARON (21.1.13 版)^[229] 求解初始粗网格上的全局优化问题, 其在可接受的时间内可以较好地求解 $K \approx 200$ 的问题 (2.8). CMGOPT 框架中用于求解加密网格上问题的局部优化算法则指定为 PALM-I 算法, 其中初始点由插值算子构造 (见算法 3.4), 子问题由半光滑 Newton 法^[180] 非精确求解. 我们将在后文指定框架中的初始网格离散、网格加密方式与加密次数 ℓ_{\max} .

算法 3.4: 插值算子 $\mathcal{I}_\ell^{\ell+1}$ 作用的计算.

输入: 加密前后的网格 $\mathcal{T}^{(\ell)}$ (离散规模为 $K_\ell \in \mathbb{N}$) 与 $\mathcal{T}^{(\ell+1)}$ (离散规模为 $K_{\ell+1} \in \mathbb{N}$), 网格 $\mathcal{T}^{(\ell)}$ 上的变量 $Y_i^{(\ell)} \in \mathbb{R}^{K_\ell \times K_\ell}$, 加密前后网格上的离散单电子密度 $\rho^{(\ell)} \in \mathbb{R}^{K_\ell}$ 与 $\rho^{(\ell+1)} \in \mathbb{R}^{K_{\ell+1}}$.

```

1 for  $j = 1, \dots, K_\ell$  do
2   for  $j' = 1, \dots, K_\ell$  do
3     寻找  $\mathcal{I}_j^{(\ell+1)}$ , 使得  $e_j^{(\ell)} = \bigcup_{u \in \mathcal{I}_j^{(\ell+1)}} e_{j_u}^{(\ell+1)}$ .
4     寻找  $\mathcal{I}_{j'}^{(\ell+1)}$ , 使得  $e_{j'}^{(\ell)} = \bigcup_{v \in \mathcal{I}_{j'}^{(\ell+1)}} e_{j'_v}^{(\ell+1)}$ .
5     置  $y_{i,j_u j'_v}^{(\ell+1)} := \rho_j^{(\ell)} y_{i,jj'}^{(\ell)} / \rho_{j_u}^{(\ell+1)}$  ( $u \in \mathcal{I}_j^{(\ell+1)}$ ,  $v \in \mathcal{I}_{j'}^{(\ell+1)}$ ).
6   end
7 end
输出:  $Y_i^{(\ell+1)} \in \mathbb{R}^{K_{\ell+1} \times K_{\ell+1}}$ .
```

表 3.3 不同问题规模 K 对应的 β Table 3.3 The values of β corresponding to different problem sizes K

K	(0, 10)	[10, 36)	[36, 80)	[80, 160)	[160, 320)
β	2^2	2^1	2^0	2^{-2}	2^{-3}
K	[320, 640)	[640, 1280)	[1280, 2560)	[2560, 5120)	[5120, ∞)
β	2^{-4}	2^{-5}	2^{-6}	2^{-7}	2^{-8}

根据经验, 使 ℓ_1 罚函数精确性成立的临界罚参数按 $\mathcal{O}(1/K)$ 方式随 K 递减. 因此, 对于不同的 K , 我们依照表 3.3 选取罚参数 β . PALM-I 算法的邻近参数固定为 $\sigma_{i,k} \equiv \tilde{\sigma} = 10^{-3}$ (对任意 $k \geq 0$). 半光滑 Newton 法在首次调用时从原点出发, 而在之后的迭代中从之前退出时的对偶变量热启动.

当如下三条中的一个满足时, 我们终止 PALM-I 算法: (1) 相邻迭代点间距离

$$\sqrt{\tilde{\sigma} \sum_{i=2}^N \|Y_i^{(k+1)} - Y_i^{(k)}\|^2} \leq 10^{-4};$$

(2) 相邻两次迭代的绝对目标函数值差小于 10^{-8} ; (3) 迭代次数超过 10^6 . 当如下两条中的一个满足时, 我们终止半光滑 Newton 法: (1) 残差函数不大于 $\epsilon_k \equiv 10^{-9}$; (2) 迭代次数超过 10^4 .

我们所关注的指标有如下四个: (1) 算法输出的目标函数值 (f_{out}); (2) 电子位置间的近似映射 $\{\hat{\mathcal{T}}_i\}_{i=2}^N : \mathbb{R}^d \rightarrow \mathbb{R}^d$, 定义如下 (可参见 (1.23) 式):

$$\hat{\mathcal{T}}_i(\mathbf{d}_j) := \sum_{1 \leq k \leq K} \mathbf{d}_k y_{i,jk}, \quad j = 1, \dots, K, \quad i = 2, \dots, N; \quad (3.32)$$

表 3.4 在一维三电子体系上, CMGOPT 框架在每层网格上输出的目标函数值与映射误差。其中 “-” 表示初始网格上没有初始点, “err_map_s” 与 “err_map_e” 分别表示初始点与 PALM-I 算法输出解的映射误差

Table 3.4 The objective values and mapping errors at each level given by the CMGOPT framework on the one-dimensional systems with three electrons. The notation “-” means that no initial point is available for the initial mesh, the columns “err_map_s” and “err_map_e” list the mapping errors evaluated at the initial points and output solutions of the PALM-I method, respectively

ℓ	体系一				体系二				体系三			
	K	f_{out}	err_map _s	err_map _e	K	f_{out}	err_map _s	err_map _e	K	f_{out}	err_map _s	err_map _e
0	12	18.114	-	0.031	12	10.695	-	0.034	12	5.935	-	0.040
1	24	18.911	0.049	0.013	24	11.301	0.053	0.016	24	6.275	0.053	0.018
2	48	19.004	0.022	0.009	48	11.362	0.026	0.011	48	6.346	0.027	0.013
3	96	19.019	0.014	0.004	96	11.370	0.016	0.007	96	6.356	0.019	0.012
4	192	19.021	0.007	0.003	192	11.372	0.011	0.004	192	6.360	0.013	0.001
5	384	19.022	0.007	0.002	384	11.373	0.006	0.002	384	6.361	0.003	0.000
6	768	19.022	0.004	0.001	768	11.373	0.003	0.000	768	6.361	0.001	0.000

(3) 映射 $\{\hat{\mathcal{T}}_i\}_{i=2}^N$ 的误差 $\text{err_map} \geq 0$, 其在 Monge 拟设正确时 (例如对一维体系^[100]) 按如下公式计算:

$$\text{err_map} := \frac{1}{K |\Omega|} \sum_{k=1}^K \sum_{i=2}^N \|\mathcal{T}_i(\mathbf{d}_k) - \hat{\mathcal{T}}_i(\mathbf{d}_k)\|;$$

(4) 向量 $\hat{\lambda} \in \mathbb{R}^K$ 作为 SCE 势的近似, 定义如下:

$$\hat{\lambda} := \bar{\lambda} - \min_{k=1}^K \{\bar{\lambda}_k\} \cdot \mathbf{1}_K, \quad (3.33)$$

其中

$$\bar{\lambda} := \frac{1}{N-1} \sum_{i=2}^N \lambda_{i,\rho} \in \mathbb{R}^K,$$

$\lambda_{i,\rho} \in \mathbb{R}^K$ 为半光滑 Newton 法输出的对应于约束 $Y_i^\top \rho = \rho$ 的对偶变量 ($i = 2, \dots, N$).

3.5.2.4 一维体系上的数值模拟

我们模拟表 3.2 中的六个一维体系. 在 CMGOPT 框架中, 我们使用等质量剖分生成初始网格, 随后一致加密网格, 即按尺寸平分每个粗网格单元.

对三电子体系, CMGOPT 框架的初始网格数为 $K_0 = 12$, 加密次数 $\ell_{\max} = 6$. 值得一提的是, 若直接求解最后一层网格对应的离散 MMOT (1.18), 我们面临的变量数将达 4.53×10^8 , 而问题 (2.8) 所涉及的变量数为 1.18×10^6 .

我们在表 3.4 中列出 CMGOPT 框架在每层网格上输出的 f_{out} 与 err_map , 其中 “-” 表示初始网格上没有初始点, “err_map_s” 表示初始点的映射误差, 而

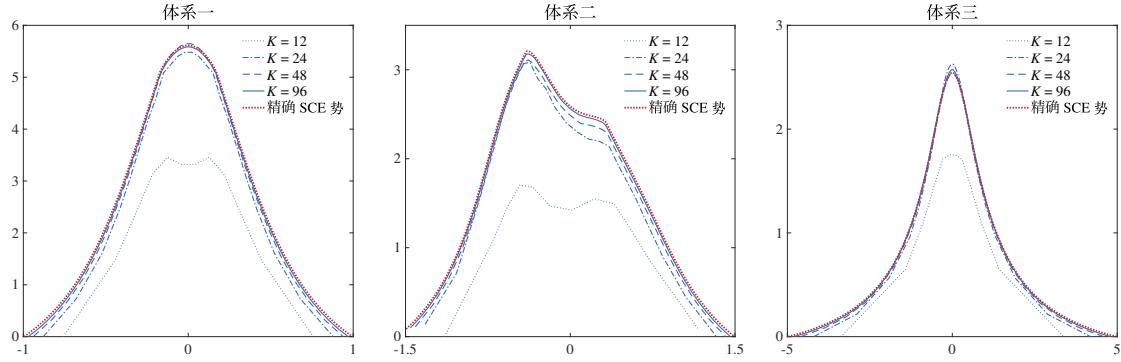


图 3.8 在一维三电子体系上, CMGOPT 框架在第 $\ell = 0 \sim 3$ 层网格上输出的 $\hat{\lambda}$ 以及精确 SCE 势. 从左至右分别为体系一、二、三的 $\hat{\lambda}$ 与 u_{SCE} , 其中蓝色点线、点划线、虚线、实线分别表示 CMGOPT 框架在第 $\ell = 0 \sim 3$ 层网格上输出的 $\hat{\lambda}$, 红色点线表示精确 SCE 势

Figure 3.8 The vectors $\hat{\lambda}$ at levels $0 \sim 3$ given by the CMGOPT framework as well as the exact SCE potentials on the one-dimensional systems with three electrons. From left to right are respectively $\hat{\lambda}$ and u_{SCE} for systems 1, 2, and 3. The blue dotted, dashdotted, dashed, and solid lines represent $\hat{\lambda}$ output by the CMGOPT framework at levels $0 \sim 3$, respectively, and the red dotted lines refer to the exact SCE potentials

“err_map_e” 表示 PALM-I 算法输出解的映射误差. 随着网格的逐层加密, CMGOPT 框架输出的目标函数值趋于收敛, PALM-I 算法输出解的映射误差逐渐趋向 0. 此外, 每层初始点处的映射误差也同样趋于 0. 我们还在图 3.8 中画出了 CMGOPT 框架在第 $\ell = 0 \sim 3$ 层网格上输出的 $\hat{\lambda}$, 并将它们与精确 SCE 势作对比. 不难发现, 随着网格的逐层加密, CMGOPT 框架输出的 $\hat{\lambda}$ 逐渐接近精确 SCE 势. 最后, 我们将第 $\ell = 0, 2, 4, 6$ 层的近似映射 $\{\hat{\mathcal{T}}_i\}_{i=2}^N$ 展示在了图 3.9 中. 所有这些结果都展现了 CMGOPT 框架以及其中插值算子(见算法 3.4)的有效性. 我们取得的数值结果可以完美地贴合理论预测 [100].

我们再考虑三个一维七电子体系. 对这三个体系, CMGOPT 框架的初始网格数为 $K_0 = 14$, 加密次数 $\ell_{\max} = 6$. 若我们直接求解最后一层网格对应的离散 MMOT (1.18), 则我们面临的变量数将达 4.64×10^{20} , 而问题 (2.8) 所涉及的变量数仅为 4.82×10^6 .

我们在表 3.5 中列出 CMGOPT 框架在每层网格上输出的 f_{out} 与 err_map. 随着网格的逐层加密, CMGOPT 框架输出的目标函数值趋于收敛, PALM-I 算法输出解的映射误差逐渐趋向 0. 此外, 每层初始点处的映射误差也同样趋于 0. 我们还在图 3.10 中画出了 CMGOPT 框架在第 $\ell = 0 \sim 3$ 层网格上输出的 $\hat{\lambda}$, 并将它们与精确 SCE 势作对比. 随着网格的逐层加密, CMGOPT 框架输出的 $\hat{\lambda}$ 逐渐接近精确 SCE 势. 最后, 我们将第 $\ell = 0, 2, 4, 6$ 层的近似映射 $\{\hat{\mathcal{T}}_i\}_{i=2}^N$ 展示在了图 3.11 中. 所有这些结果再次展现了 CMGOPT 框架以及其中插值算子的有效性. 我们取得的数值结果完美地贴合理论预测 [100]. 需要说明的是, 在表 3.5 中, 对于同一层网格, 会出现 err_map_e 高于 err_map_s 的情况(例如体系四的第二层). 此时, PALM-I 算法在不过分破坏初始点质量的同时降低了约束违反度.

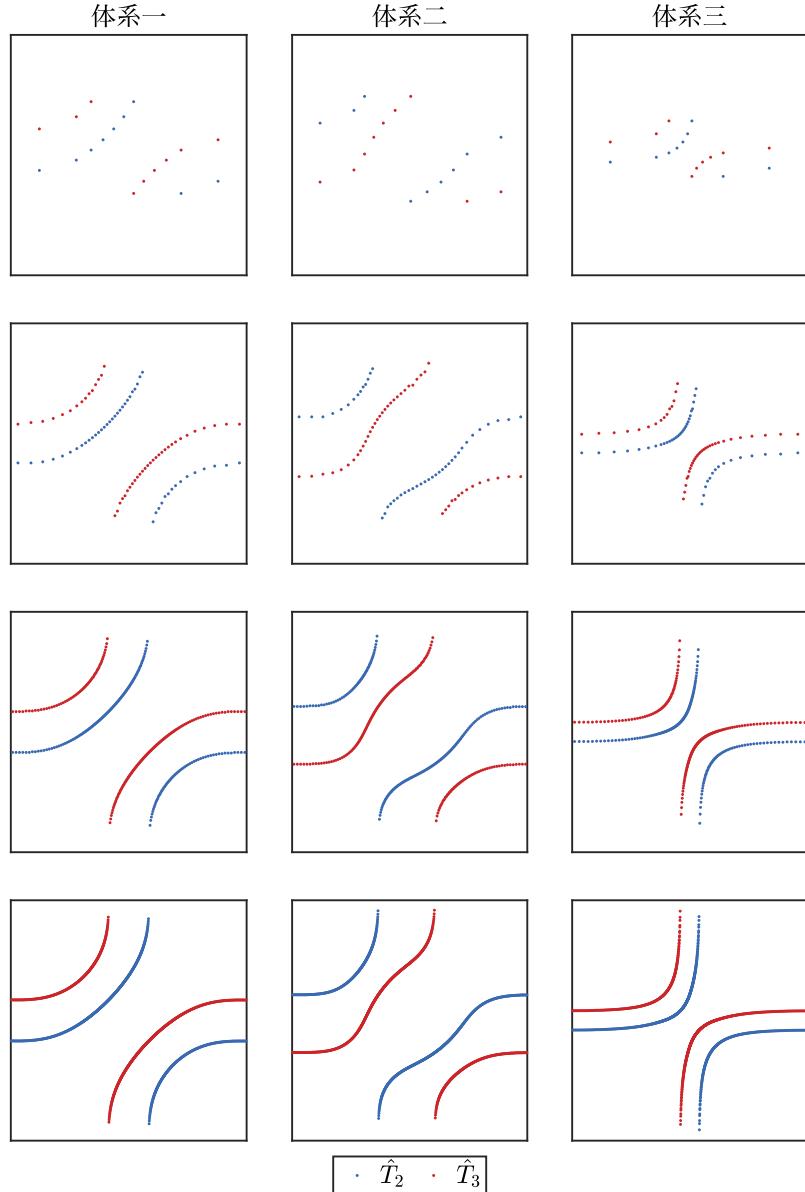


图 3.9 在一维三电子体系上, CMGOPT 框架在第 $\ell = 0, 2, 4, 6$ 层网格上输出的近似映射 $\{\hat{\mathcal{T}}_i\}_{i=2}^N$. 从左至右分别为体系一、二、三的近似映射. 其中蓝点与红点分别表示单元重心在 $\hat{\mathcal{T}}_2$ 与 $\hat{\mathcal{T}}_3$ 下的像

Figure 3.9 The approximate mappings at levels 0, 2, 4, 6 given by the CMGOPT framework on the one-dimensional systems with three electrons. From left to right are respectively the approximate mappings for systems 1, 2, and 3. The blue and red dots represent the images of barycenters under $\hat{\mathcal{T}}_2$ and $\hat{\mathcal{T}}_3$, respectively

相较从随机初始点出发调用 PALM-I 算法求解 CMGOPT 框架每层的问题, 我们的插值算子可帮助 PALM-I 算法在更短的时间内算到更好的目标函数值. 我们在表 3.6 中展示了相关的比较, 其中随机初始点使用 MATLAB 自带函数 “rand” 生成.

表 3.5 在一维七电子体系上, CMGOPT 框架在每层网格上输出的目标函数值与映射误差. 其中“-”表示初始网格没有初始点, “err_map_s”与“err_map_e”分别表示初始点与 PALM-I 算法输出解的映射误差

Table 3.5 The objective values and mapping errors at each level given by the CMGOPT framework on the one-dimensional systems with seven electrons. The notation “-” means that no initial point is available for the initial mesh, the columns “err_map_s” and “err_map_e” list the mapping errors evaluated at the initial points and output solutions of the PALM-I method, respectively

ℓ	体系四				体系五				体系六			
	K	f_{out}	err_map _s	err_map _e	K	f_{out}	err_map _s	err_map _e	K	f_{out}	err_map _s	err_map _e
0	14	173.951	-	0.052	14	151.891	-	0.039	14	111.964	-	0.030
1	28	181.474	0.045	0.019	28	158.797	0.037	0.028	28	117.223	0.030	0.010
2	56	181.929	0.018	0.025	56	158.507	0.023	0.026	56	117.050	0.011	0.008
3	112	181.989	0.019	0.012	112	158.317	0.019	0.011	112	116.914	0.007	0.008
4	224	181.954	0.014	0.013	224	158.267	0.008	0.008	224	116.876	0.008	0.006
5	448	181.942	0.007	0.002	448	158.255	0.010	0.004	448	116.864	0.005	0.003
6	896	181.939	0.002	0.001	896	158.254	0.005	0.002	896	116.861	0.003	0.001

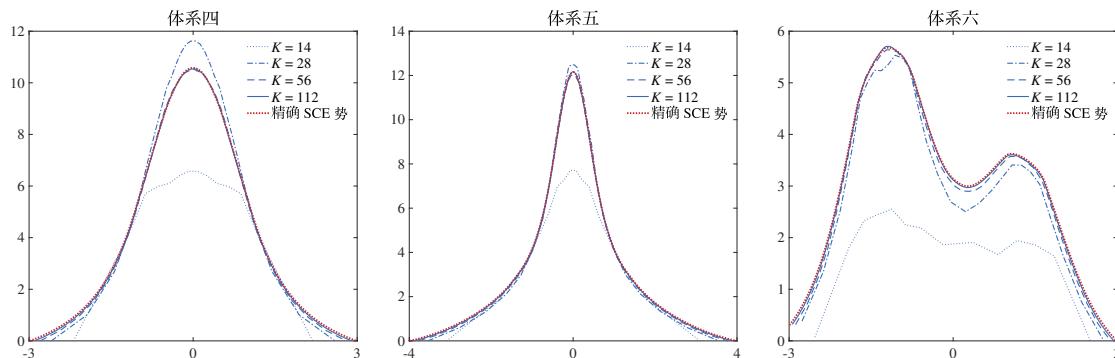


图 3.10 在一维七电子体系上, CMGOPT 框架在第 $\ell = 0 \sim 3$ 层网格上输出的 $\hat{\lambda}$ 以及精确 SCE 势. 从左至右分别为体系四、五、六的 $\hat{\lambda}$ 与 u_{SCE} . 其中蓝色点线、点划线、虚线、实线分别表示 CMGOPT 框架在第 $\ell = 0 \sim 3$ 层网格上输出的 $\hat{\lambda}$, 红色点线表示精确 SCE 势

Figure 3.10 The vectors $\hat{\lambda}$ at levels $0 \sim 3$ given by the CMGOPT framework as well as the exact SCE potentials on the one-dimensional systems with seven electrons. From left to right are respectively $\hat{\lambda}$ and u_{SCE} for systems 4, 5, and 6. The blue dotted, dashdotted, dashed, and solid lines represent $\hat{\lambda}$ output by the CMGOPT framework at levels $0 \sim 3$, respectively, and the red dotted lines refer to the exact SCE potentials

3.5.2.5 二维体系上的数值模拟

我们接着模拟两个二维三电子体系. 在 CMGOPT 框架中, 我们使用有限元软件 FreeFEM^[230] 生成非均匀初始网格. 在生成该网格时, 我们尽量让 ρ 的每个分量差不多. 在随后的步骤中, 我们一致加密网格.

不难看出, 在体系七中, 有两个电子位于 Ω 的左半部分 (由中心为 $[-1.5, 0]^T$ 的 Gauss 型函数表示), 另一个电子位于 Ω 的右半部分 (由中心为 $[1.5, 0]^T$ 的 Gauss

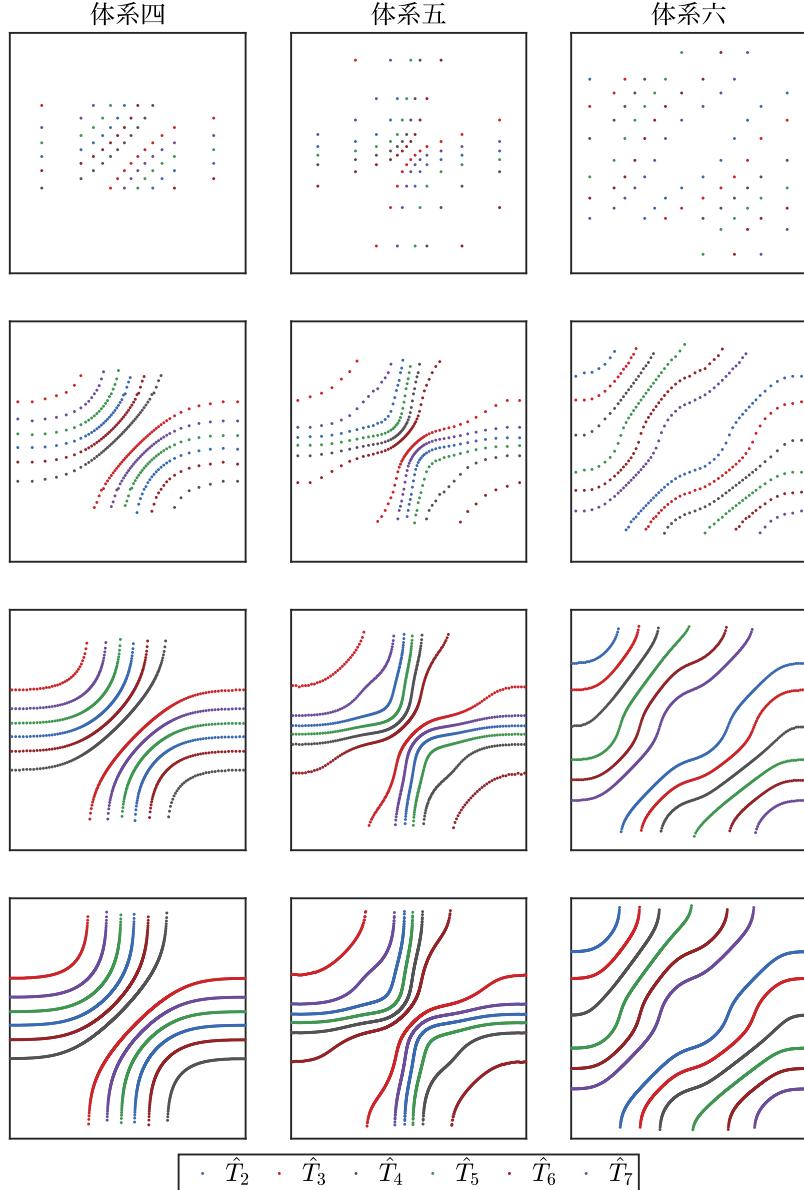


图 3.11 在一维七电子体系上, CMGOPT 框架在第 $\ell = 0, 2, 4, 6$ 层网格上输出的近似映射 $\{\hat{\mathcal{I}}_i\}_{i=2}^N$. 从左至右分别为体系四、五、六的近似映射. 其中不同颜色的点分别表示单元重心在 $\{\hat{\mathcal{I}}_i\}_{i=2}^N$ 下的像

Figure 3.11 The approximate mappings at levels 0, 2, 4, 6 given by the CMGOPT framework on the one-dimensional systems with seven electrons. From left to right are respectively the approximate mappings for systems 4, 5, and 6. The dots in different colors represent the images of barycenters under $\{\hat{\mathcal{I}}_i\}_{i=2}^N$

型函数表示); 在体系八中, 三个电子分别位于三个不同的位点 $[-1.032, -0.84]^\top$ 、 $[0, 0.96]^\top$ 、 $[1.032, -0.84]^\top$ (分别由三个 Gauss 型函数表示). 对这两个体系, CMGOPT 框架的初始网格数分别为 $K_0 = 240$ 与 $K_0 = 170$, 加密次数 $\ell_{\max} = 4$. 初始离散网格可见图 3.13 的第一行. 若我们直接求解最后一层网格对应的离散 MMOT (1.18), 则我们面临的变量数将分别为 3.62×10^{12} 与 1.29×10^{12} , 而问题 (2.8) 所涉及的变量数仅为 4.72×10^8 与 2.37×10^8 .

表 3.6 在一维三电子体系 ($K = 768$) 上, 从随机初始点与由插值算子构造的初始点出发, PALM-I 算法收敛到的目标函数值与所需运行时间. 随机初始化的结果取 10 次模拟的平均值

Table 3.6 The objective values output and running time needed by the PALM-I method starting at the initial points constructed from random initialization and interpolation operator on the one-dimensional systems with three electrons ($K = 768$). The results associated with random initialization are averaged over 10 trials

初始化方式	体系一		体系二		体系三	
	f_{out}	运行时间(秒)	f_{out}	运行时间(秒)	f_{out}	运行时间(秒)
随机初始化	19.037	823.98	11.379	716.02	6.396	538.23
插值算子(算法 3.4)	19.022	197.20	11.373	209.30	6.361	98.55

表 3.7 在二维三电子体系上, CMGOPT 框架在每层网格上输出的目标函数值

Table 3.7 The objective values at each level given by the CMGOPT framework on the two-dimensional systems with three electrons

ℓ	体系七		体系八	
	K	f_{out}	K	f_{out}
0	240	9.503	170	9.491
1	960	9.577	680	9.533
2	3840	9.598	2720	9.543
3	15360	9.604	10880	9.546
4	61440	9.606	43520	9.547

我们在表 3.7 中列出 CMGOPT 框架在每层网格上输出的 f_{out} . 由于对这两个体系暂无最优解构造, 因此我们不列出映射误差. 随着网格的逐层加密, CMGOPT 框架输出的目标函数值趋于收敛. 我们还在图 3.12 中画出了 CMGOPT 框架在第 $\ell = 0 \sim 3$ 层网格上输出的 $\hat{\lambda}$. 不难发现, 随着网格的逐层加密, CMGOPT 框架输出的 $\hat{\lambda}$ 也趋于收敛. 最后, 我们通过绘制近似映射 $\{\hat{\mathcal{T}}_i\}_{i=2}^N$ 的切片 (slices) 评估解的质量 (见图 3.13). 具体地, 我们画出了某些子区域 $\tilde{\Omega} \subseteq \Omega$ 中单元的重心在近似映射下的像. 对于体系七, 图 3.13 表明若第一个电子位于区域左边 Gauss 型函数中心附近, 则第三个电子会相应位于右边 Gauss 型函数中心附近, 而第二个电子出现的区域会完全包裹第一个电子, 但满足 Coulomb 排斥效应; 若第一个电子位于区域右边 Gauss 型函数中心附近, 则另外两个电子会徘徊在左边 Gauss 型函数中心附近, 但彼此之间保持距离. 对于体系八, 图 3.13 表明若一个电子在其中一个 Gauss 型函数中心附近, 则另外两个电子会相应徘徊在另外两个 Gauss 型函数中心附近. 上面的结果表明充分展现了 CMGOPT 框架的有效性. 我们取得的数据结果与物理直观相合. 值得一提的是, 我们在图 3.13 中首次可视化了二维情形电子位置之间的映射.

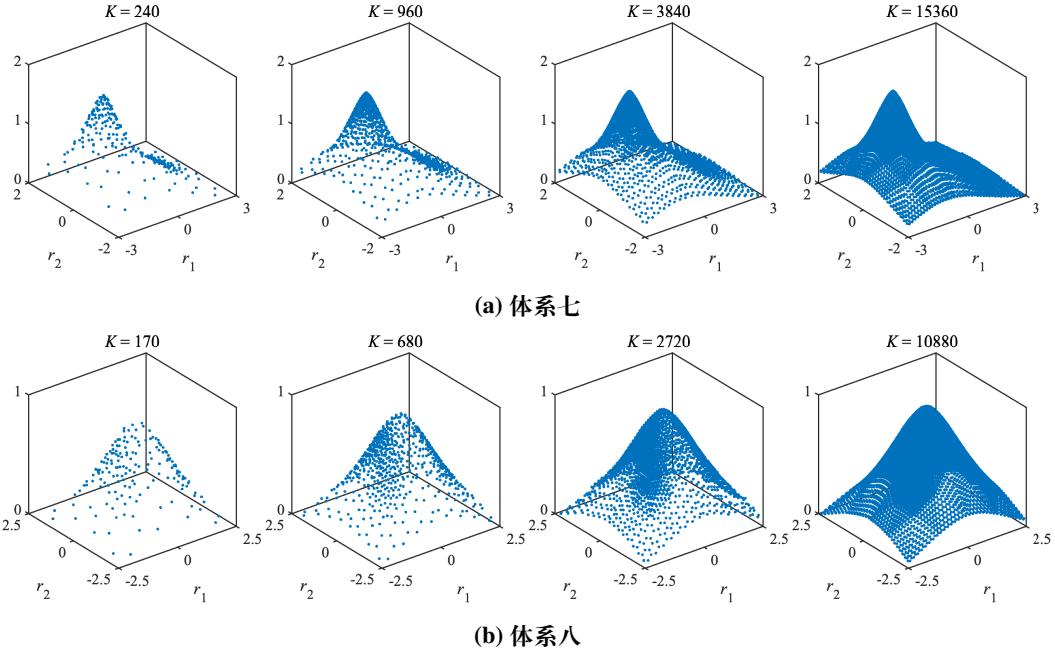


图 3.12 在二维三电子体系上, CMGOPT 框架在第 $\ell = 0 \sim 3$ 层网格上输出的 $\hat{\lambda}$. (a) 体系七.
(b) 体系八

Figure 3.12 The vectors $\hat{\lambda}$ at levels $0 \sim 3$ given by the CMGOPT framework on two-dimensional systems with three electrons. (a) System 7. (b) System 8

尽管暂无已有工作表明 Monge 拟设在体系七与八上的正确性, 但我们的数值结果为此提供了一些正面的证据, 例如图 3.13 所示. 此外, 我们还记录了解 $\{Y_i\}_{i=2}^N$ 每行的非零元素个数以及对应的重心距离, 分别定义为

$$\text{nnz}_{ij} := |\{k : y_{i,jk} > 0\}|, \quad d_{ij} := \sum_{\substack{k: y_{i,jk} > 0 \\ k': y_{i,jk'} > 0}} |\mathbf{d}_k - \mathbf{d}_{k'}|, \quad j = 1, \dots, K, \quad i = 2, \dots, N.$$

对 $i = 2, \dots, N$ 与 $j = 1, \dots, K$, nnz_{ij} 表示在固定第一个电子所在单元 e_j 时, 第 i 个电子可能出现的单元个数, 而 d_{ij} 则是这些单元重心两两之间的距离. 我们调用 MATLAB 自带函数 “hist3”, 将 $\{\text{nnz}_{ij}\}_{ij}$ 与 $\{d_{ij}\}_{ij}$ 的联合频数百分比分布绘制在图 3.14 中. 可以看出, 在我们算到的解中, 电子位置的局域化 (localization) 非常显著: 当第一个电子的位置被固定时, 仅有少数位置可供第 i 个电子选择, 且这些位置十分接近 ($i = 2, \dots, N$). 这说明类 Monge 拟设中的耦合函数接近退化为 Monge 拟设中的映射.

3.6 本章小结

在本章中, 我们考虑了具有块状结构的优化问题, 并研究了 PALM-I 算法的理论性质. 强关联电子体系计算中的问题 (2.8) 是此类问题的特例. 我们的研究动机来源于在许多实际应用中, 人们倾向或不得不使用不可行方法求解 PALM 算法的子问题. 随之而来的目标函数值序列非单调性无法避免, 给算法理论性质的

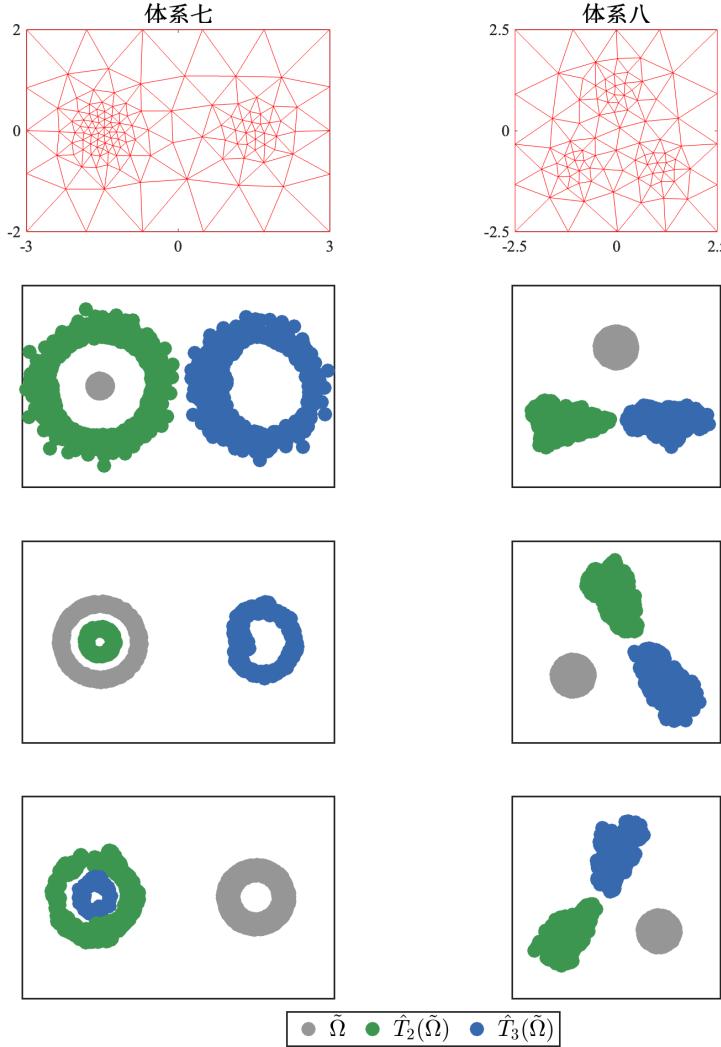


图 3.13 在二维三电子体系上, 初始离散网格 (第一行) 与 CMGOPT 框架在第 $\ell = 3$ 层网格上输出的近似映射 $\{\hat{\mathcal{T}}_i\}_{i=2}^N$ (余下三行) 切片. 从左至右分别为体系七与八的初始离散网格与近似映射. 其中灰色、绿色与蓝色部分分别表示原像 $\tilde{\Omega} \subseteq \Omega$ 及其在 $\hat{\mathcal{T}}_2$ 、 $\hat{\mathcal{T}}_3$ 下的像

Figure 3.13 The initial meshes (the first row) and approximate mappings at level 3 given by the CMGOPT framework (the remaining three rows) on the two-dimensional systems with three electrons. From left to right are respectively the initial meshes and approximate mappings for systems 7 and 8. The gray, blue, and green parts represent the pre-images $\tilde{\Omega} \subseteq \Omega$ and their images under $\hat{\mathcal{T}}_2$ and $\hat{\mathcal{T}}_3$, respectively

分析带来了巨大的困难. 已有工作需要为子问题的非精确求解设置无法实现的准则. 为此, 我们首先基于子问题的误差界提出了可实现的非精确求解准则, 并依此构造了单调下降的代理序列, 从而允许非单调目标函数值序列的存在. 我们首次在可实现的条件下, 建立了 PALM-I 算法的全局依子 (点) 列收敛性与渐进收敛速度.

在数值实验中, PALM-I 算法相较 PALM-E 算法与 PALM-F 算法具有显著的效率优势. 为加速大规模问题 (2.8) 的求解, 我们引入了 CMGOPT 框架, 并将 PALM-I 算法作为其中的局部优化算法. 以此, 我们模拟了一维、二维强关联电子

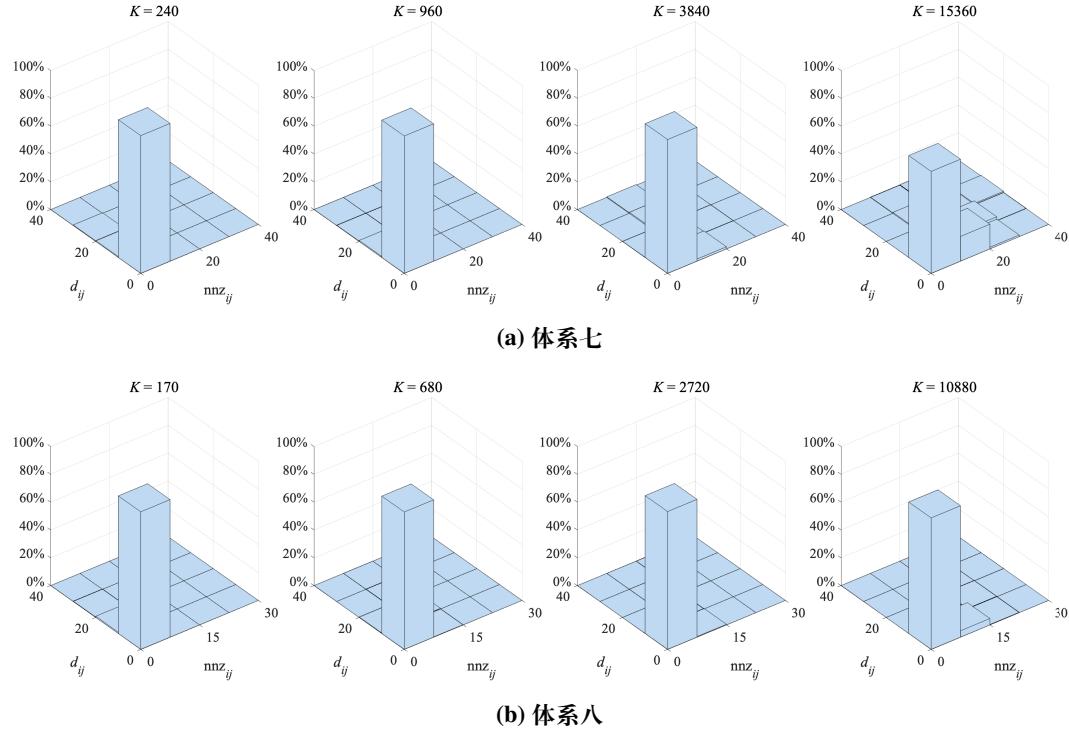


图 3.14 在二维三电子体系上,解 $\{Y_i\}_{i=2}^N$ 的 $\{\text{nnz}_{ij}\}_{ij}$ 与 $\{d_{ij}\}_{ij}$ 联合频数百分比分布图. (a) 体系七. (b) 体系八

Figure 3.14 The joint frequency percentage distributions of $\{\text{nnz}_{ij}\}_{ij}$ and $\{d_{ij}\}_{ij}$ of the solutions $\{Y_i\}_{i=2}^N$ on the two-dimensional systems with three electrons. (a) System 7. (b) System 8

体系, 取得了符合理论预测与物理直观的数值结果. 我们还首次可视化了二维情形下电子位置之间的映射. 这些结果充分展现了 CMGOPT 框架、插值算子以及 PALM-I 算法的有效性.

第4章 求解运输多胞体上分块矩阵优化问题的块坐标下降型算法

在本章中, 我们考虑运输多胞体上的分块矩阵优化问题. 强关联电子计算中的问题 (2.8) 也是此类问题的特例. 我们为其设计了完全无需全矩阵的块坐标下降型算法, 并分析了算法的理论性质. 我们在数值实验中 (1) 对比了 PALM-I 算法与新算法的数值表现; (2) 将新算法作为 CMGOPT 框架 (见框架 3.3) 中的优化算法, 模拟了一维至三维的强关联电子体系.

4.1 问题描述与研究现状

4.1.1 问题描述

我们考虑如下运输多胞体上的分块矩阵优化问题:

$$\begin{aligned} \min_{\{X_i\}_{i=1}^s} \quad & f(X_1, \dots, X_s), \\ \text{s. t.} \quad & X_i \in \mathcal{U}(\mathbf{a}_i, \mathbf{b}_i), \quad i = 1, \dots, s, \end{aligned} \tag{4.1}$$

其中对 $i = 1, \dots, s$ ($s \in \mathbb{N}$), $\mathbf{a}_i \in \mathbb{R}_+^{m_i}$ ($m_i \in \mathbb{N}$), $\mathbf{b}_i \in \mathbb{R}_+^{n_i}$ ($n_i \in \mathbb{N}$), $X_i \in \mathbb{R}^{m_i \times n_i}$, 集合

$$\mathcal{U}(\mathbf{a}_i, \mathbf{b}_i) := \left\{ T \in \mathbb{R}_+^{m_i \times n_i} : T \mathbf{1}_{n_i} = \mathbf{a}_i, T^\top \mathbf{1}_{m_i} = \mathbf{b}_i \right\}$$

是运输多胞体. 目标函数 $f : \bigtimes_{i=1}^s \mathbb{R}^{m_i \times n_i} \rightarrow \mathbb{R}$ 在可行域上分块 Lipschitz 光滑, 可能非凸. 问题 (4.1) 在许多领域都具有应用. 例如, 求多个离散概率分布的 Wasserstein 重心^[231] 在统计^[232] 与机器学习^[233] 领域受到了广泛关注, 而问题 (4.1) 可作为其中的子问题. 它还出现在监督学习里的标签分布学习任务^[234,235] 中, 其可捕捉不同标签的相对重要性. 特别地, 强关联电子体系计算中的问题 (2.8) 在经过变量替换 $X_i := \text{Diag}(\boldsymbol{\rho}) Y_i \in \mathbb{R}^{K \times K}$ ($i = 2, \dots, N$) 后也具有问题 (4.1) 的形式.

4.1.2 研究现状

目前, 求解具有块状结构优化问题最流行的算法之一是块坐标下降 (block coordinate descent, BCD) 型算法. 这些算法可以充分利用问题的分块结构, 其中子问题仅涉及单个变量块自由度, 比原问题容易求解得多. 具有代表性的 BCD 型算法包括 BCD 算法^[236–238]、分块条件梯度 (block conditional gradient, BCG) 算法^[239,240]、PALM 算法^[173] 以及它们的随机版本^[239,241–245], 其中随机性体现在梯度的计算或更新的次序. 然而对于问题 (4.1), 已有的这些 BCD 型算法都需要显式地存储或操作矩阵变量, 因此在每步迭代需要至少平方增长的复杂度. 这在矩阵维数 $\{m_i\}_{i=1}^s$ 、 $\{n_i\}_{i=1}^s$ 较大时会带来沉重的存储与计算代价. 以强关联电子体系应用中的问题 (2.8) 为例, $m_i (= n_i)$ 表示离散规模, 在实际应用中可达数十万甚至上百万.

当 s 等于 1 且 f 线性时, 问题 (4.1) 退化为经典最优运输问题^[97] 的 Kantorovich 松弛形式 ($m_i = m, n_i = n$). 最优运输问题最早由 Monge 于 18 世纪提出^[246]. Kantorovich 在 20 世纪提出了该问题的 Kantorovich 松弛形式^[247]. 因最优运输问题在流体力学、机器学习、图像处理等领域的广泛应用^[96], 其求解算法层出不穷. 传统的求解方法包括求解 Monge-Ampère 方程^[248,249] 与直接调用线性规划求解器^[250,251]. 前者需运输费用函数为平方欧氏距离, 而后者在每次迭代需要立方增长的计算复杂度. 当前, 最受欢迎的求解方法是基于熵正则的方法^[96,252]. 在使用 Sinkhorn 算法^[253,254] 求解子问题时, 这些方法近似求解问题的计算复杂度为 $\mathcal{O}(t_{\max}mn)$, 其中 $t_{\max} \in \mathbb{N}$ 是迭代次数. 近年来, 受大规模应用的驱动, 有学者提出了一些 Sinkhorn 算法的变体, 用于消除原有算法的平方增长复杂度, 包括基于低秩近似的^[255] 和基于矩阵逐元素随机近似的^[256] 变体. 特别地, 后者本质上求解的是一个限制 (restricted) 最优运输问题:

$$\min_X \langle \tilde{C}, X \rangle, \text{ s. t. } X \in \mathcal{U}(\mathbf{a}, \mathbf{b}), X_{I^c} = 0, \quad (4.2)$$

其中 $\tilde{C} \in \mathbb{R}^{m \times n}$ 是“有效”(effective) 费用矩阵(可见后文 (4.14) 式定义), $X \in \mathbb{R}^{m \times n}$, $\mathbf{a} \in \mathbb{R}^m$, $\mathbf{b} \in \mathbb{R}^n$, 集合 $I \subseteq \{(j, k) : j = 1, \dots, m, k = 1, \dots, n\}$ 是在一开始通过随机采样得到的指标集. 采样依据的概率分布与 \mathbf{a} 和 \mathbf{b} 有关. 约束 “ $X_{I^c} = 0$ ” 要求矩阵 X 中指标在 I^c 中的那些元素是 0. 有了这个额外的约束, 矩阵 X 中参与计算和更新的元素就只有 $|I|$ 个了. 如果 $|I| = o(mn)$, 则算法的计算复杂度就可被显著降低. 不过, 对于具有块状结构的非凸问题 (4.1), 这一基于采样的策略是否可以使用仍不得而知. 一种较为直接的方式是将采样策略结合到 BCD 型算法里去. 这样一来, 子问题将具有与问题 (4.2) 类似的形式. 但随着迭代的进行, 采样导致的误差将不断积聚. 此时, 该如何保证算法的收敛性便成了一个难题.

4.1.3 本章主要内容

在本章中, 我们结合采样策略, 为问题 (4.1) 的求解设计了全新的 BCD 型算法, 其在迭代过程中完全不需要存储或计算全矩阵. 在大规模应用中, 这可为我们节省巨大的存储量与计算量. 借助矩阵逐元素随机近似理论, 我们建立了由采样导致的误差的上界, 从而证明了迭代点的平均稳定性违反度在 $\sum_{i=1}^s (m_i + n_i)$ 趋于无穷大时收敛到 0 的概率趋于 1. 我们的工作首次为矩阵逐元素随机近似在分块非凸问题上的应用提供了理论保证. 在数值实验中, 我们将新算法作为 CMGOPT 框架(见框架 3.3) 中的局部优化算法, 模拟了一维至三维的强关联电子体系(其中一个可描述解离的三维锂化氢体系^[257]), 进而首次可视化了三维情形电子位置间的映射.

4.2 算法设计

本节, 我们从 BCG 算法与 PALM 算法出发, 结合最优运输与矩阵逐元素随机近似中的工具, 为问题 (4.1) 设计两类算法.

算法 4.1: 求解问题 (4.1) 的 ERALM 算法.

输入: 初始点 $(X_1^{(0)}, \dots, X_s^{(0)}) \in \bigtimes_{i=1}^s \mathbb{R}^{m_i \times n_i}$, $\mathbf{a}_i \in \mathbb{R}^{m_i}$, $\mathbf{b}_i \in \mathbb{R}^{n_i}$ ($i = 1, \dots, s$),
最大迭代次数 $t_{\max} \in \mathbb{N}$.

```

1 置  $t := 0$ .
2 while 终止准则未满足且  $t < t_{\max}$  do
3   for  $i = 1, \dots, s$  do
4     选取正则化参数  $\lambda_{i,t} > 0$  与步长  $\alpha_{i,t} \in (0, 1]$ .
5     按照 (4.4) 式计算费用矩阵  $C_i^{(t)} \in \mathbb{R}^{m_i \times n_i}$ .
6     求解子问题 (4.5) 或 (4.6) 得到  $\tilde{X}_i^{(t+1)} \in \mathbb{R}^{m_i \times n_i}$ .
7     更新变量块  $X_i^{(t+1)} := (1 - \alpha_{i,t})X_i^{(t)} + \alpha_{i,t}\tilde{X}_i^{(t+1)} \in \mathbb{R}^{m_i \times n_i}$ .
8   end
9   置  $t := t + 1$ .
10 end
```

输出: $(X_1^{(t)}, \dots, X_s^{(t)}) \in \bigtimes_{i=1}^s \mathbb{R}^{m_i \times n_i}$.

4.2.1 熵正则交替线性化极小化算法

在每次迭代中, BCG 算法通过求解如下子问题获取搜索方向:

$$\min_{X_i} \left\langle C_i^{(t)}, X_i - X_i^{(t)} \right\rangle, \quad \text{s. t. } X_i \in \mathcal{U}(\mathbf{a}_i, \mathbf{b}_i), \quad (4.3)$$

其中

$$C_i^{(t)} := \nabla_{X_i} f(X_{<i}^{(t+1)}, X_{\geq i}^{(t)}) \in \mathbb{R}^{m_i \times n_i}. \quad (4.4)$$

这里, $X_{<i}$ 与 $X_{\geq i}$ 的定义与 (3.2) 式类似. 不难发现, 问题 (4.3) 可看作以 $C_i^{(t)}$ 为费用矩阵的最优运输问题. 如之前我们所提到的, 若将问题 (4.3) 当作一般的线性规划用内点法求解, 我们在每次迭代均需要立方增长的计算复杂度. 受熵正则最优运输的启发, 我们在目标函数中增加熵正则项, 得到如下子问题:

$$\min_{X_i} \left\langle C_i^{(t)}, X_i - X_i^{(t)} \right\rangle + \lambda_{i,t} h(X_i), \quad \text{s. t. } X_i \mathbf{1}_{n_i} = \mathbf{a}_i, \quad X_i^\top \mathbf{1}_{m_i} = \mathbf{b}_i. \quad (4.5)$$

这里, $\lambda_{i,t} > 0$ 是正则化参数. 对任意 $T = (t_{ij}) \in \mathbb{R}_+^{m \times n}$, 其(负)熵^[258] 定义为

$$h(T) := \sum_{ij} t_{ij} (\ln t_{ij} - 1).$$

在热力学中, 熵被作为体系混乱度的度量. 在信息论中, 熵则被用来度量信息的不确定性. 根据 h 的定义, 我们无需在问题 (4.5) 中添加非负约束. 将 BCG 算法中的子问题换成问题 (4.5), 我们即可得到熵正则交替线性化极小化 (entropy regularized alternating linearized minimization, ERALM) 算法 (见算法 4.1).

注意到在问题 (4.5) 中变量个数随 m_i 和 n_i 平方增长, 而等式约束个数线性增长. 因此, 当 m_i 或 n_i 较大时, 从对偶的角度求解之更加划算. 根据文献^[96], 问题

(4.5) 的对偶问题是

$$\min_{\tilde{\mathbf{u}}_i, \tilde{\mathbf{v}}_i} q_i(\tilde{\mathbf{u}}_i, \tilde{\mathbf{v}}_i; \lambda_{i,t}, \Psi_i^{(t)}) := \lambda_{i,t} \exp\left(\frac{\tilde{\mathbf{u}}_i}{\lambda_{i,t}}\right) \Psi_i^{(t)} \exp\left(\frac{\tilde{\mathbf{v}}_i}{\lambda_{i,t}}\right) - \tilde{\mathbf{u}}_i^\top \mathbf{a}_i - \tilde{\mathbf{v}}_i^\top \mathbf{b}_i, \quad (4.6)$$

其中 $\tilde{\mathbf{u}}_i \in \mathbb{R}^{m_i}$ 和 $\tilde{\mathbf{v}}_i \in \mathbb{R}^{n_i}$ 是分别对应于等式约束 $X_i \mathbf{1}_{n_i} = \mathbf{a}_i$ 和 $X_i^\top \mathbf{1}_{m_i} = \mathbf{b}_i$ 的对偶变量,

$$\Psi_i^{(t)} := \exp\left(-\frac{C_i^{(t)}}{\lambda_{i,t}}\right) \in \mathbb{R}^{m_i \times n_i}$$

称为问题 (4.5) 的 (Gibbs) 核 (kernel) 矩阵, “ $\exp(\cdot)$ ”是对向量、矩阵逐元素取自然指数的操作. 由于问题 (4.6) 具有块状结构, 因此我们自然会考虑使用 BCD 算法求解之. 在最优运输领域, 该算法以 Sinkhorn 算法的别称为人熟知^[253,254]. 给定初始 $\tilde{\mathbf{v}}_i^{(t,0)} \in \mathbb{R}^{n_i}$, Sinkhorn 算法在每次迭代中按如下方式更新对偶变量, 直至满足特定的终止准则:

$$\begin{aligned} \tilde{\mathbf{u}}_i^{(t,l+1)} &:= \lambda_{i,t} \ln\left(\mathbf{a}_i \oslash \left(\Psi_i^{(t)} \exp\left(\tilde{\mathbf{v}}_i^{(t,l)} / \lambda_{i,t}\right)\right)\right), \\ \tilde{\mathbf{v}}_i^{(t,l+1)} &:= \lambda_{i,t} \ln\left(\mathbf{b}_i \oslash \left(\Psi_i^{(t)\top} \exp\left(\tilde{\mathbf{u}}_i^{(t,l+1)} / \lambda_{i,t}\right)\right)\right). \end{aligned} \quad (4.7)$$

这里, l 表示 Sinkhorn 算法的迭代次数, “ $\ln(\cdot)$ ”是对向量、矩阵逐元素取自然对数的操作. 令

$$\check{\mathbf{u}}_i^{(t,l)} := \exp\left(\frac{\tilde{\mathbf{u}}_i^{(t,l)}}{\lambda_{i,t}}\right) \in \mathbb{R}^{m_i}, \quad \check{\mathbf{v}}_i^{(t,l)} := \exp\left(\frac{\tilde{\mathbf{v}}_i^{(t,l)}}{\lambda_{i,t}}\right) \in \mathbb{R}^{n_i},$$

Sinkhorn 算法的更新格式 (4.7) 可等价地写成

$$\check{\mathbf{u}}_i^{(t,l+1)} := \mathbf{a}_i \oslash \left(\Psi_i^{(t)} \check{\mathbf{v}}_i^{(t,l)}\right), \quad \check{\mathbf{v}}_i^{(t,l+1)} := \mathbf{b}_i \oslash \left(\Psi_i^{(t)\top} \check{\mathbf{u}}_i^{(t,l+1)}\right). \quad (4.8)$$

更新格式 (4.8) 仅涉及矩阵-向量乘积以及向量逐元素相除的计算, 因此具有较高的并行可扩展性^[252], 被人们广泛使用. 由于问题 (4.5) 与 (4.6) 强凸, 我们可由 BCD 算法的理论证明 Sinkhorn 算法的线性收敛速度^[259]. 在实际使用中, 热启动可给 Sinkhorn 算法带来显著的加速^[260].

4.2.2 熵正则交替线性化极小化算法的采样版本

ERALM 算法尽管可使用高可扩展的 Sinkhorn 算法求解子问题, 但仍需要显式地操作、存储矩阵变量以及计算费用矩阵 $C_i^{(t)}$ (见 (4.4) 式). 下面, 我们使用一个稀疏矩阵 $\hat{\Psi}_i^{(t)} \in \mathbb{R}^{m_i \times n_i}$ 近似 $\Psi_i^{(t)} = (\psi_{i,jk}^{(t)})$, 进而降低计算与存储代价.

我们之所以会想到使用稀疏矩阵近似, 是因为观察到问题 (4.5) 的最优解具有如下乘积形式的表达式:

$$\tilde{X}_i^{(t+1,\star)} := \text{Diag}\left(\exp\left(\frac{\tilde{\mathbf{u}}_i^{(t,\star)}}{\lambda_{i,t}}\right)\right) \Psi_i^{(t)} \text{Diag}\left(\exp\left(\frac{\tilde{\mathbf{v}}_i^{(t,\star)}}{\lambda_{i,t}}\right)\right) \in \mathbb{R}^{m_i \times n_i}, \quad (4.9)$$

其中 $\tilde{\mathbf{u}}_i^{(t,\star)} \in \mathbb{R}^{m_i}$ 和 $\tilde{\mathbf{v}}_i^{(t,\star)} \in \mathbb{R}^{n_i}$ 是对偶问题 (4.6) 的一个最优解. (4.9) 式表明, 若 $\psi_{i,jk}^{(t)}$ 为 0, 则必有 $\tilde{x}_{i,jk}^{(t+1,\star)}$ 为 0. 此外, 已有学者证明问题 (4.3) 必定存在稀疏最优解^[261]. 当 f 多仿射时, 根据第 2 章的命题 2.6, 问题 (4.1) 甚至存在只有 $\mathcal{O}(\sum_{i=1}^s (m_i + n_i))$ 个非零元的稀疏最优解. 这些说明在 $\Psi_i^{(t)}$ 中只有一小部分的元素是真正被需要的. 如果我们在实际计算中使用一个稀疏矩阵代替它, 结合 Sinkhorn 算法的迭代格式 (4.8), 算法的计算与存储代价将会显著下降.

为此, 我们只需在算法迭代的过程中估计最优解的稀疏模式. 我们采用矩阵逐元素随机近似^[262–266]. 矩阵逐元素随机近似在文献^[267] 中首次提出, 之后在文献^[262,264,265] 中被发展或推广. 其基本想法是使用某种采样算法随机抽取矩阵的一小部分元素, 进而构造矩阵的稀疏近似. 在概率意义上, 所构造的稀疏近似与原矩阵的距离可以充分小. 采样参照的概率分布与原矩阵相关. 特别地, 重要性采样从方差缩减的角度出发, 构造采样的概率分布^[268–271]. 矩阵逐元素随机近似由于可以显著降低计算与空间复杂度, 现已被应用于许多大规模计算问题, 例如特征向量的近似计算^[267]、最优运输问题的求解^[256]、Gromov-Wasserstein 距离的计算^[266] 等.

具体地, 我们对每个元素独立地使用 Poisson 采样^[256,265,272]. 相较于有放回采样, Poisson 采样在一些场景下展现出了更高的精度^[273]. 根据重要性采样理论^[266], 最优采样概率分布应当满足 $p_{i,jk}^{(t,\star)} \propto \tilde{x}_{i,jk}^{(t+1,\star)}$. 然而, 根据 (4.9) 式, 如果我们没有预先知道关于 $\tilde{\mathbf{u}}_i^{(t,\star)}$ 和 $\tilde{\mathbf{v}}_i^{(t,\star)}$ 的信息, $\tilde{X}_i^{(t+1,\star)}$ 完全无法计算. 一种简单的替代方案是依据前一个迭代点的值分布构造采样概率, 即 $p_{i,jk}^{(t)\prime} \propto x_{i,jk}^{(t)}$. 当算法接近最优解时, 这一想法是十分合理的. 此外, 为了避免因为采样而严重错估真实的稀疏模式, 我们在 $p_{i,jk}^{(t)\prime}$ 与 $p_{i,jk}'' \propto \sqrt{a_{i,j} b_{i,k}}$ 之间做插值^[274,275]. 最终采样概率为

$$p_{i,jk}^{(t)} := \gamma p_{i,jk}^{(t)\prime} + (1 - \gamma)p_{i,jk}'' = \gamma \frac{x_{i,jk}^{(t)}}{\sum_{j',k'} x_{i,j'k'}^{(t)}} + (1 - \gamma) \frac{\sqrt{a_{i,j} b_{i,k}}}{\sum_{j',k'} \sqrt{a_{i,j'} b_{i,k'}}}, \quad (4.10)$$

其中 $j \in \{1, \dots, m_i\}$, $k \in \{1, \dots, n_i\}$, $\gamma \in [0, 1]$ 表示插值因子.

给定一采样参数 $\hat{n}_i \in \mathbb{N}$, 我们依据 Poisson 采样原理按如下方式构造核矩阵的稀疏近似 $\hat{\Psi}_i^{(t)} = (\hat{\psi}_{i,jk}^{(t)}) \in \mathbb{R}^{m_i \times n_i}$:

$$\hat{\psi}_{i,jk}^{(t)} := \begin{cases} \frac{\psi_{i,jk}^{(t)}}{p_{i,jk}^{(t)\#}}, & \text{以概率 } p_{i,jk}^{(t)\#} := \min \left\{ 1, \hat{n}_i \cdot p_{i,jk}^{(t)} \right\}; \\ 0, & \text{否则,} \end{cases} \quad (4.11)$$

其中分母保证了随机近似的无偏性 (可见后文定理 4.9 的证明). 我们将采样得到的指标集记为 $\mathcal{I}_i^{(t)} \subseteq \{(j, k) : j = 1, \dots, m_i, k = 1, \dots, n_i\}$. 注意到

$$\mathbb{E} \left(|\mathcal{I}_i^{(t)}| \right) = \sum_{j,k} p_{i,jk}^{(t)\#} \leq \hat{n}_i \sum_{j,k} p_{i,jk}^{(t)} = \hat{n}_i.$$

尽管 $|\mathcal{I}_i^{(t)}|$ 的数值具有不确定性, 但它以较高概率位于其期望附近, 从而可以较高概率被参数 \hat{n}_i 控制^[272].

将对偶问题 (4.6) 中的 $\Psi_i^{(t)}$ 替换成 $\hat{\Psi}_i^{(t)}$, 我们就有如下全新子问题:

$$\min_{\tilde{\mathbf{u}}_i, \tilde{\mathbf{v}}_i} q_i(\tilde{\mathbf{u}}_i, \tilde{\mathbf{v}}_i; \lambda_{i,t}, \hat{\Psi}_i^{(t)}). \quad (4.12)$$

它是如下问题的对偶问题:

$$\begin{aligned} & \min_{X_i} \left\langle \hat{C}_i^{(t)}, X_i - X_i^{(t)} \right\rangle + \lambda_{i,t} h(X_i), \\ & \text{s. t. } X_i \mathbf{1}_{n_i} = \mathbf{a}_i, X_i^\top \mathbf{1}_{m_i} = \mathbf{b}_i, (X_i)_{(\mathcal{I}_i^{(t)})^c} = 0. \end{aligned} \quad (4.13)$$

这里, $\hat{C}_i^{(t)} = (\hat{c}_{i,jk}^{(t)}) \in \mathbb{R}^{m_i \times n_i}$ 是“有效”费用矩阵, 定义为

$$\hat{c}_{i,jk}^{(t)} := \begin{cases} c_{i,jk}^{(t)} + \lambda_{i,t} \ln \left(|p_{i,jk}^{(t)\#}| \right), & \text{若 } (j, k) \in \mathcal{I}_i^{(t)}; \\ c_{i,jk}^{(t)}, & \text{否则.} \end{cases} \quad (4.14)$$

实际上, $\hat{C}_i^{(t)}$ 只有指标在集合 $\mathcal{I}_i^{(t)}$ 中的元素需要定义. 尽管问题 (4.13) 和 (4.12) 与问题 (4.5) 和 (4.6) 形式不同, 但我们仍然可以用 Sinkhorn 算法求解它们. 我们只需把 (4.8) 式中的 $\Psi_i^{(t)}$ 替换成 $\hat{\Psi}_i^{(t)}$.

我们将加入矩阵逐元素随机近似的 ERALM 算法简称为 S-ERALM 算法. 其伪代码可见算法 4.2. 由前所述, 采样抽取指标个数的期望由 $\{\hat{n}_i\}_{i=1}^s$ 控制. 从问题 (4.13), 我们可以看出, 矩阵逐元素随机近似本质上给原始子问题增加了随机置零约束. 这使得 S-ERALM 算法 (以及后文的算法 4.4) 与已有随机块坐标下降型算法有根本上的不同. 也正是这一不同之处, 使这些算法的迭代过程完全无需全矩阵的参与.

4.2.3 基于 Kullback-Leibler 散度的交替线性化极小化算法

在每次迭代中, PALM 算法通过求解如下子问题更新变量块:

$$\min_{X_i} \left\langle C_i^{(t)}, X_i - X_i^{(t)} \right\rangle + \frac{\mu_{i,t}}{2} \|X_i - X_i^{(t)}\|^2, \text{ s. t. } X_i \in \mathcal{U}(\mathbf{a}_i, \mathbf{b}_i), \quad (4.15)$$

其中 $\mu_{i,t} > 0$ 为邻近参数, $C_i^{(t)}$ 如 (4.4) 式定义. 求解子问题 (4.15) 等价于计算 $X_i^{(t)} - C_i^{(t)}/\mu_{i,t}$ 到 $\mathcal{U}(\mathbf{a}_i, \mathbf{b}_i)$ 上的投影. 目前, 对此最高效的算法是半光滑 Newton 法^[180], 其并行可扩展性较差. 事实上, 除了欧式邻近项, 我们还可以使用其他类型的距离度量, 例如更一般的 Bregman 距离^[276]. 它已被广泛用于优化算法的设计, 例如求解凸优化问题的 Bregman 邻近点算法^[260,277] 和求解目标函数相对光滑的非凸优化问题的 Bregman 距离 PALM 算法^[278,279]. 受此启发, 在本节中, 我们将问题 (4.15) 目标函数中的欧式邻近项换成由熵 h 诱导的 Bregman 距离, 即 Kullback-Leibler (KL) 散度 (divergence)^[280], 得到如下子问题:

$$\min_{X_i} \left\langle C_i^{(t)}, X_i - X_i^{(t)} \right\rangle + \mu_{i,t} \text{KL}(X_i; X_i^{(t)}), \text{ s. t. } X_i \mathbf{1}_{n_i} = \mathbf{a}_i, X_i^\top \mathbf{1}_{m_i} = \mathbf{b}_i. \quad (4.16)$$

算法 4.2: 求解问题 (4.1) 的 S-ERALM 算法.

输入: 初始点 $(X_1^{(0)}, \dots, X_s^{(0)}) \in \bigtimes_{i=1}^s \mathbb{R}^{m_i \times n_i}$, $\mathbf{a}_i \in \mathbb{R}^{m_i}$, $\mathbf{b}_i \in \mathbb{R}^{n_i}$ ($i = 1, \dots, s$),

插值因子 $\gamma \in [0, 1]$, 采样参数 $\{\hat{n}_i\}_{i=1}^s \subseteq \mathbb{N}$, 最大迭代次数 $t_{\max} \in \mathbb{N}$.

```

1 置  $t := 0$ .
2 while 终止准则未满足且  $t < t_{\max}$  do
3   for  $i = 1, \dots, s$  do
4     选取正则化参数  $\lambda_{i,t} > 0$  与步长  $\alpha_{i,t} \in (0, 1]$ .
5     依照 (4.10) 式中的概率分布  $P_i^{(t)} = (p_{i,jk}^{(t)}) \in \mathbb{R}_+^{m_i \times n_i}$  与 Poisson 采样原
        理随机抽取指标集  $\mathcal{I}_i^{(t)} \subseteq \{(j, k) : j = 1, \dots, m_i, k = 1, \dots, n_i\}$ .
6     按照 (4.11) 式构造稀疏核矩阵  $\hat{\Psi}_i^{(t)} = (\hat{\psi}_{i,jk}^{(t)}) \in \mathbb{R}^{m_i \times n_i}$ .
7     求解子问题 (4.12) 或 (4.13) 得到  $\tilde{X}_i^{(t+1)} \in \mathbb{R}^{m_i \times n_i}$ .
8     更新变量块  $X_i^{(t+1)} := (1 - \alpha_{i,t})X_i^{(t)} + \alpha_{i,t}\tilde{X}_i^{(t+1)} \in \mathbb{R}^{m_i \times n_i}$ .
9   end
10  置  $t := t + 1$ .
11 end
输出:  $(X_1^{(t)}, \dots, X_s^{(t)}) \in \bigtimes_{i=1}^s \mathbb{R}^{m_i \times n_i}$ .
```

这里, 对任意 $T = (t_{ij}), T' = (t'_{ij}) \in \mathbb{R}_+^{m \times n}$, 它们之间的 KL 散度定义为

$$\text{KL}(T; T') := \sum_{i,j} \left[t_{ij} \left(\ln t_{ij} - \ln t'_{ij} \right) - (t_{ij} - t'_{ij}) \right].$$

若存在 (i, j) 使得 $t_{ij} > 0$ 和 $t'_{ij} = 0$, 则 $\text{KL}(T; T') = \infty$. KL 散度常被用来度量概率分布之间的差异^[96]. 基于子问题 (4.16), 我们可得到基于 KL 散度的交替线性化极小化 (KL divergence-based alternating linearized minimization, KLALM) 算法 (可见算法 4.3).

同之前一样, 我们写出问题 (4.16) 的对偶问题:

$$\min_{\mathbf{u}_i, \mathbf{v}_i} q_i(\mathbf{u}_i, \mathbf{v}_i; \mu_{i,t}, \Phi_i^{(t)}), \quad (4.17)$$

其中 q_i 定义在 (4.6) 式中, $\mathbf{u}_i \in \mathbb{R}^{m_i}$ 和 $\mathbf{v}_i \in \mathbb{R}^{n_i}$ 分别是对应于等式约束 $X_i \mathbf{1}_{n_i} = \mathbf{a}_i$ 和 $X_i^\top \mathbf{1}_{m_i} = \mathbf{b}_i$ 的对偶变量,

$$\Phi_i^{(t)} := \exp \left(-\frac{C_i^{(t)}}{\mu_{i,t}} \right) \odot X_i^{(t)} \in \mathbb{R}^{m_i \times n_i}, \quad (4.18)$$

是对应的核矩阵. 不难看出, 只要将 (4.8) 式中的 $\Psi_i^{(t)}$ 替换为 $\Phi_i^{(t)}$, Sinkhorn 算法即可用于求解问题 (4.16) 与 (4.17).

算法 4.3: 求解问题 (4.1) 的 KLALM 算法.

输入: 初始点 $(X_1^{(0)}, \dots, X_s^{(0)}) \in \times_{i=1}^s \mathbb{R}^{m_i \times n_i}$, $\mathbf{a}_i \in \mathbb{R}^{m_i}$, $\mathbf{b}_i \in \mathbb{R}^{n_i}$ ($i = 1, \dots, s$),
最大迭代次数 $t_{\max} \in \mathbb{N}$.

```

1 置  $t := 0$ .
2 while 终止条件未满足且  $t < t_{\max}$  do
3   for  $i = 1, \dots, s$  do
4     选取邻近参数  $\mu_{i,t} > 0$ .
5     按照 (4.4) 式计算费用矩阵  $C_i^{(t)} \in \mathbb{R}^{m_i \times n_i}$ .
6     求解子问题 (4.16) 或 (4.17) 得到  $X_i^{(t+1)} \in \mathbb{R}^{m_i \times n_i}$ .
7   end
8   置  $t := t + 1$ .
9 end
```

4.2.4 基于 Kullback-Leibler 散度的交替线性化极小化算法的采样版本

我们下面考虑 KLALM 算法的采样版本. 由于子问题 (4.16) 的最优解具有如下乘积形式的表达式:

$$X_i^{(t+1,\star)} := \text{Diag} \left(\exp \left(\frac{\mathbf{u}_i^{(t,\star)}}{\mu_{i,t}} \right) \right) \Phi_i^{(t)} \text{Diag} \left(\exp \left(\frac{\mathbf{v}_i^{(t,\star)}}{\mu_{i,t}} \right) \right) \in \mathbb{R}^{m_i \times n_i}, \quad (4.19)$$

其中 $(\mathbf{u}_i^{(t,\star)}, \mathbf{v}_i^{(t,\star)}) \in \mathbb{R}^{m_i} \times \mathbb{R}^{n_i}$ 是问题 (4.17) 的一个最优解, 因此我们可类比第 4.2.2 节, 在依照 (4.10) 式定义的概率分布 $P_i^{(t)} = (p_{i,jk}^{(t)}) \in \mathbb{R}^{m_i \times n_i}$ 抽取指标集 $\mathcal{I}_i^{(t)}$ 后, 使用如下稀疏矩阵 $\hat{\Phi}_i^{(t)} = (\hat{\varphi}_{i,jk}^{(t)}) \in \mathbb{R}^{m_i \times n_i}$ 近似核矩阵 $\Phi_i^{(t)} = (\varphi_{i,jk}^{(t)})$:

$$\hat{\varphi}_{i,jk}^{(t)} := \begin{cases} \frac{\varphi_{i,jk}^{(t)}}{|P_{i,jk}^{(t)\#}|}, & \text{若 } (j, k) \in \mathcal{I}_i^{(t)}; \\ 0, & \text{否则,} \end{cases} \quad (4.20)$$

将问题 (4.17) 中的 $\Phi_i^{(t)}$ 换成 $\hat{\Phi}_i^{(t)}$ 后, 我们即有全新的子问题

$$\min_{\mathbf{u}_i, \mathbf{v}_i} q_i(\mathbf{u}_i, \mathbf{v}_i; \mu_{i,t}, \hat{\Phi}_i^{(t)}). \quad (4.21)$$

它是如下问题的对偶问题:

$$\begin{aligned} \min_{X_i} & \quad \left\langle \hat{C}_i^{(t)}, X_i - X_i^{(t)} \right\rangle + \mu_{i,t} \text{KL}(X_i; X_i^{(t)}), \\ \text{s. t.} & \quad X_i \mathbf{1}_{n_i} = \mathbf{a}_i, X_i^\top \mathbf{1}_{m_i} = \mathbf{b}_i, (X_i)_{(\mathcal{I}_i^{(t)})^c} = 0. \end{aligned} \quad (4.22)$$

将 (4.8) 式中的 $\Psi_i^{(t)}$ 替换为 $\hat{\Phi}_i^{(t)}$, Sinkhorn 算法即可用于求解问题 (4.21) 与 (4.22).

现在, 我们指出在 KLALM 算法里加入采样和在 ERALM 算法里加入采样的不同之处: 若在 KLALM 算法中的每一次迭代均进行采样, 则估计的稀疏模式将

会越来越糟糕, 甚至会导致问题 (4.22) 不可行. 不妨回顾 $\Phi_i^{(t)}$ 的定义 (4.18) 式. 可以看出, 相较于 ERALM 算法, KLALM 算法的前一个迭代点可以直接影响后一个迭代点的稀疏模式: 若 $x_{i,jk}^{(t)}$ 为 0, 则 $\varphi_{i,jk}^{(t)}$ 为 0, 进而 $x_{i,jk}^{(t+1,\star)}$ 为 0. 若子问题被精确求解且我们在连续两次迭代 (例如第 t 与 $t+1$ 次迭代) 中均进行采样, 则在第 $t+1$ 次迭代中第 i 个变量块待优化的元素个数就是 $|\mathcal{I}_i^{(t-1)} \cap \mathcal{I}_i^{(t)}|$, 不会大于 $|\mathcal{I}_i^{(t-1)}|$ 和 $|\mathcal{I}_i^{(t)}|$. 在实际计算中, 我们使用 Sinkhorn 算法 (4.8) 非精确求解子问题. 然而, 其更新格式 (4.8) 同样会使 $X_i^{(t+1)}$ 继承 $X_i^{(t)}$ 的稀疏模式. 总的来说, 若在 KLALM 算法中的每一次迭代均进行采样, 则待优化的变量数会单调下降. 此时, 计算结果将完全取决于第一次采样前的状态. 而若待优化变量数下降过多, 子问题甚至可能不可行.

为了尽量消除上面的问题, 我们仅在某次 (例如第 $\hat{t}+1 \in \mathbb{N}$ 次) 迭代中进行采样. 对于 $t < \hat{t}$, 算法计算核矩阵 $\Phi_i^{(t)}$ 的所有元素. 在第 $\hat{t}+1$ 次迭代过后, 我们可以期望 $X_i^{(\hat{t}+1)}$ 已经在一定程度上捕捉到了稀疏模式. 之后对于 $t > \hat{t}$, 我们就固定待优化的指标集 $\mathcal{I}_i^{(t)} \equiv \mathcal{I}_i^{(\hat{t})}$, 不再做新的采样.

我们将加入矩阵逐元素随机近似的 KLALM 算法 (简称为 S-KLALM 算法) 总结在算法 4.4 中.

4.2.5 单步计算代价比较

我们将四个新算法 ERALM、S-ERALM、KLALM 和 S-KLALM 单步计算代价的对比展示在表 4.1 中. 其中, 我们关注主要的计算组成, 包括 (稀疏) 核矩阵计算、矩阵逐元素采样以及 Sinkhorn 算法的迭代, $t_{\max} \in \mathbb{N}$ 和 $l_{\max} \in \mathbb{N}$ 分别是外层算法的最大迭代次数与内层 Sinkhorn 算法的最大迭代次数.

注. 我们在表 4.1 中并没有显式写出计算核矩阵的代价. 这是因为在不同实际应用中, 其每个元素的计算代价差别很大. 例如, 若考虑目标函数仿射的问题 (4.1), 则每个元素的计算代价至多是 $\mathcal{O}(1)$. 而若考虑问题 (2.8), 则每个元素的计算代价是 $\mathcal{O}(K)$, 其中 $K \in \mathbb{N}$ 为离散规模.

在辅以热启动技术时, Sinkhorn 算法通常只需几步即可给出较高精度的解. 因此, 根据表 4.1, 当 $t > \hat{t}$ 时, S-KLALM 算法具有最低的单步计算代价. 假定 $\hat{n}_i = \lfloor (m_i + n_i)^{1+\tau} \rfloor$ ($\tau \in (0, 1)$), 则 S-KLALM 算法的优势将随着 τ 趋于 0 或 $m_i + n_i$ 趋于无穷大 ($i = 1, \dots, s$) 而愈加明显.

4.3 收敛性分析

本节, 我们分析 ERALM 算法与 S-ERALM 算法的理论性质. 由于 KL 散度缺乏局部 Lipschitz 光滑性, 目前我们还无法分析 KLALM 算法与 S-KLALM 算法的理论性质. 尽管如此, 我们的工作仍然首次为矩阵逐元素随机近似在分块非凸问题上的应用提供了理论保证.

算法 4.4: 求解问题 (4.1) 的 S-KLALM 算法.

输入: 初始点 $(X_1^{(0)}, \dots, X_s^{(0)}) \in \bigtimes_{i=1}^s \mathbb{R}^{m_i \times n_i}$, $\mathbf{a}_i \in \mathbb{R}^{m_i}$, $\mathbf{b}_i \in \mathbb{R}^{n_i}$ ($i = 1, \dots, s$),
插值因子 $\gamma \in [0, 1]$, 采样参数 $\{\hat{n}_i\}_{i=1}^s \subseteq \mathbb{N}$, 采样迭代数 $\hat{t} \in \mathbb{N}$, 最大迭代次数 $t_{\max} \in \mathbb{N}$.

```

1 置  $t := 0$ .
2 while 终止条件未满足且  $t < t_{\max}$  do
3   for  $i = 1, \dots, s$  do
4     选取邻近参数  $\mu_{i,t} > 0$ .
5     if  $t = \hat{t}$  then
6       | 依照 (4.10) 式中的概率分布  $P_i^{(t)} = (p_{i,jk}^{(t)}) \in \mathbb{R}_+^{m_i \times n_i}$  与 Poisson 采
      | 样原理随机抽取指标集  $\mathcal{I}_i^{(t)} \subseteq \{(j, k) : j = 1, \dots, m_i, k = 1, \dots,$ 
      |  $n_i\}$ .
7     end
8     if  $t < \hat{t}$  then
9       | 按照 (4.18) 式计算  $\hat{\Phi}_i^{(t)} := \Phi_i^{(t)}$ .
10    else
11      | 按照 (4.20) 式构造稀疏核矩阵  $\hat{\Phi}_i^{(t)} = (\hat{\varphi}_{i,jk}^{(t)}) \in \mathbb{R}^{m_i \times n_i}$ , 其中使用
      |  $\mathcal{I}_i^{(t)} = \mathcal{I}_i^{(\hat{t})}$  和  $P_i^{(t)} = P_i^{(\hat{t})}$ .
12    end
13    求解子问题 (4.21) 或 (4.22) 得到  $X_i^{(t+1)} \in \mathbb{R}^{m_i \times n_i}$ .
14  end
15  置  $t := t + 1$ .
16 end
输出:  $(X_1^{(t)}, \dots, X_s^{(t)}) \in \bigtimes_{i=1}^s \mathbb{R}^{m_i \times n_i}$ .
```

我们所取得的所有理论结果均以子问题被精确求解为前提. 这一假设在未来有望通过第 3 章的分析方法消除. 此外, 我们还假设 S-ERALM 算法中的子问题 (4.13) 都是可行的. 尽管目前我们无法提供使后者成立的充分条件, 但在数值实验中, 我们发现无需仔细选取采样参数 (例如, $\hat{n}_i = \lfloor (m_i + n_i)^{1.5} \rfloor$), 这一假设即可成立.

为表征问题 (4.1) 在某一点处的稳定性违反度, 我们定义如下残差函数: 对任意 $X := (X_1, \dots, X_s) \in \bigtimes_{i=1}^s \mathbb{R}^{m_i \times n_i}$,

$$R_i(X) := \max_{T \in \mathcal{U}(\mathbf{a}_i, \mathbf{b}_i)} \left\langle \nabla_{X_i} f(X), X_i - T \right\rangle, \quad i = 1, \dots, s. \quad (4.23)$$

此外, 记 $R := \sum_{i=1}^s R_i$. 不难验证, 对任何 $X \in \bigtimes_{i=1}^s \mathcal{U}(\mathbf{a}_i, \mathbf{b}_i)$, 有 $R(X) \geq 0$. 而且 $X \in \bigtimes_{i=1}^s \mathcal{U}(\mathbf{a}_i, \mathbf{b}_i)$ 是问题 (4.1) 的 KKT 点当且仅当 $R(X) = 0$. 因此, 对于由 ERALM 算法与 S-ERALM 算法产生的迭代点 $X^{(t)}$, $R(X^{(t)})$ 可用于度量在该处的稳定性违反度.

表 4.1 四个新算法单步计算代价的比较

Table 4.1 Comparisons on single-iteration computational costs among four new methods

单步迭代的主要计算组成	ERALM 算法	KLALM 算法
核矩阵计算	$\sum_{i=1}^s m_i n_i$ 个元素	$\sum_{i=1}^s m_i n_i$ 个元素
Sinkhorn 算法迭代	$l_{\max} \times \sum_{i=1}^s m_i n_i$	$l_{\max} \times \sum_{i=1}^s m_i n_i$
单步迭代的主要计算组成	S-ERALM 算法	S-KLALM 算法
矩阵逐元素采样	$\mathcal{O}\left(\sum_{i=1}^s m_i n_i\right)$	$\mathcal{O}\left(\sum_{i=1}^s m_i n_i\right) (t = \hat{t})$
核矩阵计算	$\sum_{i=1}^s \hat{n}_i$ 个元素	$\begin{cases} \sum_{i=1}^s m_i n_i & \text{个元素 } (t < \hat{t}) \\ \sum_{i=1}^s \hat{n}_i & \text{个元素 } (t \geq \hat{t}) \end{cases}$
Sinkhorn 算法迭代	$l_{\max} \times \sum_{i=1}^s \hat{n}_i$	$\begin{cases} l_{\max} \times \sum_{i=1}^s m_i n_i & (t < \hat{t}) \\ l_{\max} \times \sum_{i=1}^s \hat{n}_i & (t \geq \hat{t}) \end{cases}$

我们假设目标函数 f 在可行域上是分块 Lipschitz 光滑的. 这对许多应用是自动成立的(可见问题 (2.8)).

条件 4.1. 目标函数 f 在 $\bigtimes_{i=1}^s \mathcal{U}(\mathbf{a}_i, \mathbf{b}_i)$ 上是分块 Lipschitz 光滑的, 即存在 $L \geq 0$, 使得对 $i = 1, \dots, s$,

$$\left\| \nabla_{X_i} f(X) - \nabla_{X_i} f(X') \right\| \leq L \|X - X'\|, \quad \forall X, X' \in \bigtimes_{i=1}^s \mathcal{U}(\mathbf{a}_i, \mathbf{b}_i).$$

我们首先分析 ERALM 算法的理论性质. 从文献^[281]引理 1 的证明, 我们不难得到如下引理, 其刻画了由熵正则带来的最优值偏差.

引理 4.2 (熵正则带来的最优值偏差). 设 $W \in \mathbb{R}^{m \times n}$, $\mathbf{a} \in \mathbb{R}^m$, $\mathbf{b} \in \mathbb{R}^n$, $\lambda > 0$, $\mathcal{I} \subseteq \{(j, k) : j = 1, \dots, m, k = 1, \dots, n\}$. 假设 T' 和 $T'' \in \mathbb{R}^{m \times n}$ 分别是

$$\min_T \langle W, T \rangle, \quad \text{s. t. } T \in \mathcal{U}(\mathbf{a}, \mathbf{b}), \quad T_{\mathcal{I}^c} = 0$$

和

$$\min_T \langle W, T \rangle + \lambda h(T), \quad \text{s. t. } T \in \mathcal{U}(\mathbf{a}, \mathbf{b}), \quad T_{\mathcal{I}^c} = 0$$

的最优解. 则

$$0 \leq \langle W, T'' - T' \rangle \leq -\lambda h(\mathbf{a}\mathbf{b}^\top).$$

证明. 由文献^[281]引理 1 的证明以及对任意 $T \in \mathcal{U}(\mathbf{a}, \mathbf{b})$, $0 \geq h(T) \geq h(\mathbf{a}\mathbf{b}^\top)$ 即可得证. \square

我们再给出集合 $\mathcal{U}(\mathbf{a}_i, \mathbf{b}_i)$ 大小的一个上界 ($i = 1, \dots, s$).

引理 4.3 (集合 $\mathcal{U}(\mathbf{a}_i, \mathbf{b}_i)$ 的大小). 对 $i = 1, \dots, s$, 有如下不等式成立:

$$\|T - T'\| \leq 2d_i, \quad \forall T, T' \in \mathcal{U}(\mathbf{a}_i, \mathbf{b}_i),$$

其中

$$d_i := \min \left\{ \sqrt{m_i} \|\mathbf{a}_i\|_\infty, \sqrt{n_i} \|\mathbf{b}_i\|_\infty \right\}.$$

证明. 对任意 $T \in \mathcal{U}(\mathbf{a}_i, \mathbf{b}_i)$,

$$\begin{aligned} \|T\| &= \sqrt{\sum_{j=1}^{m_i} \sum_{t=1}^{n_i} t_{jk}^2} \leq \sqrt{\sum_{j=1}^{m_i} \left(\sum_{t=1}^{n_i} t_{jk} \right)^2} = \|T \mathbf{1}_{n_i}\| = \|\mathbf{a}_i\| \\ &= \sqrt{\sum_{j=1}^{m_i} a_{i,j}^2} \leq \sqrt{m_i \|\mathbf{a}_i\|_\infty^2} = \sqrt{m_i} \|\mathbf{a}_i\|_\infty. \end{aligned}$$

类似地, 我们也有 $\|T\| \leq \sqrt{n_i} \|\mathbf{b}_i\|_\infty$. 于是, $\|T\| \leq d_i$. 根据范数的三角不等式, 即可得证. \square

基于条件 4.1 和引理 4.3, 我们可证明残差函数 R_i 具有所谓的类 Lipschitz 连续性.

引理 4.4 (残差函数的类 Lipschitz 连续性). 设 $X', X'' \in \bigtimes_{i=1}^s \mathcal{U}(\mathbf{a}_i, \mathbf{b}_i)$. 假设条件 4.1 成立且对某个 $i \in \{1, \dots, s\}$, 有 $X'_i = X''_i$. 则

$$|R_i(X') - R_i(X'')| \leq 2d_i L \|X' - X''\|.$$

证明. 令 \bar{X}'_i 和 \bar{X}''_i 分别为

$$\min_{X_i \in \mathcal{U}(\mathbf{a}_i, \mathbf{b}_i)} \left\langle \nabla_{X_i} f(X'), X_i \right\rangle \quad \text{和} \quad \min_{X_i \in \mathcal{U}(\mathbf{a}_i, \mathbf{b}_i)} \left\langle \nabla_{X_i} f(X''), X_i \right\rangle$$

的最优解. 直接计算可得

$$\begin{aligned} R_i(X') &= \left\langle \nabla_{X_i} f(X'), X'_i - \bar{X}'_i \right\rangle \quad ((4.23) \text{ 式}) \\ &= \left\langle \nabla_{X_i} f(X''), X'_i - \bar{X}'_i \right\rangle + \left\langle \nabla_{X_i} f(X') - \nabla_{X_i} f(X''), X'_i - \bar{X}'_i \right\rangle \\ &\leq \left\langle \nabla_{X_i} f(X''), X'_i - \bar{X}'_i \right\rangle + 2d_i L \|X' - X''\| \quad (\text{条件 4.1 和引理 4.3}) \\ &= \left\langle \nabla_{X_i} f(X''), X''_i - \bar{X}''_i \right\rangle + 2d_i L \|X' - X''\| \quad (\text{条件 } X'_i = X''_i) \\ &\leq \left\langle \nabla_{X_i} f(X''), X''_i - \bar{X}''_i \right\rangle + 2d_i L \|X' - X''\| \quad (\bar{X}''_i \text{ 的定义}) \\ &= R_i(X'') + 2d_i L \|X' - X''\|. \quad ((4.23) \text{ 式}) \end{aligned}$$

类似地, 还可以证明

$$R_i(X'') \leq R_i(X') + 2d_i L \|X' - X''\|.$$

结合二者即可得证. \square

利用上面的工具, 我们给出 ERALM 算法的平均稳定性违反度上界.

定理 4.5 (ERALM 算法的平均稳定性违反度上界). 假设条件 4.1 成立. 令 $\{X^{(t)}\}$ 为 ERALM 算法产生的迭代点序列, 其中

$$t_{\max} \geq \frac{f(X^{(0)}) - \underline{f}}{2\bar{d}^2 s L (2\sqrt{s} + 1)}, \quad \alpha_{1,t} = \dots = \alpha_{s,t} \equiv \alpha := \frac{1}{\bar{d}} \sqrt{\frac{f(X^{(0)}) - \underline{f}}{2sL(2\sqrt{s} + 1)t_{\max}}}, \quad (4.24)$$

$\lambda_{1,t} = \dots = \lambda_{s,t} \equiv \lambda$ ($0 \leq t \leq t_{\max}$), 子问题 (4.5) 被精确求解, $\underline{f} \in \mathbb{R}$ 不大于问题 (4.1) 的最优值, $\bar{d} := \max_{i=1}^s d_i$. 则

$$0 \leq \frac{1}{t_{\max}} \sum_{t=0}^{t_{\max}-1} R(X^{(t)}) \leq 2\bar{d}(2s+1) \sqrt{\frac{L(f(X^{(0)}) - \underline{f})}{t_{\max}}} + s\lambda\bar{h}, \quad (4.25)$$

其中 $\bar{h} := -\min_{i=1}^s h(\mathbf{a}_i \mathbf{b}_i^\top)$.

证明. 对 $i = 1, \dots, s$ 和任意 $0 \leq t \leq t_{\max} - 1$,

$$\begin{aligned} f(X_{\leq i}^{(t+1)}, X_{>i}^{(t)}) &\leq f(X_{<i}^{(t+1)}, X_{\geq i}^{(t)}) + \left\langle C_i^{(t)}, X_i^{(t+1)} - X_i^{(t)} \right\rangle + \frac{L}{2} \|X_i^{(t+1)} - X_i^{(t)}\|^2 \\ &= f(X_{<i}^{(t+1)}, X_{\geq i}^{(t)}) + \alpha \left\langle C_i^{(t)}, \tilde{X}_i^{(t+1)} - X_i^{(t)} \right\rangle + \frac{\alpha^2 L}{2} \|\tilde{X}_i^{(t+1)} - X_i^{(t)}\|^2 \\ &\leq f(X_{<i}^{(t+1)}, X_{\geq i}^{(t)}) + \alpha \left\langle C_i^{(t)}, \tilde{X}_i^{(t+1)} - X_i^{(t)} \right\rangle + 2(d_i \alpha)^2 L \\ &\leq f(X_{<i}^{(t+1)}, X_{\geq i}^{(t)}) - \alpha \lambda h(\mathbf{a}_i \mathbf{b}_i^\top) - \alpha R_i(X_{<i}^{(t+1)}, X_{\geq i}^{(t)}) + 2(d_i \alpha)^2 L, \end{aligned}$$

其中第一个不等式使用了条件 4.1, 第二个不等式使用了引理 4.3, 最后一个不等式使用了引理 4.2 和 (4.23) 式. 我们可从上述关系式进一步得到

$$\alpha R_i(X_{<i}^{(t+1)}, X_{\geq i}^{(t)}) \leq f(X_{<i}^{(t+1)}, X_{\geq i}^{(t)}) - f(X_{\leq i}^{(t+1)}, X_{>i}^{(t)}) - \alpha \lambda h(\mathbf{a}_i \mathbf{b}_i^\top) + 2(d_i \alpha)^2 L. \quad (4.26)$$

注意到

$$\begin{aligned} &\left| R_i(X_{<i}^{(t+1)}, X_{\geq i}^{(t)}) - R_i(X^{(t)}) \right|^2 \\ &\leq 4(d_i L)^2 \|X^{(t+1)} - X^{(t)}\|^2 = 4(d_i L)^2 \sum_{i=1}^s \|X_i^{(t+1)} - X_i^{(t)}\|^2 \\ &= 4(d_i L \alpha)^2 \sum_{i=1}^s \|\tilde{X}_i^{(t+1)} - X_i^{(t)}\|^2 \leq 16\bar{d}^4 L^2 \alpha^2 s, \end{aligned}$$

其中第一个不等式使用了引理 4.4, 最后一个使用了引理 4.3. 结合上述不等式和 (4.26) 式, 我们有

$$\begin{aligned} \alpha R_i(X^{(t)}) &= \alpha \left[R_i(X_{<i}^{(t+1)}, X_{\geq i}^{(t)}) + \left(R_i(X^{(t)}) - R_i(X_{<i}^{(t+1)}, X_{\geq i}^{(t)}) \right) \right] \\ &\leq f(X_{<i}^{(t+1)}, X_{\geq i}^{(t)}) - f(X_{\leq i}^{(t+1)}, X_{>i}^{(t)}) - \alpha \lambda h(\mathbf{a}_i \mathbf{b}_i^\top) + 4\bar{d}^2 \alpha^2 L \sqrt{s} + 2(d_i \alpha)^2 L \end{aligned}$$

$$\leq f(X_{<i}^{(t+1)}, X_{\geq i}^{(t)}) - f(X_{\leq i}^{(t+1)}, X_{>i}^{(t)}) + \alpha \lambda \bar{h} + 2\bar{d}^2 \alpha^2 L (2\sqrt{s} + 1).$$

将上述不等式按指标 i 从 1 累加到 s , 再同时在两边除以 α , 可得

$$R(X^{(t)}) \leq \frac{f(X^{(t)}) - f(X^{(t+1)})}{\alpha} + s\lambda\bar{h} + 2\bar{d}^2 s L (2\sqrt{s} + 1)\alpha.$$

将上述不等式按指标 t 从 0 累加到 $t_{\max} - 1$, 再同时在两边除以 t_{\max} , 即有

$$\frac{1}{t_{\max}} \sum_{t=0}^{t_{\max}-1} R(X^{(t)}) \leq \frac{f(X^{(0)}) - f(X^{(t_{\max})})}{t_{\max}\alpha} + s\lambda\bar{h} + 2\bar{d}^2 s L (2\sqrt{s} + 1)\alpha.$$

注意到 $f(X^{(t_{\max})}) \geq \underline{f}$, α 的定义 (4.24) 式以及 $2s(2\sqrt{s} + 1) < (2s + 1)^2$ 即可得证. 顺带一提, 在 (4.24) 式中 t_{\max} 的下界是为了保证 $\alpha \leq 1$. \square

注. 在定理 4.5 以及之后的推论 4.7、定理 4.9、推论 4.10 中, 假设 $\alpha_{1,t} = \dots = \alpha_{s,t}$ 与 $\lambda_{1,t} = \dots = \lambda_{s,t}$ 是为了提升可读性. 这些理论结果可以方便地推广至 $\alpha_{i,t}$ 与 $\lambda_{i,t}$ 随 $i \in \{1, \dots, s\}$ 和 $0 \leq t \leq t_{\max}$ 变化的情形.

在定理 4.5 的基础上, 我们只需恰当选取参数和借助下面的条件 4.6, 即可证明当 $\sum_{i=1}^s (m_i + n_i)$ 趋于无穷大时, (4.25) 式右端项趋于 0. 这一渐进性质对于大规模的实际应用而言是尤其重要的.

条件 4.6. (1) 存在 $\underline{f} \in \mathbb{R}$ 使得对于任意 $\{m_i\}_{i=1}^s, \{n_i\}_{i=1}^s \subseteq \mathbb{N}$, 问题 (4.1) 的最优值都以 \underline{f} 为下界.

(2) 存在 $q > 0$ 使得对于任意 $\{m_i\}_{i=1}^s, \{n_i\}_{i=1}^s \subseteq \mathbb{N}$, $\mathbf{a}_i^\top \mathbf{1}_{m_i} = \mathbf{b}_i^\top \mathbf{1}_{n_i} = 1$, $\max_j a_{i,j} \leq q \cdot \min_j a_{i,j}$, $\max_k b_{i,k} \leq q \cdot \min_k b_{i,k}$.

(3) 存在 $\theta \geq 0$ 使得对于任意 $\{m_i\}_{i=1}^s, \{n_i\}_{i=1}^s \subseteq \mathbb{N}$, 块 Lipschitz 常数 $L = \mathcal{O}(\sum_{i=1}^s (m_i + n_i)^\theta)$.

(4) 存在 $\xi \geq 0$ 使得对于任意 $\{m_i\}_{i=1}^s, \{n_i\}_{i=1}^s \subseteq \mathbb{N}$, $\max_{i=1}^s (m_i + n_i) / \min_{i=1}^s (m_i + n_i) \leq \xi$.

注. 条件 4.6 中的 (1) 和 (2) 受实际应用启发. 例如, 在处理强关联电子体系时, 问题 (2.8) 的最优值具有天然的下界 $\underline{f} = 0$, 而边际 \mathbf{a}_i 和 \mathbf{b}_i 则是单电子密度的离散, 具有恒定的总质量. 特别地, 在归一化 f 后, 我们总可以假设边际的总质量为 1. (3) 和 (4) 则是为了理论分析的方便. 它们均可在一定程度上被减弱.

推论 4.7 (ERALM 算法的渐进性质). 假设条件 4.1 与 4.6 (1)–(3) 成立. 令 $\{X^{(t)}\}$ 为 ERALM 算法产生的迭代点序列, 其中

$$\begin{aligned} t_{\max} &\geq \max \left\{ \Omega \left(\sum_{i=1}^s (m_i + n_i)^\eta \right), \frac{f(X^{(0)}) - \underline{f}}{2\bar{d}^2 s L (2\sqrt{s} + 1)} \right\}, \quad f(X^{(0)}) \leq M, \\ \alpha_{1,t} &= \dots = \alpha_{s,t} \equiv \alpha, \quad \lambda_{1,t} = \dots = \lambda_{s,t} \equiv \lambda = o \left(\frac{1}{\sum_{i=1}^s \ln(m_i n_i)} \right) \end{aligned}$$

对于 $0 \leq t \leq t_{\max}$ 均成立, 子问题 (4.5) 被精确求解, $\eta (> \theta)$ 和 M 是与 $\{m_i\}_{i=1}^s$ 和 $\{n_i\}_{i=1}^s$ 无关的常数, α 定义在 (4.24) 式中. 则随着 $\sum_{i=1}^s (m_i + n_i)$ 趋于无穷大, $\sum_{t=0}^{t_{\max}-1} R(X^{(t)})/t_{\max}$ 趋于 0.

证明. 根据条件 4.6 (1)–(3), 不难推得

$$\begin{aligned}\bar{d} &= \max_{i=1}^s \min\{\sqrt{m_i} \|\mathbf{a}_i\|_\infty, \sqrt{n_i} \|\mathbf{b}_i\|_\infty\} = \Theta\left(\max_{i=1}^s \min\left\{\frac{1}{\sqrt{m_i}}, \frac{1}{\sqrt{n_i}}\right\}\right), \\ \bar{h} &= \max_{i=1}^s \sum_{j,k} a_{i,j} b_{i,k} (1 - \ln a_{i,j} b_{i,k}) = \Theta\left(\sum_{i=1}^s \ln(m_i n_i)\right).\end{aligned}$$

基于上述, 当 $t_{\max} = \Omega(\sum_{i=1}^s (m_i + n_i)^\eta)$ 且 $\eta > \theta$ 和 M 与 $\{m_i\}_{i=1}^s$ 和 $\{n_i\}_{i=1}^s$ 无关时, 不等式 (4.25) 右端第一项随着 $\sum_{i=1}^s (m_i + n_i)$ 趋于无穷大而趋于 0, 而当 $\lambda = o(1/\sum_{i=1}^s \ln(m_i n_i))$ 时, 第二项也趋于 0. 证毕. \square

由于加入了随机采样, S-ERALM 算法的分析要比 ERALM 算法复杂得多. 我们需要借助矩阵逐元素随机近似理论和如下假设条件.

条件 4.8. (1) 存在常数 $\nu \in (1/2, 1]$, $c_1, c_2, \hat{c}_2 > 0$, 使得对任意 $0 \leq t \leq t_{\max}$ 和 $i = 1, \dots, s$,

$$\left\| \Psi_i^{(t)} \right\|_2 \geq \frac{(m_i + n_i)^\nu}{c_1}, \quad \kappa(\Psi_i^{(t)}) \leq c_2, \quad \kappa(\hat{\Psi}_i^{(t)}) \leq \hat{c}_2;$$

(2) 插值因子 γ 小于 1, 且存在 $\varepsilon > 0$, 使得对 $i = 1, \dots, s$,

$$\frac{1}{\max_{j,k,t} p_{i,jk}^{(t)}} \geq \hat{n}_i \geq \frac{8(m_i + n_i)^{1-2\nu} \ln^4(m_i + n_i)}{(1-\gamma)w_i \ln^4(1+\varepsilon)},$$

其中

$$w_i := \min_{j,k} \frac{\sqrt{a_{i,j} b_{i,k}}}{\sum_{j',k'} \sqrt{a_{i,j'} b_{i,k'}}}.$$

注. 条件 4.8 的存在是为了理论分析的方便. 特别地, 因为 $w_i \leq 1/(m_i n_i)$, 从 (2) 我们可推出 \hat{n}_i 的如下下界:

$$\frac{8(m_i + n_i)^{1-2\nu} \ln^4(m_i + n_i)}{(1-\gamma)w_i \ln^4(1+\varepsilon)} \geq \frac{8}{(1-\gamma) \ln^4(1+\varepsilon)} \frac{m_i n_i}{(m_i + n_i)^{2\nu-1}} \ln^4(m_i + n_i).$$

因为 $\nu \in (1/2, 1]$, 该下界的阶要严格小于 $m_i n_i$ 的阶. 条件 $\hat{n}_i p_{i,jk}^{(t)} \leq 1$ 则常见于使用 Poisson 采样的文献^[273,282].

在条件 4.6 (1) 之下, 边际 \mathbf{a}_i 和 \mathbf{b}_i 中各元素的阶相同. 此时, 在选取合适的 γ 后, $w_i = \Theta(1/(m_i n_i))$, $p_{i,jk}^{(t)} = \Theta(1/(m_i n_i))$. 若再有 ε 和 γ 与 $\{m_i\}_{i=1}^s$ 和 $\{n_i\}_{i=1}^s$ 无关, 我们就只需选取 $\hat{n}_i = \Theta(m_i n_i \ln^4(m_i + n_i)/(m_i + n_i)^{2\nu-1})$ 即可使条件 4.8 (2) 成立.

我们再定义辅助序列

$$\begin{aligned}\bar{X}_i^{(t+1)} &= \arg \min_{X_i} \left\langle C_i^{(t)}, X_i \right\rangle, \text{ s. t. } X_i \in \mathcal{U}(\mathbf{a}_i, \mathbf{b}_i), \\ \check{X}_i^{(t+1)} &= \arg \min_{X_i} \left\langle C_i^{(t)}, X_i \right\rangle + \lambda_{i,t} h(X_i), \text{ s. t. } X_i \in \mathcal{U}(\mathbf{a}_i, \mathbf{b}_i),\end{aligned}$$

其中 $C_i^{(t)}$ 的定义可见 (4.4) 式 ($i = 1, \dots, s, 0 \leq t \leq t_{\max} - 1$).

定理 4.9 (S-ERALM 算法的平均稳定性违反度上界). 假设条件 4.1 成立. 令 $\{X^{(t)}\}$ 为 S-ERALM 算法产生的迭代点序列, 其中 $\alpha_{1,t} = \dots = \alpha_{s,t} \equiv \alpha$, $\lambda_{1,t} = \dots = \lambda_{s,t} \equiv \hat{\lambda}$ ($0 \leq t \leq t_{\max}$), 子问题 (4.13) 可行且被精确求解, 条件 4.8 成立, α 定义在 (4.24) 式中. 则对任意 $\zeta > 0$ 和 $\iota > 0$, 只要 $m_i + n_i > \max\{152, e^{\sqrt{c_3}}\}$ ($i = 1, \dots, s$),

$$\begin{aligned}0 \leq \frac{1}{t_{\max}} \sum_{t=0}^{t_{\max}-1} R(X^{(t)}) &\leq 2\bar{d}(2s+1) \sqrt{\frac{L(f(X^{(0)}) - f)}{t_{\max}}} + 2s\hat{\lambda}\bar{h} \quad (4.27) \\ &+ \hat{\lambda}\bar{d} \sum_{i=1}^s \sqrt{\hat{n}_i + \iota \cdot m_i n_i} \ln \frac{1}{(1-\gamma)w_i \hat{n}_i} + \hat{\lambda} \sum_{i=1}^s \frac{\hat{c}_2 c_3}{\ln^2(m_i + n_i) - c_3}\end{aligned}$$

成立的概率就不小于

$$\prod_{i=1}^s \left\{ \left[1 - 2 \exp \left(-\frac{16\zeta^2}{\varepsilon^4} \ln^4(m_i + n_i) \right) \right] [1 - \exp(-2\iota^2 m_i n_i)] \right\}^{t_{\max}},$$

其中 $c_3 := c_1(1 + \varepsilon + \zeta) \ln^2(1 + \varepsilon)$.

证明. 对 $i = 1, \dots, s$ 和任意 $0 \leq t \leq t_{\max} - 1$,

$$\begin{aligned}&\left(f(X_{\leq i}^{(t+1)}, X_{>i}^{(t)}) - f(X_{<i}^{(t+1)}, X_{\geq i}^{(t)}) \right) / \alpha - 2d_i^2 L \alpha \quad (4.28) \\ &\leq \left(f(X_{\leq i}^{(t+1)}, X_{>i}^{(t)}) - f(X_{<i}^{(t+1)}, X_{\geq i}^{(t)}) \right) / \alpha - L \alpha \left\| \tilde{X}_i^{(t+1)} - X_i^{(t)} \right\|^2 / 2 \\ &= \left(f(X_{\leq i}^{(t+1)}, X_{>i}^{(t)}) - f(X_{<i}^{(t+1)}, X_{\geq i}^{(t)}) - L \left\| X_i^{(t+1)} - X_i^{(t)} \right\|^2 / 2 \right) / \alpha \\ &\leq \left\langle C_i^{(t)}, X_i^{(t+1)} - X_i^{(t)} \right\rangle / \alpha = \left\langle C_i^{(t)}, \tilde{X}_i^{(t+1)} - X_i^{(t)} \right\rangle \\ &= \left\langle \hat{C}_i^{(t)}, \tilde{X}_i^{(t+1)} \right\rangle + \left\langle C_i^{(t)} - \hat{C}_i^{(t)}, \tilde{X}_i^{(t+1)} \right\rangle - \left\langle C_i^{(t)}, X_i^{(t)} \right\rangle \\ &= \underbrace{\left\langle \hat{C}_i^{(t)}, \tilde{X}_i^{(t+1)} \right\rangle}_{I_1} - \underbrace{\left\langle C_i^{(t)}, \tilde{X}_i^{(t+1)} \right\rangle}_{I_2} + \underbrace{\left\langle C_i^{(t)} - \hat{C}_i^{(t)}, \tilde{X}_i^{(t+1)} \right\rangle}_{I_3} - R_i(X_{<i}^{(t+1)}, X_{\geq i}^{(t)}),\end{aligned}$$

其中第一个不等式使用了引理 4.3, 第二个不等式使用了条件 4.1, 最后一个等式使用了残差函数的定义 (4.23).

下面, 我们来推导 I_1 和 I_2 的上界. 首先, 对于 I_2 ,

$$\left\langle C_i^{(t)} - \hat{C}_i^{(t)}, \tilde{X}_i^{(t+1)} \right\rangle \leq d_i \left\| \hat{C}_i^{(t)} - C_i^{(t)} \right\| = d_i \hat{\lambda} \sqrt{\sum_{j,k:(j,k) \in \mathcal{I}_i^{(t)}} \ln^2(\hat{n}_i \cdot p_{i,jk}^{(t)})}$$

$$\leq d_i \hat{\lambda} \sqrt{|\mathcal{I}_i^{(t)}| \ln^2 \frac{1}{(1-\gamma)w_i \hat{n}_i}} = d_i \hat{\lambda} \sqrt{|\mathcal{I}_i^{(t)}|} \ln \frac{1}{(1-\gamma)w_i \hat{n}_i}.$$

其中第一个不等式使用了引理 4.3 的证明, 第二个不等式使用了 (4.10) 式和条件 4.8 (2), 第一个等式使用了 (4.14) 式和条件 4.8 (2). 根据 Hoeffding 不等式, 对任意 $\iota > 0$,

$$\text{Prob}\left(\left|\mathcal{I}_i^{(t)}\right| \geq \hat{n}_i + \iota \cdot m_i n_i\right) \leq \exp(-2\iota^2 m_i n_i).$$

因此, 以不小于 $1 - \exp(-2\iota^2 m_i n_i)$ 的概率, 我们有

$$\left\langle C_i^{(t)} - \hat{C}_i^{(t)}, \tilde{X}_i^{(t+1)} \right\rangle \leq d_i \hat{\lambda} \sqrt{\hat{n}_i + \iota \cdot m_i n_i} \ln \frac{1}{(1-\gamma)w_i \hat{n}_i}. \quad (4.29)$$

对于 I_1 ,

$$\begin{aligned} & \left\langle \hat{C}_i^{(t)}, \tilde{X}_i^{(t+1)} \right\rangle - \left\langle C_i^{(t)}, \tilde{X}_i^{(t+1)} \right\rangle \\ & \leq \left\langle \hat{C}_i^{(t)}, \tilde{X}_i^{(t+1)} \right\rangle - \left\langle C_i^{(t)}, \check{X}_i^{(t+1)} \right\rangle - \hat{\lambda} h(\mathbf{a}_i \mathbf{b}_i^\top) \\ & = \left[\left\langle \hat{C}_i^{(t)}, \tilde{X}_i^{(t+1)} \right\rangle + \hat{\lambda} h(\tilde{X}_i^{(t+1)}) \right] - \left[\left\langle C_i^{(t)}, \check{X}_i^{(t+1)} \right\rangle + \hat{\lambda} h(\check{X}_i^{(t+1)}) \right] \\ & \quad + \hat{\lambda} \left[h(\check{X}_i^{(t+1)}) - h(\tilde{X}_i^{(t+1)}) \right] - \hat{\lambda} h(\mathbf{a}_i \mathbf{b}_i^\top) \\ & \leq q_i(\tilde{\mathbf{u}}_i^{(t,\star)}, \tilde{\mathbf{v}}_i^{(t,\star)}; \hat{\lambda}, \hat{\Psi}_i^{(t)}) - q_i(\tilde{\mathbf{u}}_i^{(t,\star)}, \tilde{\mathbf{v}}_i^{(t,\star)}; \hat{\lambda}, \Psi_i^{(t)}) - 2\hat{\lambda} h(\mathbf{a}_i \mathbf{b}_i^\top) \\ & \leq \hat{c}_2 \hat{\lambda} \frac{\left\| \hat{\Psi}_i^{(t)} - \Psi_i^{(t)} \right\|_2}{\left\| \Psi_i^{(t)} \right\|_2} \left| 1 - \frac{\left\| \hat{\Psi}_i^{(t)} - \Psi_i^{(t)} \right\|_2}{\left\| \Psi_i^{(t)} \right\|_2} \right|^{-1} - 2\hat{\lambda} h(\mathbf{a}_i \mathbf{b}_i^\top), \end{aligned} \quad (4.30)$$

其中第一个不等式使用了引理 4.2 (令 $\mathcal{I} = \mathcal{I}_i^{(t)}$), 第二个不等式使用了 $h(\check{X}_i^{(t+1)}) \leq 0$, $h(\tilde{X}_i^{(t+1)}) \geq h(\mathbf{a}_i \mathbf{b}_i^\top)$, q_i 的定义 (4.6) 以及 $(\tilde{\mathbf{u}}_i^{(t,\star)}, \tilde{\mathbf{v}}_i^{(t,\star)})$ 的最优性, 最后一个不等式使用了文献 [266] 中的结论. 为推导 (4.30) 式右端第一项的上界, 我们借助文献 [267] 定理 3.1. 具体来讲, 对任意 $j \in \{1, \dots, m_i\}$ 和 $k \in \{1, \dots, n_i\}$, 根据 (4.10) 式和条件 4.8, 我们有

$$\begin{aligned} \mathbb{E} \left(\frac{\hat{\psi}_{i,jk}^{(t)}}{\left\| \Psi_i^{(t)} \right\|_2} \right) &= \hat{n}_i \cdot p_{i,jk}^{(t)} \cdot \frac{\psi_{i,jk}^{(t)}}{\hat{n}_i \cdot p_{i,jk}^{(t)}} \cdot \frac{1}{\left\| \Psi_i^{(t)} \right\|_2} = \frac{\psi_{i,jk}^{(t)}}{\left\| \Psi_i^{(t)} \right\|_2}, \\ \text{Var} \left(\frac{\hat{\psi}_{i,jk}^{(t)}}{\left\| \Psi_i^{(t)} \right\|_2} \right) &< \frac{(\psi_{i,jk}^{(t)})^2}{\hat{n}_i \cdot p_{i,jk}^{(t)} \left\| \Psi_i^{(t)} \right\|_2^2} \leq \frac{1}{(1-\gamma)w_i \hat{n}_i \left\| \Psi_i^{(t)} \right\|_2^2} \\ &\leq \frac{c_1^2 (m_i + n_i)^{-2\nu}}{(1-\gamma)w_i \hat{n}_i} \leq \frac{c_1^2 \ln^4(1+\varepsilon)}{8(m_i + n_i) \ln^4(m_i + n_i)}. \end{aligned}$$

此外, $\hat{\psi}_{i,jk}^{(t)} / \left\| \Psi_i^{(t)} \right\|_2$ 有上界

$$\frac{\psi_{i,jk}^{(t)}}{\hat{n}_i \cdot p_{i,jk}^{(t)} \left\| \Psi_i^{(t)} \right\|_2} \leq \frac{1}{(1-\gamma)w_i \hat{n}_i \left\| \Psi_i^{(t)} \right\|_2} \leq \frac{c_1 (m_i + n_i)^{-\nu}}{(1-\gamma)w_i \hat{n}_i}$$

$$\leq \frac{c_1 \ln^4(1 + \varepsilon)}{8(m_i + n_i)^{1-\nu} \ln^4(m_i + n_i)}.$$

于是, 根据文献^[267] 定理 3.1, 对任取 $\zeta > 0$, 只要 $m_i + n_i \geq 152$,

$$\text{Prob}\left(\frac{\|\hat{\Psi}_i^{(t)} - \Psi_i^{(t)}\|_2}{\|\Psi_i^{(t)}\|_2} \geq c_1(1 + \varepsilon + \zeta) \frac{\ln^2(1 + \varepsilon)}{\ln^2(m_i + n_i)}\right) < 2 \exp\left(-\frac{16\zeta^2}{\varepsilon^4} \ln^4(m_i + n_i)\right).$$

这等价于

$$\frac{\|\hat{\Psi}_i^{(t)} - \Psi_i^{(t)}\|_2}{\|\Psi_i^{(t)}\|_2} < c_1(1 + \varepsilon + \zeta) \frac{\ln^2(1 + \varepsilon)}{\ln^2(m_i + n_i)},$$

进而

$$\frac{\|\hat{\Psi}_i^{(t)} - \Psi_i^{(t)}\|_2}{\|\Psi_i^{(t)}\|_2} \left|1 - \frac{\|\hat{\Psi}_i^{(t)} - \Psi_i^{(t)}\|_2}{\|\Psi_i^{(t)}\|_2}\right|^{-1} < \frac{c_1(1 + \varepsilon + \zeta) \ln^2(1 + \varepsilon)}{\ln^2(m_i + n_i) - c_1(1 + \varepsilon + \zeta) \ln^2(1 + \varepsilon)} \quad (4.31)$$

成立的概率不小于 $1 - 2 \exp(-16\zeta^2 \ln^4(m_i + n_i)/\varepsilon^4)$.

结合 (4.28)–(4.31) 式, 以不小于

$$\left[1 - 2 \exp\left(-\frac{16\zeta^2}{\varepsilon^4} \ln^4(m_i + n_i)\right)\right] [1 - \exp(-2\lambda^2 m_i n_i)]$$

的概率, 有下面的关系式成立:

$$\begin{aligned} f(X_{\leq i}^{(t+1)}, X_{>i}^{(t)}) &\leq f(X_{<i}^{(t+1)}, X_{\geq i}^{(t)}) - \alpha R_i(X_{<i}^{(t+1)}, X_{\geq i}^{(t)}) + 2(d_i \alpha)^2 L \\ &\quad - 2\alpha \hat{\lambda} h(\mathbf{a}_i \mathbf{b}_i^\top) + d_i \alpha \hat{\lambda} \sqrt{\hat{n}_i + \lambda \cdot m_i n_i} \ln \frac{1}{(1 - \gamma) w_i \hat{n}_i} \\ &\quad + \alpha \hat{\lambda} \frac{c_1 \hat{c}_2 (1 + \varepsilon + \zeta) \ln^2(1 + \varepsilon)}{\ln^2(m_i + n_i) - c_1 (1 + \varepsilon + \zeta) \ln^2(1 + \varepsilon)}, \end{aligned}$$

进而

$$\begin{aligned} \alpha R_i(X_{<i}^{(t+1)}, X_{\geq i}^{(t)}) &\leq f(X_{<i}^{(t+1)}, X_{\geq i}^{(t)}) - f(X_{\leq i}^{(t+1)}, X_{>i}^{(t)}) + 2(d_i \alpha)^2 L \\ &\quad - 2\alpha \hat{\lambda} h(\mathbf{a}_i \mathbf{b}_i^\top) + d_i \alpha \hat{\lambda} \sqrt{\hat{n}_i + \lambda \cdot m_i n_i} \ln \frac{1}{(1 - \gamma) w_i \hat{n}_i} \\ &\quad + \alpha \hat{\lambda} \frac{c_1 \hat{c}_2 (1 + \varepsilon + \zeta) \ln^2(1 + \varepsilon)}{\ln^2(m_i + n_i) - c_1 (1 + \varepsilon + \zeta) \ln^2(1 + \varepsilon)}. \end{aligned}$$

之后, 再经过与定理 4.5 证明中类似的推导, 并结合 α 与 c_3 的定义, 即可得证. \square

推论 4.10 (S-ERALM 算法的渐进性质). 假设条件 4.1 与 4.6 成立. 令 $\{X^{(t)}\}$ 为 S-ERALM 算法产生的迭代点序列, 其中

$$\begin{aligned} t_{\max} &= \Theta\left(\sum_{i=1}^s (m_i + n_i)^\eta\right) \text{ 满足 } t_{\max} \geq \frac{f(X^{(0)}) - f}{2\bar{d}^2 s L (2\sqrt{s} + 1)}, \quad f(X^{(0)}) \leq M, \\ \alpha_{1,t} = \dots = \alpha_{s,t} &\equiv \alpha, \quad \lambda_{1,t} = \dots = \lambda_{s,t} \equiv \hat{\lambda} = o\left(\frac{1}{\sum_{i=1}^s \sqrt{m_i n_i} \ln(m_i + n_i)}\right), \\ \hat{n}_i &= \Theta\left(\frac{m_i n_i}{(m_i + n_i)^{2\nu-1}} \ln^4(m_i + n_i)\right) \end{aligned}$$

对于 $0 \leq t \leq t_{\max}$ 均成立, 子问题 (4.13) 可行且被精确求解, 条件 4.8 成立, c_1 、 c_2 、 \hat{c}_2 、 ε 、 $\eta (> \theta)$ 、 γ 、 ν 、 ξ 和 M 是与 $\{m_i\}_{i=1}^s$ 和 $\{n_i\}_{i=1}^s$ 无关的常数, α 定义在 (4.24) 式中. 则随着 $\sum_{i=1}^s (m_i + n_i)$ 趋于无穷大, $\sum_{t=0}^{t_{\max}-1} R(X^{(t)})/t_{\max}$ 以趋于 1 的概率趋于 0.

证明. 根据条件 4.6, 我们可推出 \bar{d} 、 \bar{h} 的界 (可见推论 4.7 的证明) 以及 $w_i = \Theta(1/(m_i n_i))$ ($i = 1, \dots, s$). 基于此, 当 $t_{\max} = \Omega(\sum_{i=1}^s (m_i + n_i)^\eta)$ 且 $\eta > \theta$ 和 M 与 $\{m_i\}_{i=1}^s$ 和 $\{n_i\}_{i=1}^s$ 无关时, 不等式 (4.27) 右端第一项随 $\sum_{i=1}^s (m_i + n_i)$ 趋于无穷大而趋于 0. 由 \hat{n}_i 的选取方式, 对任意 $\iota > 0$, 因为

$$\begin{aligned} &\sum_{i=1}^s \sqrt{\hat{n}_i + \iota \cdot m_i n_i} \ln \frac{1}{(1-\gamma)w_i \hat{n}_i} \\ &= \mathcal{O}\left(\sum_{i=1}^s \left[\frac{\sqrt{m_i n_i}}{(m_i + n_i)^{\nu-1/2}} \ln^2(m_i + n_i) + \sqrt{m_i n_i} \right] \ln \frac{(m_i + n_i)^{2\nu-1}}{\ln^4(m_i + n_i)}\right), \end{aligned}$$

所以当 $\hat{\lambda} = o(1/(\sum_{i=1}^s \sqrt{m_i n_i} \ln(m_i + n_i)))$ 且 ε 、 ν 和 γ 与 $\{m_i\}_{i=1}^s$ 和 $\{n_i\}_{i=1}^s$ 无关时, 不等式 (4.27) 右端第二和第三项趋于 0. 因为 c_1 、 c_2 和 \hat{c}_2 与 $\{m_i\}_{i=1}^s$ 和 $\{n_i\}_{i=1}^s$ 无关, 所以不等式 (4.27) 右端最后一项也趋于 0.

最后, 因为

$$\begin{aligned} &\prod_{i=1}^s \left\{ \left[1 - 2 \exp\left(-\frac{16\zeta^2}{\varepsilon^4} \ln^4(m_i + n_i)\right) \right] [1 - \exp(-2\iota^2 m_i n_i)] \right\}^{t_{\max}} \\ &\geq \prod_{i=1}^s \left\{ \left[1 - \frac{2}{(m_i + n_i)^{16\zeta^2/\varepsilon^4}} \right] [1 - \exp(-2\iota^2 m_i n_i)] \right\}^{t_{\max}}, \end{aligned}$$

所以根据条件 4.6 (4), 只要 $\zeta > \sqrt{\eta\varepsilon^2}/4$, 概率就随 $\sum_{i=1}^s (m_i + n_i)$ 趋于无穷大而趋于 1. 证毕. \square

4.4 数值实验

本节通过强关联电子体系的数值模拟展示算法的有效性与优势. 我们首先通过一维强关联电子体系的模拟, 对比本章的四个新算法以及 PALM-I 算法的

表 4.2 待模拟的一维、二维和三维强关联电子体系. 第二列为未归一化的单电子密度 ρ , 第三列为截断区域 Ω , 第四列为体系所含电子数 N

Table 4.2 One/two/three-dimensional systems used for simulations. The second column lists the unnormalized single-particle densities ρ , the third gives the truncated domains Ω , and the last indicates the numbers of electrons N in systems

体系编号	单电子密度 ρ	截断区域 Ω	电子数 N
一维体系			
1	$\cos(\pi r) + 1$	$[-1, 1]$	3
2	$2\rho_6(r; -0.5) + 1.5\rho_4(r; 0.5)$	$[-1.5, 1.5]$	3
3	$\rho_{1/\sqrt{\pi}}(r; 0)$	$[-2, 2]$	7
4	$\rho_4(r; -2) + \rho_4(r; -1.5) + \rho_4(r; -1) + \rho_4(r; -0.5) + \rho_4(r; 2/3) + \rho_4(r; 4/3) + \rho_4(r; 2)$	$[-3, 3]$	7
二维体系			
5	$\rho_3(\mathbf{r}; [0, 0.96]^\top) + \rho_3(\mathbf{r}; [1.032, -0.84]^\top) + \rho_3(\mathbf{r}; [-1.032, -0.84]^\top)$	$[-3, 3]^2$	3
6	$2\rho_3(\mathbf{r}; [0, 1.2]^\top) + \rho_3(\mathbf{r}; [1.29, -1.05]^\top) + \rho_3(\mathbf{r}; [-1.29, -1.05]^\top)$	$[-3, 3]^2$	4
三维体系			
7	$\rho_3(\mathbf{r}; [-1, -1, -1]^\top) + \rho_3(\mathbf{r}; [1, 1, -1]^\top) + \rho_3(\mathbf{r}; [-1, 1, 1]^\top)$	$[-2, 2]^3$	3
8	$3\rho_4(\mathbf{r}; [-1, 0, 0]^\top) + \rho_4(\mathbf{r}; [1, 0, 0]^\top)$	$[-2, 2] \times [-1, 1]^2$	4

数值表现. 我们使用所设计的四个新算法求解经过变量替换 $X_i := \text{Diag}(\boldsymbol{\rho})Y_i$ ($i = 2, \dots, N$) 后的问题 (2.8):

$$\begin{aligned} \min_{\{X_i\}_{i=2}^N} \quad & \sum_{i=2}^N \langle X_i, C + \beta \text{Diag}(\boldsymbol{\rho})^{-1} \rangle + \sum_{2 \leq i < j}^N \langle X_i, \text{Diag}(\boldsymbol{\rho})^{-1} X_j C + \beta \text{Diag}(\boldsymbol{\rho})^{-2} X_j \rangle, \\ \text{s. t.} \quad & X_i \in \mathcal{U}(\boldsymbol{\rho}, \boldsymbol{\rho}) \subseteq \mathbb{R}^{K \times K}, \quad i = 2, \dots, N. \end{aligned}$$

之后, 我们将数值表现较好的新算法作为 CMGOPT 框架 (见框架 3.3) 中的优化算法, 模拟二维、三维强关联电子体系. 最后, 我们测试新算法对离散规模与电子个数 (变量块数) 的标度.

所有的数值实验均在包含两颗 Intel Xeon Gold 6242R CPU (3.10 GHz $\times 20 \times 2$) 的工作站上实现. 工作站的运行内存为 510 GB, 操作系统为 Ubuntu 20.04.5. 我们使用 MATLAB R2019b 实现所有的算法.

4.4.1 待模拟强关联电子体系

我们考虑八个一维、二维和三维强关联电子体系, 在表 4.2 中列出它们的单电子密度、截断区域和电子数. 其中函数 $\rho_\alpha(\cdot; \mathbf{c})$ 的定义可见 (3.31) 式. 我们还在图 4.1 中示出表 4.2 里列出的单电子密度.

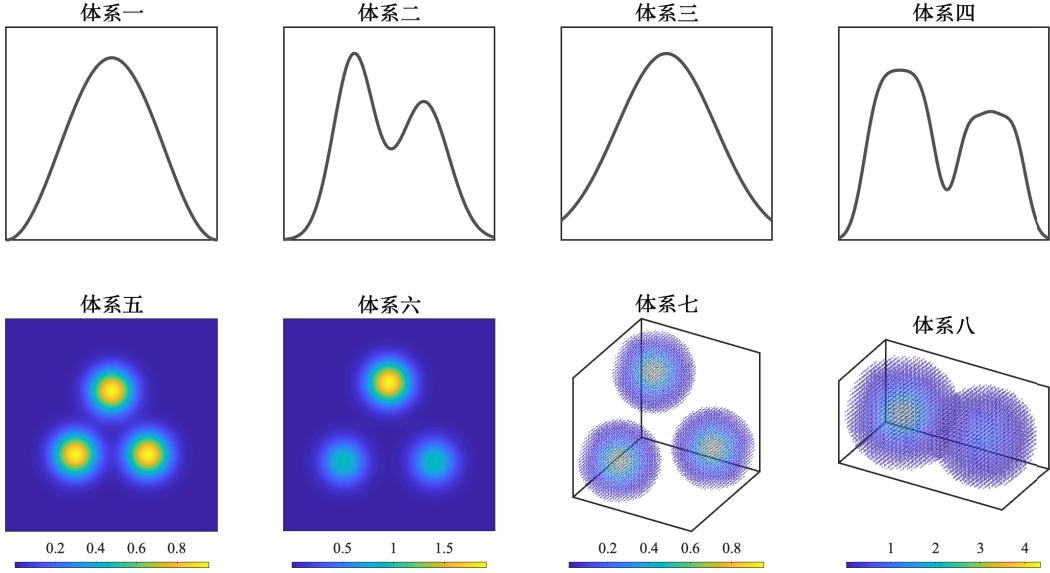


图 4.1 表 4.2 所列单电子密度的可视化. 对于三维体系, 我们仅展示单电子密度大于 0.01 的区域

Figure 4.1 The illustrations of the single-particle densities listed in Table 4.2. For the three-dimensional systems, we only show the regions where the values of the single-particle densities are larger than 0.01

4.4.2 实验设置

根据第 3 章中的数值结果, 我们固定罚参数 $\beta = 1$. 在 ERALM 算法与 S-ERALM 算法中, 我们采用下降型步长 $\alpha_{i,t} = 1/(t+1)^{0.75}$ ($i = 2, \dots, N$). 在 S-ERALM 算法与 S-KLALM 算法中, 我们设置 $\hat{n}_i = \lfloor K^{1.5} \rfloor$ ($i = 2, \dots, N$). 本章所设计的四个新算法按如下方式自适应调整正则化参数或邻近参数:

$$\lambda_{i,t} = \sigma \|\tilde{\mathbf{v}}_i^{(t)}\|_\infty / (20 \ln(K)), \quad \mu_{i,t} = \sigma \|\mathbf{v}_i^{(t)}\|_\infty / (20 \ln(K)), \quad i = 2, \dots, N, \quad (4.32)$$

其中 σ 默认取值为 1. 我们将 γ 和 \hat{t} 的选取放在后文第 4.4.3 节中. 需要指明的是, 若上述参数经过更加仔细的选取, 所设计的算法将具有更好的数值表现. 本章所设计的算法均使用 Sinkhorn 算法 (4.8) 求解子问题, 并辅以热启动. 为避免 Sinkhorn 算法在运行过程中遭遇严重数值下溢以及为加速其收敛^[283], 我们去除 $\boldsymbol{\rho}$ 中小于 $0.1\% \|\boldsymbol{\rho}\|_\infty$ 的分量. 这样的处理方式也是合理的, 因为区域的质量越小, 重要性就越低. 在后文中, 为记号方便, 我们仍然用 $\boldsymbol{\rho}$ 表示截断后的向量.

我们在 $\left\| X_i^{(t+1)} \mathbf{1}_K - \boldsymbol{\rho} \right\|_\infty$ 小于 10^{-6} 或迭代次数超过 $t_{\max} = 20$ 时终止 Sinkhorn 算法 ($i = 2, \dots, N$), 在

$$\Delta^{(t)} := \frac{1}{N-1} \sum_{i=2}^N \left\| \text{Diag}(\boldsymbol{\rho})^{-1} (X_i^{(t)} - X_i^{(t-1)}) \right\| \quad (4.33)$$

小于给定阈值 $tol > 0$ 或迭代次数超过 t_{\max} 时终止外层算法. 我们会在后文给出 tol 和 t_{\max} 的具体取值.

我们所关注的指标有如下五个: (1) 算法输出的目标函数值 (f_{out}); (2) 电子位置间的近似映射 $\{\hat{\mathcal{T}}_i\}_{i=2}^N : \mathbb{R}^d \rightarrow \mathbb{R}^d$, 类似于 (3.32) 式, 定义如下:

$$\hat{\mathcal{T}}_i(\mathbf{d}_j) := \sum_{1 \leq k \leq K} \mathbf{d}_k x_{i,jk} / \rho_j, \quad j = 1, \dots, K, \quad i = 2, \dots, N;$$

(3) 向量 $\hat{\lambda} \in \mathbb{R}^K$ 作为 SCE 势的近似, 定义类似 (3.33) 式. 此时, $\lambda_{i,\rho}$ 为 Sinkhorn 算法输出的对应于约束 $X_i^\top \mathbf{1}_K = \rho$ 的对偶变量 ($i = 2, \dots, N$); (4) 目标函数值误差 $\text{err_obj} \geq 0$ 和 SCE 势误差 $\text{err_sce} \geq 0$, 其在 Monge 拟设正确时 (例如对一维体系^[100]) 按如下公式计算:

$$\text{err_obj} := \left| \frac{f_{\text{out}} - f^*}{f^*} \right|, \quad \text{err_sce} := \frac{\|\hat{\lambda} - \lambda^*\|_\infty}{\|\lambda^*\|_\infty}.$$

这里, $f^* \in \mathbb{R}$ 为问题 (2.8) 的最优值, $\lambda^* \in \mathbb{R}^K$ 表示由 SCE 势在网格重心的取值构成的向量; (5) 算法运行 CPU 时间 $T > 0$ (秒).

4.4.3 算法比较

我们在表 4.2 中的一维体系上测试并比较 PALM-I 算法和本章的四个算法. 具体来讲, 我们首先测试使用不同采样概率的 S-ERALM 算法和使用不同 \hat{t} 的 S-KLALM 算法. 基于这些测试的结果, 我们为 γ 和 \hat{t} 选取默认取值. 之后, 我们比较本章中的四个新算法. 最后, 我们将其中数值表现较好者与 PALM-I 算法对比, 并用于后文在二维、三维体系上的模拟.

4.4.3.1 使用不同采样概率的 S-ERALM 算法的对比

我们考虑随机生成的采样概率和基于重要性采样的概率 (可见 (4.10) 式), 其中 $\gamma \in \{0.1, 0.3, 0.5, 0.7, 0.9, 0.99, 0.999\}$. 我们在体系一上进行测试, 使用等质量剖分生成网格 ($K = 90$). 我们调用 MATLAB 中的 “rand” 函数生成 10 个随机初始点 (后文同不再赘述). 终止参数为 $tol = 5 \times 10^{-3}$ 和 $t_{\max} = \infty$. 我们将不同采样概率下 S-ERALM 算法的平均 err_obj 、 err_sce 和 T 汇总在表 4.3 中.

尽管缺乏理论支撑, 但相较于随机生成的采样概率, 基于重要性采样的概率确实可以让 S-ERALM 算法在更短的 CPU 时间内找到更好的解. 此外, 我们还发现增大 γ 的取值可以让 S-ERALM 算法更早地满足终止准则, 但在超过某个阈值后输出解的质量便开始下滑. 为平衡解质量和运行效率, 我们在后续的实验中选取 S-ERALM 算法和 S-KLALM 算法中的 γ 为 0.99.

4.4.3.2 使用不同 \hat{t} 的 S-KLALM 算法的对比

我们在体系一上测试 $\hat{t} \in \{0, 5, 10\}$ 的 S-KLALM 算法. 我们采用等质量剖分生成网格, 离散规模 $K \in \{90, 180, 360, 720\}$. 对于每一对 (K, \hat{t}) , 我们让算法从 10 个不同的随机初始点出发. 终止参数为 $tol = 10^{-3} \times \sqrt{2^{\log_2(K/90)}}$ ¹ 和 $t_{\max} = \infty$. 我们将不同 (K, \hat{t}) 下 S-KLALM 算法的平均 err_obj 、 err_sce 和 T 汇总于图 4.2 中.

¹根据引理 4.3, $\mathcal{U}(\rho, \rho)$ 的大小以 $\mathcal{O}(1/\sqrt{K})$ 缩减. 由 (4.33) 式, 我们令 tol 随 K 增大而增大.

表 4.3 在不同采样概率下 S-ERALM 算法模拟体系一 ($K = 90$, 等质量剖分) 时的平均 err_obj、err_sce 和 T. 其中, 在“采样概率”列, “随机”表示随机生成采样概率, 其余表示使用不同 γ 的重要性采样概率

Table 4.3 The achieved err_obj, err_sce, and required T averaged over 10 trials given by the S-ERALM methods with different sampling probabilities on system 1 ($K = 90$, equimass discretization). In the column “Sampling probability”, “Random” refers to the results with randomly generated sampling probabilities, while others stand for the results with importance sampling-based probabilities using different values of γ

采样概率	err_obj	err_sce	T
随机	0.4184	0.86	128.84
$\gamma = 0.1$	0.3098	0.69	120.16
$\gamma = 0.3$	0.1729	0.54	93.30
$\gamma = 0.5$	0.1044	0.42	65.10
$\gamma = 0.7$	0.0693	0.39	48.13
$\gamma = 0.9$	0.0597	0.37	35.05
$\gamma = 0.99$	0.0525	0.36	21.76
$\gamma = 0.999$	0.1118	0.34	5.66

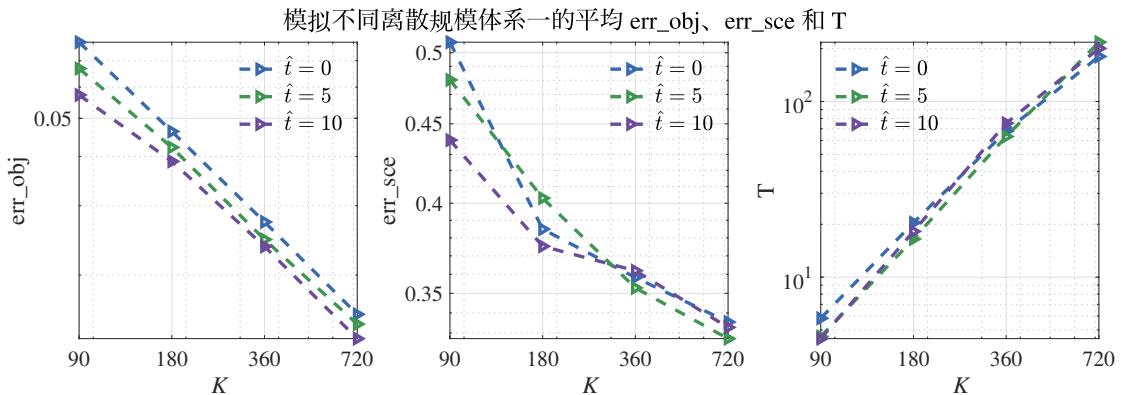


图 4.2 在不同 (K, \hat{t}) 下, S-KLALM 算法模拟体系一(等质量剖分)时的平均 err_obj、err_sce 和 T. 其中, 带有右向三角形标记的蓝色、绿色和紫色虚线分别表示 S-KLALM 算法在 $\hat{t} = 0, 5, 10$ 时所取得的结果. 左图: err_obj. 中图: err_sce. 右图: T

Figure 4.2 The achieved err_obj, err_sce, and required T averaged over 10 trials for each pair of (K, \hat{t}) given by the S-KLALM methods with different values of \hat{t} on system 1 (equimass discretization). The blue, green, and purple dashed lines with right-pointing triangle markers are the results of the S-KLALM methods with $\hat{t} = 0, 5, 10$, respectively. Left: err_obj. Middle: err_sce. Right: T

从图 4.2, 我们发现 $\hat{t} > 0$ 确实可以在差不多的 CPU 时间内使 S-KLALM 算法输出更高质量的解. 但随着离散规模 K 逐渐增大, 这一优势变得愈加不明显. 由于所关心的应用通常涉及大规模问题, 并且 $\hat{t} > 0$ 表明在头几次迭代中算法需要显式存储和计算全矩阵, 因此在后续的数值实验中, 我们选取 S-KLALM 算法

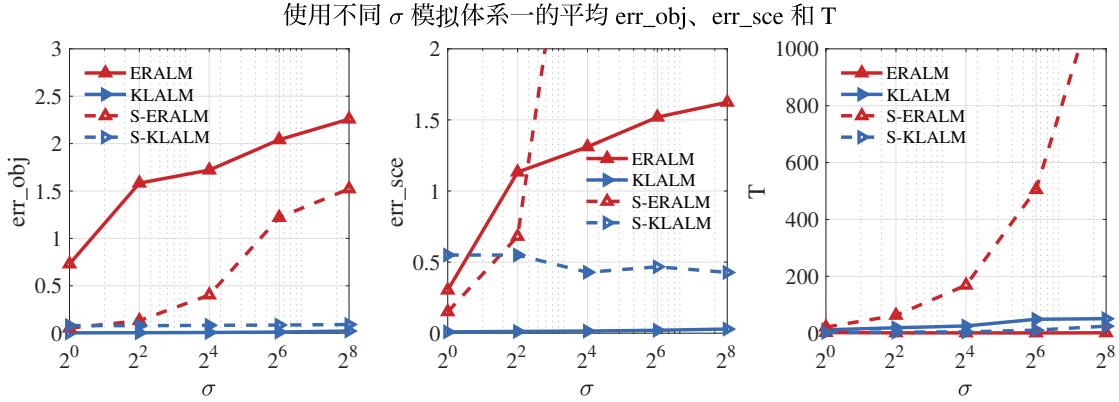


图 4.3 不同 σ 下, ERALM 算法、KLALM 算法、S-ERALM 算法与 S-KLALM 算法模拟体系一(等质量剖分, $K = 90$)时平均 err_obj、err_sce 和 T. 其中, 带有三角型标记的红色实线和虚线分别表示 ERALM 算法和 S-ERALM 算法的结果, 带有右向三角形标记的蓝色实线和虚线分别表示 KLALM 算法和 S-KLALM 算法的结果. 左图: err_obj. 中图: err_sce. 右图: T

Figure 4.3 The achieved err_obj, err_sce, and required T averaged over 10 trials for each value of σ given by the ERALM, KLALM, S-ERALM, and S-KLALM methods on system 1 with $K = 90$ (equimass discretization). The red solid and dashed lines with triangle markers represent the results of the ERALM and S-ERALM methods, respectively. The blue solid and dashed lines with right-pointing triangle markers represent the results of the KLALM and S-KLALM methods, respectively. Left: err_obj. Middle: err_sce. Right: T

中的 \hat{t} 为 0.

4.4.3.3 ERALM 算法、S-ERALM 算法、KLALM 算法与 S-KLALM 算法的对比

我们首先在体系一上比较四个算法在 $\sigma \in \{2^0, 2^2, 2^4, 2^6, 2^8\}$ 时的数值表现. 我们使用等质量剖分, 固定离散规模为 $K = 90$. 对每个 σ 的取值, 我们让算法从 10 个不同的随机初始点出发. 终止参数为 $tol = 5 \times 10^{-3}/\sqrt{\sigma}$ 和 $t_{max} = \infty$. 我们将不同 σ 下四个算法的平均 err_obj、err_sce 和 T 汇总在图 4.3 中.

从图 4.3, 我们发现 (1) 随着 σ 取值的增大, ERALM 算法和 S-ERALM 算法的目标函数值误差显著升高. 这与第 4.3 节中的理论结果相合; (2) 不论 σ 取值如何, KLALM 算法和 S-KLALM 算法输出解的质量变化不大. 这充分反映了基于 KL 散度的算法对于邻近参数选取的鲁棒性. 由于只有当 σ 取得足够小时 (S-)ERALM 算法才可以算到高质量解, 而太小的 σ 会导致严重的数值下溢, 因此这一参数鲁棒性尤为重要.

在图 4.3 中, 我们还注意到相较于 S-KLALM 算法, S-ERALM 算法在 σ 较小时可以输出更高质量的解. 不过, 下面我们将指出, 这一优势会随着 K 逐渐增大而彻底消失. 我们在体系一上测试了 $\sigma = 1$ 时的 S-ERALM 算法和 S-KLALM 算法, 采用等质量网格剖分, 离散规模 $K \in \{90, 180, 360, 720\}$. 对于每个 K 的取值, 我们让算法从 10 个不同的随机初始点出发. 终止参数为 $tol = 10^{-3} \times \sqrt{2}^{\log_2(K/90)}$

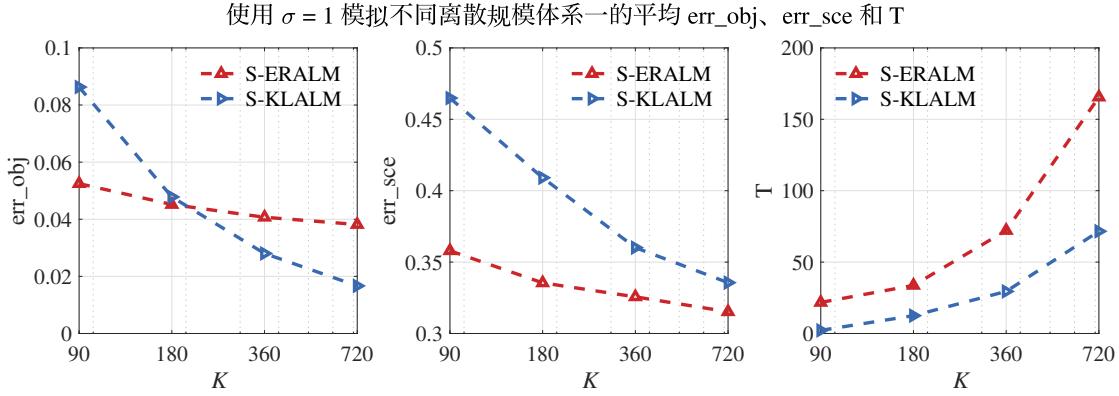


图 4.4 在不同 K 下, S-ERALM 算法与 S-KLALM 算法 ($\sigma = 1$) 模拟体系一(等质量剖分)时的平均 err_obj、err_sce 和 T. 其中, 带有三角形标记的红色虚线表示 S-ERALM 算法的结果, 带有右向三角形标记的蓝色虚线表示 S-KLALM 算法的结果. 左图: err_obj. 中图: err_sce. 右图: T

Figure 4.4 The achieved err_obj, err_sce, and required T averaged over 10 trials for each value of K given by the S-ERALM and S-KLALM methods with $\sigma = 1$ on system 1 (equimass discretization). The red dashed lines with triangle markers represent the results of the S-ERALM method. The blue dashed lines with right-pointing triangle markers represent the results of the S-KLALM method. Left: err_obj. Middle: err_sce. Right: T

和 $t_{\max} = \infty$. 我们将不同 K 下两个算法的平均 err_obj、err_sce 和 T 汇总于图 4.4 中.

从图 4.4, 我们不难发现随着问题规模的增大, S-ERALM 算法对于 S-KLALM 算法的精度优势逐渐减小甚至完全消失, 且所需 CPU 时间增长更加迅速. 这与表 4.1 是一致的, 即 S-ERALM 算法每次迭代均需进行矩阵逐元素采样, 其具有二次复杂度.

最后, 我们比较 KLALM 算法与 S-KLALM 算法. 由于本质上是在求解一个自由度受限的问题, 因此 S-KLALM 算法要想在解的质量上超越 KLALM 算法是不现实的. 然而, 若计算资源有限, 我们可以用 S-KLALM 算法在较少的 CPU 时间内获取较高质量的解. 为佐证这一说法, 我们在体系一上比较 KLALM 算法与 S-KLALM 算法, 使用等质量剖分, 考虑离散规模 $K \in \{90, 180, 360, 720, 1440, 2880\}$. 对每个 K 的取值, 我们让算法从 10 个不同的随机初始点出发. 终止参数为 $tol = 10^{-3} \times \sqrt{2^{\log_2(K/90)}}$ 和 $t_{\max} = \infty$. 我们将不同 K 下两个算法的平均 err_obj 随 CPU 时间的收敛曲线绘制在图 4.5 中, 其中 T_{inter} 表示两个算法的收敛曲线最后一次相交的 CPU 时间.

从图 4.5 不难发现, KLALM 算法输出的目标函数值误差只有在 CPU 时间超过 T_{inter} 后才会优于 S-KLALM 算法. 此外, 若我们增大 K , T_{inter} 会以立方速度增长. 三次多项式拟合结果可见图 4.6. 这表明 (1) 若问题规模较小 (例如, K 在 10^2 量级), KLALM 算法在可接受的 CPU 时间内能算得高质量的解; (2) 若问题规模较大 (例如, K 在 10^3 甚至更高量级) 且计算资源有限, S-KLALM 算法则更加实用.

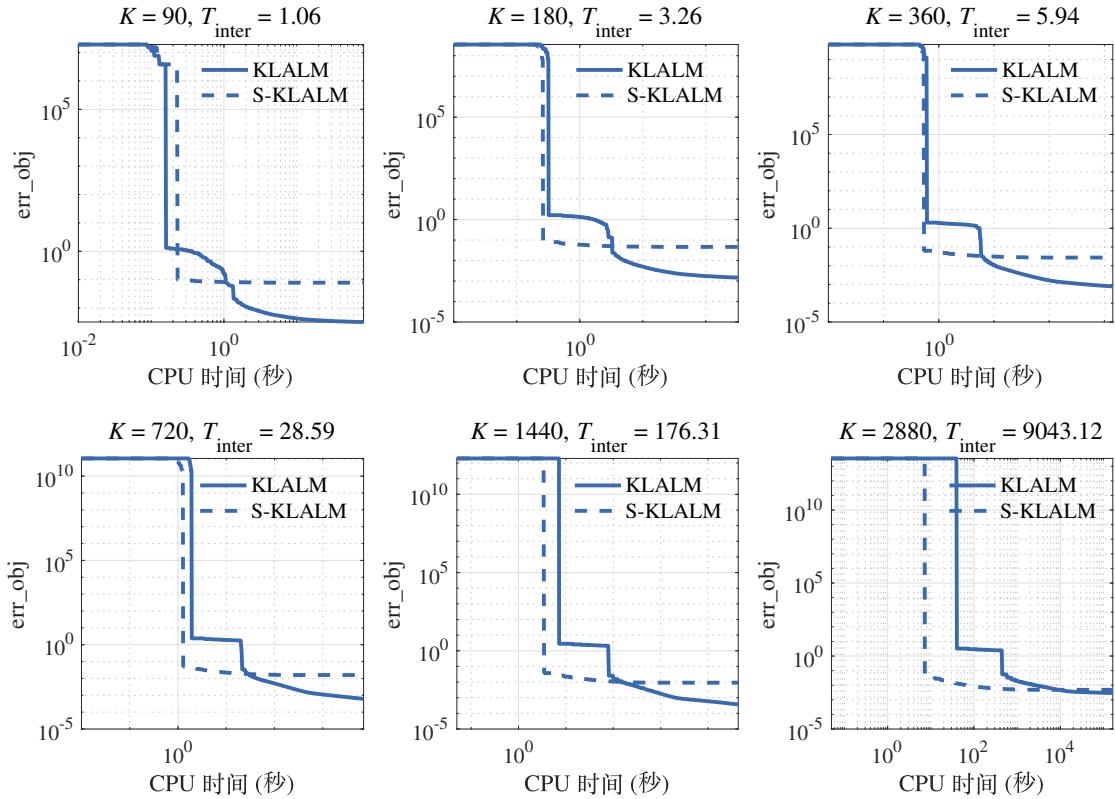


图 4.5 在不同 K 下, KLALM 算法与 S-KLALM 算法在模拟体系一(等质量剖分)时的平均 err_obj 随 CPU 时间的收敛曲线. 其中, 蓝色实线和虚线分别表示 KLALM 算法与 S-KLALM 算法的结果, T_{inter} 代表两个算法的曲线最后一次相交时的 CPU 时间

Figure 4.5 The convergence curves of err_obj along with the CPU time averaged over 10 trials for each value of K given by the KLALM and S-KLALM methods on system 1 (equimass discretization). The blue solid and dashed lines stand for the results of the KLALM and S-KLALM methods, respectively. The notation T_{inter} refers to the CPU time where the curves of two methods intersect for the last time

基于上面的数值结果, 我们将 KLALM 算法与 S-KLALM 算法用于和 PALM-I 算法的比较以及二维、三维体系的模拟.

4.4.3.4 PALM-I 算法、KLALM 算法与 S-KLALM 算法的对比

我们在表 4.2 中的所有一维体系(等质量剖分)上比较 PALM-I 算法、KLALM 算法与 S-KLALM 算法的数值表现. 在 PALM-I 算法中, 我们按 (4.32) 式设置邻近参数. 对于体系一和二, 我们考虑离散规模 $K \in \{90, 180, 360, 720\}$. 对每个 K 的取值, 我们让算法从 10 个不同的随机初始点出发. 终止参数为 $\text{tol} = 10^{-3} \times \sqrt{2^{\log_2(K/90)}}$ 和 $t_{\max} = \infty$. 对于体系三和四, 我们考虑离散规模 $K \in \{140, 280, 560, 1120\}$. 对每个 K 的取值, 我们让算法从 10 个不同的随机初始点出发. 终止参数为 $\text{tol} = 10^{-3} \times \sqrt{2^{\log_2(K/140)}}$ 和 $t_{\max} = \infty$. 我们将三个算法在四个体系不同 K 下的平均 err_obj、err_sce 和 T 汇总于图 4.7 与 4.8 中.

在同样的邻近参数与终止参数下, PALM-I 算法容易早熟, 而 KLALM 算法

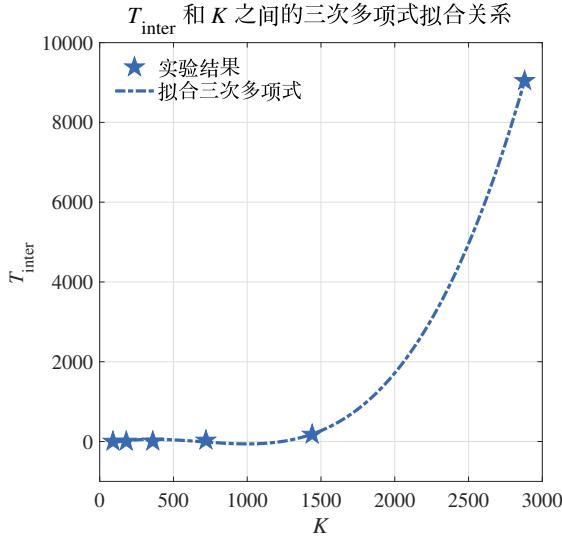


图 4.6 在体系一(等质量剖分) 上 T_{inter} 和 K 的三次多项式拟合结果. 拟合得到的多项式为 $9.09 \times 10^{-7}K^3 - 1.86 \times 10^{-3}K^2 + 1.00K - 106.38$. 其中, 蓝色五角星为实验中得到的 T_{inter} , 蓝色点划线表示拟合的三次关系

Figure 4.6 Cubic polynomial fitted relation between T_{inter} and K on system 1 (equimass discretization). The fitted polynomial is $9.09 \times 10^{-7}K^3 - 1.86 \times 10^{-3}K^2 + 1.00K - 106.38$. The blue pentagons are the obtained T_{inter} in experiments and the blue dashdotted line is the fitted relation

在一维体系上获得了显著更低的目标函数值误差. 此外, KLALM 算法子问题最优解的乘积表达式 (4.19) 使得消除全矩阵成为可能, 从而衍生出具有更好标度的 S-KLALM 算法.

4.4.4 二维、三维强关联电子体系计算

在本小节中, 我们将 KLALM 算法与 S-KLALM 算法作为 CMGOPT 框架(见框架 3.3) 中的优化算法, 模拟二维、三维强关联体系. 与第 3 章不同的是, 我们将 KLALM 算法作为 CMGOPT 框架中的高精度优化算法, 而将具有更好标度的 S-KLALM 算法用于求解加密网格上的问题. 我们将这样的算法称为 S-KLALM-CMG 算法. 在 CMGOPT 框架中, 我们采用均匀剖分生成初始网格, 并在随后的步骤中一致加密网格. 对于二维体系, CMGOPT 框架的初始网格数在密度向量截断前为 $K_0 = 900$, 加密次数为 $\ell_{\max} = 3$. 对于三维体系七和八, CMGOPT 框架的初始网格数在密度向量截断前分别为 $K_0 = 1728$ 和 $K_0 = 1000$, 加密次数均为 $\ell_{\max} = 2$. 对于第 ℓ 层网格上的问题, 算法的终止参数为 $t_{\max} = 10^4$ 和

$$tol = \begin{cases} 5 \times 10^{-3}, & \ell = 0; \\ 10^{-2} \times (\sqrt{2^d})^{\log_2(K/K_0)}, & \ell > 0. \end{cases}$$

由于这些体系目前还没有显式的最优解构造, 因此我们记录算法输出的目标函数值 f_{out} 和 $\hat{\lambda}$, 并通过绘制近似映射 $\{\hat{\mathcal{T}}_i\}_{i=2}^N$ 的切片评估解的质量. 我们将数值结果汇总在表 4.4 以及图 4.9 和 4.10 中, 其中 $K_{\text{trunc}} \in \mathbb{N}$ 表示截断后 \mathbf{p} 的维数.

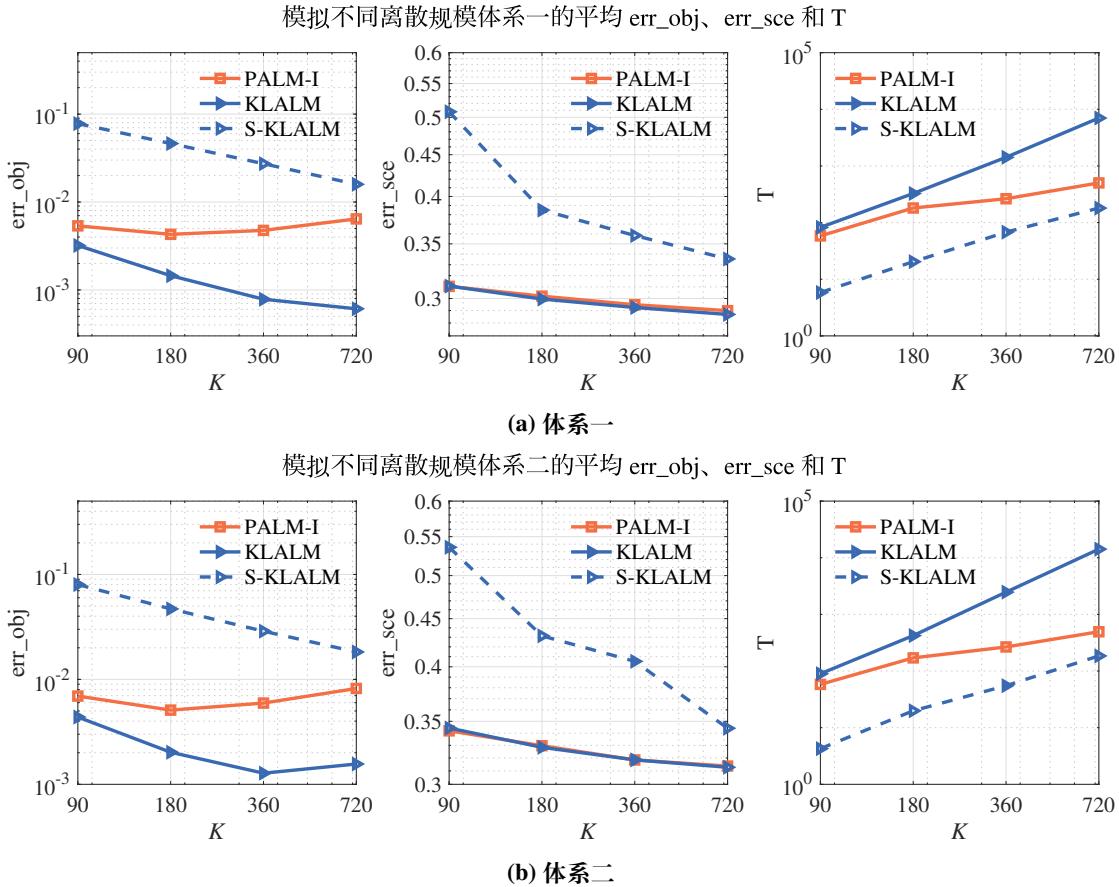


图 4.7 在不同 K 下, PALM-I 算法、KLALM 算法与 S-KLALM 算法在模拟一维三电子体系 (等质量剖分) 时的平均 err_obj、err_sce 和 T. 其中, 带有方形标记的橙色实线代表 PALM-I 算法的结果, 带有右向三角形标记的蓝色实线和虚线分别代表 KLALM 算法和 S-KLALM 算法的结果. 由左至右: err_obj、err_sce 和 T. (a) 体系一. (b) 体系二

Figure 4.7 The achieved err_obj, err_sce, and required T averaged over 10 trials for each value of K given by the PALM, KLALM, and S-KLALM methods on the one-dimensional systems with three electrons (equimass discretization). The orange solid lines with square markers are the results of the PALM method. The blue solid and dashed lines with right-pointing triangle markers are the results of the KLALM and S-KLALM methods, respectively. From left to right: err_obj, err_sce, and T. (a) System 1. (b) System 2

表 4.4 和图 4.9 展现了 CMGOPT 框架的收敛性及其中插值算子的有效性. 图 4.10 则表明由解得到的近似映射符合物理直观, 从而间接说明取得的解具有较高质量. 具体来讲, 对于体系五, 图 4.10 (a) 表明若其中一个电子在一个 Gauss 型函数中心附近, 则另外两个电子会相应位于另外两个 Gauss 型函数中心附近. 对于体系六, 图 4.10 (b) 表明若其中一个电子位于上方 Gauss 型函数中心附近但与中心保持一定距离, 则另外三个电子会相应位于三个 Gauss 型函数中心附近, 其中一个所在的区域被第一个电子出现的区域包裹, 满足 Coulomb 排斥效应以及预先设定的质量比. 对于体系七, 图 4.10 (c) 与 (a) 反映的现象类似. 对于体系八, 图 4.10 (d) 表明若其中一个电子位于 Gauss 型函数中心 $[1, 0, 0]^T$ 附近, 则另外三个电子将位于另一个 Gauss 型函数中心 $[-1, 0, 0]^T$ 附近. 它们所在的区域正好将一

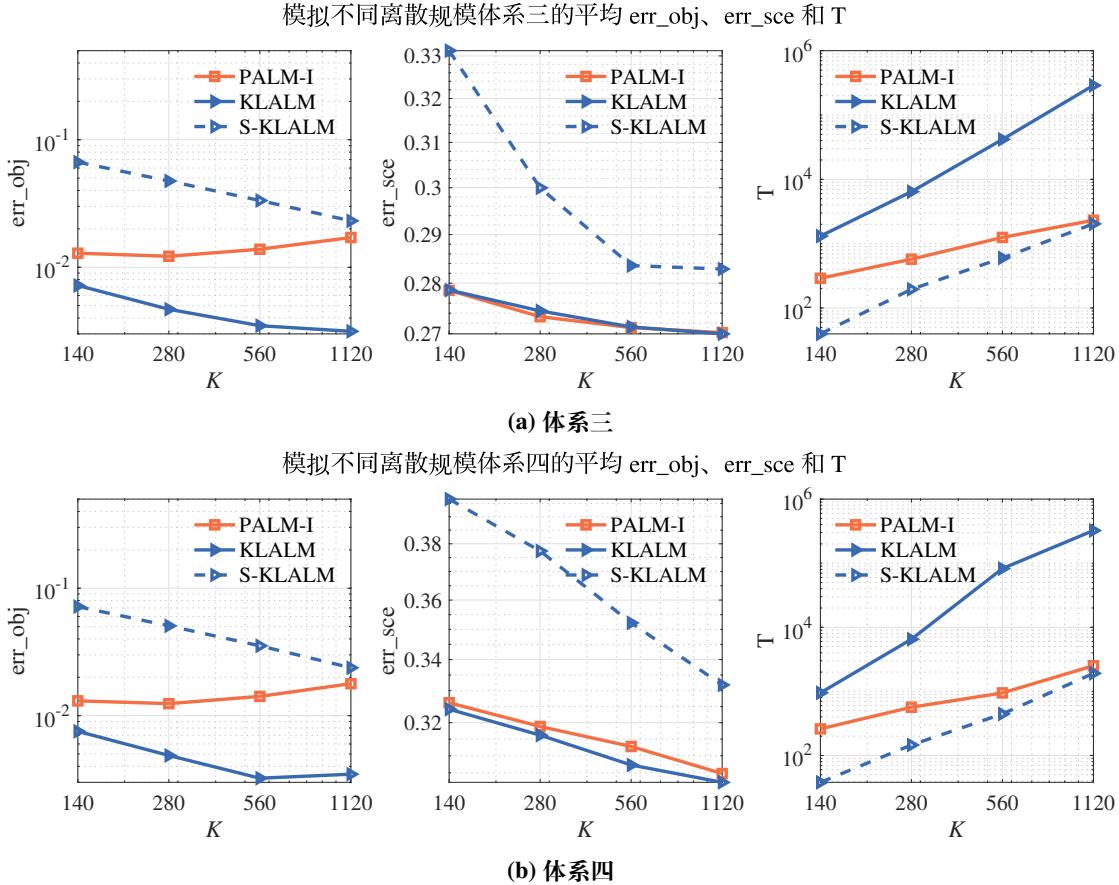


图 4.8 在不同 K 下, PALM-I 算法、KLALM 算法与 S-KLALM 算法在模拟一维七电子体系 (等质量剖分) 时的平均 err_obj、err_sce 和 T. 其中, 带有方形标记的橙色实线代表 PALM-I 算法的结果, 带有右向三角形标记的蓝色实线和虚线分别代表 KLALM 算法和 S-KLALM 算法的结果. 由左至右: err_obj、err_sce 和 T. (a) 体系三. (b) 体系四

Figure 4.8 The achieved err_obj, err_sce, and required T averaged over 10 trials for each value of K given by the PALM, KLALM, and S-KLALM methods on the one-dimensional systems with seven electrons (equimass discretization). The orange solid lines with square markers are the results of the PALM method. The blue solid and dashed lines with right-pointing triangle markers are the results of the KLALM and S-KLALM methods, respectively. From left to right: err_obj, err_sce, and T. (a) System 3. (b) System 4.

个球体划分为三个部分, 满足 Coulomb 排斥效应与预先设定的质量比.

值得一提的是, 我们在图 4.10(c) 和 (d) 中首次可视化了三维情形下电子位置之间的近似映射. 相比于文献^[91] 和第 3 章, 本章充分利用低标度的算法, 在更大规模的问题上取得了合理的数值结果.

4.4.5 算法标度测试

最后, 我们测试 KLALM 算法和 S-KLALM 算法对于 K 和 N 的标度. 我们考虑单电子密度与体系一类型相同的体系, 变化其中的电子个数和离散规模, 采用等质量剖分.

为测试算法对于 K 的标度, 我们固定 $N = 3$, 考虑 $K \in \{90, 180, 360, 720\}$.

表 4.4 在二维、三维体系上, S-KLALM-CMG 算法在每层网格上输出的目标函数值

Table 4.4 The objective values for each level given by the S-KLALM-CMG method on the two/three-dimensional systems

ℓ	体系五			体系六		
	K	K_{trunc}	f_{out}	K	K_{trunc}	f_{out}
0	900	424	1.1339	900	408	3.0690
1	3600	1622	1.1337	3600	1534	3.0690
2	14400	6410	1.1335	14400	6068	3.0677
3	57600	25562	1.1334	57600	24176	3.0667

ℓ	体系七			体系八		
	K	K_{trunc}	f_{out}	K	K_{trunc}	f_{out}
0	1728	780	1.0202	1000	720	4.6193
1	13824	5628	1.0209	8000	5272	4.6716
2	110592	42936	1.0209	64000	40764	4.6833

对 K 的每个取值, 我们让算法从 10 个不同的随机初始点出发. 邻近参数被固定为 $\mu_{i,t} \equiv 0.05$. 终止参数为 $tol = 10^{-3}$ 和 $t_{\max} = \infty$. 我们将不同 K 下两个算法的平均 err_obj、err_sce 和 T 汇总于图 4.11 (a) 中. 根据表 4.1, 我们拟合 $\ln(T)$ 与 $\ln(K)$ 之间的线性关系, 得到

$$\text{KLALM: } \ln(T) \approx 2.59 \ln(K) - 7.22;$$

$$\text{S-KLALM: } \ln(T) \approx 2.21 \ln(K) - 8.31.$$

为测试算法对于 N 的标度, 我们固定 $K = 144$, 考虑 $N \in \{3, 6, 12, 24, 48\}$. 对 N 的每个取值, 我们让算法从 10 个不同的随机初始点出发. 由于 $\|\mathbf{v}_i^{(t)}\|_\infty$ 会随 N 增大显著变化, 我们固定邻近参数为 $\mu_{i,t} \equiv 20/\ln(K)$. 终止参数为 5×10^{-3} 和 $t_{\max} = \infty$. 我们将不同 N 下两个算法的平均 err_obj、err_sce 和 T 汇总于图 4.11 (b) 中. 类似地, 我们得到了如下关系:

$$\text{KLALM: } \ln(T) \approx 1.31 \ln(N) + 0.32;$$

$$\text{S-KLALM: } \ln(T) \approx 1.06 \ln(N) + 0.49.$$

从图 4.11, 我们发现 KLALM 算法与 S-KLALM 算法对于 N 的标度与表 4.1 ($\tau = 0.5$) 中所列大致相符, 而对于 K 的标度则与理论估计有显著差距. 这是因为对于不同的 K , 算法所需要的迭代次数有较大的区别, 而表 4.1 中所列的是算法的单步计算复杂度. 此外, 表 4.1 中所列的复杂度仅为上界估计. 利用稀疏矩阵, 算法在实际运行时的复杂度可能更低.

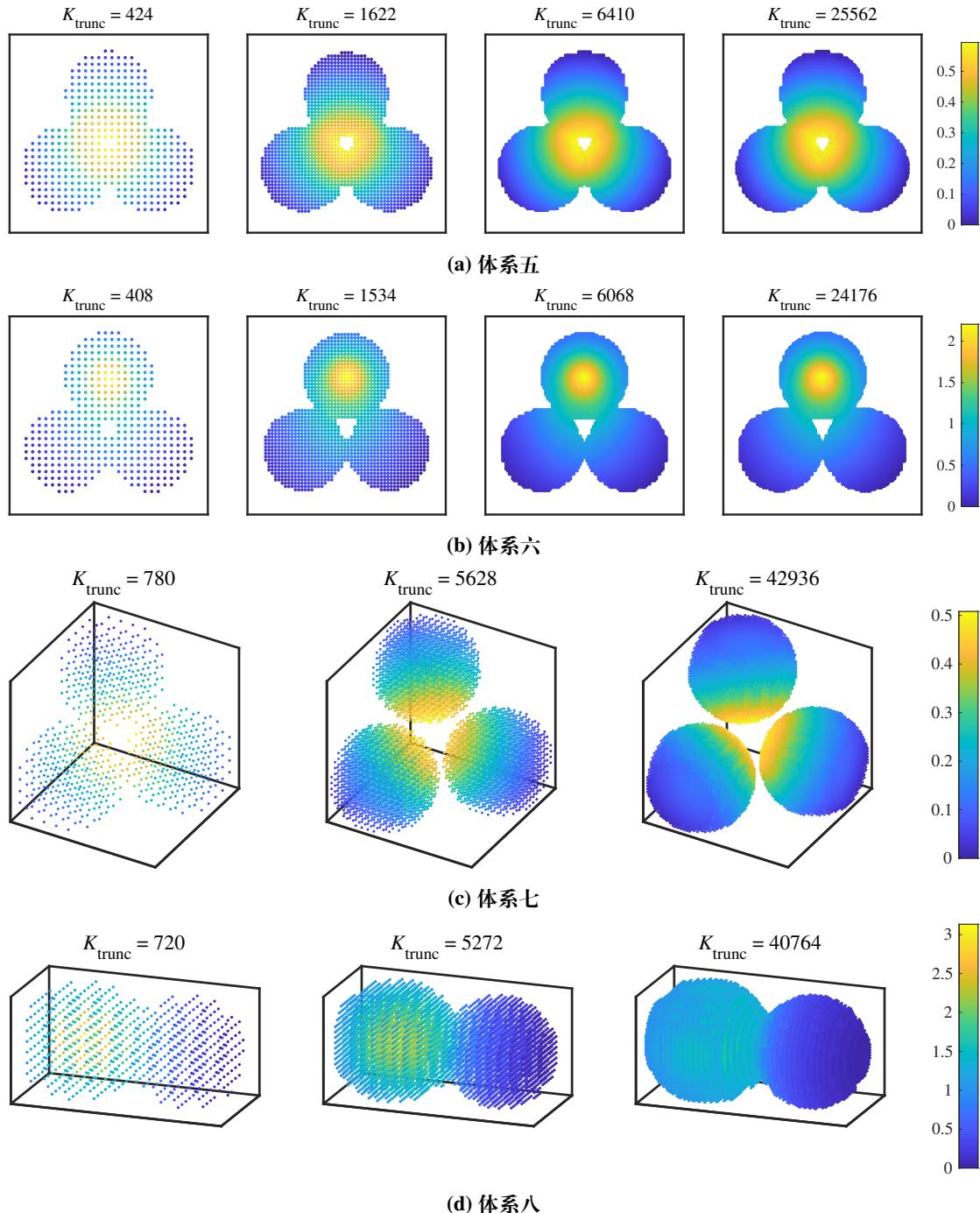


图 4.9 在二维、三维体系上, S-KLALM-CMG 算法在每层网格上输出的 $\hat{\lambda}$. (a) 体系五. (b) 体系六. (c) 体系七. (d) 体系八

Figure 4.9 The vectors $\hat{\lambda}$ at each level given by the S-KLALM-CMG method on the two/three-dimensional systems. (a) System 5. (b) System 6. (c) System 7. (d) System 8

4.5 本章小结

在本章中, 我们考虑了运输多胞体上的分块矩阵优化问题, 并为其设计了块坐标下降型算法. 强关联电子体系计算中的问题 (2.8) 也是此类问题的特例. 现有块坐标下降型算法均需要显式地存储与操作全矩阵变量, 这一缺陷使它们不适合

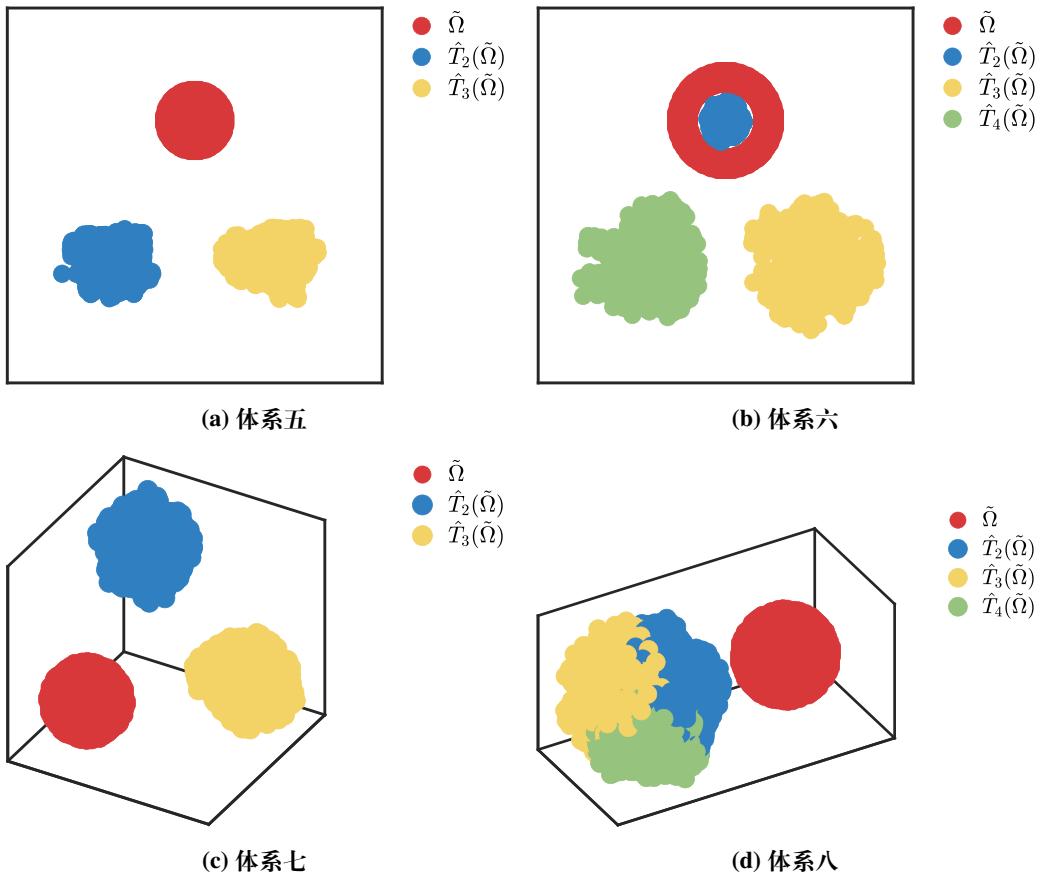


图 4.10 在二维、三维体系上, S-KLALM-CMG 算法在最后一层网格上输出的近似映射 $\{\hat{\mathcal{T}}_i\}_{i=2}^N$ 切片. 其中, 红色部分表示原像 $\tilde{\Omega} \subseteq \Omega$, 其他颜色表示原像在 $\{\hat{\mathcal{T}}_i\}_{i=2}^N$ 下的像.
(a) 体系一. (b) 体系二. (c) 体系三. (d) 体系四

Figure 4.10 The approximate mappings at the last level given by the S-KLALM-CMG method on the two/three-dimensional systems. The red part represents the pre-images $\tilde{\Omega} \subseteq \Omega$, while the parts in other colors are the images of $\tilde{\Omega}$ under $\{\hat{\mathcal{T}}_i\}_{i=2}^N$. (a) System 5. (b) System 6. (c) System 7. (d) System 8

于求解大规模问题. 为此, 我们设计了基于熵正则的 ERALM 算法与基于 KL 散度的 KLALM 算法, 并利用其中子问题最优解的乘积表达式, 进一步设计了它们的采样版本 S-ERALM 算法与 S-KLALM 算法. 由于子问题仅涉及采样得到的一小部分指标集上的自由度, S-ERALM 算法与 S-KLALM 算法完全免去了全矩阵的使用, 具有更低的计算标度. 借助矩阵逐元素随机近似理论, 我们证明了 ERALM 算法与 S-ERALM 算法的平均稳定性违反度会随着问题规模趋于无穷大而 (以趋于 1 的概率) 收敛到 0. 据我们所知, 我们的工作首次为矩阵逐元素近似在分块非凸问题上的应用提供了理论保证.

在一维强关联电子体系上, 我们发现尽管基于 KL 散度的算法缺少理论支撑, 但其数值表现对邻近参数的选取不敏感. 相较于 KLALM 算法, S-KLALM 算法以精度上的损失换取了标度上显著的改进. 我们还将 KLALM 算法与 S-KLALM 算法作为 CMGOPT 框架中的优化算法, 模拟了二维、三维强关联电子体系. 相比于第 3 章, 本章充分利用低标度的算法, 在更大规模的问题上取得了合理的数值

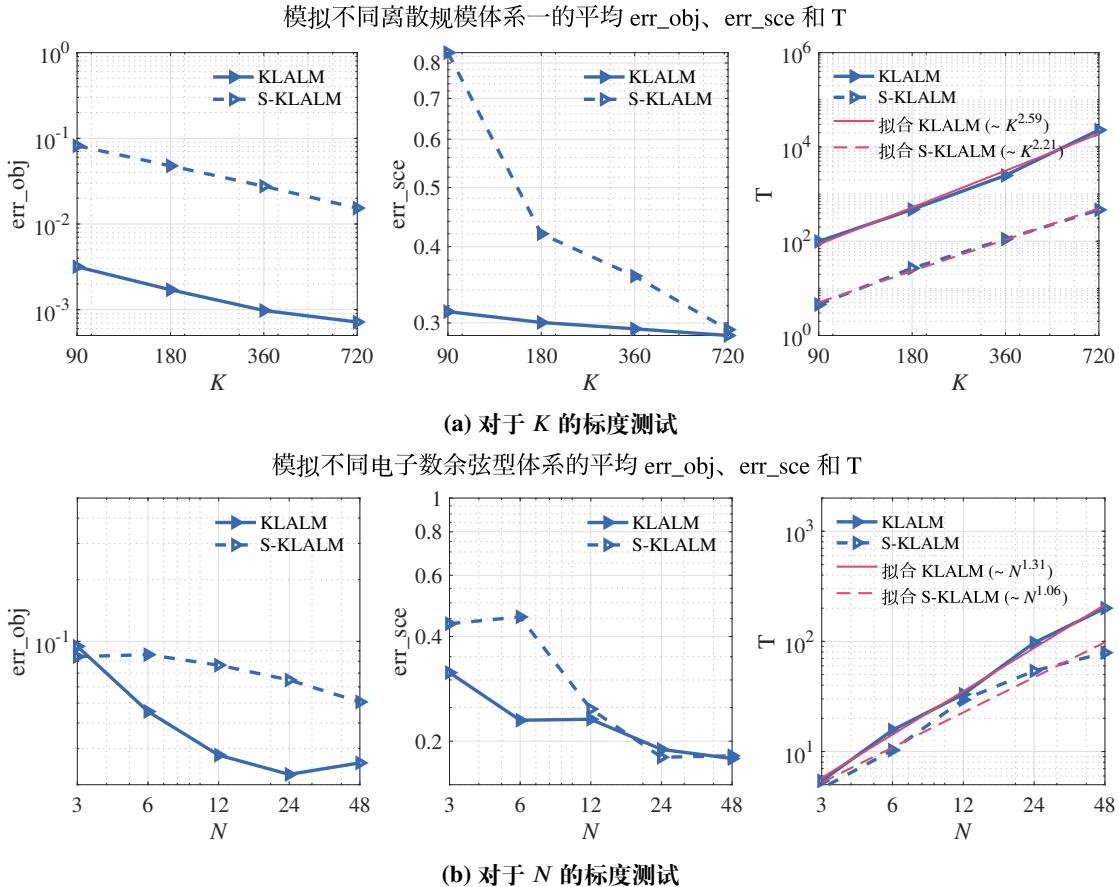


图 4.11 在不同 K 和 N 下, KLALM 算法与 S-KLALM 算法在模拟一维余弦型体系 (等质量剖分) 时的平均 err_obj、err_sce 和 T. 其中, 蓝色实线与虚线分别表示 KLALM 算法与 S-KLALM 算法的结果, 粉色实线与虚线分别代表所拟合的 KLALM 算法与 S-KLALM 算法的 T 与 K 或 T 与 N 之间的关系. 左图: err_obj. 中图: err_sce. 右图: T. (a) 对于 K 的标度测试. 对于 KLALM 算法, $T \sim K^{2.6}$. 对于 S-KLALM 算法, $T \sim K^{2.2}$. (b) 对于 N 的标度测试. 对于 KLALM 算法, $T \sim N^{1.3}$. 对于 S-KLALM 算法, $T \sim N^{1.1}$

Figure 4.11 The achieved err_obj, err_sce, and required T averaged over 10 trials for each value of K and N given by the KLALM and S-KLALM methods on the one-dimensional cosine-type system (equimass discretization). The blue solid and dashed lines represent the results of the KLALM and S-KLALM methods, respectively. The pink solid and dashed lines denote the fitted relations between T and K or T and N when using the KLALM and S-KLALM methods, respectively. Left: err_obj. Middle: err_sce. Right: T. (a) Scalability tests with respect to K . For the KLALM method, $T \sim K^{2.6}$. For the S-KLALM method, $T \sim K^{2.2}$. (b) Scalability tests with respect to N . For the KLALM method, $T \sim N^{1.3}$. For the KLALM method, $T \sim N^{1.1}$

结果, 并首次可视化了三维情形电子位置之间的映射.

第 5 章 求解行列式约束优化问题的投影梯度下降算法

在本章中, 我们考虑行列式约束优化问题. 此类问题与第 1 章的第一性原理固定晶格体积晶体结构弛豫问题 (1.37) 紧密相关. 我们为其设计了使用缩放算子保持迭代点可行性的投影梯度下降 (PGD) 算法, 并证明了算法的理论性质. 在数值实验中, 我们 (1) 将 PGD 算法推广至求解问题 (1.37); (2) 在一个基准算例集上对比了 CG 算法、QN 算法与 PGD 算法的效率与鲁棒性; (3) 将 PGD 算法用于计算难弛豫的高熵合金 AlCoCrFeNi 的物态方程, 并将计算结果与已有实验测定数据进行对比.

5.1 问题描述与研究现状

我们考虑如下行列式约束优化问题:

$$\min_X f(X), \text{ s. t. } \det(X) = V, \quad (5.1)$$

其中 $X \in \mathbb{R}^{n \times n}$ ($n \in \mathbb{N}$), $V \in \mathbb{R}$ 为给定常数, $f : \mathbb{R}^{n \times n} \rightarrow \mathbb{R}$ 可微未必凸. 在固定原子位置时, 第一性原理固定晶格体积的晶体结构弛豫问题 (1.37) 具有问题 (5.1) 的形式. 问题 (1.37) 在结构物态方程计算中举足轻重^[12,13]. 常用材料模拟软件 (例如 VASP^[122,123]、QE^[124] 等) 均实现了该功能.

由于可行域的非凸性 (可见图 1.2), 传统优化中的许多工具 (如正交投影) 均无法直接使用. 这给算法的设计与理论分析带来了巨大挑战. 目前, 尚无已有工作针对行列式约束优化问题设计全局收敛算法.

在第一性原理固定晶格体积晶体结构弛豫应用中, 人们通常使用 CG 算法与 QN 算法求解对应的问题. 其中, CG 算法相较 QN 算法更加稳定, 而 QN 算法在局部最优附近收敛更快. 特别地, 在 VASP 中实现的 CG 算法与 QN 算法通过如下缩放算子 (也可见 (1.38) 式) 的作用保持迭代点的可行性¹:

$$\mathcal{P}_V(X) := \sqrt[n]{\frac{V}{\det(X)}} X. \quad (5.2)$$

该缩放算子尽管不像正交投影对迭代点做最小的改变, 但充分地利用了行列式的性质, 单值且有解析的表达式. 在实际使用中, CG 算法与 QN 算法在效率上并不令人满意. 在使用 \mathcal{P}_V 保持可行性时, 这些算法也暂无全局收敛性保证. 这些缺陷使第一性原理固定晶格体积晶体结构弛豫成为下游应用的计算瓶颈.

5.1.1 本章主要内容

在本章中, 我们为问题 (5.1) 的求解设计了 PGD 算法, 并首次在使用缩放算子 (5.2) 保持迭代点可行性的条件下, 证明了算法的全局依子列收敛性. 在理论分

¹在第一性原理固定晶格体积晶体结构弛豫应用中, (5.2) 式的 n 取为 3.

析中, 我们主要借助可行域切锥的代数结构与非单调线搜索. 在数值实验中, 我们将 PGD 算法推广至求解问题 (1.37). 在含有 223 个来自不同类别结构的基准算例集上, PGD 算法在效率与鲁棒性方面均显著优于 CG 算法与 QN 算法. 我们还使用 PGD 算法计算了难驰豫的高熵合金 AlCrCrFeNi 的物态方程, 得到了与已有实验测定数据相合的计算结果.

5.2 算法设计

我们所设计的 PGD 算法是求解问题 (5.1) 的可行线搜索型算法. 下面, 我们依次介绍其主要步骤, 包括搜索方向与步长的确定等.

PGD 算法以负梯度到可行域在当前迭代点切锥上的正交投影为搜索方向:

$$\mathbf{D}^{(k)} := \mathcal{P}_{\mathcal{T}^{(k)}}(-\nabla f(\mathbf{X}^{(k)})) = -\nabla f(\mathbf{X}^{(k)}) + \frac{\langle Y^{(k)}, \nabla f(\mathbf{X}^{(k)}) \rangle}{\|Y^{(k)}\|^2} Y^{(k)} \in \mathbb{R}^{n \times n}, \quad (5.3)$$

其中对任意集合 \mathcal{A} , $\mathcal{P}_{\mathcal{A}}$ 为到 \mathcal{A} 上的正交投影算子, 定义为

$$\mathcal{P}_{\mathcal{A}}(\mathbf{x}) \in \arg \min_{\mathbf{y} \in \mathcal{A}} \|\mathbf{x} - \mathbf{y}\|,$$

矩阵 $Y := X^{-T} \in \mathbb{R}^{n \times n}$, 可行域在 $X^{(k)}$ 处的切锥²

$$\mathcal{T}^{(k)} := \left\{ \mathbf{D} \in \mathbb{R}^{n \times n} : \langle Y^{(k)}, \mathbf{D} \rangle = 0 \right\}.$$

之后, 在一定条件下, 我们会证明 (5.3) 式中的正交投影可消除目标函数值的一阶增长量, 从而为 PGD 算法的收敛性分析奠定基础.

在确定好搜索方向 $\mathbf{D}^{(k)}$ 后, 从某个 $\alpha_{k-1,0} > 0$ 出发, 我们搜索满足如下非单调线搜索准则的步长 $\alpha_k := \alpha_{k-1,l} > 0$:

$$f(X^{(k,l)}) \leq \bar{f}_k - \eta \cdot \alpha_{k-1,l} \|\mathbf{D}^{(k)}\|^2, \quad (5.4)$$

其中

$$X^{(k,l)} := \mathcal{P}_{\mathcal{V}}(X^{(k)} + \alpha_{k-1,l} \mathbf{D}^{(k)}), \quad (5.5)$$

$\{\bar{f}_k\}$ 是代理序列, 按照如下方式递归更新: 从 $\bar{f}_0 := f(X^{(0)})$, $q_0 := 1$ 出发, 对任意 $k \geq 1$,

$$\bar{f}_k := \frac{\bar{f}_{k-1} + \mu_{k-1} q_{k-1} f(X^{(k)})}{1 + \mu_{k-1} q_{k-1}}, \quad (5.6)$$

$$q_k := 1 + \mu_{k-1} q_{k-1}, \quad \mu_{k-1} \in [0, 1],$$

$\eta > 0$ 是一个给定常数. 不难验证, 对任意 $k \geq 0$, \bar{f}_k 是 $\{f(X^{(t)})\}_{t=0}^k$ 的凸组合, 且恒有 $\bar{f}_k \geq f(X^{(k)})$. 因此, 相比单调线搜索准则^[1,20,236], 非单调准则 (5.4) 更容易

²此表达式由线性化可行方向锥的计算公式得到 (见定义 1.3). 在每个可行点处, 由于问题 (5.1) 满足线性独立约束规范条件, 因此可行域在该点处的切锥 (见定义 1.2) 与线性化可行方向锥 (见定义 1.3) 相等.

满足. 若按(5.5)式计算的试探点 $X^{(k,l)}$ 不满足非单调线搜索准则(5.4), 则使用回溯法减小步长:

$$\alpha_{k-1,l+1} := \delta \cdot \alpha_{k-1,l}, \quad l \geq 0, \quad (5.7)$$

并相应按(5.5)式计算试探点 $X^{(k,l+1)}$. 这里 $\delta \in (0, 1)$ 是给定回溯因子. 之后, 在一定条件下, 我们会证明回溯搜索(5.7)总是能在有限步内找到一个满足(5.4)式的步长 α_k . 而(5.4)式所要求的相对于代理序列的充分下降性则可用于证明PGD算法的全局依子列收敛性.

非单调线搜索准则(5.4)与文献^[284]中的准则十分相似. 它们的区别仅在于代理序列的更新公式(5.6). 记文献^[284]中的代理序列为 $\{\tilde{f}_k\}$, 其更新公式为

$$\begin{aligned} \tilde{f}_k &:= \frac{\lambda_{k-1} p_{k-1} \tilde{f}_{k-1} + f(X^{(k)})}{\lambda_{k-1} p_{k-1} + 1}, \\ p_k &:= \lambda_{k-1} p_{k-1} + 1, \quad \lambda_{k-1} \in [0, 1], \end{aligned} \quad (5.8)$$

其中 $\tilde{f}_0 = f(X^{(0)})$, $p_0 := 1$. 与(5.8)式相比, (5.6)式在迭代初期会将更大的权重放在过去的目标函数值上^[285]. 另一种常用的非单调线搜索准则^[286]要求

$$f(X^{(k,l)}) \leq \max_{t=\max\{0,k-\hat{k}+1\}}^k \{f(X^{(t)})\} - \hat{\eta} \cdot \alpha_{k-1,l} \|D^{(k)}\|^2, \quad (5.9)$$

其中 $\hat{k} \in \mathbb{N}$ 是存储的目标函数值个数, $\hat{\eta} > 0$ 为一给定常数. 文献^[284]中的数值实验表明, 以有限内存QN方向^[135]作为搜索方向时, 基于(5.8)式更新代理序列的非单调准则(5.4)在多数情形下较非单调准则(5.9)数值表现更好.

我们将所设计PGD算法总结在算法5.1中.

5.3 收敛性分析

在本节中, 我们分析PGD算法的全局依子列收敛性. 我们的分析需要如下两个假设条件. 它们在晶体结构弛豫应用中是合理的.

条件5.1. 目标函数 f 局部 Lipschitz 连续可微.

条件5.2. 目标函数 f 在可行域上是强制的(coercive), 即对任意满足 $\det(X) = V$ 的 $X \in \mathbb{R}^{n \times n}$, 随着 $\|X\|_2$ 趋于无穷大, $f(X)$ 也趋于无穷大.

为方便起见, 记

$$\tilde{D} := Y^\top D, \quad (5.10)$$

其中 D 的定义可参见(5.3)式. 对 $k, l \geq 0$, 记

$$\tilde{X}^{(k,l)} := X^{(k)} + \alpha_{k-1,l} D^{(k)} \in \mathbb{R}^{n \times n},$$

则

$$\tilde{X}^{(k,l)} = X^{(k)} (I_n + \alpha_{k-1,l} \tilde{D}^{(k)}). \quad (5.11)$$

算法 5.1: 求解问题 (5.1) 的 PGD 算法.

输入: 初始可行点 $X^{(0)} \in \mathbb{R}^{n \times n}$, 常数 $V > 0$, 初始试探步长上下界

$\bar{\alpha} \geq \underline{\alpha} > 0$, 回溯因子 $\delta \in (0, 1)$.

1 置 $k := 0$.

2 **while** 终止准则未满足 **do**

3 按 (5.3) 式计算搜索方向 $D^{(k)} \in \mathbb{R}^{n \times n}$.

4 选取初始试探步长 $\alpha_{k-1,0} \in [\underline{\alpha}, \bar{\alpha}]$.

5 按 (5.5) 式计算试探点 $X^{(k,0)} \in \mathbb{R}^{n \times n}$.

6 置 $l := 0$.

7 **while** 非单调线搜索准则 (5.4) 不成立 **do**

8 按 (5.7) 式更新试探步长 $\alpha_{k-1,l+1}$.

9 按 (5.5) 式更新试探点 $X^{(k,l+1)} \in \mathbb{R}^{n \times n}$.

10 置 $l := l + 1$.

11 **end**

12 记 $\alpha_k := \alpha_{k-1,l} > 0$, $X^{(k+1)} := X^{(k,l)} \in \mathbb{R}^{n \times n}$.

13 置 $k := k + 1$.

14 **end**

记水平集

$$\mathcal{L}(X^{(0)}) := \left\{ X \in \mathbb{R}^{n \times n} : f(X) \leq f(X^{(0)}), \det(X) = V \right\}. \quad (5.12)$$

我们从条件 5.1 与 5.2 不难得得到下面的引理.

引理 5.3 (水平集的有界性). 假设条件 5.1 与 5.2 成立. 则存在常数 $M > 0$, 只要 $X \in \mathcal{L}(X^{(0)})$, 就有 $\max \{ \|X\|, \|Y\|, \|\tilde{D}\| \} \leq M$.

证明. $\|X\|$ 一致上界的存在性可从条件 5.2 以及水平集 $\mathcal{L}(X^{(0)})$ 的定义 (5.12) 推知. 而 $\|Y\|$ 一致上界的存在性则只需注意到 $\det(X) = V$. 最后, 根据条件 5.1, (5.3) 式以及 (5.10) 式可推得 $\|\tilde{D}\|$ 一致上界的存在性. 证毕. \square

下面, 基于条件 5.1 和 5.2, 我们定义如下常数:

$$\begin{aligned} L_f &:= \sup \left\{ \frac{|f(X_1) - f(X_2)|}{\|X_1 - X_2\|} : \|X_i\| \leq 3M, i = 1, 2 \right\}, \\ L_{\nabla f} &:= \sup \left\{ \frac{\|\nabla f(X_1) - \nabla f(X_2)\|}{\|X_1 - X_2\|} : \|X_i\| \leq \frac{3}{2}M, i = 1, 2 \right\}, \\ \bar{M} &:= \frac{3(2^n - n - 1)}{2 - 1/2^{n-1}} M^3 L_f, \quad \bar{l} := \left\lceil \max \left(\frac{\ln \frac{(L_{\nabla f} + 2\bar{M})\bar{\alpha}}{2(1-\eta)}}{\ln \frac{1}{\delta}}, \frac{\ln(2M\bar{\alpha})}{\ln \frac{1}{\delta}} \right) \right\rceil. \end{aligned} \quad (5.13)$$

在分析 PGD 算法的全局依子列收敛性前, 我们先证明回溯搜索 (5.7) 总是有限终止的, 步长序列 $\{\alpha_k\}$ 存在正下界, PGD 生成的迭代点序列 $\{X^{(k)}\}$ 一直在 $\mathcal{L}(X^{(0)})$ 中.

引理 5.4 (PGD 算法线搜索过程的有限终止性). 假设条件 5.1 与 5.2 成立, $X^{(k)} \in \mathcal{L}(X^{(0)})$. 则回溯搜索 (5.7) 至多在 \bar{l} 步后终止, 步长有下界

$$\alpha_k \geq \beta := \underline{\alpha} \delta^{\bar{l}}, \quad (5.14)$$

其中 \bar{l} 定义在 (5.13) 式中, 新的迭代点 $X^{(k+1)} \in \mathcal{L}(X^{(0)})$.

证明. 我们证明, 只要 $l \geq \bar{l}$, 则 $f(X^{(k,l)})$ 必定充分小于 $f(X^{(k)})$, 从而非单调准则 (5.4) 成立. 我们分两步进行证明.

第一步. 估计 $f(\tilde{X}^{(k,l)}) - f(X^{(k)})$.

根据 (5.5) 式,

$$\begin{aligned} \langle \nabla f(X^{(k)}), \tilde{X}^{(k,l)} - X^{(k)} \rangle &= \langle D^{(k)}, \tilde{X}^{(k,l)} - X^{(k)} \rangle \\ &= -\frac{1}{\alpha_{k-1,l}} \|\tilde{X}^{(k,l)} - X^{(k)}\|^2. \end{aligned} \quad (5.15)$$

因为 $l \geq \bar{l}$, 不难验证 $\alpha_{k-1,l} \leq 1/(2M)$. 由此, 引理 5.3 以及 (5.11) 式, 可知

$$\alpha_{k-1,l} \|\tilde{D}^{(k)}\|_2 \leq \frac{1}{2}, \quad \|\tilde{X}^{(k,l)}\| \leq \|X^{(k)}\| (1 + \alpha_{k-1,l} \|\tilde{D}^{(k)}\|_2) \leq \frac{3}{2}M. \quad (5.16)$$

根据 (5.13) 式中 $L_{\nabla f}$ 的定义, 我们可从 (5.15) 与 (5.16) 式推得

$$f(\tilde{X}^{(k,l)}) - f(X^{(k)}) \leq \left(\frac{L_{\nabla f}}{2} - \frac{1}{\alpha_{k-1,l}} \right) \|\tilde{X}^{(k,l)} - X^{(k)}\|^2. \quad (5.17)$$

第二步: 估计 $f(X^{(k,l)}) - f(\tilde{X}^{(k,l)})$.

首先由 (5.11) 式,

$$\det(\tilde{X}^{(k,l)}) = V \det(I_n + \alpha_{k-1,l} \tilde{D}^{(k)}).$$

再根据 (5.16) 式以及矩阵任意特征值的模均不超过其最大奇异值,

$$\det(\tilde{X}^{(k,l)}) \in \left[\frac{1}{2^n}V, \frac{3^n}{2^n}V \right]. \quad (5.18)$$

借助缩放算子的定义 (5.2) 以及 (5.16) 与 (5.18) 式, $\|X^{(k,l)}\| \leq 3M$. 于是根据 (5.13) 式中 L_f 的定义,

$$\begin{aligned} f(X^{(k,l)}) - f(\tilde{X}^{(k,l)}) &\leq L_f \|X^{(k,l)} - \tilde{X}^{(k,l)}\| \\ &= L_f \left| \sqrt[n]{\det(\tilde{X}^{(k,l)})} - \sqrt[n]{V} \right| \frac{\|\tilde{X}^{(k,l)}\|}{\sqrt[n]{\det(\tilde{X}^{(k,l)})}}. \end{aligned} \quad (5.19)$$

一方面, 根据 (5.16) 与 (5.18) 式,

$$\frac{\|\tilde{X}^{(k,l)}\|}{\sqrt[n]{\det(\tilde{X}^{(k,l)})}} \leq \frac{3M}{\sqrt[n]{V}}. \quad (5.20)$$

另一方面, 再次由 (5.18) 式,

$$\begin{aligned} \left| \sqrt[n]{\det(\tilde{X}^{(k,l)})} - \sqrt[n]{V} \right| &= \frac{|\det(\tilde{X}^{(k,l)}) - V|}{\sum_{i=0}^{n-1} \det(\tilde{X}^{(k,l)})^{i/n} V^{(n-1-i)/n}} \\ &\leq \frac{|\det(\tilde{X}^{(k,l)}) - V|}{V^{(n-1)/n} \sum_{i=0}^{n-1} 1/2^i} = \frac{|\det(\tilde{X}^{(k,l)}) - V|}{V^{(n-1)/n} (2 - 1/2^{n-1})}. \end{aligned} \quad (5.21)$$

令 $\lambda_{k,1}, \dots, \lambda_{k,n}$ 为 $\alpha_{k-1,l} \tilde{D}^{(k)}$ 的特征值. 因为 $D^{(k)} \in \mathcal{T}^{(k)}$, 所以由 $\mathcal{T}^{(k)}$ 的定义, $\sum_{i=1}^n \lambda_{k,i} = 0$. 因此,

$$|\det(\tilde{X}^{(k,l)}) - V| = V |\det(I_n + \alpha_{k-1,l} \tilde{D}^{(k)}) - 1| = V \left| \sum_{j=2}^n \sum_{i_1, \dots, i_j \in \{1, \dots, n\} \text{ 两两不等}} \prod_{t=1}^j \lambda_{k,i_t} \right|,$$

其中第一个等式使用了 (5.11) 式. 注意到根据 (5.16) 式,

$$\max_{i=1}^n |\lambda_{k,i}| \leq \alpha_{k-1,l} \|\tilde{D}^{(k)}\|_2 \leq \frac{1}{2},$$

于是

$$\begin{aligned} |\det(\tilde{X}^{(k,l)}) - V| &\leq V(2^n - n - 1) \max_{i=1}^n |\lambda_{k,i}|^2 \\ &\leq V(2^n - n - 1) \|\alpha_{k-1,l} \tilde{D}^{(k)}\|_2^2 \\ &= V(2^n - n - 1) \|Y^{(k)\top} (\tilde{X}^{(k,l)} - X^{(k)})\|_2^2 \\ &\leq M^2 V(2^n - n - 1) \|\tilde{X}^{(k,l)} - X^{(k)}\|^2, \end{aligned} \quad (5.22)$$

其中等式使用了 (5.11) 式以及 $Y^{(k)}$ 的定义, 最后一个不等式使用了引理 5.3. 结合 (5.13)、(5.19)–(5.22) 式, 我们就有

$$f(X^{(k,l)}) - f(\tilde{X}^{(k,l)}) \leq \bar{M} \|\tilde{X}^{(k,l)} - X^{(k)}\|^2. \quad (5.23)$$

最后, 根据 (5.17) 与 (5.23) 式,

$$f(X^{(k,l)}) - f(X^{(k)}) \leq \left[\frac{L_{\nabla f}}{2} + \bar{M} - \frac{1}{\alpha_{k-1,l}} \right] \|\tilde{X}^{(k,l)} - X^{(k)}\|^2.$$

由非单调准则 (5.4) 与 $f(X^{(k)}) \leq \bar{f}_k$, 线搜索过程必定在

$$\frac{L_{\nabla f}}{2} + \bar{M} - \frac{1}{\alpha_{k-1,l}} \leq -\frac{\eta}{\alpha_{k-1,l}}$$

成立时终止. 根据 (5.13) 式, $l \geq \bar{l}$ 即可满足条件. 因此, 步长有下界 (5.14) 且新迭代点 $X^{(k+1)} \in \mathcal{L}(X^{(0)})$. \square

借助引理 5.4, 我们可以证明 PGD 算法的全局依子列收敛性.

定理 5.5 (PGD 算法的全局依子列收敛性). 假设条件 5.1 与 5.2 成立. 令 $\{X^{(k)}\}$ 为 PGD 算法产生的迭代点序列, 其中 $\mu_k \in [\mu_{\min}, \mu_{\max}] \subseteq (0, 1]$. 则 $\{X^{(k)}\}$ 至少存在一个聚点, 且其任何一个聚点都是问题 (5.1) 的 KKT 点.

证明. 将 (5.14) 式带入非单调准则 (5.4), 即有

$$f(X^{(k+1)}) \leq \bar{f}_k - \eta \beta \|D^{(k)}\|^2. \quad (5.24)$$

令 $E_k := \beta \|D^{(k)}\|^2$. 于是代理序列 $\{\bar{f}_k\}$ 满足如下递归关系:

$$\bar{f}_{k+1} = \frac{\bar{f}_k + \mu_k q_k f(X^{(k+1)})}{q_{k+1}} \leq \frac{\bar{f}_k + \mu_k q_k (\bar{f}_k - \eta E_k)}{q_{k+1}} = \bar{f}_k - \eta \mu_k \frac{q_k}{q_{k+1}} E_k, \quad (5.25)$$

其中不等式使用了 (5.24) 式. 因此 $\{\bar{f}_k\}$ 是单调下降的. 根据条件 5.1 和引理 5.3, f 在 $\mathcal{L}(X^{(0)})$ 上下有界. 因为对任意 $k \geq 0$, $f(X^{(k)}) \leq \bar{f}_k$, 所以 $\{\bar{f}_k\}$ 也下有界. 如此一来, 根据 (5.25) 式, 对任意 $\hat{k} \in \mathbb{N}$,

$$\sum_{k=0}^{\hat{k}} \mu_k \frac{q_k}{q_{k+1}} E_k \leq \frac{1}{\eta} \sum_{k=0}^{\hat{k}} (\bar{f}_k - \bar{f}_{k+1}) \leq \frac{1}{\eta} \left(f(X^{(0)}) - \inf_k \bar{f}_k \right) < \infty.$$

于是 $\{\mu_k q_k E_k / q_{k+1}\}$ 可和. 另外, 由代理序列的定义 (5.6) 以及 $\mu_{\max} \geq \mu_k \geq \mu_{\min} > 0$, $q_k \geq 1$, 我们有

$$\mu_k \frac{q_k}{q_{k+1}} \geq \mu_{\min} \frac{q_k}{\mu_k q_k + 1} = \frac{\mu_{\min}}{\mu_k + 1/q_k} \geq \frac{\mu_{\min}}{\mu_{\max} + 1}.$$

于是 $\{E_k\}$ 可和, 从而随着 k 趋于无穷大, $\|D^{(k)}\|$ 趋于 0.

令 $X^* \in \mathbb{R}^{n \times n}$ 为 $\{X^{(k)}\}$ 的一个聚点, 其存在性由引理 5.3 保证. 于是, 根据 (5.3) 式,

$$0 = D^* = -\nabla f(X^*) + \frac{\langle Y^*, \nabla f(X^*) \rangle}{\|Y^*\|^2} Y^*.$$

将 $\langle Y^*, \nabla f(X^*) \rangle / \|Y^*\|^2$ 视作对应于行列式约束的 Lagrange 乘子, 即知 X^* 是问题 (5.1) 的 KKT 点. \square

5.4 数值实验

我们首先将 PGD 算法推广至求解第一性原理固定晶格体积晶体结构弛豫问题 (1.37). 为比较 CG 算法、QN 算法与 PGD 算法的性能与鲁棒性, 我们构建了包含 223 个来自不同类别的结构的基准算例集. 作为一个具体的应用, 我们还使用 PGD 算法计算了难弛豫合金 AlCoCrFeNi 的物态方程, 并将计算结果与已有实验测定数据进行对比.

本章所有的数值实验均在中国科学院科学与工程计算国家重点实验室高性能计算机系统 LSSC-IV³上完成。LSSC-IV 的主体部分包含 408 个节点，操作系统为 Red Hat Enterprise Linux Server 7.3。每个节点包含两颗 Intel Xeon Gold 6140 CPU (2.30 GHz × 18)，内存为 192 GB。所有算法均在材料模拟软件 Correlated Electron System Simulation Package (CESSP)^[287–290] 上实现，采用 Fortran 90 编写，使用 Intel oneAPI 编译。

5.4.1 求解固定晶格体积晶体结构弛豫问题的投影梯度下降算法

在这一部分，我们基于算法 5.1，介绍求解问题 (1.37) 的 PGD 算法。在第 $k+1$ 次迭代中，原子位置 R 与晶格基矢 A 更新分别使用搜索方向

$$D_{\text{atom}}^{(k)} := F_{\text{atom}}^{(k)}, \quad D_{\text{latt}}^{(k)} := \mathcal{P}_{\mathcal{T}^{(k)}}(F_{\text{latt}}^{(k)}), \quad (5.26)$$

其中 $\mathcal{P}_{\mathcal{T}^{(k)}}$ 的定义可见 (5.3) 式。在第一次迭代 ($k=0$) 中，我们采用人为选取的初始试探步长 $\alpha_{\text{atom}}^{(-1,0)} > 0$ 与 $\alpha_{\text{latt}}^{(-1,0)} > 0$ ；而当 $k > 0$ 时，我们使用截断的交替 Barzilai-Borwein (ABB) 步长作为初始试探步长 $\alpha_{\text{atom}}^{(k-1,0)} > 0$ 与 $\alpha_{\text{latt}}^{(k-1,0)} > 0$ 。具体地，对于 $k > 0$ ，

$$\alpha_{\text{atom}}^{(k-1,0)} = \max \left\{ \min \left\{ \left| \alpha_{\text{atom}, \text{ABB}}^{(k)} \right|, \tau_{\text{atom}}, \bar{\alpha}_{\text{atom}} \right\}, \underline{\alpha}_{\text{atom}} \right\}, \quad (5.27)$$

$$\alpha_{\text{latt}}^{(k-1,0)} = \max \left\{ \min \left\{ \left| \alpha_{\text{latt}, \text{ABB}}^{(k)} \right|, \tau_{\text{latt}}, \bar{\alpha}_{\text{latt}} \right\}, \underline{\alpha}_{\text{latt}} \right\}, \quad (5.28)$$

其中，对于原子位置，

$$\begin{aligned} \alpha_{\text{atom}, \text{ABB}}^{(k)} &:= \begin{cases} \alpha_{\text{atom}, \text{BB1}}^{(k)}, & \text{若 } \text{mod}(k, 2) = 0; \\ \alpha_{\text{atom}, \text{BB2}}^{(k)}, & \text{若 } \text{mod}(k, 2) = 1, \end{cases} \\ \alpha_{\text{atom}, \text{BB1}}^{(k)} &:= \frac{\left\| S_{\text{atom}}^{(k-1)} \right\|^2}{\left\langle S_{\text{atom}}^{(k-1)}, Y_{\text{atom}}^{(k-1)} \right\rangle}, \quad \alpha_{\text{atom}, \text{BB2}}^{(k)} := \frac{\left\langle S_{\text{atom}}^{(k-1)}, Y_{\text{atom}}^{(k-1)} \right\rangle}{\left\| Y_{\text{atom}}^{(k-1)} \right\|^2}, \end{aligned}$$

$S_{\text{atom}}^{(k-1)} := R^{(k)} - R^{(k-1)} \in \mathbb{R}^{3 \times M}$, $Y_{\text{atom}}^{(k-1)} := F_{\text{atom}}^{(k-1)} - F_{\text{atom}}^{(k)} \in \mathbb{R}^{3 \times M}$, $\tau_{\text{atom}}^{(k)}$ 为自适应更新的截断系数 (选取方式可见后文)， $\bar{\alpha}_{\text{atom}} > 0$ 与 $\underline{\alpha}_{\text{atom}} > 0$ 分别是预设的原子位置初始试探步长的上下界。对于晶格基矢， $\alpha_{\text{latt}, \text{ABB}}^{(k)}$ 与 $\tau_{\text{latt}}^{(k)}$ 的定义是类似的，只需将其中的 R 换成 A , F_{atom} 换成 D_{latt} 。确定好搜索方向与试探步长后，按如下方式计算试探构型：

$$R^{(k,l)} := R^{(k)} + \alpha_{\text{atom}}^{(k-1,l)} D_{\text{atom}}^{(k)}, \quad A^{(k,l)} := \mathcal{P}_{\mathcal{V}}(A^{(k)} + \alpha_{\text{latt}}^{(k-1,l)} D_{\text{latt}}^{(k)}), \quad (5.29)$$

并检查是否满足非单调线搜索准则

$$E(R^{(k,l)}, A^{(k,l)}) \leq \bar{E}_k - \eta \left(\alpha_{\text{atom}}^{(k-1,l)} \left\| D_{\text{atom}}^{(k)} \right\|^2 + \alpha_{\text{latt}}^{(k-1,l)} \left\| D_{\text{latt}}^{(k)} \right\|^2 \right), \quad (5.30)$$

³LSSC-IV 简介: <http://lsec.cc.ac.cn/chinese/lsec/LSSC-IVintroduction.pdf>.

算法 5.2: 求解问题 (1.37) 的 PGD 算法.

输入: 晶格体积 $V > 0$, 初始构型 $(R^{(0)}, A^{(0)}) \in \mathbb{R}^{3 \times M} \times \mathbb{R}^{3 \times 3}$: $\det(A^{(0)}) = V$,

原子位置初始试探步长界 $\bar{\alpha}_{\text{atom}} \geq \underline{\alpha}_{\text{atom}} > 0$, 晶格基矢初始试探步长界 $\bar{\alpha}_{\text{latt}} \geq \underline{\alpha}_{\text{latt}} > 0$, 回溯因子 $\delta_{\text{atom}}, \delta_{\text{latt}} \in (0, 1)$.

```

1 置  $k := 0$ .
2 while 终止准则未满足 do
3   按 (5.26) 式计算搜索方向  $D_{\text{atom}}^{(k)} \in \mathbb{R}^{3 \times M}$  与  $D_{\text{latt}}^{(k)} \in \mathbb{R}^{3 \times 3}$ .
4   if  $k = 0$  then
5     | 选取初始试探步长  $\alpha_{\text{atom}}^{(-1,0)} \in [\underline{\alpha}_{\text{atom}}, \bar{\alpha}_{\text{atom}}]$  与  $\alpha_{\text{latt}}^{(-1,0)} \in [\underline{\alpha}_{\text{latt}}, \bar{\alpha}_{\text{latt}}]$ .
6   else
7     | 按 (5.27) 与 (5.28) 式分别计算初始试探步长  $\alpha_{\text{atom}}^{(k-1,0)}$  与  $\alpha_{\text{latt}}^{(k-1,0)}$ .
8   end
9   按 (5.29) 式计算试探构型  $(R^{(k,0)}, A^{(k,0)}) \in \mathbb{R}^{3 \times M} \times \mathbb{R}^{3 \times 3}$ .
10  置  $l := 0$ .
11  while 非单调线搜索准则 (5.30) 不成立 do
12    | 按 (5.31) 式更新试探步长  $\alpha_{\text{atom}}^{(k-1,l+1)}$  与  $\alpha_{\text{latt}}^{(k-1,l+1)}$ .
13    | 按 (5.29) 式更新试探构型  $(R^{(k,l+1)}, A^{(k,l+1)}) \in \mathbb{R}^{3 \times M} \times \mathbb{R}^{3 \times 3}$ .
14    | 置  $l := l + 1$ .
15  end
16  记  $\alpha_{\text{atom}}^{(k)} := \alpha_{\text{atom}}^{(k-1,l)} > 0$ ,  $\alpha_{\text{latt}}^{(k)} := \alpha_{\text{latt}}^{(k-1,l)} > 0$ .
17  记  $(R^{(k+1)}, A^{(k+1)}) := (R^{(k,l)}, A^{(k,l)}) \in \mathbb{R}^{3 \times M} \times \mathbb{R}^{3 \times 3}$ .
18  置  $k := k + 1$ .
19 end
```

其中, 代理能量序列 $\{\bar{E}_k\}$ 的定义类似于 (5.6) 式, $\eta > 0$ 为一给定常数. 若试探构型 $(R^{(k,l)}, A^{(k,l)})$ 不满足准则 (5.30), 则使用回溯法减小步长:

$$\alpha_{\text{atom}}^{(k-1,l+1)} := \delta_{\text{atom}} \cdot \alpha_{\text{atom}}^{(k-1,l)}, \quad \alpha_{\text{latt}}^{(k-1,l+1)} := \delta_{\text{latt}} \cdot \alpha_{\text{latt}}^{(k-1,l)}, \quad l \geq 0, \quad (5.31)$$

并再按 (5.29) 式重新计算试探构型. 这里 $\delta_{\text{atom}}, \delta_{\text{latt}} \in (0, 1)$ 是给定回溯因子.

我们将上面描述的 PGD 算法总结在算法 5.2 中.

注. 在上述算法 5.2 中, 我们为原子位置、晶格基矢的更新分别使用不同的搜索方向和步长. 这是因为势能面对于这两块变量的曲率有着显著差异. 直观上, 在构型弛豫的过程中, 晶格基矢的相对变化应当小于原子位置中的变化. 在较大规模的体系上, 这一现象会尤其显著. 因此, 算法 5.2 可理解为“预条件”的 PGD 算法.

注. 经过简单改动并增加能量泛函对原子位置的光滑性与强制性, 第 5.3 节的收敛性分析仍适用于算法 5.2.

5.4.2 实验设置

在数值实验中, 电子结构计算部分采用 KSDFT 模型 (1.12), 交换–关联泛函近似选用广义梯度近似^[291], 电子–原子核相互作用能使用投影缀加波函数^[292]处理, 倒空间采样使用 Monkhorst-Pack 网格^[293] (K 点间隔距离为 $0.12 \sim 0.20 \text{ \AA}^{-1}$), 平面波能量截断为 $500 \sim 600 \text{ eV}$, 能量曲面信息通过使用预条件自洽场迭代^[289]求解 KS 方程 (1.13) 得到.

我们在软件 CESSP 中实现了 PGD 算法. 其第一次迭代的初始试探步长选取为 $\alpha_{\text{atom}}^{(-1,0)} = 4.8 \times 10^{-2} \text{ \AA}^2/\text{eV}$ 与 $\alpha_{\text{latt}}^{(-1,0)} = 10^{-6} \text{ \AA}^2/\text{eV}$. 截断系数

$$\tau_{\text{atom}}^{(k)} := \gamma_{\text{atom}}^{(k)} \max \left\{ -\log_{10} \left(\|F_{\text{atom}}^{(k)}\| / M \right), 1 \right\},$$

$\gamma_{\text{atom}}^{(k)}$ 在 $k = 0$ 时取 1, 之后根据之前的迭代更新:

$$\gamma_{\text{atom}}^{(k)} := \begin{cases} 2\gamma_{\text{atom}}^{(k-1)}, & \text{情形一;} \\ 0.5\gamma_{\text{atom}}^{(k-1)}, & \text{情形二;} \\ \gamma_{\text{atom}}^{(k-1)}, & \text{否则,} \end{cases}$$

其中情形一指“在过去 20 次迭代中, 累计两次 $\tau_{\text{atom}}^{(k)}$ 起作用但 $\alpha_{\text{atom}}^{(k-1,0)}$ 直接满足非单调准则 (5.30)”, 情形二指“在过去 20 次迭代中累计两次 $\alpha_{\text{atom}}^{(k-1,0)}$ 无法满足非单调准则 (5.30)”. 类似可定义 $\tau_{\text{latt}}^{(k)}$, 其在 $k = 0$ 时取 10^{-3} . 只要截断系数发生变化, 我们就清空对应的历史记录. 初始试探步长的上下界设置为 $\bar{\alpha}_{\text{atom}} = 10$, $\underline{\alpha}_{\text{atom}} = 10^{-5}$, $\bar{\alpha}_{\text{latt}} = 10^{-1}$, $\underline{\alpha}_{\text{latt}} = 10^{-7}$. 在非单调线搜索准则 (5.30) 中, 取 $\eta = 10^{-4}$, $\mu_k \equiv 0.05$. 回溯因子则取为 $\delta_{\text{atom}} = 0.1$, $\delta_{\text{latt}} = 0.5$. 除了 PGD 算法, 我们还测试了软件 CESSP 中的 CG 算法与 QN 算法 (分别和 VASP 中的 CG 算法与 QN 算法的实现一致).

注. 我们依照软件 CESSP 中对 CG 算法的设置选取第一次迭代的初始试探步长. 截断系数的选取主要遵从如下原则: 若在最近的迭代中, 其对于步长频繁起作用, 则适当增大其数值; 反之, 则减小其数值或保持其不变. 基于这样的原则, 截断系数可在一定程度上缩短线搜索的过程. 初始试探步长上下界的存在主要是为了保证 PGD 算法的理论收敛性. 我们选取的上下界在实际计算中很少起作用. 非单调线搜索准则中参数的选取主要依据文献^[285]. 我们选取较小的原子位置回溯因子 δ_{atom} 以尽快结束线搜索过程并帮助接下来计算的 ABB 步长获取足够精确的局部曲率信息. 晶格基矢回溯因子 δ_{latt} 的取值不太影响 PGD 算法的数值表现, 因为 $\alpha_{\text{latt}}^{(k-1,0)}$ 通常已经足够小了.

我们在如下两个条件之一满足时终止弛豫算法: (1) KS 方程的求解次数 (#KS) 超过 1000; (2) 力终止准则

$$\begin{aligned} \|F_{\text{atom}}^{(k)}\|_{2,\infty} &:= \max_{j=1}^M \|F_{\text{atom},\cdot,j}^{(k)}\| < 0.01 \text{ eV/\AA}, \\ \|\Sigma_{\text{dev}}^{(k)}\|_{\infty} &:= \max_{i,j} |\sigma_{\text{dev},ij}^{(k)}| < M \times 0.01 \text{ eV/\AA}, \end{aligned}$$

表 5.1 第一性原理固定晶格体积晶体结构弛豫算法基准测试算例集信息

Table 5.1 Information of the benchmark set for testing the methods of *ab initio* crystal structure relaxation under a fixed unit cell volume

结构类别	结构数	包含结构举例
金属	159	金属团簇、合金、金属氧化物
有机分子	25	氨基酸、烷烃、有机盐
半导体	21	砷化镓、砷化铟、体硅
钙钛矿	7	对甲胺三碘化铅、溴化铅甲胺
吸附表面	5	硫化铅-油酸分子、二氧化钚-水分子
异质结	3	砷化镓-砷化铟
二维材料	3	硅、碳、锗-氮化硼
总计	223	

其中 $\sigma_{\text{dev},ij}^{(k)}$ 是 $\Sigma_{\text{dev}}^{(k)}$ 的第 i 行第 j 列元素 ($i, j = 1, 2, 3$). 在迭代次数超过 100 或相邻两次迭代的能量差小于 10^{-5} eV 时, 我们终止自洽场迭代. 为评估与比较弛豫算法的性能, 我们记录它们为满足力终止准则所需的 #KS 与运行时间 (秒) (CPU).

5.4.3 基准算例集上的测试

为测试 CG 算法、QN 算法与 PGD 算法, 我们构建了含有 223 个来自不同类别结构的基准算例集 (包含初始构型), 其中材料类别涵盖金属、吸附表面、异质结、半导体、钙钛矿、二维材料、有机分子七大类体系, 金属材料结构数占比超 70%. 这是因为第一性原理固定晶格体积晶体结构弛豫的一个主要应用就是金属材料物态方程的计算. 我们在表 5.1 中列出了基准算例集的基本信息. 结构中原子的个数从 2 到 215 不等. 其中一些结构来源于 Materials Project^[294] 与 Organic Materials Database^[295].

我们在所构建的基准算例集上测试 CG 算法、QN 算法与 PGD 算法. 为从整体上比较它们的数值表现, 我们采用性能剖面 (performance profile)^[296] (见图 5.1).

从图 5.1 不难看出, 在基准算例集上, PGD 算法对于 CG 算法与 QN 算法具有显著且普遍的效率优势. 具体地, PGD 算法分别在 85.2% 与 68.8% 的结构上效率优于 CG 算法与 QN 算法, 平均 CPU 加速比分别为 1.41 与 1.45. 我们还分结构类别统计了 PGD 算法对于 CG 算法的平均加速比 (见图 5.2). 其中, 平均加速最显著的三个类别是有机分子、钙钛矿和金属. 这些类别结构上的平均 CPU 加速比分别为 1.67、1.50 和 1.40.

除了效率优势, PGD 算法还表现出了更好的鲁棒性. 在基准测试中, PGD 算法在所有结构上均正常收敛, CG 算法则在 11 个结构上因搜索方向非下降无法搜寻到合适的步长而崩溃, QN 算法则在 56 个结构上发散. 理论上, PGD 算法的全局依子列收敛性由定理 5.5 保证.

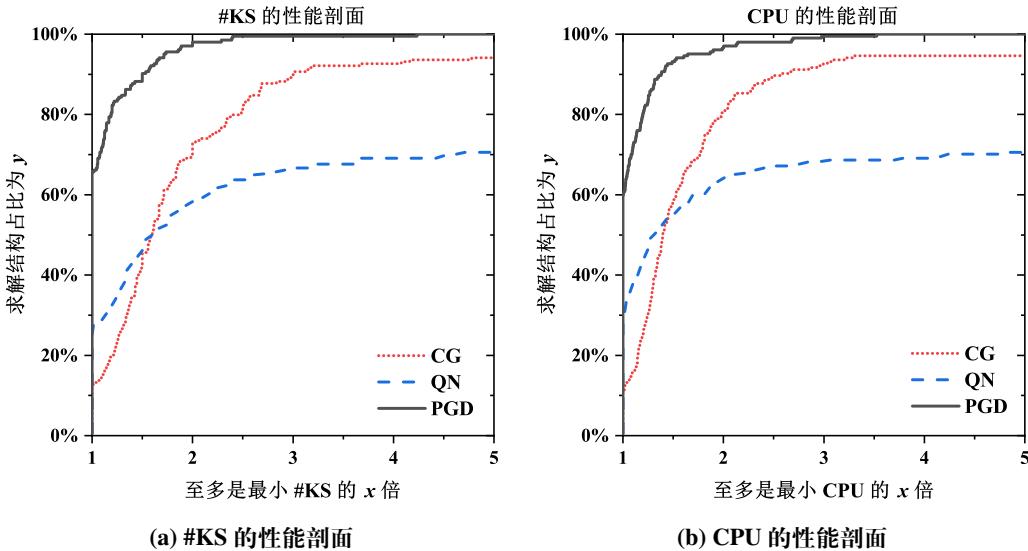


图 5.1 CG 算法、QN 算法与 PGD 算法在基准算例集上的性能剖面. 为比较公平性, 我们(1)剔除三个算法收敛到的最大与最小能量之差超过每原子 3 meV 的结构; (2)保留至少有一个算法无法在 #KS 超过 1000 之前终止的结构, 并将其 #KS 与 CPU 置为无穷大. 最终, 用于绘制性能剖面的结构数为 204. 红色点线、蓝色虚线与黑色实线分别表示 CG 算法、QN 算法与 PGD 算法的结果. (a) #KS 的性能剖面. (b) CPU 的性能剖面

Figure 5.1 The performance profiles of the CG, QN, and PGD methods. To reach a fair comparison, we (1) screen out the systems where the differences between the maximum and minimum of the converged energies per atom given by the three methods are larger than 3 meV; (2) retain the systems on which at least one method fails to converge before #KS exceeds 1000. The final number of the systems used for the performance profiles is 204. Red dotted, blue dashed, and black solid lines stand for the results of the CG, QN, and PGD methods, respectively. (a) The performance profile of #KS. (b) The performance profile of CPU

5.4.4 高熵合金物态方程计算

作为第一性原理固定晶格体积晶体结构弛豫的一个重要应用, 我们使用 PGD 算法的计算结果拟合体心立方高熵合金 AlCoCrFeNi 的静态三阶 Birch-Murnaghan (BM3) 物态方程^[297,298]:

$$E(V) = E_0 + \frac{9V_0B_0}{16} \left\{ \left[\left(\frac{V_0}{V} \right)^{2/3} - 1 \right]^3 B'_0 + \left[\left(\frac{V_0}{V} \right)^{2/3} - 1 \right]^2 \left[6 - 4 \left(\frac{V_0}{V} \right)^{2/3} \right] \right\},$$

其中, E_0 、 V_0 、 B_0 与 B'_0 分别为待拟合的平衡能量、平衡晶格体积、平衡体模量与平衡体模量对压强的导数. 已有实验表明, AlCoCrFeNi 具有优异的低温力学性质^[299]. 我们还会基于得到的静态 BM3 物态方程计算结构的热力学性质, 并通过与已有实验测定数据对比验证其正确性.

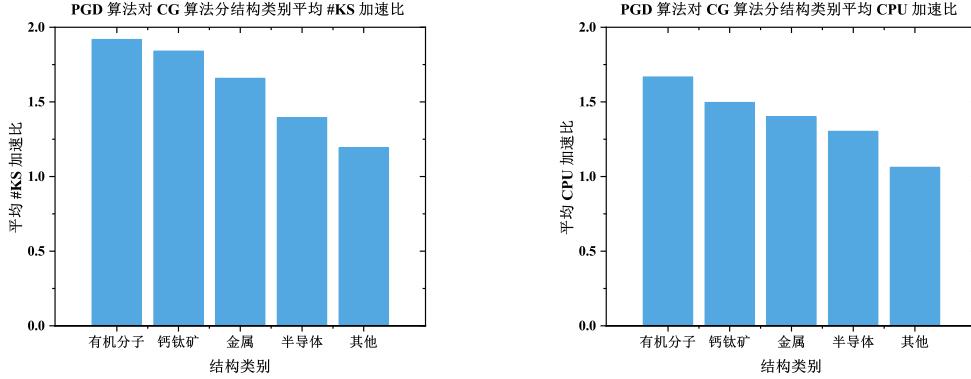


图 5.2 PGD 算法对 CG 算法分结构类别平均加速比. 为比较公平性, 我们仅保留二者均正常收敛且收敛到的能量差绝对值不超过每原子 3 meV 的结构. 此外, 我们将结构数少于 5 的类别统一并入“其他”中. 左图: 分结构类别平均 #KS 加速比. 右图: 分结构类别平均 CPU 加速比

Figure 5.2 The average speedup factors of the PGD method over the CG method by the system category. To achieve a fair comparison, we only include the systems where both methods converge normally and the converged energy differences per atom are not larger than 3 meV. Categories whose numbers of the included systems are less than 5 are merged together into “Other”. Left: Average #KS speedup factor by system category. Right: Average CPU speedup factor by system category

5.4.4.1 初始构型建模

在计算结构物态方程时, 我们通常会在一定范围内变动晶格体积, 并要求在不同体积点处弛豫得到的单胞几何结构(轴长比、轴角等)相差不大. 这样得到的物态方程才可以反映结构的某个相的性质. 由于无磁铝元素的掺杂, 在弛豫过程中, AlCoCrFeNi 结构往往会发生很强的局部晶格畸变^[300]. 这使其物态方程的计算变得异常困难. 由于目前暂无弛豫算法可以保证单胞几何结构不发生巨大改变, 因此我们需要首先为 AlCoCrFeNi 构建合理的初始构型, 尽可能减小局部晶格畸变带来的影响. 简单来说, 这需要初始构型足够无序.

为刻画由温度带来的磁结构转变, 我们需要构建 AlCoCrFeNi 的铁磁(ferromagnetic, FM)态与顺磁(paramagnetic, PM)态初始构型, 分别对应化学无序与磁无序的建模. 为此, 我们结合相似原子环境(similar atomic environment, SAE)方法^[301]与整数规划^[302]构建 $4 \times 4 \times 5$ 的体心立方超胞(supercell), 其含 160 个原子. 具体地, 我们首先用 SAE 方法建模化学无序. SAE 方法的大致想法是通过 Monte Carlo 采样极小化如下评价函数决定原子的排布:

$$\sum_{d(C_2) < r_2^c} w_2 e(C_2, \sigma) + \sum_{d(C_3) < r_3^c} w_3 e(C_3, \sigma),$$

其中 σ 是待优化的原子占位, $w_n > 0$ ($n = 2, 3$) 是权重因子, $r_n^c > 0$ ($n = 2, 3$) 是截断半径, $e(C_2, \sigma)$ 和 $e(C_3, \sigma)$ 分别代表双原子团簇 C_2 和三原子团簇 C_3 的相似度函

数, $d(C_2)$ 和 $d(C_3)$ 代表团簇的直径. 我们使用默认设置的 SAE 方法, 将第一和第二部分分别优化至 0.06 和 0.18.

接着, 我们建模磁无序. 由于上下旋的同一种元素需区别对待, 应用 SAE 方法将涉及 9 种元素构型空间的优化. 这会导致较长的计算过程. 因此, 我们转而使用整数规划, 在之前 SAE 方法生成的原子排布的基础上将上下旋均匀地分配给单胞中的磁性元素. 具体来说, 我们求解如下整数规划:

$$\begin{aligned} & \max_{\mathbf{y}} 0, \\ \text{s. t. } & [M_s/(2\ell_{\text{ax}})] \leq \sum_{i \in \mathcal{I}_{\text{ax},l,s}} y_i \leq [M_s/(2\ell_{\text{ax}})], \\ & l \in \{1, \dots, \ell_{\text{ax}}\}, \text{ax} \in \{\text{a, b, c}\}, s \in \{1, \dots, S_{\text{mag}}\}, \\ & \sum_{i \in \mathcal{I}_s} y_i = M_s/2, s \in \{1, \dots, S_{\text{mag}}\}, \\ & \mathbf{y} \in \{0, 1\}^{M_{\text{mag}}}. \end{aligned} \quad (5.32)$$

这里, $S_{\text{mag}} \in \mathbb{N}$ 表示磁元素个数 (对于 AlCoCrFeNi, $S_{\text{mag}} = 4$). 为简单起见, 我们假设第 $1, \dots, S_{\text{mag}}$ 种元素是磁元素. $M_s \in \mathbb{N}$ 表示第 s 种元素的原子个数 ($s \in \{1, \dots, S_{\text{mag}}\}$) (在本例中, $M_1 = \dots = M_4 = 32$). $M_{\text{mag}} \in \mathbb{N}$ 表示单胞中上旋磁原子的个数 (在本例中, $M_{\text{mag}} = 64$). $\ell_{\text{ax}} \in \mathbb{N}$ 表示单胞中以 ax 轴为法向的原子层数 ($\text{ax} \in \{\text{a, b, c}\}$). $\mathcal{I}_s \subseteq \{1, \dots, M\}$ 为第 s 种元素原子的指标集 ($s \in \{1, \dots, S_{\text{mag}}\}$). $\mathcal{I}_{\text{ax},l,s} \subseteq \{1, \dots, M\}$ 表示单胞中以 ax 轴为法向的第 l 个原子层中第 s 种元素原子的指标集 ($\text{ax} \in \{\text{a, b, c}\}, l \in \{1, \dots, \ell_{\text{ax}}\}, s \in \{1, \dots, S_{\text{mag}}\}$). $\mathbf{y} \in \{0, 1\}^{M_{\text{mag}}}$ 是表征上下旋元素分布的 0-1 变量: $y_i = 1$ 表示第 i 个原子上旋, $y_i = 0$ 表示第 i 个原子下旋 ($i \in \{1, \dots, M_{\text{mag}}\}$). 第一行约束限制每类元素的上(下)旋位点均匀地分布在每个原子层. 第二行约束限制每类元素的上旋与下旋位点数相等. 整数规划 (5.32) 可直接使用 MATLAB 自带函数 “intlinprog” 求解.

我们将使用 SAE 方法与整数规划建模的初始构型展示在图 5.3 中. 为说明所构建的初始磁构型的合理性, 我们在图 5.3 (d) 中绘制了磁元素的径向分布函数^[303], 其表明上旋-上旋、上旋-下旋以及下旋-下旋原子对的密度随原子对距离变化的趋势是相似的.

5.4.4.2 物态方程计算

为计算 AlCoCrFeNi 的 FM 态与 PM 态的静态 BM3 物态方程, 我们在 Wigner-Seitz 半径^[304] 范围 $[2.45\text{\AA}, 2.70\text{\AA}]$ 内按等间距选取了 18 个晶格体积点 (可见表 5.2). 对于 FM 态, 我们使用 PGD 算法弛豫 18 个体积点处的构型; 对于 PM 态, 我们直接使用 FM 态弛豫后得到的构型计算能量, 其中初始磁构型由整数规划确定 (可见图 5.3 (b) 和 (c)). 我们将计算所得以及拟合^[305] 得到的 FM 态与 PM 态的能量-体积关系绘制在图 5.4 中, 将拟合的参数汇总于表 5.3 中.

从表 5.3 可知, 使用 PGD 算法计算的结果拟合的 FM 态与 PM 态的 V_0 与 B_0 , FM 态的 E_0 以及 PM 态的 B'_0 和已有工作^[306] 相符, 而 FM 态的 B'_0 和 PM

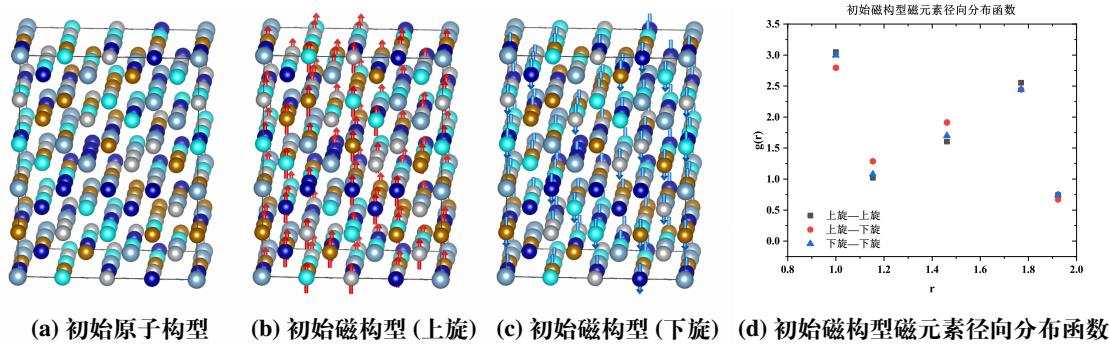


图 5.3 由 SAE 方法与整数规划生成的初始原子与磁构型. (a) 由 SAE 方法生成的初始原子构型. (b) 由整数规划选取的初始磁构型中的 64 个上旋位点 (红色箭头). (c) 由整数规划选取的初始磁构型中的 64 个下旋位点 (蓝色箭头). (d) 初始磁构型中磁元素的径向分布函数

Figure 5.3 The initial atomic and magnetic configurations generated by the SAE method and integer programming. (a) The initial atomic configuration generated by the SAE method. (b) 64 spin-up sites in the initial magnetic configuration chosen by the integer programming. (c) 64 spin-down sites in the initial magnetic configuration chosen by the integer programming. (d) The radial distribution function of the magnetic elements in the initial magnetic configuration

表 5.2 在 Wigner-Seitz 半径范围 [2.45 Å, 2.70 Å] 内按等间距选取的 18 个晶格体积点

Table 5.2 18 unit cell volumes chosen uniformly in the Wigner-Seitz radius of [2.45 Å, 2.70 Å]

体积编号	1	2	3	4	5	6	7	8	9
每原子体积 (Å³/atom)	10.29	10.45	10.60	10.76	10.92	11.08	11.24	11.40	11.57
体积序号	10	11	12	13	14	15	16	17	18
每原子体积 (Å³/atom)	11.73	11.90	12.07	12.24	12.42	12.59	12.77	12.95	13.12

态的 E_0 则有明显区别. 我们分析原因有二: (1) 相比于文献^[306], 我们采用整数规划生成了更加合理的初始磁构型; (2) 文献^[306]仅考虑了原子位置的弛豫, 而 PGD 算法则同时弛豫了原子位置与晶格基矢. 基于平均场近似^[307], 我们可以估计 AlCoCrFeNi 的 Curie 温度为

$$\frac{2}{3(1-c)k_B}(E_0^{\text{PM}} - E_0^{\text{FM}}) \approx 210.8 \text{ K},$$

其中 c 为非磁元素的浓度 (concentration) (此例中为 0.2), E_0^{PM} 和 E_0^{FM} 分别是基于 PGD 算法的结果拟合得到的 PM 态和 FM 态的平衡能量.

下面, 我们计算 AlCoCrFeNi 的热力学性质, 使用修正平均场势 (modified mean field potential) 方法^[308] 估计结构的体积-压强关系. 由于修正平均场势方法的有效性高度依赖静态物态方程的质量, 因此我们可基于此进一步验证上述构型建模与计算结果的正确性. 我们首先通过修正平均场势方法估计结构在室温条件下的平衡体积与体模量, 分别为 11.83 Å³/atom 与 148 GPa. 这与已有工作中使

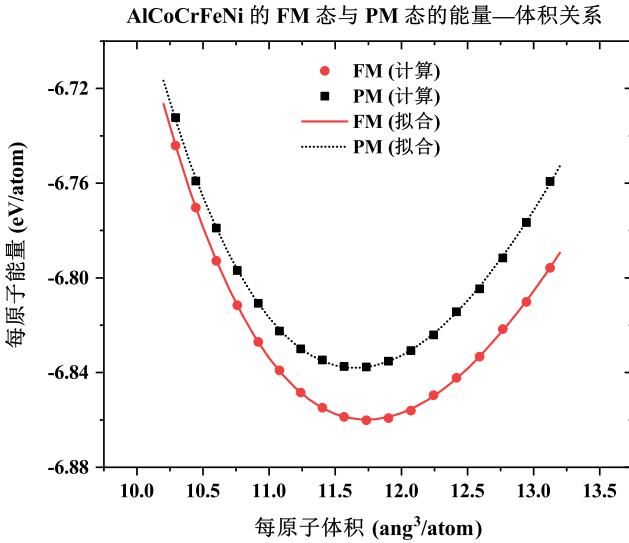


图 5.4 AlCoCrFeNi 的 FM 态与 PM 态的能量-体积关系。“计算”表示第一性原理计算结果，“拟合”表示拟合的静态 BM3 物态方程，红色圆点与实线表示 FM 态的结果，黑色方形与点线表示 PM 态的结果

Figure 5.4 Energy-volume relations on FM and PM AlCoCrFeNi. “Computation” indicates the *ab initio* results, while “Fit” is for the curves fitted to the static BM3 equation of state, red circles and solid line refer to the results of the FM state, black squares and dotted line refer to the results of the PM state

表 5.3 拟合的静态 BM3 物态方程中的参数。其中“FM/PM (PGD)”表示使用 PGD 算法弛豫的结果，“FM/PM (CG)”表示使用 CG 算法弛豫的结果，其余为已有工作的结果

Table 5.3 Fitted parameters of the static BM3 equation of state. “FM/PM (PGD)” refer to the results using the PGD method, “FM/PM (CG)” refer to the results using the CG method, the rest are the results in the existing work

数据来源	V_0 (Å ³ /atom)	E_0 (eV/atom)	B_0 (GPa)	B'_0
FM (文献 ^[306])	11.77	-6.85	157.32	8.9
FM (PGD)	11.74	-6.86	158.49	5.3
FM (CG)	11.73	-6.86	158.04	5.8
PM (文献 ^[306])	11.70	-6.82	161.64	5.3
PM (PGD)	11.65	-6.84	167.59	4.7
PM (CG)	11.65	-6.84	161.64	4.0

用 X 射线衍射 (X-ray diffraction) 与金刚石压砧 (diamond anvil cell) 实验测定的结果 $11.79 \text{ \AA}^3/\text{atom}$ 与 $150 \pm 2.5 \text{ GPa}$ ^[309] 相符。我们还计算了结构在 300 K 等温的体积-压强关系 (可见图 5.5)。在相同的压强下, 我们所估计的体积与实验测定值的最大偏差不超过 0.7%。

我们也在 AlCoCrFeNi 的 FM 态弛豫任务上比较了 PGD 算法与 CG 算法的

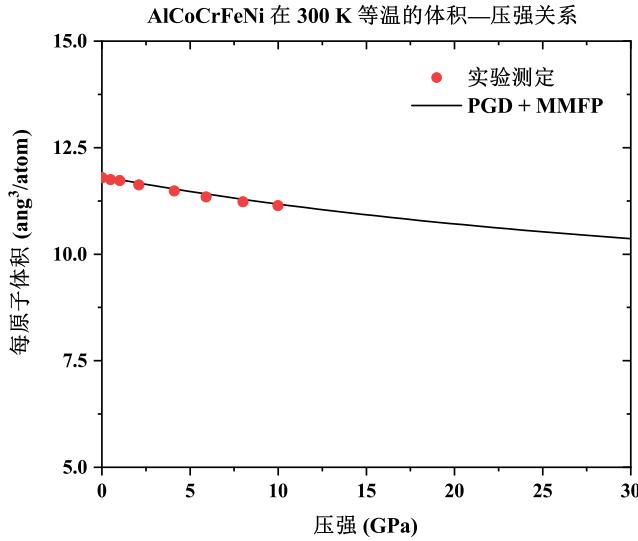


图 5.5 AlCoCrFeNi 在 300 K 等温的体积-压强关系. 其中, “实验测定” 表示使用 X 射线衍射与金刚石压砧实验测定的结果, “PGD + MMFP” 表示由修正平均场势方法基于 PGD 算法的弛豫结果给出的估计

Figure 5.5 Volume-pressure relation on AlCoCrFeNi at isotherm 300 K. “Experimental measurements” represents the X-ray diffraction and diamond anvil cell experimental results, while “PGD + MMFP” refers to the relation given by the MMFP approach based on the PGD results.

数值表现. PGD 算法在所有体积点上均正常收敛, 而 CG 算法则在四个体积点上因为线搜索方向非下降无法搜寻到合适步长而崩溃. 此外, 当压强超过 20 GPa 时, 我们发现 CG 算法在弛豫过程中会产生严重的晶格畸变 (可见图 5.6). 以 CG 在 20 GPa 以内体积点处的弛豫结果作为输入, 我们再次拟合了 AlCoCrFeNi 的静态 BM3 物态方程 (可见表 5.3). 显然, 使用 CG 算法与 PGD 算法的数据拟合的结果高度一致. 效率上, PGD 算法对于 CG 算法的平均 CPU 加速比为 1.36.

5.5 本章小结

在本章中, 我们考虑了行列式约束优化问题. 该问题在第一性原理固定晶格体积晶体结构弛豫上具有应用, 是材料结构物态方程计算的重要组成部分. 现有算法效率较低, 且因使用了缩放算子保持晶格体积, 尚无收敛性保证. 这使得该问题的求解成为下游应用的计算瓶颈. 为此, 我们结合负梯度到可行域切锥上的正交投影与缩放算子, 设计了可行的 PGD 算法. 借助可行域切锥的代数结构与非单调线搜索, 我们首次在使用缩放算子保持迭代点可行性的条件下, 证明了算法的全局依子列收敛性.

在数值实验中, 我们将 PGD 算法推广至求解第一性原理固定晶格体积晶体结构弛豫问题. 在含有 223 个来自不同类别结构的基准算例集上, PGD 算法在效率与鲁棒性上显著优于 CG 算法与 QN 算法. 我们还将新算法用于计算难弛豫的

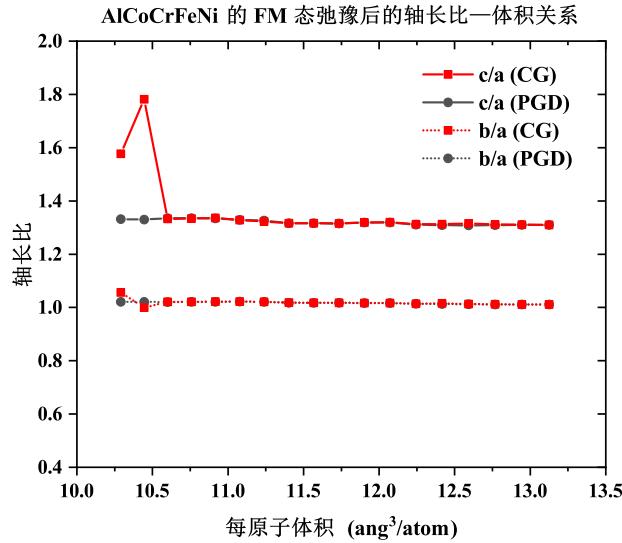


图 5.6 AlCoCrFeNi 的 FM 态弛豫后的轴长比-一体积关系. 其中, “c/a” 与 “b/a” 分别表示 c 轴和 a 轴长度之比与 b 轴与 a 轴长度之比. 红色实线与点线表示 CG 算法的结果, 黑色实线与点线表示 PGD 算法的结果. 对于 CG 算法无法正常收敛的体积点, 我们使用其最后迭代构型的轴长比

Figure 5.6 Axis length ratio-volume relation on the relaxed FM AlCoCrFeNi. “c/a” and “b/a” stand for the ratios between axes c and a and ratios between axes b and a, respectively. Red solid and dotted lines show the results of the CG method, black solid and dotted lines show the results of the PGD method. For the volumes where the CG method fails to converge normally, we use the axis length ratios of its last iterates

高熵合金 AlCoCrFeNi 的物态方程, 其中初始磁构型的构建使用了整数规划. 我们拟合的结果与已有计算结果相合, 与已有实验测定数据吻合较好.

第6章 总结与展望

本文的研究内容立足于优化与计算材料科学的交叉点.

一方面,从计算材料科学的应用背景出发,我们考虑了四类带有特殊结构的优化问题,研究了它们的模型性质,利用它们的特殊结构设计了一些具有理论保证的优化算法.我们将本文应用领域、研究内容、主要贡献及它们的关系概括在图 6.1 中.具体地说,从强关联电子体系计算出发,我们研究了带有特殊结构的

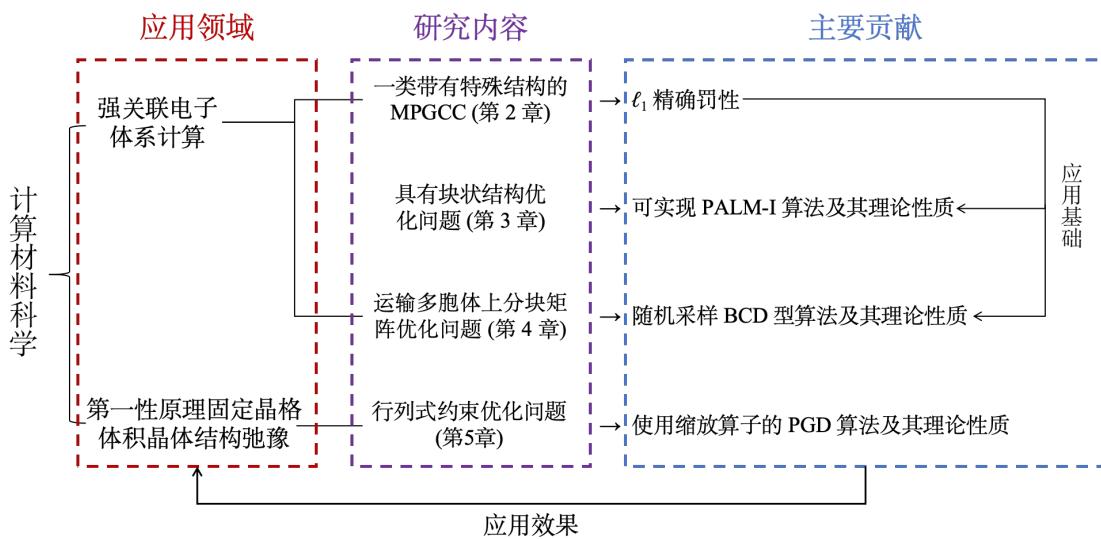


图 6.1 本文应用领域、研究内容、主要贡献及它们的关系

Figure 6.1 The application fields, research contents, main contributions of this paper and their relations

MPGCC (2.1) 和具有块状结构的优化问题 (3.1) 与 (4.1). 在第 2 章中, 为消除广义互补约束带来的困难, 我们证明了 MPGCC (2.1) 的 ℓ_1 罚函数精确性. 由于强关联电子体系计算中 MPGCC (1.26) 的 ℓ_1 罚问题是问题 (3.1) 与 (4.1) 的特例, 因此第 2 章的内容为后文算法的应用提供了理论基础. 在第 3 章中, 我们研究了求解具有块状结构优化问题的 PALM-I 算法, 首次在可实现的条件下, 证明了其全局依子 (点) 列收敛性与渐进收敛速度. 在第 4 章中, 我们为运输多胞体上的分块矩阵优化问题设计了无需使用全矩阵的块坐标下降型算法 S-ERALM 与 S-KLALM, 首次为矩阵逐元素随机近似在分块非凸问题上的应用提供了理论保证. 从第一性原理固定晶格体积晶体结构弛豫出发, 我们研究了行列式约束优化问题 (5.1). 在第 5 章中, 我们为其设计了 PGD 算法, 首次在使用缩放算子保持迭代点可行性时, 证明了算法的全局依子列收敛性.

另一方面, 我们将这些优化理论与算法应用于计算材料科学中的问题, 不仅获得了更高的求解效率, 也取得了全新的数值结果. 在第 3 与第 4 章中, 我们使用所设计的算法模拟了一维至三维的强关联电子体系. 其中, PALM-I 算法的效率显

著优于可行的 PALM 算法, S-KLALM 算法展现出了对离散规模的平方标度. 我们还将所设计的算法与 CMGOPT 框架结合, 在一维情形下取得了与理论预测相合的数值结果, 在二维与三维情形下首次可视化了电子位置之间的映射. 在第 5 章中, 我们将 PGD 算法推广至求解固定晶格体积晶体结构弛豫问题. 在包含 223 个结构的基准算例集上, 相较于常用材料模拟软件中的 CG 算法与 QN 算法, PGD 算法平均加速 1.41 与 1.45 倍, 且具有显著的鲁棒性优势. 此外, 对于难弛豫的高熵合金 AlCoCrFeNi, 我们使用 PGD 算法计算了其静态 BM3 物态方程, 取得了与已有实验测定数据吻合的计算结果.

本文的研究结果表明, 优化理论与算法在解决计算材料科学中的问题时可发挥关键的作用; 反过来, 来自计算材料科学领域的需求可促进优化理论与算法的进一步发展. 不过, 本文的研究内容仅仅是冰山一角. 下面, 我们列举一些值得继续研究的问题.

一方面, 关于本文中涉及的优化问题、理论与算法, 我们仍然有进一步改进或思考的空间. 对于 MPGCC 的 ℓ_1 罚函数精确性, 我们的证明方法较为依赖问题的特殊结构. 在未来的研究中, 我们可以考虑如何减弱对问题类的要求, 或者更一般地, 考虑 ℓ_1 罚函数精确的 MPGCC 都具有何种特殊性质. 对于 PALM-I 算法, 我们可以进一步研究其在非光滑或非凸约束优化问题上的理论性质, 也可以考虑如何进一步减弱现有理论分析中需要的条件, 如 Hoffman 型误差界 (3.10). 对于 (S-)KLALM 算法, 由于 KL 散度缺少局部 Lipschitz 光滑性, 我们并未分析其收敛性质. 在未来的研究中, 我们可以尝试从非 Lipschitz (non-Lipschitz) 优化的角度研究其理论性质. 对于非凸二次规划问题 (2.8), 第 4 章的数值结果似乎暗示随着离散规模趋于无穷大, 其一阶稳定点处的目标函数值将趋于最优值. 若这一点可以被严格证明, 我们就不需要调用昂贵的全局优化算法, 只需使用局部优化算法从任意初始点出发, 求解一个离散规模足够大的问题 (2.8) 即可. 由于该问题来源于强关联电子体系计算, 研究这一模型理论具有重要的应用价值.

另一方面, 计算材料科学中还有很多其他重要且困难的问题急需来自优化学科的解决方案. 下面以电子与原子尺度举例说明. 在电子结构计算中, 为了避免直接求解 Schrödinger 方程, 我们需要首先明确感兴趣的电子体系, 随后针对性地研究 Schrödinger 方程的近似模型, 设计求解算法. 对于强关联电子体系, 除了本文关注的 SCEDFT (1.16), 还有许多其他的模型, 例如变分 Monte Carlo、密度矩阵重整化群、动力学平均场理论、密度矩阵嵌入理论等. 在这些模型中不乏优化问题的身影. 例如, 变分 Monte Carlo 的主要数学模型是一个随机优化问题, 其目标函数具有期望形式. 由于计算期望需要近似计算高维积分且其中的分布会随着变量改变, 我们在计算中得到的目标函数值、梯度等信息天然具有误差和有偏性. 这给求解算法的设计与理论分析带来了巨大的挑战. 除了强关联材料, 我们还可以考虑其他类型材料的电子结构计算. 例如, 磁性材料电子结构计算的一个重要模型是自旋极化的 (spin-polarized) DFT, 其与不考虑自旋的 DFT 相比需要区分不同自旋电子的密度, 从而对应不同的优化问题. 在求解原子尺度的问题时, 电子结构

计算只是一个部件, 其主要作用是为了近似计算给定构型的能量. 用优化的语言来说, 它只是在算一个点的目标函数信息. 原子尺度的问题往往以搜寻满足某个性质的构型为目标, 与新材料的研发直接相关. 具有代表性的原子尺度的问题包括晶体结构预测、过渡态 (transition state) 搜索等. 其中, 晶体结构预测旨在通过最小化结构的能量, 寻找最稳定的构型, 对应于全局优化问题; 过渡态搜索旨在寻找化学反应或相变中的过渡态结构, 对应于鞍点搜索与解景观的构建. 最后, 对于跨尺度的计算, 我们还可以探究误差传播的机制, 进而在保证迭代收敛的前提下, 尽量减少计算资源的消耗.

参考文献

- [1] 袁亚湘. 非线性优化计算方法 [M]. 北京: 科学出版社, 2008.
- [2] LeSar R. Introduction to Computational Materials Science: Fundamentals to Applications [M/OL]. Cambridge: Cambridge University Press, 2013. DOI: [10.1017/CBO9781139033398](https://doi.org/10.1017/CBO9781139033398).
- [3] Schrödinger E. An undulatory theory of the mechanics of atoms and molecules [J/OL]. Physical Review, 1926, 28(6): 1049. DOI: [10.1103/PhysRev.28.1049](https://doi.org/10.1103/PhysRev.28.1049).
- [4] Bednorz J G, Müller K A. Possible high T_c superconductivity in the Ba-La-Cu-O system [J/OL]. Zeitschrift für Physik B Condensed Matter, 1986, 64(2): 189-193. DOI: [10.1007/BF01303701](https://doi.org/10.1007/BF01303701).
- [5] Capone M, Fabrizio M, Castellani C, et al. Strongly correlated superconductivity [J/OL]. Science, 2002, 296(5577): 2364-2366. DOI: [10.1126/science.1071122](https://doi.org/10.1126/science.1071122).
- [6] von Helmolt R, Wecker J, Holzapfel B, et al. Giant negative magnetoresistance in perovskite-like $\text{La}_{2/3}\text{Ba}_{1/3}\text{MnO}_x$ ferromagnetic films [J/OL]. Physical Review Letters, 1993, 71(14): 2331. DOI: [10.1103/PhysRevLett.71.2331](https://doi.org/10.1103/PhysRevLett.71.2331).
- [7] Salamon M B, Jaime M. The physics of manganites: Structure and transport [J/OL]. Reviews of Modern Physics, 2001, 73(3): 583. DOI: [10.1103/RevModPhys.73.583](https://doi.org/10.1103/RevModPhys.73.583).
- [8] Costi T, Zlatić V. Thermoelectric transport through strongly correlated quantum dots [J/OL]. Physical Review B, 2010, 81(23): 235127. DOI: [10.1103/PhysRevB.81.235127](https://doi.org/10.1103/PhysRevB.81.235127).
- [9] Kohn W, Sham L J. Self-consistent equations including exchange and correlation effects [J/OL]. Physical Review, 1965, 140(4A): A1133. DOI: [10.1103/PhysRev.140.A1133](https://doi.org/10.1103/PhysRev.140.A1133).
- [10] Imada M, Fujimori A, Tokura Y. Metal-insulator transitions [J/OL]. Reviews of Modern Physics, 1998, 70(4): 1039. DOI: [10.1103/RevModPhys.70.1039](https://doi.org/10.1103/RevModPhys.70.1039).
- [11] Cohen A J, Mori-Sánchez P, Yang W. Insights into current limitations of density functional theory [J/OL]. Science, 2008, 321(5890): 792-794. DOI: [10.1126/science.1158722](https://doi.org/10.1126/science.1158722).
- [12] Vinet P, Rose J H, Ferrante J, et al. Universal features of the equation of state of solids [J/OL]. Journal of Physics: Condensed Matter, 1989, 1(11): 1941. DOI: [10.1088/0953-8984/1/11/002](https://doi.org/10.1088/0953-8984/1/11/002).
- [13] Kittel C. Introduction to Solid State Physics [M]. 8th ed. Hoboken: John Wiley & Sons, Inc., 2005.
- [14] Oganov A R, Glass C W. Crystal structure prediction using *ab initio* evolutionary techniques: Principles and applications [J/OL]. The Journal of Chemical Physics, 2006, 124(24): 244704. DOI: [10.1063/1.2210932](https://doi.org/10.1063/1.2210932).
- [15] Cheng G, Gong X G, Yin W J. Crystal structure prediction by combining graph network and optimization algorithm [J/OL]. Nature Communications, 2022, 13: 1492. DOI: [10.1038/s41467-022-29241-4](https://doi.org/10.1038/s41467-022-29241-4).
- [16] Wang Y, Lv J, Gao P, et al. Crystal structure prediction via efficient sampling of the potential energy surface [J/OL]. Accounts of Chemical Research, 2022, 55(15): 2068-2076. DOI: [10.1021/acs.accounts.2c00243](https://doi.org/10.1021/acs.accounts.2c00243).
- [17] Xue D, Balachandran P V, Hogden J, et al. Accelerated search for materials with targeted properties by adaptive design [J/OL]. Nature Communications, 2016, 7: 11241. DOI: [10.1038/ncomms11241](https://doi.org/10.1038/ncomms11241).

- [18] Lenz M O, Purcell T A, Hicks D, et al. Parametrically constrained geometry relaxations for high-throughput materials science [J/OL]. *npj Computational Materials*, 2019, 5: 123. DOI: [10.1038/s41524-019-0254-4](https://doi.org/10.1038/s41524-019-0254-4).
- [19] 刘浩洋, 户将, 李勇锋, 等. 最优化: 建模、算法与理论 [M]. 北京: 高等教育出版社, 2020.
- [20] Nocedal J, Wright S J. Springer Series in Operations Research and Financial Engineering: Numerical Optimization [M/OL]. 2nd ed. New York: Springer, 2006. DOI: [10.1007/978-0-387-40065-5](https://doi.org/10.1007/978-0-387-40065-5).
- [21] Rockafellar R T. Princeton Landmarks in Mathematics and Physics: volume 30 Convex Analysis [M/OL]. Princeton: Princeton University Press, 1970. DOI: [10.1515/9781400873173](https://doi.org/10.1515/9781400873173).
- [22] Mordukhovich B S. A Series of Comprehensive Studies in Mathematics: volume 330 Variational Analysis and Generalized Differentiation I: Basic Theory [M/OL]. Berlin: Springer, 2006. DOI: [10.1007/3-540-31247-1](https://doi.org/10.1007/3-540-31247-1).
- [23] 王奇超, 文再文, 蓝光辉, 等. 优化算法的复杂度分析 [J]. 中国科学: 数学, 2020, 50(9): 1271-1336.
- [24] Parr R G, Yang W. Density Functional Theory of Atoms and Molecules [M]. New York: Oxford University Press, 1995.
- [25] Dreizler R M, Gross E K. Density Functional Theory: An Approach to the Quantum Many-Body Problem [M/OL]. Berlin: Springer, 1990. DOI: [10.1007/978-3-642-86105-5](https://doi.org/10.1007/978-3-642-86105-5).
- [26] Saad Y, Chelikowsky J R, Shontz S M. Numerical methods for electronic structure calculations of materials [J/OL]. *SIAM Review*, 2010, 52(1): 3-54. DOI: [10.1137/060651653](https://doi.org/10.1137/060651653).
- [27] Burke K. Perspective on density functional theory [J/OL]. *The Journal of Chemical Physics*, 2012, 136(15): 150901. DOI: [10.1063/1.4704546](https://doi.org/10.1063/1.4704546).
- [28] Becke A D. Perspective: Fifty years of density-functional theory in chemical physics [J/OL]. *The Journal of Chemical Physics*, 2014, 140(18): 18A301. DOI: [10.1063/1.4869598](https://doi.org/10.1063/1.4869598).
- [29] Lin L, Lu J, Ying L. Numerical methods for Kohn-Sham density functional theory [J/OL]. *Acta Numerica*, 2019, 28: 405-539. DOI: [10.1017/S0962492919000047](https://doi.org/10.1017/S0962492919000047).
- [30] Friesecke G, Gerolin A, Gori-Giorgi P. The strong-interaction limit of density functional theory [M/OL]//Cancès E, Friesecke G. Mathematics and Molecular Modeling: Density Functional Theory: Modeling, Mathematical Analysis, Computational Methods, and Applications. Cham: Springer, 2023: 183-266. DOI: [10.1007/978-3-031-22340-2_4](https://doi.org/10.1007/978-3-031-22340-2_4).
- [31] Born M, Oppenheimer R. Zur quantentheorie der moleküle [J/OL]. *Annalen der Physik*, 1927, 389(20): 457-484. DOI: [10.1002/andp.19273892002](https://doi.org/10.1002/andp.19273892002).
- [32] Zhislin G M. Discussion of the spectrum of the Schrödinger operator for systems of many particles [J]. Trudy Moskovskogo Matematicheskogo Obschestva, 1960, 9: 81-128.
- [33] Pauli W. Über den Zusammenhang des Abschlusses der elektronengruppen im atom mit der komplexstruktur der spektren [J/OL]. *Zeitschrift für Physik*, 1925, 31(1): 765-783. DOI: [10.1007/BF02980631](https://doi.org/10.1007/BF02980631).
- [34] Hohenberg P, Kohn W. Inhomogeneous electron gas [J/OL]. *Physical Review*, 1964, 136 (3B): B864. DOI: [10.1103/PhysRev.136.B864](https://doi.org/10.1103/PhysRev.136.B864).
- [35] Zhou A. Hohenberg-Kohn theorem for Coulomb type systems and its generalization [J/OL]. *Journal of Mathematical Chemistry*, 2012, 50(10): 2746-2754. DOI: [10.1007/s10910-012-0061-3](https://doi.org/10.1007/s10910-012-0061-3).

- [36] Zhou A. A mathematical aspect of Hohenberg-Kohn theorem [J/OL]. Science China Mathematics, 2019, 62(1): 63-68. DOI: [10.1007/s11425-018-9337-2](https://doi.org/10.1007/s11425-018-9337-2).
- [37] Levy M. Electron densities in search of Hamiltonians [J/OL]. Physical Review A, 1982, 26(3): 1200. DOI: [10.1103/PhysRevA.26.1200](https://doi.org/10.1103/PhysRevA.26.1200).
- [38] Lieb E H. Density functionals for Coulomb systems [J/OL]. International Journal of Quantum Chemistry, 1983, 24(3): 243-277. DOI: [10.1002/qua.560240302](https://doi.org/10.1002/qua.560240302).
- [39] Slater J C. The theory of complex spectra [J/OL]. Physical Review, 1929, 34(10): 1293. DOI: [10.1103/PhysRev.34.1293](https://doi.org/10.1103/PhysRev.34.1293).
- [40] Toulouse J. Review of approximations for the exchange-correlation energy in density-functional theory [M/OL]//Cancès E, Friesecke G. Mathematics and Molecular Modeling: Density Functional Theory: Modeling, Mathematical Analysis, Computational Methods, and Applications. Cham: Springer, 2023: 1-90. DOI: [10.1007/978-3-031-22340-2_1](https://doi.org/10.1007/978-3-031-22340-2_1).
- [41] Roothaan C C J. New developments in molecular orbital theory [J/OL]. Reviews of Modern Physics, 1951, 23(2): 69. DOI: [10.1103/RevModPhys.23.69](https://doi.org/10.1103/RevModPhys.23.69).
- [42] Pulay P. Convergence acceleration of iterative sequences. The case of SCF iteration [J/OL]. Chemical Physics Letters, 1980, 73(2): 393-398. DOI: [10.1016/0009-2614\(80\)80396-4](https://doi.org/10.1016/0009-2614(80)80396-4).
- [43] Pulay P. Improved SCF convergence acceleration [J/OL]. Journal of Computational Chemistry, 1982, 3(4): 556-560. DOI: [10.1002/jcc.540030413](https://doi.org/10.1002/jcc.540030413).
- [44] Cancès E. Self-consistent field algorithms for Kohn-Sham models with fractional occupation numbers [J/OL]. The Journal of Chemical Physics, 2001, 114(24): 10616-10622. DOI: [10.1063/1.1373430](https://doi.org/10.1063/1.1373430).
- [45] Marks L D, Luke D. Robust mixing for *ab initio* quantum mechanical calculations [J/OL]. Physical Review B, 2008, 78(7): 075114. DOI: [10.1103/PhysRevB.78.075114](https://doi.org/10.1103/PhysRevB.78.075114).
- [46] Gao W, Yang C, Meza J C. Solving a class of nonlinear eigenvalue problems by Newton's method [R/OL]. Berkeley, CA, United States: Lawrence Berkeley National Laboratory, 2009. <https://www.osti.gov/biblio/965775>. DOI: [10.2172/965775](https://doi.org/10.2172/965775).
- [47] Yang C, Gao W, Meza J C. On the convergence of the self-consistent field iteration for a class of nonlinear eigenvalue problems [J/OL]. SIAM Journal on Matrix Analysis and Applications, 2009, 30(4): 1773-1788. DOI: [10.1137/080716293](https://doi.org/10.1137/080716293).
- [48] Cancès E, Le Bris C. On the convergence of SCF algorithms for the Hartree-Fock equations [J/OL]. ESAIM: Mathematical Modelling and Numerical Analysis, 2000, 34(4): 749-774. DOI: [10.1051/m2an:2000102](https://doi.org/10.1051/m2an:2000102).
- [49] Liu X, Wang X, Wen Z, et al. On the convergence of the self-consistent field iteration in Kohn-Sham density functional theory [J/OL]. SIAM Journal on Matrix Analysis and Applications, 2014, 35(2): 546-558. DOI: [10.1137/130911032](https://doi.org/10.1137/130911032).
- [50] Liu X, Wen Z, Wang X, et al. On the analysis of the discretized Kohn-Sham density functional theory [J/OL]. SIAM Journal on Numerical Analysis, 2015, 53(4): 1758-1785. DOI: [10.1137/140957962](https://doi.org/10.1137/140957962).
- [51] Bai Z, Li R C, Lu D. Sharp estimation of convergence rate for self-consistent field iteration to solve eigenvector-dependent nonlinear eigenvalue problems [J/OL]. SIAM Journal on Matrix Analysis and Applications, 2022, 43(1): 301-327. DOI: [10.1137/20M136606X](https://doi.org/10.1137/20M136606X).
- [52] Rutishauser H. Simultaneous iteration method for symmetric matrices [J/OL]. Numerische Mathematik, 1970, 16(3): 205-223. DOI: [10.1007/BF02219773](https://doi.org/10.1007/BF02219773).

- [53] Knyazev A V. Toward the optimal preconditioned eigensolver: Locally optimal block preconditioned conjugate gradient method [J/OL]. SIAM Journal on Scientific Computing, 2001, 23(2): 517-541. DOI: [10.1137/S1064827500366124](https://doi.org/10.1137/S1064827500366124).
- [54] Liu X, Wen Z, Zhang Y. Limited memory block Krylov subspace optimization for computing dominant singular value decompositions [J/OL]. SIAM Journal on Scientific Computing, 2013, 35(3): A1641-A1668. DOI: [10.1137/120871328](https://doi.org/10.1137/120871328).
- [55] Mauri F, Galli G, Car R. Orbital formulation for electronic-structure calculations with linear system-size scaling [J/OL]. Physical Review B, 1993, 47(15): 9973. DOI: [10.1103/PhysRevB.47.9973](https://doi.org/10.1103/PhysRevB.47.9973).
- [56] Lin L, Lu J, Ying L, et al. Fast algorithm for extracting the diagonal of the inverse matrix with application to the electronic structure analysis of metallic systems [J/OL]. Communications in Mathematical Sciences, 2009, 7(3): 755-777. DOI: [10.4310/CMS.2009.v7.n3.a12](https://doi.org/10.4310/CMS.2009.v7.n3.a12).
- [57] Wen Z, Yang C, Liu X, et al. Trace-penalty minimization for large-scale eigenspace computation [J/OL]. Journal of Scientific Computing, 2016, 66(3): 1175-1203. DOI: [10.1007/s10915-015-0061-0](https://doi.org/10.1007/s10915-015-0061-0).
- [58] Lu J, Thicke K. Orbital minimization method with ℓ_1 regularization [J/OL]. Journal of Computational Physics, 2017, 336: 87-103. DOI: [10.1016/j.jcp.2017.02.005](https://doi.org/10.1016/j.jcp.2017.02.005).
- [59] Absil P A, Mahony R, Sepulchre R. Optimization Algorithms on Matrix Manifolds [M/OL]. Princeton: Princeton University Press, 2008. DOI: [10.1515/9781400830244](https://doi.org/10.1515/9781400830244).
- [60] Schneider R, Rohwedder T, Neelov A, et al. Direct minimization for calculating invariant subspaces in density functional computations of the electronic structure [J/OL]. Journal of Computational Mathematics, 2009, 27(2-3): 360-387. https://global-sci.org/intro/article_detail/jcm/8577.html.
- [61] Wen Z, Yin W. A feasible method for optimization with orthogonality constraints [J/OL]. Mathematical Programming, 2013, 142(1-2): 397-434. DOI: [10.1007/s10107-012-0584-1](https://doi.org/10.1007/s10107-012-0584-1).
- [62] Zhang X, Zhu J, Wen Z, et al. Gradient type optimization methods for electronic structure calculations [J/OL]. SIAM Journal on Scientific Computing, 2014, 36(3): C265-C289. DOI: [10.1137/130932934](https://doi.org/10.1137/130932934).
- [63] Dai X, Liu Z, Zhang L, et al. A conjugate gradient method for electronic structure calculations [J/OL]. SIAM Journal on Scientific Computing, 2017, 39(6): A2702-A2740. DOI: [10.1137/16M1072929](https://doi.org/10.1137/16M1072929).
- [64] Hu J, Milzarek A, Wen Z, et al. Adaptive quadratically regularized Newton method for Riemannian optimization [J/OL]. SIAM Journal on Matrix Analysis and Applications, 2018, 39(3): 1181-1207. DOI: [10.1137/17M1142478](https://doi.org/10.1137/17M1142478).
- [65] Hu J, Jiang B, Lin L, et al. Structured quasi-Newton methods for optimization with orthogonality constraints [J/OL]. SIAM Journal on Scientific Computing, 2019, 41(4): A2239-A2269. DOI: [10.1137/18M121112X](https://doi.org/10.1137/18M121112X).
- [66] Dai X, Zhang L, Zhou A. Adaptive step size strategy for orthogonality constrained line search methods [A/OL]. arXiv preprint, 2019, arXiv:1906.02883. <https://arxiv.org/abs/1906.02883>.
- [67] Dai X, Zhang L, Zhou A. Practical Newton methods for electronic structure calculations [A/OL]. arXiv preprint, 2020, arXiv:2001.09285. <https://arxiv.org/abs/2001.09285>.
- [68] Ordejón P, Drabold D A, Grumbach M P, et al. Unconstrained minimization approach for electronic computations that scales linearly with system size [J/OL]. Physical Review B, 1993, 48(19): 14646. DOI: [10.1103/PhysRevB.48.14646](https://doi.org/10.1103/PhysRevB.48.14646).

- [69] Weber V, VandeVondele J, Hutter J, et al. Direct energy functional minimization under orthogonality constraints [J/OL]. *The Journal of Chemical Physics*, 2008, 128(8): 084113. DOI: [10.1063/1.2841077](https://doi.org/10.1063/1.2841077).
- [70] Gao B, Liu X, Yuan Y X. Parallelizable algorithms for optimization problems with orthogonality constraints [J/OL]. *SIAM Journal on Scientific Computing*, 2019, 41(3): A1949-A1983. DOI: [10.1137/18M1221679](https://doi.org/10.1137/18M1221679).
- [71] Dai X, Wang Q, Zhou A. Gradient flow based Kohn-Sham density functional theory model [J/OL]. *Multiscale Modeling & Simulation*, 2020, 18(4): 1621-1663. DOI: [10.1137/19M1276170](https://doi.org/10.1137/19M1276170).
- [72] Dai X, Zhang L, Zhou A. Convergent and orthogonality preserving schemes for approximating the Kohn-Sham orbitals [J/OL]. *Numerical Mathematics: Theory, Methods and Applications*, 2023, 16(1): 1-25. DOI: [10.4208/nmtma.OA-2022-0026](https://doi.org/10.4208/nmtma.OA-2022-0026).
- [73] Xiao N, Liu X, Yuan Y X. Exact penalty function for $\ell_{2,1}$ norm minimization over the Stiefel manifold [J/OL]. *SIAM Journal on Optimization*, 2021, 31(4): 3097-3126. DOI: [10.1137/20M1354313](https://doi.org/10.1137/20M1354313).
- [74] Liu X, Xiao N, Yuan Y X. A penalty-free infeasible approach for a class of nonsmooth optimization problems over the Stiefel manifold [J/OL]. *Journal of Scientific Computing*, 2024, 90(30): 1-29. DOI: [10.1007/s10915-024-02495-4](https://doi.org/10.1007/s10915-024-02495-4).
- [75] Gao B, Hu G, Kuang Y, et al. An orthogonalization-free parallelizable framework for all-electron calculations in density functional theory [J/OL]. *SIAM Journal on Scientific Computing*, 2022, 44(3): B723-B745. DOI: [10.1137/20M1355884](https://doi.org/10.1137/20M1355884).
- [76] Xiao N, Liu X, Toh K C. Dissolving constraints for Riemannian optimization [J/OL]. *Mathematics of Operations Research*, 2024, 49(1): 1-651, C2. DOI: [10.1287/moor.2023.1360](https://doi.org/10.1287/moor.2023.1360).
- [77] McMillan W L. Ground state of liquid He⁴ [J/OL]. *Physical Review*, 1965, 138(2A): A442. DOI: [10.1103/PhysRev.138.A442](https://doi.org/10.1103/PhysRev.138.A442).
- [78] Čížek J. On the correlation problem in atomic and molecular systems. Calculation of wavefunction components in Ursell-type expansion using quantum-field theoretical methods [J/OL]. *The Journal of Chemical Physics*, 1966, 45(11): 4256-4266. DOI: [10.1063/1.1727484](https://doi.org/10.1063/1.1727484).
- [79] White S R. Density matrix formulation for quantum renormalization groups [J/OL]. *Physical Review Letters*, 1992, 69(19): 2863. DOI: [10.1103/PhysRevLett.69.2863](https://doi.org/10.1103/PhysRevLett.69.2863).
- [80] Georges A, Kotliar G, Krauth W, et al. Dynamical mean-field theory of strongly correlated fermion systems and the limit of infinite dimensions [J/OL]. *Reviews of Modern Physics*, 1996, 68(1): 13. DOI: [10.1103/RevModPhys.68.13](https://doi.org/10.1103/RevModPhys.68.13).
- [81] Seidl M, Perdew J P, Levy M. Strictly correlated electrons in density-functional theory [J/OL]. *Physical Review A*, 1999, 59(1): 51. DOI: [10.1103/PhysRevA.59.51](https://doi.org/10.1103/PhysRevA.59.51).
- [82] Seidl M. Strong-interaction limit of density-functional theory [J/OL]. *Physical Review A*, 1999, 60(6): 4387. DOI: [10.1103/PhysRevA.60.4387](https://doi.org/10.1103/PhysRevA.60.4387).
- [83] Seidl M, Perdew J P, Kurth S. Simulation of all-order density-functional perturbation theory, using the second order and the strong-correlation limit [J/OL]. *Physical Review Letters*, 2000, 84(22): 5070. DOI: [10.1103/PhysRevLett.84.5070](https://doi.org/10.1103/PhysRevLett.84.5070).
- [84] Seidl M, Perdew J P, Kurth S. Density functionals for the strong-interaction limit [J/OL]. *Physical Review A*, 2000, 62(1): 012502. DOI: [10.1103/PhysRevA.62.012502](https://doi.org/10.1103/PhysRevA.62.012502).
- [85] Knizia G, Chan G K L. Density matrix embedding: A simple alternative to dynami-

- cal mean-field theory [J/OL]. Physical Review Letters, 2012, 109(18): 186404. DOI: [10.1103/PhysRevLett.109.186404](https://doi.org/10.1103/PhysRevLett.109.186404).
- [86] Liu Z F, Burke K. Adiabatic connection for strictly correlated electrons [J/OL]. The Journal of Chemical Physics, 2009, 131(12): 124124. DOI: [10.1063/1.3239472](https://doi.org/10.1063/1.3239472).
- [87] Gori-Giorgi P, Seidl M, Vignale G. Density-functional theory for strongly interacting electrons [J/OL]. Physical Review Letters, 2009, 103(16): 166402. DOI: [10.1103/PhysRevLett.103.166402](https://doi.org/10.1103/PhysRevLett.103.166402).
- [88] Gori-Giorgi P, Seidl M. Density functional theory for strongly-interacting electrons: Perspectives for physics and chemistry [J/OL]. Physical Chemistry Chemical Physics, 2010, 12(43): 14405-14419. DOI: [10.1039/C0CP01061H](https://doi.org/10.1039/C0CP01061H).
- [89] Malet F, Gori-Giorgi P. Strong correlation in Kohn-Sham density functional theory [J/OL]. Physical Review Letters, 2012, 109(24): 246402. DOI: [10.1103/PhysRevLett.109.246402](https://doi.org/10.1103/PhysRevLett.109.246402).
- [90] Malet F, Mirtschink A, Cremon J C, et al. Kohn-Sham density functional theory for quantum wires in arbitrary correlation regimes [J/OL]. Physical Review B, 2013, 87(11): 115146. DOI: [10.1103/PhysRevB.87.115146](https://doi.org/10.1103/PhysRevB.87.115146).
- [91] Chen H, Friesecke G, Mendl C B. Numerical methods for a Kohn-Sham density functional model based on optimal transport [J/OL]. Journal of Chemical Theory and Computation, 2014, 10(10): 4360-4368. DOI: [10.1021/ct500586q](https://doi.org/10.1021/ct500586q).
- [92] Di Marino S, Gerolin A. Optimal transport losses and Sinkhorn algorithm with general convex regularization [A/OL]. arXiv preprint, 2020, arXiv:2007.00976. <https://arxiv.org/abs/2007.00976>.
- [93] Chen H, Friesecke G. Pair densities in density functional theory [J/OL]. Multiscale Modeling & Simulation, 2015, 13(4): 1259-1289. DOI: [10.1137/15M1014024](https://doi.org/10.1137/15M1014024).
- [94] Buttazzo G, De Pascale L, Gori-Giorgi P. Optimal-transport formulation of electronic density-functional theory [J/OL]. Physical Review A, 2012, 85(6): 062502. DOI: [10.1103/PhysRevA.85.062502](https://doi.org/10.1103/PhysRevA.85.062502).
- [95] Cotar C, Friesecke G, Klüppelberg C. Density functional theory and optimal transportation with Coulomb cost [J/OL]. Communications on Pure and Applied Mathematics, 2013, 66(4): 548-599. DOI: [10.1002/cpa.21437](https://doi.org/10.1002/cpa.21437).
- [96] Peyré G, Cuturi M, et al. Computational optimal transport: With applications to data science [J/OL]. Foundations and Trends® in Machine Learning, 2019, 11(5-6): 355-607. DOI: [10.1561/2200000073](https://doi.org/10.1561/2200000073).
- [97] Villani C. Graduate Studies in Mathematics: volume 58 Topics in Optimal Transportation [M]. Providence: American Mathematical Society, 2003.
- [98] Lin T, Ho N, Cuturi M, et al. On the complexity of approximating multimarginal optimal transport [J/OL]. Journal of Machine Learning Research, 2022, 23(65): 2835-2877. <https://jmlr.org/papers/v23/19-843.html>.
- [99] Friesecke G, Vögler D. Breaking the curse of dimension in multi-marginal Kantorovich optimal transport on finite state spaces [J/OL]. SIAM Journal on Mathematical Analysis, 2018, 50(4): 3996-4019. DOI: [10.1137/17M1150025](https://doi.org/10.1137/17M1150025).
- [100] Colombo M, De Pascale L, Di Marino S. Multimarginal optimal transport maps for one-dimensional repulsive costs [J/OL]. Canadian Journal of Mathematics, 2015, 67(2): 350-368. DOI: [10.4153/CJM-2014-011-x](https://doi.org/10.4153/CJM-2014-011-x).
- [101] Seidl M, Gori-Giorgi P, Savin A. Strictly correlated electrons in density-functional theory:

- A general formulation with applications to spherical densities [J/OL]. Physical Review A, 2007, 75(4): 042511. DOI: [10.1103/PhysRevA.75.042511](https://doi.org/10.1103/PhysRevA.75.042511).
- [102] Colombo M, Stra F. Counterexamples in multimarginal optimal transport with Coulomb cost and spherically symmetric data [J/OL]. Mathematical Models and Methods in Applied Sciences, 2016, 26(06): 1025-1049. DOI: [10.1142/S021820251650024X](https://doi.org/10.1142/S021820251650024X).
- [103] Gerolin A. Multimarginal Optimal Transport and Potential Optimization Problems for Schrödinger Operators [D]. Pisa: Università di Pisa, 2016.
- [104] Seidl M, Di Marino S, Gerolin A, et al. The strictly-correlated electron functional for spherically symmetric systems revisited [A/OL]. arXiv preprint, 2017, arXiv:1702.05022. <https://arxiv.org/abs/1702.05022>.
- [105] Bindini U, De Pascale L, Kausamo A. On Seidl-type maps for multi-marginal optimal transport with Coulomb cost [A/OL]. arXiv preprint, 2020, arXiv:2011.05063. <https://arxiv.org/abs/2011.05063>.
- [106] Flegel M L, Kanzow C. On the Guignard constraint qualification for mathematical programs with equilibrium constraints [J/OL]. Optimization, 2005, 54(6): 517-534. DOI: [10.1080/02331930500342591](https://doi.org/10.1080/02331930500342591).
- [107] Friesecke G, Schulz A S, Vögler D. Genetic column generation: Fast computation of high-dimensional multimarginal optimal transport problems [J/OL]. SIAM Journal on Scientific Computing, 2022, 44(3): A1632-A1654. DOI: [10.1137/21M140732X](https://doi.org/10.1137/21M140732X).
- [108] Friesecke G, Penka M. Convergence proof for the GenCol algorithm in the case of two-marginal optimal transport [A/OL]. arXiv preprint, 2023, arXiv:2303.07137. <https://arxiv.org/abs/2303.07137>.
- [109] Friesecke G, Mendl C B, Pass B, et al. N-density representability and the optimal transport limit of the Hohenberg-Kohn functional [J/OL]. The Journal of Chemical Physics, 2013, 139(16): 164109. DOI: [10.1063/1.4821351](https://doi.org/10.1063/1.4821351).
- [110] Khoo Y, Ying L. Convex relaxation approaches for strictly correlated density functional theory [J/OL]. SIAM Journal on Scientific Computing, 2019, 41(4): B773-B795. DOI: [10.1137/18M1207478](https://doi.org/10.1137/18M1207478).
- [111] Khoo Y, Lin L, Lindsey M, et al. Semidefinite relaxation of multimarginal optimal transport for strictly correlated electrons in second quantization [J/OL]. SIAM Journal on Scientific Computing, 2020, 42(6): B1462-B1489. DOI: [10.1137/20M1310977](https://doi.org/10.1137/20M1310977).
- [112] Alfonsi A, Coyaud R, Ehrlicher V, et al. Approximation of optimal transport problems with marginal moments constraints [J/OL]. Mathematics of Computation, 2021, 90(328): 689-737. DOI: [10.1090/mcom/3568](https://doi.org/10.1090/mcom/3568).
- [113] Alfonsi A, Coyaud R, Ehrlicher V. Constrained overdamped Langevin dynamics for symmetric multimarginal optimal transportation [J/OL]. Mathematical Models and Methods in Applied Sciences, 2022, 32(03): 403-455. DOI: [10.1142/S0218202522500105](https://doi.org/10.1142/S0218202522500105).
- [114] Yang L, Sun D, Toh K C. SDPNAL+: A majorized semismooth Newton-CG augmented Lagrangian method for semidefinite programming with nonnegative constraints [J/OL]. Mathematical Programming Computation, 2015, 7(3): 331-366. DOI: [10.1007/s12532-015-0082-6](https://doi.org/10.1007/s12532-015-0082-6).
- [115] Mendl C B, Lin L. Kantorovich dual solution for strictly correlated electrons in atoms and molecules [J/OL]. Physical Review B, 2013, 87(12): 125106. DOI: [10.1103/PhysRevB.87.125106](https://doi.org/10.1103/PhysRevB.87.125106).

- [116] Benamou J D, Carlier G, Nenna L. A numerical method to solve multi-marginal optimal transport problems with Coulomb cost [M/OL]//Glowinski R, Osher S J, Yin W. Scientific Computation: Splitting Methods in Communication, Imaging, Science, and Engineering. Cham: Springer, 2017: 577-601. DOI: [10.1007/978-3-319-41589-5_17](https://doi.org/10.1007/978-3-319-41589-5_17).
- [117] 黄昆, 韩汝琦. 固体物理学 [M]. 北京: 高等教育出版社, 1998.
- [118] Hellmann H. Einführung in die Quantenchemie [M]. Leipzig: Franz Deuticke, 1937.
- [119] Feynman R P. Forces in molecules [J/OL]. Physical Review, 1939, 56(4): 340. DOI: [10.1103/PhysRev.56.340](https://doi.org/10.1103/PhysRev.56.340).
- [120] Atalla V. Density-Functional Theory and Beyond for Organic Electronic Materials [D/OL]. Berlin: Technischen Universität Berlin, 2013. https://depositonce.tu-berlin.de/bitstream/11303/4156/1/atalla_victor.pdf.
- [121] Knuth F. Strain and Stress: Derivation, Implementation, and Application to Organic Crystals [D/OL]. Berlin: Freie Universität Berlin, 2015. https://pure.mpg.de/rest/items/item_2228528_5/component/file_2228530/content.
- [122] Kresse G, Furthmüller J. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set [J/OL]. Computational Materials Science, 1996, 6 (1): 15-50. DOI: [10.1016/0927-0256\(96\)00008-0](https://doi.org/10.1016/0927-0256(96)00008-0).
- [123] Kresse G, Furthmüller J. Efficient iterative schemes for *ab initio* total-energy calculations using a plane-wave basis set [J/OL]. Physical Review B, 1996, 54(16): 11169. DOI: [10.1103/PhysRevB.54.11169](https://doi.org/10.1103/PhysRevB.54.11169).
- [124] Giannozzi P, Baroni S, Bonini N, et al. QUANTUM ESPRESSO: A modular and open-source software project for quantum simulations of materials [J/OL]. Journal of Physics: Condensed Matter, 2009, 21(39): 395502. DOI: [10.1088/0953-8984/21/39/395502](https://doi.org/10.1088/0953-8984/21/39/395502).
- [125] Hestenes M R, Stiefel E, et al. Methods of conjugate gradients for solving linear systems [J/OL]. Journal of Research of the National Bureau of Standards, 1952, 49(6): 409-436. <https://nvlpubs.nist.gov/nistpubs/jres/049/6/V49.N06.A08.pdf>.
- [126] Fletcher R, Reeves C M. Function minimization by conjugate gradients [J/OL]. The Computer Journal, 1964, 7(2): 149-154. DOI: [10.1093/comjnl/7.2.149](https://doi.org/10.1093/comjnl/7.2.149).
- [127] Polak E, Ribiere G. Note sur la convergence de méthodes de directions conjuguées [J/OL]. Revue Française d’Informatique et de Recherche Opérationnelle. Série Rouge, 1969, 3(16): 35-43. http://www.numdam.org/item/M2AN_1969__3_1_35_0.pdf.
- [128] Polyak B T. The conjugate gradient method in extremal problems [J/OL]. USSR Computational Mathematics and Mathematical Physics, 1969, 9(4): 94-112. DOI: [10.1016/0041-5553\(69\)90035-4](https://doi.org/10.1016/0041-5553(69)90035-4).
- [129] Powell M J D. Restart procedures for the conjugate gradient method [J/OL]. Mathematical Programming, 1977, 12(1): 241-254. DOI: [10.1007/BF01593790](https://doi.org/10.1007/BF01593790).
- [130] Dai Y H, Yuan Y X. A nonlinear conjugate gradient method with a strong global convergence property [J/OL]. SIAM Journal on Optimization, 1999, 10(1): 177-182. DOI: [10.1137/S1052623497318992](https://doi.org/10.1137/S1052623497318992).
- [131] Broyden C G. The convergence of a class of double-rank minimization algorithms 1. General considerations [J/OL]. IMA Journal of Applied Mathematics, 1970, 6(1): 76-90. DOI: [10.1093/imamat/6.1.76](https://doi.org/10.1093/imamat/6.1.76).
- [132] Fletcher R. A new approach to variable metric algorithms [J/OL]. The Computer Journal, 1970, 13(3): 317-322. DOI: [10.1093/comjnl/13.3.317](https://doi.org/10.1093/comjnl/13.3.317).

- [133] Goldfarb D. A family of variable metric updates derived by variational means [J/OL]. Mathematics of Computation, 1970, 24(109): 23-26. DOI: [10.1090/S0025-5718-1970-0258249-6](https://doi.org/10.1090/S0025-5718-1970-0258249-6).
- [134] Shanno D F. Conditioning of quasi-Newton methods for function minimization [J/OL]. Mathematics of Computation, 1970, 24(111): 647-656. DOI: [10.1090/S0025-5718-1970-0274029-X](https://doi.org/10.1090/S0025-5718-1970-0274029-X).
- [135] Liu D C, Nocedal J. On the limited memory BFGS method for large scale optimization [J/OL]. Mathematical Programming, 1989, 45(1-3): 503-528. DOI: [10.1007/BF01589116](https://doi.org/10.1007/BF01589116).
- [136] Luo Z Q, Pang J S, Ralph D. Mathematical Programs with Equilibrium Constraints [M/OL]. Cambridge: Cambridge University Press, 1996. DOI: [10.1017/CBO9780511983658](https://doi.org/10.1017/CBO9780511983658).
- [137] Hu J, Mitchell J E, Pang J S, et al. On the global solution of linear programs with linear complementarity constraints [J/OL]. SIAM Journal on Optimization, 2008, 19(1): 445-471. DOI: [10.1137/07068463x](https://doi.org/10.1137/07068463x).
- [138] Hu J, Mitchell J E, Pang J S, et al. On linear programs with linear complementarity constraints [J/OL]. Journal of Global Optimization, 2012, 53(1): 29-51. DOI: [10.1007/s10898-010-9644-3](https://doi.org/10.1007/s10898-010-9644-3).
- [139] Jeroslow R G. The polynomial hierarchy and a simple model for competitive analysis [J/OL]. Mathematical Programming, 1985, 32(2): 146-164. DOI: [10.1007/BF01586088](https://doi.org/10.1007/BF01586088).
- [140] Chung S J. NP-completeness of the linear complementarity problem [J/OL]. Journal of Optimization Theory and Applications, 1989, 60(3): 393-399. DOI: [10.1007/bf00940344](https://doi.org/10.1007/bf00940344).
- [141] Hansen P, Jaumard B, Savard G. New branch-and-bound rules for linear bilevel programming [J/OL]. SIAM Journal on Scientific and Statistical Computing, 1992, 13(5): 1194-1217. DOI: [10.1137/0913069](https://doi.org/10.1137/0913069).
- [142] Outrata J V. Optimality conditions for a class of mathematical programs with equilibrium constraints [J/OL]. Mathematics of Operations Research, 1999, 24(3): 627-644. DOI: [10.1287/moor.24.3.627](https://doi.org/10.1287/moor.24.3.627).
- [143] Scheel H, Scholtes S. Mathematical programs with complementarity constraints: Stationarity, optimality, and sensitivity [J/OL]. Mathematics of Operations Research, 2000, 25(1): 1-22. DOI: [10.1287/moor.25.1.1.15213](https://doi.org/10.1287/moor.25.1.1.15213).
- [144] Flegel M L, Kanzow C. Abadie-type constraint qualification for mathematical programs with equilibrium constraints [J/OL]. Journal of Optimization Theory and Applications, 2005, 124(3): 595-614. DOI: [10.1007/s10957-004-1176-x](https://doi.org/10.1007/s10957-004-1176-x).
- [145] Ye J J. Necessary and sufficient optimality conditions for mathematical programs with equilibrium constraints [J/OL]. Journal of Mathematical Analysis and Applications, 2005, 307(1): 350-369. DOI: [10.1016/j.jmaa.2004.10.032](https://doi.org/10.1016/j.jmaa.2004.10.032).
- [146] Flegel M L, Kanzow C. A direct proof for M-stationarity under MPEC-GCQ for mathematical programs with equilibrium constraints [M/OL]//Dempe S, Kalashnikov V. Springer Optimization and Its Applications: volume 2 Optimization with Multivalued Mappings: Theory, Applications, and Algorithms. Boston: Springer, 2006: 111-122. DOI: [10.1007/0_387_34221-4_6](https://doi.org/10.1007/0_387_34221-4_6).
- [147] Guo L, Chen X. Mathematical programs with complementarity constraints and a non-Lipschitz objective: Optimality and approximation [J/OL]. Mathematical Programming, 2021, 185(1-2): 455-485. DOI: [10.1007/s10107-019-01435-7](https://doi.org/10.1007/s10107-019-01435-7).
- [148] Luo Z Q, Pang J S, Ralph D, et al. Exact penalization and stationarity conditions of mathematical programs with equilibrium constraints [J/OL]. Mathematical Programming, 1996, 75(1): 19-76. DOI: [10.1007/BF02592205](https://doi.org/10.1007/BF02592205).

- [149] Le Thi H A, Pham Dinh T, Ngai H V. Exact penalty and error bounds in DC programming [J/OL]. *Journal of Global Optimization*, 2012, 52(3): 509-535. DOI: [10.1007/s10898-011-9765-3](https://doi.org/10.1007/s10898-011-9765-3).
- [150] Labb   M, Marcotte P, Savard G. A bilevel model of taxation and its application to optimal highway pricing [J/OL]. *Management Science*, 1998, 44(12-part-1): 1608-1622. DOI: [10.1287/mnsc.44.12.1608](https://doi.org/10.1287/mnsc.44.12.1608).
- [151] Brotcorne L, Labb   M, Marcotte P, et al. A bilevel model and solution algorithm for a freight tariff-setting problem [J/OL]. *Transportation Science*, 2000, 34(3): 289-302. DOI: [10.1287/trsc.34.3.289.12299](https://doi.org/10.1287/trsc.34.3.289.12299).
- [152] Labb   M, Marcotte P, Savard G. On a class of bilevel programs [M/OL]//Di Pillo G, Giannessi F. *Applied Optimization*: volume 36 Nonlinear Optimization and Related Topics. Boston: Springer, 2000: 183-206. DOI: [10.1007/978-1-4757-3226-9_10](https://doi.org/10.1007/978-1-4757-3226-9_10).
- [153] Brotcorne L, Labb   M, Marcotte P, et al. A bilevel model for toll optimization on a multi-commodity transportation network [J/OL]. *Transportation Science*, 2001, 35(4): 345-358. DOI: [10.1287/trsc.35.4.345.10433](https://doi.org/10.1287/trsc.35.4.345.10433).
- [154] Bard J F, Plummer J, Sourie J C. A bilevel programming approach to determining tax credits for biofuel production [J/OL]. *European Journal of Operational Research*, 2000, 120(1): 30-46. DOI: [10.1016/S0377-2217\(98\)00373-7](https://doi.org/10.1016/S0377-2217(98)00373-7).
- [155] C  t   J P, Marcotte P, Savard G. A bilevel modelling approach to pricing and fare optimisation in the airline industry [J/OL]. *Journal of Revenue and Pricing Management*, 2003, 2(1): 23-36. DOI: [10.1057/palgrave.rpm.5170046](https://doi.org/10.1057/palgrave.rpm.5170046).
- [156] Bouhtou M, Erbs G, Minoux M. Pricing and resource allocation for point-to-point telecommunication services in a competitive market: A bilevel optimization approach [M/OL]//Raghavan S, Anandalingam G. *Operations Research/Computer Science Interfaces Series*: volume 33 *Telecommunications Planning: Innovations in Pricing, Network Design and Management*. Boston: Springer, 2006: 1-16. DOI: [10.1007/0-387-29234-9_1](https://doi.org/10.1007/0-387-29234-9_1).
- [157] Bouhtou M, Erbs G, Minoux M. Joint optimization of pricing and resource allocation in competitive telecommunications networks [J/OL]. *Networks*, 2007, 50(1): 37-49. DOI: [10.1002/net.20164](https://doi.org/10.1002/net.20164).
- [158] Bouhtou M, Erbs G. A continuous optimization model for a joint problem of pricing and resource allocation [J/OL]. *RAIRO-Operations Research*, 2009, 43(2): 115-143. DOI: [10.1051/ro/2009008](https://doi.org/10.1051/ro/2009008).
- [159] Hoffman A J. On approximate solutions of systems of linear inequalities [J/OL]. *Journal of Research of the National Bureau of Standards*, 1952, 49(4): 263-265. DOI: [10.1142/9789812796936_0018](https://doi.org/10.1142/9789812796936_0018).
- [160] Pang J S. Error bounds in mathematical programming [J/OL]. *Mathematical Programming*, 1997, 79(1-3): 299-332. DOI: [10.1007/BF02614322](https://doi.org/10.1007/BF02614322).
- [161] Liu G, Han J, Zhang J. Exact penalty functions for convex bilevel programming problems [J/OL]. *Journal of Optimization Theory and Applications*, 2001, 110(3): 621-643. DOI: [10.1023/A:1017592429235](https://doi.org/10.1023/A:1017592429235).
- [162] Anandalingam G, White D. A solution method for the linear static Stackelberg problem using penalty functions [J/OL]. *IEEE Transactions on Automatic Control*, 1990, 35(10): 1170-1173. DOI: [10.1109/9.58565](https://doi.org/10.1109/9.58565).
- [163] White D J, Anandalingam G. A penalty function approach for solving bi-level linear programs [J/OL]. *Journal of Global Optimization*, 1993, 3(4): 397-419. DOI: [10.1007/BF01096412](https://doi.org/10.1007/BF01096412).

- [164] Campêlo M, Dantas S, Scheimberg S. A note on a penalty function approach for solving bilevel linear programs [J/OL]. *Journal of Global Optimization*, 2000, 16(3): 245. DOI: [10.1023/A:1008308218364](https://doi.org/10.1023/A:1008308218364).
- [165] Campêlo M, Scheimberg S. Theoretical and computational results for a linear bilevel problem [M/OL]//Hadjisavvas N, Pardalos P M. Nonconvex Optimization and Its Applications: volume 54 Advances in Convex Analysis and Global Optimization: Honoring the Memory of C. Caratheodory (1873-1950). Boston: Springer, 2001: 269-281. DOI: [10.1007/978-1-4613-0279-7_14](https://doi.org/10.1007/978-1-4613-0279-7_14).
- [166] Bertsimas D, Tsitsiklis J N. Introduction to Linear Optimization [M]. Belmont: Athena Scientific, 1997.
- [167] Drenick R. Multilinear programming: Duality theories [J/OL]. *Journal of Optimization Theory and Applications*, 1992, 72(3): 459-486. DOI: [10.1007/BF00939837](https://doi.org/10.1007/BF00939837).
- [168] Gao W, Goldfarb D, Curtis F E. ADMM for multiaffine constrained optimization [J/OL]. *Optimization Methods and Software*, 2020, 35(2): 257-303. DOI: [10.1080/10556788.2019.1683553](https://doi.org/10.1080/10556788.2019.1683553).
- [169] Kučera R. Convergence rate of an optimization algorithm for minimizing quadratic functions with separable convex constraints [J/OL]. *SIAM Journal on Optimization*, 2008, 19(2): 846-862. DOI: [10.1137/060670456](https://doi.org/10.1137/060670456).
- [170] He S, Li Z, Zhang S. Approximation algorithms for homogeneous polynomial optimization with quadratic constraints [J/OL]. *Mathematical Programming*, 2010, 125(2): 353-383. DOI: [10.1007/s10107-010-0409-z](https://doi.org/10.1007/s10107-010-0409-z).
- [171] Bonettini S, Prato M, Rebegoldi S. A block coordinate variable metric linesearch based proximal gradient method [J/OL]. *Computational Optimization and Applications*, 2018, 71(1): 5-52. DOI: [10.1007/s10589-018-0011-5](https://doi.org/10.1007/s10589-018-0011-5).
- [172] Liu B, Jiang C, Li G, et al. Topology optimization of structures considering local material uncertainties in additive manufacturing [J/OL]. *Computer Methods in Applied Mechanics and Engineering*, 2020, 360: 112786. DOI: [10.1016/j.cma.2019.112786](https://doi.org/10.1016/j.cma.2019.112786).
- [173] Bolte J, Sabach S, Teboulle M. Proximal alternating linearized minimization for nonconvex and nonsmooth problems [J/OL]. *Mathematical Programming*, 2014, 146(1-2): 459-494. DOI: [10.1007/s10107-013-0701-9](https://doi.org/10.1007/s10107-013-0701-9).
- [174] Razaviyayn M, Hong M, Luo Z Q. A unified convergence analysis of block successive minimization methods for nonsmooth optimization [J/OL]. *SIAM Journal on Optimization*, 2013, 23(2): 1126-1153. DOI: [10.1137/120891009](https://doi.org/10.1137/120891009).
- [175] Hua X, Yamashita N. Block coordinate proximal gradient methods with variable Bregman functions for nonsmooth separable optimization [J/OL]. *Mathematical Programming*, 2016, 160(1-2): 1-32. DOI: [10.1007/s10107-015-0969-z](https://doi.org/10.1007/s10107-015-0969-z).
- [176] Xu Y, Yin W. A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion [J/OL]. *SIAM Journal on Imaging Sciences*, 2013, 6(3): 1758-1789. DOI: [10.1137/120887795](https://doi.org/10.1137/120887795).
- [177] Yang Y, Pesavento M, Luo Z Q, et al. Inexact block coordinate descent algorithms for non-smooth nonconvex optimization [J/OL]. *IEEE Transactions on Signal Processing*, 2019, 68: 947-961. DOI: [10.1109/TSP.2019.2959240](https://doi.org/10.1109/TSP.2019.2959240).
- [178] Frankel P, Garrigos G, Peypouquet J. Splitting methods with variable metric for Kurdyka-Łojasiewicz functions and general convergence rates [J/OL]. *Journal of Optimization Theory and Applications*, 2015, 165(3): 874-900. DOI: [10.1007/s10957-014-0642-3](https://doi.org/10.1007/s10957-014-0642-3).

- [179] Ochs P. Unifying abstract inexact convergence theorems and block coordinate variable metric iPiano [J/OL]. SIAM Journal on Optimization, 2019, 29(1): 541-570. DOI: [10.1137/17M1124085](https://doi.org/10.1137/17M1124085).
- [180] Li X, Sun D, Toh K C. On the efficient computation of a generalized Jacobian of the projector over the Birkhoff polytope [J/OL]. Mathematical Programming, 2020, 179(1-2): 419-446. DOI: [10.1007/s10107-018-1342-9](https://doi.org/10.1007/s10107-018-1342-9).
- [181] Powell M J D, Yuan Y X. A trust region algorithm for equality constrained optimization [J/OL]. Mathematical Programming, 1991, 49(1-3): 189-211. DOI: [10.1007/BF01588787](https://doi.org/10.1007/BF01588787).
- [182] Jia Z, Cai X, Han D. Comparison of several fast algorithms for projection onto an ellipsoid [J/OL]. Journal of Computational and Applied Mathematics, 2017, 319: 320-337. DOI: [10.1016/j.cam.2017.01.008](https://doi.org/10.1016/j.cam.2017.01.008).
- [183] Dai Y H. Fast algorithms for projection on an ellipsoid [J/OL]. SIAM Journal on Optimization, 2006, 16(4): 986-1006. DOI: [10.1137/040613305](https://doi.org/10.1137/040613305).
- [184] Mangasarian O L, De Leone R. Error bounds for strongly convex programs and (super) linearly convergent iterative schemes for the least 2-norm solution of linear programs [J/OL]. Applied Mathematics and Optimization, 1988, 17(1): 1-14. DOI: [10.1007/BF01448356](https://doi.org/10.1007/BF01448356).
- [185] Bertsekas D P. A note on error bounds for convex and nonconvex programs [J/OL]. Computational Optimization and Applications, 1999, 12(1-3): 41-51. DOI: [10.1023/A:1008659512824](https://doi.org/10.1023/A:1008659512824).
- [186] Sun T, Jiang H, Cheng L, et al. A convergence framework for inexact nonconvex and non-smooth algorithms and its applications to several iterations [A/OL]. arXiv preprint, 2017, arXiv:1709.04072. <https://arxiv.org/abs/1709.04072>.
- [187] Li X, Milzarek A, Qiu J. Convergence of random reshuffling under the Kurdyka–Łojasiewicz inequality [J/OL]. SIAM Journal on Optimization, 2023, 33(2): 1092-1120. DOI: [10.1137/21M1468048](https://doi.org/10.1137/21M1468048).
- [188] Yang L. Proximal gradient method with extrapolation and line search for a class of non-convex and non-smooth problems [J/OL]. Journal of Optimization Theory and Applications, 2024, 200(1): 68-103. DOI: [10.1007/s10957-023-02348-4](https://doi.org/10.1007/s10957-023-02348-4).
- [189] Attouch H, Bolte J. On the convergence of the proximal algorithm for nonsmooth functions involving analytic features [J/OL]. Mathematical Programming, 2009, 116(1-2): 5-16. DOI: [10.1007/s10107-007-0133-5](https://doi.org/10.1007/s10107-007-0133-5).
- [190] Łojasiewicz S. Sur la géométrie semi-et sous-analytique [J/OL]. Annales de l’Institut Fourier, 1993, 43(5): 1575-1595. DOI: [10.5802/aif.1384](https://doi.org/10.5802/aif.1384).
- [191] Kurdyka K. On gradients of functions definable in o-minimal structures [J/OL]. Annales de l’Institut Fourier, 1998, 48(3): 769-783. DOI: [10.5802/aif.1638](https://doi.org/10.5802/aif.1638).
- [192] Bolte J, Daniilidis A, Lewis A. The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems [J/OL]. SIAM Journal on Optimization, 2007, 17(4): 1205-1223. DOI: [10.1137/050644641](https://doi.org/10.1137/050644641).
- [193] Attouch H, Bolte J, Redont P, et al. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka–Łojasiewicz inequality [J/OL]. Mathematics of Operations Research, 2010, 35(2): 438-457. DOI: [10.1287/moor.1100.0449](https://doi.org/10.1287/moor.1100.0449).
- [194] Łojasiewicz S. Une propriété topologique des sous-ensembles analytiques réels [C]//Les Équations aux Dérivées Partielles: volume 117 Éditions du Centre National de la Recherche Scientifique. 1963: 87-89.

- [195] Chill R. On the Łojasiewicz–Simon gradient inequality [J/OL]. Journal of Functional Analysis, 2003, 201(2): 572-601. DOI: [10.1016/S0022-1236\(02\)00102-7](https://doi.org/10.1016/S0022-1236(02)00102-7).
- [196] Polyak B T. Translations Series in Mathematics and Engineering: Introduction to Optimization [M]. New York: Optimization Software, Inc., Publications Division, 1987.
- [197] Xu J. Theory of Multilevel Methods [M]. Ithaca: Cornell University, 1989.
- [198] Briggs W L, Henson V E, McCormick S F. Other Titles in Applied Mathematics: A Multigrid Tutorial [M/OL]. 2nd ed. Philadelphia: Society for Industrial and Applied Mathematics, 2000. DOI: [10.1137/1.9780898719505](https://doi.org/10.1137/1.9780898719505).
- [199] Trottenberg U, Oosterlee C W, Schuller A. Multigrid [M]. London: Academic Press, 2001.
- [200] Xu J, Zikatanov L. Algebraic multigrid methods [J/OL]. Acta Numerica, 2017, 26: 591-721. DOI: [10.1017/S0962492917000083](https://doi.org/10.1017/S0962492917000083).
- [201] Borzì A, Schulz V. Computational Science & Engineering: Computational Optimization of Systems Governed by Partial Differential Equations [M/OL]. Philadelphia: Society for Industrial and Applied Mathematics, 2011. DOI: [10.1137/1.9781611972054](https://doi.org/10.1137/1.9781611972054).
- [202] Schöberl J, Simon R, Zulehner W. A robust multigrid method for elliptic optimal control problems [J/OL]. SIAM Journal on Numerical Analysis, 2011, 49(4): 1482-1503. DOI: [10.1137/100783285](https://doi.org/10.1137/100783285).
- [203] Takacs S, Zulehner W. Convergence analysis of all-at-once multigrid methods for elliptic control problems under partial elliptic regularity [J/OL]. SIAM Journal on Numerical Analysis, 2013, 51(3): 1853-1874. DOI: [10.1137/120880884](https://doi.org/10.1137/120880884).
- [204] Takacs S. A robust all-at-once multigrid method for the Stokes control problem [J/OL]. Numerische Mathematik, 2015, 130(3): 517-540. DOI: [10.1007/s00211-014-0674-5](https://doi.org/10.1007/s00211-014-0674-5).
- [205] Dreyer T, Maar B, Schulz V. Multigrid optimization in applications [J/OL]. Journal of Computational and Applied Mathematics, 2000, 120(1-2): 67-84. DOI: [10.1016/S0377-0427\(00\)00304-6](https://doi.org/10.1016/S0377-0427(00)00304-6).
- [206] Maar B, Schulz V. Interior point multigrid methods for topology optimization [J/OL]. Structural and Multidisciplinary Optimization, 2000, 19(3): 214-224. DOI: [10.1007/s001580050104](https://doi.org/10.1007/s001580050104).
- [207] Hu J, Luo H, Zhang Z. A fast solver for generalized optimal transport problems based on dynamical system and algebraic multigrid [J/OL]. Journal of Scientific Computing, 2023, 97(6): 1-30. DOI: [10.1007/s10915-023-02272-9](https://doi.org/10.1007/s10915-023-02272-9).
- [208] Nash S G. A multigrid approach to discretized optimization problems [J/OL]. Optimization Methods and Software, 2000, 14(1-2): 99-116. DOI: [10.1080/10556780008805795](https://doi.org/10.1080/10556780008805795).
- [209] Lewis R M, Nash S G. Model problems for the multigrid optimization of systems governed by differential equations [J/OL]. SIAM Journal on Scientific Computing, 2005, 26(6): 1811-1837. DOI: [10.1137/S1064827502407792](https://doi.org/10.1137/S1064827502407792).
- [210] Gratton S, Sartenaer A, Toint P L. Recursive trust-region methods for multiscale nonlinear optimization [J/OL]. SIAM Journal on Optimization, 2008, 19(1): 414-444. DOI: [10.1137/050623012](https://doi.org/10.1137/050623012).
- [211] Wen Z, Goldfarb D. A line search multigrid method for large-scale nonlinear optimization [J/OL]. SIAM Journal on Optimization, 2010, 20(3): 1478-1503. DOI: [10.1137/08071524X](https://doi.org/10.1137/08071524X).
- [212] Ziems J C, Ulbrich S. Adaptive multilevel inexact SQP methods for PDE-constrained optimization [J/OL]. SIAM Journal on Optimization, 2011, 21(1): 1-40. DOI: [10.1137/080743160](https://doi.org/10.1137/080743160).

- [213] Frandi E, Papini A. Improving direct search algorithms by multilevel optimization techniques [J/OL]. Optimization Methods and Software, 2015, 30(5): 1077-1094. DOI: [10.1080/10556788.2015.1014555](https://doi.org/10.1080/10556788.2015.1014555).
- [214] Chen C, Wen Z, Yuan Y X. A general two-level subspace method for nonlinear optimization [J/OL]. Journal of Computational Mathematics, 2018, 36(6): 881-902. DOI: [10.4208/jcm.1706-m2016-0721](https://doi.org/10.4208/jcm.1706-m2016-0721).
- [215] Deuflhard P. Cascadic conjugate gradient methods for elliptic partial differential equations: Algorithm and numerical results [M/OL]//Keyes D E, Xu J. Contemporary Mathematics: volume 180 Domain Decomposition Methods in Scientific and Engineering Computing. Providence: American Mathematical Society, 1994: 29-42. DOI: [10.1090/conm/180/01954](https://doi.org/10.1090/conm/180/01954).
- [216] Bornemann F A, Deuflhard P. The cascadic multigrid method for elliptic problems [J/OL]. Numerische Mathematik, 1996, 75(2): 135-152. DOI: [10.1007/s002110050234](https://doi.org/10.1007/s002110050234).
- [217] Bornemann F A, Krause R. Classical and cascadic multigrid-A methodical comparison [M/OL]//Bjørstad P E, Espedal M S, Keyes D E. Proceedings of the 9th International Conference on Domain Decomposition Methods. Chichester: John Wiley & Sons Ltd., 1997: 64-71. <http://www.ddm.org/DD9/Bornemann.pdf>.
- [218] Borzì A, Hohenester U. Multigrid optimization schemes for solving Bose–Einstein condensate control problems [J/OL]. SIAM Journal on Scientific Computing, 2008, 30(1): 441-462. DOI: [10.1137/070686135](https://doi.org/10.1137/070686135).
- [219] Wu X, Wen Z, Bao W. A regularized Newton method for computing ground states of Bose–Einstein condensates [J/OL]. Journal of Scientific Computing, 2017, 73(1): 303-329. DOI: [10.1007/s10915-017-0412-0](https://doi.org/10.1007/s10915-017-0412-0).
- [220] Tian T, Cai Y, Wu X, et al. Ground states of spin–F Bose–Einstein condensates [J/OL]. SIAM Journal on Scientific Computing, 2020, 42(4): B983-B1013. DOI: [10.1137/19M1271117](https://doi.org/10.1137/19M1271117).
- [221] Zhou D, Chen H, Ho C H, et al. A multilevel method for many-electron Schrödinger equations based on the atomic cluster expansion [J/OL]. SIAM Journal on Scientific Computing, 2024, 46(1): A105-A129. DOI: [10.1137/23M1565887](https://doi.org/10.1137/23M1565887).
- [222] Chen J, García-Cervera C J. An efficient multigrid strategy for large-scale molecular mechanics optimization [J/OL]. Journal of Computational Physics, 2017, 342: 29-42. DOI: [10.1016/j.jcp.2017.04.035](https://doi.org/10.1016/j.jcp.2017.04.035).
- [223] Xia Q, Shi T. A cascadic multilevel optimization algorithm for the design of composite structures with curvilinear fiber based on Shepard interpolation [J/OL]. Composite Structures, 2018, 188: 209-219. DOI: [10.1016/j.compstruct.2018.01.013](https://doi.org/10.1016/j.compstruct.2018.01.013).
- [224] Liu J, Yin W, Li W, et al. Multilevel optimal transport: A fast approximation of Wasserstein–1 distances [J/OL]. SIAM Journal on Scientific Computing, 2021, 43(1): A193-A220. DOI: [10.1137/18M1219813](https://doi.org/10.1137/18M1219813).
- [225] Schmitzer B. A sparse algorithm for dense optimal transport [M/OL]//Aujol J F, Nikolova M, Papadakis N. Lecture Notes in Computer Science: volume 9087 Scale Space and Variational Methods in Computer Vision: 5th International Conference, SSVM 2015, Lège-Cap Ferret, France, May 31 - June 4, 2015, Proceedings. Cham: Springer, 2015: 629-641. DOI: [10.1007/978-3-319-18461-6_50](https://doi.org/10.1007/978-3-319-18461-6_50).
- [226] Gerber S, Maggioni M. Multiscale strategies for computing optimal transport [J/OL]. Journal of Machine Learning Research, 2017, 18(72): 2440-2471. <http://jmlr.org/papers/v18/16-108.html>.

- [227] Hickernell F J, Yuan Y X. A simple multistart algorithm for global optimization [J]. *Operations Research Transactions*, 1999, 1(2): 1-12.
- [228] Fourer R, Gay D M, Kernighan B W. A modeling language for mathematical programming [J/OL]. *Management Science*, 1990, 36(5): 519-554. DOI: [10.1287/mnsc.36.5.519](https://doi.org/10.1287/mnsc.36.5.519).
- [229] Tawarmalani M, Sahinidis N V. A polyhedral branch-and-cut approach to global optimization [J/OL]. *Mathematical Programming*, 2005, 103(2): 225-249. DOI: [10.1007/s10107-005-0581-8](https://doi.org/10.1007/s10107-005-0581-8).
- [230] Hecht F. New development in FreeFem++ [J/OL]. *Journal of Numerical Mathematics*, 2012, 20(3-4): 251-266. DOI: [10.1515/jnum-2012-0013](https://doi.org/10.1515/jnum-2012-0013).
- [231] Carlier G, Oberman A, Oudet E. Numerical methods for matching for teams and Wasserstein barycenters [J/OL]. *ESAIM: Mathematical Modelling and Numerical Analysis*, 2015, 49(6): 1621-1642. DOI: [10.1051/m2an/2015033](https://doi.org/10.1051/m2an/2015033).
- [232] Bigot J, Klein T. Consistent estimation of a population barycenter in the Wasserstein space [A/OL]. arXiv preprint, 2012, arXiv:1212.2562. <https://arxiv.org/abs/1212.2562>.
- [233] Cuturi M, Doucet A. Fast computation of Wasserstein barycenters [C/OL]//Xing E P, Jebara T. *Proceedings of Machine Learning Research*: volume 32 Proceedings of the 31st International Conference on Machine Learning. PMLR, 2014: 685-693. <http://proceedings.mlr.press/v32/cuturi14.html>.
- [234] Geng X. Label distribution learning [J/OL]. *IEEE Transactions on Knowledge and Data Engineering*, 2016, 28(7): 1734-1748. DOI: [10.1109/TKDE.2016.2545658](https://doi.org/10.1109/TKDE.2016.2545658).
- [235] Zhao P, Zhou Z H. Label distribution learning by optimal transport [C/OL]//McIlraith S A, Weinberger K Q. *Proceedings of the AAAI Conference on Artificial Intelligence*: volume 32 Proceedings of the 32nd AAAI Conference on Artificial Intelligence. AAAI Press, 2018: 4506-4513. DOI: [10.1609/aaai.v32i1.11609](https://doi.org/10.1609/aaai.v32i1.11609).
- [236] Bertsekas D P. Nonlinear Programming [M]. 3rd ed. Belmont: Athena Scientific, 2016.
- [237] Fercoq O, Richtárik P. Optimization in high dimensions via accelerated, parallel, and proximal coordinate descent [J/OL]. *SIAM Review*, 2016, 58(4): 739-771. DOI: [10.1137/16M1085905](https://doi.org/10.1137/16M1085905).
- [238] Wright S J. Coordinate descent algorithms [J/OL]. *Mathematical Programming*, 2015, 151 (1): 3-34. DOI: [10.1007/s10107-015-0892-3](https://doi.org/10.1007/s10107-015-0892-3).
- [239] Beck A, Pauwels E, Sabach S. The cyclic block conditional gradient method for convex optimization problems [J/OL]. *SIAM Journal on Optimization*, 2015, 25(4): 2024-2049. DOI: [10.1137/15M1008397](https://doi.org/10.1137/15M1008397).
- [240] Braun G, Carderera A, Combettes C W, et al. Conditional gradient methods [A/OL]. arXiv preprint, 2022, arXiv:2211.14103. <https://arxiv.org/abs/2211.14103>.
- [241] Chen C, Li M, Liu X, et al. Extended ADMM and BCD for nonseparable convex minimization models with quadratic coupling terms: Convergence analysis and insights [J/OL]. *Mathematical Programming*, 2019, 173(1-2): 37-77. DOI: [10.1007/s10107-017-1205-9](https://doi.org/10.1007/s10107-017-1205-9).
- [242] Driggs D, Tang J, Liang J, et al. A stochastic proximal alternating minimization for nonsmooth and nonconvex optimization [J/OL]. *SIAM Journal on Imaging Sciences*, 2021, 14(4): 1932-1970. DOI: [10.1137/20M1387213](https://doi.org/10.1137/20M1387213).
- [243] Hertrich J, Steidl G. Inertial stochastic PALM and applications in machine learning [J/OL]. *Sampling Theory, Signal Processing, and Data Analysis*, 2022, 20(1): 4. DOI: [10.1007/s43670-022-00021-x](https://doi.org/10.1007/s43670-022-00021-x).

- [244] Lacoste-Julien S, Jaggi M, Schmidt M, et al. Block-coordinate Frank-Wolfe optimization for structural SVMs [C/OL]//Dasgupta S, McAllester D. Proceedings of Machine Learning Research: volume 28 Proceedings of the 30th International Conference on Machine Learning. PMLR, 2013: 53-61. <https://proceedings.mlr.press/v28/lacoste-julien13.html>.
- [245] Sun R, Luo Z Q, Ye Y. On the efficiency of random permutation for ADMM and coordinate descent [J/OL]. Mathematics of Operations Research, 2020, 45(1): 233-271. DOI: [10.1287/moor.2019.0990](https://doi.org/10.1287/moor.2019.0990).
- [246] Monge G. Mémoire sur la théorie des déblais et des remblais [J]. Histoire de l'Académie Royale des Sciences, 1781: 666-704.
- [247] Kantorovitch L. On the translocation of masses [J]. Comptes Rendus (Doklady) de l'Académie des Sciences de l'URSS (Nouvelle Série), 1942, 37: 199-201.
- [248] Brenier Y. A homogenized model for vortex sheets [J/OL]. Archive for Rational Mechanics and Analysis, 1997, 138(4): 319-353. DOI: [10.1007/s002050050044](https://doi.org/10.1007/s002050050044).
- [249] Benamou J D, Brenier Y, Guittet K. The Monge-Kantorovitch mass transfer and its computational fluid mechanics formulation [J/OL]. International Journal for Numerical Methods in Fluids, 2002, 40(1-2): 21-30. DOI: [10.1002/fld.264](https://doi.org/10.1002/fld.264).
- [250] Rubner Y, Guibas L J, Tomasi C. The earth mover's distance, multi-dimensional scaling, and color-based image retrieval [C]//Proceedings of the ARPA Image Understanding Workshop. ARPA, 1997: 661-668.
- [251] Pele O, Werman M. Fast and robust earth mover's distances [C/OL]//IEEE 12th International Conference on Computer Vision. IEEE, 2009: 460-467. DOI: [10.1109/ICCV.2009.5459199](https://doi.org/10.1109/ICCV.2009.5459199).
- [252] Cuturi M. Sinkhorn distances: Lightspeed computation of optimal transport [C/OL]//Burges C J, Bottou L, Welling M, et al. Advances in Neural Information Processing Systems: volume 26. Curran Associates, Inc., 2013: 2292-2300. https://papers.nips.cc/paper_files/paper/2013/hash/af21d0c97db2e27e13572cbf59eb343d-Abstract.html.
- [253] Sinkhorn R. A relationship between arbitrary positive matrices and doubly stochastic matrices [J/OL]. Annals of Mathematical Statistics, 1964, 35: 876-879. DOI: [10.1214/aoms/1177703591](https://doi.org/10.1214/aoms/1177703591).
- [254] Sinkhorn R, Knopp P. Concerning nonnegative matrices and doubly stochastic matrices [J/OL]. Pacific Journal of Mathematics, 1967, 21(2): 343-348. DOI: [10.2307/2314570](https://doi.org/10.2307/2314570).
- [255] Altschuler J, Bach F, Rudi A, et al. Massively scalable Sinkhorn distances via the Nyström method [C/OL]//Wallach H, Larochelle H, Beygelzimer A, et al. Advances in Neural Information Processing Systems: volume 32. Curran Associates, Inc., 2019: 4427-4437. https://papers.nips.cc/paper_files/paper/2019/hash/f55cadb97eaff2ba1980e001b0bd9842-Abstract.html.
- [256] Li M, Yu J, Li T, et al. Importance sparsification for Sinkhorn algorithm [J/OL]. Journal of Machine Learning Research, 2023, 24(247): 1-44. <http://jmlr.org/papers/v24/22-1311.html>.
- [257] Filatov M. Spin-restricted ensemble-referenced Kohn-Sham method: Basic principles and application to strongly correlated ground and excited states of molecules [J/OL]. Wiley Interdisciplinary Reviews: Computational Molecular Science, 2015, 5(1): 146-167. DOI: [10.1002/wcms.1209](https://doi.org/10.1002/wcms.1209).
- [258] Shannon C E. A mathematical theory of communication [J/OL]. The Bell system Technical Journal, 1948, 27(3): 379-423. DOI: [10.1002/j.1538-7305.1948.tb01338.x](https://doi.org/10.1002/j.1538-7305.1948.tb01338.x).
- [259] Luo Z Q, Tseng P. On the convergence rate of dual ascent methods for linearly constrained

- convex minimization [J/OL]. Mathematics of Operations Research, 1993, 18(4): 846-867. DOI: [10.1287/moor.18.4.846](https://doi.org/10.1287/moor.18.4.846).
- [260] Xie Y, Wang X, Wang R, et al. A fast proximal point method for computing exact Wasserstein distance [C/OL]//Adams R P, Gogate V. Proceedings of Machine Learning Research: volume 115 Proceedings of the 35th Uncertainty in Artificial Intelligence Conference. PMLR, 2020: 433-453. <http://proceedings.mlr.press/v115/xie20b.html>.
- [261] Hosseini B, Steinerberger S. Intrinsic sparsity of Kantorovich solutions [J/OL]. Comptes Rendus. Mathématique, Tome, 2022, 360(G10): 1173-1175. DOI: [10.5802/crmath.392](https://doi.org/10.5802/crmath.392).
- [262] Drineas P, Zouzias A. A note on element-wise matrix sparsification via a matrix-valued Bernstein inequality [J/OL]. Information Processing Letters, 2011, 111(8): 385-389. DOI: [10.1016/j.ipl.2011.01.010](https://doi.org/10.1016/j.ipl.2011.01.010).
- [263] Achlioptas D, Karnin Z S, Liberty E. Near-optimal entrywise sampling for data matrices [C/OL]//Burgers C J, Bottou L, Welling M, et al. Advances in Neural Information Processing Systems: volume 26. Curran Associates, Inc., 2013: 1565-1573. https://proceedings.neurips.cc/paper_files/paper/2013/hash/6e0721b2c6977135b916ef286bcb49ec-Abstract.html.
- [264] Kundu A, Drineas P, Magdon-Ismail M. Recovering PCA and sparse PCA via hybrid- (ℓ_1, ℓ_2) sparse sampling of data elements [J/OL]. Journal of Machine Learning Research, 2017, 18(75): 1-34. <https://jmlr.csail.mit.edu/papers/v18/16-258.html>.
- [265] Braverman V, Krauthgamer R, Krishnan A R, et al. Near-optimal entrywise sampling of numerically sparse matrices [C/OL]//Belkin M, Kpotufe S. Proceedings of Machine Learning Research: volume 134 Proceedings of the 34th Conference on Learning Theory. PMLR, 2021: 759-773. <https://proceedings.mlr.press/v134/braverman21b.html>.
- [266] Li M, Yu J, Xu H, et al. Efficient approximation of Gromov-Wasserstein distance using importance sparsification [J/OL]. Journal of Computational and Graphical Statistics, 2023, 32(4): 1512-1523. DOI: [10.1080/10618600.2023.2165500](https://doi.org/10.1080/10618600.2023.2165500).
- [267] Achlioptas D, McSherry F. Fast computation of low-rank matrix approximations [J/OL]. Journal of the ACM (JACM), 2007, 54(2): Art. 9, 19. DOI: [10.1145/1219092.1219097](https://doi.org/10.1145/1219092.1219097).
- [268] Liu J S. Metropolized independent sampling with comparisons to rejection sampling and importance sampling [J/OL]. Statistics and Computing, 1996, 6(2): 113-119. DOI: [10.1007/BF00162521](https://doi.org/10.1007/BF00162521).
- [269] Liu J S. Springer Series in Statistics: Monte Carlo Strategies in Scientific Computing [M/OL]. New York: Springer, 2004. DOI: [10.1007/978-0-387-76371-2](https://doi.org/10.1007/978-0-387-76371-2).
- [270] Owen A B. Monte Carlo Theory, Methods and Examples [M]. Stanford: Stanford University, 2013.
- [271] Elvira V, Martino L. Advances in importance sampling [J/OL]. Wiley StatsRef: Statistics Reference Online, 2021, 1: 1-14. DOI: [10.1002/9781118445112.stat08284](https://doi.org/10.1002/9781118445112.stat08284).
- [272] Ai M, Wang F, Yu J, et al. Optimal subsampling for large-scale quantile regression [J/OL]. Journal of Complexity, 2021, 62: 101512. DOI: [10.1016/j.jco.2020.101512](https://doi.org/10.1016/j.jco.2020.101512).
- [273] Wang J, Zou J, Wang H. Sampling with replacement vs Poisson sampling: A comparative study in optimal subsampling [J/OL]. IEEE Transactions on Information Theory, 2022, 68 (10): 6605-6630. DOI: [10.1109/TIT.2022.3176955](https://doi.org/10.1109/TIT.2022.3176955).
- [274] Ma P, Mahoney M, Yu B. A statistical perspective on algorithmic leveraging [C/OL]//Xing E P, Jebara T. Proceedings of Machine Learning Research: volume 32 Proceedings of the 31st International Conference on Machine Learning. PMLR, 2014: 91-99. <http://proceedings.mlr.press/v32/ma14.html>.

- [275] Yu J, Wang H, Ai M, et al. Optimal distributed subsampling for maximum quasi-likelihood estimators with massive data [J/OL]. *Journal of the American Statistical Association*, 2022, 117(537): 265-276. DOI: [10.1080/01621459.2020.1773832](https://doi.org/10.1080/01621459.2020.1773832).
- [276] Brègman L M. Relaxation method for finding a common point of convex sets and its application to optimization problems [J]. *Doklady Akademii Nauk SSSR*, 1966, 171: 1019-1022.
- [277] Yang L, Toh K C. Bregman proximal point algorithm revisited: A new inexact version and its inertial variant [J/OL]. *SIAM Journal on Optimization*, 2022, 32(3): 1523-1554. DOI: [10.1137/20M1360748](https://doi.org/10.1137/20M1360748).
- [278] Li Q, Zhu Z, Tang G, et al. Provable Bregman-divergence based methods for nonconvex and non-Lipschitz problems [A/OL]. arXiv preprint, 2019, arXiv:1904.09712. <https://arxiv.org/abs/1904.09712>.
- [279] Ahookhosh M, Hien L T K, Gillis N, et al. Multi-block Bregman proximal alternating linearized minimization and its application to orthogonal nonnegative matrix factorization [J/OL]. *Computational Optimization and Applications*, 2021, 79(3): 681-715. DOI: [10.1007/s10589-021-00286-3](https://doi.org/10.1007/s10589-021-00286-3).
- [280] Kullback S, Leibler R A. On information and sufficiency [J/OL]. *The Annals of Mathematical Statistics*, 1951, 22(1): 79-86. DOI: [10.1214/aoms/1177729694](https://doi.org/10.1214/aoms/1177729694).
- [281] Kerdoncuff T, Emonet R, Sebban M. Sampled Gromov Wasserstein [J/OL]. *Machine Learning*, 2021, 110(8): 2151-2186. DOI: [10.1007/s10994-021-06035-1](https://doi.org/10.1007/s10994-021-06035-1).
- [282] Wang H, Zou J. A comparative study on sampling with replacement vs Poisson sampling in optimal subsampling [C/OL]//Banerjee A, Fukumizu K. *Proceedings of Machine Learning Research*: volume 130 Proceedings of the 24th International Conference on Artificial Intelligence and Statistics. PMLR, 2021: 289-297. <https://proceedings.mlr.press/v130/wang21a.html>.
- [283] Dvurechensky P, Gasnikov A, Kroshnin A. Computational optimal transport: Complexity by accelerated gradient descent is better than by Sinkhorn's algorithm [C/OL]//Dy J, Krause A. *Proceedings of Machine Learning Research*: volume 80 Proceedings of the 35th International Conference on Machine Learning. PMLR, 2018: 1367-1376. <http://proceedings.mlr.press/v80/dvurechensky18a.html>.
- [284] Zhang H, Hager W W. A nonmonotone line search technique and its application to unconstrained optimization [J/OL]. *SIAM Journal on Optimization*, 2004, 14(4): 1043-1056. DOI: [10.1137/S1052623403428208](https://doi.org/10.1137/S1052623403428208).
- [285] Hu Y, Gao X, Zhao Y, et al. Force-based gradient descent method for *ab initio* atomic structure relaxation [J/OL]. *Physical Review B*, 2022, 106(10): 104101. DOI: [10.1103/PhysRevB.106.104101](https://doi.org/10.1103/PhysRevB.106.104101).
- [286] Grippo L, Lampariello F, Lucidi S. A nonmonotone line search technique for Newton's method [J/OL]. *SIAM Journal on Numerical Analysis*, 1986, 23(4): 707-716. DOI: [10.1137/0723046](https://doi.org/10.1137/0723046).
- [287] Fang J, Gao X, Song H, et al. On the existence of the optimal order for wavefunction extrapolation in Born-Oppenheimer molecular dynamics [J/OL]. *The Journal of Chemical Physics*, 2016, 144(24): 244103. DOI: [10.1063/1.4954234](https://doi.org/10.1063/1.4954234).
- [288] Gao X, Mo Z, Fang J, et al. Parallel 3-dim fast Fourier transforms with load balancing of the plane waves [J/OL]. *Computer Physics Communications*, 2017, 211: 54-60. DOI: [10.1016/j.cpc.2016.07.001](https://doi.org/10.1016/j.cpc.2016.07.001).

- [289] Zhou Y, Wang H, Liu Y, et al. Applicability of Kerker preconditioning scheme to the self-consistent density functional theory calculations of inhomogeneous systems [J/OL]. Physical Review E, 2018, 97(3): 033305. DOI: [10.1103/PhysRevE.97.033305](https://doi.org/10.1103/PhysRevE.97.033305).
- [290] Fang J, Gao X, Song H. Implementation of the projector augmented-wave method: The use of atomic datasets in the standard PAW-XML format [J/OL]. Communications in Computational Physics, 2019, 26(4): 1196-1223. DOI: [10.4208/cicp.OA-2018-0302](https://doi.org/10.4208/cicp.OA-2018-0302).
- [291] Perdew J P, Burke K, Ernzerhof M. Generalized gradient approximation made simple [J/OL]. Physical Review Letters, 1996, 77(18): 3865. DOI: [10.1103/PhysRevLett.77.3865](https://doi.org/10.1103/PhysRevLett.77.3865).
- [292] Jollet F, Torrent M, Holzwarth N. Generation of projector augmented-wave atomic data: A 71 element validated table in the XML format [J/OL]. Computer Physics Communications, 2014, 185(4): 1246-1254. DOI: [10.1016/j.cpc.2013.12.023](https://doi.org/10.1016/j.cpc.2013.12.023).
- [293] Monkhorst H J, Pack J D. Special points for Brillouin-zone integrations [J/OL]. Physical Review B, 1976, 13(12): 5188. DOI: [10.1103/PhysRevB.13.5188](https://doi.org/10.1103/PhysRevB.13.5188).
- [294] Jain A, Ong S P, Hautier G, et al. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation [J/OL]. APL Materials, 2013, 1(1): 011002. DOI: [10.1063/1.4812323](https://doi.org/10.1063/1.4812323).
- [295] Borysov S S, Geilhufe R M, Balatsky A V. Organic materials database: An open-access online database for data mining [J/OL]. PLoS ONE, 2017, 12(2): e0171501. DOI: [10.1371/journal.pone.0171501](https://doi.org/10.1371/journal.pone.0171501).
- [296] Dolan E D, Moré J J. Benchmarking optimization software with performance profiles [J/OL]. Mathematical Programming, 2002, 91(2): 201-213. DOI: [10.1007/s101070100263](https://doi.org/10.1007/s101070100263).
- [297] Murnaghan F D. The compressibility of media under extreme pressures [J/OL]. Proceedings of the National Academy of Sciences, 1944, 30(9): 244-247. DOI: [10.1073/pnas.30.9.244](https://doi.org/10.1073/pnas.30.9.244).
- [298] Birch F. Finite elastic strain of cubic crystals [J/OL]. Physical Review, 1947, 71(11): 809. DOI: [10.1103/PhysRev.71.809](https://doi.org/10.1103/PhysRev.71.809).
- [299] Zhang Y, Zuo T T, Tang Z, et al. Microstructures and properties of high-entropy alloys [J/OL]. Progress in Materials Science, 2014, 61: 1-93. DOI: [10.1016/j.pmatsci.2013.10.001](https://doi.org/10.1016/j.pmatsci.2013.10.001).
- [300] Song H, Tian F, Hu Q M, et al. Local lattice distortion in high-entropy alloys [J/OL]. Physical Review Materials, 2017, 1(2): 023404. DOI: [10.1103/PhysRevMaterials.1.023404](https://doi.org/10.1103/PhysRevMaterials.1.023404).
- [301] Tian F, Lin D Y, Gao X, et al. A structural modeling approach to solid solutions based on the similar atomic environment [J/OL]. The Journal of Chemical Physics, 2020, 153(3): 034101. DOI: [10.1063/5.0014094](https://doi.org/10.1063/5.0014094).
- [302] Conforti M, Cornuéjols G, Zambelli G. Graduate Texts in Mathematics: volume 271 Integer Programming [M/OL]. Cham: Springer, 2014. DOI: [10.1007/978-3-319-11008-0](https://doi.org/10.1007/978-3-319-11008-0).
- [303] Chandler D. Introduction to Modern Statistical Mechanics [M]. New York: Oxford University Press, 1987.
- [304] Girifalco L A. Monographs on the Physics and Chemistry of Materials: volume 58 Statistical Mechanics of Solids [M]. New York: Oxford University Press, 2003.
- [305] Otero-de-la Roza A, Luaña V. Gibbs2: A new version of the quasi-harmonic model code. I. Robust treatment of the static data [J/OL]. Computer Physics Communications, 2011, 182(8): 1708-1720. DOI: [10.1016/j.cpc.2011.04.016](https://doi.org/10.1016/j.cpc.2011.04.016).
- [306] Wu J, Yang Z, Xian J, et al. Structural and thermodynamic properties of the high-entropy alloy AlCoCrFeNi based on first-principles calculations [J/OL]. Frontiers in Materials, 2020, 7: 590143. DOI: [10.3389/fmats.2020.590143](https://doi.org/10.3389/fmats.2020.590143).

- [307] Ge H, Song H, Shen J, et al. Effect of alloying on the thermal-elastic properties of 3d high-entropy alloys [J/OL]. Materials Chemistry and Physics, 2018, 210: 320-326. DOI: [10.1016/j.matchemphys.2017.10.046](https://doi.org/10.1016/j.matchemphys.2017.10.046).
- [308] Song H F, Liu H F. Modified mean-field potential approach to thermodynamic properties of a low-symmetry crystal: Beryllium as a prototype [J/OL]. Physical Review B, 2007, 75(24): 245126. DOI: [10.1103/PhysRevB.75.245126](https://doi.org/10.1103/PhysRevB.75.245126).
- [309] Cheng B, Zhang F, Lou H, et al. Pressure-induced phase transition in the AlCoCrFeNi high-entropy alloy [J/OL]. Scripta Materialia, 2019, 161: 88-92. DOI: [10.1016/j.scriptamat.2018.10.020](https://doi.org/10.1016/j.scriptamat.2018.10.020).

附录一 带 Coulomb 费用多边际最优运输问题的离散化

我们介绍在类 Monge 拟设下 MMOT (1.18) 的离散化. 我们重新陈述该问题如下:

$$\inf_{\{\gamma_i\}_{i=2}^N} \sum_{i=2}^N \int_{\Omega^2} \frac{\rho(\mathbf{r})\gamma_i(\mathbf{r}, \mathbf{r}')}{\|\mathbf{r} - \mathbf{r}'\|} d\mathbf{r} d\mathbf{r}' + \sum_{2 \leq i < j}^N \int_{\Omega^3} \frac{\rho(\mathbf{r})\gamma_i(\mathbf{r}, \mathbf{r}')\gamma_j(\mathbf{r}, \mathbf{r}'')}{\|\mathbf{r}' - \mathbf{r}''\|} d\mathbf{r} d\mathbf{r}' d\mathbf{r}'',$$

s. t. $\int_{\Omega} \gamma_i(\cdot, \mathbf{r}_i) d\mathbf{r}_i = 1, \int_{\Omega} \gamma_i(\mathbf{r}_1, \cdot) \rho(\mathbf{r}_1) d\mathbf{r}_1 = \rho, \gamma_i \geq 0, i = 2, \dots, N.$ (附 1.1)

注意到与之前的问题 (1.25) 相比, 问题 (附 1.1) 中函数的积分区间被截断成了一个有界集合 $\Omega \subseteq \mathbb{R}^d$. 这在实际使用中是合理的.

我们首先使用有限元网格 $\mathcal{T} = \{e_k\}_{k=1}^K$ 将 Ω 划分成 $K \in \mathbb{N}$ 个互不重叠的单元, 即满足 $\cup_{k=1}^K e_k = \Omega$ 且当 $k \neq k'$, $e_k \cap e_{k'} = \emptyset$. 我们使用有限个 Dirac 测度近似单电子密度 ρ : $\rho = \sum_{k=1}^K \varrho_k \delta_{\mathbf{d}_k}$, 其中

$$\varrho_k := \int_{e_k} \rho(\mathbf{r}) d\mathbf{r}, \quad k = 1, \dots, K,$$

$\mathbf{d}_k \in \mathbb{R}^d$ 是单元 e_k 的重心 (barycenter) ($k = 1, \dots, K$), 将向量 $\mathbf{p} := [\varrho_1, \dots, \varrho_K] \in \mathbb{R}_+^K$ 作为 ρ 的离散. 相应地, 两体 Coulomb 势则离散为矩阵 $C = (c_{ij}) \in \mathbb{R}^{K \times K}$, 定义为

$$c_{ij} := \begin{cases} \|\mathbf{d}_i - \mathbf{d}_j\|^{-1}, & \text{若 } i \neq j; \\ 0, & \text{否则,} \end{cases} \quad (\text{附 1.2})$$

而耦合函数 γ_i 则离散成矩阵 $Y_i = (y_{i,jk})_{jk} \in \mathbb{R}^{K \times K}$, 定义为

$$y_{i,jk} := \frac{1}{|e_j|} \int_{e_j} \int_{e_k} \gamma_i(\mathbf{r}, \mathbf{r}') d\mathbf{r}' d\mathbf{r}, \quad j, k = 1, \dots, K, \quad i = 2, \dots, N.$$

注意在矩阵 C 的定义 (附 1.2) 中, 我们将所有的对角元置为 0. 这是因为若按非对角元公式计算对角元, 我们会遇到未定式 $1/0$, 从而导致算法数值不稳定. 从问题 (附 1.1) 看, 我们需要额外要求耦合函数 γ_i ($i = 2, \dots, N$) 满足

$$\text{若 } \mathbf{r} = \mathbf{r}', \gamma_i(\mathbf{r}, \mathbf{r}') = 0; \quad \text{若 } i \neq j, \mathbf{r}' = \mathbf{r}'', \gamma_i(\mathbf{r}, \mathbf{r}')\gamma_j(\mathbf{r}, \mathbf{r}'') = 0.$$

物理上, 前者要求第 1 个电子与另一个电子同时位于同一点的概率为 0, 而后者则要求在固定第 1 个电子的位置时, 任意其他两个不同的电子位于同一点的概率为 0. 这对应于给矩阵 Y_i ($i = 2, \dots, N$) 增加如下约束:

$$\text{Tr}(Y_i) = 0, \quad i = 2, \dots, N; \quad \langle Y_i, Y_j \rangle = 0, \quad i, j = 2, \dots, N : i \neq j.$$

基于上述定义, 我们便可将问题 (附 1.1) 离散为如下 MPGCC:

$$\begin{aligned} \min_{\{Y_i\}_{i=2}^N} \quad & \sum_{i=2}^N \langle Y_i, \text{Diag}(\boldsymbol{\rho})C \rangle + \sum_{2 \leq i < j}^N \langle Y_i, \text{Diag}(\boldsymbol{\rho})Y_j C \rangle, \\ \text{s. t.} \quad & Y_i \mathbf{1}_K = \mathbf{1}_K, \quad Y_i^\top \boldsymbol{\rho} = \boldsymbol{\rho}, \quad \text{Tr}(Y_i) = 0, \quad Y_i \geq 0, \quad i = 2, \dots, N, \\ & \langle Y_i, Y_j \rangle = 0, \quad \forall i, j \in \{2, \dots, N\} : i \neq j. \end{aligned}$$

致 谢

在这漫长的五年里,我走过了一个充满挑战与收获的学术之旅。回首往昔,仿佛一幅绚丽多彩的画卷展现在我眼前,每一笔每一刻都深深地刻在了心底。如同黎明初露,我踏入博士研究的征程,怀揣着对知识的渴求和对未知世界的探索。五载光阴,我品味着学识的甘甜,感受着智慧的涌动。

感谢我的导师、恩师刘歆研究员。我与刘老师结识于2018年孟夏。之后,我有幸加入刘老师的课题组,在他的悉心指导下,踏上了学术之路。多年来,刘老师不仅以恩师之情,更以朋友之谊,倾囊相授,分享着他丰富的学术和生活经验,让我收获了熠熠生辉的研究生生涯。在刘老师的引领下,我得以沉淀专业知识,提升独立思考能力。刘老师十分注重培养我的科研品质,从细节到全局,让我在尽情钻研的同时,不至于迷失研究目标。在生活中,刘老师利用点滴时间锻炼身体的习惯深深地感染了我。追随他的脚步,我收获了强健的体魄。最重要的是,刘老师教会了我勇敢面对任何困难,积极迎接挑战,使我在今后的科研和人生道路上始终笃定前行。

感谢北京师范大学的陈华杰老师和北京应用物理与计算数学研究所的高兴誉老师在合作期间对我的悉心指导。在陈老师的带领下,我初次涉足计算材料科学领域,感受到了知识的广阔与深邃。陈老师以其严谨认真的学术态度深深地打动了我。而高老师则总能从应用的角度出发,给我带来新的问题和挑战,同时分享着自己在汇报和处世方面的宝贵经验,让我受益匪浅。

感谢袁亚湘老师。袁老师在讨论班上的建议高屋建瓴,每次都让我收获良多。袁老师渊博的学识与高尚的品格永远是我学习的榜样。感谢戴彧虹老师。戴老师在讨论班上总是能提出富有启发性的观点,使我的科研工作更加完善。感谢刘亚锋老师、马俊杰老师以及北京邮电大学的孙聪老师在我的研究过程中给予了宝贵的建议和指导。感谢高斌老师如兄长般长期以来对我的照顾和帮助。

感谢刘颖、钱莹、丁如娟、陈瑾、李雨霏、刘霞、尹永华、陆凤彬、魏敏、武晓梦、任疆等办公室老师们的帮助。你们让我的学习生活顺利而又丰富多彩。

感谢肖纳川、吴宇宸、杨沐明、陈雅丹、赵浩天、张瑞进、吉振远、刘为、黄磊、姜博鸥、陈圣杰、张吾帅君、王磊、汪思维等课题组的师兄师姐们。你们对我的关心与照顾让我感到无比温暖。其中,尤其感谢肖纳川师兄和王磊师兄。与你们的讨论让我快速成长。感谢一同入学的谢鹏程、陈硕、裴骞、王圣超、赵成、张跃和武哲宇,能够与你们共同学习和进步,是我莫大的幸运。感谢张亦、章煜海、胡雨婷、吴鹏举、刘上琳、王子岳、彭任锋、李博文、汤宇杨、李冠达、姜林硕、郑浩然、范熙来、杨俨、胡威、刁若渝、张宇航、岳艺双、张思远、罗舟行、徐勐、李新鹏、李雨芯、刘亚琛、王兆维、王宇扬、黄辰飞、张博洋等课题组师弟师妹们的陪伴与帮助。感谢徐加樑、刘泽显、戴金雨、王姝、李成梁、张旭、于腾腾、王嘉妮、张凯丽、章丽、魏奇远、张帆、沈欣、Bastien Vieublé、彭真等课

题组博士后们的关爱与照顾. 特别地, 感谢张思远、苏园茗、王宇扬、姜林硕和郑浩然审阅了我的论文初稿并提出了许多宝贵的建议.

感谢中国人民大学李梦雨、西北工业大学尹俊磊、北京科技大学杨真、北京师范大学徐歌等同学. 与你们的合作与交流不仅拓展了我的学术视野, 也让我深刻体会到了其他学科领域的魅力.

感谢与我同期入学的郭仲琨、高宇、王一甲、廖钰蕾、李子逸、胡苗、林仲烁、李翔、冯晓东、王紫菁、时间、党同贺、李瑞旸、刘士勤、高子文、任涛、郭健、刘其芳、纪政平、张瑞等同学. 愿我们都能拥有光明的前程. 其中, 特别感谢我的室友郭仲琨和高宇. 有你们的陪伴, 我的生活变得轻松愉快, 少了很多烦恼.

感谢我的好朋友张越. 我们尽管相隔甚远, 但每次相见都能毫无保留地敞开心扉, 彼此倾诉心声, 真诚地为对方着想, 共同畅想未来的工作和生活. 希望你今后一切顺利. 愿我们的友谊永不褪色.

我要深深感谢我的家人, 特别是我的母亲. 多年来, 你们始终如一地给予我无微不至的关爱、支持和照顾. 你们的爱永远陪伴、激励着我, 让我勇敢面对生活中的一切困难和挫折. 愿我的努力能够不辜负你们的期望.

最后, 感谢我的女朋友缪钰鑫. 我们在一起已经走过了将近六年. 是你, 在我无助之时帮我排解忧愁, 使我变得心胸豁达, 开始享受生活的乐趣. 未来的路上, 我愿与你共享每一个珍贵的瞬间, 感受人世间的繁华与美好.

行文至此, 落笔为终. “长风破浪会有时, 直挂云帆济沧海”.

胡雨宽

2024 年仲夏于北京

作者简历及攻读学位期间发表的学术论文与其他相关学术成果

作者简历:

胡雨宽,江西南昌人,中国科学院数学与系统科学研究院博士研究生.

2015年9月至2019年6月,就读于同济大学数学科学学院数学与应用数学专业,获理学学士学位.

2019年9月至2024年6月,在中国科学院数学与系统科学研究院攻读博士学位,导师为刘歆研究员.

已发表(或正式接收)的学术论文:

- [1] **Hu Y**, Gao X, Zhao Y, et al. Force-based gradient descent method for *ab initio* atomic structure relaxation [J/OL]. Physical Review B, 2022, 106(10): 104101. DOI: [10.1103/PhysRevB.106.104101](https://doi.org/10.1103/PhysRevB.106.104101).
- [2] **Hu Y**, Chen H, Liu X. A global optimization approach for multimarginal optimal transport problems with Coulomb cost [J/OL]. SIAM Journal on Scientific Computing, 2023, 45(3): A1214-A1238. DOI: [10.1137/21M1455164](https://doi.org/10.1137/21M1455164).
- [3] **Hu Y**, Liu X. The convergence properties of infeasible inexact proximal alternating linearized minimization [J/OL]. Science China Mathematics, 2023, 66(10): 2385-2410. DOI: [10.1007/s11425-022-2074-7](https://doi.org/10.1007/s11425-022-2074-7).
- [4] **Hu Y**, Liu X. The exactness of the ℓ_1 penalty function for a class of mathematical programs with generalized complementarity constraints [J/OL]. Fundamental Research, 2023, published online. DOI: [10.1016/j.fmre.2023.04.006](https://doi.org/10.1016/j.fmre.2023.04.006).
- [5] **Hu Y**, Li M, Liu X, et al. Sampling-based methods for multi-block optimization problems over transport polytopes [J]. Mathematics of Computation, accepted. <https://arxiv.org/abs/2306.16763>.

已投稿的学术论文:

- [1] **Hu Y**, Yin J, Gao X, et al. Projected gradient descent algorithm for *ab initio* crystal structure relaxation under a fixed unit cell volume. arXiv preprint, 2024, arXiv:2405.02934. <https://arxiv.org/abs/2405.02934>.

已获得的专利和软件著作权:

- [1] 高兴誉, 刘歆, 宋海峰, **胡雨宽**, 方俊, 杨真, 赵亚帆, 王丽芳, 刘海风. 原子结构弛豫的非单调线搜索方法及装置: 202111534901.5 [P]. 2022-12-02.
- [2] 高兴誉, 刘歆, 宋海峰, **胡雨宽**, 陈欣, 王越超, 方俊, 王丽芳, 张乐. 固定晶格

体积晶体结构弛豫的计算方法及装置: 202211210741.3 [P]. 2023-07-21.

- [3] 高兴誉, 刘歆, 胡雨宽, 宋海峰, 陈欣, 王越超, 王丽芳. 基于非单调梯度型算法的晶体结构弛豫软件 ProME-SuRe: 2023SR1558824 [软件]. 2023-12-04.

参加的研究项目:

- (1) 2019 年 12 月至 2021 年 12 月, 晶体结构弛豫的高效局域优化算法.
- (2) 2022 年 12 月至 2023 年 11 月, 原子结构弛豫优化算法研究与程序研制.

获奖情况:

- (1) 2019 年 9 月, 中国科学院数学与系统科学研究院优秀新生奖学金二等奖.
- (2) 2020 年 6 月, 中国科学院大学优秀学生干部.
- (3) 2020 年 9 月, 斯伦贝谢奖学金.
- (4) 2021 年 9 月, 中国科学院数学与系统科学研究院华罗庚奖学金.
- (5) 2021 年 12 月, 中国运筹学会数学规划分会研究生论坛优秀论文.
- (6) 2022 年 9 月, 中国科学院数学与系统科学研究院华罗庚奖学金.
- (7) 2023 年 3 月, 第十届国际工业与应用数学大会 Financial Support Program.
- (8) 2023 年 9 月, 中国科学院数学与系统科学研究院院长奖学金特等奖.
- (9) 2023 年 11 月, 中国科学院朱李月华奖学金.
- (10) 2023 年 11 月, 北京数学会首届青年优秀论文.