中国科学院大学

University of Chinese Academy of Sciences

博士学位论文

低秩张量优化问题的几何方法

作者姓名：彭任锋

指导教师：袁亚湘 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

2026 年 6 月

Geometric methods for low-rank tensor optimization

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

PENG Renfeng

Supervisor: Professor YUAN Ya-xiang

Academy of Mathematics and Systems Science

Chinese Academy of Sciences

June, 2026

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

张量, 即多维数组, 是向量与矩阵的高阶推广. 低秩张量优化是指变量为张量且满足低秩约束的最优化问题, 在高维数据分析、科学计算以及量子信息等领域具有广泛的应用. 然而, 低秩张量优化相比于成熟的低秩矩阵优化, 仍面临诸多的挑战. 一方面, 相比于低秩矩阵, 不同的张量分解会导出不同的张量秩, 这些秩的定义和性质各异; 另一方面, 低秩张量集合比低秩矩阵集合的几何结构更复杂, 很多低秩矩阵优化的方法无法直接推广到张量, 必须设计新的几何工具和优化方法. 本文从低秩张量集合的几何视角出发, 在理论分析、算法设计与应用三个方面系统研究了低秩张量优化问题.

首先我们研究乘积流形优化问题, 旨在探索度量选择对黎曼优化方法性能的影响, 并探索如何通过精心设计度量以加速算法收敛. 研究乘积流形的动机在于, 在张量分解中参数通常由多个因子组成, 从而整体参数空间可自然建模为一个乘积流形. 我们为此提出了一个黎曼预条件优化框架. 在该框架中, 我们引入由算子定义的预条件度量, 该算子目标在于逼近目标函数黎曼 Hessian 算子的对角块, 从而改善问题的条件数. 进一步地, 针对典型相关分析和截断奇异值分解问题, 我们构造了与问题结构相匹配的新的预条件黎曼度量、提出新的黎曼预条件方法, 并在理论上证明了所提出方法在这些问题上的加速效果. 实验结果表明: 精细地设计黎曼度量的确能够显著提升黎曼优化方法的性能.

作为上述框架的直接应用, 我们提出了一类基于张量环分解的张量补全问题的黎曼预条件算法. 我们通过利用精确块对角预条件思想, 在张量环分解中由各核张量的模态二展平矩阵所构成的乘积空间上构造了一个新的黎曼度量, 并提出了预条件度量下的黎曼梯度法和黎曼共轭梯度法. 在算法实现方面, 我们充分利用了张量结构, 采用了一种经济的计算策略, 避免了梯度计算过程中大规模矩阵的显式构造与运算, 从而显著降低了计算成本. 在人工合成数据集、电影评分数据、高光谱图像以及高维函数上的数值实验结果表明, 所提出的算法在性能上优于现有方法.

接下来, 我们对 Tucker 张量代数簇 (即有界秩 Tucker 张量构成的集合) 及其上面的优化问题展开系统研究. 相比于已被充分研究的矩阵代数簇的几何结构, Tucker 张量代数簇的几何性质是更加复杂的. 我们给出了 Tucker 张量代数簇切锥的显式参数化表示, 并利用其几何结构, 针对在 Tucker 张量代数簇上的优化问题, 提出了具有理论收敛性保证的、基于梯度相关搜索方向的线搜索方法, 避免了计算无显式表达的度量投影. 在实际应用中, 低秩张量优化普遍面临秩参数难以可靠选取的问题. 为此, 我们结合所得到的几何结构, 提出了一种 Tucker 秩自适应方法, 该方法能够在迭代过程中自动识别合适的秩参数, 同时仍然有收敛性保证. 在张量补全问题的数值实验结果表明, 所提出的方法在恢复性能方面优于其他代表性的方法. 此外, 秩自适应方法在不同秩参数设定下均表现出最优性能,

并且确实能够识别出合适的 Tucker 秩参数.

此外, 我们研究了单位 Frobenius 范数的低秩张量的几何与应用. 事实上, 这类张量是科学计算和量子物理等领域中的基础研究对象, 它们能够用于表示归一化的特征向量以及纯量子态. 尽管张量链分解为处理高维问题提供了一种强有力的低秩分解形式, 但该分解形式本身并不能内在地保证单位范数约束. 为此, 我们引入了归一化张量链分解 (NTT), 其目标是在张量链格式下, 用单位范数张量来逼近给定的张量. 本文证明了固定秩 NTT 张量的集合构成一个光滑流形, 并推导了其对应的几何结构, 从而为几何优化方法的提出奠定了理论基础. 在此基础上, 我们将其应用于低秩张量恢复问题、高维特征值问题、稳定子秩的估计以及量子信道最小输出 Rényi-2 熵的计算. 数值实验结果表明, 所提出的基于 NTT 的方法在计算效率和可扩展性方面均表现出显著优势.

关键词：张量分解, 低秩张量优化, 流形优化, 切锥, 预条件方法, 秩自适应策略

Abstract

Tensors are referred to as multidimensional arrays, which can be viewed as a higher-order generalization of vectors and matrices. Low-rank tensor optimization refers to optimization problems where the tensor variable satisfies low-rank constraints, and it plays an important role in data analysis, scientific computing, and quantum information theory. However, in contrast with the well-studied low-rank matrix optimization, low-rank tensor optimization remains significantly more challenging. On the one hand, different tensor decompositions induce different notions of tensor rank, whose definitions and properties vary substantially. On the other hand, the geometric structure of low-rank tensor sets is much more intricate than low-rank matrix counterparts. Therefore, the low-rank tensor optimization methods cannot be simply generalized from low-rank matrix optimization, necessitating new geometric tools and optimization methods. To this end, this dissertation investigates low-rank tensor optimization problems from a geometric perspective, focusing on theoretical analysis, optimization methods, and applications.

We first study optimization on product manifolds, and explore how the performance of a Riemannian optimization method varies with different metrics and how to exquisitely construct a metric to accelerate a method. In fact, the parameters in tensor decompositions naturally enjoy product structures. To this end, we propose a general framework for optimization on product manifolds endowed with a preconditioned metric. The metric is constructed by an operator that aims to approximate the diagonal blocks of the Riemannian Hessian of the cost function, and to eliminate ill-conditioning. Specifically, we tailor new preconditioned metrics and adapt Riemannian methods to the canonical correlation analysis and the truncated singular value decomposition problems, which provably accelerate the Riemannian methods. Numerical results among these applications verify that a delicate metric does accelerate the Riemannian methods.

As a direct application of the proposed framework, we propose Riemannian preconditioned algorithms for the tensor completion problem via tensor ring decomposition. By using the exact-block preconditioning, we develop a new Riemannian metric on the product space of the mode-2 unfolding matrices of the core tensors in tensor ring decomposition. We propose the Riemannian gradient descent and Riemannian conjugate gradient algorithms under the proposed metric. In practice, we exploit the tensor structure and adopt an economical procedure to avoid large matrix formulation and computation in gradients, which significantly reduces the computational cost. Numerical experiments on various synthetic and real-world datasets—movie ratings, hyperspectral images, and high-dimensional functions—suggest that the proposed algorithms have

better performance than other candidates.

Next, we embark on an exploration of Tucker tensor varieties—the set of tensors with bounded Tucker rank—and associated optimization problems. The geometry of Tucker tensor varieties is notably more intricate than the well-explored geometry of matrix varieties. We give an explicit parametrization of the tangent cone of Tucker tensor varieties and leverage its geometry to develop provable gradient-related line-search methods for optimization on Tucker tensor varieties. The search directions are computed from approximate projections which circumvents the calculation of intractable metric projections. In practice, low-rank tensor optimization suffers from the difficulty of choosing a reliable rank parameter. To this end, we incorporate the established geometry and propose a Tucker rank-adaptive method that is capable of identifying an appropriate rank during iterations while the convergence is also guaranteed. Numerical experiments on tensor completion reveal that the proposed methods are in favor of recovering performance over other state-of-the-art methods. Moreover, the rank-adaptive method performs the best across various rank parameter selections and is indeed able to find an appropriate rank.

We also study the geometry and applications of low-rank tensors with unit Frobenius norm. In fact, tensors with unit Frobenius norm are fundamental objects in many fields, including scientific computing and quantum physics, which are able to represent normalized eigenvectors and pure quantum states. While the tensor train decomposition provides a powerful low-rank format for tackling high-dimensional problems, it does not intrinsically enforce the unit-norm constraint. To address this, we introduce the normalized tensor train (NTT) decomposition, which aims to approximate a tensor by unit-norm tensors in tensor train format. We prove that the set of fixed-rank NTT tensors forms a smooth manifold, and the corresponding Riemannian geometry is derived, paving the way for geometric methods. We propose NTT-based methods for low-rank tensor recovery, high-dimensional eigenvalue problems, estimation of stabilizer rank, and calculation of the minimum output Rényi-2 entropy of quantum channels. Numerical experiments demonstrate the superior efficiency and scalability of the proposed NTT-based methods.

Key Words: Tensor decomposition, low-rank tensor optimization, manifold optimization, tangent cones, preconditioned methods, rank-adaptive strategies

目 录

图目录

表目录

符号列表

字符

$\mathbb{N}$	自然数集
$\mathbb{R}$	实数集
$\mathbb{C}$	复数集
$\mathbb{R}^{n_1\times n_2\times\cdots\times n_d}$	欧氏空间
$a, b, c, \ldots$	标量
$\mathbf{a}, \mathbf{b}, \mathbf{c}, \ldots$	向量
$\mathbf{A}, \mathbf{B}, \mathbf{C}, \ldots$	矩阵
$\mathbf{I}$	单位矩阵
$\mathcal{A}, \mathcal{B}, \mathcal{C}, \ldots$	张量
$\mathrm{St}(p, n)$	Stiefel 流形, 大小为 $n \times p$ 的列正交矩阵集合
$\mathrm{Gr}(p, n)$	Grassmann 流形, n 维空间中的 p 维子空间集合

算子

∇	梯度
∇^2	Hessian 算子
$\lambda_{\max}(\mathbf{A})$	矩阵 $\mathbf{A}$ 的最大特征值
$\lambda_{\min}(\mathbf{A})$	矩阵 $\mathbf{A}$ 的最小特征值
$\sigma_{\max}(\mathbf{A})$	矩阵 $\mathbf{A}$ 的最大奇异值
$\sigma_{\min}(\mathbf{A})$	矩阵 $\mathbf{A}$ 的最小奇异值
$\mathbf{A}^\dagger$	矩阵 $\mathbf{A}$ 的 Moore–Penrose 伪逆

缩写

SVD	奇异值分解
ALS	交替最小二乘方法
TT	张量链分解
TR	张量环分解

第 1 章　绪论

1.1　研究背景

优化 (Optimization) 是应用数学中的一个核心研究方向, 主要关注在给定约束条件或无约束情形下, 如何选择合适的决策变量, 使目标函数在某种意义下达到最优, 例如实现收益最大化或风险最小化. 作为一门内容体系庞大的学科, 优化涵盖了丰富的理论与方法, 并且实际生活中的很多问题, 均可建模为优化问题. 通常在刻画实际问题时, 首先需要结合具体应用背景, 对问题的目标和约束条件进行合理的分析以建立恰当的数学模型. 在模型构建完成后, 可以运用优化理论设计高效且可靠的算法以求解相应的最优解. 与此同时, 为了确保算法在实际应用中的可靠性, 还需对其收敛性、计算复杂度等理论性质进行分析, 并通过数值实验测试其实际性能. 随着计算能力的提升与应用需求的不断扩展, 优化理论与算法在人工智能、数据挖掘、管理科学、图像与信号处理、金融经济以及生物医学等诸多领域扮演着越来越重要的角色, 成为了求解实际问题的重要数学工具. 因此, 优化在当代应用数学及相关交叉学科中的地位愈发重要. 具体而言, 优化问题可统一表示为如下的形式

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & x \in C \subseteq \mathbb{R}^n, \end{aligned} \tag{1-1}$$

这里 x 为决策变量, $f : \mathbb{R}^n \to \mathbb{R}$ 为目标函数, $C \subseteq \mathbb{R}^n$ 为约束集或可行域. 称可行域包含的点 $x \in C$ 为可行解或可行点. “s.t.” 为英文 “subject to” 的缩写, 即 x 需要满足的条件. 如果我们希望在可行域 C 上求解目标函数 $f(x)$ 的最大值, 则仅需将问题 (1-1) 中的记号 “min” 相应地替换为 “max”. 事实上, 任意最大化问题都可以通过在目标函数前引入负号转化为等价的最小化问题. 因此, 本文仅需讨论最小化问题.

在问题 (1-1) 中, 变量 x 通常被视为 $\mathbb{R}^n$ 空间中的向量. 然而, 在具体应用中, 优化变量的形式并不局限于向量, 还可以是矩阵、张量, 甚至是抽象流形上的元素. 当可行域满足 $C = \mathbb{R}^n$ 时, 问题 (1-1) 被称为无约束优化问题; 否则称其为约束优化问题. 此外, 若目标函数 f 为凸函数且可行域 C 为凸集, 则称问题 (1-1) 为凸优化问题; 若不满足上述条件, 则称为非凸优化问题.

定义 1.1 (全局最优点)**.** 称 $x^* \in C$ 为问题 (1-1)的全局最优点, 若 $f(x^*) \leq f(x)$ 对任意的 $x \in C$ 均成立.

定义 1.2 (局部最优点)**.** 称 $x^* \in C$ 为问题 (1-1)的局部最优点, 若存在某个关于 x^* 的邻域 U, 使得 $f(x^*) \leq f(x)$ 对任意的 $x \in C \cap U$ 均成立. 进一步, 若 $f(x^*) < f(x)$ 对任意的 $x \in C \cap U$ 均成立, 则称 $x^* \in C$ 为问题 (1-1)的严格局部最优点.

值得注意的是, 全局最优点一定是局部最优点, 反之则不一定成立. 当问题 (1-1)是凸问题时, 局部最优点一定也是全局最优点.

根据目标函数的性质、可行域的结构以及应用背景的不同, 优化问题 (1-1) 可以进一步被划分为多种类别, 例如线性规划、二次规划、半定规划、多项式优化、全局优化、整数规划、组合优化、锥优化、流形优化、非光滑优化、稀疏优化、无导数优化、鲁棒优化、随机优化、双层优化、分布式优化、多目标优化、以及低秩矩阵张量优化等. 感兴趣的读者可以参考相关领域的经典专著及文献 [1–5].

在本文中, 我们关心的优化变量为张量. 张量是多维数组, 是向量与矩阵的高阶推广, 见图 1-1. 相比于矩阵, 张量有更多的维度, 可以记录数据的不同属性与方向. Kolda [6] 指出, "传统大数据研究多集中于矩阵, 但现实数据往往具有超过两个维度的多模态变化, 因此更适合用张量 (多维数组) 来表示". 张量数据的一个典型应用是人脸数据分析 [7, 8], 目标是从多个人脸数据中提取人脸的主要特征. 相比于将二维图像展开成向量再进行矩阵形式的主成分分析, 将人脸数据直接建模为张量并进行高阶主成分分析, 可以更自然地保留不同模态 (如身份、表情、姿态、光照) 的结构信息, 从而在特征提取上达到更好的效果; 见图 1-2.

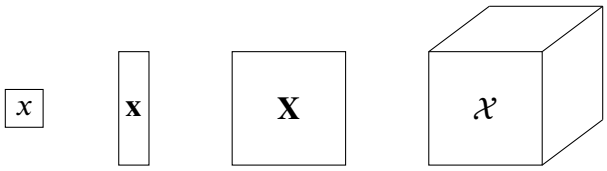


图 1-1 标量、向量、矩阵以及张量的示意图.

Figure 1-1 Scalar, vector, matrix, and tensor.

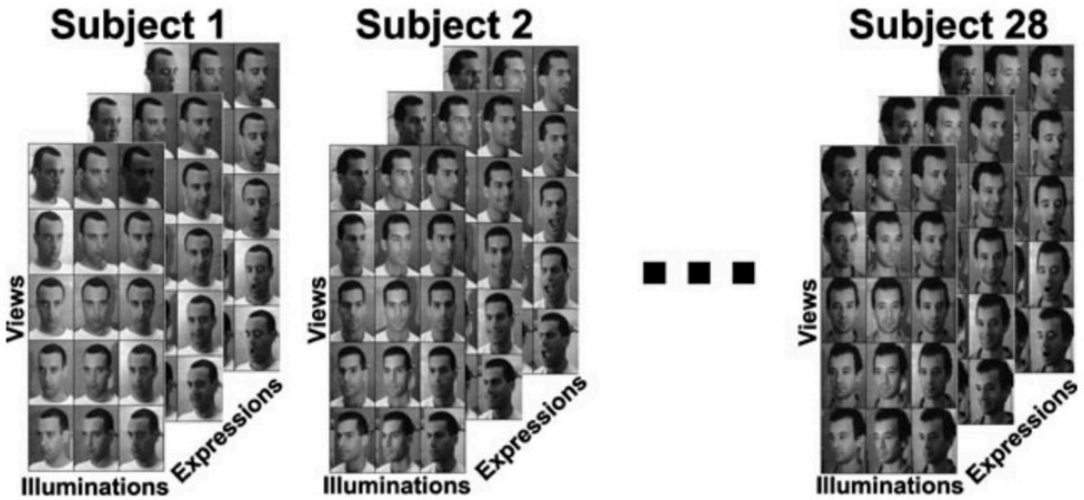


图 1-2 张量脸数据. 图片来源于 [7].

Figure 1-2 Tensor faces in [7].

然而, 在实际的计算中, 完整存储一个高阶的张量在计算上是不可行的, 因为其存储量会随着张量阶数呈指数增长, 会面临维数灾难 (curse of dimensionality) 的问题. Kressner [9] 指出, "完整存储一个 *50* 阶的张量, 即使是大小为 $2\times2\times\cdots\times2$ 的张量, 也需要 *9Pb* 的空间. 因此, 我们需要利用低秩张量分解的手段降低储存

量”. 而低秩张量分解, 不仅仅能降低储存量, 更能提取出高阶张量中最关键的信息. 在数学层面上，这种低秩性也具有理论基础, 若一个矩阵的行和列由隐变量模型 (latent variable model) 生成, 则它天生具有低秩性 [10]. 这说明低秩并非人为施加的假设, 而是源于数据生成机制的内在规律, 这也为低秩矩阵张量的应用提供了理论支撑.

近年来, 随着大数据时代的到来, 低秩张量在其他诸多应用中展现了其显著的有效性, 包括图像与信号处理 [8, 11–13]、推荐系统 [14–17]、张量方程 [18]、数理金融 [19]、量子计算 [20]、以及高维偏微分方程的数值求解 [21, 22]. 关于低秩矩阵张量方法的系统性综述与更多相关工作, 可参见文献 [9, 23].

本文聚焦于如下的低秩张量优化问题,

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & x \in \{\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d} : \operatorname{Rank}(\mathcal{X}) \le \mathbf{r}\}. \end{aligned} \tag{1-2}$$

这类问题的决策变量为低秩张量, 并带有低秩约束 $\operatorname{Rank}(\mathcal{X}) \le \mathbf{r}$, 这里当 $d \ge 3$ 时, Rank 代表某种张量的秩, 详见 1.2节中张量分解以及张量秩的介绍. 值得注意的是, 当 $d = 2$ 时, Rank 代表矩阵的秩, 此时问题 (1-2) 退化为低秩矩阵优化问题. 我们将低秩矩阵与低秩张量优化问题统称为低秩优化问题.

低秩张量优化在实际应用当中扮演着重要的角色. 例如, 在推荐系统中, 用户对物品、电影的评分数据通常可以表示为一个高维稀疏矩阵或张量, 而这些数据往往由少数潜在因素 (如用户兴趣、年龄特征、物品属性、电影类别等) 所驱动，而这些因素都可以被归类为几个典型的类别 (genre), 因此数据能呈现出明显的低秩结构. 利用这一结构, 可以通过低秩矩阵或张量分解建模、求解低秩优化问题有效地进行缺失数据预测. 一个著名的例子是 Netflix Prize[1]，其目标是通过分析用户评分数据来显著提升电影推荐系统的准确率，而最终获胜的方案正是基于低秩矩阵分解方法 [24]. 近几年, 低秩张量优化方法在深度学习, 特别是大语言模型中的应用也逐渐受到关注. Shampoo [25] 等优化方法利用张量结构构造预条件子以加速训练; 低秩微调 (Low-Rank Adaptation, LoRA) [26] 通过利用低秩矩阵分解对大模型权重进行调整, 实现了在有限资源下的高效微调; 文献 [27] 利用 Tucker 分解对注意力机制中的参数进行压缩, 以降低模型的存储和计算开销; 张量链分解被用于构造低秩微调方法 [28], 从而在保持模型性能的同时大幅减少训练的参数量; 文献 [29] 利用张量环分解构造矩阵乘积算子, 从而实现从小模型参数到大模型参数的结构化映射; 张量环分解也可以用于构造神经网络 [30] 并应用在图像分类等任务中. 这些例子显示, 低秩张量优化通过捕捉数据的内在低秩结构, 不仅能够降低计算复杂度, 还能提升算法的稳健性和泛化能力.

我们下面简要介绍论文所需的一些基本概念. 首先, 我们将在1.2节中介绍张量以及典型的张量分解形式. 然后, 1.3节介绍了低秩矩阵与张量构成的集合. 接下来, 基于低秩矩阵张量集合, 我们在1.4节中介绍低秩矩阵与张量优化现有的理论与方法. 最后, 1.5节中展示了本文主要工作的概述.

[1]参见: https://web.archive.org/web/20061106031103/http://www.netflixprize.com/.

1.2 张量分解及其应用简介

什么是张量? 张量一词 (拉丁语: tendere, tensus) 在 1846 年由哈密顿 [31] 第一次在数学中使用. 1898 年, Voigt [32] 第一次提出张量是标量, 向量和矩阵的高阶推广. 在不同的领域中对于张量的定义是有不同的. 在物理中, 张量经常用来指代张量场 (如黎曼曲率张量, 度量张量等) [33]. 通常, 在代数教材中 [34, 35], 一个 d 阶张量 $\mathcal{A}$ 定义为 d 个实向量空间的张量积中的元素, 即 $V_1 \otimes V_2 \otimes \cdots \otimes V_d$, 在选定了 $V_1, V_2, \cdots, V_d$ 中的基以后, $\mathcal{A}$ 可以被坐标化, 即可以用一个 d 维数组表示 $\mathcal{A}$, 不加区分地, 我们仍记为 $\mathcal{A}$. 我们用 $\mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$ 代表全体 d 维实数组 $\mathcal{A} = (a_{i_1,i_2,\ldots,i_d})_{i_1=1,i_2=1,\ldots,i_d=1}^{n_1,n_2,\ldots,n_d}$ 和 $\mathcal{B}$ 在加法和数乘

$$(a_{i_1,i_2,\ldots,i_d}) + (b_{i_1,i_2,\ldots,i_d}) = (a_{i_1,i_2,\ldots,i_d} + b_{i_1,i_2,\ldots,i_d}),\ \lambda(a_{i_1,i_2,\ldots,i_d}) = (\lambda a_{i_1,i_2,\ldots,i_d})$$

下构成的向量空间. 该定义可以被拓展到复数域 $\mathbb{C}$. 在本文中, 我们采用张量是矩阵向量的高阶推广的定义. d 阶张量也被称为 d 维超矩阵 [36], 或 d 路阵列 [37]. 由于当 $V_1, V_2, \ldots, V_d$ 为有限维空间时, 它们与 $\mathbb{R}^{n_1}, \mathbb{R}^{n_2}, \ldots, \mathbb{R}^{n_d}$ 同构, 故采用这种定义是合理的. 这一节中我们将介绍张量的基本运算以及典型的张量分解.

1.2.1 张量计算的基本符号

张量计算中将涉及如下概念 [37, 38]. 记索引集合 $\{1, 2, \ldots, n\}$ 为 $[n]$. 首先, 定义如下索引映射 π_k:

$$\pi_k : (i_1, \ldots, i_{k-1}, i_{k+1}, \ldots, i_d) \mapsto 1 + \sum_{\ell \neq k, \ell=1}^{d} (i_\ell - 1) J_\ell,$$

其中 $J_\ell = \prod_{m=1, m\neq k}^{\ell-1} n_m$, $k \in [d]$. 张量 $\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$ 的模态 k 展平矩阵 (mode-k unfolding) 记为 $\mathbf{X}_{(k)} \in \mathbb{R}^{n_k \times n_{-k}}$, 其中 $n_{-k} := \prod_{i \neq k} n_i$. $\mathcal{X}$ 的第 $(i_1, i_2, \ldots, i_d)$ 个元素对应于矩阵 $\mathbf{X}_{(k)}$ 的第 (i_k, j) 个元素, 其中 $j = \pi_k(i_1, \ldots, i_{k-1}, i_{k+1}, \ldots, i_d)$. 类似地, 索引集合 Ω 的模态 k 展平矩阵定义为 $\Omega_{(k)} := \{(i_k, \pi_k(i_1, \ldots, i_{k-1}, i_{k+1}, \ldots, i_d)) : (i_1, i_2, \ldots, i_d) \in \Omega\}$. 两个张量 $\mathcal{X}, \mathcal{Y} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$ 之间的内积定义为 $\langle \mathcal{X}, \mathcal{Y} \rangle := \sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2} \cdots \sum_{i_d=1}^{n_d} \mathcal{X}(i_1, i_2 \ldots, i_d) \mathcal{Y}(i_1, i_2 \ldots, i_d)$. 张量 $\mathcal{X}$ 的 Frobenius 范数定义为 $\|\mathcal{X}\|_{\mathrm{F}} := \sqrt{\langle \mathcal{X}, \mathcal{X} \rangle}$.

张量化算子将矩阵 $\mathbf{X}_k \in \mathbb{R}^{n_k \times n_{-k}}$ 映射为张量 $\mathrm{ten}_{(k)}(\mathbf{X}_k) \in \mathbb{R}^{n_1 \times \cdots \times n_d}$, 其定义为 $\mathrm{ten}_{(k)}(\mathbf{X}_k)(i_1, \ldots, i_d) = \mathbf{X}_k(i_k, 1 + \sum_{\ell \neq k, \ell=1}^{d} (i_\ell - 1) J_\ell)$, 其中 $(i_1, \ldots, i_d) \in [n_1] \times \cdots \times [n_d]$. 注意到, 在 $n_1, \ldots, n_d$ 固定的情况下, 有 $\mathrm{ten}_{(k)}(\mathbf{X}_{(k)}) = \mathcal{X}$ 成立. 因此, 张量化算子是可逆的. 张量 $\mathcal{X}$ 与矩阵 $\mathbf{A} \in \mathbb{R}^{M \times n_k}$ 的 k 模态积记为 $\mathcal{X} \times_k \mathbf{A} \in \mathbb{R}^{n_1 \times \cdots \times M \times \cdots \times n_d}$, 其中 $\mathcal{X} \times_k \mathbf{A}$ 的第 $(i_1, \ldots, i_{k-1}, j, i_{k+1}, \ldots, i_d)$ 个元素定义为 $\sum_{i_k=1}^{n_k} x_{i_1 \ldots i_d} a_{j i_k}$. 并且有 $(\mathcal{X} \times_k \mathbf{A})_{(k)} = \mathbf{A}\mathbf{X}_{(k)}$. 给定 $\mathbf{u}_1 \in \mathbb{R}^{n_1} \backslash \{0\}, \ldots, \mathbf{u}_d \in \mathbb{R}^{n_d} \backslash \{0\}$, 大小为 $n_1 \times \cdots \times n_d$ 的秩 1 张量可由外积定义为 $\mathcal{V} := \mathbf{u}_1 \circ \cdots \circ \mathbf{u}_d$, 等价地, 其元素表示为 $v_{i_1,\ldots,i_d} := u_{1,i_1} \cdots u_{d,i_d}$. 两个矩阵 $\mathbf{A} \in \mathbb{R}^{m_1 \times n_1}$ 和 $\mathbf{B} \in \mathbb{R}^{m_2 \times n_2}$ 的 Kronecker 积定义为一个 $(m_1 m_2)$ 行 $(n_1 n_2)$ 列的矩阵, 记为 $\mathbf{A} \otimes \mathbf{B} := (a_{ij}\mathbf{B})_{ij}$. 向量 $\mathbf{e}_i \in \mathbb{R}^n$ 定义为 $n \times n$ 单位

矩阵 $\mathbf{I}_n$ 的第 i 列. 给定两个向量 $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, 若对所有 $i \in [d]$ 有 $x_i \leq y_i$(或 $x_i < y_i$), 则分别记为 $\mathbf{x} \leq \mathbf{y}$(或 $\mathbf{x} < \mathbf{y}$). 一个张量 $\mathcal{U}_k \in \mathbb{C}^{r_{k-1} \times n_k \times r_k}$ 可以重排为其左展平和右展平矩阵, 分别定义为 $\mathrm{L}(\mathcal{U}_k) \in \mathbb{C}^{(r_{k-1}n_k) \times r_k}$ 和 $\mathrm{R}(\mathcal{U}_k) \in \mathbb{C}^{r_{k-1} \times (n_k r_k)}$. 给定张量 $\mathcal{X} \in \mathbb{C}^{n_1 \times n_2 \times n_3}$ 以及矩阵 $\mathbf{A} \in \mathbb{C}^{n_1 \times n_1}$、$\mathbf{B} \in \mathbb{C}^{n_2 \times n_2}$、$\mathbf{C} \in \mathbb{C}^{n_3 \times n_3}$, 我们给出关于左展开与右展开的如下等式关系.

$$\begin{aligned} \mathrm{L}(\mathcal{X} \times_1 \mathbf{A} \times_2 \mathbf{B} \times_3 \mathbf{C}) &= (\mathbf{B} \otimes \mathbf{A})\mathrm{L}(\mathcal{X})\mathbf{C}^\top, \\ \mathrm{R}(\mathcal{X} \times_1 \mathbf{A} \times_2 \mathbf{B} \times_3 \mathbf{C}) &= \mathbf{A}\mathrm{R}(\mathcal{X})(\mathbf{C} \otimes \mathbf{B})^\top. \end{aligned} \tag{1-3}$$

集合 $\mathrm{St}(r, n) := \{\mathbf{X} \in \mathbb{R}^{n \times r} : \mathbf{X}^\top \mathbf{X} = \mathbf{I}_r\}$ 称为 Stiefel 流形, 其中 $\mathbf{I}_r$ 表示 $r \times r$ 的单位矩阵. 正交群定义为 $\mathcal{O}(n) := \{\mathbf{Q} \in \mathbb{R}^{n \times n} : \mathbf{Q}^\top \mathbf{Q} = \mathbf{Q}\mathbf{Q}^\top = \mathbf{I}_n\}$. 集合 $\mathrm{St}_{\mathbb{C}}(p, n) = \{\mathbf{X} \in \mathbb{C}^{n \times p} : \mathbf{X}^\dagger \mathbf{X} = \mathbf{I}_p\}$ 被称为复 Stiefel 流形, 其中 $\mathbf{X}^\dagger$ 表示矩阵 $\mathbf{X}$ 的共轭转置. 矩阵 $\mathbf{X}$ 的复共轭记为 $\mathrm{conj}(\mathbf{X}) = (\mathrm{conj}(x_{i,j}))_{i,j}$.

1.2.2　典型的张量分解形式

与矩阵秩不同, 不同的张量分解形式会诱导出多种张量上秩的定义. 典型的张量分解形式包括标准多元分解 (CANDECOMP/PARAFAC 分解, 简称 CP 分解) [39]、Tucker 分解 [40]、张量链分解 (tensor train decomposition) [41] 以及张量环分解 (tensor ring decomposition) [42]; 相关综述可参见文献 [38].

张量的标准多元分解　首先我们介绍张量的标准多元分解. 张量多元分解形式 (polyadic form) 的概念起源于 Hitchcock [39, 43], Cattell [44, 45] 在 1944 年提出了多路模型的概念. 在经历了多次推广以后, 这个分解形式开始流行起来. 1970 年, 由 Carroll 和 Chang [46] 在心理测量学中以标准分解形式 (canonical decomposition) 和 Harshman [47] 以平行因子 (parallel factors) 模型, 第三次引入了标准多元分解形式的概念. 自此之后 CP 分解形式开始吸引众多的科研人员. Möcks [48] 于 1988 年在大脑图像学中提出了地形学成分模型, 其数学本质即为 CP 分解形式.

定义 1.3 (张量的标准多元分解 [43])**.** 标准多元分解形式旨在将一个张量表示为几个秩 1 张量的和, 即给定 $\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$, 其标准多元分解为

$$\mathcal{X} = [\![\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_d]\!]_{\mathrm{CP}} := \sum_{r=1}^{R} \mathbf{a}_{1,r} \circ \mathbf{a}_{2,r} \circ \cdots \circ \mathbf{a}_{d,r}, \tag{1-4}$$

即 R 个秩 1 张量和的形式, 如图1-3所示, 这里 $\mathbf{A}_k = [\mathbf{a}_{k,1}, \mathbf{a}_{k,2}, \ldots, \mathbf{a}_{k,R}]$ 为因子矩阵. 从元素上来看,

$$x_{i_1, i_2, \ldots, i_d} = \sum_{r=1}^{R} \mathbf{a}_{1,r}(i_1)\mathbf{a}_{2,r}(i_2) \cdots \mathbf{a}_{d,r}(i_d).$$

通过标准多元分解, 可以引出张量秩的概念. Hitchcock [43] 在 1927 年首先提出用能生成 $\mathcal{X}$ 的最少的秩 1 张量的个数作为张量的秩, 记为 $\mathrm{rank}(\mathcal{X})$, 即

$$\mathrm{rank}(\mathcal{X}) := \min\{R \in \mathbb{N} : \mathcal{X} = [\![\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_d]\!]_{\mathrm{CP}},\ \mathbf{A}_k \in \mathbb{R}^{n_k \times R}\}.$$

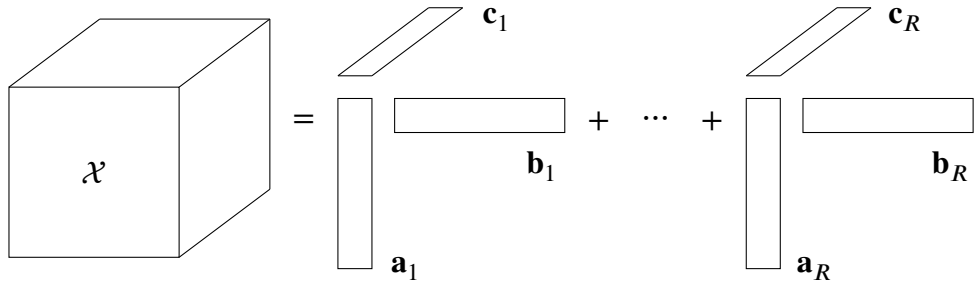


图 1-3 一个三阶张量的标准多元分解.

Figure 1-3 CP decomposition for a third-order tensor.

在 50 年以后 Kruskal [49] 独立地提出了同样的概念. 然而, 张量秩和矩阵秩有着诸多的不同之处. 其中一个不同之处是, 张量秩在数域 $\mathbb{R}$ 和 $\mathbb{C}$ 下可能是不同的. 另一个区别在于, 对于任何一个张量, 张量秩的计算问题是 NP-难的, 这也说明基于定义的直接计算方法在实际中是不可行的. 事实上, Håstad [50] 证明了在有理数域以及有限域中计算三阶张量的秩是一个 NP-难问题. Hillar 和 Lim [51] 证明了在实数域和复数域上完成张量 CP 分解是一个 NP-难问题. Shitov [52] 证明了计算在任意整环上的张量秩可以多项式时间内归约到求解多项式系统的零点的问题. 更多关于张量秩的综述见 [38].

张量的 Tucker 分解 1966 年, Tucker [53] 给出了 Tucker 分解形式, 此后 Levin [54] 和 Tucker [40, 53] 将该分解形式完备化. Kapteyn 等人 [55] 从高维形式主成分分析的角度阐述 Tucker 分解. 数学上, Tucker 分解将一个张量分解形式为一个核张量与在每一个模态与之相乘的矩阵的乘积.

定义 1.4 (Tucker 分解 [40]). 给定一个张量 $\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$, 其 Tucker 分解定义为

$$\mathcal{X} = \mathcal{G} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \cdots \times_d \mathbf{U}_d = \mathcal{G} \times_{k=1}^{d} \mathbf{U}_k,$$

其中 $\mathcal{G} \in \mathbb{R}^{r_1 \times r_2 \times \cdots \times r_d}$ 为核张量, $\mathbf{U}_k \in \mathrm{St}(r_k, n_k)$ 为具有正交列单位向量的因子矩阵, 并且 $r_k = \mathrm{rank}(\mathbf{X}_{(k)})$.

图 1-4 展示了一个三阶张量的 Tucker 分解. 注意, 对于张量 $\mathcal{X} = \mathcal{G} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \cdots \times_d \mathbf{U}_d$ 其模态 k 展平矩阵满足

$$\mathbf{X}_{(k)} = \mathbf{U}_k \mathbf{G}_{(k)} \left(\mathbf{U}_d \otimes \cdots \otimes \mathbf{U}_{k+1} \otimes \mathbf{U}_{k-1} \otimes \cdots \otimes \mathbf{U}_1 \right)^\top = \mathbf{U}_k \mathbf{G}_{(k)} ((\mathbf{U}_j)^{\otimes j \neq k})^\top,$$

其中 $(\mathbf{U}_j)^{\otimes j \neq k} := \mathbf{U}_d \otimes \cdots \otimes \mathbf{U}_{k+1} \otimes \mathbf{U}_{k-1} \otimes \cdots \otimes \mathbf{U}_1$ for $k \in [d]$. 特别地, 对于一个 d 阶张量 $\mathcal{A}$, 有如下等价关系,

$$\mathcal{A} \in \bigotimes_{k=1}^{d} \mathrm{span}(\mathbf{U}_k) \iff \mathcal{A} = \mathcal{C} \times_{k=1}^{d} \mathbf{U}_k, \tag{1-5}$$

其中 $\mathcal{C} \in \mathbb{R}^{r_1 \times r_2 \times \cdots \times r_d}$.

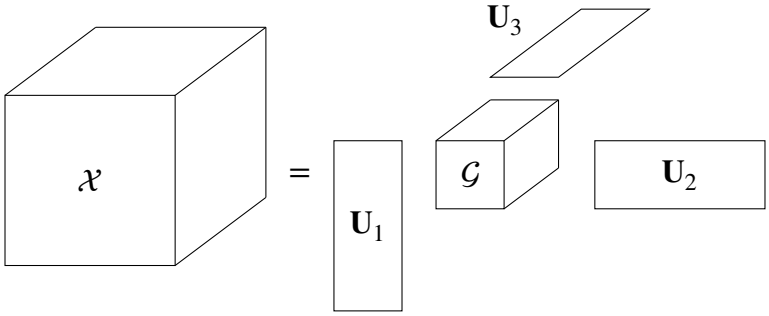


图 1-4 一个三阶张量的 Tucker 分解.

Figure 1-4 Tucker decomposition of a third-order tensor.

Tucker 分解引入了一种不同的张量的秩. Kruskal [56] 首先提出了张量 "n-秩" 的概念, 它也被称为多线性秩 (multilinear rank) [57]. 在 De Lathauwer 等人 [58] 提出多线性 SVD 后, n-秩的概念获得了更多的关注. 张量的 n-秩也被称为 Tucker 秩, 对于一个张量 $\mathcal{X}$, 其 Tucker 秩定义为

$$\operatorname{rank}_{\mathrm{tc}}(\mathcal{X}) := \mathbf{r} = (r_1, r_2, \dots, r_d) = (\operatorname{rank}(\mathbf{X}_{(1)}), \operatorname{rank}(\mathbf{X}_{(2)}), \dots, \operatorname{rank}(\mathbf{X}_{(d)})).$$

值得注意的是, 张量的 Tucker 秩是一个数组.

张量链分解　2011 年, Oseledets [41] 提出了张量链 (tensor train, TT) 分解. 它在计算物理中也被称为矩阵乘积态 (matrix product states, MPS) [20, 59]. TT 分解形式的核心思想是将一个 d 阶张量 $\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$ 分解为 d 个 3 阶张量.

定义 1.5 (张量链分解 [41])**.** 给定一个 d 阶张量 $\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$, 张量链分解将 $\mathcal{X}$ 分解为 d 个核张量 $\mathcal{U}_k \in \mathbb{R}^{r_{k-1} \times n_k \times r_k}$, 其中 $\mathcal{X}$ 中的第 $(i_1, i_2, \dots, i_d)$ 个元素可以表示为

$$\mathcal{X}_{i_1, i_2, \dots, i_d} = \mathbf{U}_1(i_1)\mathbf{U}_2(i_2) \cdots \mathbf{U}_d(i_d), \tag{1-6}$$

这里 $\mathbf{U}_k(i_k) = \mathcal{U}_k(:, i_k, :)$ 为 $\mathcal{U}_k$ 的横向切面, $k \in [d]$, $r_0 = r_d = 1$. 注意对于复数域 $\mathbb{C}$ 张量链分解的定义类似.

事实上, $\mathcal{U}_1$ 和 $\mathcal{U}_d$ 均为矩阵, 但是为了形式上的统一性, 我们仍然使用张量的符号表示它们. 下面我们引入第 k 展平矩阵以及接口矩阵的概念. 对于张量 $\mathcal{X} \in \mathbb{C}^{n_1 \times n_2 \times \cdots \times n_d}$, 其第 k 展平矩阵定义为 $\mathbf{X}_{\langle k \rangle} \in \mathbb{C}^{(n_1 n_2 \cdots n_k) \times (n_{k+1} n_{k+2} \cdots n_d)}$, $k \in [d-1]$ 其中

$$\mathbf{X}_{\langle k \rangle}\left(i_1 + \sum_{j=2}^{k}(i_j - 1)\prod_{\ell=1}^{j-1} n_\ell,\ i_{k+1} + \sum_{j=k+2}^{d}(i_j - 1)\prod_{\ell=k+1}^{j-1} n_\ell\right) = \mathcal{X}(i_1, i_2, \dots, i_d).$$

张量 $\mathcal{X}$ 的接口矩阵 $\mathbf{X}_{\leq k}$ 与 $\mathbf{X}_{\geq k+1}$ 定义为

$$\mathbf{X}_{\leq k}(i_1 + \sum_{j=2}^{k}(i_j - 1)\prod_{\ell=1}^{j-1} n_\ell, :) = \mathbf{U}_1(i_1)\mathbf{U}_2(i_2)\cdots\mathbf{U}_k(i_k),$$

$$\mathbf{X}_{\geq k+1}(i_{k+1} + \sum_{j=k+2}^{d}(i_j - 1)\prod_{\ell=k+1}^{j-1} n_\ell, :) = (\mathbf{U}_{k+1}(i_{k+1})\mathbf{U}_{k+2}(i_{k+2})\cdots\mathbf{U}_d(i_d))^\top.$$

由此可得 $\mathbf{X}_{\langle k\rangle} = \mathbf{X}_{\leq k}\mathbf{X}_{\geq k+1}^\top$ 并且接口矩阵可以通过如下递推方式构造:

$$\mathbf{X}_{\leq k} = (\mathbf{I}_{n_k} \otimes \mathbf{X}_{\leq k-1})\mathrm{L}(\mathcal{U}_k) \quad 和 \quad \mathbf{X}_{\geq k+1} = (\mathbf{X}_{\geq k+2} \otimes \mathbf{I}_{n_{k+1}})\mathrm{R}(\mathcal{U}_{k+1})^\top. \tag{1-7}$$

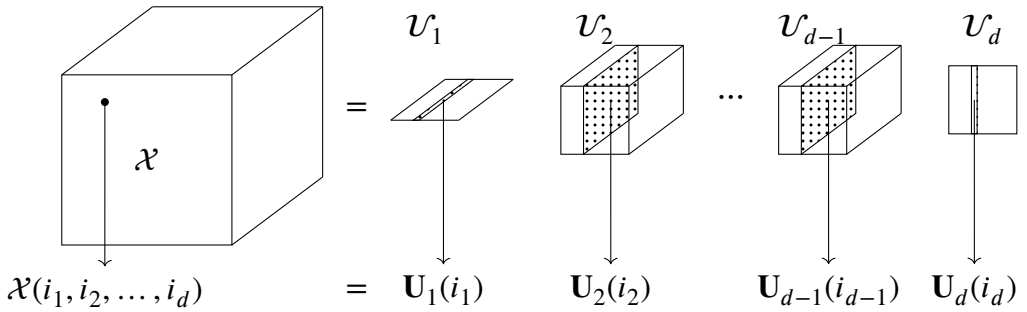


图 1-5 一个张量的张量链分解.

Figure 1-5 Tensor train decomposition of a tensor.

张量链分解的示意图如图 1-5所示. 张量 $\mathcal{X}$ 的张量链秩 (TT 秩) 定义为

$$\mathrm{rank}_{\mathrm{TT}}(\mathcal{X}) := (1, \mathrm{rank}(\mathbf{X}_{\langle 1\rangle}), \mathrm{rank}(\mathbf{X}_{\langle 2\rangle}), \ldots, \mathrm{rank}(\mathbf{X}_{\langle d-1\rangle}), 1).$$

若张量 $\mathcal{X} = [\![\mathcal{U}_1, \mathcal{U}_2, \ldots, \mathcal{U}_d]\!]$ 满足对所有 $j \in [k-1]$, 有 $\mathrm{L}(\mathcal{U}_j) \in \mathrm{St}_{\mathbb{C}}(r_j, r_{j-1}n_j)$, 且对所有 $j = k+1, k+2, \ldots, d$, 有 $\mathrm{R}(\mathcal{U}_j)^\top \in \mathrm{St}_{\mathbb{C}}(r_{j-1}, n_j r_j)$, 则称 $\mathcal{X}$ 为 k-正交张量. 特别地, 当 $k = d$ 或 $k = 1$ 时, $\mathcal{X}$ 分别称为左正交或右正交张量. 由 [60, §3.1] 可知, 任意张量 $\mathcal{X}$ 都可以通过 QR 分解实现左正交或右正交化. 正交化在 TT 张量的运算中扮演着十分重要的角色, 详见第5章.

接下来我们介绍张量环 (tensor ring, TR) 分解.

定义 1.6 (张量环分解 [42])**.** 给定一个 d 阶张量 $\mathcal{X} \in \mathbb{R}^{n_1\times\cdots\times n_d}$, 张量环分解, 记为

$$\mathcal{X} = [\![\mathcal{U}_1, \mathcal{U}_2, \ldots, \mathcal{U}_d]\!],$$

将 $\mathcal{X}$ 表示为 d 个三阶核张量的乘积形式, 其中 $\mathcal{U}_k \in \mathbb{R}^{r_k\times n_k\times r_{k+1}}$, $k \in [d]$, 并约定 $r_{d+1} = r_1$. 具体而言, $\mathcal{X}$ 在位置 $(i_1, i_2, \ldots, i_d)$ 处的元素可以表示为 d 个矩阵乘积的迹, 即

$$\mathcal{X}(i_1, i_2, \ldots, i_d) = \mathrm{tr}\left(\mathbf{U}_1(i_1)\mathbf{U}_2(i_2)\cdots\mathbf{U}_d(i_d)\right),$$

其中 $\mathbf{U}_k(i_k) = \mathcal{U}_k(:, i_k, :) \in \mathbb{R}^{r_k\times r_{k+1}}$ 表示核张量 $\mathcal{U}_k$ 在第 i_k 个索引处对应的侧向切片矩阵, $i_k \in [n_k]$.

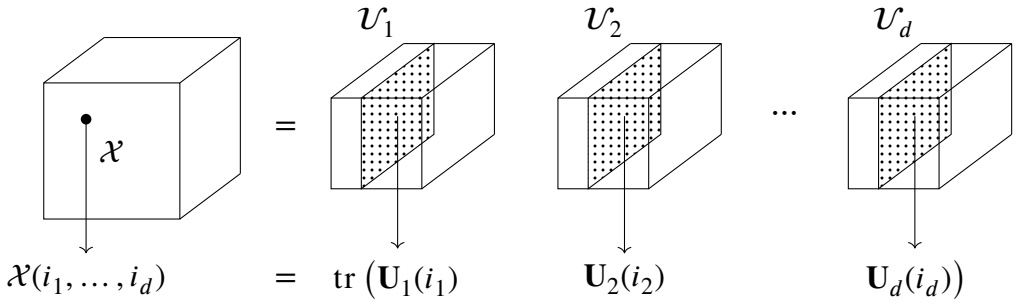


图 1-6 一个张量的张量环分解.

Figure 1-6 Illustration of tensor ring decomposition of a tensor.

元组 $\mathbf{r} = (r_1, r_2, \ldots, r_d)$ 被称为张量环秩. 图 1-6 展示了一个张量的张量环分解.

根据迹算子的循环不变性, 我们可以将张量环分解中张量 $\mathcal{X}$ 的 $(i_1, i_2, \ldots, i_d)$ 元素重写为如下形式:

$$\begin{aligned}&\operatorname{tr}(\mathbf{U}_1(i_1)\cdots\mathbf{U}_d(i_d)) = \operatorname{tr}\left(\mathbf{U}_k(i_k)\cdots\mathbf{U}_d(i_d)\mathbf{U}_1(i_1)\cdots\mathbf{U}_{k-1}(i_{k-1})\right)\\&= \left\langle \operatorname{vec}(\mathbf{U}_k(i_k)), \operatorname{vec}\left(\left(\mathbf{U}_{k+1}(i_{k+1})\cdots\mathbf{U}_d(i_d)\mathbf{U}_1(i_1)\cdots\mathbf{U}_{k-1}(i_{k-1})\right)^{\mathsf{T}}\right)\right\rangle,\end{aligned}$$

其中, $\operatorname{vec}(\cdot)$ 表示矩阵张量的列向量化算子. 事实上, $\operatorname{vec}(\mathbf{U}_k(i_k))^{\mathsf{T}}$ 正是核张量 $\mathcal{U}_k$ 的模态 2 展平矩阵 $\left(\mathcal{U}_k\right)_{(2)} \in \mathbb{R}^{n_k \times r_k r_{k+1}}$ 的第 i_k 行. 此外, 对于给定的矩阵 $\mathbf{W}_k \in \mathbb{R}^{n_k \times (r_k r_{k+1})}$(其中 $n_1, \ldots, n_d$ 和 $r_1, \ldots, r_d$ 固定), 模态二张量化算子将 $\mathbf{W}_k$ 映射为一个张量 $\operatorname{ten}_{(2)}(\mathbf{W}_k) \in \mathbb{R}^{r_k \times n_k \times r_{k+1}}$, 其定义为

$$\operatorname{ten}_{(2)}(\mathbf{W}_k)(i_1, i_2, i_3) := \mathbf{W}_k(i_2, i_1 + (i_3 - 1)r_k),$$

其中 $(i_1, i_2, i_3) \in [r_k] \times [n_k] \times [r_{k+1}]$. 可以验证, $(\operatorname{ten}_{(2)}(\mathbf{W}_k))_{(2)} = \mathbf{W}_k$ 恒成立, 因此第二张量化算子是可逆的. 下面给出子链张量的定义.

定义 1.7 (子链张量)**.** 子链张量 $\mathcal{U}_{\neq k} \in \mathbb{R}^{r_k \times n_{-k} \times r_{k+1}}$ 通过其侧向切片矩阵来定义, 即

$$\mathbf{U}_{\neq k}\left(\pi_k(i_1, \ldots, i_{k-1}, i_{k+1}, \ldots, i_d)\right) := \left(\prod_{j=k+1}^{d} \mathbf{U}_j(i_j) \prod_{j=1}^{k-1} \mathbf{U}_j(i_j)\right)^{\mathsf{T}}, \tag{1-8}$$

其中 $(i_1, \ldots, i_{k-1}, i_{k+1}, \ldots, i_d) \in [n_1] \times \cdots \times [n_{k-1}] \times [n_{k+1}] \times \cdots \times [n_d]$, 且 $k \in [d]$.

利用核张量 $\mathcal{U}_k$ 的模态 2 展平矩阵 $\left(\mathcal{U}_k\right)_{(2)}$ 以及定义 1.7 中的子链张量, 可以将具有张量环分解形式的张量 $\mathcal{X}$ 的模态 k 展平矩阵表示为两个较小矩阵的乘积 (见 [42, Theorem 3.5]), 即

$$\mathbf{X}_{(k)} = \mathbf{W}_k \mathbf{W}_{\neq k}^{\mathsf{T}},$$

其中 $\mathbf{W}_k := \left(\mathcal{U}_k\right)_{(2)} \in \mathbb{R}^{n_k \times r_k r_{k+1}}$ 且 $\mathbf{W}_{\neq k} := \left(\mathcal{U}_{\neq k}\right)_{(2)} \in \mathbb{R}^{n_{-k} \times r_k r_{k+1}}$, 上式对所有 $k \in [d]$ 均成立.

1.3 低秩矩阵与张量构成的集合

设 m, n, r 为正整数, 且满足 $r \leq \min\{m, n\}$. 给定矩阵 $\mathbf{X} \in \mathbb{R}^{m\times n}$, 其像空间及其正交补分别定义为 $\operatorname{span}(\mathbf{X}) := \{\mathbf{X}\mathbf{y} : \mathbf{y} \in \mathbb{R}^n\} \subseteq \mathbb{R}^m$ 和 $\operatorname{span}(\mathbf{X})^\perp := \{\mathbf{y} \in \mathbb{R}^m : \langle \mathbf{x}, \mathbf{y}\rangle = 0 \text{ 对所有的 } \mathbf{x} \in \operatorname{span}(\mathbf{X})\}$. 接下来我们介绍低秩矩阵、Tucker 与 TT 张量所构成的集合及其相应的几何性质.

1.3.1 低秩矩阵构成的集合

秩为 r 的矩阵构成的集合 $\mathbb{R}_r^{m\times n} := \{\mathbf{X} \in \mathbb{R}^{m\times n} : \operatorname{rank}(\mathbf{X}) = r\}$ 是一个光滑流形 (详见 [61, 62]), 我们称之为矩阵流形. 秩不超过 r 的矩阵构成的集合即为 $\mathbb{R}_{\leq r}^{m\times n} := \{\mathbf{X} \in \mathbb{R}^{m\times n} : \operatorname{rank}(\mathbf{X}) \leq r\}$. 由于它能通过 $(r+1)$ 阶余子式等于 0 构造, 因此 $\mathbb{R}_{\leq r}^{m\times n}$ 是一个代数簇, 我们称之为矩阵代数簇.

接下来我们介绍矩阵流形与矩阵代数簇的几何性质. 切锥与法锥在矩阵代数簇 $\mathbb{R}_{\leq r}^{m\times n}$ 上的优化问题中起着至关重要的作用. 因此, 我们在此给出矩阵代数簇的切锥与法锥的显式表达式 (参见 [14, Proposition 2.1] 以及 [63, Theorem 3.2]), 如下所示.

命题 1.1. *给定矩阵 $\mathbf{X} \in \mathbb{R}_{\underline{r}}^{m\times n}$, 它的秩为 $\underline{r} \leq r$. 矩阵 $\mathbf{X}$ 的一个紧奇异值分解 (thin SVD) 为 $\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^\top$, 这里 $\mathbf{U} \in \operatorname{St}(\underline{r}, m)$, $\mathbf{V} \in \operatorname{St}(\underline{r}, n)$ 以及 $\Sigma = \operatorname{diag}(\sigma_1, \sigma_2, \ldots, \sigma_{\underline{r}})$ 满足 $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{\underline{r}} > 0$. 则对于所有满足 $\operatorname{span}(\mathbf{U}^\perp) = \operatorname{span}(\mathbf{U})^\perp$ 的 $\mathbf{U}^\perp \in \operatorname{St}(m - \underline{r}, m)$ 以及满足 $\operatorname{span}(\mathbf{V}^\perp) = \operatorname{span}(\mathbf{V})^\perp$ 的 $\mathbf{V}^\perp \in \operatorname{St}(n - \underline{r}, n)$, 我们有*

$$
\begin{aligned}
\mathrm{T}_{\mathbf{X}}\mathbb{R}_{\underline{r}}^{m\times n} &= \left\{ \begin{bmatrix} \mathbf{U} & \mathbf{U}^\perp \end{bmatrix} \begin{bmatrix} \mathbb{R}^{\underline{r}\times\underline{r}} & \mathbb{R}^{\underline{r}\times(n-\underline{r})} \\ \mathbb{R}^{(m-\underline{r})\times\underline{r}} & 0 \end{bmatrix} \begin{bmatrix} \mathbf{V} & \mathbf{V}^\perp \end{bmatrix}^\top \right\}, \\
\mathrm{N}_{\mathbf{X}}\mathbb{R}_{\underline{r}}^{m\times n} &= \left\{ \begin{bmatrix} \mathbf{U} & \mathbf{U}^\perp \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & \mathbb{R}^{(m-\underline{r})\times(n-\underline{r})} \end{bmatrix} \begin{bmatrix} \mathbf{V} & \mathbf{V}^\perp \end{bmatrix}^\top \right\}, \\
\mathrm{T}_{\mathbf{X}}\mathbb{R}_{\leq r}^{m\times n} &= \left\{ \begin{bmatrix} \mathbf{U} & \mathbf{U}^\perp \end{bmatrix} \begin{bmatrix} \mathbb{R}^{\underline{r}\times\underline{r}} & \mathbb{R}^{\underline{r}\times(n-\underline{r})} \\ \mathbb{R}^{(m-\underline{r})\times\underline{r}} & \mathbb{R}_{\leq(r-\underline{r})}^{(m-\underline{r})\times(n-\underline{r})} \end{bmatrix} \begin{bmatrix} \mathbf{V} & \mathbf{V}^\perp \end{bmatrix}^\top \right\}, \\
\mathrm{N}_{\mathbf{X}}\mathbb{R}_{\leq r}^{m\times n} &= \begin{cases} \mathrm{N}_{\mathbf{X}}\mathbb{R}_r^{m\times n}, & \text{若 } \underline{r} = r; \\ \{0\}, & \text{若 } \underline{r} < r \end{cases}
\end{aligned}
$$

成立.

注意到满足 $(\mathbf{U}^\perp)^\top\mathbf{U} = 0$ 且 $(\mathbf{V}^\perp)^\top\mathbf{V} = 0$ 的 $\mathbf{U}^\perp$ 与 $\mathbf{V}^\perp$ 并非唯一, 但命题 1.1 中的结论与具体选取的 $\mathbf{U}^\perp$ 和 $\mathbf{V}^\perp$ 无关. 事实上, 从张量积的意义上看, 切空间与法空间可以唯一地表示为

$$
\begin{aligned}
\mathrm{T}_{\mathbf{X}}\mathbb{R}_{\underline{r}}^{m\times n} &= \operatorname{span}(\mathbf{U}) \otimes \operatorname{span}(\mathbf{V}) + \operatorname{span}(\mathbf{U})^\perp \otimes \operatorname{span}(\mathbf{V}) + \operatorname{span}(\mathbf{U}) \otimes \operatorname{span}(\mathbf{V})^\perp, \\
\mathrm{N}_{\mathbf{X}}\mathbb{R}_{\underline{r}}^{m\times n} &= \operatorname{span}(\mathbf{U})^\perp \otimes \operatorname{span}(\mathbf{V})^\perp.
\end{aligned}
$$

切空间与法空间的直和构成整个欧氏空间 $\mathbb{R}^{m\times n}$, 即

$$\begin{aligned}\mathbb{R}^{m\times n} &= \mathrm{T}_{\mathbf{X}}\mathbb{R}_{\underline{r}}^{m\times n} + \mathrm{N}_{\mathbf{X}}\mathbb{R}_{\underline{r}}^{m\times n}\\ &= \left\{ \begin{bmatrix}\mathbf{U} & \mathbf{U}^{\perp}\end{bmatrix} \begin{bmatrix}\blacksquare & \blacksquare\\ \blacksquare & \end{bmatrix} \begin{bmatrix}\mathbf{V} & \mathbf{V}^{\perp}\end{bmatrix}^{\top} \right\} + \left\{ \begin{bmatrix}\mathbf{U} & \mathbf{U}^{\perp}\end{bmatrix} \begin{bmatrix} & \\ & \blacksquare\end{bmatrix} \begin{bmatrix}\mathbf{V} & \mathbf{V}^{\perp}\end{bmatrix}^{\top} \right\},\end{aligned}$$

其中阴影方块表示任意矩阵, 而空白方块表示元素全为零的矩阵.

度量投影　给定矩阵 $\mathbf{A} = \sum_{k=1}^{I} \sigma_k \mathbf{u}_k \mathbf{v}_k^{\top} \in \mathbb{R}_I^{m\times n}$, 其中 $\sigma_1 \geq \cdots \geq \sigma_I > 0$, $\mathbf{u}_k \in \mathbb{R}^m$ 与 $\mathbf{v}_k \in \mathbb{R}^n$ 为 $\mathbf{A}$ 的奇异向量. 矩阵 $\mathbf{A}$ 到矩阵代数簇 $\mathbb{R}_{\leq r}^{m\times n}$ 上的度量投影定义为

$$\mathrm{P}_{\leq r}(\mathbf{A}) := \operatorname*{arg\,min}_{\mathbf{X}\in\mathbb{R}_{\leq r}^{m\times n}} \|\mathbf{A}-\mathbf{X}\|_{\mathrm{F}}^2.$$

由 Eckart–Young 定理可知, 该度量投影是存在的, 且可表示为

$$\mathrm{P}_{\leq r}(\mathbf{A}) = \begin{cases} \mathbf{A}, & \text{if } I \leq r,\\ \sum_{k=1}^{r} \sigma_k \mathbf{u}_k \mathbf{v}_k^{\top}, & \text{if } I > r. \end{cases}$$

需要注意的是, 当奇异值 σ_r 与 σ_{r+1} 相等时, $\mathrm{P}_{\leq r}(\mathbf{A})$ 并非唯一, 但在实际计算中始终可以选取 $\sum_{k=1}^{r} \sigma_k \mathbf{u}_k \mathbf{v}_k^{\top}$ 作为其一个代表. 当 $\operatorname{rank}(\mathbf{A}) \geq r$ 时, 投影到固定秩流形 $\mathbb{R}_r^{m\times n}$ 上的度量投影 $\mathrm{P}_r(\mathbf{A})$ 等于 $\mathrm{P}_{\leq r}(\mathbf{A})$. 然而, 当 $\operatorname{rank}(\mathbf{A}) < r$ 时, 矩阵 $\mathbf{A}$ 可以被秩为 r 的矩阵任意逼近, 因此 $\mathrm{P}_r(\mathbf{A})$ 不存在.

此外, 根据命题 1.1, 对于任意矩阵 $\mathbf{A} \in \mathbb{R}^{m\times n}$, 其到切空间与切锥的正交投影分别为

$$\begin{aligned}\mathrm{P}_{\mathrm{T}_{\mathbf{X}}\mathbb{R}_{\underline{r}}^{m\times n}}\mathbf{A} &= \mathrm{P}_{\mathbf{U}}\mathbf{A}\mathrm{P}_{\mathbf{V}} + \mathrm{P}_{\mathbf{U}}^{\perp}\mathbf{A}\mathrm{P}_{\mathbf{V}} + \mathrm{P}_{\mathbf{U}}\mathbf{A}\mathrm{P}_{\mathbf{V}}^{\perp},\\ \mathrm{P}_{\mathrm{T}_{\mathbf{X}}\mathbb{R}_{\leq r}^{m\times n}}\mathbf{A} &= \mathrm{P}_{\mathrm{T}_{\mathbf{X}}\mathbb{R}_{\underline{r}}^{m\times n}}\mathbf{A} + \mathrm{P}_{\leq(r-\underline{r})}\left(\mathrm{P}_{\mathbf{U}}^{\perp}\mathbf{A}\mathrm{P}_{\mathbf{V}}^{\perp}\right).\end{aligned} \tag{1-9}$$

在实际计算中, $\mathrm{P}_{\mathbf{U}}^{\perp}\mathbf{A}\mathrm{P}_{\mathbf{V}}^{\perp}$ 可以通过如下方式高效计算: $\mathrm{P}_{\mathbf{U}}^{\perp}\mathbf{A}\mathrm{P}_{\mathbf{V}}^{\perp} = \mathbf{A} - \mathbf{U}(\mathbf{U}^{\top}\mathbf{A}) - (\mathbf{A}\mathbf{V})\mathbf{V}^{\top} + \mathbf{U}(\mathbf{U}^{\top}\mathbf{A}\mathbf{V})\mathbf{V}^{\top}$.

1.3.2　低秩 Tucker 张量构成的集合

在本节中, 我们首先介绍固定秩 Tucker 张量的集合构成的流形, 然后介绍往低秩集合上的度量投影.

固定秩 Tucker 张量集合　由于 $\mathcal{M}_{\mathbf{r}} = \{\mathcal{X} \in \mathbb{R}^{n_1\times n_2\times\cdots\times n_d} : \operatorname{rank}_{\mathrm{tc}}(\mathcal{X}) = \mathbf{r}\}$ 构成一个光滑流形, 其维数为 $\dim(\mathcal{M}_{\mathbf{r}}) = r_1 r_2 \cdots r_d + \sum_{k=1}^{d}(n_k r_k - r_k^2)$, Koch 和 Lubich [64] 给出了该流形在点 $\mathcal{X}$ 处的切空间, 其刻画如下:

$$\mathrm{T}_{\mathcal{X}}\mathcal{M}_{\mathbf{r}} = \left\{ \begin{array}{l} \dot{\mathcal{G}} \times_1 \mathbf{U}_1 \cdots \times_d \mathbf{U}_d + \sum_{k=1}^{d} \mathcal{G} \times_k \dot{\mathbf{U}}_k \times_{j\neq k} \mathbf{U}_j : \\ \dot{\mathcal{G}} \in \mathbb{R}^{r_1\times r_2\times\cdots\times r_d}, \dot{\mathbf{U}}_k \in \mathbb{R}^{n_k\times r_k}, \dot{\mathbf{U}}_k^{\top}\mathbf{U}_k = 0 \end{array} \right\}, \tag{1-10}$$

其中 $\mathcal{G} \times_k \dot{\mathbf{U}}_k \times_{j\neq k} \mathbf{U}_j = \mathcal{G} \times_1 \mathbf{U}_1 \cdots \times_{k-1} \mathbf{U}_{k-1} \times_k \dot{\mathbf{U}}_k \times_{k+1} \mathbf{U}_{k+1} \cdots \times_d \mathbf{U}_d$.

尽管张量的 Tucker 分解并不唯一 [38, §4.3], 上述对 $\mathrm{T}_{\mathcal{X}}\mathcal{M}_{\mathbf{r}}$ 的参数化并不依赖于某一特定的 Tucker 分解. 具体而言, 设 $\mathcal{X}$ 的另一种 Tucker 分解为 $\mathcal{X} = \check{\mathcal{G}} \times_{k=1}^{d} \check{\mathbf{U}}_k$ 并通过式 (1-10) 构造相应的切空间 $\check{\mathrm{T}}_{\mathcal{X}}\mathcal{M}_{\mathbf{r}}$, 其中 $\check{\mathcal{G}} \in \mathbb{R}^{r_1 \times r_2 \times \cdots \times r_d}$, $\check{\mathbf{U}}_k \in \mathrm{St}(r_k, n_k)$, $k \in [d]$. 只需证明 $\mathrm{T}_{\mathcal{X}}\mathcal{M}_{\mathbf{r}} = \check{\mathrm{T}}_{\mathcal{X}}\mathcal{M}_{\mathbf{r}}$. 注意到 $\operatorname{span}(\mathbf{U}_k) = \operatorname{span}(\mathbf{X}_{(k)}) = \operatorname{span}(\check{\mathbf{U}}_k)$, 因此存在 $\mathbf{Q}_k \in \mathcal{O}(r_k)$, 使得 $\check{\mathbf{U}}_k = \mathbf{U}_k\mathbf{Q}_k$ 和 $\check{\mathcal{G}} = \mathcal{G} \times_{i=1}^{d} \mathbf{Q}_i^{\top}$ 对 $k \in [d]$ 成立. 对任意 $\mathcal{V} \in \mathrm{T}_{\mathcal{X}}\mathcal{M}_{\mathbf{r}}$, 有

$$\begin{aligned}
\mathcal{V} &= \dot{\mathcal{G}} \times_{k=1}^{d} \mathbf{U}_k + \sum_{k=1}^{d} \mathcal{G} \times_k \dot{\mathbf{U}}_k \times_{j \neq k} \mathbf{U}_j \\
&= (\dot{\mathcal{G}} \times_{k=1}^{d} \mathbf{Q}_k^{\top}) \times_{k=1}^{d} (\mathbf{U}_k\mathbf{Q}_k) + \sum_{k=1}^{d} (\mathcal{G} \times_{i=1}^{d} \mathbf{Q}_i^{\top}) \times_k (\dot{\mathbf{U}}_k\mathbf{Q}_k) \times_{j \neq k} (\mathbf{U}_j\mathbf{Q}_j) \\
&= (\dot{\mathcal{G}} \times_{k=1}^{d} \mathbf{Q}_k^{\top}) \times_{k=1}^{d} \check{\mathbf{U}}_k + \sum_{k=1}^{d} \check{\mathcal{G}} \times_k (\dot{\mathbf{U}}_k\mathbf{Q}_k) \times_{j \neq k} \check{\mathbf{U}}_j.
\end{aligned}$$

由于 $\dot{\mathcal{G}} \times_{k=1}^{d} \mathbf{Q}_k^{\top} \in \mathbb{R}^{r_1 \times r_2 \times \cdots \times r_d}$, $\dot{\mathbf{U}}_k\mathbf{Q}_k \in \mathbb{R}^{n_k \times r_k}$, 并且

$$(\dot{\mathbf{U}}_k\mathbf{Q}_k)^{\top}\check{\mathbf{U}}_k = \mathbf{Q}_k^{\top}\dot{\mathbf{U}}_k^{\top}\mathbf{U}_k\mathbf{Q}_k = 0,$$

可知 $\mathcal{V} \in \check{\mathrm{T}}_{\mathcal{X}}\mathcal{M}_{\mathbf{r}}$, 即满足式 (1-10) 的刻画. 因此, $\mathrm{T}_{\mathcal{X}}\mathcal{M}_{\mathbf{r}} \subseteq \check{\mathrm{T}}_{\mathcal{X}}\mathcal{M}_{\mathbf{r}}$. 反向包含关系同理成立, 从而二者相等.

度量投影 给定张量 $\mathcal{A} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$, 其到有界秩 Tucker 张量集合 $\mathcal{M}_{\leq \mathbf{r}} := \{\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d} : \operatorname{rank}_{\mathrm{tc}}(\mathcal{X}) \leq \mathbf{r}\}$ 的度量投影定义为

$$\mathrm{P}_{\leq \mathbf{r}}(\mathcal{A}) := \operatorname*{arg\,min}_{\mathcal{X} \in \mathcal{M}_{\leq \mathbf{r}}} \|\mathcal{A} - \mathcal{X}\|_{\mathrm{F}}^2. \tag{1-11}$$

与矩阵情形不同, 一般情况下 $\mathrm{P}_{\leq \mathbf{r}}(\mathcal{A})$ 并不存在显式表达 [58]. 然而, 可以利用高阶奇异值分解 (higher-order singular value decomposition, HOSVD) 构造一个拟最优解. 具体而言, HOSVD 依次将最佳秩 r_k 逼近算子 $\mathrm{P}_{\leq r_k}^{k}$ 作用在 $\mathcal{A}$ 的每一个模态上, 即

$$\mathrm{P}_{\leq \mathbf{r}}^{\mathrm{HO}}(\mathcal{A}) := \mathrm{P}_{\leq r_d}^{d}(\mathrm{P}_{\leq r_{d-1}}^{d-1} \cdots (\mathrm{P}_{\leq r_1}^{1}(\mathcal{A}))), \tag{1-12}$$

其中 $\mathrm{P}_{\leq r_k}^{k}(\mathcal{A}) := \mathrm{ten}_{(k)}(\bar{\mathbf{U}}_k\bar{\mathbf{U}}_k^{\top}\mathbf{A}_{(k)})$, $\bar{\mathbf{U}}_k$ 为 $\mathbf{A}_{(k)}$ 的前 r_k 个奇异向量, 并且 $\mathrm{P}_{\leq \mathbf{r}}^{\mathrm{HO}}$ 与算子 $\{\mathrm{P}_{\leq r_k}^{k}\}_{k=1}^{d}$ 的施加顺序无关 [65, §3]. 由于如下拟优性不等式成立:

$$\|\mathcal{A} - \mathrm{P}_{\leq \mathbf{r}}^{\mathrm{HO}}(\mathcal{A})\|_{\mathrm{F}} \leq \sqrt{d}\|\mathcal{A} - \mathrm{P}_{\leq \mathbf{r}}(\mathcal{A})\|_{\mathrm{F}}. \tag{1-13}$$

该结果可参见 [66, Lemma 2.6], 因此 HOSVD 可以作为到 $\mathcal{M}_{\leq \mathbf{r}}$ 的一种近似投影. 此外, 我们还可以证明, HOSVD 在 $\mathcal{X}$ 的邻域内也是 $\mathcal{M}_{\leq \mathbf{r}}$ 上的一个收缩映射 (retraction).

命题 1.2. 映射 $\mathrm{R}_{\mathcal{X}}^{\mathrm{HO}} : \mathrm{T}_{\mathcal{X}}\mathcal{M}_{\leq \mathbf{r}} \to \mathcal{M}_{\leq \mathbf{r}}$, $\mathrm{R}_{\mathcal{X}}^{\mathrm{HO}}(\mathcal{V}) := \mathrm{P}_{\leq \mathbf{r}}^{\mathrm{HO}}(\mathcal{X}+\mathcal{V})$ 是 $\mathcal{M}_{\leq \mathbf{r}}$ 上的一个收缩映射.

证明. 我们只需证明 $\lim_{t\to 0^+}(\mathrm{R}_{\mathcal{X}}^{\mathrm{HO}}(t\mathcal{V}) - \mathcal{X} - t\mathcal{V})/t = 0$. 由于 $\mathcal{V} \in \mathrm{T}_{\mathcal{X}}\mathcal{M}_{\leq \mathbf{r}}$, 我们根据 [67, Proposition 2] 可得存在一条解析曲线 $\gamma : [0, \epsilon) \to \mathcal{M}_{\leq \mathbf{r}}$ 使得 $\gamma(0) = \mathcal{X}$ 以及 $\dot{\gamma}(0) = \mathcal{V}$ 成立.

于是, 由于 $\mathrm{P}_{\leq \mathbf{r}}(\mathcal{X}+t\mathcal{V})$ 是点 $\mathcal{X}+t\mathcal{V}$ 到集合 $\mathcal{M}_{\leq \mathbf{r}}$ 上的正交投影, 我们有

$$\|\mathcal{X}+t\mathcal{V} - \mathrm{P}_{\leq \mathbf{r}}(\mathcal{X}+t\mathcal{V})\|_{\mathrm{F}} \leq \|\mathcal{X}+t\mathcal{V} - \gamma(t)\|_{\mathrm{F}}.$$

利用拟最优性 (1-13), 我们可以得到

$$\|\mathcal{X}+t\mathcal{V} - \mathrm{R}_{\mathcal{X}}^{\mathrm{HO}}(t\mathcal{V})\|_{\mathrm{F}} \leq \sqrt{d}\|\mathcal{X}+t\mathcal{V} - \mathrm{P}_{\leq \mathbf{r}}(\mathcal{X}+t\mathcal{V})\|_{\mathrm{F}} \leq \sqrt{d}\|\mathcal{X}+t\mathcal{V} - \gamma(t)\|_{\mathrm{F}} = o(t),$$

因此 $\mathrm{R}_{\mathcal{X}}^{\mathrm{HO}}$ 是一个收缩映射. □

最后, 给定张量 $\mathcal{A} \in \mathbb{R}^{n_1\times n_2\times\cdots\times n_d}$, 其到切空间 $\mathrm{T}_{\mathcal{X}}\mathcal{M}_{\mathbf{r}}$ 的正交投影可表示为 [64, §2.3]

$$\mathrm{P}_{\mathrm{T}_{\mathcal{X}}\mathcal{M}_{\mathbf{r}}}\mathcal{A} = \mathcal{A}\times_{k=1}^{d}\mathrm{P}_{\mathbf{U}_k} + \sum_{k=1}^{d}\mathcal{G}\times_k\left(\mathrm{P}_{\mathbf{U}_k}^{\perp}\left(\mathcal{A}\times_{j\neq k}\mathbf{U}_j^{\top}\right)_{(k)}\mathbf{G}_{(k)}^{\dagger}\right)\times_{j\neq k}\mathbf{U}_j, \tag{1-14}$$

其中 $\mathbf{G}_{(k)}^{\dagger} = \mathbf{G}_{(k)}^{\top}(\mathbf{G}_{(k)}\mathbf{G}_{(k)}^{\top})^{-1}$ 为 $\mathcal{G}$ 的模态 k 展平矩阵 $\mathbf{G}_{(k)}$ 的 Moore–Penrose 伪逆.

1.3.3　低秩 TT 张量构成的集合

固定秩 TT 张量集合　固定秩 TT 张量所构成的集合为

$$\mathcal{M}_{\mathbf{r}}^{\mathrm{TT}} = \left\{\mathcal{X} \in \mathbb{C}^{n_1\times n_2\times\cdots\times n_d} : \mathrm{rank}_{\mathrm{TT}}(\mathcal{X}) = \mathbf{r}\right\}, \tag{1-15}$$

该集合是 $\mathbb{C}^{n_1\times n_2\times\cdots\times n_d}$ 的一个复子流形; 参见 [68, 定理 14]. 集合 $\mathcal{M}_{\mathbf{r}}^{\mathrm{TT}}$ 在点 $\mathcal{X} = [\![\mathcal{U}_1, \mathcal{U}_2, \ldots, \mathcal{U}_d]\!] \in \mathcal{M}_{\mathbf{r}}^{\mathrm{TT}}$ 处的切空间有如下的参数化 (参见 [68, 69]):

$$\mathrm{T}_{\mathcal{X}}\mathcal{M}_{\mathbf{r}}^{\mathrm{TT}} = \left\{ \begin{array}{l} [\![\dot{\mathcal{U}}_1, \mathcal{U}_2, \mathcal{U}_3, \ldots, \mathcal{U}_d]\!] \\ +\,[\![\mathcal{U}_1, \dot{\mathcal{U}}_2, \mathcal{U}_3, \ldots, \mathcal{U}_d]\!] \\ \vdots \\ +\,[\![\mathcal{U}_1, \mathcal{U}_2, \ldots, \mathcal{U}_{d-1}, \dot{\mathcal{U}}_d]\!] \end{array} : \begin{array}{c} \dot{\mathcal{U}}_k \in \mathbb{C}^{r_{k-1}\times n_k\times r_k}, k \in [d], \\ \mathrm{L}(\dot{\mathcal{U}}_k)^{\dagger}\mathrm{L}(\mathcal{U}_k) = 0, k \in [d-1] \end{array} \right\}. \tag{1-16}$$

度量投影　由 [69, Theorem 1] 可知, 若 $\mathcal{X}$ 满足 $\mathrm{rank}_{\mathrm{TT}}(\mathcal{X}) = \mathbf{r} = (r_0, r_1, \ldots, r_d)$ 且 $r_0 = r_d = 1$, 则可通过 d 次奇异值分解得到其 TT 分解 $\mathcal{X} = [\![\mathcal{U}_1, \mathcal{U}_2, \ldots, \mathcal{U}_d]\!]$ (等价地, $\mathcal{X}(i_1, i_2, \ldots, i_d) = \mathbf{U}_1(i_1)\mathbf{U}_2(i_2)\cdots\mathbf{U}_d(i_d)$, $i_k \in [n_k]$, $k \in [d]$) 其中核张量 $\mathcal{U}_k \in \mathbb{C}^{r_{k-1}\times n_k\times r_k}$. 该过程称为 TT-SVD 算法, 见 [41, Algorithm 1]. 具体而言, TT-SVD 算法从第一展平矩阵 $\mathbf{X}_{\langle 1\rangle}$ 出发, 依次执行: 1) 将张量重排为矩阵; 2) 对该矩

阵进行 SVD; 3) 将分解结果重排以得到核张量 $\mathcal{U}_k$ 及一个维数更小的张量; 详见图 5-2. TT-SVD 算子 $\mathrm{P}_{\mathbf{r}}^{\mathrm{TTSVD}}$ 满足如下拟优性估计:

$$\|\mathrm{P}_{\mathbf{r}}^{\mathrm{TTSVD}}(\mathcal{A})-\mathcal{A}\|_{\mathrm{F}} \leq \sqrt{d-1}\|\mathrm{P}_{\mathbf{r}}^{\mathrm{TT}}(\mathcal{A})-\mathcal{A}\|_{\mathrm{F}}, \tag{1-17}$$

其中 $\mathrm{P}_{\mathbf{r}}^{\mathrm{TT}}(\mathcal{A})$ 表示 $\mathcal{A}$ 在 TT 格式下的最佳秩 $\mathbf{r}$ 近似.

1.4 低秩矩阵与张量优化简介

由于低秩张量优化与低秩矩阵优化有着紧密的联系, 我们在本节中将回顾已有的低秩矩阵与低秩张量优化中的工作. 总体而言, 针对低秩张量优化问题 (1-2), 主要存在三类不同的几何方法: 流形优化方法, 代数簇优化方法以及参数化方法; 详见图 1-7.

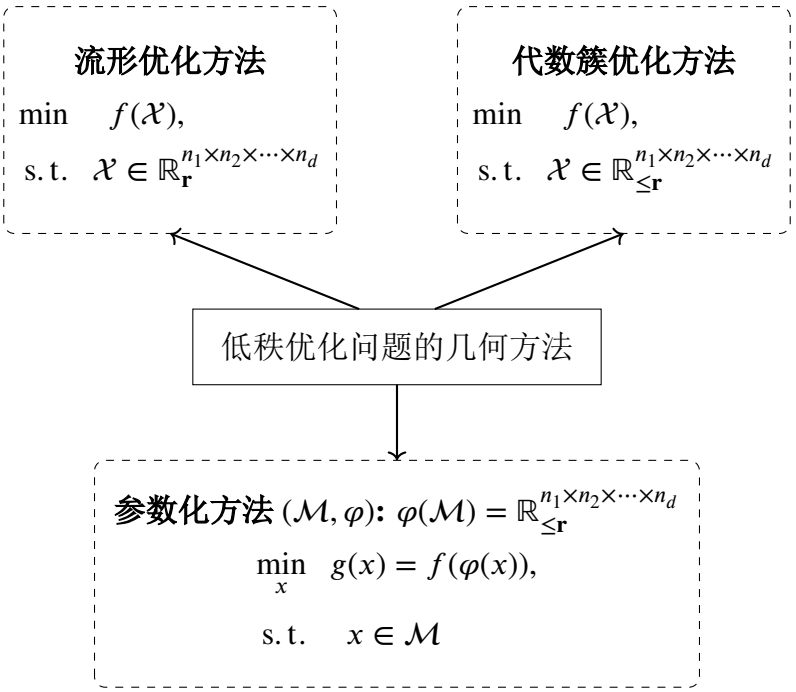


图 1-7 低秩矩阵与张量优化的三类几何方法.

Figure 1-7 Illustration of three types of geometric methods for low-rank optimization.

1.4.1 流形优化方法

在本节中, 我们首先引入黎曼流形的基本记号以及几何对象 (如切空间, 收缩映射, 向量传输算子等), 然后我们介绍流形优化方法. 特别地, 我们将介绍低秩矩阵张量优化问题的流形优化方法.

黎曼流形的基本记号 乘积流形 $\mathcal{M}$ 定义为若干流形的笛卡尔积, 即

$$\mathcal{M} = \mathcal{M}_1 \times \mathcal{M}_2 \times \cdots \times \mathcal{M}_K.$$

值得注意的是当 $K = 1$ 时, 乘积流形退化为单个流形, 下述数学记号仍然良定义. 假设 $\mathcal{M}$ 嵌入在欧氏空间 $\mathcal{E} = \mathcal{E}_1 \times \mathcal{E}_2 \times \cdots \times \mathcal{E}_K$ 中, 该空间称为 $\mathcal{M}$ 的环境空间 (ambient space). 在点 $x = (x_1, x_2, \ldots, x_K)$ 处, $\mathcal{M}$ 的切空间记为 $\mathrm{T}_x\mathcal{M} = \mathrm{T}_{x_1}\mathcal{M}_1 \times \mathrm{T}_{x_2}\mathcal{M}_2 \times \cdots \times \mathrm{T}_{x_K}\mathcal{M}_K$, 相应的切向量记为 $\eta = (\eta_1, \eta_2, \ldots, \eta_K)$. 设每个流形 $\mathcal{M}_k$ 都配备一个黎曼度量 g^k. 则乘积流形 $\mathcal{M}$ 上的黎曼度量可以定义为

$$g_x(\xi, \eta) := g^1_{x_1}(\xi_1, \eta_1) + g^2_{x_2}(\xi_2, \eta_2) + \cdots + g^K_{x_K}(\xi_K, \eta_K),$$

其中 $\xi, \eta \in \mathrm{T}_x\mathcal{M}$. 该度量诱导的范数定义为 $\|\eta\|_x := \sqrt{g_x(\eta, \eta)}$. 给定向量 $\bar{\eta} = (\bar{\eta}_1, \bar{\eta}_2, \ldots, \bar{\eta}_K) \in \mathrm{T}_x\mathcal{E} \simeq \mathcal{E}$, 关于度量 g 到切空间 $\mathrm{T}_x\mathcal{M}$ 的正交投影算子定义为 $\Pi_{g,x}(\bar{\eta}) := (\Pi_{g^1,x_1}(\bar{\eta}_1), \Pi_{g^2,x_2}(\bar{\eta}_2), \ldots, \Pi_{g^K,x_K}(\bar{\eta}_K))$, 其中 Π_{g^k,x_k} 表示在度量 g^k 下到 $\mathrm{T}_{x_k}\mathcal{M}_k$ 的正交投影算子, 这里 $k = 1, 2, \ldots, K$. 记 $\mathrm{T}\mathcal{M} := \cup_{x\in\mathcal{M}} \mathrm{T}_x\mathcal{M}$ 为 $\mathcal{M}$ 的切丛 (tangent bundle). 一个光滑映射 $\mathrm{R} : \mathrm{T}\mathcal{M} \to \mathcal{M}$ 若满足 $\mathrm{R}_x(0_x) = x$ 和 $\mathrm{DR}_x(0_x) = \mathrm{I}_x$ 则称其为一个收缩映射 (retraction). 其中 $0_x \in \mathrm{T}_x\mathcal{M}$ 表示零切向量, $\mathrm{DR}_x(0_x)$ 是 R_x 在 0_x 处的微分, $\mathrm{I}_x : \mathrm{T}_x\mathcal{M} \to \mathrm{T}_x\mathcal{M}$ 是切空间上 $\mathrm{T}_x\mathcal{M}$ 的恒等算子. 乘积流形 $\mathcal{M}$ 上的收缩映射可以逐分量定义为 $\mathrm{R}_x(\eta) := (\mathrm{R}^1_{x_1}(\eta_1), \mathrm{R}^2_{x_2}(\eta_2), \ldots, \mathrm{R}^K_{x_K}(\eta_K))$, 其中 R^k 是流形 $\mathcal{M}_k$ 上的一个收缩映射. 向量传输算子 (vector transport) 记为 $\mathcal{T}_{y\leftarrow x} : \mathrm{T}_x\mathcal{M} \to \mathrm{T}_y\mathcal{M}$ 其中 $x, y \in \mathcal{M}$.

设 $f : \mathcal{M} \to \mathbb{R}$ 是一个光滑函数. 在度量 g 下, f 的黎曼梯度记为 $\mathrm{grad}_g f(x)$, 其定义为对任意 $\eta \in \mathrm{T}_x\mathcal{M}$ 唯一满足 $g_x(\mathrm{grad}_g f(x), \eta) = \mathrm{D}f(x)[\eta]$ 的切向量, 其中 $\mathrm{D}f(x)[\eta]$ 表示 f 在 x 处沿方向 η 的微分. f 在点 x 处、相对于度量 g 的黎曼 Hessian 算子定义为 $\mathrm{Hess}_g f(x)[\eta] := \nabla_\eta \mathrm{grad}_g f$, 其中 ∇ 表示 $\mathcal{M}$ 上的 Levi–Civita 联络. 若 $\mathcal{M}$ 是欧氏空间 $\mathcal{E}$ 的一个黎曼子流形, 则由 [70, Corollary 5.1.6] 可得

$$\mathrm{Hess}_{\mathrm{e}} f(x)[\eta] = \Pi_{\mathrm{e},x}(\mathrm{D}\bar{G}(x)[\eta]), \tag{1-18}$$

其中 $\bar{G}$ 是 $\mathrm{grad}_{\mathrm{e}} f(x)$ 在 $\mathcal{M}$ 邻域内的一个光滑延拓, $\mathrm{grad}_{\mathrm{e}} f(x)$ 和 $\mathrm{Hess}_{\mathrm{e}} f(x)$ 分别表示在欧氏度量下 f 的黎曼梯度与黎曼 Hessian 算子.

流形优化方法概述　通过组合上述的几何工具, 我们在算法 1和 2中分别展示黎曼梯度法与黎曼共轭梯度法的算法框架. RGD 和 RCG 算法的收敛性可以参考文献 [71, 72].

我们发现 RGD 和 RCG 算法中的黎曼梯度取决于给定的度量 g. 也就是说, 不同的黎曼度量决定了不同的黎曼梯度. 此外, RGD 和 RCG 算法每步迭代的计算量也取决于度量. 因此, 选择一个合适的黎曼度量可以提升流形优化方法的算法表现.

定义 1.8 (一阶稳定点)**.** 设 f 是定义在赋予度量 g 的流形 $\mathcal{M}$ 上的一个光滑函数. 若点 $x^* \in \mathcal{M}$ 满足 $\mathrm{grad}_g f(x^*) = 0$, 则称 x^* 为函数 f 的一个一阶稳定点.

需要注意的是, 黎曼梯度的定义依赖于度量 g, 而函数 f 的一阶稳定点集合却与度量的选择无关; 见下述命题.

算法 1 黎曼梯度法 (RGD)

输入: 黎曼流形 $(\mathcal{M}, g)$, 初始点 $x^{(0)} \in \mathcal{M}$, $t = 0$.

1: **while** 停机准则未被满足 **do**
2: 计算 $\eta^{(t)} = -\mathrm{grad}_g f(x^{(t)})$.
3: 计算步长 $s^{(t)}$.
4: 更新 $x^{(t+1)} = \mathrm{R}_{x^{(t)}}(s^{(t)}\eta^{(t)})$; $t = t + 1$.
5: **end while**

输出: $x^{(t)} \in \mathcal{M}$.

算法 2 黎曼共轭梯度法 (RCG)

输入: 黎曼流形 $(\mathcal{M}, g)$, 初始点 $x^{(0)} \in \mathcal{M}$, $t = 0$, $\beta^{(0)} = 0$.

1: **while** 停机准则未被满足 **do**
2: 计算 $\eta^{(t)} = -\mathrm{grad}_g f(x^{(t)}) + \beta^{(t)} \mathcal{T}_{x^{(t)} \leftarrow x^{(t-1)}} \eta^{(t-1)}$, 这里 $\beta^{(t)}$ 是共轭参数.
3: 计算步长 $s^{(t)}$.
4: 更新 $x^{(t+1)} = \mathrm{R}_{x^{(t)}}(s^{(t)}\eta^{(t)})$; $t = t + 1$.
5: **end while**

输出: $x^{(t)} \in \mathcal{M}$.

命题 1.3. *设 f 是定义在流形 $\mathcal{M}$ 上的一个光滑函数. 考虑两个黎曼流形 $(\mathcal{M}, g)$ 与 $(\mathcal{M}, \tilde{g})$, 则对任意 $x \in \mathcal{M}$, 有*

$$g_x(\mathrm{grad}_g f(x), \mathrm{grad}_{\tilde{g}} f(x)) \geq 0 \quad \text{且} \quad \tilde{g}_x(\mathrm{grad}_g f(x), \mathrm{grad}_{\tilde{g}} f(x)) \geq 0.$$

等号成立当且仅当 x 是一个一阶稳定点. 此外, $\mathrm{grad}_g f(x) = 0$ 当且仅当 $\mathrm{grad}_{\tilde{g}} f(x) = 0$.

证明. 只需证明第一个不等式, 另一个可采用相同的方法证明. 设 $(\mathcal{U}, \varphi)$ 是流形 $\mathcal{M}$ 的一个坐标图, E_i 表示第 i 个坐标向量场. 对于向量场 $\zeta = \sum_i \alpha_i E_i$ 和 $\chi = \sum_i \beta_i E_i$, 根据黎曼度量 g 的定义, 有

$$g_x(\zeta_x, \chi_x) = \sum_{i,j} g_{ij}\alpha_i\beta_j = \hat{\zeta}_{\hat{x}}^{\top} \mathbf{G}_{\hat{x}} \hat{\chi}_{\hat{x}},$$

其中 $\hat{x} := \varphi(x)$, $\hat{\zeta}_{\hat{x}} := \mathrm{D}\varphi(\varphi^{-1}(\hat{x}))[\zeta_x]$, $\hat{\chi}_{\hat{x}} := \mathrm{D}\varphi(\varphi^{-1}(\hat{x}))[\chi_x]$, 而矩阵 $\mathbf{G}_{\hat{x}}$ 的第 (i, j) 个元素为 $g_{ij} := g(E_i, E_j)$.

记 $\zeta_x := \mathrm{grad}_g f(x)$, $\chi_x := \mathrm{grad}_{\tilde{g}} f(x)$. 由坐标表示形式 (参见 [71, §3.6]) 可知

$$\hat{\zeta}_{\hat{x}} = \mathbf{G}_{\hat{x}}^{-1} \nabla \hat{f}(\hat{x}), \qquad \hat{\chi}_{\hat{x}} = \tilde{\mathbf{G}}_{\hat{x}}^{-1} \nabla \hat{f}(\hat{x}),$$

其中 $\hat{f}(\hat{x}) := f \circ \varphi^{-1}(\hat{x})$, $\nabla \hat{f}$ 表示 $\hat{f}$ 的欧氏梯度. 于是可得

$$g_x(\mathrm{grad}_g f(x), \mathrm{grad}_{\tilde{g}} f(x)) = \hat{\zeta}_{\hat{x}}^{\top} \mathbf{G}_{\hat{x}} \hat{\chi}_{\hat{x}} = (\nabla \hat{f}(\hat{x}))^{\top} \tilde{\mathbf{G}}_{\hat{x}}^{-1} \nabla \hat{f}(\hat{x}) \geq 0.$$

等号成立当且仅当 $\nabla \hat{f}(\hat{x}) = 0$, 即 $\mathrm{grad}_g f(x) = \mathrm{grad}_{\tilde{g}} f(x) = 0$. 此外, 若 $\mathrm{grad}_g f(x) = 0$, 则有 $\hat{\xi}_{\hat{x}} = 0$, 从而

$$\hat{\chi}_{\hat{x}} = \tilde{\mathbf{G}}_{\hat{x}}^{-1} \nabla \hat{f}(\hat{x}) = \tilde{\mathbf{G}}_{\hat{x}}^{-1} \mathbf{G}_{\hat{x}} \mathbf{G}_{\hat{x}}^{-1} \nabla \hat{f}(\hat{x}) = \tilde{\mathbf{G}}_{\hat{x}}^{-1} \mathbf{G}_{\hat{x}} \hat{\xi}_{\hat{x}} = 0,$$

即 $\mathrm{grad}_{\tilde{g}} f(x) = 0$. □

函数 f 的二阶稳定点定义如下.

定义 1.9 (二阶稳定点)**.** 设 f 是定义在赋予度量 g 的流形 $\mathcal{M}$ 上的一个光滑函数. 若 $x^* \in \mathcal{M}$ 是 f 的一个一阶稳定点, 并且 $\mathrm{Hess}_g f(x^*)$ 为半正定, 则称 x^* 为 f 的一个二阶稳定点. 进一步地, 若 $\mathrm{Hess}_g f(x^*)$ 为正定, 则 x^* 是问题 (2-1) 的一个局部极小点.

需要注意的是, 二阶稳定点的集合在不同度量意义下也是不变的, 参见 [70, Proposition 6.3]. 具体而言, 若 x^* 是函数 f 的一个二阶稳定点, 则对于不同的度量 g 与 $\tilde{g}$, 有 $\mathrm{Hess}_g f(x^*)$ 为半正定当且仅当 $\mathrm{Hess}_{\tilde{g}} f(x^*)$ 为半正定.

流形优化方法的局部收敛性　我们从条件数的角度给出黎曼梯度法的局部收敛性质. 在算法 1 中, 我们采用 Armijo 回溯线搜索来计算步长.

定义 1.10 (Armijo 回溯线搜索)**.** 设 f 是定义在赋予度量 g 的流形 $\mathcal{M}$ 上的光滑函数, $x \in \mathcal{M}$ 为当前点, $\eta \in \mathrm{T}_x\mathcal{M}$ 为一个切向量, $s_0 > 0$ 为初始步长, 且 $\rho, a \in (0, 1)$ 为给定常数. Armijo 回溯线搜索的目标是找到最小的非负整数 ℓ, 使得当 $s = \rho^\ell s_0$ 时, 下述条件成立:

$$f(x) - f(\mathrm{R}_x(s\eta)) \geq -sa g_x(\mathrm{grad}_g f(x), \eta). \tag{1-19}$$

在流形优化中, 局部极小点 x^* 处黎曼 Hessian 算子的条件数对于黎曼方法的局部收敛速率起着至关重要的作用; 参见 [71, Theorem 4.5.6] 和 [70, Theorem 4.20]. 黎曼 Hessian 算子 $\mathrm{Hess}_g f(x^*)$ 的条件数定义如下:

$$\kappa_g(\mathrm{Hess}_g f(x^*)) := \frac{\lambda_{\max}(\mathrm{Hess}_g f(x^*))}{\lambda_{\min}(\mathrm{Hess}_g f(x^*))} = \frac{\sup_{\eta \in \mathrm{T}_{x^*}\mathcal{M}} q_{x^*}(\eta)}{\inf_{\eta \in \mathrm{T}_{x^*}\mathcal{M}} q_{x^*}(\eta)}, \tag{1-20}$$

其中 $\lambda_{\min}(\mathrm{Hess}_g f(x^*))$ 和 $\lambda_{\max}(\mathrm{Hess}_g f(x^*))$ 分别表示 $\mathrm{Hess}_g f(x^*)$ 的最小特征值和最大特征值, 并且

$$q_{x^*}(\eta) := \frac{g_{x^*}(\eta, \mathrm{Hess}_g f(x^*)[\eta])}{g_{x^*}(\eta, \eta)} \tag{1-21}$$

表示 Rayleigh 商, 其依赖于度量 g. 随后, 可以仿照 [71, Theorem 4.5.6] 的分析方法, 证明在乘积流形上采用 Armijo 回溯线搜索 (1-19) 的 RGD 方法的局部收敛速率.

定理 1.4. 设 $\{x^{(t)}\}_{t=0}^{\infty}$ 是由算法 *1* 在回溯线搜索 (1-19) 下生成并收敛到某一局部极小点 x^* 的无限序列. 则存在 $T > 0$, 使得对所有 $t > T$, 有

$$\frac{f(x^{(t)}) - f(x^*)}{f(x^{(t-1)}) - f(x^*)} \leq 1 - \min\left\{ 2as_0\lambda_{\min}(\mathrm{Hess}_g f(x^*)), \frac{4a(1-a)\rho}{\kappa_g(\mathrm{Hess}_g f(x^*))} \right\}.$$

值得注意的是, 由式 (1-20) 可知, 不同的度量将导致不同的 $\lambda_{\min}(\mathrm{Hess}_g f(x^*))$ 和 $\kappa_g(\mathrm{Hess}_g f(x^*))$, 从而影响算法的局部收敛速率. 更具体地说, 较小的条件数通常意味着 RGD 方法具有更快的局部收敛速度.

低秩矩阵张量优化已有的流形优化方法 低秩矩阵优化的第一类几何方法是在光滑流形 $\mathbb{R}_r^{m\times n}$ 上最小化目标函数 f, 从而忽略秩亏矩阵集合 $\mathbb{R}_{\leq r}^{m\times n}\backslash\mathbb{R}_r^{m\times n}$, 即求解问题

$$\begin{aligned} \min \quad & f(\mathbf{X}), \\ \text{s.t.} \quad & \mathbf{X} \in \mathbb{R}_r^{m\times n}. \end{aligned}$$

这个问题可以采用流形优化方法求解; 相关工作可参见, 例如 [14, 73]. 然而, 由于 $\mathbb{R}_r^{m\times n}$ 不是闭集, 当算法收敛到边界 $\mathbb{R}_{\leq r}^{m\times n}\backslash\mathbb{R}_r^{m\times n}$ 上的点时, 经典流形优化理论中建立的收敛性结果 (例如 [74]) 将不再适用.

由于张量有多种不同的分解形式, 以及相应的几何错综复杂, 低秩张量优化问题相比于矩阵情形更具有挑战. 针对 Tucker 分解, 由于固定秩的 Tucker 张量构成的集合 $\mathcal{M}_{\mathbf{r}}$ 是 $\mathbb{R}^{n_1\times n_2\times\cdots\times n_d}$ 的嵌入子流形 [57], 低秩优化问题可以类似表述为在固定 Tucker 秩的张量流形上进行优化. 例如, Koch 和 Lubich [64] 给出了 $\mathcal{M}_{\mathbf{r}}$ 的切空间元素的显式表达式. Kressner 等人 [15] 提出了黎曼共轭梯度法 (GeomCG). 针对分层 Tucker 分解, Uschmajew 和 Vandereycken [57] 证明了固定秩的分层 Tucker 张量构成的集合同样是一个流形, 并建立了其微分几何理论. Da Silva 等人 [75] 利用了分层 Tucker 张量的微分几何, 设计了流形优化方法并应用于张量补全问题. 针对张量链分解, Holtz 等人 [69] 证明了固定秩的 TT 张量所构成的集合是一个流形, 并给出了切空间的显式参数化. Steinlechner [60] 基于此提出了黎曼共轭梯度法, 并应用于张量补全问题. 针对求解低秩矩阵与张量优化问题的几何优化方法, 感兴趣的读者可以参见综述 [23]. 然而, 与矩阵情形类似的是, 固定秩张量流形同样并非闭集.

1.4.2 代数簇上的优化方法

不同于在固定秩流形上进行优化, 另一类方法是直接在集合 $\mathcal{M}_{\leq\mathbf{r}}$ 上最小化目标函数 f, 即

$$\begin{aligned} \min \quad & f(\mathcal{X}), \\ \text{s.t.} \quad & \mathcal{X} \in \mathcal{M}_{\leq\mathbf{r}}. \end{aligned} \tag{1-22}$$

给定 $\mathbb{R}^{n_1\times n_2\times\cdots\times n_d}$ 的一个非空子集 C, C 在点 $\mathcal{X}\in C$ 处的 Bouligand 切锥 (tangent cone) 定义为

$$\mathrm{T}_{\mathcal{X}}C := \{\mathcal{V}\in\mathbb{R}^{n_1\times n_2\times\cdots\times n_d} : \exists t^{(i)}\to 0,\ C\text{中的点列}\mathcal{X}^{(i)}\to\mathcal{X},\ \text{s.t.}\ \frac{\mathcal{X}^{(i)}-\mathcal{X}}{t^{(i)}}\to\mathcal{V}\}.$$

称集合 $\mathrm{N}_{\mathcal{X}}C = (\mathrm{T}_{\mathcal{X}}C)^\circ := \{\mathcal{N}\in\mathbb{R}^{n_1\times n_2\times\cdots\times n_d} : \langle\mathcal{N},\mathcal{V}\rangle\le 0 \text{ 对所有的 } \mathcal{V}\in\mathrm{T}_{\mathcal{X}}C\}$ 为 C 在点 $\mathcal{X}$ 处的法锥. 注意, 当 C 为一个流形时, 其切锥 $\mathrm{T}_{\mathcal{X}}C$(法锥 $\mathrm{N}_{\mathcal{X}}C$) 分别退化为 C 在 $\mathcal{X}$ 处的切空间和法空间. 若映射 $\mathrm{R}:\bigcup_{\mathcal{X}\in C}\{\mathcal{X}\}\times\mathrm{T}_{\mathcal{X}}C\to C$ 若对任意 $\mathcal{X}\in C$ 和 $\mathcal{V}\in\mathrm{T}_{\mathcal{X}}C$ 满足 $\lim_{t\to 0^+}(\mathrm{R}_{\mathcal{X}}(t\mathcal{V})-\mathcal{X}-t\mathcal{V})/t=0$ 则称其为一个收缩映射 [76, §3.1.2]. 接下来我们介绍关于问题 (1-22)的一阶稳定点的定义.

定义 1.11. 令 $C=\mathcal{M}_{\le\mathbf{r}}$, 若对所有 $\mathcal{V}\in\mathrm{T}_{\mathcal{X}}\mathcal{M}_{\le\mathbf{r}}$ 都有 $\langle\nabla f(\mathcal{X}),\mathcal{V}\rangle\ge 0$ 则称点 $\mathcal{X}\in\mathcal{M}_{\le\mathbf{r}}$ 是优化问题 (1-22) 的一个稳定点. 该条件等价于 $-\nabla f(\mathcal{X})\in\mathrm{N}_{\mathcal{X}}\mathcal{M}_{\le\mathbf{r}}$ 或等价地, 投影负梯度满足 $\|\mathrm{P}_{\mathrm{T}_{\mathcal{X}}\mathcal{M}_{\le\mathbf{r}}}(-\nabla f(\mathcal{X}))\|_{\mathrm{F}}=0$.

接下来, 我们介绍求解问题 (1-22)的已有方法. 针对矩阵情形, Jain 等人 [77] 提出了用于低秩矩阵回归的迭代硬阈值方法. Schneider 和 Uschmajew [63] 考虑在 $\mathbb{R}_r^{m\times n}$ 的闭包上最小化目标函数 f, 即在矩阵代数簇 $\mathbb{R}_{\le r}^{m\times n}=\{\mathbf{X}\in\mathbb{R}^{m\times n} : \mathrm{rank}(\mathbf{X})\le r\}$ 上进行优化, 并提出了投影梯度法. 在该方法中, 第 t 次迭代以步长 $s^{(t)}$ 按如下方式更新:

$$\mathbf{X}^{(t+1)}=\mathrm{P}_{\le r}\left(\mathbf{X}^{(t)}+s^{(t)}\,\mathrm{P}_{\mathrm{T}_{\mathbf{X}^{(t)}}\mathbb{R}_{\le r}^{m\times n}}(-\nabla f(\mathbf{X}^{(t)}))\right),$$

该更新过程涉及两次度量投影, 分别投影到矩阵代数簇以及切锥上. 投影梯度法的收敛性是在假设目标函数 f 满足 Łojasiewicz 不等式的条件下证明的. Zhou 等人 [78] 在 $\mathbb{R}_{\le r}^{m\times n}$ 上设计了一种黎曼秩自适应方法, 并证明了其收敛性. 近期, Hosseini 和 Uschmajew [76] 提出了一种用于一般实代数簇优化的梯度采样方法. Gao 和 Absil [79] 利用切锥的几何刻画, 发展了一种用于低秩矩阵补全的黎曼秩自适应方法, 该方法在低秩半定规划中同样表现出良好的效果 [80]. 更近期地, Olikier 和 Absil [81] 在投影梯度法的基础上引入了一系列秩减机制, 提出了一种一阶算法, 并证明其任意聚点都是稳定点. 此外, Olikier 等人 [82] 进一步构建了一个用于一般具有 Whitney 分层矩阵集合的一阶优化框架.

针对张量链分解, Kutschan [83] 研究了张量链分解代数簇 (即有界 TT 秩张量构成的集合) 的切锥显式表达式. Vermeylen 等人 [84, 85] 将其应用于设计基于张量链分解的秩自适应算法. 而针对张量 Tucker 分解, 这方面的工作是非常稀少的, Luo 和 Qi [86] 研究了问题 (1-22)的稳定性条件, 如何在有界秩 Tucker 张量集合上设计算法是一个有趣但具有挑战的问题. 此外, 在张量代数簇上进行优化的一大困难在于其内在的非光滑性, 这给收敛性分析带来了显著挑战. 具体而言, 当算法的聚点是秩亏点, 即属于 $\mathbb{R}_{\le\mathbf{r}}^{n_1\times n_2\times\cdots\times n_d}\backslash\mathbb{R}_{\mathbf{r}}^{n_1\times n_2\times\cdots\times n_d}$ 时, 已建立的张量代数簇优化中的局部收敛性结果将不再成立; 参见 [63, 87].

1.4.3 参数化方法

第三类方法基于参数化思想, 其目标是规避代数簇所固有的非光滑性. 例如, 可以通过一个流形 $\mathcal{M}$ 及一个光滑映射 $\varphi : \mathcal{M} \to \mathcal{M}_{\leq \mathbf{r}}$ 对可行集 $\mathcal{M}_{\leq \mathbf{r}}$ 进行参数化, 使得 $\varphi(\mathcal{M}) = \mathcal{M}_{\leq \mathbf{r}}$. 由此, 在 $\mathcal{M}_{\leq \mathbf{r}}$ 上最小化 f 的问题可等价转化为在流形 $\mathcal{M}$ 上最小化复合函数 $f \circ \varphi$, 即求解

$$\min_{x \in \mathcal{M}} g(x) = f(\varphi(x)). \tag{1-23}$$

光滑流形结构使得我们可以对问题 (1-23) 直接应用流形优化方法求解 (相关综述可参见 [70, 71]). 在理论层面, 若 $x \in \mathcal{M}$ 是 $f \circ \varphi$ 的一个二阶稳定点, 则 $\varphi(x) \in \mathbb{R}^{m \times n}_{\leq r}$ 是 f 的一阶稳定点; 详见 [88]. 在同一思路下, Levin 等人 [89] 提出了一个用于在可行集 $\mathbb{R}^{m \times r} \times \mathbb{R}^{n \times r}$ 上最小化 $f \circ \varphi$ 的黎曼信赖域方法, 称为“RTR-LR”. 近年来, 一种称为去奇异化的参数化方法被提出并研究 (见 [90–92]). 该参数化方法引入了一个定义在 Grassmann 流形 $\mathrm{Gr}(n-r, n)$ 上的松弛变量 $\mathbf{P}$, 用于刻画矩阵 $\mathbf{X}$ 的零空间, 从而构造出如下光滑流形:

$$\mathcal{M}(\mathbb{R}^{m \times n}, r) = \{(\mathbf{X}, \mathbf{P}) : \mathbf{XP} = 0, \mathbf{X} \in \mathbb{R}^{m \times n}, \mathbf{P} \in \mathrm{Gr}(n-r, n)\}$$

并定义映射 $\varphi : (\mathbf{X}, \mathbf{P}) \mapsto \mathbf{X}$. Rebjock 和 Boumal [92] 在该流形上采用流形优化方法, 并给出了全局与局部收敛性保证. Yang 等人 [93] 进一步提出了一个基于去奇异化的框架, 用于在带有正交不变约束的情形下对 $\mathbb{R}^{m \times n}_{\leq r}$ 进行优化.

我们同样可以通过参数化张量代数簇求解低秩张量优化问题 (1-22). 在张量优化问题中, 由于 CP 张量相关的低秩逼近等优化问题是 NP 难的 [33], 并且固定秩的 CP 张量不再具有流形结构 [94], 一个退而求其次的方法是通过秩 1 张量构成的流形 (该流形也被称为 Segre 流形) 的乘积, 参数化固定秩的 CP 张量, 并在该乘积流形上极小化目标函数. Swijsen 等人 [95] 将这种思想应用于求解张量补全问题. 针对 Tucker 分解, $\mathcal{M}_{\leq \mathbf{r}}$ 中的张量可以通过 Tucker 分解参数化为

$$\mathcal{M}^{\text{Tucker}} = \mathbb{R}^{r_1 \times r_2 \times \cdots \times r_d} \times \mathbb{R}^{n_1 \times r_1} \times \mathbb{R}^{n_2 \times r_2} \times \cdots \times \mathbb{R}^{n_d \times r_d}.$$

Kasai 和 Mishra [96] 通过构造如下商流形, 对固定秩流形 $\mathcal{M}_{\mathbf{r}}$ 进行了参数化:

$$\mathcal{M}^{\text{quotient}} = \mathbb{R}^{r_1 \times \cdots \times r_d} \times \mathrm{St}(r_1, n_1) \times \cdots \times \mathrm{St}(r_d, n_d)/(\mathcal{O}(r_1) \times \cdots \times \mathcal{O}(r_d)).$$

近年来, 有界秩矩阵的去奇异化为非光滑代数簇提供了一种光滑参数化方式, 从而显著促进了流形优化方法的应用. 这一思想也启发了探索面向有界秩张量的去奇异化方法; 详见 [97].

1.4.4 低秩优化中的其他类型的方法

低秩优化中有一类非常重要的问题: 低秩恢复. 给定矩阵 $\mathbf{X}^*$ 上的观测 $\mathscr{A}(\mathbf{X}^*)$, 低秩恢复的目标是通过观测恢复完整的矩阵 $\mathbf{X}^*$, 这里 $\mathscr{A} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$ 是一个

线性映射. 事实上, 低秩矩阵张量补全是低秩恢复的一个特例. 低秩恢复问题中, 一类很重要的方法是秩最小化方法, 即在约束 $\mathscr{A}(\mathbf{X}) = \mathscr{A}(\mathbf{X}^*)$ 下最小化矩阵 $\mathbf{X}$ 的秩, 以恢复 $\mathbf{X}^*$. 该问题在一般情形下是 NP-hard 的 [98]. 作为替代方案, Fazel [99] 提出了核范数最小化方法, 即最小化 $\mathbf{X}$ 的奇异值之和. 核范数是矩阵秩的一个凸松弛, 因此核范数最小化问题是一个凸优化问题. Recht 等人 [100] 证明了在满足有限等距性质的假设下, 核范数最小化能够恢复低秩矩阵; 该性质最初由 Candès 和 Tao [101] 在压缩感知中提出. Candès、Recht 和 Tao [102–104] 给出了 $\mathbf{X} = \mathbf{X}^*$ 为核范数最小化问题唯一解的充分条件, 并证明当从 $[n_1] \times [n_2]$ 中以均匀随机方式选取足够多的索引集 Ω 时 (即矩阵补全情形), 这些条件以高概率成立. 近期, Ding 和 Chen [105] 利用留一法分析对矩阵补全中所需采样索引数量 $|\Omega|$ 给出了更优的估计. 需要指出的是, 核范数最小化通常要求完整存储矩阵 $\mathbf{X}$. 针对基于 Tucker 秩的核范数最小化问题, Gandy 等人 [106] 提出了交替方向乘子法. Yuan 和 Zhang [107] 为此类核范数最小化问题建立了理论保证. Barak 和 Moitra [108] 研究了含噪声的低秩张量补全问题, 并提出了一种基于平方和分层的算法. 需要注意的是, 核范数最小化在张量情形下同样需要存储完整大小的张量.

此外, 近期有许多的工作研究在其他额外约束条件下研究了低秩矩阵的几何结构. 例如, Cason 等人 [109] 研究了具有单位 Frobenius 范数约束的低秩矩阵集合的几何性质; Rakhuba 和 Oseledets [110] 考虑了在带有单位范数约束的固定秩矩阵集合上计算最小特征值的问题. 针对同时计算多个特征值的情形, Krumnow 等人 [111] 在 Stiefel 流形与低秩矩阵集合的交集上提出了一种迹最小化方法. 近期, Yang 等人 [93] 分析了在正交不变约束下低秩矩阵代数簇的几何结构.

1.5　本文主要内容

本文主要从理论、算法与应用三个角度, 系统地研究了低秩优化问题. 本文剩余部分的内容安排如下所示.

在第二章中, 我们研究乘积流形优化问题, 旨在探索度量选择对流形优化方法性能的影响, 并探索如何通过精心设计度量以加速算法收敛. 我们为此提出了一个黎曼预条件优化框架. 在该框架中, 我们引入由算子定义的预条件度量, 该算子目标在于逼近目标函数黎曼 Hessian 算子的对角块, 从而改善问题的条件数. 围绕这一思想, 我们提出了三类预条件算子的构造方式: 精确块对角预条件、左右预条件以及高斯–牛顿型预条件. 进一步地, 针对典型相关分析和截断奇异值分解问题, 我们构造了与问题结构相匹配的新的预条件黎曼度量, 并在理论上证明了所提出方法在这些问题上的加速效果. 我们将高斯–牛顿型预条件引入到张量环补全问题中, 并给出了一种高效的流形优化算法. 通过在上述应用中的系统数值实验, 本文验证了所提出方法的有效性, 实验结果表明: 精细地设计黎曼度量的确能够显著提升流形优化方法的性能.

在第三章中, 我们提出了一类基于张量环分解的张量补全问题的黎曼预条件算法. 在张量环分解中, 我们在由各核张量的模态 2 展平矩阵所构成的乘积空间

上构造了一种新的黎曼度量. 该度量的构造目标是通过其对角块来近似目标函数的 Hessian, 并提出了预条件度量下的黎曼梯度法和黎曼共轭梯度法. 我们证明了这两种算法均全局收敛到一个稳定点. 在算法实现方面, 我们充分利用了张量结构, 采用了一种经济的计算策略, 避免了梯度计算过程中大规模矩阵的显式构造与运算, 从而显著降低了计算成本. 在人工合成数据集和多种真实数据集 (包括电影评分数据、高光谱图像以及高维函数) 上的数值实验结果表明, 所提出的算法在性能上优于现有方法.

在第四章中, 我们对 Tucker 张量代数簇 (即有界秩 Tucker 张量构成的集合) 展开系统研究. 相比于已被充分研究的矩阵代数簇的几何结构, Tucker 张量代数簇的几何性质是更加复杂的. 我们给出了 Tucker 张量代数簇切锥的显式参数化表示, 并利用其几何结构, 针对在 Tucker 张量代数簇上的优化问题, 提出了具有理论收敛性保证的、基于梯度相关搜索方向的线搜索方法. 我们通过将负梯度近似投影到切锥得到搜索方向, 避免了计算无显式表达的度量投影. 在实际应用中, 低秩张量优化普遍面临秩参数难以可靠选取的问题. 为此, 我们结合所得到的几何结构, 提出了一种 Tucker 秩自适应方法, 该方法能够在迭代过程中自动识别合适的秩参数, 同时仍然有收敛性保证. 在合成数据集和真实数据集上的张量补全问题的数值实验结果表明, 所提出的方法在恢复性能方面优于其他代表性的方法. 此外, 秩自适应方法在不同秩参数设定下均表现出最优性能, 并且确实能够识别出合适的 Tucker 秩参数.

在第五章中, 我们研究了单位 Frobenius 范数的低秩张量的几何与应用. 事实上, 这类型的张量是科学计算和量子物理等领域中的基础研究对象, 它们能够用于表示归一化的特征向量以及纯量子态. 尽管张量链分解为处理高维问题提供了一种强有力的低秩表示格式, 但该分解形式本身并不能内在地保证单位范数约束. 为此, 我们引入了归一化张量链分解 (NTT), 其目标是在张量列格式下, 用单位范数张量来逼近给定的张量. NTT 分解的低秩结构不仅能够节省存储和计算开销, 同时还能够保持张量所固有的单位范数结构. 本文证明了固定秩 NTT 张量的集合构成一个光滑流形, 并推导了其对应的几何结构, 从而为几何优化方法的提出奠定了理论基础. 在此基础上, 我们将其应用于低秩张量恢复问题、高维特征值问题、稳定子秩的估计以及量子信道最小输出 Rényi-2 熵的计算. 数值实验结果表明, 所提出的基于 NTT 的方法在计算效率和可扩展性方面均表现出显著优势.

最后一章总结了全文的主要内容, 并探讨了在后续工作中值得继续研究的问题.

第 2 章　乘积流形上的预条件方法

2.1　引言

我们考虑定义在乘积流形上的优化问题:

$$\min_{x\in\mathcal{M}} f(x), \tag{2-1}$$

其中 f 是一个光滑的目标函数, 搜索空间 $\mathcal{M}$ 是一个乘积流形, 即

$$\mathcal{M} := \mathcal{M}_1 \times \mathcal{M}_2 \times \cdots \times \mathcal{M}_K,$$

其中 $\mathcal{M}_k$ 是一个光滑流形 ($k = 1, 2, \ldots, K$), K 为正整数. 乘积流形上的优化在诸多问题中具有广泛应用, 包括奇异值分解 [112]、联合近似张量对角化问题 [113]、脑电信号 (EEG) 协方差矩阵的降维 [114], 以及典型相关分析 [115]. 此外, 相较于直接在完整规模的矩阵或张量上进行优化, 矩阵与张量分解能将矩阵或张量分解为若干较小的块, 使得我们能够利用乘积流形上的优化问题 (2-1)建模低秩优化问题 [17, 95, 96, 116–118].

相关工作与研究动机　流形优化通过利用黎曼流形 $\mathcal{M}$ 的几何结构来设计算法, 已在诸多领域展现出应用潜力. 针对问题 (2-1), 可以构造多种流形优化方法, 例如黎曼梯度法和黎曼共轭梯度法. 流形优化的综述可参见文献 [70, 71].

由于不同的度量会诱导不同的黎曼梯度, 从而产生不同的流形优化方法, 人们自然会关心: 流形优化方法的性能在多大程度上依赖于度量 g 的选择? 此外, 在局部极小点 x^* 处, 目标函数的黎曼 Hessian 算子的条件数 $\kappa := \kappa_g(\mathrm{Hess}_g f(x^*))$ 会影响流形优化中梯度类方法的局部收敛性质. 例如, 在欧几里得情形 (即 $\mathcal{M} = \mathbb{R}^n$) 下, 针对对称正定线性系统, 最速下降法和共轭梯度法的渐近局部线性收敛因子分别为 $(\kappa-1)/(\kappa+1)$ 和 $(\sqrt{\kappa}-1)/(\sqrt{\kappa}+1)$, 见 [4, Theorem 3.3, Theorem 5.5]. 在一般情形下, 黎曼梯度法的渐近局部线性收敛率已被证明为 $1 - 1/\mathcal{O}(\kappa)$, 参见 [119, Chapter 7, Theorem 4.2]、[71, Theorem 4.5.6] 以及 [70, Theorem 4.20]. 需要注意的是, 选择合适的黎曼度量 g 可以显著降低条件数. 基于上述观察, 一个自然的问题是:

是否可以通过 “精细” 选择黎曼度量, 从而加速流形优化方法?

下面的例子给出了一个肯定的回答.

例 2.1. 考虑如下优化问题:

$$\min\ f(\mathbf{x}) := -\mathbf{b}^\top\mathbf{x}, \quad \text{s.t. } \mathbf{x} \in \mathcal{M}_{\mathbf{B}} := \{\mathbf{x} \in \mathbb{R}^n :\ \mathbf{x}^\top\mathbf{B}\mathbf{x} = 1\},$$

其中 $\mathbf{B} \in \mathbb{R}^{n\times n}$ 为对称正定矩阵, $\mathbf{b} \in \mathbb{R}^n$. 约束集 $\mathcal{M}_{\mathbf{B}}$ 是一个椭球面. 该问题具有显式解 $\mathbf{x}^* = \mathbf{B}^{-1}\mathbf{b}/\|\mathbf{B}^{-1}\mathbf{b}\|_{\mathbf{B}}$, 其中 $\|\mathbf{x}\|_{\mathbf{B}}^2 := \mathbf{x}^\top\mathbf{B}\mathbf{x}$. 我们考察一族黎曼度量

$$g_{\lambda,\mathbf{x}}(\xi,\eta) := \langle \xi, (\lambda\mathbf{I}_n + (1-\lambda)\mathbf{B})\eta\rangle,$$

其中 ξ 和 η 为切向量, $\lambda \in \mathbb{R}$ 使得 $\lambda\mathbf{I}_n + (1-\lambda)\mathbf{B}$ 为正定矩阵. 在该族度量下, f 在 $\mathbf{x} \in \mathcal{M}_{\mathbf{B}}$ 处的黎曼梯度为

$$\operatorname{grad}_{g_\lambda} f(\mathbf{x}) = -(\mathbf{I}_n - \frac{\mathbf{B}_\lambda^{-1}\mathbf{B}\mathbf{x}\mathbf{x}^\top\mathbf{B}}{\mathbf{x}^\top\mathbf{B}\mathbf{B}_\lambda^{-1}\mathbf{B}\mathbf{x}})\mathbf{B}_\lambda^{-1}\mathbf{b} = -\mathbf{B}_\lambda^{-1}\mathbf{b} + \frac{\mathbf{x}^\top\mathbf{B}\mathbf{B}_\lambda^{-1}\mathbf{b}}{\mathbf{x}^\top\mathbf{B}\mathbf{B}_\lambda^{-1}\mathbf{B}\mathbf{x}}\mathbf{B}_\lambda^{-1}\mathbf{B}\mathbf{x},$$

这是通过命题 2.4 得出的. 随后, 在度量 g_λ 下, RGD 的更新公式为 $\mathbf{x}^{(t+1)} = \bar{\mathbf{x}}^{(t)}/\|\bar{\mathbf{x}}^{(t)}\|_{\mathbf{B}}$, 其中 $\bar{\mathbf{x}}^{(t)} = \mathbf{x}^{(t)} - s^{(t)}\operatorname{grad}_{g_\lambda} f(\mathbf{x}^{(t)})$, 此处我们采用了极分解作为收缩映射 [115, (3.3)]. f 在 $\mathbf{x}^*$ 处沿 $\eta \in \mathrm{T}_{\mathbf{x}^*}\mathcal{M}_{\mathbf{B}}$ 方向的黎曼 Hessian 为:

$$\operatorname{Hess}_{g_\lambda} f(\mathbf{x}^*)[\eta] = \Pi_{g_\lambda,\mathbf{x}^*}(\operatorname{Dgrad}_{g_\lambda} f(\mathbf{x}^*)[\eta]) = \Pi_{g_\lambda,\mathbf{x}^*}(\|\mathbf{B}^{-1}\mathbf{b}\|_{\mathbf{B}}\mathbf{B}_\lambda{}^{-1}\mathbf{B}\eta),$$

这是因为 $\operatorname{grad}_{g_\lambda} f(\mathbf{x}^*) = \Pi_{g_\lambda,\mathbf{x}^*}(-\mathbf{B}_\lambda^{-1}\mathbf{b}) = 0$. 于是 Rayleigh 商 (1-21) 由下式给出:

$$q(\eta) = \|\mathbf{B}^{-1}\mathbf{b}\|_{\mathbf{B}} \cdot \frac{\langle \eta, \mathbf{B}\eta\rangle}{\langle \eta, \mathbf{B}_\lambda\eta\rangle} \quad \text{对于} \eta \in \mathrm{T}_{\mathbf{x}^*}\mathcal{M}_{\mathbf{B}}.$$

因此, 我们可以按照与命题 2.5 相同的方式计算 $\operatorname{Hess}_{g_\lambda} f(\mathbf{x}^*)$ 的条件数. 注意到如果 $\lambda = 0$, Rayleigh 商将退化为常数 $\|\mathbf{B}^{-1}\mathbf{b}\|_{\mathbf{B}}$, 因此 $\kappa_{g_0}(\operatorname{Hess}_{g_0} f(\mathbf{x}^*)) = 1$.

图 2-1 展示了该度量族对黎曼梯度法 (RGD) 以及 $\operatorname{Hess}_{g_\lambda} f(\mathbf{x}^*)$ 条件数的影响. 左图比较了在欧氏度量 $g_{1,\mathbf{x}}(\xi,\eta) = \langle \xi, \eta\rangle$ 与非欧氏度量 $g_{0,\mathbf{x}}(\xi,\eta) = \langle \xi, \mathbf{B}\eta\rangle$ 下 RGD 生成的迭代序列, 可以观察到在度量 g_0 下算法具有更快的收敛速度. 此外, 右图表明条件数随度量的选择而变化, 且度量 g_0 对应的条件数最小.

我们来梳理通过构造合适度量以提升流形优化方法性能的已有工作. 例如, Mishra 和 Sepulchre 在 [120] 中提出了黎曼预条件方法, 用于求解可行集具有流形结构的等式约束优化问题. 该工作中采用的非欧氏度量来源于拉格朗日函数的欧氏 Hessian 算子, 然而在实际应用中, 欧氏 Hessian 算子的显式构造往往是代价高昂的. 作为一种替代方案, 在矩阵与张量补全问题中, 研究者采用块对角近似来构造黎曼度量 [17, 96, 117, 118, 121]. 具体而言, 利用张量分解所固有的块结构, 通过提取目标函数 Hessian 算子的对角块来构造度量, 并证明了在这些度量下的流形优化方法具有良好的计算效率. 近年来, Shustin 与 Avron [115, 122] 进一步提出了一种用于广义 Stiefel 流形的预条件度量, 该度量通过利用局部极小点处的黎曼 Hessian 算子来构造.

此外, 还有一些工作通过其他方式将预条件思想引入流形优化中. Boumal 与 Absil [116] 在矩阵补全问题中构造了用于近似黎曼 Hessian 算子的预条件算子. Kressner 等人 [18] 通过构造类拉普拉斯算子, 提出了用于求解张量方程的预条件 Richardson 迭代和近似 Newton 方法. 近期, Tong 等人 [123] 提出了用于低秩矩阵 ScaledGD 方法. Bian 等人 [124] 提出了一种用于低秩矩阵恢复的预条件黎曼梯度法. Hamed 与 Hosseini [125] 则在新的黎曼度量下, 提出了一种用于低多线性秩逼近的黎曼坐标下降方法.

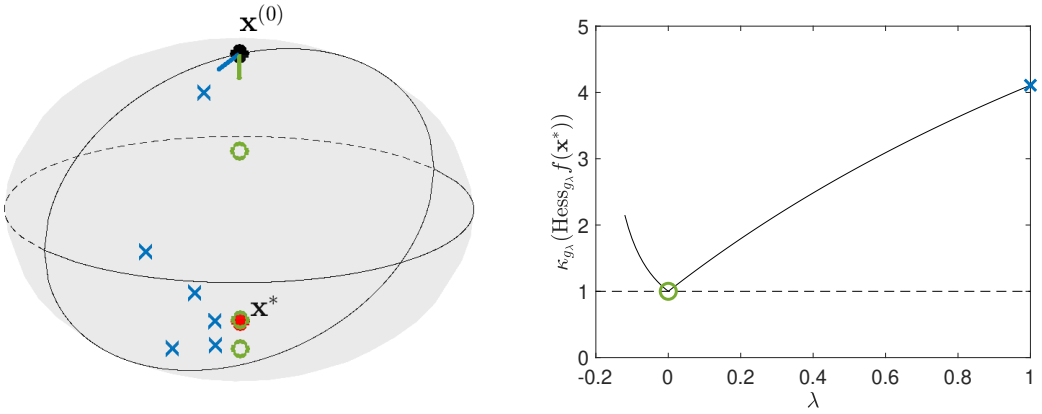
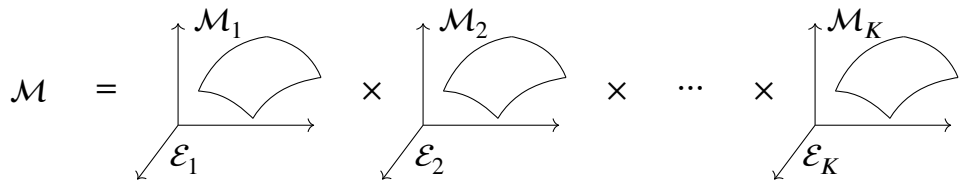


图 2-1 左图: 在 $\mathbf{B} = \mathrm{diag}(2^2, 3^2, 1)$ 且 $\mathbf{b} = (1, 1, 1)$ 的情形下, 黎曼梯度法在两种不同度量下生成的迭代序列. 右图: 当 $\lambda \in (-1/8, 1]$ 时, $\mathrm{Hess}_{g_\lambda} f(\mathbf{x}^*)$ 的条件数. 蓝色标记表示欧几里得度量, 绿色标记表示缩放度量.

Figure 2-1 Left: sequences generated by the Riemannian gradient descent method under two metrics for $\mathbf{B} = \mathrm{diag}(2^2, 3^2, 1)$ and $\mathbf{b} = (1, 1, 1)$. Right: the condition number of $\mathrm{Hess}_{g_\lambda} f(\mathbf{x}^*)$ for $\lambda \in (-1/8, 1]$. Blue marker: the Euclidean metric; green marker: the scaled metric.

本章主要内容　我们针对乘积流形上的优化问题 (2-1), 系统研究了通过构造黎曼预条件度量来提升流形优化方法性能的理论与方法. 首先, 针对乘积流形 $\mathcal{M} = \mathcal{M}_1 \times \mathcal{M}_2 \times \cdots \times \mathcal{M}_K$, 我们提出了一种统一的预条件度量构造框架, 即通过在切空间上设计自伴随、正定的线性算子 $\bar{\mathcal{H}}(x)$, 使所构造的黎曼度量能够有效近似目标函数的二阶信息, 从而改善黎曼 Hessian 算子的条件数并加速算法收敛. 图 2-2展示了我们如何在乘积流形上构造黎曼度量.



$$
\begin{aligned}
g_x(\xi, \eta) &= g^1_{x_1}(\xi_1, \eta_1) + g^2_{x_2}(\xi_2, \eta_2) + \cdots + g^K_{x_K}(\xi_K, \eta_K) \\
&= \langle \xi_1, \bar{\mathcal{H}}_1(x)[\eta_1] \rangle + \langle \xi_2, \bar{\mathcal{H}}_2(x)[\eta_2] \rangle + \cdots + \langle \xi_K, \bar{\mathcal{H}}_K(x)[\eta_K] \rangle
\end{aligned}
$$

图 2-2 乘积流形 $\mathcal{M}$ 上的预条件度量的构造.

Figure 2-2 A new metric on the product manifold $\mathcal{M}$.

具体而言, 本章提出了三类预条件策略: 1) 精确块对角预条件方法, 利用黎曼 Hessian 算子的对角块构造度量, 避免显式计算完整 Hessian 算子；2) 左右预条件方法, 在对角块不正定的情形下, 通过构造正定算子对线性项进行近似；3) 高斯–牛顿型预条件方法, 面向最小二乘问题, 基于一阶导数构造相应的黎曼度量. 基于上述预条件度量, 我们构造并分析了黎曼梯度法和黎曼共轭梯度法, 给出了与条件数相关的收敛性结果, 并表明该框架能够统一解释多种已有的黎曼预条件

方法. 表2-1 展示了如何用我们的框架解释已有的预条件方法.

表 2-1 已有的能被预条件度量框架解释的工作. “∗”: 非奇异矩阵与张量; RGN: 黎曼高斯牛顿法; CCA: 典型相关分析.

Table 2-1 Existing and our works interpreted by preconditioned metrics. “∗”: non-sigular matrices or tensors; RGN: Riemannian Gauss–Newton; CCA: canonical correlation analysis.

问题	方法	搜索空间 $\mathcal{M}$	类别
矩阵补全 [121]	RGD, RCG, RTR	$\mathbb{R}_*^{m\times r}\times\mathbb{R}_*^{n\times r}$	精确块对角
矩阵回归 [123]	ScaledGD	$\mathbb{R}_*^{m\times r}\times\mathbb{R}_*^{n\times r}$	精确块对角
Tucker 张量补全 [120]	RCG	$\times_{k=1}^{3}\mathrm{St}(r_k,n_k)\times\mathbb{R}^{r_1\times r_2\times r_3}$	精确块对角
CP 张量补全 [117]	RGD, RCG	$\times_{k=1}^{K}\mathbb{R}^{n_k\times r}$	精确块对角
TT 张量补全 [118]	RGD, RCG, RGN	$\times_{k=1}^{K}\mathbb{R}_*^{r_{k-1}\times n_k\times r_k}$	精确块对角
TR 张量补全 [17]	RGD, RCG	$\times_{k=1}^{K}\mathbb{R}^{n_k\times r_{k-1}r_k}$	精确块对角
CCA [115, 126]	RCG	$\mathrm{St}_{\Sigma_{xx}}(m,d_x)\times\mathrm{St}_{\Sigma_{yy}}(m,d_y)$	左右预条件
CCA (本章)	RGD, RCG	$\mathrm{St}_{\Sigma_{xx}}(m,d_x)\times\mathrm{St}_{\Sigma_{yy}}(m,d_y)$	左右预条件
SVD (本章)	RGD, RCG	$\mathrm{St}(p,m)\times\mathrm{St}(p,n)$	左右预条件
TR 张量补全 (本章)	高斯牛顿法	$\times_{k=1}^{K}\mathbb{R}^{n_k\times r_{k-1}r_k}$	高斯–牛顿型

在应用方面, 我们将所提出的预条件框架应用于典型相关分析、截断奇异值分解以及张量环补全问题. 在这些问题中, 我们构造了新的黎曼预条件度量, 并计算了相应问题在局部极小点处黎曼 Hessian 算子的条件数. 数值实验结果表明, 所提出的预条件度量能够显著改善条件数, 从而有效加速流形优化算法, 同时保持与现有方法相当的计算复杂度.

2.2 在乘积流形上设计预条件度量

我们首先提出一个在乘积流形 $\mathcal{M}$ 上构造预条件度量的一般框架, 其核心思想是构造一个算子 $\bar{\mathcal{H}}(x)$, 用于近似黎曼 Hessian 算子的对角块. 随后, 我们进一步提出三种具体方法来构造算子 $\bar{\mathcal{H}}(x)$.

一般而言, 我们首先通过在环境空间 $\mathcal{E}$ 上设计一个作用于切丛 $\mathrm{T}\mathcal{E}$ 的自伴且正定的线性算子 $\bar{\mathcal{H}}$, 从而为 $\mathcal{E}$ 赋予一个度量 $\bar{g}$. 该度量的目标是近似目标函数的二阶信息, 即

$$\bar{g}_x(\xi,\eta)=\langle\xi,\bar{\mathcal{H}}(x)[\eta]\rangle\approx\langle\xi,\mathrm{Hess}_{\mathrm{e}}f(x)[\eta]\rangle\quad\text{对所有}\xi,\eta\in\mathrm{T}_x\mathcal{M}. \tag{2-2}$$

随后, 基于黎曼子流形的观点, $\mathcal{M}$ 上的黎曼度量 g 由 $\mathcal{E}$ 上的度量 $\bar{g}$ 自然诱导得到. 由于 $\mathcal{M}=\mathcal{M}_1\times\mathcal{M}_2\times\cdots\times\mathcal{M}_K$ 是一个乘积流形, 根据 [70, Example 5.19], 函数 f 在点 $x=(x_1,x_2,\ldots,x_K)$ 处、沿方向 $\eta=(\eta_1,\eta_2,\ldots,\eta_K)$ 的黎曼 Hessian 算子具有

如下的分块结构:

$$\begin{aligned}\mathrm{Hess}_{\mathrm{e}} f(x)[\eta] = (&H_{11}(x)[\eta_1] + H_{12}(x)[\eta_2] + \cdots + H_{1K}(x)[\eta_K],\\ &H_{21}(x)[\eta_1] + H_{22}(x)[\eta_2] + \cdots + H_{2K}(x)[\eta_K],\\ &\qquad\vdots\\ &H_{K1}(x)[\eta_1] + H_{K2}(x)[\eta_2] + \cdots + H_{KK}(x)[\eta_K]),\end{aligned} \tag{2-3}$$

这里当 $i = j$ 时, $H_{ij}(x)[\eta_j]$ 表示 $f(x_1, \ldots, x_{i-1}, \cdot, x_{i+1}, \ldots, x_K)$ 在点 x_i 处沿方向 η_i 的欧几里得 Hessian 算子, 即 $H_{ij}(x)[\eta_j] := \mathrm{Hess}_{\mathrm{e}} f(x_1, \ldots, x_{i-1}, \cdot, x_{i+1}, \ldots, x_K)(x_i)[\eta_i]$; 当 $i \neq j$ 时, $H_{ij}(x)[\eta_j]$ 表示映射 G_i 关于第 j 个变量的导数作用在 η_j 上, 即 $H_{ij}(x)[\eta_j] := \mathrm{D}G_i(x_1, \ldots, x_{j-1}, \cdot, x_{j+1}, \ldots, x_K)(x_j)[\eta_j]$. $f(x_1, \ldots, x_{i-1}, \cdot, x_{i+1}, \ldots, x_K)$ 表示将函数 f 限制在流形 $\mathcal{M}_i$ 上所得到的函数. 算子 $G_i : \mathcal{M} \to \mathrm{T}_{x_i}\mathcal{M}_i$ 给出了上述限制函数在点 x_i 处的黎曼梯度. 将 G_i 限制在 $\mathcal{M}_j$ 上, 可得到映射 $G_i(x_1, \ldots, x_{j-1}, \cdot, x_{j+1}, \ldots, x_K) : \mathcal{M}_j \to \mathrm{T}_{x_i}\mathcal{M}_i$. 基于黎曼子流形的理论, 函数 f 在 $x \in \mathcal{M}$ 处关于度量 g 的黎曼梯度可以按照 [71, (3.37)] 中的方法进行计算.

命题 2.1. *设 $(\mathcal{M}, g)$ 为 $(\mathcal{E}, \bar{g})$ 的一个嵌入子流形. 给定函数 $f : \mathcal{M} \to \mathbb{R}$ 以及其光滑延拓 $\bar{f} : \mathcal{E} \to \mathbb{R}$, 则函数 f 在 $x \in \mathcal{M}$ 处的黎曼梯度可通过下式计算:*

$$\mathrm{grad}_g f(x) = \Pi_{g,x}(\bar{\mathcal{H}}(x)^{-1}[\nabla \bar{f}(x)]),$$

这里 $\Pi_{g,x} : \mathrm{T}_x\mathcal{E} \simeq \mathcal{E} \to \mathrm{T}_x\mathcal{M}$ 为关于度量 g 的往 $\mathrm{T}_x\mathcal{M}$ 上正交投影算子, 以及 $\nabla \bar{f}(x)$ 为 $\bar{f}$ 的欧氏梯度.

结合式 (2-2) 与命题 2.1 可以看出, 算子 $\bar{\mathcal{H}}(x)$ 在计算黎曼梯度的过程中起到了预条件的作用. 因此, 我们将度量 g 称为预条件度量, 并将 $\bar{\mathcal{H}}$ 称为相应的预条件子. 采用预条件度量的思想可以被视为一种加速流形优化方法的通用框架. 接下来, 我们将基于该框架设计若干构造算子 $\bar{\mathcal{H}}(x)$ 的具体方法.

2.2.1　精确块对角预条件

相比于通过计算式 (2-3) 中的所有分块 $H_{ij}(x)$ 以获取完整的黎曼 Hessian 算子 $\mathrm{Hess}_{\mathrm{e}} f(x)$, 我们希望在算法效率与计算代价之间进行权衡, 为此, 我们仅利用对角块 $H_{11}, H_{22}, \ldots, H_{KK}$ 来构造一种更加经济的度量.

回顾一下, 乘积流形 $\mathcal{M}$ 上的黎曼度量可以表示为各分量流形上度量之和, 即对于 $\xi = (\xi_1, \xi_2, \ldots, \xi_K), \eta = (\eta_1, \eta_2, \ldots, \eta_K) \in \mathrm{T}_x\mathcal{M}$, 有 $g_x(\xi, \eta) = \sum_{k=1}^{K} g^k_{x_k}(\xi_k, \eta_k)$. 设 $\bar{H}_{kk}(x) : \mathrm{T}_{x_k}\mathcal{E}_k \to \mathrm{T}_{x_k}\mathcal{E}_k$ 是对角块 H_{kk} 在环境空间上的一个光滑延拓, 其中 $k = 1, 2, \ldots, K$. 如果 $\bar{H}_{11}, \bar{H}_{22}, \ldots, \bar{H}_{KK}$ 在环境空间 $\mathcal{E}$ 上都是正定的, 则可以直接利用这些对角块来构造算子 $\bar{\mathcal{H}}$, 即

$$\bar{\mathcal{H}}(x)[\eta] = (\bar{\mathcal{H}}_1(x)[\eta_1], \ldots, \bar{\mathcal{H}}_K(x)[\eta_K]) = (\bar{H}_{11}(x)[\eta_1], \ldots, \bar{H}_{KK}(x)[\eta_K]). \tag{2-4}$$

因此, 度量

$$\begin{aligned} g_x(\xi,\eta) &= g^1_{x_1}(\xi_1,\eta_1) + g^2_{x_2}(\xi_2,\eta_2) + \cdots + g^K_{x_K}(\xi_K,\eta_K) \\ &= \langle \xi_1, \bar{H}_{11}(x)[\eta_1]\rangle + \langle \xi_2, \bar{H}_{22}(x)[\eta_2]\rangle + \cdots + \langle \xi_K, \bar{H}_{KK}(x)[\eta_K]\rangle \end{aligned}$$

在 $\mathcal{M}$ 上是良定的黎曼度量, 从而得到所谓的精确块对角 (exact block-diagonal) 预条件方法. 需要指出的是, 当至少有一个对角块 $\bar{H}_{11}, \bar{H}_{22}, \ldots, \bar{H}_{KK}$ 不正定时, 上述精确块对角预条件方法将不再适用. 在实际应用中, 可以通过引入正则化项 $\delta_k \mathcal{I}_k(x)$ 来解决这一问题, 其中 $\mathcal{I}_k(x) : \mathrm{T}_{x_k}\mathcal{E}_k \to \mathrm{T}_{x_k}\mathcal{E}_k$ 为恒等算子, $\delta_k > 0$ 为正则化参数, 从而保证算子 $\bar{H}_{kk}(x) + \delta_k \mathcal{I}_k(x)$ 为正定.

与包含 $\eta_1, \eta_2, \ldots, \eta_K$ 之间交叉项的黎曼 Hessian 算子 (2-3) 不同, 式 (2-4) 中的算子 $\bar{H}(x)$ 具有严格的块对角结构. 因此, 结合命题 2.1, 函数 $f : \mathcal{M} \to \mathbb{R}$ 在点 $x \in \mathcal{M}$ 处的黎曼梯度可以在各个分块上分别计算.

命题 2.2. *设 $\mathcal{M} = \mathcal{M}_1 \times \mathcal{M}_2 \times \cdots \times \mathcal{M}_K$ 是一个具有度量 g 的乘积流形. 给定一个函数 $f : \mathcal{M} \to \mathbb{R}$ 以及其光滑延拓 $\bar{f} : \mathcal{E} \to \mathbb{R}$, f 在 x 处的黎曼梯度为*

$$\begin{aligned} \mathrm{grad}_g f(x) = \big(&\Pi_{g^1,x_1}(\bar{H}_1(x)^{-1}[\partial_1 \bar{f}(x)]), \\ &\Pi_{g^2,x_2}(\bar{H}_2(x)^{-1}[\partial_2 \bar{f}(x)]), \\ &\qquad\vdots \\ &\Pi_{g^K,x_K}(\bar{H}_K(x)^{-1}[\partial_K \bar{f}(x)])\big), \end{aligned} \tag{2-5}$$

这里 Π_{g^k,x_k} 为关于度量 g^k 的往 $\mathrm{T}_{x_k}\mathcal{M}_k$ 上正交投影算子, $k = 1, 2, \ldots, K$, 以及 $\partial_k \bar{f}(x)$ 为函数 f 关于变量 x_k 的部分导数.

值得注意的是, 利用对角块来构造合适的度量与数值线性代数中的块雅克比 (block-Jacobi) 预条件方法 [127] 密切相关. 具体而言, 给定一个对称正定矩阵 $\mathbf{A} \in \mathbb{R}^{n\times n}$, 块雅克比预条件的目标是构造一个可逆的块对角矩阵

$$\mathbf{D} = \begin{bmatrix} \mathbf{D}_{11} & & & \\ & \mathbf{D}_{22} & & \\ & & \ddots & \\ & & & \mathbf{D}_{KK} \end{bmatrix} \in \mathbb{R}^{n\times n},$$

其中 $\mathbf{D}_{kk} \in \mathbb{R}^{n_k \times n_k}$, $k = 1, 2, \ldots, K$, 且满足 $n_1 + n_2 + \cdots + n_K = n$. 该预条件子使得条件数 $\kappa_2(\mathbf{DAD}^\top) := \lambda_{\max}(\mathbf{DAD}^\top)/\lambda_{\min}(\mathbf{DAD}^\top)$ 降低. 现在我们考虑二次函数的最小化问题 $\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) := \frac{1}{2}\mathbf{x}^\top\mathbf{Ax}$. 我们可以在乘积流形 $\mathbb{R}^n = \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \times \cdots \times \mathbb{R}^{n_K}$ 上构造如下预条件度量:

$$g_{\mathbf{x}}(\xi,\eta) := \sum_{k=1}^{K} \langle \xi_k, (\mathbf{D}_{kk}^\top \mathbf{D}_{kk})^{-1}\eta_k\rangle = \langle \xi, (\mathbf{D}^\top\mathbf{D})^{-1}\eta\rangle.$$

给定 $\mathbf{x} \in \mathbb{R}^n$, 由命题 2.2 以及黎曼 Hessian 算子的定义可得 $\mathrm{grad}_g f(\mathbf{x}) = (\mathbf{D}^\top\mathbf{D})\mathbf{A}\mathbf{x}$, $\mathrm{Hess}_g f(\mathbf{x}) = (\mathbf{D}^\top\mathbf{D})\mathbf{A}$. 因此, 其 Rayleigh 商可表示为

$$q_{\mathbf{x}}(\eta) = \frac{g_{\mathbf{x}}(\eta, \mathrm{Hess}_g f(\mathbf{x})[\eta])}{g_{\mathbf{x}}(\eta,\eta)} = \frac{\langle \eta, \mathbf{A}\eta\rangle}{\langle \eta, (\mathbf{D}^\top\mathbf{D})^{-1}\eta\rangle} \xlongequal{\tilde{\eta}:=\mathbf{D}^{-\top}\eta} \frac{\langle \tilde{\eta}, (\mathbf{D}\mathbf{A}\mathbf{D}^\top)\tilde{\eta}\rangle}{\langle \tilde{\eta}, \tilde{\eta}\rangle},$$

其中 $\eta \in \mathrm{T}_{\mathbf{x}}\mathbb{R}^n \simeq \mathbb{R}^n$. 由式 (1-20) 可知

$$\kappa_g(\mathrm{Hess}_g f(\mathbf{x})) = \frac{\sup_{\eta\in\mathrm{T}_{\mathbf{x}}\mathcal{M}} q_{\mathbf{x}}(\eta)}{\inf_{\eta\in\mathrm{T}_{\mathbf{x}}\mathcal{M}} q_{\mathbf{x}}(\eta)} = \frac{\lambda_{\max}(\mathbf{D}\mathbf{A}\mathbf{D}^\top)}{\lambda_{\min}(\mathbf{D}\mathbf{A}\mathbf{D}^\top)} = \kappa_2(\mathbf{D}\mathbf{A}\mathbf{D}^\top).$$

因此, 旨在降低 $\kappa_2(\mathbf{D}\mathbf{A}\mathbf{D}^\top)$ 的块雅克比预条件, 与在 $\mathbb{R}^n$ 上选取合适的预条件度量以降低函数 f 的黎曼 Hessian 算子条件数 $\kappa_g(\mathrm{Hess}_g f(\mathbf{x}))$ 在本质上是等价的. 此外, 矩阵补全与张量补全中的多种预条件策略也可以解释为本文提出的精确块对角预条件方法；具体细节见第 2.5 节.

2.2.2　左右预条件方法

一般而言, 在块对角预条件中, $\bar{H}_{11}, \bar{H}_{22}, \ldots, \bar{H}_{KK}$ 在环境空间 $\mathcal{E}$ 上并不一定是正定的. 因此, 我们需要对这些项构造合适的近似.

具体地, 我们假设 $\mathcal{E}$ 是一个由矩阵构成的直积空间, 并且 $H_{kk}(x)[\eta_k]$ 中包含一个线性项 $\bar{\mathbf{M}}_{k,1}(x)\eta_k\bar{\mathbf{M}}_{k,2}(x)$, 其中 $\bar{\mathbf{M}}_{k,1}(x)$ 和 $\bar{\mathbf{M}}_{k,2}(x)$ 在固定 $x \in \mathcal{E}$ 时为方阵. 结合黎曼 Hessian 算子 (2-3) 的表达式, 我们通过该线性项 $\bar{\mathbf{M}}_{k,1}(x)\eta_k\bar{\mathbf{M}}_{k,2}(x)$ 来近似对角块 $H_{kk}(x)$, 具体做法是构造一个算子 $\bar{H}_k(x) : \mathrm{T}_{x_k}\mathcal{E}_k \to \mathrm{T}_{x_k}\mathcal{E}_k$, 使其满足

$$\langle \xi_k, \bar{H}_k(x)[\eta_k]\rangle = \langle \xi_k, \mathbf{M}_{k,1}(x)\eta_k\mathbf{M}_{k,2}(x)\rangle \approx \langle \xi_k, H_{kk}(x)[\eta_k]\rangle \quad \text{对于} \xi_k, \eta_k \in \mathrm{T}_{x_k}\mathcal{M}_k,$$

其中 $\mathbf{M}_{k,j}(x) = (\mathrm{sym}(\bar{\mathbf{M}}_{k,j}(x))^2 + \delta\mathbf{I})^{1/2}$, $j = 1, 2$, $\mathrm{sym}(\mathbf{A}) := (\mathbf{A} + \mathbf{A}^\top)/2$, 并且 $\delta > 0$ 用以保证正定性. 在此基础上, 我们在 $\mathrm{T}_x\mathcal{M}$ 上定义如下的左右预条件的度量

$$\begin{aligned} g_x(\xi,\eta) &= g^1_{x_1}(\xi_1,\eta_1) + \cdots + g^K_{x_K}(\xi_K,\eta_K) \\ &= \langle \xi_1, \mathbf{M}_{1,1}(x)\eta_1\mathbf{M}_{1,2}(x)\rangle + \cdots + \langle \xi_K, \mathbf{M}_{K,1}(x)\eta_K\mathbf{M}_{K,2}(x)\rangle. \end{aligned} \tag{2-6}$$

注意到, 对于所有 $x \in \mathcal{E}$, $\mathbf{M}_{k,j}(x)$ 都是光滑且正定的, 因此 (2-6) 给出了一个良定义的黎曼度量. 相应的黎曼梯度也可以通过命题 2.2 来计算, 因为算子 $\bar{H}$ 是逐块定义的.

我们采用上述左右预条件策略来加速典型相关分析 (CCA) 中的流形优化方法, 具体细节见 2.3 节. 在实际应用中, 为了节省计算成本, 也可以只采用左预条件或右预条件. 2.4 节中展示了我们如何为截断奇异值分解 (SVD) 构造右预条件. 值得注意的是, 如果在 (2-6) 中算子 $\mathbf{M}_{k,j}$ 的选取不当, 则在该度量下的黎曼方法甚至可能比在欧几里得度量下表现更差；参见 2.3.4 节. 尽管如此, 针对 CCA 和 SVD 所设计的预条件度量确实能够改善 $\mathrm{Hess} f(x)$ 的条件数, 从而加速流形优化方法；参见命题 2.7.

2.2.3 高斯–牛顿型预条件方法

如果 (2-1) 中的目标函数满足 $f(x) := \frac{1}{2}\|F(x)\|_{\mathrm{F}}^2$, 其中 $F : \mathcal{M} \to \mathbb{R}^n$ 是光滑函数且 $\mathrm{D}F(x)$ 为单射, 则可以考虑算子 $\bar{\mathcal{H}}(x) = (\mathrm{D}F(x))^* \circ \mathrm{D}F(x)$ 以近似 $\mathrm{Hess}_{\mathrm{e}} f(x)$, 并构造预条件度量如下:

$$g_x(\xi, \eta) = \langle \xi, \bar{\mathcal{H}}(x)[\eta] \rangle = \langle \xi, ((\mathrm{D}F(x))^* \circ \mathrm{D}F(x))[\eta] \rangle \approx \langle \xi, \mathrm{Hess}_{\mathrm{e}} f(x)[\eta] \rangle, \tag{2-7}$$

其中 $(\mathrm{D}F(x))^*$ 是 $\mathrm{D}F(x)$ 的伴随算子. 高斯–牛顿型预条件不再是块对角预条件, 因为 $\bar{\mathcal{H}}(x)$ 包含了 $\eta_1, \eta_2, \ldots, \eta_K$ 之间的交叉项. 因此, 黎曼梯度可以直接由命题 2.1 得到.

事实上, 使用度量 g 的黎曼梯度法恰好就是黎曼高斯–牛顿方法 [71, §8.4.1], 其中在 $x^{(t)} \in \mathcal{M}$ 处的搜索方向 $\eta^{(t)} \in \mathrm{T}_{x^{(t)}}\mathcal{M}$ 由以下高斯–牛顿方程计算:

$$\langle \mathrm{D}F(x^{(t)})[\xi], \mathrm{D}F(x^{(t)})[\eta^{(t)}] \rangle + \langle \mathrm{D}F(x^{(t)})[\xi], F(x^{(t)}) \rangle = 0 \quad \text{对于所有的} \xi \in \mathrm{T}_{x^{(t)}}\mathcal{M},$$

或 $((\mathrm{D}F(x^{(t)}))^* \circ \mathrm{D}F(x^{(t)}))[\eta^{(t)}] = -(\mathrm{D}F(x^{(t)}))^*[F(x^{(t)})]$. 由 $\mathrm{D}F(x^{(t)})$ 的单射性可得

$$\eta^{(t)} = -((\mathrm{D}F(x^{(t)}))^* \circ \mathrm{D}F(x^{(t)}))^{-1}[(\mathrm{D}F(x^{(t)}))^*[F(x^{(t)})]],$$

这也是以下最小二乘问题的解:

$$\min_{\eta \in \mathrm{T}_{x^{(t)}}\mathcal{M}} \frac{1}{2}\langle \mathrm{D}F(x^{(t)})[\eta], \mathrm{D}F(x^{(t)})[\eta] \rangle + \langle \mathrm{D}F(x^{(t)})[\eta], F(x^{(t)}) \rangle. \tag{2-8}$$

由于 $\langle \mathrm{D}F(x^{(t)})[\eta^{(t)}], F(x^{(t)}) \rangle = \mathrm{D}f(x^{(t)})[\eta^{(t)}] = \mathrm{D}\bar{f}(x^{(t)})[\eta^{(t)}] = \langle \nabla \bar{f}(x^{(t)}), \eta^{(t)} \rangle$, 其中 $\bar{f} : \mathcal{E} \to \mathbb{R}$ 是 f 的光滑延拓, 式(2-8) 等价于

$$\min_{\eta \in \mathrm{T}_{x^{(t)}}\mathcal{M}} \frac{1}{2}\langle \bar{\mathcal{H}}(x^{(t)})[\eta], \eta \rangle + \langle \nabla \bar{f}(x^{(t)}), \eta \rangle. \tag{2-9}$$

根据 [120] 可得知方程(2-9) 的解为 $\eta^{(t)} = -\mathrm{grad}_g f(x^{(t)})$. 换句话说, 黎曼高斯–牛顿方法可以被理解为使用度量 g 的黎曼梯度法. 因此, 我们将该框架称为高斯–牛顿型预条件, 它可以应用于张量补全问题；具体细节见 2.5 节.

注. 取定 $\bar{\mathcal{H}}(x)$ 为函数 f 在点 $x \in \mathcal{M}$ 处欧氏度量下的黎曼 Hessian 算子 $\mathrm{Hess}_{\mathrm{e}} f(x)$. 若 $\mathrm{Hess}_{\mathrm{e}} f(x)$ 对称正定, 则度量 $g_x(\xi, \eta) = \langle \xi, \mathrm{Hess}_{\mathrm{e}} f(x)[\eta] \rangle$ 被称为 Hessian 度量, 参见 [128]. 此时黎曼梯度的反方向 $-\mathrm{grad}_g f(x)$ 正是在欧氏度量意义下的黎曼牛顿方向; 参考 [129, Proposition 4.1] 以及 [120, Proposition 2.1]. 注意我们此处提出的高斯牛顿类的预条件方法是与 Hessian 度量不同的, 这是因为式 (2-7) 仅利用了黎曼 Hessian 算子的部分信息.

2.3 在典型相关分析的应用

在本节中, 我们将我们提出的预条件框架应用于典型相关分析 (CCA) 问题. 我们将提出一个新的左右预条件子, 并证明在新度量下的黎曼 Hessian 算子的条

件数的确相比于其他度量变得更小. 我们通过数值实验验证了新的度量加速了流形优化方法.

考虑两个数据矩阵 $\mathbf{X} \in \mathbb{R}^{n\times d_x}$ 以及 $\mathbf{Y} \in \mathbb{R}^{n\times d_y}$, 它们由 n 个样本以及分别 d_x, d_y 个变量构成. CCA 的目标是选择 m 个权重 $\mathbf{u}_1, \dots, \mathbf{u}_m \in \mathbb{R}^{d_x}$ 与 $\mathbf{v}_1, \dots, \mathbf{v}_m \in \mathbb{R}^{d_y}$ 使得变换后的数据矩阵 $\mathbf{XU}$ 以及 $\mathbf{YV}$ 有着最高的相关性, 这里 $\mathbf{U} = [\mathbf{u}_1, \dots, \mathbf{u}_m]$ 以及 $\mathbf{V} = [\mathbf{v}_1, \dots, \mathbf{v}_m]$. CCA 可以被写为建模在两个广义 Stiefel 流形乘积上的优化问题, 即

$$\min_{\mathbf{U},\mathbf{V}} f(\mathbf{U},\mathbf{V}) := -\mathrm{tr}(\mathbf{U}^\top \Sigma_{xy}\mathbf{V}\mathbf{N}), \quad \text{s.t. } (\mathbf{U},\mathbf{V}) \in \mathcal{M} = \mathcal{M}_1 \times \mathcal{M}_2, \tag{2-10}$$

这里 $\Sigma_{xx} := \mathbf{X}^\top\mathbf{X} + \lambda_x \mathbf{I}_{d_x}$, $\Sigma_{yy} := \mathbf{Y}^\top\mathbf{Y} + \lambda_y \mathbf{I}_{d_y}$, $\lambda_x, \lambda_y \geq 0$ 为正则化参数, $\Sigma_{xy} := \mathbf{X}^\top\mathbf{Y}$, $\mathcal{M}_1 := \mathrm{St}_{\Sigma_{xx}}(m, d_x) = \{\mathbf{U} \in \mathbb{R}^{d_x\times m} : \mathbf{U}^\top\Sigma_{xx}\mathbf{U} = \mathbf{I}_m\}$ 和 $\mathcal{M}_2 := \mathrm{St}_{\Sigma_{yy}}(m, d_y)$ 为广义 Stiefel 流形, 以及 $\mathbf{N} := \mathrm{diag}(\mu_1, \mu_2, \dots, \mu_m)$ 满足 $\mu_1 > \mu_2 > \cdots > \mu_m > 0$. 问题 (2-10)中的目标函数 f 也被称为 von Neumann 目标函数 [130]. 问题 (2-10) 有如下的显式解

$$(\mathbf{U}^*, \mathbf{V}^*) = (\Sigma_{xx}^{-1/2}\bar{\mathbf{U}}, \Sigma_{yy}^{-1/2}\bar{\mathbf{V}}), \tag{2-11}$$

这里 $\bar{\mathbf{U}} := [\bar{\mathbf{u}}_1, \dots, \bar{\mathbf{u}}_m]$ 和 $\bar{\mathbf{V}} := [\bar{\mathbf{v}}_1, \dots, \bar{\mathbf{v}}_m]$ 为矩阵 $\Sigma_{xx}^{-1/2}\Sigma_{xy}\Sigma_{yy}^{-1/2}$ 的前 m 个左右奇异向量. 矩阵 $\Sigma_{xx}^{-1/2}\Sigma_{xy}\Sigma_{yy}^{-1/2}$ 最大的 m 个奇异值 $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_m > 0$ 被称为典型相关系数. 我们的目标是通过在 $\mathcal{M}$ 上提出新的预条件度量, 用流形优化方法求解问题 (2-10).

2.3.1　左预条件方法

给定切向量 $\xi, \eta \in \mathrm{T}_{(\mathbf{U},\mathbf{V})}\mathcal{M}$, Shustin 和 Avron 在文献 [115, §4.2] 中提出为流形 $\mathcal{M}$ 赋予如下的黎曼度量

$$g_{(\mathbf{U},\mathbf{V})}(\xi, \eta) := \langle \xi_1, \Sigma_{xx}\eta_1\rangle + \langle \xi_2, \Sigma_{yy}\eta_2\rangle, \tag{2-12}$$

其中, 切空间 $\mathrm{T}_{(\mathbf{U},\mathbf{V})}\mathcal{M}$ 的定义是 $\mathrm{T}_{(\mathbf{U},\mathbf{V})}\mathcal{M} \simeq \mathrm{T}_{\mathbf{U}}\mathcal{M}_1 \times \mathrm{T}_{\mathbf{V}}\mathcal{M}_2$ 以及

$$\mathrm{T}_{\mathbf{U}}\mathcal{M}_1 = \{\mathbf{U}\boldsymbol{\Omega}_1 + \mathbf{U}_{\Sigma_{xx}\perp}\mathbf{K}_1 : \boldsymbol{\Omega}_1 \in \mathbb{R}^{m\times m}, \boldsymbol{\Omega}_1^\top = -\boldsymbol{\Omega}_1, \mathbf{K}_1 \in \mathbb{R}^{(d_x-m)\times m}\} \tag{2-13}$$

是广义 Stiefel 流形 $\mathcal{M}_1$ 的切空间, 其维数为 $md_x - m(m+1)/2$. 矩阵 $\mathbf{U}_{\Sigma_{xx}\perp} \in \mathbb{R}^{d_x\times(d_x-m)}$ 满足 $(\mathbf{U}_{\Sigma_{xx}\perp})^\top\Sigma_{xx}\mathbf{U}_{\Sigma_{xx}\perp} = \mathbf{I}_{d_x-m}$ 和 $\mathbf{U}^\top\Sigma_{xx}\mathbf{U}_{\Sigma_{xx}\perp} = 0$. 切空间 $\mathrm{T}_{\mathbf{V}}\mathcal{M}_2$ 的定义类似.

在我们的统一框架下, 这一度量等价于设 (2-6) 中的算子为 $\bar{\mathcal{H}}_1(\mathbf{U},\mathbf{V})[\eta_1] = \Sigma_{xx}\eta_1$ 和 $\bar{\mathcal{H}}_2(\mathbf{U},\mathbf{V})[\eta_2] = \Sigma_{yy}\eta_2$, 它们具有左预条件的效果. 相对于度量 g, 将向量 $\bar{\eta} \in \mathrm{T}_{(\mathbf{U},\mathbf{V})}\mathcal{E} \simeq \mathcal{E}$ 投影到切空间 $\mathrm{T}_{(\mathbf{U},\mathbf{V})}\mathcal{M}$ 上的正交投影算子可表示为 $\Pi_{g,(\mathbf{U},\mathbf{V})}(\bar{\eta}) = \big(\bar{\eta}_1 - \mathbf{U}\,\mathrm{sym}(\mathbf{U}^\top\Sigma_{xx}\bar{\eta}_1), \bar{\eta}_2 - \mathbf{V}\,\mathrm{sym}(\mathbf{V}^\top\Sigma_{yy}\bar{\eta}_2)\big)$, 这里 $\mathcal{E} = \mathbb{R}^{d_x\times m} \times \mathbb{R}^{d_y\times m}$ 是流形 $\mathcal{M}$ 的环境空间. 因此, 由 (2-5) 可得黎曼梯度为

$$\begin{aligned}\mathrm{grad}_g f(\mathbf{U},\mathbf{V}) = (&-\Sigma_{xx}^{-1}\Sigma_{xy}\mathbf{V}\mathbf{N} + \mathbf{U}\,\mathrm{sym}(\mathbf{U}^\top\Sigma_{xy}\mathbf{V}\mathbf{N}),\\ &-\Sigma_{yy}^{-1}\Sigma_{xy}^\top\mathbf{U}\mathbf{N} + \mathbf{V}\,\mathrm{sym}(\mathbf{V}^\top\Sigma_{xy}^\top\mathbf{U}\mathbf{N})).\end{aligned} \tag{2-14}$$

由于流形优化方法的局部收敛速度与条件数 $\kappa_g(\mathrm{Hess}_g f(\mathbf{U}^*,\mathbf{V}^*))$ 密切相关 (参见定理 1.4), 我们首先利用 (1-18) 计算函数 f 在点 $(\mathbf{U},\mathbf{V})$ 处沿方向 $\eta=(\eta_1,\eta_2)\in \mathrm{T}_{(\mathbf{U},\mathbf{V})}\mathcal{M}$ 的黎曼 Hessian 算子, 得到

$$\begin{aligned}\mathrm{Hess}_g f(\mathbf{U},\mathbf{V})[\eta] = \Pi_{g,(\mathbf{U},\mathbf{V})}\big(&\eta_1\,\mathrm{sym}(\mathbf{U}^\top\Sigma_{xy}\mathbf{V}\mathbf{N}) + \mathbf{U}\,\mathrm{sym}(\eta_1^\top\Sigma_{xy}\mathbf{V}\mathbf{N})\\ &+\mathbf{U}\,\mathrm{sym}(\mathbf{U}^\top\Sigma_{xy}\eta_2\mathbf{N}) - \Sigma_{xx}^{-1}\Sigma_{xy}\eta_2\mathbf{N},\\ &\eta_2\,\mathrm{sym}(\mathbf{V}^\top\Sigma_{xy}^\top\mathbf{U}\mathbf{N}) + \mathbf{V}\,\mathrm{sym}(\eta_2^\top\Sigma_{xy}^\top\mathbf{U}\mathbf{N})\\ &+\mathbf{V}\,\mathrm{sym}(\mathbf{V}^\top\Sigma_{xy}^\top\eta_1\mathbf{N}) - \Sigma_{yy}^{-1}\Sigma_{xy}^\top\eta_1\mathbf{N}\big),\end{aligned} \tag{2-15}$$

进而我们可以计算 $\mathrm{Hess}_g f(\mathbf{U}^*,\mathbf{V}^*)$ 的条件数.

命题 2.3. 设 $\sigma_1>\sigma_2>\cdots>\sigma_{m+1}\geq\cdots\geq\sigma_{\min\{d_x,d_y\}}$ 为矩阵 $\Sigma_{xx}^{-1/2}\Sigma_{xy}\Sigma_{yy}^{-1/2}$ 的奇异值. 则有

$$\kappa_g(\mathrm{Hess}_g f(\mathbf{U}^*,\mathbf{V}^*)) = \frac{\max\left\{\frac{1}{2}(\mu_1+\mu_2)(\sigma_1+\sigma_2),\mu_1(\sigma_1+\sigma_{m+1})\right\}}{\min\{\min_{i,j\in[m],i\neq j}\frac{1}{2}(\mu_i-\mu_j)(\sigma_i-\sigma_j),\mu_m(\sigma_m-\sigma_{m+1})\}}.$$

证明. 由于 $(\mathbf{U}^*,\mathbf{V}^*)$ 是 f 的一个稳定点, 由 $(\mathbf{U}^*)^\top\Sigma_{xy}\mathbf{V}^*=\Sigma$ 以及 $\mathrm{grad}_g f(\mathbf{U}^*,\mathbf{V}^*)=0$ 可知

$$\Sigma_{xx}^{-1}\Sigma_{xy}\mathbf{V}^*=\mathbf{U}^*\Sigma \quad 和 \quad \Sigma_{yy}^{-1}\Sigma_{xy}^\top\mathbf{U}^*=\mathbf{V}^*\Sigma, \tag{2-16}$$

其中 $\Sigma=\mathrm{diag}(\sigma_1,\sigma_2,\ldots,\sigma_m)$, 并且 $\sigma_1>\sigma_2>\cdots>\sigma_m$ 是矩阵 $\Sigma_{xx}^{-1/2}\Sigma_{xy}\Sigma_{yy}^{-1/2}$ 的前 m 个最大奇异值. 因此, 由 (2-13)、(2-15) 以及 (2-16) 可得

$$\begin{aligned}g_{(\mathbf{U}^*,\mathbf{V}^*)}(\eta,\mathrm{Hess}_g f(\mathbf{U}^*,\mathbf{V}^*)[\eta]) =& \langle\eta_1,\Sigma_{xx}\eta_1\Sigma\mathbf{N}\rangle + \langle\eta_1,\Sigma_{xx}\mathbf{U}^*\,\mathrm{sym}(\eta_1^\top\Sigma_{xy}\mathbf{V}^*\mathbf{N})\rangle\\ &+\langle\eta_1,\Sigma_{xx}\mathbf{U}^*\,\mathrm{sym}((\mathbf{U}^*)^\top\Sigma_{xy}\eta_2\mathbf{N})\rangle - \langle\eta_1,\Sigma_{xy}\eta_2\mathbf{N}\rangle\\ &+\langle\eta_2,\Sigma_{yy}\eta_2\Sigma\mathbf{N}\rangle + \langle\eta_2,\Sigma_{yy}\mathbf{V}^*\,\mathrm{sym}(\eta_2^\top\Sigma_{xy}^\top\mathbf{U}^*\mathbf{N})\rangle\\ &+\langle\eta_2,\Sigma_{yy}\mathbf{V}^*\,\mathrm{sym}((\mathbf{V}^*)^\top\Sigma_{xy}^\top\eta_1\mathbf{N})\rangle - \langle\eta_2,\Sigma_{xy}^\top\eta_1\mathbf{N}\rangle\\ =& \langle\eta_1,\Sigma_{xx}\eta_1\Sigma\mathbf{N}\rangle - 2\langle\eta_1,\Sigma_{xy}\eta_2\mathbf{N}\rangle + \langle\eta_2,\Sigma_{yy}\eta_2\Sigma\mathbf{N}\rangle\end{aligned}$$

其中 $\eta=(\eta_1,\eta_2)\in\mathrm{T}_{(\mathbf{U}^*,\mathbf{V}^*)}\mathcal{M}$.

我们证明的目标是计算 $\mathrm{Hess}_g f(\mathbf{U}^*,\mathbf{V}^*)$ 的条件数. 为此, 我们利用 (1-21) 计算 $\mathrm{Hess}_g f(\mathbf{U}^*,\mathbf{V}^*)$ 的 Rayleigh 商, 并分别估计其上界与下界. 首先, Rayleigh 商可

写为

$$
\begin{aligned}
q(\eta) &= \frac{\langle \eta_1, \Sigma_{xx}\eta_1\Sigma\mathbf{N}\rangle - 2\langle \eta_1, \Sigma_{xy}\eta_2\mathbf{N}\rangle + \langle \eta_2, \Sigma_{yy}\eta_2\Sigma\mathbf{N}\rangle}{\langle \eta_1, \Sigma_{xx}\eta_1\rangle + \langle \eta_2, \Sigma_{yy}\eta_2\rangle} \\
&= \frac{\langle \tilde{\eta}_1, \tilde{\eta}_1\Sigma\mathbf{N}\rangle - 2\langle \tilde{\eta}_1, \Sigma_{xx}^{-1/2}\Sigma_{xy}\Sigma_{yy}^{-1/2}\tilde{\eta}_2\mathbf{N}\rangle + \langle \tilde{\eta}_2, \tilde{\eta}_2\Sigma\mathbf{N}\rangle}{\langle \tilde{\eta}_1, \tilde{\eta}_1\rangle + \langle \tilde{\eta}_2, \tilde{\eta}_2\rangle} \\
&= \frac{\begin{bmatrix}\mathrm{vec}(\tilde{\eta}_1)^\top & \mathrm{vec}(\tilde{\eta}_2)^\top\end{bmatrix}\begin{bmatrix}\Sigma\mathbf{N}\otimes\mathbf{I}_{d_x} & -\mathbf{N}\otimes\mathbf{M} \\ -\mathbf{N}\otimes\mathbf{M}^\top & \Sigma\mathbf{N}\otimes\mathbf{I}_{d_y}\end{bmatrix}\begin{bmatrix}\mathrm{vec}(\tilde{\eta}_1) \\ \mathrm{vec}(\tilde{\eta}_2)\end{bmatrix}}{\langle \tilde{\eta}_1, \tilde{\eta}_1\rangle + \langle \tilde{\eta}_2, \tilde{\eta}_2\rangle} \\
&= \frac{\sum_{i=1}^{m}\mu_i\begin{bmatrix}(\tilde{\eta}_1(:,i))^\top & (\tilde{\eta}_2(:,i))^\top\end{bmatrix}\begin{bmatrix}\sigma_i\mathbf{I}_{d_x} & -\mathbf{M} \\ -\mathbf{M}^\top & \sigma_i\mathbf{I}_{d_y}\end{bmatrix}\begin{bmatrix}\tilde{\eta}_1(:,i) \\ \tilde{\eta}_2(:,i)\end{bmatrix}}{\langle \tilde{\eta}_1, \tilde{\eta}_1\rangle + \langle \tilde{\eta}_2, \tilde{\eta}_2\rangle},
\end{aligned}
\tag{2-17}
$$

其中 $\mathbf{M} = \Sigma_{xx}^{-1/2}\Sigma_{xy}\Sigma_{yy}^{-1/2}$, $\tilde{\eta}_1 = \Sigma_{xx}^{1/2}\eta_1$, $\tilde{\eta}_2 = \Sigma_{yy}^{1/2}\eta_2$. 利用 (2-13), 我们可以将 $\tilde{\eta}$ 表示为

$$
\begin{aligned}
\tilde{\eta} &= (\Sigma_{xx}^{\frac{1}{2}}\eta_1, \Sigma_{yy}^{\frac{1}{2}}\eta_2) = (\Sigma_{xx}^{\frac{1}{2}}\mathbf{U}^*\boldsymbol{\Omega}_1 + \Sigma_{xx}^{\frac{1}{2}}\mathbf{U}^*_{\Sigma_{xx}\perp}\mathbf{K}_1, \Sigma_{yy}^{\frac{1}{2}}\mathbf{V}^*\boldsymbol{\Omega}_2 + \Sigma_{yy}^{\frac{1}{2}}\mathbf{V}^*_{\Sigma_{yy}\perp}\mathbf{K}_2) \\
&= (\bar{\mathbf{U}}\boldsymbol{\Omega}_1 + \bar{\mathbf{U}}_\perp\mathbf{K}_1, \bar{\mathbf{V}}\boldsymbol{\Omega}_2 + \bar{\mathbf{V}}_\perp\mathbf{K}_2),
\end{aligned}
\tag{2-18}
$$

其中 $\bar{\mathbf{U}} = \Sigma_{xx}^{\frac{1}{2}}\mathbf{U}^* \in \mathrm{St}(m, d_x)$, $\bar{\mathbf{V}} = \Sigma_{yy}^{\frac{1}{2}}\mathbf{V}^* \in \mathrm{St}(m, d_y)$, $\bar{\mathbf{U}}_\perp = \mathbf{U}^*_{\Sigma_{xx}\perp} \in \mathrm{St}(d_x - m, d_x)$ 和 $\bar{\mathbf{V}}_\perp = \Sigma_{yy}^{\frac{1}{2}}\mathbf{V}^*_{\Sigma_{yy}\perp} \in \mathrm{St}(d_y - m, d_y)$ 满足 $\bar{\mathbf{U}}^\top\bar{\mathbf{U}}_\perp = 0$, $\bar{\mathbf{V}}^\top\bar{\mathbf{V}}_\perp = 0$. 此外, 由 (2-11)有 $\mathbf{M} = [\bar{\mathbf{U}}\ \bar{\mathbf{U}}_\perp]\,\mathrm{diag}(\sigma_1, \ldots, \sigma_r, 0, \ldots, 0)[\bar{\mathbf{V}}\ \bar{\mathbf{V}}_\perp]^\top$, 其中 $r = \mathrm{rank}(\mathbf{M})$. 将 (2-18) 代入 (2-17), 可得

$$
\begin{aligned}
q(\eta) &= \frac{\sum_{i=1}^{m}\mu_i\begin{bmatrix}[\bar{\mathbf{U}}\ \bar{\mathbf{U}}_\perp]\bar{\boldsymbol{\Omega}}_1(:,i) \\ [\bar{\mathbf{V}}\ \bar{\mathbf{V}}_\perp]\bar{\boldsymbol{\Omega}}_2(:,i)\end{bmatrix}^\top\begin{bmatrix}\sigma_i\mathbf{I}_{d_x} & -\mathbf{M} \\ -\mathbf{M}^\top & \sigma_i\mathbf{I}_{d_y}\end{bmatrix}\begin{bmatrix}[\bar{\mathbf{U}}\ \bar{\mathbf{U}}_\perp]\bar{\boldsymbol{\Omega}}_1(:,i) \\ [\bar{\mathbf{V}}\ \bar{\mathbf{V}}_\perp]\bar{\boldsymbol{\Omega}}_2(:,i)\end{bmatrix}}{\langle \tilde{\eta}_1, \tilde{\eta}_1\rangle + \langle \tilde{\eta}_2, \tilde{\eta}_2\rangle} \\
&= \frac{\sum_{i=1}^{m}\mu_i\Big(-\sum_{j=1}^{r}2\sigma_j\bar{\boldsymbol{\Omega}}_1(j,i)\bar{\boldsymbol{\Omega}}_2(j,i) + \sum_{j=1}^{d_x}\sigma_i\bar{\boldsymbol{\Omega}}_1(j,i)^2 + \sum_{j=1}^{d_y}\sigma_i\bar{\boldsymbol{\Omega}}_2(j,i)^2\Big)}{\|\bar{\boldsymbol{\Omega}}_1\|_{\mathrm{F}}^2 + \|\bar{\boldsymbol{\Omega}}_2\|_{\mathrm{F}}^2},
\end{aligned}
\tag{2-19}
$$

其中 $\bar{\boldsymbol{\Omega}}_\ell := \begin{bmatrix}\boldsymbol{\Omega}_\ell \\ \mathbf{K}_\ell\end{bmatrix}$, $\ell = 1, 2$.

随后, 利用 $\bar{\boldsymbol{\Omega}}_\ell(j,i)^2 = \bar{\boldsymbol{\Omega}}_\ell(i,j)^2$, $\ell = 1, 2$, $i, j \in [m]$, 我们对 (2-19) 中的项进行

重新分组并得到

$$
\begin{aligned}
q(\eta) &\leq \frac{\sum_{i=1}^{m}\mu_i\Big(\sum_{j=1}^{r}\sigma_j(\bar{\boldsymbol{\Omega}}_1(j,i)^2+\bar{\boldsymbol{\Omega}}_2(j,i)^2)+\sum_{j=1}^{d_x}\sigma_i\bar{\boldsymbol{\Omega}}_1(j,i)^2+\sum_{j=1}^{d_y}\sigma_i\bar{\boldsymbol{\Omega}}_2(j,i)^2\Big)}{\|\bar{\boldsymbol{\Omega}}_1\|_{\mathrm{F}}^2+\|\bar{\boldsymbol{\Omega}}_2\|_{\mathrm{F}}^2} \\
&= \frac{\sum_{i=1}^{m}\Big(\sum_{j=1}^{r}\bar{s}_{ij}(\bar{\boldsymbol{\Omega}}_1(j,i)^2+\bar{\boldsymbol{\Omega}}_2(j,i)^2)+\sum_{j=r+1}^{d_x}\mu_i\sigma_i\bar{\boldsymbol{\Omega}}_1(j,i)^2+\sum_{j=r+1}^{d_y}\mu_i\sigma_i\bar{\boldsymbol{\Omega}}_2(j,i)^2\Big)}{\|\bar{\boldsymbol{\Omega}}_1\|_{\mathrm{F}}^2+\|\bar{\boldsymbol{\Omega}}_2\|_{\mathrm{F}}^2} \\
&\leq \max\left\{(\mu_1+\mu_2)(\sigma_1+\sigma_2)/2,\mu_1(\sigma_1+\sigma_{m+1})\right\},
\end{aligned}
\tag{2-20}
$$

其中 $\bar{s}_{ij} := \begin{cases}(\mu_i+\mu_j)(\sigma_i+\sigma_j)/2, & j=1,2,\ldots,m;\\ \mu_i(\sigma_i+\sigma_j), & j=m+1,m+2,\ldots,r\end{cases}$, $i=1,2,\ldots,m$. 等号成立当且仅当以下条件同时满足: 1) 对所有 $i\in[m], j\in[r]$, 有 $\bar{\boldsymbol{\Omega}}_1(j,i)=-\bar{\boldsymbol{\Omega}}_2(j,i)$; 2) 对所有 $i\in[m], j=r+1,\ldots,d_x$, 有 $\bar{\boldsymbol{\Omega}}_1(j,i)^2=0$; 3) 对所有 $i\in[m], j=r+1,\ldots,d_y$, 有 $\bar{\boldsymbol{\Omega}}_2(j,i)^2=0$; 4) 除 (i^*,j^*) 外, $\bar{\boldsymbol{\Omega}}_1(j,i)=\bar{\boldsymbol{\Omega}}_2(j,i)=0$, 其中 $(i^*,j^*)\in\arg\max_{i\in[m],j\in[r],i\neq j}\bar{s}_{ij}\subseteq\{(1,2),(2,1),(1,m+1)\}$.

类似地, 我们可以以同样的方法计算 Rayleigh 商的下界, 从而得到

$$
\begin{aligned}
q(\eta) &\geq \frac{\sum_{i=1}^{m}\mu_i\Big(-\sum_{j=1}^{r}\sigma_j(\bar{\boldsymbol{\Omega}}_1(j,i)^2+\bar{\boldsymbol{\Omega}}_2(j,i)^2)+\sum_{j=1}^{d_x}\sigma_i\bar{\boldsymbol{\Omega}}_1(j,i)^2+\sum_{j=1}^{d_y}\sigma_i\bar{\boldsymbol{\Omega}}_2(j,i)^2\Big)}{\|\bar{\boldsymbol{\Omega}}_1\|_{\mathrm{F}}^2+\|\bar{\boldsymbol{\Omega}}_2\|_{\mathrm{F}}^2} \\
&= \frac{\sum_{i=1}^{m}\Big(\sum_{j=1}^{r}\underline{s}_{ij}(\bar{\boldsymbol{\Omega}}_1(j,i)^2+\bar{\boldsymbol{\Omega}}_2(j,i)^2)+\sum_{j=r+1}^{d_x}\mu_i\sigma_i\bar{\boldsymbol{\Omega}}_1(j,i)^2+\sum_{j=r+1}^{d_y}\mu_i\sigma_i\bar{\boldsymbol{\Omega}}_2(j,i)^2\Big)}{\|\bar{\boldsymbol{\Omega}}_1\|_{\mathrm{F}}^2+\|\bar{\boldsymbol{\Omega}}_2\|_{\mathrm{F}}^2} \\
&\geq \min\{\min_{i,j\in[m],i\neq j}(\mu_i-\mu_j)(\sigma_i-\sigma_j)/2,\mu_m(\sigma_m-\sigma_{m+1})\},
\end{aligned}
\tag{2-21}
$$

这里 $\underline{s}_{ij} := \begin{cases}(\mu_i-\mu_j)(\sigma_i-\sigma_j)/2, & j=1,2,\ldots,m;\\ \mu_i(\sigma_i-\sigma_j), & j=m+1,m+2,\ldots,r\end{cases}$ for $i=1,2,\ldots,m$. 等号成立当且仅当: 1) 对所有 $i\in[m], j\in[r]$, 有 $\bar{\boldsymbol{\Omega}}_1(j,i)=\bar{\boldsymbol{\Omega}}_2(j,i)$; 2) 对所有 $i\in[m], j=r+1,\ldots,d_x$, 有 $\bar{\boldsymbol{\Omega}}_1(j,i)^2=0$; 3) 对所有 $i\in[m], j=r+1,\ldots,d_y$, 有 $\bar{\boldsymbol{\Omega}}_2(j,i)^2=0$; 4) 除 (i^*,j^*) 外, $\bar{\boldsymbol{\Omega}}_1(j,i)=\bar{\boldsymbol{\Omega}}_2(j,i)=0$ 其中 $(i^*,j^*)\in\arg\min_{i\in[m],j\in[r],i\neq j}\underline{s}_{ij}$.

由于 (2-20)–(2-21) 中的不等式都是紧的, 我们完成了证明. □

该命题的证明可参见文献 [131, Theorem 5]. 不过, 为了便于后续命题 2.5与命题 2.8 的证明, 我们给出了命题 2.3 的一种不同的证明. 具体而言, 命题 2.5与命题 2.8 的证明思路遵循相同的步骤: 1) 在给定的度量下计算黎曼 Hessian 算子 $\mathrm{Hess}_g f(\mathbf{U}^*,\mathbf{V}^*)[\eta]$; 2) 计算 Rayleigh 商 $q(\eta)$; 3) 将切空间的参数化代入 $q(\eta)$ 中, 求其最大值与最小值. 需要指出的是, 当 $m=1$ 时, 命题 2.3中的结论可化简为

$\kappa_g(\mathrm{Hess}_g f(\mathbf{U}^*,\mathbf{V}^*)) = (\sigma_1+\sigma_2)/(\sigma_1-\sigma_2)$, 这与文献 [115, Lemma 4.1] 中的结果一致.

2.3.2　新的左右预条件子

通过观察式 (2-15) 中的二阶信息, 我们的目标是对式 (2-15) 中的对角块进行近似, 并提出一种新的度量, 使得 (2-6) 中的算子同时具有左、右预条件效果. 为此, 我们利用第 2.2.2 节中介绍的左右预条件思想, 提出如下新的黎曼度量

$$g_{\mathrm{new},(\mathbf{U},\mathbf{V})}(\xi,\eta) := \langle \xi_1, \Sigma_{xx}\eta_1\mathbf{M}_{1,2}\rangle + \langle \xi_2, \Sigma_{yy}\eta_2\mathbf{M}_{2,2}\rangle, \tag{2-22}$$

这里 $\mathbf{M}_{1,2} := (\mathrm{sym}(\mathbf{U}^\top\Sigma_{xy}\mathbf{V}\mathbf{N})^2+\delta\mathbf{I}_m)^{1/2}$, $\mathbf{M}_{2,2} := (\mathrm{sym}(\mathbf{V}^\top\Sigma_{xy}^\top\mathbf{U}\mathbf{N})^2+\delta\mathbf{I}_m)^{1/2}$, 以及 $\delta>0$. 相应的正交投影算子 $\Pi_{\mathrm{new},(\mathbf{U},\mathbf{V})}$ 由下面的命题给出.

命题 2.4. *在新的度量 (2-22)下, $\bar\eta\in \mathrm{T}_{(\mathbf{U},\mathbf{V})}\mathcal{E}\simeq\mathcal{E}$ 往切空间 $\mathrm{T}_{(\mathbf{U},\mathbf{V})}\mathcal{M}$ 上的正交投影为*

$$\Pi_{\mathrm{new},(\mathbf{U},\mathbf{V})}(\bar\eta) = (\Pi_{\mathrm{new},\mathbf{U}}(\bar\eta_1),\Pi_{\mathrm{new},\mathbf{V}}(\bar\eta_2)) = (\bar\eta_1-\mathbf{U}\mathbf{S}_1\mathbf{M}_{1,2}^{-1},\bar\eta_2-\mathbf{V}\mathbf{S}_2\mathbf{M}_{2,2}^{-1}), \tag{2-23}$$

这里 $\mathbf{S}_1$, $\mathbf{S}_2$ 为 Lyapunov 方程 $\mathbf{M}_{1,2}^{-1}\mathbf{S}_1+\mathbf{S}_1\mathbf{M}_{1,2}^{-1}=2\,\mathrm{sym}(\mathbf{U}^\top\Sigma_{xx}\bar\eta_1)$ 和 $\mathbf{M}_{2,2}^{-1}\mathbf{S}_2+\mathbf{S}_2\mathbf{M}_{2,2}^{-1}=2\,\mathrm{sym}(\mathbf{V}^\top\Sigma_{yy}\bar\eta_2)$ 的唯一解.

证明. 我们只需证明 $\Pi_{\mathrm{new},\mathbf{U}}(\bar\eta_1)=\bar\eta_1-\mathbf{U}\mathbf{S}_1\mathbf{M}_{1,2}^{-1}$, 另外一个结论同理可得. 我们回忆式 (2-13)中给出的 $\mathcal{M}_1$ 的切空间. 在度量 (2-22) 下往切空间 $\mathrm{T}_{\mathbf{U}}\mathcal{M}_1$ 的正交补为

$$(\mathrm{T}_{\mathbf{U}}\mathcal{M}_1)^\perp = \{\mathbf{U}\mathbf{S}_1\mathbf{M}_{1,2}^{-1} : \mathbf{S}_1\in\mathbb{R}^{m\times m},\ \mathbf{S}_1=\mathbf{S}_1^\top\}, \tag{2-24}$$

这是由于 $\{\mathbf{U}\mathbf{S}_1\mathbf{M}_{1,2}^{-1} : \mathbf{S}_1\in\mathbb{R}^{m\times m},\ \mathbf{S}_1=\mathbf{S}_1^\top\}$ 的维数是 $m(m+1)/2$ 以及对所有满足 $\mathbf{S}_1=\mathbf{S}_1^\top$ 和 $\boldsymbol{\Omega}_1=-\boldsymbol{\Omega}_1^\top$ 的 $\mathbf{S}_1,\boldsymbol{\Omega}_1,\mathbf{K}_1$ 有 $\mathrm{tr}((\mathbf{U}\mathbf{S}_1\mathbf{M}_{1,2}^{-1})^\top\Sigma_{xx}(\mathbf{U}\boldsymbol{\Omega}_1+\mathbf{U}_{\Sigma_{xx}\perp}\mathbf{K}_1)\mathbf{M}_{1,2})=0$. 此外, 根据直和关系 $\mathrm{T}_{\mathbf{U}}\mathcal{M}_1\oplus(\mathrm{T}_{\mathbf{U}}\mathcal{M}_1)^\perp=\mathrm{T}_{\mathbf{U}}\mathbb{R}^{d_x\times m}\simeq\mathbb{R}^{d_x\times m}$, 有对 $\bar\eta_1\in\mathbb{R}^{d_x\times m}$ 的唯一正交分解

$$\bar\eta_1 = \Pi_{\mathrm{new},\mathbf{U}}(\bar\eta_1)+\Pi_{\mathrm{new},\mathbf{U}}^\perp(\bar\eta_1) = (\mathbf{U}\boldsymbol{\Omega}_1+\mathbf{U}_{\Sigma_{xx}\perp}\mathbf{K}_1)+\mathbf{U}\mathbf{S}_1\mathbf{M}_{1,2}^{-1}, \tag{2-25}$$

也就是说 $\Pi_{\mathrm{new},\mathbf{U}}(\bar\eta_1)=\bar\eta_1-\Pi_{\mathrm{new},\mathbf{U}}^\perp(\bar\eta_1)=\bar\eta_1-\mathbf{U}\mathbf{S}_1\mathbf{M}_{1,2}^{-1}$. 为了给出 $\mathbf{S}_1$ 满足的表达式, 我们在式 (2-25) 两边同时左乘 $\mathbf{U}^\top\Sigma_{xx}$ 得到 $\mathbf{U}^\top\Sigma_{xx}\bar\eta_1=\boldsymbol{\Omega}_1+\mathbf{S}_1\mathbf{M}_{1,2}^{-1}$. 将 $\mathbf{U}^\top\Sigma_{xx}\bar\eta_1$ 和 $(\mathbf{U}^\top\Sigma_{xx}\bar\eta_1)^\top$ 相加可得 $\mathbf{S}_1\mathbf{M}_{1,2}^{-1}+\mathbf{M}_{1,2}^{-1}\mathbf{S}_1=\mathbf{U}^\top\Sigma_{xx}\bar\eta_1+\bar\eta_1^\top\Sigma_{xx}\mathbf{U}$, 根据 [132, Theorem 2.4.4.1] 可知该方程有唯一解. □

由命题 2.2与 2.4 得知 f 在 $(\mathbf{U},\mathbf{V})\in\mathcal{M}$ 处的黎曼梯度为

$$\mathrm{grad}_{\mathrm{new}} f(\mathbf{U},\mathbf{V}) = -((\Sigma_{xx}^{-1}\Sigma_{xy}\mathbf{V}\mathbf{N}+\mathbf{U}\mathbf{S}_1)\mathbf{M}_{1,2}^{-1},(\Sigma_{yy}^{-1}\Sigma_{xy}^\top\mathbf{U}\mathbf{N}+\mathbf{V}\mathbf{S}_2)\mathbf{M}_{2,2}^{-1}). \tag{2-26}$$

由于 $\mathbf{M}_{1,2}, \mathbf{M}_{2,2} \in \mathbb{R}^{m\times m}$ 并且 $m \ll \min\{d_x, d_y\}$, 在新度量 (2-22)下计算黎曼梯度的计算量与度量 (2-12)是差不多的. 接下来, f 在 $(\mathbf{U}^*, \mathbf{V}^*)$ 处的黎曼 Hessian 算子将切向量 η 映射为

$$\begin{aligned}\mathrm{Hess}_{\mathrm{new}} f(\mathbf{U}^*, \mathbf{V}^*)[\eta] &= \Pi_{\mathrm{new},(\mathbf{U}^*,\mathbf{V}^*)}(\mathrm{D}\bar{G}_{\mathrm{new}}(\mathbf{U}^*, \mathbf{V}^*)[\eta]) \\ &= \Pi_{\mathrm{new},(\mathbf{U}^*,\mathbf{V}^*)}(-\Sigma_{xx}^{-1}\Sigma_{xy}\eta_2\mathbf{N}\mathbf{M}_{1,2}^{-1} + \Sigma_{xx}^{-1}\Sigma_{xy}\mathbf{V}^*\mathbf{N}\mathbf{M}_{1,2}^{-1}\dot{\mathbf{M}}_{1,2}\mathbf{M}_{1,2}^{-1} \\ &\qquad - \eta_1\mathbf{S}_1\mathbf{M}_{1,2}^{-1} - \mathbf{U}^*\dot{\mathbf{S}}_1\mathbf{M}_{1,2}^{-1} + \mathbf{U}^*\mathbf{S}_1\mathbf{M}_{1,2}^{-1}\dot{\mathbf{M}}_{1,2}\mathbf{M}_{1,2}^{-1}, \\ &\qquad - \Sigma_{yy}^{-1}\Sigma_{xy}^{\top}\eta_1\mathbf{N}\mathbf{M}_{2,2}^{-1} + \Sigma_{yy}^{-1}\Sigma_{xy}^{\top}\mathbf{U}^*\mathbf{N}\mathbf{M}_{2,2}^{-1}\dot{\mathbf{M}}_{2,2}\mathbf{M}_{2,2}^{-1} \\ &\qquad - \eta_2\mathbf{S}_2\mathbf{M}_{2,2}^{-1} - \mathbf{V}^*\dot{\mathbf{S}}_2\mathbf{M}_{2,2}^{-1} + \mathbf{V}^*\mathbf{S}_2\mathbf{M}_{2,2}^{-1}\dot{\mathbf{M}}_{2,2}\mathbf{M}_{2,2}^{-1}),\end{aligned}$$

这里 $\bar{G}_{\mathrm{new}} : \mathcal{E} \to \mathbb{R}$ 为 $\mathrm{grad}_{\mathrm{new}} f$ 的光滑延拓, $\dot{\mathbf{M}}_{1,2} := \mathrm{D}\mathbf{M}_{1,2}(\mathbf{U}^*, \mathbf{V}^*)[\eta]$, $\dot{\mathbf{M}}_{2,2} := \mathrm{D}\mathbf{M}_{2,2}(\mathbf{U}^*, \mathbf{V}^*)[\eta]$, 对称矩阵 $\dot{\mathbf{S}}_1$ 和 $\dot{\mathbf{S}}_2$ 满足下面的 Lyapunov 方程

$$\begin{aligned}\mathrm{sym}(\dot{\mathbf{M}}_{1,2}\mathbf{S}_1 + \mathbf{M}_{1,2}\dot{\mathbf{S}}_1 + \dot{\mathbf{M}}_{1,2}\Sigma\mathbf{N} + \mathbf{M}_{1,2}(\eta_1^{\top}\Sigma_{xy}\mathbf{V}^* + (\mathbf{U}^*)^{\top}\Sigma_{xy}\eta_2)) &= 0, \\ \mathrm{sym}(\dot{\mathbf{M}}_{2,2}\mathbf{S}_2 + \mathbf{M}_{2,2}\dot{\mathbf{S}}_2 + \dot{\mathbf{M}}_{2,2}\Sigma\mathbf{N} + \mathbf{M}_{2,2}(\eta_2^{\top}\Sigma_{xy}^{\top}\mathbf{U}^* + (\mathbf{V}^*)^{\top}\Sigma_{xy}^{\top}\eta_1)) &= 0.\end{aligned}$$

最后, 我们通过计算 $\mathrm{Hess}_g f(\mathbf{U}^*, \mathbf{V}^*)$ 的条件数来说明度量 (2-22) 的效果, 具体结果见下述命题. 该命题的证明可采用与命题 2.3 的证明类似思路完成.

命题 2.5. *设 $\sigma_1 > \sigma_2 > \cdots > \sigma_{m+1} \geq \cdots \geq \sigma_{\min\{d_x,d_y\}}$ 为矩阵 $\Sigma_{xx}^{-1/2}\Sigma_{xy}\Sigma_{yy}^{-1/2}$ 的奇异值. 则在局部极小点 $(\mathbf{U}^*, \mathbf{V}^*)$ 处, 其条件数可表示为*

$$\kappa_{\mathrm{new}}(\mathrm{Hess}_{\mathrm{new}} f(\mathbf{U}^*, \mathbf{V}^*)) = \frac{\max\{\max\limits_{i,j\in[m],i\neq j} \frac{(\mu_i+\mu_j)(\sigma_i+\sigma_j)}{\sqrt{\mu_i^2\sigma_i^2+\delta}+\sqrt{\mu_j^2\sigma_j^2+\delta}}, \max\limits_{i\in[m]} \frac{\mu_i(\sigma_i+\sigma_{m+1})}{\sqrt{\mu_i^2\sigma_i^2+\delta}}\}}{\min\{\min\limits_{i,j\in[m],i\neq j} \frac{(\mu_i-\mu_j)(\sigma_i-\sigma_j)}{\sqrt{\mu_i^2\sigma_i^2+\delta}+\sqrt{\mu_j^2\sigma_j^2+\delta}}, \min\limits_{i\in[m]} \frac{\mu_i(\sigma_i-\sigma_{m+1})}{\sqrt{\mu_i^2\sigma_i^2+\delta}}\}}. \tag{2-27}$$

证明. 参照命题 2.3 的证明思路, 我们在新提出的度量 (2-22) 下计算 Rayleigh 商 (1-21), 并分别估计其上界与下界. 为此, 我们首先计算并化简黎曼 Hessian 算子 $\mathrm{Hess}_{\mathrm{new}} f(\mathbf{U}^*, \mathbf{V}^*)[\eta]$.

通过求解命题 2.4 中的 Lyapunov 方程, 可以得到 $\mathbf{S}_1 = \mathbf{S}_2 = -\Sigma\mathbf{N}$, 其中 $\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_m)$. 由于 $(\mathbf{U}^*, \mathbf{V}^*)$ 是 f 的一个稳定点, 根据命题 1.3 有黎曼梯度等于 0, 即 $\mathrm{grad}_{\mathrm{new}} f(\mathbf{U}^*, \mathbf{V}^*) = 0$ 于是我们有 $\Sigma_{xx}^{-1}\Sigma_{xy}\mathbf{V}^* = \mathbf{U}^*\Sigma$ 和 $\Sigma_{yy}^{-1}\Sigma_{xy}^{\top}\mathbf{U}^* = \mathbf{V}^*\Sigma$. 因此我们可以将黎曼 Hessian $\mathrm{Hess}_{\mathrm{new}} f(\mathbf{U}^*, \mathbf{V}^*)[\eta]$ 化简为

$$\begin{aligned}\mathrm{Hess}_{\mathrm{new}} f(\mathbf{U}^*, \mathbf{V}^*)[\eta] &= \Pi_{\mathrm{new},(\mathbf{U}^*,\mathbf{V}^*)}(\mathrm{D}\bar{G}_{\mathrm{new}}(\mathbf{U}^*, \mathbf{V}^*)[\eta]) \\ &= \Pi_{\mathrm{new},(\mathbf{U}^*,\mathbf{V}^*)}(-\Sigma_{xx}^{-1}\Sigma_{xy}\eta_2\mathbf{N}(\mathbf{M}_{1,2}^*)^{-1} + \eta_1\Sigma\mathbf{N}(\mathbf{M}_{1,2}^*)^{-1}, \\ &\qquad -\Sigma_{yy}^{-1}\Sigma_{xy}^{\top}\eta_1\mathbf{N}(\mathbf{M}_{2,2}^*)^{-1} + \eta_2\Sigma\mathbf{N}(\mathbf{M}_{2,2}^*)^{-1}),\end{aligned}$$

这里 $\mathbf{M}_{1,2}^* = \mathbf{M}_{2,2}^* = (\Sigma^2\mathbf{N}^2 + \delta\mathbf{I}_m)^{1/2}$ 为对角矩阵, 并且我们使用了 (2-24)中关于 $(\mathrm{T}_{\mathbf{U}^*}\mathcal{M}_1)^{\perp}$ 和 $(\mathrm{T}_{\mathbf{V}^*}\mathcal{M}_2)^{\perp}$ 的参数化.

接下来计算 Rayleigh 商. 由 (2-13) 可知,

$$\begin{aligned} q(\eta) &= \frac{g_{\mathrm{new},(\mathbf{U}^*,\mathbf{V}^*)}(\eta, \mathrm{Hess}_{\mathrm{new}} f(\mathbf{U}^*,\mathbf{V}^*)[\eta])}{g_{\mathrm{new},(\mathbf{U}^*,\mathbf{V}^*)}(\eta,\eta)} \\ &= \frac{\langle \eta_1, \Sigma_{xx}\eta_1 \Sigma\mathbf{N}\rangle - 2\langle \eta_1, \Sigma_{xy}\eta_2\mathbf{N}\rangle + \langle \eta_2, \Sigma_{yy}\eta_2\Sigma\mathbf{N}\rangle}{\langle \eta_1, \Sigma_{xx}\eta_1\mathbf{M}^*_{1,2}\rangle + \langle \eta_2, \Sigma_{yy}\eta_2\mathbf{M}^*_{2,2}\rangle} \end{aligned}$$

对任意 $\eta = (\eta_1, \eta_2) \in \mathrm{T}_{(\mathbf{U}^*,\mathbf{V}^*)}\mathcal{M}$ 成立, 其中, 我们利用了 $\dot{\mathbf{S}}_1$ 和 $\dot{\mathbf{S}}_2$ 为对称矩阵, 以及由 (2-13)得知 $\langle \eta_1, \Sigma_{xx}\mathbf{U}^*\dot{\mathbf{S}}_1\rangle = \langle \eta_2, \Sigma_{yy}\mathbf{V}^*\dot{\mathbf{S}}_2\rangle = 0$.

可以注意到, $q(\eta)$ 的分母与 (2-17) 中的 $\langle \eta_1, \Sigma_{xx}\eta_1\rangle + \langle \eta_2, \Sigma_{yy}\eta_2\rangle$ 的唯一区别在于权重的不同, 而 $\mathbf{M}^*_{1,2}$ 与 $\mathbf{M}^*_{2,2}$ 均为对角矩阵. 因此, 我们可以完全参照命题2.3中的分析方法, 对 $q(\eta)$ 的上界与下界进行估计, 从而得到上述结论. □

条件数的降低　我们通过条件数的降低来说明所提出的黎曼度量 (2-22) 的效果. 由于黎曼 Hessian 算子的条件数被降低, 在定理 1.4 的意义下, 黎曼度量 (2-22) 的确加速黎曼方法的收敛. 为此, 我们首先给出引理 2.6, 用于简化 (2-27) 中的条件数表达式;随后, 我们在命题 2.7 中证明 $\kappa_{\mathrm{new}}(\mathrm{Hess}_{\mathrm{new}} f(\mathbf{U}^*,\mathbf{V}^*)) \le \kappa_g(\mathrm{Hess}_g f(\mathbf{U}^*,\mathbf{V}^*))$.

引理 2.6. *记 1)* $\bar{\upsilon}_{ij}(\delta) := (\mu_i+\mu_j)(\sigma_i+\sigma_j)/(\sqrt{\mu_i^2\sigma_i^2+\delta}+\sqrt{\mu_j^2\sigma_j^2+\delta})$ *和* $\bar{\upsilon}_{i,m+1}(\delta) = \mu_i(\sigma_i+\sigma_{m+1})/\sqrt{\mu_i^2\sigma_i^2+\delta}$*; 2)* $\underline{\upsilon}_{ij}(\delta) := (\mu_i-\mu_j)(\sigma_i-\sigma_j)/(\sqrt{\mu_i^2\sigma_i^2+\delta}+\sqrt{\mu_j^2\sigma_j^2+\delta})$ *和* $\underline{\upsilon}_{i,m+1}(\delta) := \mu_i(\sigma_i-\sigma_{m+1})/\sqrt{\mu_i^2\sigma_i^2+\delta}$ *对* $i,j \in [m]$*. 则对所有满足* $m \ge 3$ *的* $1 \le i < j < k \le m+1$*, 我们有*

$$\bar{\upsilon}_{ij}(0) > \bar{\upsilon}_{ik}(0) \qquad 和 \qquad \underline{\upsilon}_{ij}(0) < \underline{\upsilon}_{ik}(0).$$

证明. 我们发现若 $\mu_{m+1} = 0$, 则对 $i \in [m]$ 和 $j \in [m+1]$ 有 $\bar{\upsilon}_{ij}(0) = (\mu_i+\mu_j)(\sigma_i+\sigma_j)/(\mu_i\sigma_i+\mu_j\sigma_j)$ 以及 $\underline{\upsilon}_{ij}(0) = (\mu_i-\mu_j)(\sigma_i-\sigma_j)/(\mu_i\sigma_i+\mu_j\sigma_j)$ 成立. 首先, 我们来证明 $\bar{\upsilon}_{ij}(0) > \bar{\upsilon}_{ik}(0)$. 由于 $\mu_i > \mu_j > \mu_k$ 和 $\sigma_i > \sigma_j > \sigma_k$, 我们有

$$\begin{aligned} \bar{\upsilon}_{ij}(0) - \bar{\upsilon}_{ik}(0) &= \frac{(\mu_i\sigma_j+\mu_j\sigma_i)(\mu_i\sigma_i+\mu_k\sigma_k) - (\mu_i\sigma_k+\mu_k\sigma_i)(\mu_i\sigma_i+\mu_j\sigma_j)}{(\mu_i\sigma_i+\mu_j\sigma_j)(\mu_i\sigma_i+\mu_k\sigma_k)} \\ &= \frac{(\mu_i^2-\mu_j\mu_k)\sigma_i(\sigma_j-\sigma_k) + \mu_i(\mu_j-\mu_k)(\sigma_i^2-\sigma_j\sigma_k)}{(\mu_i\sigma_i+\mu_j\sigma_j)(\mu_i\sigma_i+\mu_k\sigma_k)} > 0. \end{aligned}$$

因此, $\bar{\upsilon}_{ij}(0) > \bar{\upsilon}_{ik}(0)$ 成立. 根据 $\bar{\upsilon}_{ij}(0) + \underline{\upsilon}_{ij}(0) = 2$ 和 $\bar{\upsilon}_{ik}(0) + \underline{\upsilon}_{ik}(0) = 2$, 同理可证 $\underline{\upsilon}_{ij}(0) < \underline{\upsilon}_{ik}(0)$. □

接下来, 根据引理 2.6 以及变量 $\bar{\upsilon}_{ij}$ 和 $\underline{\upsilon}_{ij}$ 关于 $\delta \in [0,\infty)$ 的连续性可以得知, 存在一个常数 $\bar{\delta}_1 > 0$, 使得

$$\bar{\upsilon}_{ij}(\delta) > \bar{\upsilon}_{ik}(\delta) \qquad 和 \qquad \underline{\upsilon}_{ij}(\delta) < \underline{\upsilon}_{ik}(\delta)$$

对所有的 $1 \le i < j < k \le m+1$ 和 $\delta \in (0, \bar{\delta}_1)$ 成立. 于是, 我们可以简化条件数

$$\kappa_{\text{new}}(\text{Hess}_{\text{new}} f(\mathbf{U}^*, \mathbf{V}^*)) = \frac{\max_{i\in[m], j\in[m+1], i\neq j} \bar{\upsilon}_{ij}(\delta)}{\min_{i\in[m], j\in[m+1], i\neq j} \underline{\upsilon}_{ij}(\delta)} = \frac{\max_{i\in[m]} \bar{\upsilon}_{i,i+1}(\delta)}{\min_{i\in[m]} \underline{\upsilon}_{i,i+1}(\delta)}. \quad (2\text{-}28)$$

我们的目标是证明 $\kappa_{\text{new}}(\text{Hess}_{\text{new}} f(\mathbf{U}^*, \mathbf{V}^*)) \le \kappa_g(\text{Hess}_g f(\mathbf{U}^*, \mathbf{V}^*))$ 对 $m \ge 2$ 成立. 注意当 $m = 1$ 时 $\kappa_{\text{new}}(\text{Hess}_{\text{new}} f(\mathbf{U}^*, \mathbf{V}^*)) = \kappa_g(\text{Hess}_g f(\mathbf{U}^*, \mathbf{V}^*))$, 这是因为式 (2-22) 中的右预条件子退化为标量, 从而不再具有预条件的效果.

命题 2.7. *假设 $m \ge 2$. 则存在一个常数 $\bar{\delta} > 0$, 使得*

$$\kappa_{\text{new}}(\text{Hess}_{\text{new}} f(\mathbf{U}^*, \mathbf{V}^*)) \le \kappa_g(\text{Hess}_g f(\mathbf{U}^*, \mathbf{V}^*))$$

对所有式 (2-22)中的 $\delta \in (0, \bar{\delta})$ 成立.

证明. 若 $\arg\max_{i\in[m]} \bar{\upsilon}_{i,i+1}(0) = \{i^*\}$ 对单个的 $i^* \in [m]$ 成立, 则根据 $\bar{\upsilon}_{i,i+1}(0) + \underline{\upsilon}_{i,i+1}(0) = 2$ 可知 $\arg\min_{i\in[m]} \underline{\upsilon}_{i,i+1}(0) = \arg\max_{i\in[m]} \bar{\upsilon}_{i,i+1}(0) = \{i^*\}$. 由于 $\bar{\upsilon}_{ij}(\delta)$ 和 $\underline{\upsilon}_{ij}(\delta)$ 是关于 δ 连续的, 存在 $\bar{\delta} > 0$, 使得 $\{i^*\} = \arg\max_{i\in[m]} \bar{\upsilon}_{i,i+1}(\delta)$ 和 $\{i^*\} = \arg\min_{i\in[m]} \underline{\upsilon}_{i,i+1}(\delta)$ 对所有的 $\delta \in [0, \bar{\delta})$ 恒成立. 接下来, 我们可以得到

$$\begin{aligned}
\kappa_{\text{new}}(\text{Hess}_{\text{new}} f(\mathbf{U}^*, \mathbf{V}^*)) &= \frac{\bar{\upsilon}_{i^*,i^*+1}(\delta)}{\underline{\upsilon}_{i^*,i^*+1}(\delta)} = \frac{\frac{1}{2}(\mu_{i^*} + \mu_{i^*+1})(\sigma_{i^*} + \sigma_{i^*+1})}{\frac{1}{2}(\mu_{i^*} - \mu_{i^*+1})(\sigma_{i^*} - \sigma_{i^*+1})} \\
&\le \frac{\max\{\frac{1}{2}(\mu_1 + \mu_2)(\sigma_1 + \sigma_2), \mu_1(\sigma_1 + \sigma_{m+1})\}}{\min\{\min\limits_{i,j\in[m], i\neq j} \frac{1}{2}(\mu_i - \mu_j)(\sigma_i - \sigma_j), \mu_m(\sigma_m - \sigma_{m+1})\}} \\
&= \kappa_g(\text{Hess}_g f(\mathbf{U}^*, \mathbf{V}^*)).
\end{aligned}$$

若 $\{i_1^*, i_2^*\} \subseteq \arg\max_{i\in[m]} \bar{\upsilon}_{i,i+1}(0)$ 对某组 $i_1^* < i_2^*$ 成立, 则根据 $\bar{\upsilon}_{ij}$ 和 $\underline{\upsilon}_{ij}$ 的连续性得知, 存在 $\bar{\delta}_{i_1^*, i_2^*} > 0$, 使得 $\bar{\upsilon}_{i_1^*,i_1^*+1}(\delta) > \bar{\upsilon}_{i_2^*,i_2^*+1}(\delta)$ 和 $\underline{\upsilon}_{i_1^*,i_1^*+1}(\delta) > \underline{\upsilon}_{i_2^*,i_2^*+1}(\delta)$. 其依据在于: 两者在 $\delta = 0$ 时相等, 即 $\bar{\upsilon}_{i_1^*,i_1^*+1}(0) = \bar{\upsilon}_{i_2^*,i_2^*+1}(0)$, $\underline{\upsilon}_{i_1^*,i_1^*+1}(0) = \underline{\upsilon}_{i_2^*,i_2^*+1}(0)$, 且导数对于 $i_2^* < m$ 满足

$$\begin{aligned}
\bar{\upsilon}'_{i_1^*,i_1^*+1}(0) &= -\frac{\bar{\upsilon}_{i_1^*,i_1^*+1}(0)}{2\mu_{i_1^*}\sigma_{i_1^*}\mu_{i_1^*+1}\sigma_{i_1^*+1}} > -\frac{\bar{\upsilon}_{i_2^*,i_2^*+1}(0)}{2\mu_{i_2^*}\sigma_{i_2^*}\mu_{i_2^*+1}\sigma_{i_2^*+1}} = \bar{\upsilon}'_{i_2^*,i_2^*+1}(0), \\
\underline{\upsilon}'_{i_1^*,i_1^*+1}(0) &= -\frac{\underline{\upsilon}_{i_1^*,i_1^*+1}(0)}{2\mu_{i_1^*}\sigma_{i_1^*}\mu_{i_1^*+1}\sigma_{i_1^*+1}} > -\frac{\underline{\upsilon}_{i_2^*,i_2^*+1}(0)}{2\mu_{i_2^*}\sigma_{i_2^*}\mu_{i_2^*+1}\sigma_{i_2^*+1}} = \underline{\upsilon}'_{i_2^*,i_2^*+1}(0),
\end{aligned}$$

以及对于 $i_2^* = m$ 满足

$$\begin{aligned}
\bar{\upsilon}'_{i_1^*,i_1^*+1}(0) &= -\frac{\bar{\upsilon}_{i_1^*,i_1^*+1}(0)}{2\mu_{i_1^*}\sigma_{i_1^*}\mu_{i_1^*+1}\sigma_{i_1^*+1}} > -\frac{\bar{\upsilon}_{m,m+1}(0)}{2\mu_m^2\sigma_m^2} = \bar{\upsilon}'_{m,m+1}(0), \\
\underline{\upsilon}'_{i_1^*,i_1^*+1}(0) &= -\frac{\underline{\upsilon}_{i_1^*,i_1^*+1}(0)}{2\mu_{i_1^*}\sigma_{i_1^*}\mu_{i_1^*+1}\sigma_{i_1^*+1}} > -\frac{\underline{\upsilon}_{m,m+1}(0)}{2\mu_m^2\sigma_m^2} = \underline{\upsilon}'_{m,m+1}(0).
\end{aligned}$$

因此, 存在 $\bar{\delta} \in (0, \min\{\bar{\delta}_{i,j} : i,j \in \arg\max_{i\in[m]} \bar{v}_{i,i+1}(0)\})$ 和 $i^*, j^* \in [m]$, 使得如下的命题成立: 1) $i^* = \arg\max_{i\in[m]} \bar{v}_{i,i+1}(\delta)$; 2) 对所有的 $\delta \in [0, \bar{\delta})$ 有 $j^* = \arg\min_{i\in[m]} \underline{v}_{i,i+1}(\delta)$; 3) $i^* < j^*$. 最终, 我们可以得到

$$
\begin{aligned}
&\kappa_{\text{new}}(\text{Hess}_{\text{new}} f(\mathbf{U}^*, \mathbf{V}^*)) = \frac{\bar{v}_{i^*,i^*+1}(\delta)}{\underline{v}_{j^*,j^*+1}(\delta)} \\
&= \frac{(\mu_{i^*} + \mu_{i^*+1})(\sigma_{i^*} + \sigma_{i^*+1})}{(\mu_{j^*} - \mu_{j^*+1})(\sigma_{j^*} - \sigma_{j^*+1})} \cdot \frac{\sqrt{\mu_{j^*}^2\sigma_{j^*}^2 + 1} + \sqrt{\mu_{j^*+1}^2\sigma_{j^*+1}^2 + 1}}{\sqrt{\mu_{i^*}^2\sigma_{i^*}^2 + 1} + \sqrt{\mu_{i^*+1}^2\sigma_{i^*+1}^2 + 1}} \\
&< \frac{\max\{\frac{1}{2}(\mu_1 + \mu_2)(\sigma_1 + \sigma_2), \mu_1(\sigma_1 + \sigma_{m+1})\}}{\min\{\min\limits_{i,j\in[m],i\neq j} \frac{1}{2}(\mu_i - \mu_j)(\sigma_i - \sigma_j), \mu_m(\sigma_m - \sigma_{m+1})\}} \\
&= \kappa_g(\text{Hess}_g f(\mathbf{U}^*, \mathbf{V}^*)).
\end{aligned}
$$

于是结论得证. □

最后需要指出的是, 参数 $\delta > 0$ 在理论上保证了 (2-22) 的确构成一个黎曼度量；在实际计算中, 我们可以选取一个足够小的正数, 例如 $\delta = 10^{-15}$.

2.3.3　求解典型相关分析的 RGD 和 RCG 方法

通过采用黎曼度量 (2-22), 我们将黎曼梯度法 (算法 1) 和黎曼共轭梯度方法 (算法 2) 应用于求解 CCA 问题, 并于算法 3 与 4 中给出其具体形式.

算法 3 求解 CCA 问题的黎曼梯度法

输入: 赋予度量 (2-22)的流形 $\mathcal{M}$, 初始值 $(\mathbf{U}^{(0)}, \mathbf{V}^{(0)}) \in \mathcal{M}$, $t = 0$.

1: **while** 停机准则未被满足 **do**
2: 　通过 (2-26) 计算 $\eta^{(t)} = -\text{grad}_g f(\mathbf{U}^{(t)}, \mathbf{V}^{(t)})$.
3: 　通过 Armijo 回溯线搜索 (1-19) 计算步长 $s^{(t)}$.
4: 　更新 $\mathbf{U}^{(t+1)} = \Sigma_{xx}^{-1/2}\,\text{qf}(\Sigma_{xx}^{1/2}(\mathbf{U}^{(t)} + \eta_1^{(t)}))$, $\mathbf{V}^{(t+1)} = \Sigma_{yy}^{-1/2}\,\text{qf}(\Sigma_{yy}^{1/2}(\mathbf{V}^{(t)} + \eta_2^{(t)}))$; $t = t + 1$.
5: **end while**

输出: $(\mathbf{U}^{(t)}, \mathbf{V}^{(t)}) \in \mathcal{M}$.

值得注意的是: 1) 我们采用关于 Σ_{xx} 和 Σ_{yy} 的广义 QR 分解 [133] 作为收缩映射, 即对于 $\eta \in \text{T}_{(\mathbf{U},\mathbf{V})}\mathcal{M}$ 有

$$
\text{R}_{(\mathbf{U},\mathbf{V})}(\eta) := (\Sigma_{xx}^{-\frac{1}{2}}\,\text{qf}(\Sigma_{xx}^{\frac{1}{2}}(\mathbf{U} + \eta_1)), \Sigma_{yy}^{-\frac{1}{2}}\,\text{qf}(\Sigma_{yy}^{\frac{1}{2}}(\mathbf{V} + \eta_2))),
$$

其中 $\text{qf}(\mathbf{X})$ 表示 QR 分解 $\mathbf{QR} = \mathbf{X}$ 中的 $\mathbf{Q}$ 因子. 在实际计算中, 收缩映射可以按照 [133] 更高效地计算为 $\text{R}_{(\mathbf{U},\mathbf{V})}(\eta) = ((\mathbf{U} + \eta_1)\mathbf{R}_1^{-1}, (\mathbf{V} + \eta_2)\mathbf{R}_2^{-1})$, 这里 $\mathbf{R}_1^\top\mathbf{R}_1 = (\mathbf{U} + \eta_1)^\top\Sigma_{xx}(\mathbf{U} + \eta_1)$ 和 $\mathbf{R}_2^\top\mathbf{R}_2 = (\mathbf{V} + \eta_2)^\top\Sigma_{yy}(\mathbf{V} + \eta_2)$ 为乔列斯基 (Cholesky) 分解;

算法 4 求解 CCA 问题的黎曼共轭梯度法

输入: 赋予度量 (2-22)的流形 $\mathcal{M}$, 初始值 $(\mathbf{U}^{(0)}, \mathbf{V}^{(0)}) \in \mathcal{M}$, $t = 0$, $\beta^{(0)} = 0$.

1: **while** 停机准则未被满足 **do**

2: 通过 (2-26) 计算 $\eta^{(t)} = -\mathrm{grad}_g f(\mathbf{U}^{(t)}, \mathbf{U}^{(t)}) + \beta^{(t)} \Pi_{g,(\mathbf{U}^{(t)},\mathbf{V}^{(t)})}(\eta^{(t-1)})$.

3: 通过 Armijo 回溯线搜索 (1-19) 计算步长 $s^{(t)}$.

4: 更新 $\mathbf{U}^{(t+1)} = \Sigma_{xx}^{-1/2} \mathrm{qf}(\Sigma_{xx}^{1/2}(\mathbf{U}^{(t)} + \eta_1^{(t)}))$, $\mathbf{V}^{(t+1)} = \Sigma_{yy}^{-1/2} \mathrm{qf}(\Sigma_{yy}^{1/2}(\mathbf{V}^{(t)} + \eta_2^{(t)}))$; $t = t + 1$.

5: **end while**

输出: $(\mathbf{U}^{(t)}, \mathbf{V}^{(t)}) \in \mathcal{M}$.

2) 算法 4 中的向量传输由 (2-23) 中的投影算子定义, 即对于 $\eta \in \mathrm{T}_{(\mathbf{U}^{(t-1)},\mathbf{V}^{(t-1)})}\mathcal{M}$, 向量传输算子定义为 $\mathcal{T}_{t \leftarrow t-1}(\eta) = \Pi_{g,(\mathbf{U}^{(t)},\mathbf{V}^{(t)})}(\eta)$.

2.3.4 数值验证

算法 3 与 4 基于工具箱 Manopt v7.1.0 [134] 实现, 该工具箱是一个流形优化方法的 Matlab 库. 停止准则采用 Manopt 中的默认设置. 所有实验均在一台 MacBook Pro 2019 上完成, 系统为 MacOS Ventura 13.3, 处理器为 2.4 GHz 的 8 核 Intel Core i9, 内存为 32GB, Matlab 版本为 R2020b. 相关代码已公开, 可见 https://github.com/JimmyPeng1998/popman.

表 2-2 CCA 实验中对比的度量.

Table 2-2 Compared metrics in CCA.

算子	(E)	(L1)	(L2)	(L12)	(LR12)
$\bar{\mathcal{H}}_1(\mathbf{U},\mathbf{V})[\eta_1]$	η_1	$\Sigma_{xx}\eta_1$	η_1	$\Sigma_{xx}\eta_1$	$\Sigma_{xx}\eta_1\mathbf{M}_{1,2}$
$\bar{\mathcal{H}}_2(\mathbf{U},\mathbf{V})[\eta_2]$	η_2	η_2	$\Sigma_{yy}\eta_2$	$\Sigma_{yy}\eta_2$	$\Sigma_{yy}\eta_2\mathbf{M}_{2,2}$

我们在不同度量下测试了 RGD 和 RCG 方法的性能, 即在

$$g_{(\mathbf{U},\mathbf{V})}(\xi,\eta) = \langle \xi_1, \bar{\mathcal{H}}_1(\mathbf{U},\mathbf{V})[\eta_1] \rangle + \langle \xi_2, \bar{\mathcal{H}}_2(\mathbf{U},\mathbf{V})[\eta_2] \rangle$$

中选取五种不同的 $\bar{\mathcal{H}}_1, \bar{\mathcal{H}}_2$; 见表 2-2. 欧氏度量记为 “(E)”. “(L1)” 和 “(L2)” 表示仅在 $\mathcal{M} = \mathcal{M}_1 \times \mathcal{M}_2$ 的其中一个分量上引入预条件度量的情形. 文献 [115] 提出的度量 (2-12) 记为 “(L12)”. 我们提出的度量 (2-22) 记为 “(LR12)”, 其在左右两侧同时起到预条件的作用. 我们取参数 $d_x = 800$、$d_y = 400$、$n = 30000$、$m = 5$、$\delta = 10^{-15}$、$\lambda_x = \lambda_y = 10^{-6}$, 并令 $\mathbf{N} = \mathrm{diag}(m, m-1, \ldots, 1)$. 数据矩阵 $\mathbf{X}$ 和 $\mathbf{Y}$ 的元素均独立同分布地采样自区间 $[0,1]$ 上的均匀分布. 方法的性能通过以下指标进行评估: 残差 $(f(\mathbf{U},\mathbf{V}) - f_{\min})$, 梯度范数 “gnorm”, 以及子空间距离 $D(\mathbf{U},\mathbf{U}^*) := \|\mathbf{U}\mathbf{U}^\top - \mathbf{U}^*(\mathbf{U}^*)^\top\|_{\mathrm{F}}$ 和 $D(\mathbf{V},\mathbf{V}^*) := \|\mathbf{V}\mathbf{V}^\top - \mathbf{V}^*(\mathbf{V}^*)^\top\|_{\mathrm{F}}$, 这里 $f_{\min} = f(\mathbf{U}^*,\mathbf{V}^*)$ 和 $(\mathbf{U}^*,\mathbf{V}^*)$ 的定义由 (2-11) 给出.

数值结果汇总于图 2-3、图 2-4 以及表 2-3. 我们有如下的观察: 1) 所提出的度量 (2-22) 的确提升了 RGD 和 RCG 方法的性能, 这是因为该度量更充分地利用了二阶信息；2) 图 2-4 表明, 算法 3 与 4 的单次迭代计算时间与 RGD(L12) 和 RCG(L12) 相当；3) 表 2-3 显示, 与其他方法相比, RGD(LR12) 和 RCG(LR12) 在达到停止准则时所需的迭代次数更少、耗时更短. 所得子空间距离均小于 10^{-8}, 因此可以认为所提出方法生成的序列收敛到了正确的子空间.

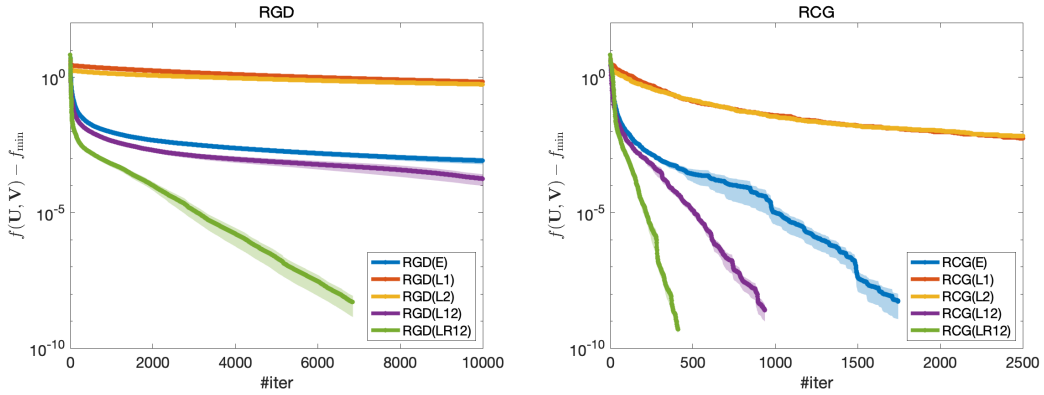


图 2-3 CCA 问题在 $d_x = 800$、$d_y = 400$ 且 $m = 5$ 情形下的数值结果. 左图: RGD. 右图: RCG. 每种方法均进行 10 次独立重复实验.

Figure 2-3 Numerical results for CCA problem for $d_x = 800$, $d_y = 400$, and $m = 5$. Left: RGD. Right: RCG. Each method is tested for 10 runs.

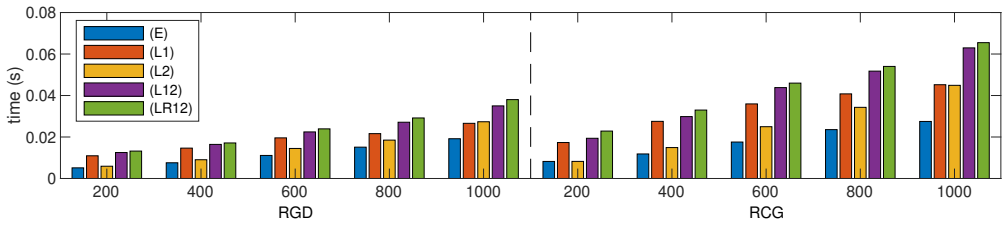


图 2-4 CCA 问题下, 不同度量所导出的 RGD(左图) 和 RCG(右图) 算法的每步迭代时间. 参数设置为 $d_x = 800$, $m = 5$, 以及 $d_y = 200, 400, \ldots, 1000$.

Figure 2-4 Computation time per iteration for RGD (left) and RCG (right) under different metrics for CCA problem for $d_x = 800$, $m = 5$, and $d_y = 200, 400, \ldots, 1000$.

此外, 我们通过 Manopt 中的函数 hessianspectrum 数值计算黎曼 Hessian 算子的条件数: 五种度量下的 $\kappa_g(\mathrm{Hess}_g f(\mathbf{U}^*, \mathbf{V}^*))$ 分别为 $2.10 \cdot 10^4$(E)、$1.43 \cdot 10^7$(L1)、$1.52 \cdot 10^7$(L2)、$1.12 \cdot 10^4$(L12) 以及 $2.38 \cdot 10^3$(LR12). 可以直接验证, 这些数值结果与命题 2.3、2.5 以及命题 2.7 中的理论结论一致. 我们观察到, 在所提出的度量 (LR12) 下, 黎曼 Hessian 算子的条件数在所有度量选择中最小, 这也体现在数值实验中 RGD(LR12) 和 RCG(LR12) 的表现优于其他方法.

表 2-3 CCA 问题在 $d_x = 800$、$d_y = 400$ 且 $m = 5$ 情形下的收敛性结果.

Table 2-3 Convergence results of the CCA problem for $d_x = 800$, $d_y = 400$, and $m = 5$.

度量	方法	迭代步数	时间 (秒)	gnorm	$D(\mathbf{U},\mathbf{U}^*)$	$D(\mathbf{V},\mathbf{V}^*)$	κ_g
(E)	RGD	10000	249.11	5.95e-02	2.69e-05	2.66e-05	2.10e+04
	RCG	1745	31.03	1.70e-05	4.01e-10	3.89e-10	
(L1)	RGD	10000	255.33	1.02e+00	4.12e-04	4.07e-04	1.43e+07
	RCG	2500	74.13	4.94e-02	2.85e-04	2.79e-04	
(L2)	RGD	10000	245.81	8.20e-01	4.13e-04	4.05e-04	1.52e+07
	RCG	2500	56.16	6.90e-02	2.93e-04	2.90e-04	
(L12)	RGD	10000	274.91	4.67e-04	9.68e-07	9.57e-07	1.12e+04
	RCG	937	30.39	8.82e-07	1.68e-09	1.65e-09	
(LR12)	RGD	6607	195.03	1.34e-06	**7.47e-16**	**7.46e-16**	**2.38e+03**
	RCG	**410**	**15.38**	**8.49e-07**	4.63e-09	4.53e-09	

2.4 在截断奇异值分解中的应用

在本节中, 我们考虑截断奇异值分解问题. 具体而言, 给定矩阵 $\mathbf{A} \in \mathbb{R}^{m\times n}$, 当 $p < \min\{m,n\}$ 时, 矩阵 $\mathbf{A}$ 的最大的 p 个奇异值对应的奇异向量 $(\mathbf{U}^*,\mathbf{V}^*)$ 是如下优化问题的全局极小解,

$$\min_{\mathbf{U},\mathbf{V}} f(\mathbf{U},\mathbf{V}) := -\mathrm{tr}(\mathbf{U}^\top\mathbf{A}\mathbf{V}\mathbf{N}),\ \text{s.t.}\ (\mathbf{U},\mathbf{V}) \in \mathcal{M} := \mathrm{St}(p,m)\times\mathrm{St}(p,n), \tag{2-29}$$

其中 $\mathrm{St}(p,m) := \mathbf{U}\in\mathbb{R}^{m\times p} : \mathbf{U}^\top\mathbf{U} = \mathbf{I}_p$ 表示 Stiefel 流形, 且 $\mathbf{N} := \mathrm{diag}\,\mu_1,\ldots,\mu_p$, 满足 $\mu_1 > \mu_2 > \cdots > \mu_p > 0$. Sato 和 Iwai [112] 提出了 RGD 和 RCG 方法来求解问题 (2-29), 其搜索空间配备的是欧氏度量. 我们将所提出的框架应用于求解 (2-29), 通过为 $\mathcal{M}$ 赋予一种非欧氏度量来加速流形优化方法的收敛.

2.4.1 一个新的预条件度量

注意到, 函数 f 在点 $(\mathbf{U},\mathbf{V})$ 处沿着 $\eta = (\eta_1,\eta_2) \in \mathrm{T}_{(\mathbf{U},\mathbf{V})}\mathcal{M}$ 的黎曼 Hessian 算子在欧氏度量下可表示为

$$\begin{aligned}\mathrm{Hess}_{\mathrm{e}} f(\mathbf{U},\mathbf{V})[\eta] = (&\eta_1\mathbf{M}_1 - \mathbf{A}\eta_2\mathbf{N} - \mathbf{U}\,\mathrm{sym}(\mathbf{U}^\top(\eta_1\mathbf{M}_1 - \mathbf{A}\eta_2\mathbf{N})),\\ &\eta_2\mathbf{M}_2 - \mathbf{A}^\top\eta_1\mathbf{N} - \mathbf{V}\,\mathrm{sym}(\mathbf{V}^\top(\eta_2\mathbf{M}_2 - \mathbf{A}^\top\eta_1\mathbf{N}))).\end{aligned}$$

该结果可以参考文献 [112, Proposition 3.5], 其中 $\mathbf{M}_1 := \mathrm{sym}(\mathbf{U}^\top\mathbf{A}\mathbf{V}\mathbf{N})$, $\mathbf{M}_2 := \mathrm{sym}(\mathbf{V}^\top\mathbf{A}^\top\mathbf{U}\mathbf{N})$. 利用黎曼 Hessian 算子的对角块结构, 以及第 2.2.1 节中介绍的左右预条件思想, 我们在 $\mathcal{M}$ 上定义如下新的预条件度量:

$$g_{\mathrm{new},(\mathbf{U},\mathbf{V})}(\xi,\eta) := \langle\xi_1,\eta_1\mathbf{M}_{1,2}\rangle + \langle\xi_2,\eta_2\mathbf{M}_{2,2}\rangle \quad \text{对所有的}\ \xi,\eta\in\mathrm{T}_{(\mathbf{U},\mathbf{V})}\mathcal{M}, \tag{2-30}$$

其中 $\mathbf{M}_{1,2} = (\mathrm{sym}(\mathbf{U}^\top\mathbf{A}\mathbf{V}\mathbf{N})^2 + \delta\mathbf{I}_p)^{1/2}$, $\mathbf{M}_{2,2} = (\mathrm{sym}(\mathbf{V}^\top\mathbf{A}^\top\mathbf{U}\mathbf{N})^2 + \delta\mathbf{I}_p)^{1/2}$, 和 $\delta > 0$. 需要指出的是, 左预条件矩阵被选取为单位矩阵. 关于度量 (2-30) 的投影算子可表示为

$$\Pi_{\mathrm{new},(\mathbf{U},\mathbf{V})}(\bar\eta) = (\bar\eta_1 - \mathbf{U}\mathbf{S}_1\mathbf{M}_{1,2}^{-1}, \bar\eta_2 - \mathbf{V}\mathbf{S}_2\mathbf{M}_{2,2}^{-1}), \tag{2-31}$$

其中 $\bar{\eta} \in \mathrm{T}_{(\mathbf{U},\mathbf{V})}(\mathbb{R}^{m\times p}\times\mathbb{R}^{n\times p}) \simeq \mathbb{R}^{m\times p}\times\mathbb{R}^{n\times p}$, $\mathbf{S}_1$, $\mathbf{S}_2$ 分别是如下 Lyapunov 方程的唯一解 $\mathbf{M}_{1,2}^{-1}\mathbf{S}_1 + \mathbf{S}_1\mathbf{M}_{1,2}^{-1} = 2\,\mathrm{sym}(\mathbf{U}^\top\bar{\eta}_1)$, $\mathbf{M}_{2,2}^{-1}\mathbf{S}_2 + \mathbf{S}_2\mathbf{M}_{2,2}^{-1} = 2\,\mathrm{sym}(\mathbf{V}^\top\bar{\eta}_2)$. 于是, 由命题 2.2与式 (2-31) 可得

$$\mathrm{grad}_{\mathrm{new}} f(\mathbf{U},\mathbf{V}) = (\mathbf{AVNM}_{1,2}^{-1} - \mathbf{US}_1\mathbf{M}_{1,2}^{-1}, \mathbf{A}^\top\mathbf{UNM}_{2,2}^{-1} - \mathbf{VS}_2\mathbf{M}_{2,2}^{-1}). \tag{2-32}$$

上述结果的推导方式与 CCA 情形下命题 2.4 的推导过程完全类似. 需要注意的是, 由于 $\mathbf{M}_{1,2}, \mathbf{M}_{2,2} \in \mathbb{R}^{p\times p}$ 且 $p \ll \min\{m,n\}$, 黎曼梯度 (2-32) 的计算成本与欧氏度量下的计算成本是同阶的.

下面的命题刻画了新度量 (2-30) 的作用效果. 该命题可通过在命题 2.3 与命题 2.5 中令 $\Sigma_{xx} = \mathbf{I}_{d_x}$、$\Sigma_{yy} = \mathbf{I}_{d_y}$、$\Sigma_{xy} = \mathbf{A}$、$d_x = m$ 以及 $d_y = n$ 得到, 证明过程与前述情形类似.

命题 2.8. *设 $\sigma_1 > \sigma_2 > \cdots > \sigma_p > \sigma_{p+1} \geq \cdots \geq \sigma_{\min\{m,n\}}$ 为矩阵 $\mathbf{A}$ 的奇异值, $\mathbf{U}^*$ 和 $\mathbf{V}^*$ 为 $\mathbf{A}$ 的前 p 个左、右奇异向量, 则有*

$$\kappa_{\mathrm{e}}(\mathrm{Hess}_{\mathrm{e}} f(\mathbf{U}^*,\mathbf{V}^*)) = \frac{\max\left\{\frac{1}{2}(\mu_1+\mu_2)(\sigma_1+\sigma_2), \mu_1(\sigma_1+\sigma_{p+1})\right\}}{\min\{\min_{i,j\in[p],i\neq j}\frac{1}{2}(\mu_i-\mu_j)(\sigma_i-\sigma_j), \mu_p(\sigma_p-\sigma_{p+1})\}},$$

$$\kappa_{\mathrm{new}}(\mathrm{Hess}_{\mathrm{new}} f(\mathbf{U}^*,\mathbf{V}^*)) = \frac{\max\{\max\limits_{i,j\in[p],i\neq j}\frac{(\mu_i+\mu_j)(\sigma_i+\sigma_j)}{\sqrt{\mu_i^2\sigma_i^2+\delta}+\sqrt{\mu_j^2\sigma_j^2+\delta}}, \max\limits_{i\in[p]}\frac{\mu_i(\sigma_i+\sigma_{p+1})}{\sqrt{\mu_i^2\sigma_i^2+\delta}}\}}{\min\{\min\limits_{i,j\in[p],i\neq j}\frac{(\mu_i-\mu_j)(\sigma_i-\sigma_j)}{\sqrt{\mu_i^2\sigma_i^2+\delta}+\sqrt{\mu_j^2\sigma_j^2+\delta}}, \min\limits_{i\in[p]}\frac{\mu_i(\sigma_i-\sigma_{p+1})}{\sqrt{\mu_i^2\sigma_i^2+\delta}}\}}.$$

此外, 新度量 (2-30) 的确能够改善黎曼 Hessian 算子的条件数, 即

$$\kappa_{\mathrm{new}}(\mathrm{Hess}_{\mathrm{new}} f(\mathbf{U}^*,\mathbf{V}^*)) \leq \kappa_{\mathrm{e}}(\mathrm{Hess}_{\mathrm{e}} f(\mathbf{U}^*,\mathbf{V}^*)).$$

2.4.2　求解截断奇异值分解的 RGD 和 RCG 方法

我们采用黎曼梯度法 (算法 1) 和黎曼共轭梯度法 (算法 2) 来求解 SVD 问题 (2-29). 通过为 $\mathcal{M}$ 赋予非欧氏度量 (2-30), 我们给出求解 SVD 问题的 RGD 和 RCG 算法, 即算法 5 与算法 6. 这里我们基于 QR 分解给出收缩映射, 即对于 $\eta \in \mathrm{T}_{(\mathbf{U},\mathbf{V})}\mathcal{M}$, 定义 $\mathrm{R}_{(\mathbf{U},\mathbf{V})}(\eta) := (\mathrm{qf}(\mathbf{U}+\eta_1), \mathrm{qf}(\mathbf{V}+\eta_2))$. 在算法 6 中所使用的向量传输由投影算子 (2-31) 给出.

2.4.3　数值验证

我们将算法 5 与算法 6 的性能, 与文献 [112] 中在欧氏度量下的 RGD 与 RCG 方法进行比较. 由于所提出的预条件度量 (2-30) 具有右预条件的效果, 我们将其记为 “(R12)”.

我们设置 $m = 1000$、$n = 500$、$p = 10$, 以及 $\mathbf{N} = \mathrm{diag}(p, p-1, \ldots, 1)$. 矩阵 $\mathbf{A}$ 构造为 $\mathbf{A} = \mathbf{U}^*\Sigma(\mathbf{V}^*)^\top$, 其中 $\mathbf{U}^* \in \mathbb{R}^{m\times p}$ 和 $\mathbf{V}^* \in \mathbb{R}^{n\times p}$ 的元素首先从区间 $[0,1]$ 上的均匀分布中独立同分布采样, 然后通过 QR 分解对 $\mathbf{U}^*$ 和 $\mathbf{V}^*$ 进行正交化. 我们

算法 5 求解截断奇异值分解的 RGD 方法

输入: 赋予度量 (2-30)的流形 $\mathcal{M}$, 初始值 $(\mathbf{U}^{(0)}, \mathbf{V}^{(0)}) \in \mathcal{M}$, $t = 0$.

1: **while** 停机准则未被满足 **do**
2: 通过 (2-32)计算 $\eta^{(t)} = -\mathrm{grad}_g f(\mathbf{U}^{(t)}, \mathbf{U}^{(t)})$.
3: 通过 Armijo 回溯线搜索 (1-19) 计算步长 $s^{(t)}$.
4: 更新 $\mathbf{U}^{(t+1)} = \mathrm{qf}(\mathbf{U}^{(t)} + s^{(t)}\eta_1^{(t)})$, $\mathbf{V}^{(t+1)} = \mathrm{qf}(\mathbf{V}^{(t)} + s^{(t)}\eta_2^{(t)})$; $t = t + 1$.
5: **end while**

输出: $(\mathbf{U}^{(t)}, \mathbf{V}^{(t)}) \in \mathcal{M}$.

算法 6 求解截断奇异值分解的 RCG 方法

输入: 赋予度量 (2-30)的流形 $\mathcal{M}$, 初始值 $(\mathbf{U}^{(0)}, \mathbf{V}^{(0)}) \in \mathcal{M}$, $t = 0$, $\beta^{(0)} = 0$.

1: **while** 停机准则未被满足 **do**
2: 通过 (2-32)计算 $\eta^{(t)} = -\mathrm{grad}_g f(\mathbf{U}^{(t)}, \mathbf{U}^{(t)}) + \beta^{(t)}\Pi_{g,(\mathbf{U}^{(t)},\mathbf{V}^{(t)})}(\eta^{(t-1)})$.
3: 通过 Armijo 回溯线搜索 (1-19) 计算步长 $s^{(t)}$.
4: 更新 $\mathbf{U}^{(t+1)} = \mathrm{qf}(\mathbf{U}^{(t)} + s^{(t)}\eta_1^{(t)})$, $\mathbf{V}^{(t+1)} = \mathrm{qf}(\mathbf{V}^{(t)} + s^{(t)}\eta_2^{(t)})$; $t = t + 1$.
5: **end while**

输出: $(\mathbf{U}^{(t)}, \mathbf{V}^{(t)}) \in \mathcal{M}$.

取 $\Sigma := \mathrm{diag}(1, \gamma, \gamma^2, \ldots, \gamma^{p-1})$ 且 $\gamma = 1/1.5$. RGD 与 RCG 的实现方式与 2.3 节中一致.

数值结果如图 2-5、图 2-6 以及表 2-4 所示. 我们得到与 2.3 节中实验类似的结论. 首先, 由于我们提出的度量更好地利用了二阶信息, 导致新的流形优化方法在迭代次数上显著优于 RGD(E) 与 RCG(E). 其次, 算法 5 与 6 的单步迭代计算时间分别与 RGD(E) 和 RCG(E) 相当. 第三, 表 2-4 表明, 在 RGD(R12) 与 RCG(R12) 中, 子空间距离均小于 10^{-6}, 这说明所提出方法生成的序列收敛到了正确的子空间.

此外, 我们还计算了在两种度量下 $\mathrm{Hess} f(\mathbf{U}^*, \mathbf{V}^*)$ 的条件数. 由 $\mathbf{A}$ 的构造方式以及命题 2.8 可得

$$\kappa(\mathrm{Hess}_{\mathrm{e}} f(\mathbf{U}^*, \mathbf{V}^*)) = \frac{(\mu_1 + \mu_2)(\gamma + 1)}{(\mu_{p-1} - \mu_p)(\gamma^{p-2} - \gamma^{p-1})} = \frac{153389}{63} \approx 2.43 \times 10^3,$$
$$\kappa(\mathrm{Hess}_{\mathrm{new}} f(\mathbf{U}^*, \mathbf{V}^*)) = \frac{(\mu_1 + \mu_2)(1 + \gamma)}{(\mu_1 - \mu_2)(1 - \gamma)} = 95,$$

其结果与表 2-4 中给出的数值结果完全一致. 因此, 更小的条件数表明所提出的方法具有更快的收敛速度.

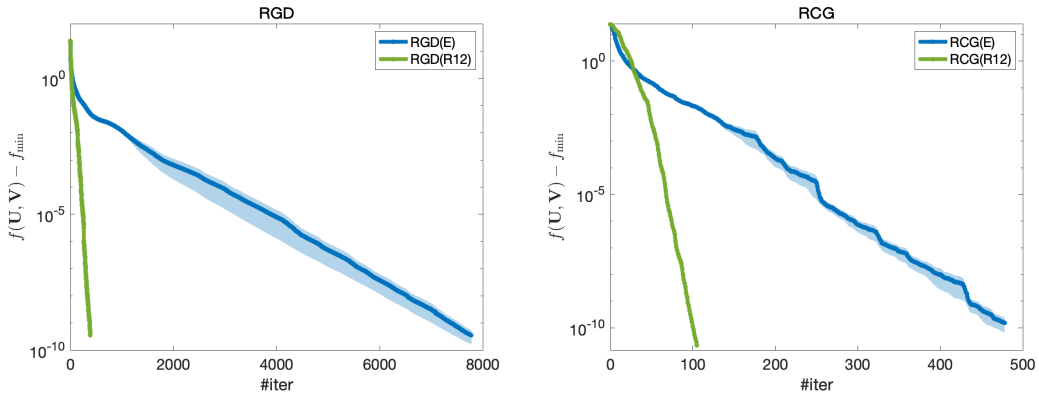


图 2-5 SVD 问题在 $m = 1000$, $n = 500$, 和 $p = 10$ 情形下的数值结果. 左图: RGD. 右图: RCG. 每种方法均进行 10 次独立重复实验.

Figure 2-5 Numerical results for the SVD problem for $m = 1000$, $n = 500$, and $p = 10$. Left: RGD. Right: RCG. Each method is tested for 10 runs.

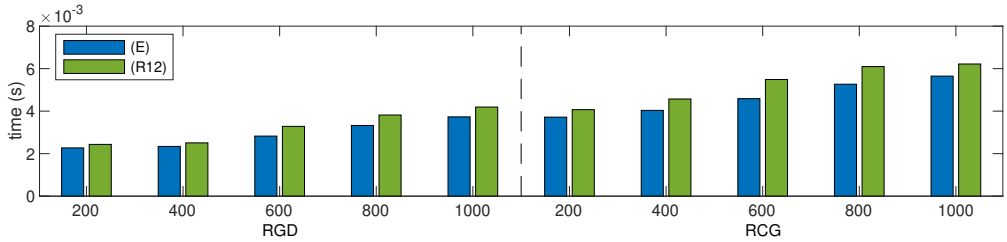


图 2-6 SVD 问题下, 不同度量所导出的 RGD(左图) 和 RCG(右图) 算法的每步迭代时间. 参数设置为 $m = 1000$, $p = 10$, 和 $n = 200, 400, \ldots, 1000$.

Figure 2-6 Average computation time per iteration for RGD (left) and RCG (right) under the Euclidean and proposed metric for $m = 1000$, $p = 10$, and $n = 200, 400, \ldots, 1000$.

表 2-4 SVD 问题在 $m = 1000$, $n = 500$, 和 $p = 10$ 情形下的收敛性结果.

Table 2-4 Convergence results of the SVD problem for $m = 1000$, $n = 500$, and $p = 10$.

度量	方法	迭代步数	迭代时间 (秒)	gnorm	$D(\mathbf{U},\mathbf{U}^*)$	$D(\mathbf{V},\mathbf{V}^*)$	κ_g
(E)	RGD	7781	117.29	9.64e-07	4.53e-05	4.53e-05	2.43e+03
	RCG	478	5.44	8.54e-07	2.00e-05	2.00e-05	
(R12)	RGD	387	3.41	8.72e-07	**2.38e-15**	**1.38e-15**	**9.50e+01**
	RCG	**105**	**1.45**	**7.88e-07**	3.26e-07	3.83e-07	

2.5 在矩阵张量补全中的应用

在本节中, 我们研究矩阵与张量补全问题. 给定一个在索引集 $\Omega \subseteq [n_1] \times [n_2] \times \cdots \times [n_d]$ 上部分元素可知的张量 $\mathcal{A} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$, 张量补全的目标是仅利用 Ω 上的观测恢复原始张量 $\mathcal{A}$ 的所有元素. 需要指出的是, 当 $d = 2$ 时, 该问题即退化为矩阵补全问题.

张量补全问题存在多种不同的建模方式. 其中一类方法基于核范数最小化, 例如 [104, 135]. 这类方法通常需要在完整的张量上进行计算. 相比之下, 基于张量分解的方法利用了张量的低秩结构, 有效降低了搜索空间中的参数量. 因此, 基于张量分解来建模张量补全问题在计算上更加经济. 基于张量分解的张量补全问题是一个定义在乘积流形上的优化问题

$$\min f(x) := \frac{1}{2p} \left\| \mathrm{P}_\Omega(\tau(x) - \mathcal{A}) \right\|_{\mathrm{F}}^2, \ \text{s.t.}\ x \in \mathcal{M} = \mathcal{M}_1 \times \mathcal{M}_2 \times \cdots \times \mathcal{M}_K, \quad (2\text{-}33)$$

这里 $p := |\Omega|/(n_1 n_2 \cdots n_d)$ 表示采样率, P_Ω 表示到索引集合 Ω 上的投影算子, 即对 $\mathcal{X} \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ 而言, 当 $(i_1, \ldots, i_d) \in \Omega$ 时 $\mathrm{P}_\Omega(\mathcal{X})(i_1, \ldots, i_d) = \mathcal{X}(i_1, \ldots, i_d)$, 否则 $\mathrm{P}_\Omega(\mathcal{X})(i_1, \ldots, i_d) = 0$, 以及 $\tau(x)$ 表示由分量 $x_k \in \mathcal{M}_k$ ($k \in [K]$) 确定的张量分解, 其中 $x = (x_1, x_2, \ldots, x_K)$.

由于直接计算欧氏 Hessian 算子 $\nabla^2 f(x)$ 通常较为复杂, Kasai 和 Mishra [96] 在基于 Tucker 分解的张量补全问题中, 引入了一种基于 $\nabla^2 f(x)$ 的分块对角近似的预条件度量. 近年来, 这一思想在其他张量格式的低秩张量近似与补全问题中得到了广泛发展, 例如 [17, 117, 118, 136], 具体可参见表 2-1. 总的来说, 这类度量是通过构造算子 $\bar{\mathcal{H}}(x)$ 得到的, 该算子基于目标函数 $\phi(x) := \frac{1}{2}\|\tau(x) - \mathcal{A}\|_{\mathrm{F}}^2$ 的 Hessian 算子的对角块, 即

$$\bar{\mathcal{H}}(x)[\eta] := (\partial_{11}^2 \phi(x)[\eta_1], \ldots, \partial_{KK}^2 \phi(x)[\eta_K]) \quad \text{对于} \eta = (\eta_1, \eta_2, \ldots, \eta_K) \in \mathrm{T}_x\mathcal{M},$$

这里 $\partial_{kk}^2 \phi(x)[\eta_k] := \lim_{h \to 0} (\partial_k \phi(x_1, \ldots, x_{k-1}, x_k + h\eta_k, x_{k+1}, \ldots, x_K) - \partial_k \phi(x))/h$ 以及 $k \in [K]$. 需要注意的是, 这种预条件策略与 2.2.1节中所讨论的精确分块对角预条件是一致的. 另一方面, 由于 (2-33) 中的目标函数 f 具有最小二乘结构, 我们也可以采用 2.2.3节中介绍的高斯–牛顿型预条件方法来求解问题 (2-33).

2.5.1 针对张量环张量补全问题的高斯–牛顿方法

由于张量环分解在张量补全问题中已被证明是有效的 [17], 我们考虑如下张量环补全问题

$$\min_{\mathcal{U}_k \in \mathbb{R}^{r_{k-1} \times n_k \times r_k}} \quad f(\mathcal{U}_1, \ldots, \mathcal{U}_d) := \frac{1}{2p} \left\| \mathrm{P}_\Omega([\![\mathcal{U}_1, \ldots, \mathcal{U}_d]\!]) - \mathrm{P}_\Omega(\mathcal{A}) \right\|_{\mathrm{F}}^2, \quad (2\text{-}34)$$

其中 $[\![\mathcal{U}_1, \ldots, \mathcal{U}_d]\!]$ 表示张量环分解 [42]. 具体而言, 设 $\mathcal{X} = [\![\mathcal{U}_1, \ldots, \mathcal{U}_d]\!]$ 其中 $\mathcal{U}_k \in \mathbb{R}^{r_{k-1} \times n_k \times r_k}$, $k \in [d]$ 和 $r_0 = r_d$, 则 $\mathcal{X}$ 的第 $(i_1, i_2, \ldots, i_d)$ 个元素定义为

$$\mathcal{X}(i_1, i_2, \ldots, i_d) := \mathrm{tr}(\mathbf{U}_1(i_1)\mathbf{U}_2(i_2) \cdots \mathbf{U}_d(i_d)),$$

这里 $\mathbf{U}_k(i_k) := \mathcal{U}_k(:, i_k, :) \in \mathbb{R}^{r_{k-1} \times r_k}$ 表示张量 $\mathcal{U}_k$ 的第 i_k 个侧向切片, $i_k \in [n_k]$. 问题 (2-34) 可通过引入 1.2节中的记号 $\mathbf{W}_k := (\mathcal{U}_k)_{(2)}$ 和 $\mathbf{W}_{\neq_k} := (\mathcal{U}_{\neq k})_{(2)}$ 进行重写. 因此, 问题 (2-34) 可等价改写为

$$\begin{aligned} \min_{\vec{\mathbf{W}}} \quad & f(\vec{\mathbf{W}}) := \frac{1}{2p} \| \mathrm{P}_\Omega(\tau(\vec{\mathbf{W}}) - \mathcal{A})\|_{\mathrm{F}}^2 \\ \text{s.t.} \quad & \vec{\mathbf{W}} \in \mathcal{M} = \mathbb{R}^{n_1 \times r_0 r_1} \times \mathbb{R}^{n_2 \times r_1 r_2} \times \cdots \times \mathbb{R}^{n_d \times r_{d-1} r_d}, \end{aligned} \tag{2-35}$$

其中映射 τ 定义为 $\tau : \vec{\mathbf{W}} \mapsto [\![\mathrm{ten}_{(2)}(\mathbf{W}_1), \mathrm{ten}_{(2)}(\mathbf{W}_2), \ldots, \mathrm{ten}_{(2)}(\mathbf{W}_d)]\!]$, $\mathrm{ten}_{(2)}(\cdot)$ 表示模态 2 张量化算子.

注意到 (2-35) 中函数 f 具有最小二乘结构, 即: $f(\vec{\mathbf{W}}) = \frac{1}{2}\|F(\vec{\mathbf{W}})\|_{\mathrm{F}}^2$, 这里 $F(\vec{\mathbf{W}}) = \mathrm{P}_\Omega(\tau(\vec{\mathbf{W}}) - \mathcal{A})/\sqrt{p}$ 是一个光滑函数, 我们采用高斯–牛顿型预条件方法求解问题 (2-35). 由于搜索空间 $\mathcal{M}$ 是平直的, 在度量 (2-7) 下的 RGD 方法本质上等价于欧氏高斯–牛顿方法 (参见 [4, §10.3]). 我们在算法 7 中给出了高斯–牛顿方法.

算法 7 求解张量环张量补全的高斯牛顿方法 (TR-GN)

输入: 赋予度量 g 的流形 $\mathcal{M}$, 初始值 $\vec{\mathbf{W}}^{(0)} \in \mathcal{M}$, $t = 0$.

1: **while** 停机准则未被满足 **do**
2: 　　通过求解 (2-8) 计算 $\eta^{(t)}$.
3: 　　更新 $\vec{\mathbf{W}}^{(t+1)} = \vec{\mathbf{W}}^{(t)} + \eta^{(t)}$; $t = t + 1$.
4: **end while**

输出: $\vec{\mathbf{W}}^{(t)} \in \mathcal{M}$.

在算法实现上, 我们回顾算法 7中的搜索方向 $\eta^{(t)}$ 是根据如下的最小二乘问题得到的

$$\arg\min_{\eta \in \mathrm{T}_{\vec{\mathbf{W}}}\mathcal{M}} \|\mathrm{D}F(\vec{\mathbf{W}})[\eta] + F(\vec{\mathbf{W}})\|_{\mathrm{F}}^2. \tag{2-36}$$

根据映射 τ 的多线性性可以得到式 (2-36)中的方向导数 $\mathrm{D}F(\vec{\mathbf{W}})[\eta]$ 为

$$\begin{aligned} \mathrm{D}F(\vec{\mathbf{W}})[\eta] &= \lim_{h \to 0} \frac{\mathrm{P}_\Omega(\tau(\vec{\mathbf{W}} + h\eta) - \mathcal{A}) - \mathrm{P}_\Omega(\tau(\vec{\mathbf{W}}) - \mathcal{A})}{\sqrt{p}h} \\ &= \frac{1}{\sqrt{p}} \sum_{k=1}^{d} \mathrm{P}_\Omega(\tau(\mathbf{W}_1, \ldots, \mathbf{W}_{k-1}, \eta_k, \mathbf{W}_{k+1}, \ldots, \mathbf{W}_d)). \end{aligned}$$

接下来我们得到

$$
\begin{aligned}
&\|\mathrm{D}F(\vec{\mathbf{W}})[\eta] + F(\vec{\mathbf{W}})\|_{\mathrm{F}}^2 \\
&= \frac{1}{p}\sum_{i=1}^{n_1 n_2 \cdots n_d} \langle \mathrm{P}_\Omega(\mathcal{B}_i), \sum_{k=1}^{d} \tau(\mathbf{W}_1, \ldots, \mathbf{W}_{k-1}, \eta_k, \mathbf{W}_{k+1}, \ldots, \mathbf{W}_d) + \tau(\vec{\mathbf{W}}) - \mathcal{A}\rangle^2 \\
&= \frac{1}{p}\sum_{i=1}^{n_1 n_2 \cdots n_d} \Big(\sum_{k=1}^{d} \Big\langle \mathrm{P}_{\Omega_{(k)}}((\mathcal{B}_i)_{(k)})\mathbf{W}_{\neq k}, \eta_k \Big\rangle + \Big\langle \mathrm{P}_\Omega(\mathcal{B}_i), \tau(\vec{\mathbf{W}}) - \mathcal{A} \Big\rangle \Big)^2 \\
&= \frac{1}{p}\sum_{(i_1,\ldots,i_d)\in\Omega} \Big(\sum_{k=1}^{d} \eta_k(i_k, :)^{\mathsf{T}} \mathrm{vec}((\prod_{j=k+1}^{d} \mathbf{U}_j(i_j) \prod_{j=1}^{k-1} \mathbf{U}_j(i_j))^{\mathsf{T}}) + \mathcal{S}(i_1, \ldots, i_d) \Big)^2,
\end{aligned}
$$

这里 $\{\mathcal{B}_i\}_{i=1}^{n_1 n_2 \cdots n_d}$ 的定义为若 $i = \sum_{j=1}^{d}(i_j - 1)\prod_{\ell=1}^{j-1} n_\ell$ 则 $(\mathcal{B}_i)(i_1, i_2, \ldots, i_d) = 1$, 否则 $(\mathcal{B}_i)(i_1, i_2, \ldots, i_d) = 0$, 以及 $\mathcal{S} := \mathrm{P}_\Omega(\tau(\vec{\mathbf{W}}) - \mathcal{A})$ 是残差张量. 值得注意的是对于 $i = \sum_{j=1}^{d}(i_j - 1)\prod_{\ell=1}^{j-1} n_\ell$, 若 $(i_1, i_2, \ldots, i_d) \notin \Omega$ 则有 $\mathrm{P}_{\Omega_{(k)}}((\mathcal{B}_i)_{(k)}) = 0$. 最终, 问题 (2-36) 是一个变量数为 $\sum_{k=1}^{d} n_k r_{k-1} r_k$ 的最小二乘问题.

由于张量环分解的结构较为复杂, 我们将与命题 2.3、命题 2.5 以及命题 2.8 类似的条件数分析结果留待未来工作中展开. 尽管如此, 若高斯–牛顿方法 (算法 7) 生成的序列收敛到某个 $\vec{\mathbf{W}}^* \in \mathcal{M}$ 且满足 $F(\vec{\mathbf{W}}^*) = 0$, 则该高斯–牛顿方法具有超线性收敛性；关于黎曼高斯–牛顿方法可参见 [71, §8.4.1], 关于欧氏高斯–牛顿方法可参见 [4, §10.3].

2.5.2 数值验证

我们将算法 7 与文献 [17] 中的黎曼梯度法 (TR-RGD) 和黎曼共轭梯度法 (TR-RCG) 进行比较. 后两种方法采用的度量是 $g_{\vec{\mathbf{W}}}(\xi, \eta) := \sum_{k=1}^{d}\langle \xi_k, \eta_k(\mathbf{W}_{\neq k}^{\mathsf{T}}\mathbf{W}_{\neq k} + \delta \mathbf{I}_{r_{k-1}r_k})\rangle$ 其中 $\xi, \eta \in \mathrm{T}_{\vec{\mathbf{W}}}\mathcal{M}$, 以及 $\delta > 0$ 为常数. TR-RGD、TR-RCG 以及 TR-GN 方法的代码均公开于 https://github.com/JimmyPeng1998/LRTCTR.

张量 $\mathcal{A} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$ 由 $\mathcal{A} = \tau(\vec{\mathbf{W}}^*)$ 构造, 其中 $\vec{\mathbf{W}}^* \in \mathcal{M}$ 的各个元素均从区间 $[0, 1]$ 上的均匀分布中采样. 初始值 $\vec{\mathbf{W}}^{(0)} \in \mathcal{M}$ 采用相同方式生成. 给定采样率 p, 我们从 $[n_1] \times [n_2] \times \cdots \times [n_d]$ 中随机选取 $p n_1 n_2 \cdots n_d$ 个样本以构造采样集合 Ω. 我们在实验中设置 $d = 3$, $n_1 = n_2 = n_3 = 100$, $p = 0.05$, 张量环秩参数取 $\mathbf{r}^* = (1, 1, 1), (2, 2, 2), \ldots, (8, 8, 8)$, 并令 $\delta = 10^{-15}$.

我们采用所有方法的默认参数设置. TR-RGD 方法与 TR-RCG 方法的步长策略均采用 Armijo 回溯线搜索 (1-19). 共轭梯度参数选取为改进的 Hestenes–Stiefel 规则在黎曼情形下的版本 [134]. 式 (1-19) 中的参数设置为 $\rho = 0.3$, $a = 2^{-13}$, $s_0 = 1$. 每种方法的性能通过训练误差 $\varepsilon_\Omega(\vec{\mathbf{W}}^{(t)}) := \|\mathrm{P}_\Omega(\tau(\vec{\mathbf{W}}^{(t)})) - \mathrm{P}_\Omega(\mathcal{A})\|_{\mathrm{F}} / \|\mathrm{P}_\Omega(\mathcal{A})\|_{\mathrm{F}}$ 以及测试误差 $\varepsilon_\Gamma(\vec{\mathbf{W}}^{(t)})$ 来评估, 其中 Γ 为与 Ω 不同的测试集合, 且设置 $|\Gamma| = 100$. 当满足以下任一停机准则时算法终止: 1) 训练误差 $\varepsilon_\Omega(\vec{\mathbf{W}}^{(t)}) < 10^{-14}$; 2) 达到最大迭代次数 1000; 3) 相对变化率 $|(\varepsilon_\Omega(\vec{\mathbf{W}}^{(t)}) - \varepsilon_\Omega(\vec{\mathbf{W}}^{(t-1)}))/\varepsilon_\Omega(\vec{\mathbf{W}}^{(t-1)})| < \varepsilon$; 4) 步长 $s^{(t)} < 10^{-10}$.

数值结果如图 2-7 和图 2-8 所示. 一方面, 我们观察到 TR-GN 方法比 TR-RGD 和 TR-RCG 具有更快的收敛速度, 这是因为 TR-GN 利用了 $\nabla^2 f(\vec{\mathbf{W}})$ 的更多二阶信息, 而 TR-RGD 和 TR-RCG 中的预条件度量仅利用了 Hessian 算子的对角块信息. 另一方面, 图 2-8 表明, 随着 TR 秩 $\mathbf{r}^*$ 的增大, TR-GN 方法达到停止准则所需的计算时间增长速度快于 TR-RGD 和 TR-RCG.

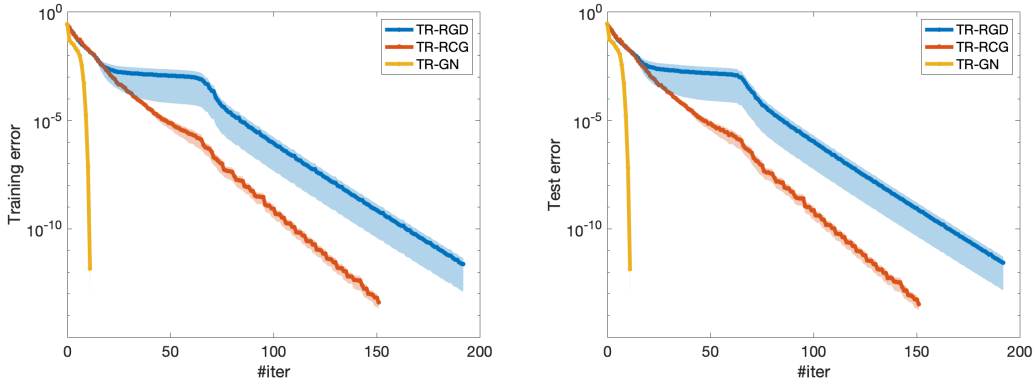


图 2-7 当 TR 秩参数为 $\mathbf{r}^* = (5,5,5)$ 时的训练误差和测试误差. 每种方法均进行 10 次独立重复实验.

Figure 2-7 Training and test errors for TR rank $\mathbf{r}^* = (5,5,5)$. Each method is tested for 10 runs.

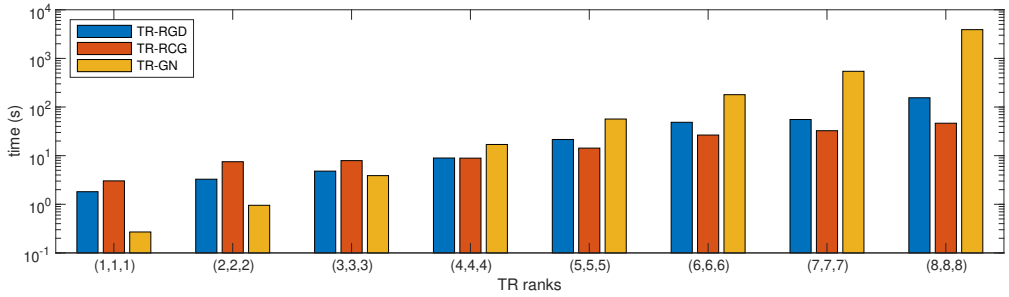


图 2-8 当 TR 秩 $\mathbf{r}^* = (1,1,1),(2,2,2),\ldots,(8,8,8)$ 时, 各方法达到停止准则所需的计算时间.

Figure 2-8 Computation time required for each method to reach the stopping criteria for TR rank $\mathbf{r}^* = (1,1,1),(2,2,2),\ldots,(8,8,8)$.

2.6　本章小结

不同的黎曼度量, 能导出不同的黎曼梯度, 进而影响流形优化方法的性能. 我们通过理论与实验证明, 精心构造的度量确实有助于加速黎曼方法. 具体而言, 我们提出了一个求解乘积流形上优化问题的一般框架, 该框架使用了预条件度量, 并提出了三种具体方法来构造算子以近似黎曼 Hessian 算子. 从概念上讲, 包括黎曼高斯牛顿方法以及数值线性代数中的块 Jacobi 预条件法在内的多种现有方法, 都可以通过选取特定度量的方式, 利用所提出的框架进行解释. 在本文中, 我们提出了三类预条件子的构造方式, 即精确预条件子、左右预条件子以及高斯–牛顿

预条件子. 在实际应用中, 预条件子的选择需结合具体问题进行权衡分析. 一般而言, 若精确预条件子或高斯–牛顿预条件子的计算代价在可接受范围内, 我们优先推荐采用这两类方法; 而当其计算开销较大时, 则更建议采用基于矩阵结构的左右预条件子. 该类预条件子作用于矩阵 (或张量) 的各个模态, 在有效改善问题条件数的同时, 不会额外引入过大的计算代价.

我们基于所提出的框架, 为典型相关分析和截断奇异值分解设计了新的预条件度量, 并通过计算局部极小点处黎曼 Hessian 算子的条件数验证了所提出度量的效果, 理论结果显示条件数确实得到了改善. 数值结果进一步验证, 精心设计的度量确实能够提升流形优化方法的性能.

第 3 章　求解张量环格式张量补全问题的黎曼预条件方法

3.1　引言

张量补全问题作为矩阵补全问题的自然推广, 其目标是基于部分观测到的张量元素来恢复原始张量. 在实际应用中, 从真实场景中采集的数据通常被认为具有潜在的低秩结构. 因此, 在矩阵补全问题中, 低秩矩阵分解方法被广泛采用, 以降低计算成本并节省存储开销. 基于同样的思想, 低秩张量分解在张量补全问题中发挥着重要作用, 其应用已广泛存在于多个领域中, 例如推荐系统 [96, 117]、图像处理 [135], 以及高维函数的插值问题 [60, 137].

在本章中, 我们考虑具有有界张量环秩 $\mathbf{r} := (r_1, r_2, \dots, r_d)$ 的张量补全问题. 给定一个仅在指标集 $\Omega \subseteq [n_1] \times \cdots \times [n_d]$ 上元素可知的 d 阶张量 $\mathcal{A} \in \mathbb{R}^{n_1 \times \cdots \times n_d}$, 这里 $k \in [d]$ 以及 $d \geq 3$ 是一个正整数. 张量补全问题建模如下:

$$\begin{aligned} \min \quad & \tfrac{1}{2} \left\| \mathrm{P}_\Omega(\llbracket \mathcal{U}_1, \mathcal{U}_2, \dots, \mathcal{U}_d \rrbracket) - \mathrm{P}_\Omega(\mathcal{A}) \right\|_{\mathrm{F}}^2 \\ \text{s.t.} \quad & (\mathcal{U}_1, \mathcal{U}_2, \dots, \mathcal{U}_d) \in \mathcal{M}_{\mathcal{U}}, \end{aligned} \tag{3-1}$$

这里 $\llbracket \mathcal{U}_1, \mathcal{U}_2, \dots, \mathcal{U}_d \rrbracket \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ 具有核张量为 $\mathcal{U}_k \in \mathbb{R}^{r_k \times n_k \times r_{k+1}}$ 的张量环分解; 具体定义详见 1.6. P_Ω 为往 Ω 上的投影, 也就是说, 若 $(i_1, \dots, i_d) \in \Omega$, 则 $\mathrm{P}_\Omega(\mathcal{X})(i_1, \dots, i_d) = \mathcal{X}(i_1, \dots, i_d)$, 否则 $\mathrm{P}_\Omega(\mathcal{X})(i_1, \dots, i_d) = 0$. 问题 (3-1) 的搜索空间为张量空间的乘积空间, 也就是

$$\mathcal{M}_{\mathcal{U}} := \mathbb{R}^{r_1 \times n_1 \times r_2} \times \mathbb{R}^{r_2 \times n_2 \times r_3} \times \cdots \times \mathbb{R}^{r_d \times n_d \times r_1}.$$

相关的研究与动机　张量补全问题具有多种不同的建模方式, 其中一类是基于核范数最小化的方法. 在矩阵补全问题中, 通常通过最小化矩阵的核范数 (它是矩阵秩的一个凸松弛) 来求解. Liu 等人 [135] 将 “核范数” 的概念推广到张量情形, 通过计算张量各个展平矩阵的核范数之和来定义张量核范数, 并采用交替方向乘子法来求解张量补全问题. 鉴于实际观测到的张量数据往往受到噪声扰动, Zhao 等人 [138] 针对基于管状核范数 (tubal nuclear norm) 的鲁棒张量补全问题进行了研究, 并提出了一种近端主要化–最小化 (proximal majorization–minimization) 算法. 上述算法通常需要以完整大小存储张量, 其参数数量为 $n_1 n_2 \cdots n_d$, 随张量阶数 d 呈指数级增长.

与直接处理完整大小的张量相比, 基于张量分解的方法 [38] 能够利用张量补全问题中的低秩结构, 从而显著减少搜索空间中的参数量并节约存储. 考虑到张量分解所具有的块结构, 可以通过在固定其余块的情况下更新某一个块来构造交替最小化方法. Jain 和 Oh [139] 针对对称三阶张量的 CP 分解提出了一种交替最小化方法. 对于 Tucker 分解, 交替最小化方法, 也称为交替最小二乘 (alternating least squares, ALS) 算法, 由 Andersson 和 Bro [140] 进行了研究, 其中每一步的

子问题均为一个最小二乘问题. 张量链分解 [41, 141], 在计算物理中也被称为矩阵乘积态 (matrix product states, MPS) [20, 59], 将一个张量分解为 d 个核张量. Grasedyck 等人 [142] 提出了用于张量链分解补全问题的交替方向拟合算法. 近年来, Zhao 等人 [42] 提出了张量环分解, 作为张量链分解的推广形式; 在计算物理中, 它也被称为具有周期边界条件的 MPS. 文献 [42] 中提出了用于求解 TR 分解问题的 ALS 算法, 并在后续工作中将其应用于张量补全问题 [143]. 总体而言, 当参数数量随张量维数 d 线性增长时, ALS 方法已被证明在张量补全问题中是有效的. 然而, 在实际应用中, 这类方法可能会面临过拟合的问题, 并且对初始化较为敏感 [144].

近年来, 基于张量流形的黎曼优化方法在求解张量补全问题中展现出良好的前景. 由于搜索空间本身是一个流形, 可以利用丰富的几何工具, 在流形上构造高效的优化算法, 并能够保证算法的收敛性; 相关综述可参见 [70, 71]. Acar 等人 [145] 基于 CP 分解提出了一种用于张量补全的欧氏非线性共轭梯度方法. Dong 等人 [117] 在 CP 分解导出的矩阵乘积空间上引入黎曼度量, 提出了相应的黎曼梯度方法和黎曼共轭梯度方法. 针对 Tucker 分解下的张量补全问题, Kressner 等人 [15] 在固定秩张量流形上提出了黎曼共轭梯度方法. Kasai 和 Mishra [96] 进一步引入了一种预条件黎曼度量, 并研究了基于 Tucker 分解的商流形几何结构, 同时提出了相应的黎曼共轭梯度法. 基于固定 TT 秩张量流形的几何性质, Steinlechner [60] 提出了黎曼共轭梯度法. 进一步地, Cai 等人 [118] 系统研究了该流形上的商几何结构, 并提出了黎曼梯度、黎曼共轭梯度以及黎曼高斯–牛顿算法.

张量环分解是张量链分解的一种推广. 一方面, TR 分解提供了更加灵活的张量秩选择. 具体而言, TR 分解中各个核张量的展平矩阵不必是满秩的, 而在 TT 分解中, 核张量的模态 1 与模态 3 展平矩阵必须满秩. 此外, TR 分解通过将 TT 分解中首尾核张量的秩约束从 1 放宽为任意正整数, 即从 TT 秩 $(1, r_2, \dots, r_d, 1)$ 放宽至 TR 秩 $(r_1, r_2, \dots, r_d)$, 从而允许在模态 1 与模态 d 方向上更充分地挖掘张量中的信息, 实现更强的表达力 [146]. 另一方面, 尽管具有有界 TR 秩的张量集合并不构成 $\mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$ 的一个黎曼子流形, TR 分解仍然保留了与 TT 分解类似的块结构. 这一结构使我们能够为搜索空间赋予流形结构, 并构造具有预条件效果的黎曼度量. 具体而言, 我们通过对问题 (3-1) 进行等价重构, 在如下由展平矩阵组成的乘积空间上发展基于 TR 分解的张量补全问题的黎曼优化方法：

$$\mathcal{M} := \mathbb{R}^{n_1 \times r_1 r_2} \times \mathbb{R}^{n_2 \times r_2 r_3} \times \cdots \times \mathbb{R}^{n_{d-1} \times r_{d-1} r_d} \times \mathbb{R}^{n_d \times r_d r_1}.$$

通过在 $\mathcal{M}$ 上引入一个非欧度量, 我们将该搜索空间 $\mathcal{M}$ 视为一个黎曼流形.

本章主要内容 我们基于张量环分解对张量补全问题进行建模, 其搜索空间为由 TR 分解中各个核张量的模态 2 展平矩阵所组成的乘积空间. 我们在该空间上设计了具有预条件效果的黎曼度量, 并提出了黎曼梯度法和黎曼共轭梯度法来求解

张量补全问题. 我们证明了由所提出算法生成的序列的任一聚点都是一个稳定点. 在计算 (黎曼) 梯度的过程中, 往往涉及大规模矩阵的构造与乘法运算, 其计算复杂度随张量阶数 d 呈指数级增长, 在实际应用中难以承受. 为提升所提出算法的计算效率, 我们利用 TR 分解的结构, 提出了一种计算复杂度关于阶数 d 多项式级别的高效算法来计算 (黎曼) 梯度.

我们在多种人工合成数据和真实数据集上, 将所提出的算法与现有方法在不同张量补全任务中进行了比较, 这些数据集包括电影评分数据、高光谱图像以及高维函数等. 数值实验结果表明, 基于 TR 分解的方法在恢复性能上优于基于其他分解的算法. 此外, 在 TR 分解框架下, 所提出的算法在性能上优于交替最小二乘算法.

3.2　张量环格式张量补全问题的等价刻画

在本节中, 我们对基于 TR 的张量补全问题 (3-1) 进行等价刻画. 回顾定义 1.7 中子链张量的性质, 我们可以将基于 TR 的张量问题表述在模态 2 展平矩阵 $\mathbf{W}_1, \dots, \mathbf{W}_d$ 上. 具体而言, 一个 TR 张量 $\mathcal{X} = [\![\mathcal{U}_1, \dots, \mathcal{U}_d]\!]$ 的更新过程涉及对各个核张量 $\mathcal{U}_1, \dots, \mathcal{U}_d$ 的更新, 而通过矩阵化操作 $\mathbf{W}_k = (\mathcal{U}_k)_{(2)}$ 这等价于更新对应的模态 2 展平矩阵 $\mathbf{W}_1, \dots, \mathbf{W}_d$. 换言之, 模态 2 展平矩阵足以承载所有与张量相关的计算. 值得注意的是, 在给定 $\mathbf{W}_1, \dots, \mathbf{W}_d$ 的情况下, 对于任意 $j \neq k$, 矩阵 $\mathbf{W}_k \mathbf{W}_{\neq k}^\top$ 与 $\mathbf{W}_j \mathbf{W}_{\neq j}^\top$ 表示的是同一个张量在不同模态下的展平矩阵. 在实际计算中, 我们从不显式地计算 $\mathbf{W}_k \mathbf{W}_{\neq k}^\top$.

随后, 张量补全问题 (3-1) 可以被重新表述为定义在 TR 分解中各个核张量的模态 2 展平矩阵所构成的乘积空间上的优化问题, 即

$$\mathcal{M} = \mathbb{R}^{n_1 \times r_1 r_2} \times \mathbb{R}^{n_2 \times r_2 r_3} \times \cdots \times \mathbb{R}^{n_{d-1} \times r_{d-1} r_d} \times \mathbb{R}^{n_d \times r_d r_1}.$$

在该乘积空间 $\mathcal{M}$ 上, Frobenius 范数定义为 $\|\vec{\mathbf{W}}\|_{\mathrm{F}} := \sqrt{\sum_{k=1}^d \|\mathbf{W}_k\|_{\mathrm{F}}^2}$, 其中 $\vec{\mathbf{W}} = \left(\mathbf{W}_1, \dots, \mathbf{W}_d\right) \in \mathcal{M}$.

在本文中, 我们关注如下张量补全问题：

$$\min_{\vec{\mathbf{W}} = (\mathbf{W}_1, \dots, \mathbf{W}_d) \in \mathcal{M}} f(\vec{\mathbf{W}}) := f_\Omega(\vec{\mathbf{W}}) + r(\vec{\mathbf{W}}). \tag{3-2}$$

目标函数 f 由两部分组成. 第一部分为残差函数

$$f_\Omega(\vec{\mathbf{W}}) := \frac{1}{2p} \left\| \mathrm{P}_\Omega(\mathcal{X}) - \mathrm{P}_\Omega(\mathcal{A}) \right\|_{\mathrm{F}}^2 = \frac{1}{2p} \left\| \mathrm{P}_{\Omega_{(k)}} \left(\mathbf{W}_k \mathbf{W}_{\neq k}^\top - \mathbf{A}_{(k)} \right) \right\|_{\mathrm{F}}^2,$$

其中 $p := |\Omega|/(n_1 n_2 \cdots n_d)$ 为采样率, $\mathbf{A}(k)$ 表示张量 $\mathcal{A}$ 的模态 k 展平矩阵, $\Omega_{(k)}$ 为采样集合 Ω 的模态 k 展平矩阵; 第二部分为正则化项 $r(\vec{\mathbf{W}})$, 我们选取

$$r(\vec{\mathbf{W}}) := \frac{\lambda}{2} \|\vec{\mathbf{W}}\|_{\mathrm{F}}^2,$$

该形式与最大间隔矩阵分解 (Maximum-margin Matrix Factorization) [147] 中的正则化方式类似, 其中 $\lambda > 0$ 为正则化参数. 该正则化项使变量 $\vec{\mathbf{W}}$ 保持在 $\mathcal{M}$ 的一个紧子集中, 并保证算法的收敛性; 相关分析可参见第 3.4节.

由于在给定 $n_1, n_2, \ldots, n_d$ 以及 $r_1, r_2, \ldots, r_d$ 的情况下, 张量化算子 $\mathrm{ten}_{(2)}(\cdot)$ 是可逆的, 因此, 问题 (3-2) 中的搜索空间与原问题 (3-1) 中的搜索空间是等价的. 这两个搜索空间分别通过矩阵化与张量化操作相互关联:即对于 $(\mathbf{W}_1, \mathbf{W}_2, \ldots, \mathbf{W}_d) \in \mathcal{M}$ 与 $(\mathcal{U}_1, \mathcal{U}_2, \ldots, \mathcal{U}_d) \in \mathcal{M}_{\mathcal{U}}$, 有矩阵化 $\mathbf{W}_k = (\mathcal{U}_k)_{(2)}$ 和张量化 $\mathcal{U}_k = \mathrm{ten}_{(2)}(\mathbf{W}_k)$, 也就是说,

$$\begin{array}{ccc} (\mathcal{U}_1, \ldots, \mathcal{U}_d) \in \mathcal{M}_{\mathcal{U}} & \xrightarrow[\text{张量化}]{\mathbf{W}_k = (\mathcal{U}_k)_{(2)}} & \mathcal{M} \ni (\mathbf{W}_1, \ldots, \mathbf{W}_d) \\ \mathbb{R}^{r_1 \times n_1 \times r_2} \times \cdots \times \mathbb{R}^{r_d \times n_d \times r_1} & \xleftarrow{\mathcal{U}_k = \mathrm{ten}_{(2)}(\mathbf{W}_k)} & \mathbb{R}^{n_1 \times r_1 r_2} \times \cdots \times \mathbb{R}^{n_d \times r_d r_1} \end{array}.$$

此外, 我们注意到, 乘积空间 $\mathcal{M}$ 中的一个元素 $\vec{\mathbf{W}} = \left(\mathbf{W}_1, \ldots, \mathbf{W}_d\right)$ 也可以通过 TR 分解表示为一个张量 $[\![\mathrm{ten}_{(2)}(\mathbf{W}_1), \ldots, \mathrm{ten}_{(2)}(\mathbf{W}_d)]\!] \in \mathbb{R}^{n_1 \times \cdots \times n_d}$. 基于此, 我们定义映射

$$\tau : \mathcal{M} \to \mathbb{R}^{n_1 \times \cdots \times n_d},\ \tau(\vec{\mathbf{W}}) := [\![\mathrm{ten}_{(2)}(\mathbf{W}_1), \ldots, \mathrm{ten}_{(2)}(\mathbf{W}_d)]\!], \tag{3-3}$$

该映射将在第 3.5 节数值实验中用于生成人工合成数据.

给定目标函数 f, 其关于 $\mathbf{W}_k$ 的一阶导数具有如下形式:

$$\partial_{\mathbf{W}_k} f(\vec{\mathbf{W}}) = \partial_{\mathbf{W}_k} f_\Omega(\vec{\mathbf{W}}) + \partial_{\mathbf{W}_k} r(\vec{\mathbf{W}}) = \frac{1}{p}\mathbf{S}_{(k)}\mathbf{W}_{\neq k} + \lambda \mathbf{W}_k \quad \text{对所有的 } k \in [d],$$

其中 $\mathcal{S} := \mathrm{P}_\Omega(\tau(\vec{\mathbf{W}})) - \mathrm{P}_\Omega(\mathcal{A})$ 称为残差张量, $\mathbf{S}_{(k)}$ 表示张量 $\mathcal{S}$ 的模态 k 展平矩阵. 因此, f_Ω 在 $\vec{\mathbf{W}} \in \mathcal{M}$ 处的欧氏梯度可表示为

$$\nabla f_\Omega(\vec{\mathbf{W}}) = \left(\frac{1}{p}\mathbf{S}_{(1)}\mathbf{W}_{\neq 1}, \frac{1}{p}\mathbf{S}_{(2)}\mathbf{W}_{\neq 2}, \ldots, \frac{1}{p}\mathbf{S}_{(d)}\mathbf{W}_{\neq d}\right). \tag{3-4}$$

此外, 还可以采用欧氏梯度下降类算法 (例如 [148]) 来求解补全问题 (3-2). 近年来, 研究者开始关注黎曼预条件算法, 即在搜索空间上引入非欧氏度量. 该度量的构造旨在通过目标函数 Hessian 算子的 “对角块” 来对其进行近似. 这类算法能够显著提升基于欧氏方法的性能, 并已成功应用于矩阵与张量补全问题中 (例如 [96, 117, 118, 121]). 然而, 将上述方法直接推广到 TR 分解情形并非易事, 因为这通常涉及大规模矩阵的构造与计算. 在下一节中, 我们将探讨如何针对张量补全问题 (3-2) 设计高效的预条件算法.

3.3 张量补全算法

我们首先在流形 $\mathcal{M}$ 上构造一种预条件黎曼度量, 并在该度量下推导相应的黎曼梯度. 随后, 提出黎曼梯度下降算法和黎曼共轭梯度法. 最后, 给出一种高效计算黎曼梯度的实现流程.

3.3.1　一个新的预条件度量

在 $\mathcal{M}$ 上构造预条件度量的核心思想, 是利用目标函数 f_Ω 的二阶信息, 从而构造一个近似牛顿方向的搜索方向. 具体而言, 我们希望构造一个算子 $\mathcal{H}(\vec{\mathbf{W}})$: $\mathrm{T}_{\vec{\mathbf{W}}}\mathcal{M} \to \mathrm{T}_{\vec{\mathbf{W}}}\mathcal{M}$ 使得

$$\langle \mathcal{H}(\vec{\mathbf{W}})[\vec{\xi}], \vec{\eta} \rangle \approx \nabla^2 f_\Omega(\vec{\mathbf{W}})[\vec{\xi}, \vec{\eta}] \tag{3-5}$$

对任意 $\vec{\xi}, \vec{\eta} \in \mathrm{T}_{\vec{\mathbf{W}}}\mathcal{M} \simeq \mathcal{M}$ 成立. 其中, $\mathrm{T}_{\vec{\mathbf{W}}}\mathcal{M}$ 表示 $\vec{\mathbf{W}} \in \mathcal{M}$ 处的切空间, $\nabla^2 f_\Omega$ 表示 f_Ω 的欧氏 Hessian 算子. 注意到

$$\begin{aligned}\mathrm{T}_{\vec{\mathbf{W}}}\mathcal{M} &= \mathrm{T}_{\mathbf{W}_1}\mathbb{R}^{n_1 \times r_1 r_2} \times \cdots \times \mathrm{T}_{\mathbf{W}_{d-1}}\mathbb{R}^{n_{d-1} \times r_{d-1} r_d} \times \mathrm{T}_{\mathbf{W}_d}\mathbb{R}^{n_d \times r_d r_1} \\ &= \mathbb{R}^{n_1 \times r_1 r_2} \times \cdots \times \mathbb{R}^{n_{d-1} \times r_{d-1} r_d} \times \mathbb{R}^{n_d \times r_d r_1}.\end{aligned}$$

因此, 一个切向量 $\vec{\xi} \in \mathrm{T}_{\vec{\mathbf{W}}}\mathcal{M}$ 可以表示为 $\vec{\xi} = (\xi_1, \xi_2, \ldots, \xi_d)$ 这里 $\xi_k \in \mathbb{R}^{n_k \times r_k r_{k+1}}$.

为此, 我们首先计算欧氏 Hessian 算子的显式形式. 对于任意 $\vec{\xi}, \vec{\eta} \in \mathrm{T}_{\vec{\mathbf{W}}}\mathcal{M}$, 有

$$\nabla^2 f_\Omega(\vec{\mathbf{W}})[\vec{\xi}, \vec{\eta}] = \sum_{k=1}^{d} \langle \partial^2_{\mathbf{W}_k, \mathbf{W}_k} f_\Omega(\vec{\mathbf{W}})[\vec{\xi}], \vec{\eta} \rangle + \sum_{\ell, m=1, \ell \neq m}^{d} \langle \partial^2_{\mathbf{W}_\ell, \mathbf{W}_m} f_\Omega(\vec{\mathbf{W}})[\vec{\xi}], \vec{\eta} \rangle.$$

通过直接计算, $\nabla^2 f_\Omega(\vec{\mathbf{W}})$ 的 "对角块" 具有如下形式:

$$\partial^2_{\mathbf{W}_k, \mathbf{W}_k} f_\Omega(\vec{\mathbf{W}})[\vec{\xi}] = \frac{1}{p} \mathrm{P}_{\Omega_{(k)}} \left(\xi_k \mathbf{W}_{\neq k}^\top \right) \mathbf{W}_{\neq k} \quad \text{对所有的} k \in [d]. \tag{3-6}$$

由于计算 "非对角块" $\partial^2_{\mathbf{W}_\ell, \mathbf{W}_m} f_\Omega(\vec{\mathbf{W}})$ 的形式较为复杂, 为了在精度与计算量之间取得平衡, 我们仅保留 "对角块" 来构造算子 $\mathcal{H}(\vec{\mathbf{W}})$. 一种直观的构造方式是直接利用上述二阶导数, 即定义

$$\mathcal{H}_\Omega(\vec{\mathbf{W}})[\vec{\xi}] := \left(\partial^2_{\mathbf{W}_1, \mathbf{W}_1} f_\Omega(\vec{\mathbf{W}})[\vec{\xi}], \ldots, \partial^2_{\mathbf{W}_d, \mathbf{W}_d} f_\Omega(\vec{\mathbf{W}})[\vec{\xi}] \right),$$

该算子依赖于具体的采样集合 Ω. 然而, 我们希望构造一个能够适用于具有子采样结构的一类张量补全问题的算子, 因此对式 (3-6) 取期望. 更具体地, 我们假设采样集合 Ω 中的索引是以概率 p 独立同分布地服从伯努利分布. 通过对 "对角块" 关于 Ω 取期望, 可以消除投影算子 P_Ω. 由此, 我们定义算子

$$\begin{aligned}\mathcal{H}(\vec{\mathbf{W}})[\vec{\xi}] &:= \left(\mathbb{E}_\Omega \left[\partial^2_{\mathbf{W}_1, \mathbf{W}_1} f_\Omega(\vec{\mathbf{W}})[\vec{\xi}] \right], \ldots, \mathbb{E}_\Omega \left[\partial^2_{\mathbf{W}_d, \mathbf{W}_d} f_\Omega(\vec{\mathbf{W}})[\vec{\xi}] \right] \right) \\ &= \left(\xi_1 \mathbf{W}_{\neq 1}^\top \mathbf{W}_{\neq 1}, \ldots, \xi_d \mathbf{W}_{\neq d}^\top \mathbf{W}_{\neq d} \right).\end{aligned} \tag{3-7}$$

需要注意的是, $\mathcal{H}_\Omega$ 与 $\mathcal{H}$ 均可用于近似目标函数的二阶信息, 其中 $\mathcal{H}$ 是 $\mathcal{H}_\Omega$ 的期望形式. 出于计算上的便利性, 本文采用算子 (3-7), 当然, 在实际应用中也可以考虑其他类型的子采样方式. 基于上述算子, 我们进一步定义一种新的度量.

定义 3.1 (预条件度量)**.** g 是 $\mathcal{M}$ 上的一个内积, 对于所有的 $\vec{\mathbf{W}} \in \mathcal{M}$ 和 $\vec{\xi}, \vec{\eta} \in \mathrm{T}_{\vec{\mathbf{W}}}\mathcal{M}$, 它的定义是

$$g_{\vec{\mathbf{W}}}(\vec{\xi}, \vec{\eta}) := \sum_{k=1}^{d} \mathrm{tr} \left(\xi_k \mathbf{H}_k(\vec{\mathbf{W}}) \eta_k^\top \right), \tag{3-8}$$

这里 $\mathbf{H}_k(\vec{\mathbf{W}}) \in \mathbb{R}^{r_k r_{k+1} \times r_k r_{k+1}}$ 是一个矩阵, 定义为

$$\mathbf{H}_k(\vec{\mathbf{W}}) := \mathbf{W}_{\neq k}^{\top} \mathbf{W}_{\neq k} + \delta \mathbf{I}_{r_k r_{k+1}},$$

$\delta > 0$ 以及单位矩阵 $\mathbf{I}_{r_k r_{k+1}} \in \mathbb{R}^{r_k r_{k+1} \times r_k r_{k+1}}$.

由于矩阵 $\mathbf{W}_{\neq k}^{\top} \mathbf{W}_{\neq k}$ 不一定是正定的, 因此引入一个平移项 $\delta \mathbf{I}_{r_k r_{k+1}}$ 以避免奇异性. 此外, 度量 g 在 $\mathcal{M}$ 上是光滑的, 从而构成 $\mathcal{M}$ 上一个良定义的黎曼度量. 由此可知, 赋予度量 g 之后, $\mathcal{M}$ 成为一个黎曼流形, 而任意切向量 $\vec{\boldsymbol{\xi}} \in \mathrm{T}_{\vec{\mathbf{W}}}\mathcal{M}$ 的范数可定义为 $\|\vec{\boldsymbol{\xi}}\|_{\vec{\mathbf{W}}} := \sqrt{g_{\vec{\mathbf{W}}}(\vec{\boldsymbol{\xi}}, \vec{\boldsymbol{\xi}})}$. 黎曼梯度 (参见 [71, §3.6]) $\mathrm{grad} f(\vec{\mathbf{W}})$ 被定义为 $\mathrm{T}_{\vec{\mathbf{W}}}\mathcal{M}$ 中对所有 $\vec{\boldsymbol{\xi}} \in \mathrm{T}_{\vec{\mathbf{W}}}\mathcal{M}$ 唯一满足

$$g_{\vec{\mathbf{W}}}\left(\mathrm{grad} f(\vec{\mathbf{W}}), \vec{\boldsymbol{\xi}}\right) = \mathrm{D} f(\vec{\mathbf{W}})[\vec{\boldsymbol{\xi}}] := \langle \nabla f(\vec{\mathbf{W}}), \vec{\boldsymbol{\xi}} \rangle$$

成立的元素. 值得注意的是

$$g_{\vec{\mathbf{W}}}(\vec{\boldsymbol{\eta}}, \vec{\boldsymbol{\xi}}) = \sum_{k=1}^{d} \mathrm{tr}\left(\boldsymbol{\eta}_k \mathbf{H}_k(\vec{\mathbf{W}}) \left(\boldsymbol{\xi}_k\right)^{\top}\right) = \langle (\mathcal{H} + \delta \mathcal{I})(\vec{\mathbf{W}})[\vec{\boldsymbol{\eta}}], \vec{\boldsymbol{\xi}} \rangle,$$

这里 $\mathcal{I}(\vec{\mathbf{W}}) : \mathrm{T}_{\vec{\mathbf{W}}}\mathcal{M} \to \mathrm{T}_{\vec{\mathbf{W}}}\mathcal{M}$ 是恒等映射. 于是我们有

$$\mathrm{grad} f(\vec{\mathbf{W}}) = (\mathcal{H} + \delta \mathcal{I})^{-1}(\vec{\mathbf{W}})[\nabla f(\vec{\mathbf{W}})].$$

结合 (3-2) 与 (3-5) 可知, $\mathrm{grad} f$ 是目标函数 f 的 Newton 方向的一种近似. 因此, 度量 (3-8) 对欧氏梯度起到了预条件作用. 基于这一观察, 我们将新的度量 (3-8) 称为定义在 $\mathcal{M}$ 上的预条件度量. 综上所述, 黎曼梯度可以由式 (3-4) 按如下方式计算.

命题 3.1 (黎曼梯度)**.** *函数 f 在 $\vec{\mathbf{W}} \in \mathcal{M}$ 处关于度量 g 的黎曼梯度为*

$$\mathrm{grad} f(\vec{\mathbf{W}}) = \left(\partial_{\mathbf{W}_1} f(\vec{\mathbf{W}}) \mathbf{H}_1^{-1}(\vec{\mathbf{W}}), \ldots, \partial_{\mathbf{W}_d} f(\vec{\mathbf{W}}) \mathbf{H}_d^{-1}(\vec{\mathbf{W}})\right). \tag{3-9}$$

3.3.2 黎曼预条件算法

基于预条件度量 (3-8) 以及黎曼梯度 (3-9), 我们提出黎曼梯度法和黎曼共轭梯度法, 用以求解张量补全问题 (3-2).

黎曼梯度法 黎曼梯度下降算法如算法 8 所示. 需要注意的是, 在流形 $\mathcal{M}$ 上的收缩映射为恒等映射. 关于步长的选取, 我们考虑以下两种策略：1) 精确线搜索. 通过求解如下优化问题得到步长

$$s_{\mathrm{exact}}^{(t)} := \arg\min_{s>0} h(s) = f(\vec{\mathbf{W}}^{(t)} + s\vec{\boldsymbol{\eta}}^{(t)}). \tag{3-10}$$

由于 h 是关于 s 的一个 $2d$ 次多项式, 其导函数 $h'(s)$ 为 $2d-1$ 次多项式, 因此 $s_{\text{exact}}^{(t)}$ 可由 $h'(s)$ 的根得到; 2) Armijo 回溯线搜索. 给定初始步长 $s_0^{(t)}>0$, 寻找最小的非负整数 ℓ, 使得当$s^{(t)}=\rho^{\ell}s_0^{(t)}>s_{\min}$ 时, 不等式

$$f(\vec{\mathbf{W}}^{(t)})-f(\vec{\mathbf{W}}^{(t)}+s^{(t)}\vec{\boldsymbol{\eta}}^{(t)})\geq -s^{(t)}ag_{\vec{\mathbf{W}}^{(t)}}\left(\text{grad}f(\vec{\mathbf{W}}^{(t)}),\vec{\boldsymbol{\eta}}^{(t)}\right) \tag{3-11}$$

成立, 其中 $\rho,a\in(0,1)$ 以及 $s_{\min}>0$ 为回溯线搜索参数. 此外, 文献 [149] 中提出的黎曼 Barzilai–Borwein(RBB) 步长, 在许多应用中有良好的数值表现. 其定义为

$$s_{\text{RBB1}}^{(t)}:=\frac{\|\vec{\mathbf{Z}}^{(t-1)}\|_{\vec{\mathbf{W}}^{(t)}}^2}{|g_{\vec{\mathbf{W}}^{(t)}}(\vec{\mathbf{Z}}^{(t-1)},\vec{\mathbf{Y}}^{(t-1)})|}\quad\text{以及}\quad s_{\text{RBB2}}^{(t)}:=\frac{|g_{\vec{\mathbf{W}}^{(t)}}(\vec{\mathbf{Z}}^{(t-1)},\vec{\mathbf{Y}}^{(t-1)})|}{\|\vec{\mathbf{Y}}^{(t-1)}\|_{\vec{\mathbf{W}}^{(t)}}^2}, \tag{3-12}$$

其中 $\vec{\mathbf{Z}}^{(t-1)}:=\vec{\mathbf{W}}^{(t)}-\vec{\mathbf{W}}^{(t-1)}$, $\vec{\mathbf{Y}}^{(t-1)}:=\text{grad}f(\vec{\mathbf{W}}^{(t)})-\text{grad}f(\vec{\mathbf{W}}^{(t-1)})$. 基于 RBB 步长的良好数值表现, 我们将 RBB 步长作为初始步长 $s_0^{(t)}$. 值得注意的是, 在实际计算中, 初始步长 $s_0^{(0)}$ 也可以通过精确线搜索 (3-10) 生成.

算法 8 求解问题 (3-2)的黎曼梯度法 (TR-RGD)

输入: $f:\mathcal{M}\to\mathbb{R}$, $\vec{\mathbf{W}}^{(0)}\in\mathcal{M}$, 阈值 $\varepsilon>0$, $t=0$; 回溯线搜索参数 $\rho,a\in(0,1)$, $s_{\min}>0$.

1: **while** 停机准则未被满足 **do**
2: 　计算搜索方向 $\vec{\boldsymbol{\eta}}^{(t)}=-\text{grad}f(\vec{\mathbf{W}}^{(t)})$.
3: 　通过精确线搜索 (3-10) 或回溯线搜索 (3-11) 计算步长 $s^{(t)}$.
4: 　更新 $\vec{\mathbf{W}}^{(t+1)}=\vec{\mathbf{W}}^{(t)}+s^{(t)}\vec{\boldsymbol{\eta}}^{(t)}$; $t=t+1$.
5: **end while**

输出: $\vec{\mathbf{W}}^{(t)}\in\mathcal{M}$.

求解张量环补全问题的一个典型方法是交替最小二乘法 [143](简称 TR-ALS). 我们在下面的注释中阐述 TR-ALS 与本文提出的 TR-RGD 算法之间的联系与区别.

注. TR-ALS 和 TR-RGD 都可以被视为线搜索类方法. 不同之处在于, 在 TR-ALS 中, 因子矩阵 $\mathbf{W}_1,\ldots,\mathbf{W}_d$ 是通过精确线搜索按顺序逐个更新的; 而在所提出的 TR-RGD 方法中, $(\mathbf{W}_1,\ldots,\mathbf{W}_d)$ 被整体视为流形 $\mathcal{M}$ 上的一个点 $\vec{\mathbf{W}}$, 并通过黎曼梯度法进行同步更新. 所提出的 TR-RGD 方法得益于一个针对张量补全问题 (3-2) 精心设计的预条件度量 (3-8). 此外, TR-RGD 允许更加灵活的步长选择策略, 例如精确线搜索以及黎曼 Barzilai–Borwein 步长. 因此, TR-RGD 方法在实际应用中具有潜在的竞争优势.

黎曼共轭梯度法 算法 9展示了求解问题 (3-2)的黎曼共轭梯度法. 针对参数 $\beta^{(t)}$, 我们考虑黎曼版本的 [134] 修正 Hestenes–Stiefel 准则 (HS+) [150], 即

$$\beta^{(t)} := \max\left\{\frac{g_{\vec{\mathbf{W}}^{(t)}}\left(\mathrm{grad} f(\vec{\mathbf{W}}^{(t)}) - \mathrm{grad} f(\vec{\mathbf{W}}^{(t-1)}), \mathrm{grad} f(\vec{\mathbf{W}}^{(t)})\right)}{g_{\vec{\mathbf{W}}^{(t)}}\left(\mathrm{grad} f(\vec{\mathbf{W}}^{(t)}) - \mathrm{grad} f(\vec{\mathbf{W}}^{(t-1)}), \vec{\boldsymbol{\eta}}^{(t-1)}\right)}, 0\right\}. \tag{3-13}$$

我们选择回溯线搜索 (3-11)作为算法 9的步长选取准则.

算法 9 求解问题 (3-2)的黎曼共轭梯度法 (TR-RCG)

输入: $f : \mathcal{M} \to \mathbb{R}$, $\vec{\mathbf{W}}^{(0)} \in \mathcal{M}$, 阈值 $\varepsilon > 0$, $t = 0$, $\vec{\boldsymbol{\eta}}^{(-1)} = 0$; 回溯线搜索参数 $\rho, a \in (0,1)$, $s_{\min} > 0$.

1: **while** 停机准则未被满足 **do**
2: 利用 (3-13)计算搜索参数 $\vec{\boldsymbol{\eta}}^{(t)} = -\mathrm{grad} f(\vec{\mathbf{W}}^{(t)}) + \beta^{(t)}\vec{\boldsymbol{\eta}}^{(t-1)}$.
3: 通过回溯线搜索 (3-11) 计算步长 $s^{(t)}$.
4: 更新 $\vec{\mathbf{W}}^{(t+1)} = \vec{\mathbf{W}}^{(t)} + s^{(t)}\vec{\boldsymbol{\eta}}^{(t)}$; $t = t + 1$.
5: **end while**

输出: $\vec{\mathbf{W}}^{(t)} \in \mathcal{M}$.

在这两种算法中, 黎曼梯度的计算占据了总体计算开销的主要部分, 因为其涉及大规模矩阵的构造与乘法运算. 由于该计算代价随张量阶数 d 呈指数级增长, 通过构造大规模矩阵直接计算黎曼梯度的方法在实际应用中是不可行的. 因此, 我们有必要设计一种能够高效计算黎曼梯度的计算方法.

3.3.3 梯度的高效计算方式

本小节我们详细研究黎曼梯度 (3-9) 的计算细节. 一般而言, $\mathrm{grad} f(\vec{\mathbf{W}}) = (\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_d)$ 的计算包含两个步骤：第一步是计算欧氏梯度 $\nabla f(\vec{\mathbf{W}}) = (\mathbf{G}_1, \ldots, \mathbf{G}_d)$, 见式 (3-4); 第二步是组装成黎曼梯度 $\mathrm{grad} f(\vec{\mathbf{W}})$, 其中

$$\mathbf{G}_k := \partial_{\mathbf{W}_k} f(\vec{\mathbf{W}}) = \frac{1}{p}\mathbf{S}_{(k)}\mathbf{W}_{\neq k} + \lambda \mathbf{W}_k,$$
$$\boldsymbol{\eta}_k := \mathbf{G}_k \mathbf{H}_k^{-1}(\vec{\mathbf{W}}) = \mathbf{G}_k(\mathbf{W}_{\neq k}^{\top}\mathbf{W}_{\neq k} + \delta \mathbf{I}_{r_k r_{k+1}})^{-1}$$

以及 $k \in [d]$.

采用直接方式计算梯度时, 主要涉及以下五个操作：1) 对 $k \in [d]$ 构造矩阵 $\mathbf{W}_{\neq k}$, 其计算量为 $2\sum_{k=1}^{d} n_{-k} R_k$ 次浮点运算, 其中 $R_k := r_k\left(\sum_{j=1, j\neq k}^{d} r_j r_{j+1}\right)$; 2) 计算稀疏张量 $\mathcal{S}$, 需要 $2|\Omega| r_1 r_2$ 次浮点运算; 3) 通过稀疏与稠密矩阵的乘法计算欧氏梯度, 该过程需要 $2|\Omega|\bar{r}$ 次浮点运算, 其中 $\bar{r} := \sum_{k=1}^{d} r_k r_{k+1}$; 4) 通过稠密矩阵的乘法计算 $\mathbf{H}_k(\vec{\mathbf{W}})$, 其计算量为 $2\sum_{k=1}^{d} n_{-k}(r_k r_{k+1})^2$; 5) 通过乔列斯基分解求解线性系统以获得黎曼梯度, 其计算量为 $\sum_{k=1}^{d}\left(2n_k(r_k r_{k+1})^2 + C_{\mathrm{chol}}(r_k r_{k+1})^3\right)$. 综合上述各项操作, 直接计算黎曼梯度的总计算量为

$$2|\Omega| d\bar{r} + 2|\Omega| r_1 r_2 + \sum_{k=1}^{d}\left(2n_{-k}(R_k + (r_k r_{k+1})^2) + 2n_k(r_k r_{k+1})^2 + C_{\mathrm{chol}}(r_k r_{k+1})^3\right)$$

次浮点运算. 若 $n_1 = \cdots = n_d = n$ 且 $r_1 = \cdots = r_d = r$, 则上述复杂度可化简为

$$2(d+1)|\Omega|r^2 + 2d(d-1)n^{d-1}r^3 + 2dn^{d-1}r^4 + 2dnr^4 + C_{\text{chol}}dr^6$$

次浮点运算. 在实际计算中, 阶为 $\mathcal{O}(n^{d-1})$ 的项占据了主导地位.

通过利用 $\mathbf{W}_{\neq k}^{\top}\mathbf{W}_{\neq k}$ 的 Kronecker 积结构, 可以显著降低总体计算成本. 算法 10 给出了在不显式构造 $\mathbf{W}_{\neq k}$ 的情况下计算 $\mathbf{W}_{\neq k}^{\top}\mathbf{W}_{\neq k}$ 的方法. 具体而言, 该矩阵乘积可表示为

$$\mathbf{W}_{\neq k}^{\top}\mathbf{W}_{\neq k} = \sum_{i=1}^{n_{-k}} \mathbf{W}_{\neq k}(:,i)\mathbf{W}_{\neq k}(:,i)^{\top} = \sum_{\mathbf{i}_{-k}} \tilde{\mathbf{w}}_k(\mathbf{i}_{-k})\tilde{\mathbf{w}}_k(\mathbf{i}_{-k})^{\top}, \tag{3-14}$$

这里 $\tilde{\mathbf{w}}_k(\mathbf{i}_{-k}) := \operatorname{vec}((\prod_{j=k+1}^{d}\mathbf{U}_j(i_j)\prod_{j=1}^{k-1}\mathbf{U}_j(i_j))^{\top})$ 对应于矩阵 $\mathbf{W}_{\neq k}$ 的行, 以及指标 $\mathbf{i}_{-k} := (i_{k+1},\ldots,i_d,i_1,\ldots,i_{k-1}) \in [n_{k+1}]\times\cdots\times[n_d]\times[n_1]\times\cdots\times[n_{k-1}]$. 通过利用等式 $\operatorname{vec}(\mathbf{CXB}^{\top}) = (\mathbf{B}\otimes\mathbf{C})\operatorname{vec}(\mathbf{X})$ (其中 $\mathbf{B},\mathbf{C},\mathbf{X}$ 为大小合适的矩阵), 我们可以得到

$$\begin{aligned}\tilde{\mathbf{w}}_k(\mathbf{i}_{-k}) &= \left(\left(\prod_{j=k+1}^{d}\mathbf{U}_j(i_j)\prod_{j=1}^{k-1}\mathbf{U}_j(i_j)\right)\otimes\mathbf{I}_{r_k}\right)\operatorname{vec}(\mathbf{U}_{k-1}(i_{k-1})^{\top}) \\ &= \prod_{j=k+1}^{d}(\mathbf{U}_j(i_j)\otimes\mathbf{I}_{r_k})\prod_{j=1}^{k-1}(\mathbf{U}_j(i_j)\otimes\mathbf{I}_{r_k})\operatorname{vec}(\mathbf{U}_{k-1}(i_{k-1})^{\top}),\end{aligned} \tag{3-15}$$

其中 $\otimes$ 表示 Kronecker 积. 将式 (3-15) 代入 (3-14), 可以递归地计算 $\mathbf{W}_{\neq k}^{\top}\mathbf{W}_{\neq k}$:

$$\begin{aligned}\tilde{\mathbf{H}}_1 &:= \sum_{i_{k-1}=1}^{n_{k-1}}\operatorname{vec}(\mathbf{U}_{k-1}(i_{k-1})^{\top})\operatorname{vec}(\mathbf{U}_{k-1}(i_{k-1})^{\top})^{\top}, \\ \tilde{\mathbf{H}}_2 &:= \sum_{i_{k-2}=1}^{n_{k-2}}(\mathbf{U}_{k-2}(i_{k-2})\otimes\mathbf{I}_{r_k})\tilde{\mathbf{H}}_1(\mathbf{U}_{k-2}(i_{k-2})^{\top}\otimes\mathbf{I}_{r_k}), \\ &\vdots \\ \tilde{\mathbf{H}}_{d-1} &:= \sum_{i_{k+1}=1}^{n_{k+1}}(\mathbf{U}_{k+1}(i_{k+1})\otimes\mathbf{I}_{r_k})\tilde{\mathbf{H}}_{d-2}(\mathbf{U}_{k+1}(i_{k+1})^{\top}\otimes\mathbf{I}_{r_k}).\end{aligned}$$

由此可得 $\mathbf{W}_{\neq k}^{\top}\mathbf{W}_{\neq k} = \tilde{\mathbf{H}}_{d-1}$. 于是利用算法 10 计算 $k\in[d]$ 的 $\mathbf{W}_{\neq k}^{\top}\mathbf{W}_{\neq k}$ 计算量为

$$\sum_{k=1}^{d}\left(2n_{k-1}(r_kr_{k-1})^2 + 2\sum_{i=1,i\neq k,k-1}^{d} n_i r_k^2 r_i r_{i+1}(r_i + r_{i+1})\right)$$

次浮点运算. 若 $n_1 = \cdots = n_d$ 且 $r_1 = \cdots = r_d$, 则黎曼梯度的计算总复杂度为

$$2d(d-1)|\Omega|r^3 + 2|\Omega|r^2 + 4d(d-2)nr^5 + 2dnr^4 + C_{\text{chol}}dr^6,$$

其中最高阶项为 $\mathcal{O}(n)$. 为验证算法 10 的加速效果, 我们在表3-3 中给出了基于电影评分真实数据集的对比实验.

算法 10 计算 $\mathbf{W}_{\neq k}^{\top}\mathbf{W}_{\neq k}$ 的高效算法

输入: $k \in [d]$, 核张量 $\mathcal{U}_l = \mathrm{ten}_{(2)}(\mathbf{W}_l)$, 切片矩阵 $\mathbf{U}_l(i_l)$, $i_l \in [n_l]$, $l \in [d]$.

1: 设置 $j = \mathrm{mod}(k-2+d, d)+1$.
2: 计算 $\tilde{\mathbf{H}}_1 = \sum_{i_j=1}^{n_j} \mathrm{vec}(\mathbf{U}_j(i_j)^{\top})\mathrm{vec}(\mathbf{U}_j(i_j)^{\top})^{\top}$.
3: **for** $l = 2, 3, \ldots, d-1$ **do**
4: 设置 $j = \mathrm{mod}(k-l-1+d, d)+1$.
5: 计算 $\tilde{\mathbf{H}}_l = \sum_{i_j=1}^{n_j}(\mathbf{U}_j(i_j)\otimes\mathbf{I}_{r_k})\tilde{\mathbf{H}}_{l-1}(\mathbf{U}_j(i_j)^{\top}\otimes\mathbf{I}_{r_k})$.
6: **end for**

输出: $\tilde{\mathbf{H}}_{d-1} = \mathbf{W}_{\neq k}^{\top}\mathbf{W}_{\neq k}$.

3.4 收敛性分析

本节分析 TR-RGD 和 TR-RCG 的全局收敛性. 设 $\{\vec{\mathbf{W}}^{(t)}\}_{t\geq 0}$ 为由算法 8 或算法 9 生成的一个无限序列. 我们证明该序列的每一个聚点都是一个稳定点; 见定理 3.3 和定理 3.4.

引理 3.2 ([74, Lemma 2.7]). *设 $\mathcal{M}' \subseteq \mathcal{M}$ 为一个紧黎曼子流形, $\mathrm{R}_x : \mathrm{T}_x\mathcal{M}' \to \mathcal{M}'$ 为一个收缩映射. 若函数 $f : \mathcal{M}' \to \mathbb{R}$ 在 $\mathcal{M}'$ 的凸包上具有 Lipschitz 连续的梯度, 则存在常数 $L > 0$, 使得对所有的 $x \in \mathcal{M}'$, $\xi \in \mathrm{T}_x\mathcal{M}'$ 有*

$$\left|f(\mathrm{R}_x(\xi)) - f(x) - g_x(\xi, \mathrm{grad}f(x))\right| \leq \frac{L}{2}g_x(\xi,\xi) \quad .$$

由于搜索空间 $\mathcal{M}$ 是平直的, 引理 3.2 中的收缩映射可取为恒等映射. 由问题 (3-2) 中的正则项 $\frac{\lambda}{2}\|\vec{\mathbf{W}}\|_{\mathrm{F}}^2$ 可知目标函数 f 具有强制性 (coercivity). 因此, 水平集 $\mathcal{L} := \{\vec{\mathbf{W}} : f(\vec{\mathbf{W}}) \leq f(\vec{\mathbf{W}}^{(0)})\}$ 是紧集. 由式 (3-10) 和 (3-11) 得到的函数值序列 $\{f(\vec{\mathbf{W}}^{(t)})\}_{t\geq 0}$ 是单调递减的. 于是, $\{\vec{\mathbf{W}}^{(t)}\}_{t\geq 0}$ 是包含在 $\mathcal{L}$ 中的有界序列, 并且满足

$$\|\vec{\mathbf{W}}^{(t)}\|_{\mathrm{F}}^2 = \frac{2\left(f(\vec{\mathbf{W}}^{(t)}) - f_{\Omega}(\vec{\mathbf{W}}^{(t)})\right)}{\lambda} \leq \frac{2f(\vec{\mathbf{W}}^{(t)})}{\lambda} \leq \frac{2f(\vec{\mathbf{W}}^{(0)})}{\lambda}.$$

因此, 序列 $\{\vec{\mathbf{W}}^{(t)}\}_{t\geq 0}$ 至少存在一个聚点. 此外, 目标函数 f 具有 Lipschitz 连续梯度. 结合引理 3.2 以及 [117, Proposition 4.3], 可以采用类似的论证方式证明所提出的黎曼梯度下降算法的全局收敛性.

定理 3.3. *设 $\{\vec{\mathbf{W}}^{(t)}\}_{t\geq 0}$ 为由算法 8 生成的一个无限序列, 则存在常数 $C > 0$, 使得 $f(\vec{\mathbf{W}}^{(t)}) - f(\vec{\mathbf{W}}^{(t+1)}) > C\|\mathrm{grad}f(\vec{\mathbf{W}}^{(t)})\|_{\vec{\mathbf{W}}^{(t)}}$. 此外, 有以下结论成立: 1) $\{\vec{\mathbf{W}}^{(t)}\}_{t\geq 0}$ 的每一个聚点都是 f 的一个稳定点; 2) 算法至多经过 $\left\lceil f(\vec{\mathbf{W}}^{(0)})/(C\epsilon^2)\right\rceil$ 次迭代, 即可返回一个满足 $\|\mathrm{grad}f(\vec{\mathbf{W}})\|_{\vec{\mathbf{W}}} < \epsilon$ 的点 $\vec{\mathbf{W}} \in \mathcal{M}$.*

下面讨论算法 9 中所提出的黎曼共轭梯度法的全局收敛性. 关于一般流形上 RCG 方法的收敛性分析, 读者可参考 [71, 72]. 这里我们沿用 [71] 中的分析框架. 为了满足 [71, Theorem 4.3.1] 的基本假设, 即搜索方向序列 $\{\vec{\boldsymbol{\eta}}^{(t)}\}_{t\geq 0}$ 与 $\{\vec{\mathbf{W}}^{(t)}\}_{t\geq 0}$ 是梯度相关的 (gradient-related), 我们对搜索方向 $\vec{\boldsymbol{\eta}}^{(t)}$ 施加如下重启策略：

$$\vec{\boldsymbol{\eta}}^{(t)} = -\mathrm{grad} f(\vec{\mathbf{W}}^{(t)}) \quad 若 \quad g_{\vec{\mathbf{W}}^{(t)}}(\vec{\boldsymbol{\eta}}^{(t-1)}, \mathrm{grad} f(\vec{\mathbf{W}}^{(t)})) \geq 0. \tag{3-16}$$

因此, $\{\vec{\boldsymbol{\eta}}^{(t)}\}_{t\geq 0}$ 是一列下降方向, 并且满足

$$\begin{aligned} & g_{\vec{\mathbf{W}}^{(t)}}(\vec{\boldsymbol{\eta}}^{(t)}, \mathrm{grad} f(\vec{\mathbf{W}}^{(t)})) \\ =& - \|\mathrm{grad} f(\vec{\mathbf{W}}^{(t)})\|^2_{\vec{\mathbf{W}}^{(t)}} + \beta^{(t)} g_{\vec{\mathbf{W}}^{(t)}}(\vec{\boldsymbol{\eta}}^{(t-1)}, \mathrm{grad} f(\vec{\mathbf{W}}^{(t)})) \\ \leq& - \|\mathrm{grad} f(\vec{\mathbf{W}}^{(t)})\|^2_{\vec{\mathbf{W}}^{(t)}}. \end{aligned} \tag{3-17}$$

此外, 由算法 9 可得

$$\|\vec{\boldsymbol{\eta}}^{(t)}\|_{\mathrm{F}} = \frac{\|\vec{\mathbf{W}}^{(t+1)} - \vec{\mathbf{W}}^{(t)}\|_{\mathrm{F}}}{s^{(t)}} \leq \frac{\|\vec{\mathbf{W}}^{(t+1)}\|_{\mathrm{F}} + \|\vec{\mathbf{W}}^{(t)}\|_{\mathrm{F}}}{s_{\min}} \leq \frac{2}{s_{\min}} \sqrt{\frac{2 f(\vec{\mathbf{W}}^{(0)})}{\lambda}}, \tag{3-18}$$

从而 $\|\vec{\boldsymbol{\eta}}^{(t)}\|_{\mathrm{F}}$ 是一致有界的. 结合 (3-17) 和 (3-18) 可知 $\{\vec{\boldsymbol{\eta}}^{(t)}\}_{t\geq 0}$ 与 $\{\vec{\mathbf{W}}^{(t)}\}_{t\geq 0}$ 是梯度相关的. 由 [71, Theorem 4.3.1], 我们得到如下结果.

定理 3.4. *设 $\{\vec{\mathbf{W}}^{(t)}\}_{t\geq 0}$ 为在采用重启策略 (3-16) 下, 由算法 9 生成的一个无限序列, 则 $\{\vec{\mathbf{W}}^{(t)}\}_{t\geq 0}$ 的每一个聚点都是 f 的一个稳定点.*

3.5　数值实验

在本节中, 我们在人工合成数据集和真实数据集 (包括电影评分数据、高光谱图像以及高维函数) 上, 对 TR-RGD(算法 8) 和 TR-RCG(算法 9) 与若干基于不同张量分解的现有算法进行了数值比较. 首先, 我们介绍所有参与比较的算法及其默认参数设置.

我们考虑文献 [60] 中基于张量链分解的黎曼共轭梯度法[1], 记为 “TT-RCG”. 对于基于 CP 分解的算法, 我们选用 Tensor-Toolbox[2] 中的 “CP-WOPT” [145], 该工具箱由 Bader 和 Kolda 提供 [151], 并采用作者推荐的有限记忆 BFGS 算法[3]. “GeomCG” [15] 是一个用于 Tucker 分解张量补全问题的黎曼共轭梯度法[4]. 此外, 我们还考虑了一种基于核范数的方法[5], 即 “HaLRTC” [135, 152]. 除非另有说明, 所有对比算法均采用其默认参数设置.

需要指出的是, 步长的选取方式有多种. 在初步数值实验中, 我们发现, 采用公式 (3-12) 中 RBB2 步长的 TR-RGD 表现优于采用 RBB1 的版本. 因此, 在后续

[1] TTeMPS 工具箱：https://www.epfl.ch/labs/anchp/index-html/software/ttemps/.

[2] Tensor-Toolbox v3.4: http://www.tensortoolbox.org/.

[3] 可从 https://github.com/stephenbeckr/L-BFGS-B-C获取.

[4] GeomCG 工具箱：https://www.epfl.ch/labs/anchp/index-html/software/geomcg/.

[5] 可从 https://github.com/andrewssobral/mctc4bmi/tree/master/algs_tc/LRTC获取.

比较中我们只考虑 RBB2 步长. “TR-RGD (exact)” 和 “TR-RGD (RBB)” 分别表示采用精确线搜索 (3-10) 的算法 8和采用 Armijo 回溯线搜索 (3-11) 并结合 RBB2 步长的算法 8. 此外, 我们还考虑了欧氏梯度下降法 (记为 “TR-GD”) 用于对比. TR-GD 的步长基于 Armijo 回溯线搜索以及标准的 BB 步长 [153]. 线搜索参数的默认设置为 $\rho = 0.4$, $a = 10^{-5}$, and $s_{\min} = 10^{-10}$. 另外, 我们还实现了用于张量补全问题 (3-2) 的交替最小二乘算法 [143, 146], 记为 “TR-ALS”.

所有算法均由 $\mathcal{X}^{(0)} = \tau(\vec{\mathbf{W}})$ 进行初始化, 其中映射 τ 定义于公式 (3-3), 而 $\vec{\mathbf{W}}$ 为随机生成的变量. 我们在采样集 Ω 上定义相对误差为

$$\varepsilon_{\Omega}(\vec{\mathbf{W}}) := \frac{\| \mathrm{P}_{\Omega}(\tau(\vec{\mathbf{W}})) - \mathrm{P}_{\Omega}(\mathcal{A})\|_{\mathrm{F}}}{\| \mathrm{P}_{\Omega}(\mathcal{A})\|_{\mathrm{F}}}.$$

我们将 $\varepsilon_{\Omega}(\vec{\mathbf{W}})$ 称为训练误差. 此外, 我们在与 Ω 不同的测试集 Γ 上计算测试误差 $\varepsilon_{\Gamma}(\vec{\mathbf{W}})$, 其中 $|\Gamma|$ 的默认取值为 100. 当满足以下任一条件时算法终止: 1) 训练误差满足 $\varepsilon_{\Omega}(\vec{\mathbf{W}}^{(t)}) < 10^{-12}$; 2) 相对变化量 $|(\varepsilon_{\Omega}(\vec{\mathbf{W}}^{(t)}) - \varepsilon_{\Omega}(\vec{\mathbf{W}}^{(t-1)}))/\varepsilon_{\Omega}(\vec{\mathbf{W}}^{(t-1)})| < \varepsilon$; 3) 梯度范数 $\|\mathrm{grad} f(\vec{\mathbf{W}})\|_{\mathrm{F}} < \varepsilon$. 该准则仅对黎曼方法启用; 4) 达到最大迭代次数; 5) 超出时间预算. 其中阈值 ε 选定为 10^{-8}. 所有实验均在一台 MacBook Pro 2019 上完成, 配置为：MacOS Ventura 13.1, 2.4 GHz 八核 Intel Core i9 处理器, 32GB 内存, Matlab R2020b. TR-RGD 与 TR-RCG 的代码可从 https://github.com/JimmyPeng1998/LRTCTR下载.

表 3-1 不同张量分解下秩参数选取方式.

Table 3-1 Rank selections for different tensor formats in numerical experiments.

形式	电影评分数据集				高光谱图像		高维函数	
	秩参数	参数量	秩参数	参数量	秩参数	参数量	秩参数	参数量
CP	36	365112	49	496958	110	67430	-	-
Tucker	(36,36,36)	411768	(60,30,18)	516060	(65,65,7)	67506	-	-
TT	(1,9,9,1)	375822	(1,10,10,1)	457100	(1,15,15,1)	78495	(1,5,5,5,1)	1200
TR	(6,6,6)	365112	(6,10,3)	483660	(7,16,7)	66577	(4,4,4,4)	1280

表 3-1介绍了在数值实验中秩参数的选取准则. 例如, 在大小为 $250\times330\times33$ 的高光谱图像上的实验中, 我们根据文献 [15] 选择了 Tucker 秩 $(65, 65, 7)$. 为了保证公平比较, 我们选取秩参数使得对于不同的张量分解形式, 参数空间的参数量相近. 为此, 我们计算总参数量:

$$65 \times 65 \times 7 + 250 \times 65 + 330 \times 65 + 33 \times 7 = 67506.$$

对于 TT 分解, 我们选择秩参数为 $(1, 15, 15, 1)$, 此时它的总参数量为

$$250 \times 15 + 15 \times 330 \times 15 + 15 \times 33 = 78495.$$

对于 TR 分解, 我们选择秩参数为 $(7, 16, 7)$. 此时它的总参数量为

$$7 \times 250 \times 16 + 16 \times 330 \times 7 + 7 \times 33 \times 7 = 66577.$$

对于 CP 分解, 我们选择秩参数为

$$R = \frac{67506}{250 + 330 + 33} \approx 110.$$

此时, 对于不同的分解而言, 他们的总参数量是相当的. 对于其他的数值实验, 选取方式与上述方法相同.

3.5.1　人工合成数据集

在本小节中, 我们研究 TR 分解下张量补全算法在无噪声观测和含噪声观测情形下的性能及恢复完整数据的能力.

无噪声观测　我们设问题 (3-2) 中的一个人工合成低秩张量为 $\mathcal{A} \in \mathbb{R}^{n_1 \times \cdots \times n_d}$, 其生成方式为

$$\mathcal{A} = \tau(\vec{\mathbf{W}}),$$

其中映射 τ 定义于公式 (3-3), $\vec{\mathbf{W}} \in \mathcal{M}$ 的每个元素均从区间 $[0, 1]$ 上均匀采样. 在实验中, 我们取 $d = 3$, $n_1 = n_2 = n_3 = 100$, 真实的 TR 秩为 $\mathbf{r}^* = (6, 6, 6)$. 给定采样率 p, 我们通过从集合 $[n_1] \times [n_2] \times \cdots \times [n_d]$ 中随机选取 $pn_1n_2 \cdots n_d$ 个索引来构造采样集 Ω. 为了获得无偏的恢复结果, 我们取正则化参数 $\lambda = 0$, 采样率 $p = 0.3$, 以及估计秩 $\mathbf{r} = (6, 6, 6)$. 时间预算设为 1800 秒, 最大迭代次数为 10000.

图 3-1 给出了无噪声情形下的数值实验结果. 首先可以观察到, 所有算法均在给定的时间预算内成功恢复了真实低秩张量. 在训练误差和测试误差方面, TR-RGD(RBB) 和 TR-RCG 的表现均优于交替最小二乘算法 (TR-ALS), 且 TR-RGD(RBB) 与 TR-RCG 的性能基本相当. 其次, 采用 RBB 步长的 TR-RGD 在效率上优于采用精确线搜索的版本, 这是因为精确线搜索需要求解一个 $2d - 1$ 次多项式的根. 然而, 在实际数值实验中, 计算该多项式系数的代价较高. 第三, 值得注意的是, 所提出的黎曼梯度算法的性能明显优于采用 BB 步长的欧氏梯度算法, 这表明所引入的新黎曼度量确实具有预条件效果.

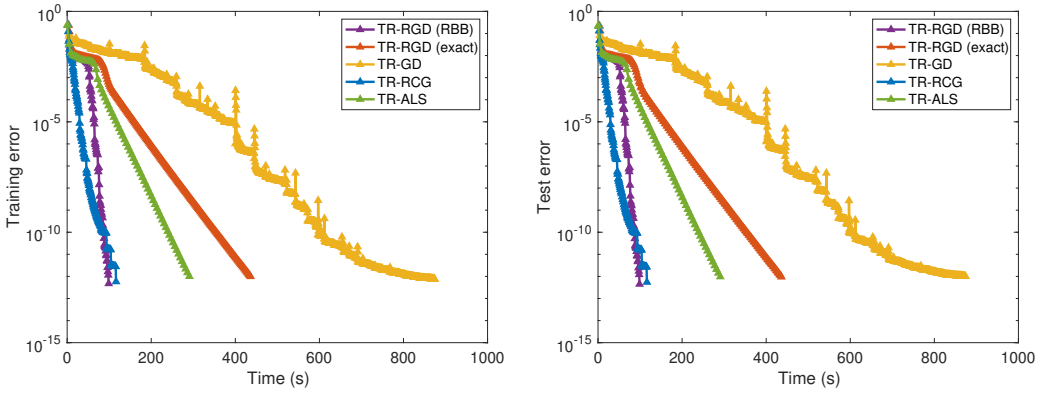


图 3-1 无噪声观测下的数值结果. 左图: 训练误差. 右图: 测试误差.

Figure 3-1 Numerical results of noiseless case. Left: training error. Right: test error.

“恢复一个低秩数据需要多少样本” 这一问题既有趣又具有挑战性. 在矩阵情形 ($d = 2$) 下, 大约需要 $\mathcal{O}(nr \log n)$ 个样本才能恢复一个低秩矩阵 [102, 154]. 我们希望通过数值实验来研究这一问题, 即探索样本量与张量大小 n 之间的关系, 研究对象为 TR 秩 $\mathbf{r}^* = (3, 3, 3)$ 的三阶张量. 为此, 我们按照前述方式随机生成 TR 秩为 $\mathbf{r} = (3, 3, 3)$ 的三阶人工合成张量 $\mathcal{A}$, 并令张量大小 $n := n_1 = n_2 = n_3$ 取自集合 $\{60, 70, \ldots, 180\}$, 采样点数 $|\Omega|$ 取自集合 $\{1000, 1500, \ldots, 20000\}$. 对于每一组 $(n, |\Omega|)$ 的组合, 我们分别运行 TR-RGD、TR-RCG 和 TR-ALS 算法各五次. 若某一算法在最大迭代次数 250 以内, 使得测试误差满足 $\varepsilon_\Gamma < 10^{-4}$, 则认为该算法成功恢复了对应的张量. 图 3-2 给出了恢复结果的相图, 其中每个方块的灰度表示成功恢复的次数, 白色方块表示在五次实验中均成功恢复. 红色曲线表示数值 $\mathcal{O}(n \log n)$. 这些相图表明, 三阶张量补全在 TR 分解下呈现出与矩阵情形相似的行为, 这一点也与其他针对 $d = 3$ 的数值实验结果一致, 例如 [15].

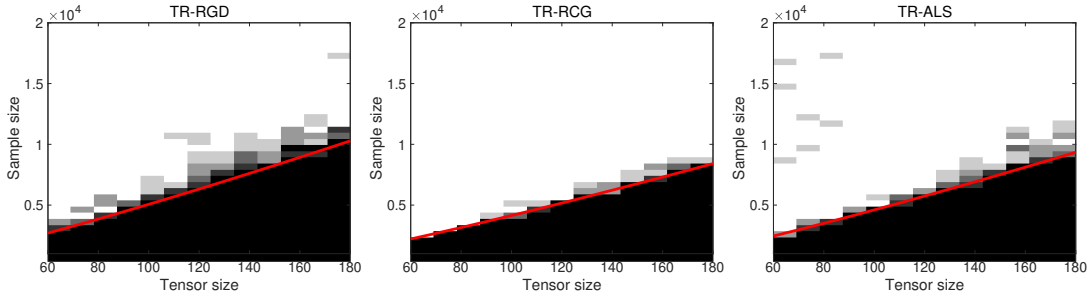


图 3-2 五次实验的恢复结果相图. 白色方块表示在五次实验中均成功恢复, 黑色方块表示在五次实验中均未成功恢复.

Figure 3-2 Phase plots of recovery results for five runs. The white block indicates successful recovery in all five runs, while the black block signifies failure of recovery in all five runs.

含噪声观测 此外, 我们研究了 TR 分解算法在不同噪声水平下的重建能力. 我们考虑如下形式的合成含噪声张量

$$\mathcal{A} := \frac{\hat{\mathcal{A}}}{\|\hat{\mathcal{A}}\|_{\mathrm{F}}} + \sigma \cdot \frac{\mathcal{E}}{\|\mathcal{E}\|_{\mathrm{F}}},$$

其中, $\hat{\mathcal{A}}$ 按照无噪声情形中的相同规则生成, 其 TR 秩为 $\mathbf{r}^* = (3, 3, 3)$, 并且 $n_1 = n_2 = n_3 = 100$. $\mathcal{E}$ 是一个张量, 其元素服从正态分布 $\mathcal{N}(0, 1)$. 参数 σ 用于刻画张量中的噪声水平. 理想情况下, 当算法终止时, 相对误差 ε_Ω 和 ε_Γ 应当接近噪声水平 σ. 我们设置 $\sigma = 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}, 10^{-8}$, 并取正则化参数 $\lambda = 10^{-12}$. 由图 3-2 可观察到, 至少需要 8000 个采样点才能恢复无噪张量, 即采样率 p 应满足 $8000/100^3 = 0.008$. 因此, 我们选择采样率 $p = 0.05$. 时间预算设为 120 秒, 最大迭代次数为 1000.

图 3-3 展示了基于 TR 分解的算法在不同噪声水平下的恢复性能. 除 TR-GD 外, 其余算法均在给定时间预算内成功恢复了潜在的低秩张量. 由于 TR 秩参数 $\mathbf{r}$

取为与数据张量 $\mathcal{A}$ 的真实 TR 秩 $\mathbf{r}^*$ 完全一致, 由表 3-2 可见, 测试误差与训练误差处于同一量级, 并且略大于训练误差.

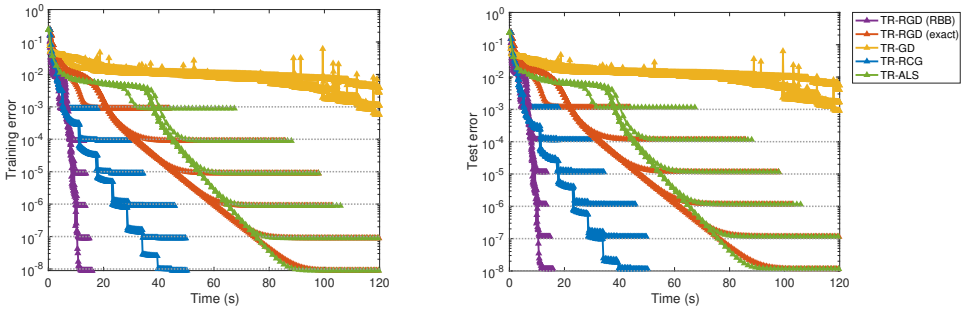


图 3-3 基于 TR 分解的算法在不同噪声水平下的恢复能力. 左图: 训练误差. 右图: 测试误差.

Figure 3-3 Reconstruction ability of TR-based algorithms under different noise levels. Left: training error. Right: test error.

表 3-2 含噪声情形下的训练和测试误差

Table 3-2 Training and test errors of noisy case.

σ	误差类型	TR-RGD (RBB)	TR-RGD (exact)	TR-RCG	TR-GD	TR-ALS
1e-3	训练误差	8.9142e-04	8.9142e-04	8.9142e-04	1.1003e-03	8.9142e-04
	测试误差	1.1471e-03	1.1471e-03	1.1472e-03	1.4066e-03	1.1470e-03
1e-4	训练误差	8.9154e-05	8.9154e-05	8.9154e-05	1.1129e-03	8.9154e-05
	测试误差	1.1458e-04	1.1459e-04	1.1461e-04	1.6735e-03	1.1458e-04
1e-5	训练误差	8.9155e-06	8.9155e-06	8.9155e-06	5.7432e-04	8.9155e-06
	测试误差	1.1457e-05	1.1457e-05	1.1461e-05	9.1646e-04	1.1457e-05
1e-6	训练误差	8.9155e-07	8.9155e-07	8.9155e-07	2.7753e-03	8.9155e-07
	测试误差	1.1457e-06	1.1457e-06	1.1458e-06	4.0237e-03	1.1458e-06
1e-7	训练误差	8.9155e-08	8.9155e-08	8.9155e-08	3.8170e-03	8.9156e-08
	测试误差	1.1456e-07	1.1456e-07	1.1457e-07	5.4727e-03	1.1462e-07
1e-8	训练误差	8.9156e-09	8.9156e-09	8.9156e-09	9.5924e-04	8.9289e-09
	测试误差	1.1450e-08	1.1455e-08	1.1451e-08	1.4597e-03	1.1512e-08

3.5.2　电影评分数据集 MovieLens 1M

我们在电影评分数据集 MovieLens 1M 数据集上考虑了一个真实世界的张量补全问题[6]. 该数据集包含从 1997 年 9 月 19 日到 1998 年 4 月 22 日期间, 6040 名用户对 3952 部电影给出的约一百万条评分记录. 我们以一周作为一个时间周期, 将电影评分数据构造成一个大小为 $6040 \times 3952 \times 150$ 的三阶张量 $\mathcal{A}$. 我们随机选取已知评分记录中的 80% 作为训练集 Ω, 其余 20% 作为测试集 Γ. 我们将所提出的 TR-RGD 和 TR-RCG 算法与其他张量补全算法进行比较, 包括 TR-GD、TR-ALS、TT-RCG、CP-WOPT 以及 geomCG. 实验中采用的参数设置如下：TR 秩取 $\mathbf{r} = (6, 10, 3)$ 和 $\mathbf{r} = (6, 6, 6)$, 正则化参数设为 $\lambda = 1$, 时间预算设为 500 秒.

[6]可从 https://grouplens.org/datasets/movielens/1m/中获取.

为了保证不同搜索空间中参数数量具有可比性, 我们选择 TT 秩为 $(1, 10, 10, 1)$ 和 $(1, 9, 9, 1)$, Tucker 秩为 $(60, 30, 18)$ 和 $(36, 36, 36)$, 以及 CP 秩为 49 和 36.

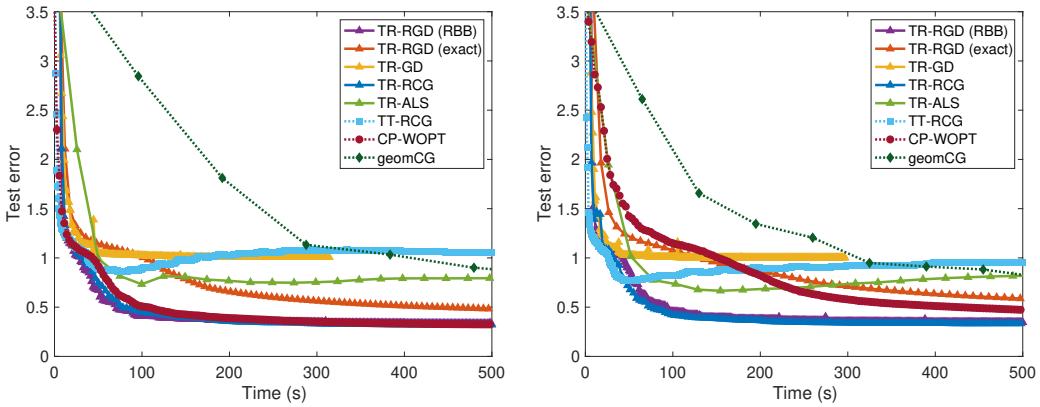


图 3-4 在 MovieLens 1M 上的测试误差. 左图: $\mathbf{r} = (6, 6, 6)$. 右图: $\mathbf{r} = (6, 10, 3)$.

Figure 3-4 Test error on MovieLens 1M dataset. Left: $\mathbf{r} = (6, 6, 6)$. Right: $\mathbf{r} = (6, 10, 3)$.

两种不同秩选择下的测试误差如图 3-4 所示. 可以观察到, 所提出的 TR-RGD(RBB) 和 TR-RCG 方法与 CP-WOPT 的性能相当, 这是因为 CP-WOPT 同样利用了二阶信息. 相比其他算法, 它们表现出更快的收敛速度以及更低的测试误差. 此外, TR-RGD(exact) 相较于 TR-RGD(RBB) 需要消耗更多的计算时间. 为简洁起见, 在后续实验中我们仅考虑 TR-RGD(RBB), 并将 “TR-RGD(RBB)” 简记为 “TR-RGD”.

相比于直接的计算, 算法 10提供了一种高效的计算梯度的方式. 表 3-3展示了在电影评分数据集 MovieLens 1M 上不同梯度计算方法的加速结果. 其中 “直接方法” 为精确计算 $\mathbf{W}_{\neq k}$ 并执行大规模矩阵乘法. 表中的结果表明所提出的高效梯度计算方法确实有助于所提出的黎曼预条件方法在大规模问题中的应用.

表 3-3 在电影评分数据集 MovieLens 1M 上不同梯度计算方法的加速结果.

Table 3-3 Speedup of efficient gradient computation on MovieLens 1M dataset.

迭代步数	时间 (秒)		加速比	在 Ω 上的相对误差 (ε_Ω)		在 Γ 上的相对误差 (ε_Γ)	
	直接方法	加速方法		直接方法	加速方法	直接方法	加速方法
1	266.6065	5.9717	-	4.0151	4.0151	4.0233	4.0233
21	4880.3810	37.7889	145.5539×	0.7636	0.7636	0.8674	0.8674
41	9400.1950	69.2155	144.6929×	0.3979	0.3979	0.5268	0.5268
61	13957.4616	100.5884	144.8814×	0.2796	0.2796	0.4084	0.4084
81	18518.4900	132.6191	144.2527×	0.2612	0.2612	0.3880	0.3880

3.5.3 高光谱图像

在本实验中, 我们考虑高光谱图像的张量补全问题. 高光谱图像可以表示为一个三阶张量 $\mathcal{A} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$. 其中, 第三个模态对应于光的 n_3 个不同波长, 而第一个模态和第二个模态分别表示在不同波长下的空间反射率分布. 我们从

Foster 提供的 “50 reduced hyperspectral reflectance images” 数据集中选取了两幅图像[7] [155]: “Ribeira Houses Shrubs”, 简称为 “Ribeira”, 其大小为 $249 \times 330 \times 33$; “Bom Jesus Bush”, 简称为 “Bush”, 其大小为 $250 \times 330 \times 33$. 我们按照给定的教程[8] 将高光谱图像转换为 RGB 图像表示. 图 3-5 展示了这两幅高光谱图像对应的 RGB 可视化结果.

图 3-5 高光谱图像. 左图: “Ribeira House Shrubs”. 右图: “Bom Jesus Bush”.

Figure 3-5 Hyperspectral images. Left: “Ribeira House Shrubs”. Right: “Bom Jesus Bush”.

为了评估图像补全的恢复性能, 我们采用峰值信噪比 (peak signal-to-noise ratio, PSNR) 来衡量两幅图像之间的相似性, 其定义为

$$\mathrm{PSNR} := 10\log_{10}\left(\frac{\max(\mathcal{A})^2}{\mathrm{MSE}}\right) = 10\log_{10}\left(n_1n_2n_3\frac{\max(\mathcal{A})^2}{\|\mathcal{X}-\mathcal{A}\|_{\mathrm{F}}^2}\right),$$

其中, $\max(\mathcal{A})$ 表示张量 $\mathcal{A}$ 中的最大像素值, MSE 为均方误差, 其定义为 $\mathrm{MSE} := \|\mathcal{X}-\mathcal{A}\|_{\mathrm{F}}^2/(n_1n_2n_3)$. 此外, 我们还展示相对误差

$$\mathrm{relerr}(\mathcal{X}) := \frac{\|\mathcal{X}-\mathcal{A}\|_{\mathrm{F}}}{\|\mathcal{A}\|_{\mathrm{F}}}.$$

在数值实验中, 我们在将恢复后的高光谱图像转换为 RGB 图像后, 计算其与原始图像之间的 PSNR. 给定采样率 p, 采样集合 Ω 的构造方式与小节 3.3.3 相同. 初始值 $\mathcal{X}^{(0)}$ 的各个元素均服从正态分布 $\mathcal{N}(0,1)$. 我们设置 TR 秩为 $\mathbf{r} = (7,16,7)$, 遵循 [15] 设置 Tucker 秩为 $(65,65,7)$, TT 秩为 $(1,15,15,1)$, 以及 CP 秩为 110. 不同张量格式对应的搜索空间规模大致相当. 此外, 我们还比较了基于核范数的算法 HaLRTC [135]. 最大迭代次数设置为 200.

两幅高光谱图像的补全数值结果如图 3-6 和表 3-4 所示. 对于 “Ribeira” 图像, 其恢复的主要难点在于建筑物周围的灌木区域, 这会导致不同算法在恢复质量上有较大的差异; 而 “Bush” 图像的恢复效果则主要取决于叶片细节的恢复情况. 图 3-6 展示了在不同采样率 $p = 0.1, 0.3, 0.5$ 下的恢复结果. 可以观察到, 基于 TR 的算法 (TR-RGD、TR-RCG 以及 TR-ALS) 相比其他方法能够恢复出更多细节

[7] 图片来源: 链接 https://figshare.manchester.ac.uk/articles/dataset/Fifty_hyperspectral_reflectance_images_of_outdoor_scenes/14877285 中的文件 hsi_34.mat 和 hsi_46.mat.

[8] https://personalpages.manchester.ac.uk/staff/david.foster/Tutorial_HSI2RGB/Tutorial_HSI2RGB.html.

信息, 例如灌木和枝叶结构. 事实上, 这一现象可以通过 PSNR 和相对误差进行定量刻画, 详见表 3-4. 总体而言, 所提出的方法在性能上略强于 TR-ALS 算法. 在大多数情况下, 所提出的 TR-RCG 算法在所有比较方法中取得了最高的 PSNR 和最低的相对误差.

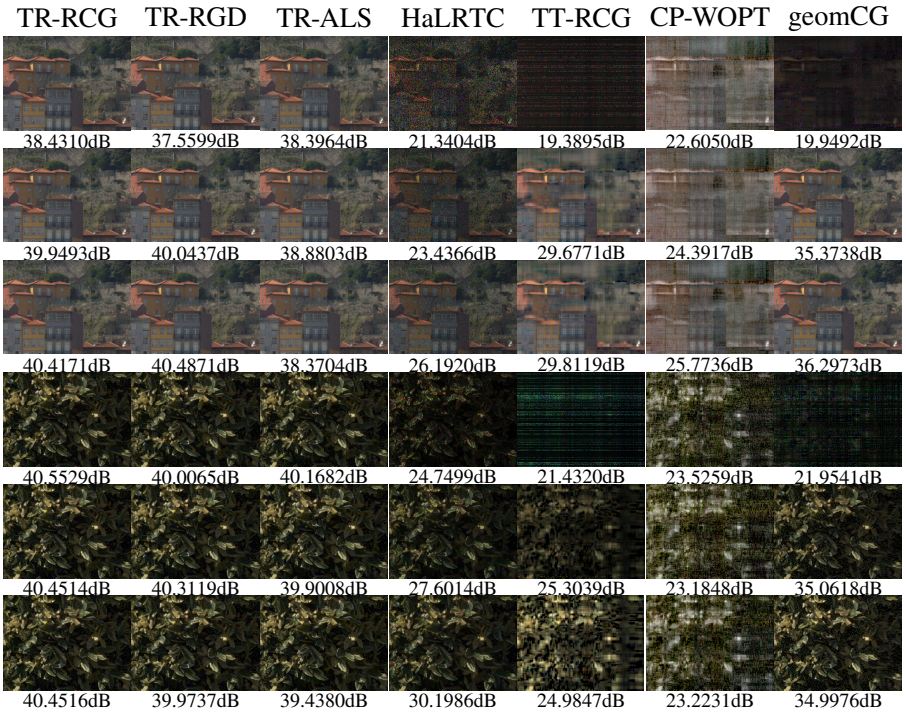


图 3-6 不同补全算法恢复结果的 RGB 表示. 前三行对应 “Ribeira” 图像在采样率 $p = 0.1, 0.3, 0.5$ 下的恢复结果 (每一行对应一个采样率); 后三行对应 “Bush” 图像在采样率 $p = 0.1, 0.3, 0.5$ 下的恢复结果 (每一行对应一个采样率). 每幅图像下方给出了对应的 PSNR 值.

Figure 3-6 RGB representations of recovered images by different completion algorithms. The first three rows represent recovery results of the “Ribeira” image, under sampling rates $p = 0.1, 0.3, 0.5$ in each row. The last three rows represent recovery results of the “Bush” image, under sampling rates $p = 0.1, 0.3, 0.5$ in each row. The PSNR is displayed under each image.

3.5.4 高维函数

存储高维函数离散后的全部函数值 $\{h : [0,1]^d \to \mathbb{R}\}$ 相当于存储一个规模巨大的张量 $\mathcal{A} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$, 这在实际中是不可接受的. 为此, 可以利用张量分解与张量补全方法, 仅通过张量部分观测的元素来恢复完整张量 $\mathcal{A}$; 相关应用可参见 [19]. 在本小节中, 我们比较基于 TR 的算法 (TR-RGD、TR-RCG、TR-ALS) 与 TT-RCG, 在由函数 h 生成的数据张量 $\mathcal{A}$ 上的补全性能. 张量 $\mathcal{A}$ 通过如下方式由 h 生成：在 $[0,1]^d$ 的第 k 个维度上, 将区间 $[0,1]$ 均匀划分为 $(n_k - 1)$ 个子区间

表 3-4 高光谱图像补全任务下的峰值信噪比与相对误差.

Table 3-4 PSNR and relative errors for completion of two hyperspectral images.

p	结果	TR-RCG	TR-RGD	TR-ALS	HaLRTC	TT-RCG	CP-WOPT	geomCG
		Ribeira House Shrubs						
0.1	PSNR	38.4310	37.5599	38.3964	21.3404	19.3895	22.6050	19.9492
	relerr	0.1058	0.1170	0.1062	0.7569	0.9476	0.6544	0.8884
0.3	PSNR	39.9493	40.0437	38.8803	23.4366	29.6771	24.3917	35.3738
	relerr	0.0888	0.0879	0.1005	0.5946	0.2899	0.5327	0.1505
0.5	PSNR	40.4171	40.4871	38.3704	26.1920	29.8119	25.7736	36.2973
	relerr	0.0842	0.0835	0.1066	0.4330	0.2854	0.4544	0.1353
		Bom Jesus Bush						
0.1	PSNR	40.5529	40.0065	40.1682	24.7499	21.4320	23.5259	21.9541
	relerr	0.1185	0.1262	0.1239	0.7310	1.0711	0.8417	1.0086
0.3	PSNR	40.4514	40.3119	39.9008	27.6014	25.3039	23.1848	35.0618
	relerr	0.1199	0.1219	0.1278	0.5265	0.6859	0.8754	0.2230
0.5	PSNR	40.4516	39.9737	39.4380	30.1986	24.9847	23.2231	34.9976
	relerr	0.1199	0.1267	0.1348	0.3904	0.7115	0.8715	0.2247

($k \in [d]$), 并定义

$$\mathcal{A}(i_1, i_2, \ldots, i_d) = h\left(\frac{i_1 - 1}{n_1 - 1}, \frac{i_2 - 1}{n_2 - 1}, \ldots, \frac{i_d - 1}{n_d - 1}\right),\ i_k \in [n_k],\ k \in [d].$$

我们考虑如下两个函数 [60, §5.4],

$$h_1 : [0,1]^d \to \mathbb{R}, \quad h_1(\mathbf{x}) := \exp(-\|\mathbf{x}\|), \quad \text{和}$$
$$h_2 : [0,1]^d \to \mathbb{R}, \quad h_2(\mathbf{x}) := \frac{1}{\|\mathbf{x}\|},$$

实验中取 $d = 4$, $n_1 = n_2 = n_3 = n_4 = 20$, 采样率设为 $p = 0.001, 0.005, 0.01, 0.05, 0.1$. 样本 Ω 的构造方式与小节 3.5.1 中相同. 我们按照 [60] 中的方法, 对 TT 和 TR 算法均采用秩增策略. 为保证不同张量形式下公平比较, TT-RCG 的最大秩设为 $(1, 5, 5, 5, 1)$, 而 TR 类算法的最大秩设为 $(4, 4, 4, 4)$; 详见表3-1. 停止准则采用 3.5 节中的设置. 此外, 由于使用了秩增策略, 当满足以下任一条件时算法也会终止: 1) 在固定秩下的搜索达到最大迭代步数 50; 2) 秩增到了最大允许的秩; 3) 在当前点处, 沿任一模态均无法接受进一步的秩提升.

恢复高维函数 h_1 和 h_2 的数值结果汇总于表 3-5中. 结果表明, 所有算法的性能总体上是相当的, 基于 TR 的算法在性能上与 TT-RCG 相当. 在所有 TR 类算法中, TR-RCG 在大多数实验中表现最优.

表 3-5 在高维函数补全上的测试误差.

Table 3-5 Test errors for high-dimensional functions.

p	$\exp(-\|\mathbf{x}\|)$				$1/\|\mathbf{x}\|$			
	TR-RGD	TR-RCG	TR-ALS	TT-RCG	TR-RGD	TR-RCG	TR-ALS	TT-RCG
0.001	8.0884e-2	7.4157e-2	7.4161e-2	1.3445e-1	1.7531e-1	1.8106e-1	1.8081e-1	2.6876e-1
0.005	7.3505e-3	8.7366e-3	9.2121e-3	1.5904e-2	3.4428e-2	2.9218e-2	3.2090e-2	1.2899e-1
0.01	6.2650e-3	9.7247e-4	1.8737e-3	4.1233e-3	2.5230e-2	1.7676e-2	1.8697e-2	3.4675e-2
0.05	3.8862e-4	1.5019e-4	1.8218e-4	2.2991e-4	3.8510e-3	3.6002e-3	5.2173e-3	3.5697e-3
0.1	1.2251e-4	5.9871e-5	6.8898e-5	8.2512e-5	7.7886e-4	2.9423e-4	6.0423e-4	7.4727e-4

3.6 本章小结

我们基于张量环分解, 提出了用于张量补全问题的黎曼预条件算法. 预条件效应来源于定义在由核张量的模态 2 展平矩阵所构成的矩阵乘积空间上的一种度量. 我们发现, 直接计算黎曼梯度需要进行大规模矩阵乘法, 在实际中代价过高. 为此, 我们采用了一种高效的计算策略, 在不显式构造大矩阵的情况下计算黎曼梯度. 所提出的算法具有全局收敛性保证, 并且在人工合成数据和真实数据集上的数值实验均展示出了良好的性能.

第 4 章　Tucker 张量代数簇上的低秩优化方法

4.1　引言

Tucker 分解, 也被称为高阶主成分分析 [55] 或者是高阶奇异值分解 [58], 提供了一种通过张量的展平矩阵研究张量低秩性的工具. 此外, Tucker 分解在矩阵情形下即为数值线性代数中非常重要的奇异值分解. 在本章中, 我们考虑如下的在有界 Tucker 秩约束下的低秩张量优化问题

$$\begin{aligned} \min_{\mathcal{X}} \quad & f(\mathcal{X}) \\ \text{s.t.} \quad & \mathcal{X} \in \mathcal{M}_{\leq \mathbf{r}} := \{\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d} : \operatorname{rank}_{\mathrm{tc}}(\mathcal{X}) \leq \mathbf{r}\}, \end{aligned} \tag{4-1}$$

这里 $f : \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d} \to \mathbb{R}$ 是一个光滑函数, $\mathbf{r} = (r_1, r_2, \ldots, r_d)$ 是 d 维整数数组, 以及张量 $\mathcal{X}$ 的 Tucker 秩被记为 $\operatorname{rank}_{\mathrm{tc}}(\mathcal{X})$. 由于可行集 $\mathcal{M}_{\leq \mathbf{r}}$ 可以通过模态 k 展平矩阵的所有 $(r_k + 1)$ 阶余子式等于 0 构造 ($k \in [d]$), 故 $\mathcal{M}_{\leq \mathbf{r}}$ 是一个代数簇. 因此我们称集合 $\mathcal{M}_{\leq \mathbf{r}}$ 为 Tucker 张量代数簇. 对于张量链分解, 同样可以构造一个不同的张量代数簇. Kutschan [83] 针对张量链分解代数簇提供了一种显式参数化. 在近期 Vermeylen 等人 [85] 针对张量补全问题, 利用张量链分解代数簇的几何结构设计出一种秩自适应方法.

低秩 Tucker 张量优化问题相比矩阵优化问题要困难得多, 其根本原因在于 Tucker 张量所具有的复杂几何结构. 一方面, 与矩阵代数簇 [63] 或固定秩 Tucker 张量流形 [57] 已被充分研究的几何性质不同, 研究者对 Tucker 张量代数簇的几何结构知之甚少, 例如 $\mathcal{M}_{\leq \mathbf{r}} \backslash \mathcal{M}_{\mathbf{r}}$ 中秩亏点处的切锥尚不清楚. 另一方面, $\mathcal{M}_{\leq \mathbf{r}}$ 可以看作沿不同模态展平得到的 d 个张量化矩阵代数簇的交集, 但 $\mathcal{M}_{\leq \mathbf{r}}$ 的切锥与各个模态展平矩阵代数簇切锥之间的关系并不明确. 因此, Tucker 张量代数簇的几何结构难以直接由矩阵代数簇的情形加以推广. 目前, 关于 Tucker 张量代数簇的几何与优化研究仍然十分有限. Luo 和 Qi [86] 通过利用法锥的一个子集研究了问题 (4-1) 的最优性条件, 但切锥的具体刻画仍然未知. Tucker 张量代数簇几何结构的不明确, 严重制约了在其上设计有效优化算法. 综上所述, 我们的研究动机在于寻求切锥的显式参数化, 并基于已有理论结果, 设计针对问题 (4-1) 的几何优化方法.

此外, 在低秩优化中, “如何选择合适的秩参数 $\mathbf{r}$” 这一问题在实际应用中同样具有重要意义. 已有研究 [78, 79, 117] 表明, 低秩优化算法的数值表现对秩参数 $\mathbf{r}$ 的选择往往十分敏感. 选择较大的秩参数 $\mathbf{r}$ 可以扩大搜索空间, 从而有可能获得更优的解; 然而, 当 $\mathbf{r}$ 过大时, 会显著增加存储和计算开销, 并且由于 $\mathcal{M}_{\mathbf{r}}$ 本身并非闭集, 优化算法还可能收敛到秩亏点. 基于上述困难, 我们进一步希望设计适用于 Tucker 张量代数簇优化的秩自适应方法, 使算法能够在迭代过程中自动选择合适的秩参数.

本章主要内容 本章围绕 Tucker 张量代数簇 $\mathcal{M}_{\leq \mathbf{r}}$ 的几何结构与优化方法展开研究. 首先, 我们给出了矩阵代数簇切锥与 Tucker 张量流形切空间的等价表述, 并在此基础上构造了 $\mathcal{M}_{\leq \mathbf{r}}$ 的切锥的显式刻画, 为在 Tucker 张量代数簇上设计算法提供了几何基础.

由于切锥上的度量投影不具有显式表达, 本章提出了一类近似投影算子 $\tilde{\mathrm{P}}$, 并据此提出了梯度相关的近似投影方法 (GRAP)

$$\mathcal{X}^{(t+1)} = \mathrm{P}^{\mathrm{HO}}_{\leq \mathbf{r}} \left(\mathcal{X}^{(t)} + s^{(t)} \tilde{\mathrm{P}}_{\mathrm{T}_{\mathcal{X}^{(t)}} \mathcal{M}_{\leq \mathbf{r}}} (-\nabla f(\mathcal{X}^{(t)})) \right),$$

该方法可视为固定秩 Tucker 张量流形上黎曼梯度方法的推广, 具有全局收敛性, 并可利用 Łojasiewicz 不等式证明局部收敛性. 进一步地, 通过利用切锥的部分结构信息, 本章提出了一种新的近似投影算子 $\hat{\mathrm{P}}$, 从而构造了无需回缩的梯度相关近似投影方法 (rfGRAP)

$$\mathcal{X}^{(t+1)} = \mathcal{X}^{(t)} + s^{(t)} \hat{\mathrm{P}}_{\mathrm{T}_{\mathcal{X}^{(t)}} \mathcal{M}_{\leq \mathbf{r}}} (-\nabla f(\mathcal{X}^{(t)})),$$

在保持可行性的同时显著降低了计算成本. 此外, 基于切锥的几何刻画, 本章提出了一种 Tucker 秩自适应方法 (TRAM), 通过在固定秩流形上的线搜索以及秩降低与秩提升策略的结合, 在迭代过程中自动识别合适的秩参数.

在人工合成数据、高光谱图像和电影评分数据上的数值实验表明, 所提出的 GRAP、rfGRAP 以及 TRAM 方法在不同秩参数选择下均优于或不逊于现有方法, 其中 TRAM 方法在实际问题中展现了良好的秩选择能力和稳定的数值性能.

4.2 张量代数簇的几何

首先, 我们回顾固定 Tucker 秩张量所构成流形的几何, 并给出新的视角. 接下来, 我们给出 Tucker 张量代数簇在任一点处切锥的显式表达式. 最后, 我们提出一种往切锥上的近似投影.

4.2.1 矩阵代数簇切锥的新参数化表示

我们提出了一种矩阵代数簇切锥的新的降维参数化表示. 具体而言, 对于任意切锥 $\mathrm{T}_{\mathbf{X}} \mathbb{R}^{m \times n}_{\leq \underline{r}}$ 中的元素 Ξ, 存在矩阵 $\mathbf{C} \in \mathbb{R}^{\underline{r} \times \underline{r}}$, $\mathbf{D} \in \mathbb{R}^{(m-\underline{r}) \times \underline{r}}$, $\mathbf{E} \in \mathbb{R}^{\underline{r} \times (n-\underline{r})}$, 以及 $\mathbf{F} \in \mathbb{R}^{(m-\underline{r}) \times (n-\underline{r})}_{\leq (r-\underline{r})}$, 使得

$$\Xi = \mathbf{U}\mathbf{C}\mathbf{V}^{\top} + \mathbf{U}^{\perp}\mathbf{D}\mathbf{V}^{\top} + \mathbf{U}\mathbf{E}(\mathbf{V}^{\perp})^{\top} + \mathbf{U}^{\perp}\mathbf{F}(\mathbf{V}^{\perp})^{\top}. \tag{4-2}$$

接下来, 我们对 Ξ 进行进一步分解. 具体地, 对 $\mathbf{F}$ 作分解 $\mathbf{F} = \tilde{\mathbf{U}}\mathbf{S}\tilde{\mathbf{V}}^{\top}$, 其中 $\tilde{\mathbf{U}} \in \mathrm{St}(r-\underline{r}, m-\underline{r})$, $\tilde{\mathbf{V}} \in \mathrm{St}(r-\underline{r}, n-\underline{r})$, $\mathbf{S} \in \mathbb{R}^{(r-\underline{r}) \times (r-\underline{r})}$, 这里 $\mathbf{S}$ 不一定是满秩的. 由于 $r - \underline{r} \leq \min\{m-\underline{r}, n-\underline{r}\}$, 存在矩阵 $\tilde{\mathbf{U}}_2 \in \mathrm{St}(m-r, m-\underline{r})$ 和 $\tilde{\mathbf{V}}_2 \in \mathrm{St}(n-r, n-\underline{r})$, 使得 $[\tilde{\mathbf{U}}\ \tilde{\mathbf{U}}_2] \in \mathcal{O}(m-\underline{r})$ 且 $[\tilde{\mathbf{V}}\ \tilde{\mathbf{V}}_2] \in \mathcal{O}(n-\underline{r})$. 事实上, 我们有 $\mathrm{span}(\tilde{\mathbf{U}}_2) = \mathrm{span}(\tilde{\mathbf{U}})^{\perp}$

以及 $\operatorname{span}(\tilde{\mathbf{V}}_2) = \operatorname{span}(\tilde{\mathbf{V}})^{\perp}$. 由此, 我们可以得到 Ξ 的一个新的等价 (降维) 参数化表示:

$$\begin{aligned}
\Xi &= \mathbf{U}\mathbf{C}\mathbf{V}^{\top} + \mathbf{U}^{\perp}\mathbf{D}\mathbf{V}^{\top} + \mathbf{U}\mathbf{E}(\mathbf{V}^{\perp})^{\top} + \mathbf{U}^{\perp}\tilde{\mathbf{U}}\mathbf{S}\tilde{\mathbf{V}}^{\top}(\mathbf{V}^{\perp})^{\top} \\
&= \mathbf{U}\mathbf{C}\mathbf{V}^{\top} + \mathbf{U}^{\perp}\mathbf{D}\mathbf{V}^{\top} + \mathbf{U}\mathbf{E}(\mathbf{V}^{\perp})^{\top} + \mathbf{U}_1\mathbf{S}\mathbf{V}_1^{\top} \\
&= \mathbf{U}\mathbf{C}\mathbf{V}^{\top} + \mathbf{U}^{\perp}[\tilde{\mathbf{U}}\ \tilde{\mathbf{U}}_2][\tilde{\mathbf{U}}\ \tilde{\mathbf{U}}_2]^{\top}\mathbf{D}\mathbf{V}^{\top} + \mathbf{U}\mathbf{E}[\tilde{\mathbf{V}}\ \tilde{\mathbf{V}}_2][\tilde{\mathbf{V}}\ \tilde{\mathbf{V}}_2]^{\top}(\mathbf{V}^{\perp})^{\top} + \mathbf{U}_1\mathbf{S}\mathbf{V}_1^{\top} \\
&= \begin{bmatrix}\mathbf{U} & \mathbf{U}_1 & \mathbf{U}_2\end{bmatrix}\begin{bmatrix}\mathbf{C} & \mathbf{E}\tilde{\mathbf{V}} & \mathbf{E}\tilde{\mathbf{V}}_2 \\ \tilde{\mathbf{U}}^{\top}\mathbf{D} & \mathbf{S} & 0 \\ \tilde{\mathbf{U}}_2^{\top}\mathbf{D} & 0 & 0\end{bmatrix}\begin{bmatrix}\mathbf{V} & \mathbf{V}_1 & \mathbf{V}_2\end{bmatrix}^{\top},
\end{aligned}$$

其中 $\mathbf{U}_1 := \mathbf{U}^{\perp}\tilde{\mathbf{U}}$, $\mathbf{U}_2 := \mathbf{U}^{\perp}\tilde{\mathbf{U}}_2$, $\mathbf{V}_1 := \mathbf{V}^{\perp}\tilde{\mathbf{V}}$, 和 $\mathbf{V}_2 := \mathbf{V}^{\perp}\tilde{\mathbf{V}}_2$. 图 4-1 给出了 $\mathrm{T}_{\mathbf{X}}\mathbb{R}^{m\times n}_{\leq r}$ 中元素的一个等价示意图, 其中虚线方块表示 $\mathbb{R}^{r\times r}$ 中的任意矩阵. 该示意对于后续构造 Tucker 张量代数簇的切锥显式表达至关重要; 参见定理 4.2.

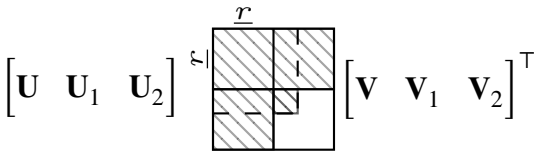


图 4-1 $\mathrm{T}_{\mathbf{X}}\mathbb{R}^{m\times n}_{\leq r}$ 中元素的示意图, 其中参数 $\mathbf{U}_1 \in \mathrm{St}(r-\underline{r}, m)$、$\mathbf{V}_1 \in \mathrm{St}(r-\underline{r}, n)$、$\mathbf{U}_2 \in \mathrm{St}(m-r, m)$、$\mathbf{V}_2 \in \mathrm{St}(n-r, n)$, 并且满足 $[\mathbf{U}\ \mathbf{U}_1\ \mathbf{U}_2] \in \mathcal{O}(m)$ 以及 $[\mathbf{V}\ \mathbf{V}_1\ \mathbf{V}_2] \in \mathcal{O}(n)$.

Figure 4-1 Illustration of an element in $\mathrm{T}_{\mathbf{X}}\mathbb{R}^{m\times n}_{\leq r}$ with parameters $\mathbf{U}_1 \in \mathrm{St}(r-\underline{r}, m)$, $\mathbf{V}_1 \in \mathrm{St}(r-\underline{r}, n)$, $\mathbf{U}_2 \in \mathrm{St}(m-r, m)$, $\mathbf{V}_2 \in \mathrm{St}(n-r, n)$ satisfying $[\mathbf{U}\ \mathbf{U}_1\ \mathbf{U}_2] \in \mathcal{O}(m)$ and $[\mathbf{V}\ \mathbf{V}_1\ \mathbf{V}_2] \in \mathcal{O}(n)$.

值得注意的是, 我们还可以通过对式 (4-2) 中的 $\mathbf{F}$ 采用满秩分解的方式, 进一步构造一种紧凑的参数化. 具体而言, 我们考虑分解 $\mathbf{F} = \hat{\mathbf{U}}\hat{\mathbf{S}}\hat{\mathbf{V}}^{\top}$, 其中 $\hat{\mathbf{U}} \in \mathrm{St}(\ell, m-\underline{r})$、$\hat{\mathbf{V}} \in \mathrm{St}(\ell, n-\underline{r})$, $\hat{\mathbf{S}} \in \mathbb{R}^{\ell\times\ell}_{\ell}$ 为一个满秩矩阵, 且 $\ell = \operatorname{rank}(\mathbf{F})$. 由此, 我们得到如下紧凑参数化形式

$$\Xi = \mathbf{U}\mathbf{C}\mathbf{V}^{\top} + \mathbf{U}^{\perp}\mathbf{D}\mathbf{V}^{\top} + \mathbf{U}\mathbf{E}(\mathbf{V}^{\perp})^{\top} + \hat{\mathbf{U}}_1\hat{\mathbf{S}}\hat{\mathbf{V}}_1^{\top}, \tag{4-3}$$

其中 $\hat{\mathbf{U}}_1 := \mathbf{U}^{\perp}\hat{\mathbf{U}}$ 和 $\hat{\mathbf{V}}_1 := \mathbf{V}^{\perp}\hat{\mathbf{V}}$. 由于 $\hat{\mathbf{S}}$ 是满秩的, 式 (4-3) 在对 $\hat{\mathbf{U}}_1$ 和 $\hat{\mathbf{V}}_1$ 施加右正交群作用的意义下是唯一的. 事实上, 我们注意到 $\mathrm{P}^{\perp}_{\mathbf{U}}\Xi\mathrm{P}^{\perp}_{\mathbf{V}} = \hat{\mathbf{U}}_1\hat{\mathbf{S}}\hat{\mathbf{V}}_1^{\top}$, 这里 $\mathrm{P}_{\mathbf{U}} := \mathbf{U}\mathbf{U}^{\top}$, $\mathrm{P}^{\perp}_{\mathbf{U}} := \mathbf{I}_m - \mathrm{P}_{\mathbf{U}}$, $\mathrm{P}_{\mathbf{V}} := \mathbf{V}\mathbf{V}^{\top}$, 以及 $\mathrm{P}^{\perp}_{\mathbf{V}} := \mathbf{I}_n - \mathrm{P}_{\mathbf{V}}$. 因此我们有

$$\operatorname{span}(\hat{\mathbf{U}}_1) = \operatorname{span}(\mathrm{P}^{\perp}_{\mathbf{U}}\Xi\mathrm{P}^{\perp}_{\mathbf{V}}) \quad \text{和} \quad \operatorname{span}(\hat{\mathbf{V}}_1) = \operatorname{span}((\mathrm{P}^{\perp}_{\mathbf{U}}\Xi\mathrm{P}^{\perp}_{\mathbf{V}})^{\top}),$$

这表明上述子空间是唯一确定的.

注. 我们注意到, 切锥还可以被分解为

$$\mathrm{T}_{\mathbf{X}}\mathbb{R}^{m\times n}_{\leq r} = \mathrm{T}_{\mathbf{X}}\mathbb{R}^{m\times n}_{\underline{r}} + \mathrm{N}_{\leq(r-\underline{r})}(\mathbf{X}),$$

其中

$$\mathrm{N}_{\leq(r-\underline{r})}(\mathbf{X}) := \left\{ \mathbf{N} \in \mathrm{N}_{\mathbf{X}}\mathbb{R}^{m\times n}_{\underline{r}} \,:\, \mathrm{rank}(\mathbf{N}) \leq (r-\underline{r}) \right\}.$$

正如 [79] 所指出的, 对于 $\mathbf{X} \in \mathbb{R}^{m\times n}_{\underline{r}}$ 且 $\underline{r} < r$, 以及任意非零向量 $\mathbf{V} \in \mathrm{N}_{\leq(r-\underline{r})}(\mathbf{X})\backslash\{0\}$, 当 $s > 0$ 时, 有

$$\mathrm{rank}(\mathbf{X} + s\mathbf{V}) \in (\underline{r}, r].$$

该性质可用于提升 $\mathbf{X}$ 的秩. 在张量情形下也可以得到类似的结论; 详见第 4.5.3 节.

4.2.2 Tucker 张量流形的切空间的等价刻画

在本节中, 我们给出 Tucker 张量流形 $\mathcal{M}_{\underline{\mathbf{r}}}$ 的切空间的等价刻画. 给定一个张量 $\mathcal{X} \in \mathbb{R}^{n_1\times n_2\times\cdots\times n_d}$ 以及 Tucker 秩 $\mathrm{rank}_{\mathrm{tc}}(\mathcal{X}) = \underline{\mathbf{r}}$ 和 Tucker 分解 $\mathcal{X} = \mathcal{G}\times_{k=1}^{d}\mathbf{U}_k$, 根据式 (1-10) 我们可以得到 $\mathrm{span}(\dot{\mathbf{U}}_k) \subseteq \mathrm{span}(\mathbf{U}_k)^{\perp}$, 也就是说集合 $\{\dot{\mathbf{U}}_k \in \mathbb{R}^{n_k\times\underline{r}_k} : \dot{\mathbf{U}}_k^{\top}\mathbf{U}_k = 0\}$ 可以被表达为 $\{\dot{\mathbf{U}}_k = \mathbf{U}_k^{\perp}\dot{\mathbf{R}}_k : \dot{\mathbf{R}}_k \in \mathbb{R}^{(n_k-\underline{r}_k)\times\underline{r}_k}\}$, 这里 $\mathbf{U}_k^{\perp}$ 的定义参见命题 1.1 以及 $k \in [d]$. 接下来, 对于任意的切向量 $\mathcal{V} \in \mathrm{T}_{\mathcal{X}}\mathcal{M}_{\underline{\mathbf{r}}}$, 我们有

$$\begin{aligned}
\mathcal{V} &= \dot{\mathcal{G}}\times_1\mathbf{U}_1\cdots\times_d\mathbf{U}_d + \sum_{k=1}^{d}\mathcal{G}\times_k\dot{\mathbf{U}}_k\times_{j\neq k}\mathbf{U}_j \\
&= \dot{\mathcal{G}}\times_1\mathbf{U}_1\cdots\times_d\mathbf{U}_d + \sum_{k=1}^{d}\mathcal{G}\times_k(\mathbf{U}_k^{\perp}\dot{\mathbf{R}}_k)\times_{j\neq k}\mathbf{U}_j \\
&= \dot{\mathcal{G}}\times_1\mathbf{U}_1\cdots\times_d\mathbf{U}_d + \sum_{k=1}^{d}(\mathcal{G}\times_k\dot{\mathbf{R}}_k)\times_k\mathbf{U}_k^{\perp}\times_{j\neq k}\mathbf{U}_j.
\end{aligned}$$

因此, 切空间 $\mathrm{T}_{\mathcal{X}}\mathcal{M}_{\underline{\mathbf{r}}}$ 可以等价地被参数化为

$$\mathrm{T}_{\mathcal{X}}\mathcal{M}_{\underline{\mathbf{r}}} = \left\{ \begin{array}{l} \dot{\mathcal{G}}\times_1\mathbf{U}_1\cdots\times_d\mathbf{U}_d + \sum_{k=1}^{d}(\mathcal{G}\times_k\dot{\mathbf{R}}_k)\times_k\mathbf{U}_k^{\perp}\times_{j\neq k}\mathbf{U}_j : \\ \dot{\mathcal{G}} \in \mathbb{R}^{\underline{r}_1\times\underline{r}_2\times\cdots\times\underline{r}_d}, \dot{\mathbf{R}}_k \in \mathbb{R}^{(n_k-\underline{r}_k)\times\underline{r}_k} \end{array} \right\}. \tag{4-4}$$

特别地, 图 4-2展示了当 $d = 3$ 时, 在 $\mathrm{T}_{\mathcal{X}}\mathcal{M}_{\underline{\mathbf{r}}}$ 中的一个切向量, 这里阴影方块 $\dot{\mathcal{G}} \in \mathbb{R}^{\underline{r}_1\times\underline{r}_2\times\cdots\times\underline{r}_d}$ 可以表示任意的张量以及白色部分表示零张量. 为简洁起见, 我们将用这些符号及颜色表达一个张量.

4.2.3 Tucker 张量代数簇的切锥

在本节中, 我们将研究 Tucker 张量代数簇 $\mathcal{M}_{\leq\mathbf{r}}$ 的切锥. 事实上, $\mathcal{M}_{\leq\mathbf{r}}$ 可以通过 d 个矩阵代数簇 $\mathbb{R}^{n_k\times n_{-k}}_{\leq r_k}$ 的展平矩阵构造, 也就是说

$$\mathcal{M}_{\leq\mathbf{r}} = \bigcap_{k=1}^{d}\mathrm{ten}_{(k)}\left(\mathbb{R}^{n_k\times n_{-k}}_{\leq r_k}\right),$$

于是我们有下面的引理.

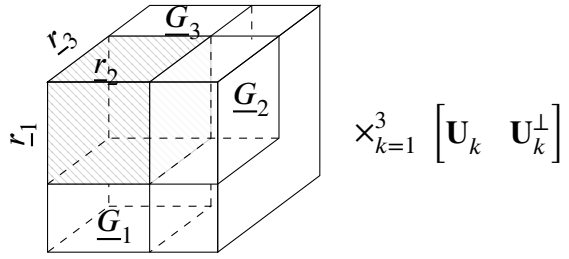


图 4-2 当 $d = 3$ 时, 在点 $\mathcal{X} = \mathcal{G} \times_{k=1}^{d} \mathbf{U}_k$ 处切空间 $\mathrm{T}_{\mathcal{X}} \mathcal{M}_{\underline{\mathbf{r}}}$ 中的切向量的图示. 这里 $\underline{G}_k := \mathcal{G} \times_k \dot{\mathbf{R}}_k$ 以及 $\dot{\mathbf{R}}_k \in \mathbb{R}^{(n_k - \underline{r}_k) \times \underline{r}_k}$ 为自由变量.

Figure 4-2 Illustration of a tangent vector in $\mathrm{T}_{\mathcal{X}} \mathcal{M}_{\underline{\mathbf{r}}}$ at $\mathcal{X} = \mathcal{G} \times_{k=1}^{d} \mathbf{U}_k$ for $d = 3$. $\underline{G}_k := \mathcal{G} \times_k \dot{\mathbf{R}}_k$ with arbitrary $\dot{\mathbf{R}}_k \in \mathbb{R}^{(n_k - \underline{r}_k) \times \underline{r}_k}$.

引理 4.1. 给定一个张量 $\mathcal{X} \in \mathcal{M}_{\leq \mathbf{r}}$, 张量代数簇 $\mathcal{M}_{\leq \mathbf{r}}$ 的切锥是沿着不同模态张量化的矩阵代数簇的切锥交集的一个子集, 也就是说,

$$\mathrm{T}_{\mathcal{X}} \mathcal{M}_{\leq \mathbf{r}} \subseteq \bigcap_{k=1}^{d} \mathrm{ten}_{(k)} \left(\mathrm{T}_{\mathbf{X}_{(k)}} \mathbb{R}_{\leq r_k}^{n_k \times n_{-k}} \right).$$

接下来, 我们给出 Tucker 张量代数簇切锥的显式参数化.

定理 4.2. 给定具有如下 *Tucker* 分解的张量 $\mathcal{X} = \mathcal{G} \times_1 \mathbf{U}_1 \cdots \times_d \mathbf{U}_d \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$ 以及 *Tucker* 秩 $\mathrm{rank_{tc}}(\mathcal{X}) = \underline{\mathbf{r}} \leq \mathbf{r}$, 在 $\mathcal{X}$ 处代数簇 $\mathcal{M}_{\leq \mathbf{r}}$ 的切锥中的任意元素可以被显式参数化为

$$\mathcal{V} = \mathcal{C} \times_{k=1}^{d} \begin{bmatrix} \mathbf{U}_k & \mathbf{U}_{k,1} \end{bmatrix} + \sum_{k=1}^{d} \mathcal{G} \times_k (\mathbf{U}_{k,2} \mathbf{R}_{k,2}) \times_{j \neq k} \mathbf{U}_j, \tag{4-5}$$

这里 $\mathcal{C} \in \mathbb{R}^{r_1 \times r_2 \times \cdots \times r_d}$, $\mathbf{R}_{k,2} \in \mathbb{R}^{(n_k - r_k) \times \underline{r}_k}$, $\mathbf{U}_{k,1} \in \mathrm{St}(r_k - \underline{r}_k, n_k)$ 和 $\mathbf{U}_{k,2} \in \mathrm{St}(n_k - r_k, n_k)$ 是对 $k \in [d]$ 满足 $[\mathbf{U}_k \ \mathbf{U}_{k,1} \ \mathbf{U}_{k,2}] \in \mathcal{O}(n_k)$ 的自由变量.

证明. 对于所有的 $\mathcal{V} = \mathcal{C} \times_{k=1}^{d} \begin{bmatrix} \mathbf{U}_k & \mathbf{U}_{k,1} \end{bmatrix} + \sum_{k=1}^{d} \mathcal{G} \times_k (\mathbf{U}_{k,2} \mathbf{R}_{k,2}) \times_{j \neq k} \mathbf{U}_j$, $t^{(i)} = 1/i$ 和 $i \in \mathbb{N}$, 我们考虑如下的序列

$$\begin{aligned} \mathcal{X}^{(i)} &= t_i \mathcal{C} \times_{k=1}^{d} \begin{bmatrix} \mathbf{U}_k + t_i \mathbf{U}_{k,2} \mathbf{R}_{k,2} & \mathbf{U}_{k,1} \end{bmatrix} + \mathcal{G} \times_{k=1}^{d} (\mathbf{U}_k + t_i \mathbf{U}_{k,2} \mathbf{R}_{k,2}) \\ &\in \bigotimes_{k=1}^{d} \mathrm{span} \left(\begin{bmatrix} \mathbf{U}_k + t_i \mathbf{U}_{k,2} \mathbf{R}_{k,2} & \mathbf{U}_{k,1} \end{bmatrix} \right) \subseteq \mathcal{M}_{\leq \mathbf{r}}, \end{aligned}$$

这里我们用到了式 (1-5) 以及 $\mathrm{rank}([\mathbf{U}_k + t_i \mathbf{U}_{k,2} \mathbf{R}_{k,2} \ \mathbf{U}_{k,1}]) \leq r_k$. 通过直接计算, 我们可以得到 $\lim_{i \to \infty} (\mathcal{X}^{(i)} - \mathcal{X}) / t^{(i)} = \mathcal{V}$ 这也就意味着 $\mathcal{V} \in \mathrm{T}_{\mathcal{X}} \mathcal{M}_{\leq \mathbf{r}}$.

接下来我们证明任意的 $\mathcal{V} \in \mathrm{T}_{\mathcal{X}} \mathcal{M}_{\leq \mathbf{r}}$ 可以通过式 (4-5)来表示. 我们记张量 $\mathcal{V}$ 的模态 k 展平矩阵为 $\Xi_k := \mathbf{V}_{(k)}$, 这里 $n_{-k} = n_1 n_2 \cdots n_d / n_k$. 通过引理 4.1 我们对所有的 $k \in [d]$ 可以得到 $\Xi_k \in \mathrm{T}_{\mathbf{X}_{(k)}} \mathbb{R}_{\leq r_k}^{n_k \times n_{-k}}$. 由于 $\mathcal{X}$ 所有的 k 模态展平矩阵 $\mathbf{X}_{(k)}$ 具有如下的分解

$$\mathbf{X}_{(k)} = \mathbf{U}_k \mathbf{G}_{(k)} \mathbf{V}_k^{\top} = \mathbf{U}_k \tilde{\mathbf{U}}_k \tilde{\Sigma}_k \tilde{\mathbf{V}}_k^{\top} \mathbf{V}_k^{\top},$$

这里 $\mathbf{G}_{(k)} = \tilde{\mathbf{U}}_k \tilde{\Sigma}_k \tilde{\mathbf{V}}_k^\top$ 是矩阵 $\mathbf{G}_{(k)}$ 的 SVD 分解, $\tilde{\mathbf{U}}_k \in \mathcal{O}(\underline{r}_k)$, 以及 $\mathbf{V}_k := (\mathbf{U}_j)^{\otimes j \neq k}$, 我们根据 $\Xi_k \in \mathrm{T}_{\mathbf{X}_{(k)}} \mathbb{R}^{n_k \times n_{-k}}_{\leq r_k}$ 以及图 4-1 的切锥元素表达可以得到, 存在变量 $\mathbf{C}_k \in \mathbb{R}^{\underline{r}_k \times \underline{r}_k}$, $\mathbf{D}_{k,1} \in \mathbb{R}^{(r_k - \underline{r}_k) \times \underline{r}_k}$, $\mathbf{D}_{k,2} \in \mathbb{R}^{(n_k - r_k) \times \underline{r}_k}$, $\mathbf{E}_{k,1} \in \mathbb{R}^{\underline{r}_k \times (r_k - \underline{r}_k)}$, $\mathbf{E}_{k,2} \in \mathbb{R}^{\underline{r}_k \times (n_{-k} - r_k)}$, $\mathbf{S}_k \in \mathbb{R}^{(r_k - \underline{r}_k) \times (r_k - \underline{r}_k)}$, $\mathbf{U}_{k,1} \in \mathrm{St}(r_k - \underline{r}_k, n_k)$, $\mathbf{U}_{k,2} \in \mathrm{St}(n_k - r_k, n_k)$, $\mathbf{V}_{k,1} \in \mathrm{St}(r_k - \underline{r}_k, n_{-k})$, $\mathbf{V}_{k,2} \in \mathrm{St}(n_{-k} - r_k, n_{-k})$ 满足 $[\mathbf{U}_k \tilde{\mathbf{U}}_k \ \mathbf{U}_{k,1} \ \mathbf{U}_{k,2}] \in \mathcal{O}(n_k)$ 和 $[\mathbf{V}_k \tilde{\mathbf{V}}_k \ \mathbf{V}_{k,1} \ \mathbf{V}_{k,2}] \in \mathcal{O}(n_{-k})$, 使得

$$\Xi_k = \begin{bmatrix} \mathbf{U}_k \tilde{\mathbf{U}}_k & \mathbf{U}_{k,1} & \mathbf{U}_{k,2} \end{bmatrix} \begin{bmatrix} \mathbf{C}_k & \mathbf{E}_{k,1} & \mathbf{E}_{k,2} \\ \mathbf{D}_{k,1} & \mathbf{S}_k & 0 \\ \mathbf{D}_{k,2} & 0 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{V}_k \tilde{\mathbf{V}}_k & \mathbf{V}_{k,1} & \mathbf{V}_{k,2} \end{bmatrix}^\top.$$

现在我们希望通过利用 Ξ_k 的结构, 恢复出式 (4-5)中变量 $\mathcal{C}$ 和 $\mathbf{R}_{k,2}$ 的表达式. 为此, 我们先证明如下的论断

$$\mathcal{W} := \mathcal{V} - \sum_{k=1}^{d} \mathcal{G} \times_k (\mathbf{U}_{k,2} \mathbf{R}_{k,2}) \times_{j \neq k} \mathbf{U}_j \in \bigotimes_{k=1}^{d} \mathrm{span}([\mathbf{U}_k \ \mathbf{U}_{k,1}]),$$

这里 $\mathbf{R}_{k,2} = \mathbf{D}_{k,2} \tilde{\Sigma}_k^{-1} \tilde{\mathbf{U}}_k^\top$. 事实上, 我们发现

$$\begin{aligned} \mathrm{P}_{\mathbf{U}_{k,2}} \mathbf{W}_{(k)} &= \mathrm{P}_{\mathbf{U}_{k,2}} \left(\mathcal{V} - \sum_{i=1}^{d} \mathcal{G} \times_i (\mathbf{U}_{i,2} \mathbf{R}_{i,2}) \times_{j \neq i} \mathbf{U}_j \right)_{(k)} \\ &= \mathrm{P}_{\mathbf{U}_{k,2}} \Xi_k - \mathbf{U}_{k,2} \mathbf{R}_{k,2} \mathbf{G}_{(k)} \mathbf{V}_k^\top \\ &= \mathbf{U}_{k,2} \mathbf{D}_{k,2} \tilde{\mathbf{V}}_k^\top \mathbf{V}_k^\top - \mathbf{U}_{k,2} \mathbf{R}_{k,2} \mathbf{G}_{(k)} \mathbf{V}_k^\top \\ &= \mathbf{U}_{k,2} \mathbf{D}_{k,2} \tilde{\Sigma}_k^{-1} \tilde{\mathbf{U}}_k^\top \tilde{\mathbf{U}}_k \tilde{\Sigma}_k \tilde{\mathbf{V}}_k^\top \mathbf{V}_k^\top - \mathbf{U}_{k,2} \mathbf{R}_{k,2} \mathbf{G}_{(k)} \mathbf{V}_k^\top \\ &= \mathbf{U}_{k,2} \mathbf{D}_{k,2} \tilde{\Sigma}_k^{-1} \tilde{\mathbf{U}}_k^\top \mathbf{G}_{(k)} \mathbf{V}_k^\top - \mathbf{U}_{k,2} \mathbf{R}_{k,2} \mathbf{G}_{(k)} \mathbf{V}_k^\top \\ &= \mathbf{U}_{k,2} \left(\mathbf{D}_{k,2} \tilde{\Sigma}_k^{-1} \tilde{\mathbf{U}}_k^\top - \mathbf{R}_{k,2} \right) \mathbf{G}_{(k)} \mathbf{V}_k^\top \\ &= 0 \end{aligned}$$

对所有的 $k \in [d]$ 成立. 这些等式来源于 $\mathbf{V}_{(k)} = \Xi_k$, $\mathrm{P}_{\mathbf{U}_{k,2}} \mathbf{U}_k = 0$, $\mathbf{V}_k = (\mathbf{U}_j)^{\otimes j \neq k}$, $\mathbf{G}_{(k)} = \tilde{\mathbf{U}}_k \tilde{\Sigma}_k \tilde{\mathbf{V}}_k^\top$, 以及 $\mathbf{R}_{k,2} = \mathbf{D}_{k,2} \tilde{\Sigma}_k^{-1} \tilde{\mathbf{U}}_k^\top$. 于是, 我们可以得到

$$\mathbf{W}_{(k)} \in \mathrm{span}(\mathbf{U}_{k,2})^\perp = \mathrm{span}([\mathbf{U}_k \tilde{\mathbf{U}}_k \ \ \mathbf{U}_{k,1}]) = \mathrm{span}([\mathbf{U}_k \ \mathbf{U}_{k,1}]),$$

因此 $\mathcal{W} \in \bigotimes_{k=1}^{d} \mathrm{span}([\mathbf{U}_k \ \mathbf{U}_{k,1}])$.

因此, 根据式 (1-5) 以及我们上述证明的论断可以得到, 存在张量 $\mathcal{C} \in \mathbb{R}^{r_1 \times r_2 \times \cdots \times r_d}$ 使得

$$\mathcal{C} \times_{k=1}^{d} \begin{bmatrix} \mathbf{U}_k & \mathbf{U}_{k,1} \end{bmatrix} = \mathcal{W} = \mathcal{V} - \sum_{k=1}^{d} \mathcal{G} \times_k (\mathbf{U}_{k,2} \mathbf{R}_{k,2}) \times_{j \neq k} \mathbf{U}_j.$$

总而言之, $\mathcal{V}$ 可以被表达为式 (4-5) 的形式. □

值得注意的是, 在定理 4.2的证明当中, 我们对 Tucker 张量代数簇切锥中的所有出现的展平矩阵, 利用了图 4-1中的切锥结构. 也就是说, 图 4-1 中新推出的矩阵代数簇切锥的表达式在推导 Tucker 张量代数簇切锥显式表达式 (4-5)中, 扮演很重要的角色. 我们在图 4-3 中对 $d = 3$ 的情形给出了切锥 $\mathrm{T}_{\mathcal{X}}\mathcal{M}_{\leq \mathbf{r}}$ 一个几何上的刻画, 这里带阴影的立方体代表了一个任意的张量 $\mathcal{C} \in \mathbb{R}^{r_1 \times \cdots \times r_d}$. 注意到式 (4-5) 当 $\underline{\mathbf{r}} = \mathbf{r}$ 时可以退化到式 (4-4) 中 Tucker 张量流形切空间的显式表达, 只需设定 $\mathbf{U}_{k,2} = \mathbf{U}_k^{\perp}$, $\mathbf{R}_{k,2} = \mathbf{R}_k$, $\mathcal{C} = \dot{\mathcal{G}}$, 并且删除变量 $\mathbf{U}_{k,1}$. 更进一步, 这些结果可以自然的退化到矩阵流形代数簇的切空间切锥结果. 事实上, 图 4-1中矩阵切锥显式表达, 可以直观地被看成图 4-3中的立方体从前往后压缩后得到的结果, 类似于“演奏手风琴” 的过程. 上述过程可以参考示意图 4-4.

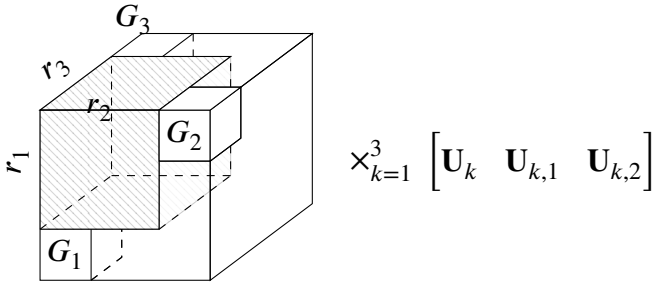


图 4-3 当 $d = 3$ 时, 在 $\mathcal{X} = \mathcal{G} \times_{k=1}^{d} \mathbf{U}_k$ 处切锥 $\mathrm{T}_{\mathcal{X}}\mathcal{M}_{\leq \mathbf{r}}$ 的几何刻画. 这里 $G_k := \mathcal{G} \times_k \mathbf{R}_{k,2}$ 以及参数 $\mathbf{R}_{k,2} \in \mathbb{R}^{(n_k - r_k) \times \underline{r}_k}$, $\mathbf{U}_{k,1} \in \mathrm{St}(r_k - \underline{r}_k, n_k)$ 和对 $k \in [d]$ 满足 $[\mathbf{U}_k\ \mathbf{U}_{k,1}\ \mathbf{U}_{k,2}] \in \mathcal{O}(n_k)$ 的参数 $\mathbf{U}_{k,2} \in \mathrm{St}(n_k - r_k, n_k)$.

Figure 4-3 Illustration of an element in $\mathrm{T}_{\mathcal{X}}\mathcal{M}_{\leq \mathbf{r}}$ at $\mathcal{X} = \mathcal{G} \times_{k=1}^{d} \mathbf{U}_k$ for $d = 3$. $G_k := \mathcal{G} \times_k \mathbf{R}_{k,2}$ with parameters $\mathbf{R}_{k,2} \in \mathbb{R}^{(n_k - r_k) \times \underline{r}_k}$, $\mathbf{U}_{k,1} \in \mathrm{St}(r_k - \underline{r}_k, n_k)$ and $\mathbf{U}_{k,2} \in \mathrm{St}(n_k - r_k, n_k)$ satisfying $[\mathbf{U}_k\ \mathbf{U}_{k,1}\ \mathbf{U}_{k,2}] \in \mathcal{O}(n_k)$ for $k \in [d]$.

切锥显式表达的唯一性　事实上, 定理 4.2所展示的切锥元素的参数化可能不是唯一的. 特别地, 我们发现对任意的 $\mathbf{Q}_{k,1} \in O(r_k - \underline{r}_k)$ 和 $\mathbf{Q}_{k,2} \in O(n_k - r_k)$,

$$\mathcal{V} = \check{\mathcal{C}} \times_{k=1}^{d} [\mathbf{U}_k\ \check{\mathbf{U}}_{k,1}] + \sum_{k=1}^{d} \mathcal{G} \times_k (\check{\mathbf{U}}_{k,2} \check{\mathbf{R}}_{k,2}) \times_{j \neq k} \mathbf{U}_j$$

是张量 $\mathcal{V} \in \mathrm{T}_{\mathcal{X}}\mathcal{M}_{\leq \mathbf{r}}$ 的一个不同的表达, 这里

$$\check{\mathcal{C}} = \mathcal{C} \times_{k=1}^{d} [\mathbf{I}_{\underline{r}_k}\ \mathbf{Q}_{k,1}^{\top}],\ \check{\mathbf{U}}_{k,1} = \mathbf{U}_{k,1}\mathbf{Q}_{k,1},\ \check{\mathbf{U}}_{k,2} = \mathbf{U}_{k,2}\mathbf{Q}_{k,2},\ \check{\mathbf{R}}_{k,2} = \mathbf{Q}_{k,2}^{\top}\mathbf{R}_{k,2}$$

以及 $k \in [d]$. 因此, 式 (4-5)对切锥元素的刻画是不唯一的. 类似于矩阵情形, 式 (4-5) 在右正交群作用于 $\mathbf{U}_{k,1}$ 和 $\mathbf{U}_{k,2}$ ($k \in [d]$) 上的意义下唯一, 如果

$$\mathrm{rank_{tc}}(\mathcal{V} \times_{k=1}^{d} \mathrm{P}_{\mathbf{U}_k}^{\perp}) = \mathrm{rank_{tc}}(\mathcal{C} \times_{k=1}^{d} [0\ \mathbf{U}_{k,1}]) = \mathbf{r} - \underline{\mathbf{r}}.$$

注意到上述的限制条件可以退化到式 (4-3)中对矩阵切锥表达唯一性的限制条件 $\mathrm{rank}(\mathrm{P}_{\mathbf{U}}^{\perp} \Xi \mathrm{P}_{\mathbf{V}}^{\perp}) = r - \underline{r}$. 更进一步, 我们通过矩阵切锥的紧参数化 (4-3)以及定理 4.2中的奇异值分解 $\mathbf{X}_{(k)} = \mathbf{U}_k \mathbf{G}_{(k)} \mathbf{V}_k^{\top} = \mathbf{U}_k \tilde{\mathbf{U}}_k \tilde{\Sigma}_k \tilde{\mathbf{V}}_k^{\top} \mathbf{V}_k^{\top}$, 可以给出 $\mathrm{T}_{\mathcal{X}}\mathcal{M}_{\leq \mathbf{r}}$ 的一个紧参数化. 值得注意的是 (4-6)中的矩阵 $\mathcal{C}$ 和 $\mathbf{U}_{k,1}$ 相比于 (4-5)参数量更少.

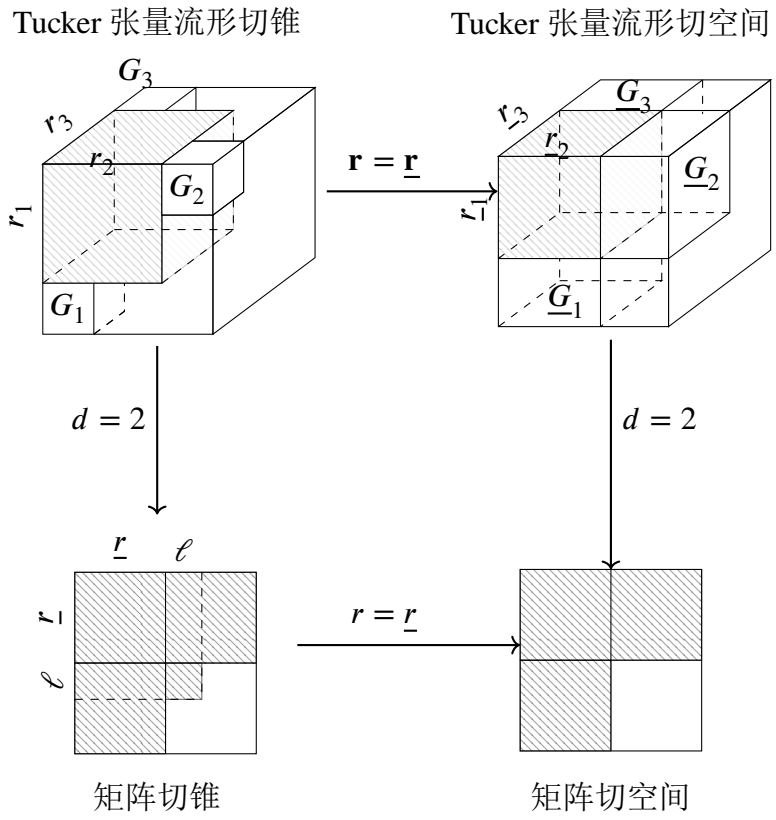


图 4-4 矩阵张量流形代数簇切锥表达的联系.

Figure 4-4 Connections of tangent cones of matrix and tensor varieties.

推论 4.3. 给定一个具有 *Tucker* 分解的张量 $\mathcal{X} = \mathcal{G} \times_1 \mathbf{U}_1 \cdots \times_d \mathbf{U}_d \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ 以及 $\operatorname{rank}_{\mathrm{tc}}(\mathcal{X}) = \underline{\mathbf{r}} \leq \mathbf{r}$, 在 $\mathcal{X}$ 处切锥 $\mathcal{M}_{\leq \mathbf{r}}$ 中的任意元素 $\mathcal{V}$ 可以被表示为

$$\mathcal{V} = \mathcal{C} \times_{k=1}^{d} \begin{bmatrix} \mathbf{U}_k & \mathbf{U}_{k,1} \end{bmatrix} + \sum_{k=1}^{d} \mathcal{G} \times_k (\mathbf{U}_{k,2} \mathbf{R}_{k,2}) \times_{j \neq k} \mathbf{U}_j, \tag{4-6}$$

这里 $\mathcal{C} \in \mathbb{R}^{(\underline{r}_1 + \ell_1) \times \cdots \times (\underline{r}_d + \ell_d)}$, $\mathbf{R}_{k,2} \in \mathbb{R}^{(n_k - \underline{r}_k - \ell_k) \times \underline{r}_k}$, $\mathbf{U}_{k,1} \in \operatorname{St}(\ell_k, n_k)$ 以及 $\mathbf{U}_{k,2} \in \operatorname{St}(n_k - \underline{r}_k - \ell_k, n_k)$ 为对 $k \in [d]$ 满足 $[\mathbf{U}_k \ \mathbf{U}_{k,1} \ \mathbf{U}_{k,2}] \in \mathcal{O}(n_k)$ 的自由变量, $\ell = (\ell_1, \ell_2, \ldots, \ell_d)$ 满足

$$\ell_k = \operatorname{rank}(\mathrm{P}^{\perp}_{\mathbf{U}_k} \mathbf{V}_{(k)} \mathrm{P}^{\perp}_{\mathbf{V}_k \tilde{\mathbf{V}}_k}) = \operatorname{rank}([0 \ \mathbf{U}_{k,1}] \mathbf{C}_{(k)} ([\mathbf{U}_j \ \mathbf{U}_{j,1}]^{\otimes j \neq k})^{\top} \mathrm{P}^{\perp}_{\mathbf{V}_k \tilde{\mathbf{V}}_k})$$

这里 $\mathbf{V}_k = (\mathbf{U}_j)^{\otimes j \neq k}$ 以及 $\tilde{\mathbf{V}}_k \in \operatorname{St}(\underline{r}_k, \underline{r}_{-k})$ 是矩阵 $\mathbf{G}_{(k)}$ 的右奇异向量. 更进一步, 式 (4-6)中的切锥表达在正交群作用在矩阵 $\mathbf{U}_{k,1}$ 和 $\mathbf{U}_{k,2}$ 的意义下是唯一的, 这里 $k \in [d]$.

证明. 我们可以通过将图 4-1中的参数化替换为 (4-3)得到式 (4-6), 这里 $\ell_k = \operatorname{rank}(\mathrm{P}^{\perp}_{\mathbf{U}_k} \mathbf{V}_{(k)} \mathrm{P}^{\perp}_{\mathbf{V}_k \tilde{\mathbf{V}}_k})$ 来自于定理 4.2中的证明.

更进一步, 我们的目标是证明 $\operatorname{span}(\mathbf{U}_{k,1}) = \operatorname{span}(\mathrm{P}^{\perp}_{\mathbf{U}_k} \mathbf{V}_{(k)} \mathrm{P}^{\perp}_{\mathbf{V}_k \tilde{\mathbf{V}}_k})$. 为此, 我们有

$$\begin{aligned}
\mathrm{P}^{\perp}_{\mathbf{U}_k} \mathbf{V}_{(k)} \mathrm{P}^{\perp}_{\mathbf{V}_k \tilde{\mathbf{V}}_k} &= \mathrm{P}^{\perp}_{\mathbf{U}_k} \Big([\mathbf{U}_k \ \mathbf{U}_{k,1}] \mathbf{C}_{(k)} ([\mathbf{U}_j \ \mathbf{U}_{j,1}]^{\otimes j \neq k})^{\top} + \mathbf{U}_{k,2} \mathbf{R}_{k,2} \mathbf{G}_{(k)} \mathbf{V}_k^{\top} \Big) \mathrm{P}^{\perp}_{\mathbf{V}_k \tilde{\mathbf{V}}_k} \\
&= [0 \ \mathbf{U}_{k,1}] \mathbf{C}_{(k)} ([\mathbf{U}_j \ \mathbf{U}_{j,1}]^{\otimes j \neq k})^{\top} \mathrm{P}^{\perp}_{\mathbf{V}_k \tilde{\mathbf{V}}_k} + \mathbf{U}_{k,2} \mathbf{R}_{k,2} \tilde{\mathbf{U}}_k \tilde{\Sigma}_k \tilde{\mathbf{V}}_k^{\top} \mathbf{V}_k^{\top} \mathrm{P}^{\perp}_{\mathbf{V}_k \tilde{\mathbf{V}}_k} \\
&= \mathbf{U}_{k,1} \underline{\mathbf{C}}_{(k)} ([\mathbf{U}_j \ \mathbf{U}_{j,1}]^{\otimes j \neq k})^{\top} \mathrm{P}^{\perp}_{\mathbf{V}_k \tilde{\mathbf{V}}_k},
\end{aligned}$$

这里 $\mathbf{G}_{(k)} = \tilde{\mathbf{U}}_k \tilde{\Sigma}_k \tilde{\mathbf{V}}_k^{\top}$ 是展平矩阵 $\mathbf{G}_{(k)}$ 的 SVD 分解, 以及 $\underline{\mathbf{C}}_{(k)} \in \mathbb{R}^{\ell_k \times (\prod_{j \neq k} (\underline{r}_j + \ell_j))}$ 由矩阵 $\mathbf{C}_{(k)}$ 的最后 ℓ_k 行组成. 利用 $\mathbf{U}_{k,1} \in \operatorname{St}(\ell_k, n_k)$ 和 $\ell_k = \operatorname{rank}(\mathrm{P}^{\perp}_{\mathbf{U}_k} \mathbf{V}_{(k)} \mathrm{P}^{\perp}_{\mathbf{V}_k \tilde{\mathbf{V}}_k})$, 我们可以得到 $\operatorname{span}(\mathbf{U}_{k,1}) = \operatorname{span}(\mathrm{P}^{\perp}_{\mathbf{U}_k} \mathbf{V}_{(k)} \mathrm{P}^{\perp}_{\mathbf{V}_k \tilde{\mathbf{V}}_k})$.

于是, 式 (4-6)中的切锥表达在正交群作用在矩阵 $\mathbf{U}_{k,1}$ 和 $\mathbf{U}_{k,2}$ 的意义下是唯一的. □

这里值得注意的是, 由于 $r_{-k} = r$ 对于 $d = 2$ 以及 $k = 1, 2$ 成立, 我们可以得到 $\tilde{\mathbf{V}}_k \in \mathcal{O}(r)$ 因此 $\mathrm{P}^{\perp}_{\mathbf{V}_k \tilde{\mathbf{V}}_k} = \mathrm{P}^{\perp}_{\mathbf{V}_k}$. 因此, 张量的紧参数化 (4-6) 与矩阵情形 (4-3) 完全相符. 虽然由于我们没有对 $\mathcal{C}$ 提出秩约束条件, 导致式 (4-5)不是一个紧凑的表达, 但是式 (4-5)仍然可以用于计算切锥投影, 详见接下来的章节.

我们从定理 4.2的证明中发现, 引理 4.1 中关于切锥的包含关系事实上是等于关系, 这一点我们在下面的推论中展示. 具体而言, 任意满足 $\mathbf{V}_{(k)} \in \mathrm{T}_{\mathbf{X}_{(k)}} \mathbb{R}^{n_k \times n_{-k}}_{\leq r_k}$ 的张量 $\mathcal{V} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$ 都可以表达为式 (4-5)的形式, 也就是说 $\mathcal{V} \in \mathrm{T}_{\mathcal{X}} \mathcal{M}_{\leq \mathbf{r}}$.

推论 4.4. 张量代数簇 $\mathcal{M}_{\leq \mathbf{r}}$ 的切锥, 等于沿着不同模态张量化的矩阵代数簇的切锥交集, 也就是说,

$$\mathrm{T}_{\mathcal{X}} \mathcal{M}_{\leq \mathbf{r}} = \bigcap_{k=1}^{d} \operatorname{ten}_{(k)} \left(\mathrm{T}_{\mathbf{X}_{(k)}} \mathbb{R}^{n_k \times n_{-k}}_{\leq r_k} \right).$$

通常情况下, 定义 1.11所给出的对问题 (4-1)的一阶稳定性条件是很难在计算上验证的 (详见 4.2.4节). 下面的命题给出了如何在每个模态都不满秩的点 $\mathcal{X}^* \in \mathcal{M}_{<\mathbf{r}} := \{\mathcal{X} \in \mathbb{R}^{n_1\times n_2\times\cdots\times n_d} : \mathrm{rank}_{\mathrm{tc}}(\mathcal{X}) < \mathbf{r}\}$ 上验证稳定性条件的等价刻画.

命题 4.5. 设 $\mathcal{X}^*$ 是问题 (4-1) 的一个一阶稳定点 $\underline{\mathbf{r}}^* := \mathrm{rank}_{\mathrm{tc}}(\mathcal{X}^*) < \mathbf{r}$. 那么, 如下的等式成立

$$\nabla f(\mathcal{X}^*) = 0.$$

证明. 由于任意的张量 $\mathcal{A} \in \mathbb{R}^{n_1\times n_2\times\cdots\times n_d}$ 可以通过秩 1 张量 $\mathbf{e}_{i_1}\circ\mathbf{e}_{i_2}\circ\cdots\circ\mathbf{e}_{i_d}$ 和 $(i_1, i_2, \ldots, i_d) \in [n_1]\times[n_2]\times\cdots\times[n_d]$ 被表达为

$$\mathcal{A} = \sum_{i_1=1}^{n_1}\cdots\sum_{i_d=1}^{n_d} y_{i_1,\ldots,i_d}\mathbf{e}_{i_1}\circ\cdots\circ\mathbf{e}_{i_d},$$

我们只需要证明 $\langle \mathbf{e}_{i_1}\circ\cdots\circ\mathbf{e}_{i_d}, \nabla f(\mathcal{X}^*)\rangle = 0$.

我们根据 $\underline{\mathbf{r}}^* < \mathbf{r}$ 可以得到 $\mathcal{X}^* + t\mathbf{e}_{i_1}\circ\cdots\circ\mathbf{e}_{i_d} \in \mathcal{M}_{\leq\mathbf{r}}$ 对所有的 t, $i_k \in [n_k]$ 和 $k \in [d]$ 成立. 因此, 我们可以得到 $\pm\mathbf{e}_{i_1}\circ\cdots\circ\mathbf{e}_{i_d} \in \mathrm{T}_{\mathcal{X}^*}\mathcal{M}_{\leq\mathbf{r}}$. 根据定义 1.11 以及 $\mathcal{X}^*$ 是一阶稳定点, 我们有 $\langle \mathbf{e}_{i_1}\circ\cdots\circ\mathbf{e}_{i_d}, \nabla f(\mathcal{X}^*)\rangle = 0$ 以及 $\nabla f(\mathcal{X}^*) = 0$. □

与矩阵情形中等式关系 $\mathbb{R}^{m\times n}_{\leq r} = \mathbb{R}^{m\times n}_{<r} \cup \mathbb{R}^{m\times n}_{r}$ 不同的是, Tucker 张量代数簇 $\mathcal{M}_{\leq\mathbf{r}}$ 不仅仅包含了集合 $\mathcal{M}_{<\mathbf{r}}$ 和 $\mathcal{M}_{\mathbf{r}}$ 中的点, 也包含了部分模态满秩部分模态亏秩的点. 命题 4.5 中的稳定性条件是限制在集合 $\mathcal{M}_{<\mathbf{r}}$ 上的.

定理 4.2给出的切锥显式表达式让我们可以得到一些新的结果. 我们回顾式 (4-5) 中的参数化:

$$\mathcal{V} = \mathcal{V}_0 + \sum_{k=1}^{d}\mathcal{V}_k := \mathcal{C}\times_{k=1}^{d}\begin{bmatrix}\mathbf{U}_k & \mathbf{U}_{k,1}\end{bmatrix} + \sum_{k=1}^{d}\mathcal{G}\times_k(\mathbf{U}_{k,2}\mathbf{R}_{k,2})\times_{j\neq k}\mathbf{U}_j. \tag{4-7}$$

注意到对于 $i \neq j$ 来说有 $\langle\mathcal{V}_i, \mathcal{V}_j\rangle = 0$, 因此 $\|\mathcal{V}\|_{\mathrm{F}}^2 = \sum_{k=0}^{d}\|\mathcal{V}_k\|_{\mathrm{F}}^2$. 令人意外的是, 沿着 $\mathcal{V}$ 中的两类方向做线搜索并不会离开 Tucker 张量代数簇: 1) 从式 (1-5) 和 Tucker 分解 $\mathcal{X} = \mathcal{G}\times_{k=1}^{d}\mathbf{U}_k$ 得知

$$\begin{aligned}\mathcal{X} + \mathcal{V}_0 &= \mathcal{G}\times_{k=1}^{d}\mathbf{U}_k + \mathcal{C}\times_{k=1}^{d}\begin{bmatrix}\mathbf{U}_k & \mathbf{U}_{k,1}\end{bmatrix}\\ &\in \bigotimes_{k=1}^{d}\mathrm{span}([\mathbf{U}_k\ \mathbf{U}_{k,1}]) \subseteq \mathcal{M}_{\leq\mathbf{r}};\end{aligned} \tag{4-8}$$

2) 对于所有的 $k \in [d]$, 我们通过 $\mathrm{rank}(\mathbf{U}_k + \mathbf{U}_{k,2}\mathbf{R}_{k,2}) \leq \underline{r}_k$ 发现

$$\begin{aligned}\mathcal{X} + \mathcal{V}_k &= \mathcal{G}\times_{i=1}^{d}\mathbf{U}_i + \mathcal{G}\times_k(\mathbf{U}_{k,2}\mathbf{R}_{k,2})\times_{j\neq k}\mathbf{U}_j\\ &= \mathcal{G}\times_k(\mathbf{U}_k + \mathbf{U}_{k,2}\mathbf{R}_{k,2})\times_{j\neq k}\mathbf{U}_j\\ &\in \mathcal{M}_{\leq\mathbf{r}}.\end{aligned} \tag{4-9}$$

值得注意的是式 (4-8) 和 (4-9) 提供了 $(d+1)$ 个在切锥 $\mathrm{T}_{\mathcal{X}}\mathcal{M}_{\leq \mathbf{r}}$ 中的免收缩映射的搜索方向 $\mathcal{V}_0, \mathcal{V}_1, \ldots, \mathcal{V}_d$, 这些方向可以用于在代数簇 $\mathcal{M}_{\leq \mathbf{r}}$ 上设计免收缩映射的线搜索方法 (即无需往代数簇 $\mathcal{M}_{\leq \mathbf{r}}$ 投影的线搜索方法); 详见4.4节. 此外, 下面的推论给出了收缩映射 $\mathrm{R}_{\mathcal{X}}^{\mathrm{HO}} : \mathcal{V} \mapsto \mathrm{P}_{\leq \mathbf{r}}^{\mathrm{HO}}(\mathcal{X}+\mathcal{V})$ 满足的界, 这个结果对证明 $\mathcal{M}_{\leq \mathbf{r}}$ 上线搜索方法的收敛性有至关重要的作用.

推论 4.6. *高阶奇异值分解 (HOSVD) 收缩映射满足对任意的 $\mathcal{X} \in \mathcal{M}_{\leq \mathbf{r}}$ 和 $\mathcal{V} \in \mathrm{T}_{\mathcal{X}}\mathcal{M}_{\leq \mathbf{r}}$ 有*

$$\|\mathcal{X} - \mathrm{P}_{\leq \mathbf{r}}^{\mathrm{HO}}(\mathcal{X}+\mathcal{V})\|_{\mathrm{F}} \leq (1 + \frac{d}{\sqrt{d+1}})\|\mathcal{V}\|_{\mathrm{F}}.$$

证明. 根据式 (1-11) 以及式(4-8)–(4-9)的推论 $\mathcal{X} + \mathcal{V}_k \in \mathcal{M}_{\leq \mathbf{r}}$, 我们可以得到对于所有的 $k = 0, 1, \ldots, d$ 有

$$\|\mathcal{X}+\mathcal{V} - \mathrm{P}_{\leq \mathbf{r}}(\mathcal{X}+\mathcal{V})\|_{\mathrm{F}}^2 = \min_{\mathcal{Y} \in \mathcal{M}_{\leq \mathbf{r}}} \|\mathcal{X}+\mathcal{V}-\mathcal{Y}\|_{\mathrm{F}}^2 \leq \|\mathcal{X}+\mathcal{V}-(\mathcal{X}+\mathcal{V}_k)\|_{\mathrm{F}}^2,$$

于是有

$$(d+1)\|\mathcal{X}+\mathcal{V} - \mathrm{P}_{\leq \mathbf{r}}(\mathcal{X}+\mathcal{V})\|_{\mathrm{F}}^2 \leq \sum_{k=0}^{d} \|\mathcal{X}+\mathcal{V}-(\mathcal{X}+\mathcal{V}_k)\|_{\mathrm{F}}^2.$$

由拟最优性式 (1-13) 以及 $\|\mathcal{V}\|_{\mathrm{F}}^2 = \sum_{k=0}^{d} \|\mathcal{V}_k\|_{\mathrm{F}}^2$ 得知

$$\begin{aligned}
\left\|\mathcal{X}+\mathcal{V} - \mathrm{P}_{\leq \mathbf{r}}^{\mathrm{HO}}(\mathcal{X}+\mathcal{V})\right\|_{\mathrm{F}}^2 &\leq d\,\|\mathcal{X}+\mathcal{V} - \mathrm{P}_{\leq \mathbf{r}}(\mathcal{X}+\mathcal{V})\|_{\mathrm{F}}^2 \\
&\leq \frac{d}{d+1}\sum_{k=0}^{d} \|\mathcal{X}+\mathcal{V}-(\mathcal{X}+\mathcal{V}_k)\|_{\mathrm{F}}^2 \\
&= \frac{d}{d+1}\sum_{k=0}^{d}\sum_{j=0, j\neq k}^{d} \|\mathcal{V}_j\|_{\mathrm{F}}^2 \\
&= \frac{d^2}{d+1}\|\mathcal{V}\|_{\mathrm{F}}^2
\end{aligned}$$

成立. 因此有如下的不等式成立

$$\|\mathcal{X} - \mathrm{P}_{\leq \mathbf{r}}^{\mathrm{HO}}(\mathcal{X}+\mathcal{V})\|_{\mathrm{F}} \leq \|\mathcal{X}+\mathcal{V} - \mathrm{P}_{\leq \mathbf{r}}^{\mathrm{HO}}(\mathcal{X}+\mathcal{V})\|_{\mathrm{F}} + \|\mathcal{V}\|_{\mathrm{F}} \leq (1 + \frac{d}{\sqrt{d+1}})\|\mathcal{V}\|_{\mathrm{F}}.$$

□

4.2.4　往切锥上的度量投影

为了在 $\mathcal{M}_{\leq \mathbf{r}}$ 上构造投影梯度类方法, 我们考虑将张量 $\mathcal{A} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$ 到 $\mathcal{X} = \mathcal{G} \times_{k=1}^{d} \mathbf{U}_k$ 处切锥 $\mathrm{T}_{\mathcal{X}}\mathcal{M}_{\leq \mathbf{r}}$ 的度量投影, 这个投影由如下的优化问题所决定:

$$\mathrm{P}_{\mathrm{T}_{\mathcal{X}}\mathcal{M}_{\leq \mathbf{r}}}\mathcal{A} := \operatorname*{arg\,min}_{\mathcal{V} \in \mathrm{T}_{\mathcal{X}}\mathcal{M}_{\leq \mathbf{r}}} \|\mathcal{A} - \mathcal{V}\|_{\mathrm{F}}^2. \tag{4-10}$$

由于切锥 $\mathrm{T}_{\mathcal{X}}\mathcal{M}_{\leq \mathbf{r}}$ 是一个闭集, 因此度量投影是存在的. 值得注意的是, 度量投影并不是唯一的, 但是在实际设计 $\mathcal{M}_{\leq \mathbf{r}}$ 上投影梯度类算法中, 我们总能选择其中一个投影保证算法能运行. 为了研究度量投影, 我们将切锥表达式 (4-5) 和式 (4-7) 代入度量投影的表达式中 (4-10), 可以得到

$$\begin{aligned}\|\mathcal{A}-\mathcal{V}\|_{\mathrm{F}}^2 &= \|\mathcal{A}-\sum_{k=0}^{d}\mathcal{V}_k\|_{\mathrm{F}}^2 = \|\mathcal{A}\|_{\mathrm{F}}^2+\sum_{k=0}^{d}\|\mathcal{V}_k\|_{\mathrm{F}}^2-2\sum_{k=0}^{d}\langle\mathcal{A},\mathcal{V}_k\rangle\\ &= \|\mathcal{A}\|_{\mathrm{F}}^2+\|\mathcal{C}\|_{\mathrm{F}}^2-2\left\langle\mathcal{A}\times_{k=1}^{d}\begin{bmatrix}\mathbf{U}_k & \mathbf{U}_{k,1}\end{bmatrix}^{\top},\mathcal{C}\right\rangle\\ &\quad+\sum_{k=1}^{d}\left(\|\mathcal{G}\times_k\mathbf{R}_{k,2}\|_{\mathrm{F}}^2-2\left\langle\mathcal{A}\times_k\mathbf{U}_{k,2}^{\top}\times_{j\neq k}\mathbf{U}_j^{\top},\mathcal{G}\times_k\mathbf{R}_{k,2}\right\rangle\right).\end{aligned}\tag{4-11}$$

为了求解问题 (4-10), 我们希望解出式 (4-11)中未知变量 $\mathcal{C}$, $\mathbf{U}_{k,1}$, $\mathbf{U}_{k,2}$ 和 $\mathbf{R}_{k,2}$, 这里 $k\in[d]$. 简而言之, 我们通过如下两个步骤计算度量投影 $\mathrm{P}_{\mathrm{T}_{\mathcal{X}}\mathcal{M}_{\leq\mathbf{r}}}\mathcal{A}$: 1) 固定 $\mathbf{U}_{k,1}$ 和 $\mathbf{U}_{k,2}$, 我们通过最小二乘问题得到 $\mathcal{C}$ 和 $\mathbf{R}_{k,2}$ 的显式表达; 2) 我们将得到的解代入式 (4-11) 并通过代入后的优化问题得到 $\mathbf{U}_{k,1}$ 满足的优化问题, 此时 $\mathbf{U}_{k,2}$ 可以天然地通过 $[\mathbf{U}_k\ \mathbf{U}_{k,1}\ \mathbf{U}_{k,2}]\in\mathcal{O}(n_k)$ 得到.

步 1: 固定 $\mathbf{U}_{k,1}$ 和 $\mathbf{U}_{k,2}$ 得到 $\mathcal{C}$ 和 $\mathbf{R}_{k,2}$ 我们通过固定 $\mathbf{U}_{k,1}$ 和 $\mathbf{U}_{k,2}$, 发现问题 (4-11) 中的目标函数是一个关于 $\mathcal{C}$ 和 $\mathbf{R}_{k,2}$ 的可分二次函数. 因此, 度量投影 (4-10) 中的变量 $\mathcal{C}$ 和 $\mathbf{R}_{k,2}$ 可以被唯一的表达为

$$\begin{aligned}\mathcal{C} &= \mathcal{A}\times_{k=1}^{d}\begin{bmatrix}\mathbf{U}_k & \mathbf{U}_{k,1}\end{bmatrix}^{\top},\\ \mathbf{R}_{k,2} &= \left(\mathcal{A}\times_k\mathbf{U}_{k,2}^{\top}\times_{j\neq k}\mathbf{U}_j^{\top}\right)_{(k)}\mathbf{G}_{(k)}^{\dagger}.\end{aligned}\tag{4-12}$$

步 2: 求解 $\mathbf{U}_{k,1}$ 我们将式 (4-12) 代入式 (4-11), 并得到

$$\begin{aligned}\|\mathcal{A}-\mathcal{V}\|_{\mathrm{F}}^2 &= \|\mathcal{A}\|_{\mathrm{F}}^2-\|\mathcal{C}\|_{\mathrm{F}}^2-\sum_{k=1}^{d}\|\mathcal{G}\times_k\mathbf{R}_{k,2}\|_{\mathrm{F}}^2\\ &= \|\mathcal{A}\|_{\mathrm{F}}^2-\|\mathcal{C}\|_{\mathrm{F}}^2-\sum_{k=1}^{d}\left\langle\mathbf{A}_{\neq k}\,\mathrm{P}_{\mathbf{G}_{(k)}^{\top}},(\mathbf{I}_{n_k}-\mathbf{U}_k\mathbf{U}_k^{\top}-\mathbf{U}_{k,1}\mathbf{U}_{k,1}^{\top})\mathbf{A}_{\neq k}\,\mathrm{P}_{\mathbf{G}_{(k)}^{\top}}\right\rangle\\ &= \|\mathcal{A}\|_{\mathrm{F}}^2-\left\|\mathcal{A}\times_{k=1}^{d}\begin{bmatrix}\mathbf{U}_k & \mathbf{U}_{k,1}\end{bmatrix}^{\top}\right\|_{\mathrm{F}}^2\\ &\quad-\sum_{k=1}^{d}\left\|\mathbf{A}_{\neq k}\,\mathrm{P}_{\mathbf{G}_{(k)}^{\top}}\right\|_{\mathrm{F}}^2+\sum_{k=1}^{d}\left\|\mathbf{U}_k^{\top}\mathbf{A}_{\neq k}\,\mathrm{P}_{\mathbf{G}_{(k)}^{\top}}\right\|_{\mathrm{F}}^2+\sum_{k=1}^{d}\left\|\mathbf{U}_{k,1}^{\top}\mathbf{A}_{\neq k}\,\mathrm{P}_{\mathbf{G}_{(k)}^{\top}}\right\|_{\mathrm{F}}^2,\end{aligned}$$

这里 $\mathbf{A}_{\neq k}:=(\mathcal{A}\times_{j\neq k}\mathbf{U}_j^{\top})_{(k)}\in\mathbb{R}^{n_k\times r_{-k}}$ 以及 $\mathrm{P}_{\mathbf{G}_{(k)}^{\top}}:=\mathbf{G}_{(k)}^{\dagger}\mathbf{G}_{(k)}$. 这里 $\mathbf{U}_{k,2}$ 被 $\mathbf{U}_k$ 和 $\mathbf{U}_{k,1}$ 消去了. 有一种替代的方案是利用 $\mathbf{U}_k$ 和 $\mathbf{U}_{k,2}$ 将变量 $\mathbf{U}_{k,1}$ 消去. 然而, 当

$r_k \ll n_k$ 时, 变量 $\mathbf{U}_{k,1} \in \mathbb{R}^{n_k \times (r_k - \underline{r}_k)}$ 的参数量少于变量 $\mathbf{U}_{k,2} \in \mathbb{R}^{n_k \times (n_k - r_k)}$. 因此, 从计算上来说将问题 (4-10) 变成

$$\begin{aligned} \min_{\mathbf{U}_{1,1},\mathbf{U}_{2,1},\ldots,\mathbf{U}_{d,1}} \quad & -\left\| \mathcal{A} \times_{k=1}^{d} \begin{bmatrix} \mathbf{U}_k & \mathbf{U}_{k,1} \end{bmatrix}^{\top} \right\|_{\mathrm{F}}^2 + \sum_{k=1}^{d} \left\| \mathbf{U}_{k,1}^{\top} \mathbf{A}_{\neq k} \mathrm{P}_{\mathbf{G}_{(k)}^{\top}} \right\|_{\mathrm{F}}^2 \\ \text{s.t.} \quad & \begin{bmatrix} \mathbf{U}_k & \mathbf{U}_{k,1} \end{bmatrix}^{\top} \begin{bmatrix} \mathbf{U}_k & \mathbf{U}_{k,1} \end{bmatrix} = \mathbf{I}_{r_k} \text{ 对所有的 } k \in [d] \end{aligned} \tag{4-13}$$

是更节省空间的. 由于问题 (4-13)的可行集是紧集, 问题 (4-13)的全局最优解是存在的.

设 $(\mathbf{U}_{1,1}^*, \mathbf{U}_{2,1}^*, \ldots, \mathbf{U}_{d,1}^*)$ 是问题 (4-13)的一个全局最优解. 我们通过式 (4-5) 和式 (4-12), 可以得到往切锥 $\mathrm{T}_{\mathcal{X}}\mathcal{M}_{\leq \mathbf{r}}$ 的度量投影的表达式

$$\mathrm{P}_{\mathrm{T}_{\mathcal{X}}\mathcal{M}_{\leq \mathbf{r}}} \mathcal{A} = \mathcal{A} \times_{k=1}^{d} \mathrm{P}_{\mathbf{S}_k^*} + \sum_{k=1}^{d} \mathcal{G} \times_k \left(\mathrm{P}_{\mathbf{S}_k^*}^{\perp} \left(\mathcal{A} \times_{j \neq k} \mathbf{U}_j^{\top} \right)_{(k)} \mathbf{G}_{(k)}^{\dagger} \right) \times_{j \neq k} \mathbf{U}_j, \tag{4-14}$$

这里 $\mathbf{S}_k^* := [\mathbf{U}_k \ \mathbf{U}_{k,1}^*]$. 我们发现投影 (4-14) 取决于投影算子 $\mathrm{P}_{\mathbf{S}_k^*}$ 而非特定的列正交矩阵 $\mathbf{U}_{k,1}^*$. 此外, 当 $\mathbf{r} = \underline{\mathbf{r}}$ 时, 式 (4-14) 可以自然地退化为式 (1-14), 即往切空间投影的显式表达式.

与矩阵代数簇的联系　值得注意的是, 式 (4-14)中的投影与已有的矩阵代数簇上的投影 (即式 (1-9)) 的结果是相合的. 具体而言, 给定一个矩阵 $\mathbf{X} \in \mathbb{R}_{\underline{r}}^{m \times n}$ 和它的 SVD $\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^{\top}$, 此时式 (4-13)会退化为

$$\max_{\mathbf{U}_{1,1},\mathbf{V}_{2,1}} \|\mathbf{U}_{1,1}^{\top} \mathbf{A} \mathbf{V}_{2,1}\|_{\mathrm{F}}^2, \quad \text{s.t.} \begin{bmatrix} \mathbf{U} & \mathbf{U}_{1,1} \end{bmatrix}^{\top} \begin{bmatrix} \mathbf{U} & \mathbf{U}_{1,1} \end{bmatrix} = \begin{bmatrix} \mathbf{V} & \mathbf{V}_{2,1} \end{bmatrix}^{\top} \begin{bmatrix} \mathbf{V} & \mathbf{V}_{2,1} \end{bmatrix} = \mathbf{I}_r. \tag{4-15}$$

事实上, 问题 (4-15) 有一个显式解 $(\mathbf{U}^*, \mathbf{V}^*)$, 它是矩阵 $\mathrm{P}_{\mathbf{U}}^{\perp} \mathbf{A} \mathrm{P}_{\mathbf{V}}^{\perp}$ 的前 $(r - \underline{r})$ 个奇异向量. 这是因为, 由于 $\mathbf{U}^{\top}\mathbf{U}^* = 0$ 和 $\mathbf{V}^{\top}\mathbf{V}^* = 0$, 导致 $(\mathbf{U}^*, \mathbf{V}^*)$ 是问题 (4-15)的一个可行解. 更进一步, 对于所有的可行解 $(\mathbf{U}_{1,1}, \mathbf{V}_{2,1})$, 我们根据 $\mathrm{P}_{\mathbf{U}}^{\perp}\mathbf{U}_{1,1} = \mathbf{U}_{1,1}$, $\mathrm{P}_{\mathbf{V}}^{\perp}\mathbf{V}_{2,1} = \mathbf{V}_{2,1}$, 和 Eckart–Young 定理得到

$$\|\mathbf{U}_{1,1}^{\top} \mathbf{A} \mathbf{V}_{2,1}\|_{\mathrm{F}}^2 = \|\mathbf{U}_{1,1}^{\top} \mathrm{P}_{\mathbf{U}}^{\perp} \mathbf{A} \mathrm{P}_{\mathbf{V}}^{\perp} \mathbf{V}_{2,1}\|_{\mathrm{F}}^2 \leq \|(\mathbf{U}^*)^{\top} \mathrm{P}_{\mathbf{U}}^{\perp} \mathbf{A} \mathrm{P}_{\mathbf{V}}^{\perp} \mathbf{V}^*\|_{\mathrm{F}}^2 = \|(\mathbf{U}^*)^{\top} \mathbf{A} \mathbf{V}^*\|_{\mathrm{F}}^2.$$

因此, $(\mathbf{U}^*, \mathbf{V}^*)$ 确实是问题 (4-15)的一个全局最优解.

我们通过式 (4-14) 得到

$$\begin{aligned} \mathrm{P}_{\mathrm{T}_{\mathbf{X}}\mathbb{R}_{\leq r}^{m \times n}} \mathbf{A} &= \mathrm{P}_{[\mathbf{U}\ \mathbf{U}^*]} \mathbf{A} \mathrm{P}_{[\mathbf{V}\ \mathbf{V}^*]} + \mathrm{P}_{[\mathbf{U}\ \mathbf{U}^*]}^{\perp} \mathbf{A} \mathrm{P}_{\mathbf{V}} + \mathrm{P}_{\mathbf{U}} \mathbf{A} \mathrm{P}_{[\mathbf{V}\ \mathbf{V}^*]}^{\perp} \\ &= \mathrm{P}_{\mathrm{T}_{\mathbf{X}}\mathbb{R}_{\underline{r}}^{m \times n}} \mathbf{A} + \mathrm{P}_{\leq (r - \underline{r})} \left(\mathrm{P}_{\mathbf{U}}^{\perp} \mathbf{A} \mathrm{P}_{\mathbf{V}}^{\perp} \right), \end{aligned}$$

这个结果与式 (1-9)中往矩阵切锥上投影的结果相容.

相比于矩阵的情形, 在 $d \geq 3$ 时求解问题 (4-13)的全局极小值点是一个计算上不可行的方案. 因此, 一个更加可行的办法是构造近似投影.

4.2.5 一个近似投影

式 (4-10) 中的投影通常是没有显式解的, 因此我们需要设计近似投影. 给定 $\mathcal{X} = \mathcal{G} \times_{k=1}^{d} \mathbf{U}_k$ 以及对任意 $k \in [d]$ 满足 $\mathrm{P}_{\mathbf{U}_k} \tilde{\mathbf{U}}_{k,1} = 0$ 的 $\tilde{\mathbf{U}}_{k,1} \in \mathrm{St}(r_k - \underline{r}_k, n_k)$, 我们通过将式 (4-14) 中的 $\mathbf{U}_{k,1}^*$ 替换为 $\tilde{\mathbf{U}}_{k,1}$ 来构造近似投影.

命题 4.7. *给定对 $k \in [d]$ 满足 $\mathrm{P}_{\mathbf{U}_k} \tilde{\mathbf{U}}_{k,1} = 0$ 的矩阵 $\tilde{\mathbf{U}}_{k,1} \in \mathrm{St}(r_k - \underline{r}_k, n_k)$, 如下定义的近似投影*

$$\tilde{\mathrm{P}}_{\mathrm{T}_{\mathcal{X}}\mathcal{M}_{\leq \mathbf{r}}} \mathcal{A} := \mathcal{A} \times_{k=1}^{d} \mathrm{P}_{\tilde{\mathbf{S}}_k} + \sum_{k=1}^{d} \mathcal{G} \times_k \left(\mathrm{P}_{\tilde{\mathbf{S}}_k}^{\perp} \left(\mathcal{A} \times_{j \neq k} \mathbf{U}_j^{\top} \right)_{(k)} \mathbf{G}_{(k)}^{\dagger} \right) \times_{j \neq k} \mathbf{U}_j, \quad (4\text{-}16)$$

满足 $\tilde{\mathrm{P}}_{\mathrm{T}_{\mathcal{X}}\mathcal{M}_{\leq \mathbf{r}}} \mathcal{A} \in \mathrm{T}_{\mathcal{X}}\mathcal{M}_{\leq \mathbf{r}}$ 和 $\langle \mathcal{A}, \tilde{\mathrm{P}}_{\mathrm{T}_{\mathcal{X}}\mathcal{M}_{\leq \mathbf{r}}} \mathcal{A} \rangle = \| \tilde{\mathrm{P}}_{\mathrm{T}_{\mathcal{X}}\mathcal{M}_{\leq \mathbf{r}}} \mathcal{A} \|_{\mathrm{F}}^2$, 这里 $\tilde{\mathbf{S}}_k := [\mathbf{U}_k \ \tilde{\mathbf{U}}_{k,1}]$.

证明. 为简化表达, 我们类似于式 (4-7), 为式 (4-16)引入记号 $\tilde{\mathrm{P}}_{\mathrm{T}_{\mathcal{X}}\mathcal{M}_{\leq \mathbf{r}}} \mathcal{A} = \tilde{\mathcal{V}}_0 + \sum_{k=1}^{d} \tilde{\mathcal{V}}_k$. 我们可以很直接地验证

$$\begin{aligned}
\tilde{\mathcal{V}}_0 &= \left(\mathcal{A} \times_{j=1}^{d} [\mathbf{U}_j \ \tilde{\mathbf{U}}_{j,1}]^{\top} \right) \times_{i=1}^{d} [\mathbf{U}_i \ \tilde{\mathbf{U}}_{i,1}], \\
\tilde{\mathcal{V}}_k &= \mathcal{G} \times_k \left(\tilde{\mathbf{U}}_{k,2} \tilde{\mathbf{U}}_{k,2}^{\top} \left(\mathcal{A} \times_{j \neq k} \mathbf{U}_j^{\top} \right)_{(k)} \mathbf{G}_{(k)}^{\dagger} \right) \times_{j \neq k} \mathbf{U}_j,
\end{aligned}$$

这里 $\tilde{\mathbf{U}}_{k,2} \in \mathrm{St}(n_k - r_k, n_k)$ 满足 $[\mathbf{U}_k \ \tilde{\mathbf{U}}_{k,1} \ \tilde{\mathbf{U}}_{k,2}] \in \mathcal{O}(n_k)$. 因此, $\tilde{\mathrm{P}}_{\mathrm{T}_{\mathcal{X}}\mathcal{M}_{\leq \mathbf{r}}} \mathcal{A}$ 是符合切锥 $\mathrm{T}_{\mathcal{X}}\mathcal{M}_{\leq \mathbf{r}}$ 中一个元素的表达式, 即 (4-5).

注意到对于 $i \neq j$ 有 $\langle \tilde{\mathcal{V}}_i, \tilde{\mathcal{V}}_j \rangle = 0$ 成立, 因此 $\| \tilde{\mathrm{P}}_{\mathrm{T}_{\mathcal{X}}\mathcal{M}_{\leq \mathbf{r}}} \mathcal{A} \|_{\mathrm{F}}^2 = \sum_{k=0}^{d} \| \tilde{\mathcal{V}}_k \|_{\mathrm{F}}^2$. 利用 $\mathrm{P}_{\tilde{\mathbf{S}}_k}^2 = \mathrm{P}_{\tilde{\mathbf{S}}_k} = \mathrm{P}_{\tilde{\mathbf{S}}_k}^{\top}$, $(\mathrm{P}_{\tilde{\mathbf{S}}_k}^{\perp})^2 = \mathrm{P}_{\tilde{\mathbf{S}}_k}^{\perp} = (\mathrm{P}_{\tilde{\mathbf{S}}_k}^{\perp})^{\top}$, $(\mathbf{G}_{(k)}^{\dagger} \mathbf{G}_{(k)})^2 = \mathbf{G}_{(k)}^{\dagger} \mathbf{G}_{(k)} = (\mathbf{G}_{(k)}^{\dagger} \mathbf{G}_{(k)})^{\top}$, 以及 $\mathbf{V}_k^{\top} \mathbf{V}_k = \mathbf{I}_{r_{-k}}$ (这里 $\mathbf{V}_k = (\mathbf{U}_j)^{\otimes j \neq k}$), 我们可以得到

$$\begin{aligned}
\left\langle \mathcal{A}, \tilde{\mathrm{P}}_{\mathrm{T}_{\mathcal{X}}\mathcal{M}_{\leq \mathbf{r}}} \mathcal{A} \right\rangle &= \left\langle \mathcal{A}, \tilde{\mathcal{V}}_0 + \sum_{k=1}^{d} \tilde{\mathcal{V}}_k \right\rangle = \langle \mathcal{A}, \mathcal{A} \times_{k=1}^{d} \mathrm{P}_{\tilde{\mathbf{S}}_k} \rangle + \sum_{k=1}^{d} \langle \mathcal{A}, \tilde{\mathcal{V}}_k \rangle \\
&= \| \tilde{\mathcal{V}}_0 \|_{\mathrm{F}}^2 + \sum_{k=1}^{d} \left\langle \mathbf{A}_{(k)}, \left(\mathrm{P}_{\tilde{\mathbf{S}}_k}^{\perp} \mathbf{A}_{(k)} \mathbf{V}_k \mathbf{G}_{(k)}^{\dagger} \right) \mathbf{G}_{(k)} \mathbf{V}_k^{\top} \right\rangle \\
&= \| \tilde{\mathcal{V}}_0 \|_{\mathrm{F}}^2 + \sum_{k=1}^{d} \left\| \left(\mathrm{P}_{\tilde{\mathbf{S}}_k}^{\perp} \mathbf{A}_{(k)} \mathbf{V}_k \mathbf{G}_{(k)}^{\dagger} \right) \mathbf{G}_{(k)} \mathbf{V}_k^{\top} \right\|_{\mathrm{F}}^2 \\
&= \| \tilde{\mathcal{V}}_0 \|_{\mathrm{F}}^2 + \sum_{k=1}^{d} \left\| \mathcal{G} \times_k \left(\mathrm{P}_{\tilde{\mathbf{S}}_k}^{\perp} \left(\mathcal{A} \times_{j \neq k} \mathbf{U}_j^{\top} \right)_{(k)} \mathbf{G}_{(k)}^{\dagger} \right) \times_{j \neq k} \mathbf{U}_j \right\|_{\mathrm{F}}^2 \\
&= \| \tilde{\mathcal{V}}_0 \|_{\mathrm{F}}^2 + \sum_{k=1}^{d} \| \tilde{\mathcal{V}}_k \|_{\mathrm{F}}^2 \\
&= \left\| \tilde{\mathrm{P}}_{\mathrm{T}_{\mathcal{X}}\mathcal{M}_{\leq \mathbf{r}}} \mathcal{A} \right\|_{\mathrm{F}}^2.
\end{aligned}$$

□

命题 4.7 直接可以导出如下的结论: $\langle \mathcal{A}, \mathrm{P}_{\mathrm{T}_{\mathcal{X}}\mathcal{M}_{\leq \mathbf{r}}}\mathcal{A}\rangle = \|\mathrm{P}_{\mathrm{T}_{\mathcal{X}}\mathcal{M}_{\leq \mathbf{r}}}\mathcal{A}\|_{\mathrm{F}}^2$ 以及 $\|\mathcal{A} - \tilde{\mathrm{P}}_{\mathrm{T}_{\mathcal{X}}\mathcal{M}_{\leq \mathbf{r}}}\mathcal{A}\|_{\mathrm{F}}^2 = \|\mathcal{A}\|_{\mathrm{F}}^2 - \|\tilde{\mathrm{P}}_{\mathrm{T}_{\mathcal{X}}\mathcal{M}_{\leq \mathbf{r}}}\mathcal{A}\|_{\mathrm{F}}^2$. 因此, 我们根据 (4-10) 可以得到

$$\begin{aligned}\|\tilde{\mathrm{P}}_{\mathrm{T}_{\mathcal{X}}\mathcal{M}_{\leq \mathbf{r}}}\mathcal{A}\|_{\mathrm{F}}^2 &= \|\mathcal{A}\|_{\mathrm{F}}^2 - \|\mathcal{A} - \tilde{\mathrm{P}}_{\mathrm{T}_{\mathcal{X}}\mathcal{M}_{\leq \mathbf{r}}}\mathcal{A}\|_{\mathrm{F}}^2 \\ &\leq \|\mathcal{A}\|_{\mathrm{F}}^2 - \|\mathcal{A} - \mathrm{P}_{\mathrm{T}_{\mathcal{X}}\mathcal{M}_{\leq \mathbf{r}}}\mathcal{A}\|_{\mathrm{F}}^2 = \|\mathrm{P}_{\mathrm{T}_{\mathcal{X}}\mathcal{M}_{\leq \mathbf{r}}}\mathcal{A}\|_{\mathrm{F}}^2. \end{aligned} \tag{4-17}$$

我们可以通过算法 11显式地计算近似投影. 特别地, 由于我们不能显式地计算出问题 (4-13)的全局极小值点, 我们随机地选取对 $k \in [d]$ 满足 $\mathbf{U}_k^{\perp}\tilde{\mathbf{U}}_{k,1} = 0$ 的参数 $\tilde{\mathbf{U}}_{k,1} \in \mathrm{St}(r_k - \underline{r}_k, n_k)$. 比如, 我们可以通过对矩阵 $[\mathbf{U}_k\ \mathbf{I}_{n_k}]$ 进行 QR 分解并随机从 Q 因子的最后 $(n_k - \underline{r}_k)$ 行随机选择 $(r_k - \underline{r}_k)$ 个列. 尽管 $\tilde{\mathbf{U}}_{k,1}$ 是随机选择的, 对函数 f 而言 $\tilde{\mathrm{P}}_{\mathrm{T}_{\mathcal{X}}\mathcal{M}_{\leq \mathbf{r}}}(-\nabla f(\mathcal{X})) \in \mathrm{T}_{\mathcal{X}}\mathcal{M}_{\leq \mathbf{r}}$ 仍然是一个下降方向, 这是因为我们在命题 4.7设定 $\mathcal{A} = -\nabla f(\mathcal{X})$ 可以得到

$$\langle -\nabla f(\mathcal{X}), \tilde{\mathrm{P}}_{\mathrm{T}_{\mathcal{X}}\mathcal{M}_{\leq \mathbf{r}}}(-\nabla f(\mathcal{X}))\rangle = \|\tilde{\mathrm{P}}_{\mathrm{T}_{\mathcal{X}}\mathcal{M}_{\leq \mathbf{r}}}(-\nabla f(\mathcal{X}))\|_{\mathrm{F}}^2 \geq 0.$$

算法 11 往 $\mathrm{T}_{\mathcal{X}}\mathcal{M}_{\leq \mathbf{r}}$ 上的近似投影算法

输入: 具有 Tucker 秩 $\mathrm{rank}_{\mathrm{tc}}(\mathcal{X}) = \underline{\mathbf{r}} \leq \mathbf{r}$ 的张量 $\mathcal{X} = \mathcal{G} \times_1 \mathbf{U}_1 \cdots \times_d \mathbf{U}_d$, 以及张量 $\mathcal{A} \in \mathbb{R}^{n_1 \times \cdots \times n_d}$.

1: 随机选择 $\tilde{\mathbf{U}}_{1,1}, \ldots, \tilde{\mathbf{U}}_{d,1}$ 使得 $\mathbf{U}_k^{\perp}\tilde{\mathbf{U}}_{k,1} = 0$.
2: 通过式 (4-16)计算近似投影.

输出: $\tilde{\mathrm{P}}_{\mathrm{T}_{\mathcal{X}}\mathcal{M}_{\leq \mathbf{r}}}\mathcal{A}$.

4.3　梯度相关近似投影方法

在本节中, 我们旨在通过利用定理 4.2所得到的切锥参数化, 来设计线搜索方法. 一个直觉上的想法是设计投影梯度法

$$\mathcal{X}^{(t+1)} = \mathrm{P}_{\leq \mathbf{r}}\left(\mathcal{X}^{(t)} + s^{(t)}\,\mathrm{P}_{\mathrm{T}_{\mathcal{X}^{(t)}}\mathcal{M}_{\leq \mathbf{r}}}(-\nabla f(\mathcal{X}^{(t)}))\right),$$

这是在低秩张量流形 $\mathcal{M}_{\mathbf{r}}$ 上极小化函数 f 的黎曼梯度法的推广; 对于矩阵情形可参考文献 [63, 82]. 然而, 度量投影 $\mathrm{P}_{\mathrm{T}_{\mathcal{X}^{(t)}}\mathcal{M}_{\leq \mathbf{r}}}$ 和 $\mathrm{P}_{\leq \mathbf{r}}$ 都不具有显式表达. 因此, 我们通过利用近似投影 (4-16) 和 HOSVD, 提出梯度相关近似投影方法.

4.3.1　算法框架

算法 12 展示了求解问题 (4-1)的梯度相关近似投影方法 (GRAP).

相比于通过求解问题 (4-14) 得到精确的度量投影, GRAP 方法将式 (4-14) 替换为近似投影 (4-16). 此外, 精确低秩逼近 $\mathrm{P}_{\leq \mathbf{r}}$ 也被 HOSVD $\mathrm{P}_{\leq \mathbf{r}}^{\mathrm{HO}}$ 所替代. 总的来说, GRAP 方法的更新公式为

$$\mathcal{X}^{(t+1)} = \mathrm{P}_{\leq \mathbf{r}}^{\mathrm{HO}}\left(\mathcal{X}^{(t)} + s^{(t)}\,\tilde{\mathrm{P}}_{\mathrm{T}_{\mathcal{X}^{(t)}}\mathcal{M}_{\leq \mathbf{r}}}(-\nabla f(\mathcal{X}^{(t)}))\right).$$

算法 12 梯度相关近似投影方法 (GRAP)

输入: 初始点 $\mathcal{X}^{(0)} \in \mathcal{M}_{\leq \mathbf{r}}$, $\omega = (0, 1]$, 回溯线搜索参数 $\rho, a \in (0, 1)$, $s_{\min} > 0$.

1: **while** 停机准则未被满足 **do**
2: 通过算法 11计算 $g^{(t)} = \tilde{\mathrm{P}}_{\mathrm{T}_{\mathcal{X}^{(t)}}\mathcal{M}_{\leq \mathbf{r}}}(-\nabla f(\mathcal{X}^{(t)}))$, 直到满足角度条件 (4-18).
3: 通过 Armijo 回溯线搜索 (4-19)选择步长 $s^{(t)}$.
4: 更新 $\mathcal{X}^{(t+1)} = \mathrm{P}^{\mathrm{HO}}_{\leq \mathbf{r}}\left(\mathcal{X}^{(t)} + s^{(t)} g^{(t)}\right)$ 以及 $t = t + 1$.
5: **end while**

输出: $\mathcal{X}^{(t)}$.

更进一步, 为保证搜索方向 $g^{(t)} \in \mathrm{T}_{\mathcal{X}^{(t)}}\mathcal{M}_{\leq \mathbf{r}}$ 是足够梯度相关的, 我们提出了如下的角度条件 (angle condition):

$$\langle -\nabla f(\mathcal{X}^{(t)}), g^{(t)} \rangle \geq \omega \, \| \mathrm{P}_{\mathrm{T}_{\mathcal{X}^{(t)}}\mathcal{M}_{\leq \mathbf{r}}}(-\nabla f(\mathcal{X}^{(t)}))\|_{\mathrm{F}} \|g^{(t)}\|_{\mathrm{F}}, \tag{4-18}$$

这里 $\omega \in (0, 1]$.

特别地, 若 $g^{(t)} = \tilde{\mathrm{P}}_{\mathrm{T}_{\mathcal{X}^{(t)}}\mathcal{M}_{\leq \mathbf{r}}}(-\nabla f(\mathcal{X}^{(t)}))$ 不满足角度条件, 我们将生成通过新的 $\tilde{\mathbf{U}}_{k,1}$ 构造不同的近似投影 11, 直到角度条件被满足. 这背后的道理是, 近似投影 (4-16) 仅取决于线性空间 $\mathrm{span}([\mathbf{U}_k \ \tilde{\mathbf{U}}_{k,1}])$ 而非特定的 $\tilde{\mathbf{U}}_{k,1}$, 并且通过遍历所有可能的 $\mathrm{span}(\tilde{\mathbf{U}}_{k,1})$ 我们能找到 $-\nabla f(\mathcal{X}^{(t)})$ 在切锥上的度量投影 (4-14), 而度量投影是满足角度条件的. 在实际计算中, 对于满足 $\mathcal{X}^{(t)} \in \mathcal{M}_{\mathbf{r}}$ 的张量, 由于此时 $g^{(t)}$ 退化为度量投影, 我们不再需要关心角度条件.

对于步长, 我们选择 Armijo 回溯线搜索. 具体而言, 给定初始步长 $s_0^{(t)} > 0$, 通过选择整数 l, 使得对于 $s^{(t)} = \rho^l s_0^{(t)} > s_{\min}$, 有

$$f(\mathcal{X}^{(t)}) - f(\mathrm{P}^{\mathrm{HO}}_{\leq \mathbf{r}}(\mathcal{X}^{(t)} + s^{(t)} g^{(t)})) \geq s^{(t)} a \langle -\nabla f(\mathcal{X}^{(t)}), g^{(t)} \rangle \tag{4-19}$$

成立, 这里 ρ, $a \in (0, 1)$, $s_{\min} > 0$ 为线搜索参数. 若 $\|g^{(t)}\|_{\mathrm{F}} \neq 0$, 则 Armijo 线搜索条件 (4-19)一定能成立. 记 $\mathcal{X}(s) = \mathrm{P}^{\mathrm{HO}}_{\leq \mathbf{r}}(\mathcal{X}^{(t)} + s g^{(t)})$. 我们根据 Taylor 展开以及命题 1.2得知

$$\begin{aligned} f(\mathcal{X}(s)) &= f(\mathcal{X}^{(t)}) + \langle \nabla f(\mathcal{X}^{(t)}), \mathrm{P}^{\mathrm{HO}}_{\leq \mathbf{r}}(\mathcal{X}^{(t)} + s g^{(t)}) - \mathcal{X}^{(t)} \rangle + o(\|\mathcal{X}(s) - \mathcal{X}^{(t)}\|_{\mathrm{F}}) \\ &= f(\mathcal{X}^{(t)}) + \langle \nabla f(\mathcal{X}^{(t)}), s g^{(t)} + o(s) \rangle + o(\|\mathcal{X}(s) - \mathcal{X}^{(t)}\|_{\mathrm{F}}) \\ &= f(\mathcal{X}^{(t)}) + s a \langle \nabla f(\mathcal{X}^{(t)}), g^{(t)} \rangle + s(1 - a) \langle \nabla f(\mathcal{X}^{(t)}), g^{(t)} \rangle + o(s). \end{aligned}$$

我们从命题 4.7可以推出 $\langle -\nabla f(\mathcal{X}^{(t)}), g^{(t)} \rangle = \|g^{(t)}\|_{\mathrm{F}}^2 > 0$, 于是对足够小的 $s > 0$ 我们有

$$\begin{aligned} f(\mathcal{X}^{(t)}) - f(\mathcal{X}(s)) &= s a \langle -\nabla f(\mathcal{X}^{(t)}), g^{(t)} \rangle + s(1 - a) \|g^{(t)}\|_{\mathrm{F}}^2 + o(s) \\ &\geq s a \langle -\nabla f(\mathcal{X}^{(t)}), g^{(t)} \rangle. \end{aligned}$$

值得注意的是 GRAP 方法是一个求解 $\mathcal{M}_{\leq \mathbf{r}}$ 上优化问题的方法, 尽管在实际当中大概率不会生成一个亏秩点, 但它对于亏秩点 $\mathcal{M}_{\leq \mathbf{r}} \backslash \mathcal{M}_{\mathbf{r}}$ 同样能给出搜索方向. 对于矩阵的情形也有类似的发现; 详见 [63].

4.3.2　全局收敛性

设 $\{\mathcal{X}^{(t)}\}_{t\geq 0}$ 是由算法 12生成的无穷序列. 我们证明算法的全局收敛性, 即稳定性度量 $\|\mathrm{P}_{\mathrm{T}_{\mathcal{X}^{(t)}}\mathcal{M}_{\leq \mathbf{r}}}(-\nabla f(\mathcal{X}^{(t)}))\|_{\mathrm{F}}$ 收敛到 0.

定理 4.8. *设 $\{\mathcal{X}^{(t)}\}_{t\geq 0}$ 是由算法 12生成的无穷序列. 假设 f 是下有界的, 且 f^* 是一个下界, 则*

$$\lim_{t\to\infty} \|\mathrm{P}_{\mathrm{T}_{\mathcal{X}^{(t)}}\mathcal{M}_{\leq \mathbf{r}}}(-\nabla f(\mathcal{X}^{(t)}))\|_{\mathrm{F}} = 0$$

成立. 更进一步, 算法 12 在最多 $\left\lceil f(\mathcal{X}^{(0)})/(s_{\min} a\omega^2\epsilon^2)\right\rceil$ 迭代步后, 返回一个满足 $\|\mathrm{P}_{\mathrm{T}_{\mathcal{X}^{(t)}}\mathcal{M}_{\leq \mathbf{r}}}(-\nabla f(\mathcal{X}^{(t)}))\|_{\mathrm{F}} < \epsilon$ 的点 $\mathcal{X}^{(t)} \in \mathcal{M}_{\leq \mathbf{r}}$.

证明. 根据式 (4-18)–(4-19) 和命题 4.7 我们可以得到

$$\begin{aligned} f(\mathcal{X}^{(t)}) - f(\mathcal{X}^{(t+1)}) &\geq s^{(t)} a\langle -\nabla f(\mathcal{X}^{(t)}), g^{(t)}\rangle \\ &\geq s_{\min} a\|g^{(t)}\|_{\mathrm{F}}^2 \\ &\geq s_{\min} a\omega^2 \|\mathrm{P}_{\mathrm{T}_{\mathcal{X}^{(t)}}\mathcal{M}_{\leq \mathbf{r}}}(-\nabla f(\mathcal{X}^{(t)}))\|_{\mathrm{F}}^2. \end{aligned}$$

于是, 我们有

$$f(\mathcal{X}^{(0)}) - f^* \geq \sum_{t=0}^{\infty} s_{\min} a\omega^2 \|\mathrm{P}_{\mathrm{T}_{\mathcal{X}^{(t)}}\mathcal{M}_{\leq \mathbf{r}}}(-\nabla f(\mathcal{X}^{(t)}))\|_{\mathrm{F}}^2,$$

因此 $\|\mathrm{P}_{\mathrm{T}_{\mathcal{X}^{(t)}}\mathcal{M}_{\leq \mathbf{r}}}(-\nabla f(\mathcal{X}^{(t)}))\|_{\mathrm{F}}$ 收敛到 0.

更进一步, 假设 $\|\mathrm{P}_{\mathrm{T}_{\mathcal{X}^{(t)}}\mathcal{M}_{\leq \mathbf{r}}}(-\nabla f(\mathcal{X}^{(t)}))\|_{\mathrm{F}} \geq \epsilon$ 对 $t = 0, 1, \ldots, T$ 成立. 我们有

$$f(\mathcal{X}^{(0)}) - f(\mathcal{X}^{(T)}) \geq s_{\min} a\omega^2\epsilon^2 T,$$

于是 $T \leq f(\mathcal{X}^{(0)})/(s_{\min} a\omega^2\epsilon^2)$. □

值得注意的是, 上述定理证明了 GRAP 方法需要 $\mathcal{O}(\varepsilon^{-2})$ 个迭代步以达到 ε-稳定点, 这个结果与黎曼优化中经典的结论相合; 可参考 [74, Theorem 2.5].

4.3.3　局部收敛性

设 $\{\mathcal{X}^{(t)}\}_{t\geq 0}$ 是由算法 12生成的无穷序列, 我们通过利用 Łojasiewicz 梯度不等式 [63, Definition 2.1] 证明局部收敛性. 我们称一个点 $\mathcal{X} \in \mathcal{M}_{\leq \mathbf{r}}$ 满足 Łojasiewicz 梯度不等式, 若存在 $\delta, L > 0$ 以及 $\theta \in (0, 1/2]$ 使得

$$|f(\mathcal{X}) - f(\mathcal{Y})|^{1-\theta} \leq L\|\mathrm{P}_{\mathrm{T}_{\mathcal{X}}\mathcal{M}_{\leq \mathbf{r}}}(-\nabla f(\mathcal{Y}))\|_{\mathrm{F}} \tag{4-20}$$

对所有满足 $\|\mathcal{Y} - \mathcal{X}\|_{\mathrm{F}} \leq \delta$ 的 $\mathcal{Y} \in \mathcal{M}_{\leq \mathbf{r}}$ 成立. 在 f 满足式 (4-20)的假设之下, 我们可以证明算法 12 的局部收敛性.

定理 4.9. 设 $\{\mathcal{X}^{(t)}\}_{t\geq 0}$ 是由算法 *12*生成的无穷序列. 假设 f 是下有界的, 且 f^* 是一个下界, 并且满足 *Łojasiewicz* 梯度不等式. 若 $\{\mathcal{X}^{(t)}\}_{t\geq 0}$ 有一个聚点 $\mathcal{X}^*$, 则 $\mathcal{X}^{(t)}$ 收敛到 $\mathcal{X}^*$. 更进一步, 若 $\mathrm{rank_{tc}}(\mathcal{X}^*)=\mathbf{r}$, 则稳定性度量满足 $\|\mathrm{P}_{\mathrm{T}_{\mathcal{X}^*}\mathcal{M}_{\leq\mathbf{r}}}(-\nabla f(\mathcal{X}^*))\|_{\mathrm{F}} = \|\mathrm{grad} f(\mathcal{X}^*)\|_{\mathrm{F}} = 0$ 以及

$$\|\mathcal{X}^{(t)}-\mathcal{X}^*\|_{\mathrm{F}} \leq C\begin{cases} \mathrm{e}^{-ct}, & \text{若 } \theta=\frac{1}{2},\\ t^{-\frac{\theta}{1-2\theta}}, & \text{若 } 0<\theta<\frac{1}{2}\end{cases}$$

对某个 $C, c>0$ 成立.

证明. 我们通过利用 [63, Theorem 2.3, Corollary 2.11] 的结论, 只需证明如下的三个论断: 1) 存在一个常数 $c>0$, 使得

$$f(\mathcal{X}^{(t)})-f(\mathcal{X}^{(t+1)}) \geq c\,\|\mathrm{P}_{\mathrm{T}_{\mathcal{X}^{(t)}}\mathcal{M}_{\leq\mathbf{r}}}(-\nabla f(\mathcal{X}^{(t)}))\|_{\mathrm{F}}\|\mathcal{X}^{(t)}-\mathcal{X}^{(t+1)}\|_{\mathrm{F}};$$

2) 若稳定性度量 $\|\mathrm{P}_{\mathrm{T}_{\mathcal{X}^{(t)}}\mathcal{M}_{\leq\mathbf{r}}}(-\nabla f(\mathcal{X}^{(t)}))\|_{\mathrm{F}}=0$, 则对充分大的 t 有 $\mathcal{X}^{(t+1)}=\mathcal{X}^{(t)}$ 成立; 3) 若 $\mathrm{rank_{tc}}(\mathcal{X}^*)=\mathbf{r}$, 则对充分大的 t 有 $\mathcal{X}^{(t)}\in\mathcal{M}_{\mathbf{r}}$ 成立.

对于第一个论断, 我们根据式 (4-18)–(4-19) 以及推论 4.6 得知

$$\begin{aligned} f(\mathcal{X}^{(t)})-f(\mathcal{X}^{(t+1)}) &\geq a\omega\,\|\mathrm{P}_{\mathrm{T}_{\mathcal{X}^{(t)}}\mathcal{M}_{\leq\mathbf{r}}}(-\nabla f(\mathcal{X}^{(t)}))\|_{\mathrm{F}}\|s^{(t)}g^{(t)}\|_{\mathrm{F}}\\ &\geq \frac{a\omega}{M}\|\mathrm{P}_{\mathrm{T}_{\mathcal{X}^{(t)}}\mathcal{M}_{\leq\mathbf{r}}}(-\nabla f(\mathcal{X}^{(t)}))\|_{\mathrm{F}}\|\mathcal{X}^{(t)}-\mathcal{X}^{(t+1)}\|_{\mathrm{F}}, \end{aligned}$$

这里 $M := 1+d/\sqrt{d+1}>0$.

更进一步, 假设 $\|\mathrm{P}_{\mathrm{T}_{\mathcal{X}^{(t)}}\mathcal{M}_{\leq\mathbf{r}}}(-\nabla f(\mathcal{X}^{(t)}))\|_{\mathrm{F}}=0$ 对某个足够大的 t 成立. 我们根据式 (4-17) 得到

$$\|g^{(t)}\|_{\mathrm{F}} = \|\tilde{\mathrm{P}}_{\mathrm{T}_{\mathcal{X}^{(t)}}\mathcal{M}_{\leq\mathbf{r}}}(-\nabla f(\mathcal{X}^{(t)}))\|_{\mathrm{F}} \leq \|\mathrm{P}_{\mathrm{T}_{\mathcal{X}^{(t)}}\mathcal{M}_{\leq\mathbf{r}}}(-\nabla f(\mathcal{X}^{(t)}))\|_{\mathrm{F}}=0,$$

于是我们有 $g^{(t)}=0$ 以及 $\mathcal{X}^{(t+1)}=\mathcal{X}^{(t)}$.

第三个论断是显然的, 因为 $\mathcal{M}_{\mathbf{r}}$ 是 $\mathcal{M}_{\leq\mathbf{r}}$ 中的一个开集. □

4.3.4 关于有界秩集合上灾难点现象的讨论

我们观察到即使点列 $\{\mathcal{X}^{(t)}\}_{t\geq 0}$ 有一个聚点 $\mathcal{X}^*$, 定理 4.8中的全局收敛性结果 $\lim_{t\to\infty}\|\mathrm{P}_{\mathrm{T}_{\mathcal{X}^{(t)}}\mathcal{M}_{\leq\mathbf{r}}}(-\nabla f(\mathcal{X}^{(t)}))\|_{\mathrm{F}}=0$ 也不能保证 $\mathcal{X}^*$ 是 f 的一个稳定点, 也就是说 $\|\mathrm{P}_{\mathrm{T}_{\mathcal{X}^*}\mathcal{M}_{\leq\mathbf{r}}}(-\nabla f(\mathcal{X}^*))\|_{\mathrm{F}}$ 可能为非 0 值; 参考如下的反例.

例 4.1. 给定 $\mathcal{A}=\mathbf{e}_1\circ\mathbf{e}_1\circ\mathbf{e}_1+\mathbf{e}_3\circ\mathbf{e}_3\circ\mathbf{e}_3\in\mathbb{R}^{n\times n\times n}$ 以及 $\mathbf{r}=(2,2,2)$, 我们考虑目标函数 $f(\mathcal{X})=\|\mathcal{X}-\mathcal{A}\|_{\mathrm{F}}^2$, 以及初始值点 $\mathcal{X}^{(0)}=\mathbf{e}_1\circ\mathbf{e}_1\circ\mathbf{e}_1+\mathbf{e}_2\circ\mathbf{e}_2\circ\mathbf{e}_2$. 则由算法 12 以及固定步长 $s^{(t)}=\alpha\in(0,1)$ 生成的点列为

$$\mathcal{X}^{(t)}=\mathbf{e}_1\circ\mathbf{e}_1\circ\mathbf{e}_1+(1-\alpha)^t\mathbf{e}_2\circ\mathbf{e}_2\circ\mathbf{e}_2,$$

这个点列收敛到 $\mathcal{X}^*=\mathbf{e}_1\circ\mathbf{e}_1\circ\mathbf{e}_1$. 根据定理 4.8, 这个点列满足稳定性度量收敛到 0, 即 $\lim_{t\to\infty}\|\mathrm{P}_{\mathrm{T}_{\mathcal{X}^{(t)}}\mathcal{M}_{\leq\mathbf{r}}}(-\nabla f(\mathcal{X}^{(t)}))\|_{\mathrm{F}}=0$. 然而, 由于 $\nabla f(\mathcal{X}^*)\neq 0$, 根据性质 4.5得知 $\mathcal{X}^*$ 并不是 f 的稳定点.

在近期的研究中, 上述现象被命名为灾难点现象 [89]. 特别地, 点 $\mathcal{X}^* \in \mathcal{M}_{\leq \mathbf{r}}$ 被称为是一个灾难点, 如果存在一个收敛到 $\mathcal{X}^*$ 的序列 $\{\mathcal{X}^{(t)}\} \subseteq \mathcal{M}_{\leq \mathbf{r}}$ 以及一个光滑函数 f, 使得

$$\lim_{t\to\infty} \| \mathrm{P}_{\mathrm{T}_{\mathcal{X}^{(t)}}\mathcal{M}_{\leq \mathbf{r}}}(-\nabla f(\mathcal{X}^{(t)}))\|_{\mathrm{F}} = 0 \text{ 但是 } \| \mathrm{P}_{\mathrm{T}_{\mathcal{X}^*}\mathcal{M}_{\leq \mathbf{r}}}(-\nabla f(\mathcal{X}^*))\|_{\mathrm{F}} > 0.$$

三元组 $(\mathcal{X}^*, \{\mathcal{X}^{(t)}\}, f)$ 被称为一个灾难点组. 我们发现和矩阵代数簇一样, Tucker 张量代数簇在秩亏点处同样的有灾难点问题.

命题 4.10. *如果 $\mathcal{X}^* \in \mathcal{M}_{\leq \mathbf{r}}$ 的 Tucker 秩满足 $\mathrm{rank}_{\mathrm{tc}}(\mathcal{X}^*) = \underline{\mathbf{r}} < \mathbf{r}$, 则 $\mathcal{X}^*$ 是一个灾难点.*

证明. 注意到张量 $\mathcal{X}^*$ 拥有如下的 Tucker 分解 $\mathcal{G}^* \times_1 \mathbf{U}_1^* \cdots \times_d \mathbf{U}_d^*$, 这里 $\mathcal{G}^* \in \mathbb{R}^{\underline{r}_1 \times \underline{r}_2 \times \cdots \times \underline{r}_d}$, $\mathbf{U}_k^* \in \mathrm{St}(\underline{r}_k, n_k)$ 以及 $k \in [d]$. 现在我们考虑第 k_0 个展平矩阵 $\mathbf{X}_{(k_0)}^* = \mathbf{U}_{k_0}^* \mathbf{G}_{(k_0)}^* (\mathbf{V}_{k_0}^*)^\top \in \mathbb{R}^{n_{k_0} \times n_{-k_0}}$, 这里 $\mathbf{V}_{k_0}^* := (\mathbf{U}_j^*)^{\otimes j \neq k_0}$. 由于 $\underline{r}_{k_0} < n_{k_0}$ 以及 $\underline{r}_{-k_0} < n_{-k_0}$, 存在矩阵 $\mathbf{u} \in \mathbb{R}^{n_{k_0}} \setminus \{0\}$ 和 $\mathbf{v} \in \mathbb{R}^{n_{-k_0}} \setminus \{0\}$, 使得 $(\mathbf{U}_{k_0}^*)^\top \mathbf{u} = 0$ 以及 $(\mathbf{V}_{k_0}^*)^\top \mathbf{v} = 0$. 我们目标是构造一个序列 $\{\mathcal{X}^{(t)}\} \subseteq \mathcal{M}_{\leq \mathbf{r}}$ 以及一个函数 f 使得 $\mathcal{X}^{(t)}$ 收敛到 $\mathcal{X}^*$ 以及 $\| \mathrm{P}_{\mathrm{T}_{\mathcal{X}^{(t)}}\mathcal{M}_{\leq \mathbf{r}}}(-\nabla f(\mathcal{X}^{(t)}))\|_{\mathrm{F}}$ 收敛到 0, 但是 $\mathcal{X}^*$ 并不是 f 的一阶稳定点.

首先, 我们考虑序列 $\mathcal{X}^{(t)} \in \mathcal{M}_{\leq \mathbf{r}}$, 它的定义是 $\mathcal{X}^{(t)} = \mathcal{G}^{(t)} \times_{k=1}^d \begin{bmatrix} \mathbf{U}_k^* & \mathbf{U}_{k,1} \end{bmatrix}$, 这里 $\mathbf{U}_{k,1} \in \mathrm{St}(r_k - \underline{r}_k, n_k)$ 以及 $\mathbf{U}_{k,1}^\top \mathbf{U}_k^* = 0$, $\mathcal{G}^{(t)}(i_1, \ldots, i_d) := \mathcal{G}^*(i_1, \ldots, i_d)$ 若 $(i_1, \ldots, i_d) \leq \underline{\mathbf{r}}$, $\mathcal{G}^{(t)}(i_1, \ldots, i_d) := \frac{1}{t}\bar{\mathcal{G}}(i_1 - \underline{r}_1, \ldots, i_d - \underline{r}_d)$ 若 $(i_1, \ldots, i_d) > \underline{\mathbf{r}}$, 否则 $\mathcal{G}^{(t)}(i_1, \ldots, i_d) := 0$; 以及 $\bar{\mathcal{G}} \in \mathbb{R}^{(r_1 - \underline{r}_1) \times (r_2 - \underline{r}_2) \times \cdots \times (r_d - \underline{r}_d)}$ 满足 $\mathrm{rank}_{\mathrm{tc}}(\bar{\mathcal{G}}) = \mathbf{r} - \underline{\mathbf{r}}$. 于是, 我们有 $\mathrm{rank}_{\mathrm{tc}}(\mathcal{X}^{(t)}) = \mathbf{r}$ 以及 $\mathcal{X}^{(t)}$ 收敛到 $\mathcal{X}^*$.

接下来, 我们构造函数 $f(\mathcal{X}) = \mathbf{u}^\top \mathbf{X}_{(k_0)}^* \mathbf{v}$. 由于 $\mathcal{X}^*$ 是秩亏的并且梯度满足 $\nabla f(\mathcal{X}) = \mathrm{ten}_{(k_0)}(\mathbf{u}\mathbf{v}^\top) \neq 0$, 由命题 4.5 得知 $\mathcal{X}^*$ 不是一个一阶稳定点. 更进一步, 对于 $\mathcal{X}^{(t)} \in \mathcal{M}_{\mathbf{r}}$, 我们有

$$\begin{aligned}
\mathrm{P}_{\mathrm{T}_{\mathcal{X}^{(t)}}\mathcal{M}_{\leq \mathbf{r}}}(-\nabla f(\mathcal{X}^{(t)})) &= \mathrm{P}_{\mathrm{T}_{\mathcal{X}^{(t)}}\mathcal{M}_{\mathbf{r}}}(-\nabla f(\mathcal{X}^{(t)})) \\
&= -\left(\mathrm{ten}_{(k_0)}(\mathbf{u}\mathbf{v}^\top) \times_{k=1}^d (\mathbf{U}_k^*)^\top\right) \times_{k=1}^d \mathbf{U}_k^* \\
&\quad - \sum_{k=1}^d \mathcal{G}^{(t)} \times_k \left(\mathrm{P}_{\mathbf{U}_k^*}^{\perp}(\mathrm{ten}_{(k_0)}(\mathbf{u}\mathbf{v}^\top) \times_{j\neq k} (\mathbf{U}_j^*)^\top)_{(k)} (\mathcal{G}_{(k)}^{(t)})^\dagger\right) \times_{j\neq k} \mathbf{U}_j^* \\
&= -\mathcal{G}^{(t)} \times_{k_0} \left(\mathrm{P}_{\mathbf{U}_{k_0}^*}^{\perp} \mathbf{u}\mathbf{v}^\top \mathbf{V}_{k_0}^* (\mathcal{G}_{(k_0)}^{(t)})^\dagger\right) \times_{j\neq k_0} \mathbf{U}_j^* \\
&= 0,
\end{aligned}$$

这里我们用到式 (1-14) 以及对所有的 $k \neq k_0$ 成立的 $\mathrm{ten}_{(k_0)}(\mathbf{u}\mathbf{v}^\top) \times_{j\neq k} (\mathbf{U}_j^*)^\top = 0$.

因此, 这个三元组 $(\mathcal{X}^*, \{\mathcal{X}^{(t)}\}, f)$ 是一个灾难点组. □

针对 Tucker 张量代数簇克服灾难点问题不在本章的讨论范围内, 我们将在未来的工作中讨论如何克服灾难点问题.

4.4 一个免收缩映射的近似投影梯度法

算法 12中引入了收缩映射 $\mathrm{R}_{\mathcal{X}}^{\mathrm{HO}}$ 保证点列的可行性 (即 Tucker 秩小于等于 $\mathbf{r}$), 收缩映射通过对 d 阶大小为 $(\underline{r}_1 + r_1) \times \cdots \times (\underline{r}_d + r_d)$ 的张量执行高阶奇异值分解 HOSVD. 我们思考是否能利用切锥 $\mathrm{T}_{\mathcal{X}}\mathcal{M}_{\leq \mathbf{r}}$ 的几何结构, 设计出合适的线搜索方向, 不需要收缩映射也能保证可行性, 进而节省计算量. 在本节中, 我们将通过设计部分投影以提出 $\mathcal{M}_{\leq \mathbf{r}}$ 上一个免收缩映射的近似投影梯度法.

4.4.1 新的部分投影

我们在此回顾, 在 $\mathcal{X} = \mathcal{G} \times_{k=1}^{d} \mathbf{U}_k$ (这里 $\mathrm{rank_{tc}}(\mathcal{X}) = \underline{\mathbf{r}}$) 处切锥 $\mathrm{T}_{\mathcal{X}}\mathcal{M}_{\leq \mathbf{r}}$ 中的任意元素 $\mathcal{V}$ 可以用参数 $(\mathcal{C}, \{\mathbf{U}_{k,1}\}_{k=1}^{d}, \{\mathbf{U}_{k,2}\}_{k=1}^{d}, \{\mathbf{R}_{k,2}\}_{k=1}^{d})$ 来表达; 见 (4-5), 具体而言可以被参数化为

$$\mathcal{V} = \mathcal{V}_0 + \sum_{k=1}^{d} \mathcal{V}_k = \mathcal{C} \times_{k=1}^{d} \begin{bmatrix} \mathbf{U}_k & \mathbf{U}_{k,1} \end{bmatrix} + \sum_{k=1}^{d} \mathcal{G} \times_k (\mathbf{U}_{k,2}\mathbf{R}_{k,2}) \times_{j \neq k} \mathbf{U}_j.$$

我们根据式 (4-8)–(4-9), 发现沿着这 $(d+1)$ 个搜索方向 $\{\mathcal{V}_k\}$ 可以避免引入收缩映射以保证可行性. 因此, 我们根据 $\{\mathcal{V}_k\}$ 的表达式, 设计出新的部分投影.

给定 $\mathcal{A} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$, 我们定义新的部分投影如下:

$$\mathrm{P}_0(\mathcal{A}) := \arg\min_{\mathcal{V}_0} \left\{ \|\mathcal{V}_0 - \mathcal{A}\|_{\mathrm{F}} : \mathcal{V}_0 = \mathcal{C} \times_{k=1}^{d} \begin{bmatrix} \mathbf{U}_k & \mathbf{U}_{k,1} \end{bmatrix} \in \mathrm{T}_{\mathcal{X}}\mathcal{M}_{\leq \mathbf{r}} \right\}, \tag{4-21}$$

$$\mathrm{P}_k(\mathcal{A}) := \arg\min_{\mathcal{V}_k} \left\{ \|\mathcal{V}_k - \mathcal{A}\|_{\mathrm{F}} : \mathcal{V}_k = \mathcal{G} \times_k (\mathbf{U}_{k,2}\mathbf{R}_{k,2}) \times_{j \neq k} \mathbf{U}_j \in \mathrm{T}_{\mathcal{X}}\mathcal{M}_{\leq \mathbf{r}} \right\}. \tag{4-22}$$

由于 (4-21) 并不具有显式解, 我们考虑它的一个近似

$$\tilde{\mathrm{P}}_0(\mathcal{A}) := \mathcal{A} \times_{k=1}^{d} \mathrm{P}_{\tilde{\mathbf{S}}_k}, \tag{4-23}$$

这个近似正好是给定满足 $\tilde{\mathbf{U}}_{k,1}^{\top}\mathbf{U}_k = 0$ 的 $\tilde{\mathbf{U}}_{k,1} \in \mathrm{St}(r_k - \underline{r}_k, n_k)$ 以后近似投影 $\tilde{\mathrm{P}}_{\mathrm{T}_{\mathcal{X}}\mathcal{M}_{\leq \mathbf{r}}}(\mathcal{A})$, 即式 (4-16)的第一项, 这里 $\tilde{\mathbf{S}}_k := [\mathbf{U}_k \ \tilde{\mathbf{U}}_{k,1}]$ 以及 $k \in [d]$. 值得注意的是部分投影 $\tilde{\mathrm{P}}_0(\mathcal{A})$ 需要事先知道 $\{\tilde{\mathbf{U}}_{k,1}\}_{k=1}^{d}$. 此外, 我们通过固定式 (4-22)中的参数 $\mathbf{U}_{k,2}$ 以及利用 $[\mathbf{U}_k \ \mathbf{U}_{k,1} \ \mathbf{U}_{k,2}] \in \mathcal{O}(n_k)$ 和 (4-14), 得到 $\mathrm{P}_k(\mathcal{A})$ 具有如下的形式

$$\mathrm{P}_k(\mathcal{A}) = \mathcal{G} \times_k \left(\mathrm{P}_{\mathbf{U}_{k,2}} \left(\mathcal{A} \times_{j \neq k} \mathbf{U}_j^{\top} \right)_{(k)} \mathbf{G}_{(k)}^{\dagger} \right) \times_{j \neq k} \mathbf{U}_j.$$

类似的, 由于 $\mathbf{U}_{k,2}$ 事先未知, 我们将投影 $\mathrm{P}_{\mathbf{U}_{k,2}}$ 替换为 $\mathrm{P}_{\mathbf{U}_k}^{\perp}$ 并得到如下的部分投影

$$\tilde{\mathrm{P}}_k(\mathcal{A}) = \mathcal{G} \times_k \left(\mathrm{P}_{\mathbf{U}_k}^{\perp} \left(\mathcal{A} \times_{j \neq k} \mathbf{U}_j^{\top} \right)_{(k)} \mathbf{G}_{(k)}^{\dagger} \right) \times_{j \neq k} \mathbf{U}_j, \tag{4-24}$$

这个部分投影与命题 4.7 中近似投影 (4-16) 中的 $\tilde{\mathcal{V}}_k$ 项是不同的, 这是因为 $\mathrm{P}_{\mathbf{U}_k}^{\perp} \neq \mathrm{P}_{\tilde{\mathbf{S}}_k}^{\perp}$.

式 (4-23)和 (4-24)中的部分投影满足 $\tilde{\mathrm{P}}_k(\mathcal{A}) \in \mathrm{T}_{\mathcal{X}}\mathcal{M}_{\leq \mathbf{r}}$ 以及 $\langle \mathcal{A}, \tilde{\mathrm{P}}_k(\mathcal{A})\rangle = \|\tilde{\mathrm{P}}_k(\mathcal{A})\|_{\mathrm{F}}^2$, 这两个论断可通过命题 4.7类似的方式证明. 此外, 我们还可以通过类似于式 (4-8)–(4-9)中的过程证明 $\mathrm{rank}_{\mathrm{tc}}(\mathcal{X} + \tilde{\mathrm{P}}_k(\mathcal{A})) \leq \mathbf{r}$, 即

$$\mathcal{X} + \tilde{\mathrm{P}}_k(\mathcal{A}) \in \mathcal{M}_{\leq \mathbf{r}}.$$

然而, 这个性质对两个不同的部分投影 $\tilde{\mathrm{P}}_j(\mathcal{A})$ 和 $\tilde{\mathrm{P}}_k(\mathcal{A})$ $(j \neq k)$ 的线性组合不再成立, 也就是说 $\mathrm{rank}_{\mathrm{tc}}(\mathcal{X} + \tilde{\mathrm{P}}_j(\mathcal{A}) + \tilde{\mathrm{P}}_k(\mathcal{A}))$ 可能会大于 $\mathbf{r}$.

4.4.2　算法以及收敛性结果

总而言之, 我们提出如下的部分投影

$$\hat{\mathrm{P}}_{\mathrm{T}_{\mathcal{X}}\mathcal{M}_{\leq \mathbf{r}}}(\mathcal{A}) := \underset{\mathcal{V} \in \{\tilde{\mathrm{P}}_0(\mathcal{A}), \ldots, \tilde{\mathrm{P}}_d(\mathcal{A})\}}{\arg\max} \|\mathcal{V}\|_{\mathrm{F}}. \tag{4-25}$$

通过利用式 (4-25)定义的部分投影, 我们在算法 13中展示了免收缩映射的梯度相关近似投影梯度法 (rfGRAP) 的算法框架. rfGRAP 方法的每步迭代公式如下

$$\mathcal{X}^{(t+1)} = \mathcal{X}^{(t)} + s^{(t)} \hat{\mathrm{P}}_{\mathrm{T}_{\mathcal{X}^{(t)}}\mathcal{M}_{\leq \mathbf{r}}}(-\nabla f(\mathcal{X}^{(t)})),$$

这里 $s^{(t)}$ 同样由 Armijo 回溯线搜索 (4-19)给出. 相比于 GRAP 方法 (即算法 12), 在算法 13中没有收缩映射.

算法 13 免收缩映射的梯度相关近似投影梯度法 (rfGRAP)

输入: 初始点 $\mathcal{X}^{(0)} \in \mathcal{M}_{\leq \mathbf{r}}$, $\omega \in (0, 1/\sqrt{d+1})$, 线搜索参数 $\rho, a \in (0,1)$, $s_{\min} > 0$.

1: **while** 停机准则未被满足 **do**
2: 选择随机的 $\tilde{\mathbf{U}}_{1,1}, \tilde{\mathbf{U}}_{2,1}, \ldots, \tilde{\mathbf{U}}_{d,1}$ 并计算 $g^{(t)} = \hat{\mathrm{P}}_{\mathrm{T}_{\mathcal{X}^{(t)}}\mathcal{M}_{\leq \mathbf{r}}}(-\nabla f(\mathcal{X}^{(t)}))$, 直到角度条件 (4-18) 被满足.
3: 通过 Armijo 回溯线搜索 (4-19)选择步长 $s^{(t)}$.
4: 更新 $\mathcal{X}^{(t+1)} = \mathcal{X}^{(t)} + s^{(t)} g^{(t)}$ 以及 $t = t + 1$.
5: **end while**

输出: $\mathcal{X}^{(t)}$

类似于算法 12, 如果 $g^{(t)} = \hat{\mathrm{P}}_{\mathrm{T}_{\mathcal{X}^{(t)}}\mathcal{M}_{\leq \mathbf{r}}}(-\nabla f(\mathcal{X}^{(t)}))$ 不满足角度条件 (4-18), 我们将使用不同的 $\tilde{\mathbf{U}}_{k,1}$ 重复进行投影, 直至角度条件被满足. 由于部分投影 (4-25) 仅利用了切锥的部分信息, 我们选择参数 $\omega \in (0, 1/\sqrt{d+1})$. 令 (4-23) 中的 $\tilde{\mathbf{U}}_{k,1}$

为 (4-13) 的全局极小解 $\mathbf{U}_{k,1}^*$, 则由 (4-23)–(4-25) 可得

$$\begin{aligned}
\|\hat{\mathrm{P}}_{\mathrm{T}_{\mathcal{X}}\mathcal{M}_{\leq\mathbf{r}}}(\mathcal{A})\|_{\mathrm{F}}^2 &= \max_{k=0,1,\ldots,d}\{\|\tilde{\mathrm{P}}_k(\mathcal{A})\|_{\mathrm{F}}^2\} \geq \frac{1}{d+1}\sum_{k=0}^{d}\|\tilde{\mathrm{P}}_k(\mathcal{A})\|_{\mathrm{F}}^2 \\
&= \frac{1}{d+1}\Big(\|\mathcal{A}\times_{k=1}^{d}\mathrm{P}_{\mathbf{S}_k^*}\|_{\mathrm{F}}^2 + \sum_{k=1}^{d}\|\mathcal{G}\times_k(\mathrm{P}_{\mathbf{U}_k}^{\perp}(\mathcal{A}\times_{j\neq k}\mathbf{U}_j^{\top})_{(k)}\mathbf{G}_{(k)}^{\dagger})\times_{j\neq k}\mathbf{U}_j\|_{\mathrm{F}}^2\Big) \\
&\geq \frac{1}{d+1}\Big(\|\mathcal{A}\times_{k=1}^{d}\mathrm{P}_{\mathbf{S}_k^*}\|_{\mathrm{F}}^2 + \sum_{k=1}^{d}\|\mathcal{G}\times_k(\mathrm{P}_{\mathbf{S}_k^*}^{\perp}(\mathcal{A}\times_{j\neq k}\mathbf{U}_j^{\top})_{(k)}\mathbf{G}_{(k)}^{\dagger})\times_{j\neq k}\mathbf{U}_j\|_{\mathrm{F}}^2\Big) \\
&= \frac{1}{d+1}\|\mathrm{P}_{\mathrm{T}_{\mathcal{X}}\mathcal{M}_{\leq\mathbf{r}}}(\mathcal{A})\|_{\mathrm{F}}^2,
\end{aligned}$$

其中 $\mathbf{S}_k^* = [\mathbf{U}_k\ \mathbf{U}_{k,1}^*]$ 和 $\mathrm{span}(\mathbf{S}_k^*)^{\perp} \subseteq \mathrm{span}(\mathbf{U}_k)^{\perp}$. 我们可以得到

$$\langle\mathcal{A}, \hat{\mathrm{P}}_{\mathrm{T}_{\mathcal{X}}\mathcal{M}_{\leq\mathbf{r}}}(\mathcal{A})\rangle = \|\hat{\mathrm{P}}_{\mathrm{T}_{\mathcal{X}}\mathcal{M}_{\leq\mathbf{r}}}(\mathcal{A})\|_{\mathrm{F}}^2 \geq \frac{1}{\sqrt{d+1}}\|\mathrm{P}_{\mathrm{T}_{\mathcal{X}}\mathcal{M}_{\leq\mathbf{r}}}(\mathcal{A})\|_{\mathrm{F}}\|\hat{\mathrm{P}}_{\mathrm{T}_{\mathcal{X}}\mathcal{M}_{\leq\mathbf{r}}}(\mathcal{A})\|_{\mathrm{F}}.$$

因此, 在算法 13 的角度条件 (4-18) 中, 将参数 ω 取为 $\omega \in (0, 1/\sqrt{d+1})$ 是合理的. 注意到, 利用 $\langle\mathcal{A}, \tilde{\mathrm{P}}_k(\mathcal{A})\rangle = \|\tilde{\mathrm{P}}_k(\mathcal{A})\|_{\mathrm{F}}^2, k = 0, 1, \ldots, d$, 我们可以得到 $\|\hat{\mathrm{P}}_{\mathrm{T}_{\mathcal{X}}\mathcal{M}_{\leq\mathbf{r}}}(\mathcal{A})\|_{\mathrm{F}} \leq \|\mathrm{P}_{\mathrm{T}_{\mathcal{X}}\mathcal{M}_{\leq\mathbf{r}}}(\mathcal{A})\|_{\mathrm{F}}$, 这一结论与 (4-17) 类似.

rfGRAP 方法的全局与局部收敛性可以采用与定理 4.8 和定理 4.9 类似的方式进行证明.

定理 4.11. *设 $\{\mathcal{X}^{(t)}\}_{t\geq0}$ 为算法 13 生成的无穷序列. 假设函数 f 在下界 f^* 上有界, 并满足 Łojasiewicz 梯度不等式, 则有*

$$\lim_{t\to\infty}\|\mathrm{P}_{\mathrm{T}_{\mathcal{X}^{(t)}}\mathcal{M}_{\leq\mathbf{r}}}(-\nabla f(\mathcal{X}^{(t)}))\|_{\mathrm{F}} = 0.$$

在至多 $\lceil f(\mathcal{X}^{(0)})/(s_{\min}a\,\omega^2\epsilon^2)\rceil$ 次迭代后, 返回满足 $\|\mathrm{P}_{\mathrm{T}_{\mathcal{X}^{(t)}}\mathcal{M}_{\leq\mathbf{r}}}(-\nabla f(\mathcal{X}^{(t)}))\|_{\mathrm{F}} < \epsilon$ 的点 $\mathcal{X}^{(t)} \in \mathcal{M}_{\leq\mathbf{r}}$. 若序列 $\{\mathcal{X}^{(t)}\}_{t\geq0}$ 存在聚点 $\mathcal{X}^*$, 则 $\mathcal{X}^{(t)}$ 收敛到 $\mathcal{X}^*$. 进一步地, 若 $\mathrm{rank}_{\mathrm{tc}}(\mathcal{X}^*) = \mathbf{r}$, 则稳定性度量 $\|\mathrm{P}_{\mathrm{T}_{\mathcal{X}^*}\mathcal{M}_{\leq\mathbf{r}}}(-\nabla f(\mathcal{X}^*))\|_{\mathrm{F}} = \|\mathrm{grad}f(\mathcal{X}^*)\|_{\mathrm{F}} = 0$ 以及存在常数 $C, c > 0$ 使得

$$\|\mathcal{X}^{(t)} - \mathcal{X}^*\|_{\mathrm{F}} \leq C\begin{cases}\mathrm{e}^{-ct}, & \text{若 } \theta = \frac{1}{2},\\ t^{-\frac{\theta}{1-2\theta}}, & \text{若 } 0 < \theta < \frac{1}{2}\end{cases}$$

成立.

4.4.3 与矩阵结果之间的联系

我们研究 rfGRAP 方法在矩阵情形下与现有方法之间的联系.

具体而言, 给定矩阵 $\mathbf{A} \in \mathbb{R}^{m\times n}$ 以及 $\mathbf{X} \in \mathbb{R}_{r}^{m\times n}$ 的奇异值分解 $\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^{\top}$, 则 (4-23)–(4-24) 中提出的部分投影 $\{\tilde{\mathrm{P}}_k(\mathbf{A})\}_{k=0}^{d}$ 在 $d = 2$ 的情况下可化简为

$$\begin{aligned}
\tilde{\mathrm{P}}_0(\mathbf{A}) &= \mathrm{P}_{[\mathbf{U}\ \mathbf{U}_1]}\mathbf{A}\mathrm{P}_{[\mathbf{V}\ \mathbf{V}_1]},\\
\tilde{\mathrm{P}}_1(\mathbf{A}) &= \mathrm{P}_{\mathbf{U}}^{\perp}\mathbf{A}\mathrm{P}_{\mathbf{V}},\\
\tilde{\mathrm{P}}_2(\mathbf{A}) &= \mathrm{P}_{\mathbf{U}}\mathbf{A}\mathrm{P}_{\mathbf{V}}^{\perp},
\end{aligned}$$

这里 $\mathbf{U}_1 \in \mathrm{St}(r-\underline{r},m), \mathbf{V}_1 \in \mathrm{St}(r-\underline{r},n)$ 即为式 (4-23)中的 $\tilde{\mathbf{U}}_{1,1}$ 和 $\tilde{\mathbf{U}}_{2,1}$, 它们由矩阵 $\mathrm{P}_{\mathbf{U}}^{\perp}\mathbf{A}\mathrm{P}_{\mathbf{V}}^{\perp}$ 的 $(r-\underline{r})$ 个左右奇异向量选取. 再给定 $\mathbf{U}_2 \in \mathrm{St}(m-r,m)$ 满足 $[\mathbf{U}\,\mathbf{U}_1\,\mathbf{U}_2] \in \mathcal{O}(m)$ 以及 $\mathbf{V}_2 \in \mathrm{St}(n-r,n)$ 满足 $[\mathbf{V}\,\mathbf{V}_1\,\mathbf{V}_2] \in \mathcal{O}(n)$. 于是, 这些部分投影可如图 4-1 所示表示为

$$\tilde{\mathrm{P}}_0(\mathbf{A}) = \begin{bmatrix}\mathbf{U} & \mathbf{U}_1 & \mathbf{U}_2\end{bmatrix} \begin{bmatrix}\mathbf{V} & \mathbf{V}_1 & \mathbf{V}_2\end{bmatrix}^{\mathsf{T}},$$

$$\tilde{\mathrm{P}}_1(\mathbf{A}) = \begin{bmatrix}\mathbf{U} & \mathbf{U}_1 & \mathbf{U}_2\end{bmatrix} \begin{bmatrix}\mathbf{V} & \mathbf{V}_1 & \mathbf{V}_2\end{bmatrix}^{\mathsf{T}},$$

$$\tilde{\mathrm{P}}_2(\mathbf{A}) = \begin{bmatrix}\mathbf{U} & \mathbf{U}_1 & \mathbf{U}_2\end{bmatrix} \begin{bmatrix}\mathbf{V} & \mathbf{V}_1 & \mathbf{V}_2\end{bmatrix}^{\mathsf{T}},$$

这与 [63, §3.4] 中的“部分投影”

$$\check{\mathrm{P}}_1(\mathbf{A}) = \begin{bmatrix}\mathbf{U} & \mathbf{U}_1 & \mathbf{U}_2\end{bmatrix} \begin{bmatrix}\mathbf{V} & \mathbf{V}_1 & \mathbf{V}_2\end{bmatrix}^{\mathsf{T}},$$

$$\check{\mathrm{P}}_2(\mathbf{A}) = \begin{bmatrix}\mathbf{U} & \mathbf{U}_1 & \mathbf{U}_2\end{bmatrix} \begin{bmatrix}\mathbf{V} & \mathbf{V}_1 & \mathbf{V}_2\end{bmatrix}^{\mathsf{T}}$$

是不同的. 因此, 所提出的部分投影 $\hat{\mathrm{P}}_{\mathrm{T}_{\mathbf{X}}\mathbb{R}^{m\times n}_{\leq r}}$ 同样可以作为矩阵簇优化中的一种新的免收缩映射搜索方向.

4.5　Tucker 秩自适应算法

在实际问题当中, 如何选择一个合适的秩参数 $\mathbf{r}$ 是一个具有挑战的问题. 一个较大的秩参数 $\mathbf{r}$ 由于可以导出一个较大的搜索空间故而可能得到一个比较好的解, 但与此同时也加大了计算量. 此外, 正如我们在例 4.1中所看到的, 即使由 GRAP 算法生成的序列满足 $\|\mathrm{P}_{\mathrm{T}_{\mathcal{X}^{(t)}}\mathcal{M}_{\leq\mathbf{r}}}(-\nabla f(\mathcal{X}^{(t)}))\|_{\mathrm{F}}$ 收敛到 0, 若聚点是一个亏秩点, 则该聚点可能不是稳定点. 类似的挑战对于在 Tucker 张量流形 $\mathcal{M}_{\mathbf{r}}$ 上的黎曼优化方法上同样存在; 详情可参见 [117, Section 5.1]. 因此, 在本节中, 我们的目标是通过利用如下包含关系

$$\mathrm{T}_{\mathcal{X}^{(t)}}\mathcal{M}_{\underline{\mathbf{r}}^{(t)}} + \mathrm{N}_{\leq\ell^{(t)}}(\mathcal{X}^{(t)}) \subseteq \mathrm{T}_{\mathcal{X}^{(t)}}\mathcal{M}_{\leq\mathbf{r}}, \tag{4-26}$$

构造搜索方向, 并设计一个能自适应调整迭代点 $\mathcal{X}^{(t)} = \mathcal{G}^{(t)} \times_{k=1}^{d} \mathbf{U}_k^{(t)}$ 秩的优化算法, 这里 $\mathrm{N}_{\leq\ell^{(t)}}(\mathcal{X}^{(t)}) := \mathcal{M}_{\leq\ell^{(t)}} \cap \left(\bigotimes_{k=1}^{d} \mathrm{span}(\mathbf{U}_k^{(t)})^{\perp}\right) \subseteq \mathrm{N}_{\mathcal{X}^{(t)}}\mathcal{M}_{\underline{\mathbf{r}}^{(t)}}$ 以及 $\ell^{(t)} \in \mathbb{N}_+^d$ 满足 $\ell^{(t)} \leq \mathbf{r} - \underline{\mathbf{r}}^{(t)}$.

总的来说, 我们首先通过黎曼优化方法让迭代点在流形 $\mathcal{M}_{\underline{\mathbf{r}}^{(t)}}$ 上更新, 详见4.5.1节. 接下来, 在4.5.2节与4.5.3节中, 我们设计秩减与秩增机制以动态调整迭代点 $\mathcal{X}^{(t)}$ 的秩. 我们将在4.5.4节中提出 Tucker 秩自适应方法 (TRAM), 并在4.5.5节中分析 TRAM 的收敛性. 最后, 我们在4.5.6节中讨论 TRAM 的实际计算细节.

4.5.1　在固定秩流形上的线搜索方法

给定一个点 $\tilde{\mathcal{X}}^{(t)} \in \mathcal{M}_{\underline{\mathbf{r}}^{(t)}}$, 我们根据式 (4-5) 发现 $\mathrm{T}_{\tilde{\mathcal{X}}^{(t)}}\mathcal{M}_{\underline{\mathbf{r}}^{(t)}} \subseteq \mathrm{T}_{\tilde{\mathcal{X}}^{(t)}}\mathcal{M}_{\leq\mathbf{r}}$. 因此, 函数 f 在 $\tilde{\mathcal{X}}^{(t)}$ 处关于流形 $\mathcal{M}_{\underline{\mathbf{r}}^{(t)}}$ 的负黎曼梯度提供了一个具有显式表达

式 (1-14)的搜索方向. 算法 14展示了在流形 $\mathcal{M}_{\underline{\mathbf{r}}^{(t)}}$ 上的黎曼梯度法 (RGD) 的框架.

算法 14 在流形 $\mathcal{M}_{\underline{\mathbf{r}}^{(t)}}$ 上的黎曼梯度法

输入: 初始点 $\mathcal{Y}^{(0)} = \tilde{\mathcal{X}}^{(t)} \in \mathcal{M}_{\underline{\mathbf{r}}^{(t)}}$; 回溯线搜索的参数 $\rho, a \in (0,1), s_{\min} > 0$.

1: **while** 停机准则未被满足 **do**
2: 通过式 (1-14)计算 $g^{(i)} = -\mathrm{grad} f(\mathcal{Y}^{(i)}) = \mathrm{P}_{\mathrm{T}_{\mathcal{Y}^{(i)}}\mathcal{M}_{\underline{\mathbf{r}}^{(t)}}}(-\nabla f(\mathcal{Y}^{(i)}))$.
3: 通过 Armijo 线搜索, 即式 (4-19)选择步长 $s^{(i)}$.
4: 更新 $\mathcal{Y}^{(i+1)} = \mathrm{P}^{\mathrm{HO}}_{\underline{\mathbf{r}}^{(t)}}\left(\mathcal{Y}^{(i)} + s^{(i)} g^{(i)}\right)$ 以及 $i = i + 1$.
5: **end while**

输出: 迭代点 $\mathcal{X}^{(t)} = \mathcal{Y}^{(i)}$.

设 $\{\mathcal{Y}^{(i)}\}_{i\geq 0}$ 是从初始点 $\mathcal{Y}^{(0)} = \tilde{\mathcal{X}}^{(t)}$ 出发由 RGD 方法生成的点列. RGD 方法通过如下的更新格式来更新 $\mathcal{Y}^{(i)}$:

$$\mathcal{Y}^{(i+1)} = \mathrm{P}^{\mathrm{HO}}_{\underline{\mathbf{r}}^{(t)}}\left(\mathcal{Y}^{(i)} - s^{(i)} \mathrm{grad} f(\mathcal{Y}^{(i)})\right).$$

为保证收敛性, 我们通过 Armijo 线搜索, 即式 (4-19)选择步长. 算法的终止准则为: 1) 探测到亏秩点, 即 $\mathcal{Y}$ 的至少一个 k 模态展平矩阵满足 $\sigma_{\min}(\mathbf{Y}^{(i)}_{(k)})/\sigma_{\max}(\mathbf{Y}^{(i)}_{(k)}) \leq \Delta$; 2) 稳定性条件, 即黎曼梯度对某个阈值 $\varepsilon_R^{(t)} > 0$ 满足 $\|\mathrm{grad} f(\mathcal{Y}^{(i)})\|_{\mathrm{F}} \leq \varepsilon_R^{(t)}$. 请注意我们总是优先检查秩亏性, 再检查稳定性条件.

4.5.2 秩减机制

给定一个算法 14返回的 $\mathcal{X}^{(t)} = \mathcal{G}^{(t)} \times_{k=1}^{d} \mathbf{U}_k^{(t)}$, 若 RGD 是由于探测到亏秩点而停止, 则我们将通过秩减机制, 手动调整迭代点的秩. 秩减机制有利于减少参数数量进而减小计算量. 算法 15展示了秩减机制的计算细节.

算法 15 秩减机制

输入: $\mathcal{X}^{(t)} = \mathcal{G}^{(t)} \times_{k=1}^{d} \mathbf{U}_k^{(t)}$, $\Delta > 0$, $\rho_1 \in (0,1)$.

1: **for** $k = 1, 2, \ldots, d$ **do**
2: 计算矩阵 $\mathbf{G}^{(t)}_{(k)}$ 的奇异值 $\sigma_{1,k} \geq \cdots \geq \sigma_{r_k^{(t)},k} > 0$.
3: **end for**
4: **repeat**
5: 对于所有的 $k = 1, \ldots, d$, 找到下标 $\hat{r}_k = \min\{i : \sigma_{i+1,k} < \Delta\sigma_{1,k}\}$.
6: 计算 $\hat{\mathcal{X}} = \mathrm{P}^{\mathrm{HO}}_{\leq \hat{\mathbf{r}}}(\mathcal{X}^{(t)})$.
7: 设置 $\Delta = \rho_1 \Delta$.
8: **until** $f(\hat{\mathcal{X}}) \leq f(\mathcal{X}^{(t)})$

输出: $\tilde{\mathcal{X}}^{(t+1)} = \hat{\mathcal{X}}$ 以及新的 Tucker 秩 $\underline{\mathbf{r}}^{(t+1)} = \hat{\mathbf{r}}$.

特别地, 我们生成点 $\mathcal{X}^{(t)}$ 的一个秩 $\hat{\mathbf{r}}$ 截断

$$\hat{r}_k := \min\{i : \sigma_{i+1,k} < \Delta\sigma_{1,k}\} \qquad \text{或者} \qquad \hat{r}_k := \underline{r}_k^{(t)} \text{ 若 } \sigma_{\underline{r}_k^{(t)},k} \geq \Delta\sigma_{1,k},$$

这里 $\sigma_{1,k} \geq \cdots \geq \sigma_{\underline{r}_k^{(t)},k}$ 是矩阵 $\mathbf{X}_{(k)}^{(t)}$ 的奇异值以及 $\Delta \in (0,1)$ 是一个阈值. 于是, 我们得到了一个被截断后的低秩张量 $\mathrm{P}_{\leq\hat{\mathbf{r}}}^{\mathrm{HO}}(\mathcal{X}^{(t)})$. 为保证收敛性, 我们自适应的通过缩放因子 $\rho_1 \in (0,1)$ 缩小 Δ 到 $\rho_1\Delta$ 直到 $f(\mathcal{X}^{(t)}) \geq f(\mathrm{P}_{\leq\hat{\mathbf{r}}}^{\mathrm{HO}}(\mathcal{X}^{(t)}))$ 成立. 此时, 我们设定

$$\tilde{\mathcal{X}}^{(t+1)} = \mathrm{P}_{\leq\hat{\mathbf{r}}}^{\mathrm{HO}}(\mathcal{X}^{(t)}) \in \mathcal{M}_{\leq\mathbf{r}} \text{ 于是 } \mathrm{rank}_{\mathrm{tc}}(\tilde{\mathcal{X}}^{(t+1)}) \leq \underline{\mathbf{r}}^{(t)}.$$

图 4-5展示了当 $d = 3$ 时, 秩减机制如何将一个秩高张量截断为一个秩低的张量.

图 4-5 当 $d = 3$ 时秩减机制的示意图.

Figure 4-5 Illustration of rank-decreasing procedure for $d = 3$.

4.5.3　秩增机制

给定算法 14 根据初始值 $\tilde{\mathcal{X}}^{(t)}$ 迭代生成的点 $\mathcal{X}^{(t)} = \mathcal{G}^{(t)} \times_{k=1}^{d} \mathbf{U}_k^{(t)}$, 如果 $\mathcal{X}^{(t)}$ 是一个 $\varepsilon_R^{(t)}$-稳定点并且 $\underline{\mathbf{r}}^{(t)} < \mathbf{r}$, 为了提高精度得到一个更精确的解, 我们可以考虑增加秩参数. 正如第4.2.1节中的注记所述, 给定一个矩阵 $\mathbf{X} \in \mathbb{R}_{\underline{r}}^{m\times n}$, 增加一个 "法空间" $\mathrm{N}_{\leq\ell}(\mathbf{X})$ 中的元素可以提高 $\mathbf{X}$ 的秩. 对于 Tucker 张量而言, 我们类似的发现对所有的定义在式 (4-26)的 $\mathcal{N}_{\leq\ell^{(t)}}^{(t)} \in \mathrm{N}_{\leq\ell^{(t)}}(\mathcal{X}^{(t)})$ 以及 $0 < \ell^{(t)} \leq \mathbf{r} - \underline{\mathbf{r}}^{(t)}$, 有

$$\underline{\mathbf{r}}^{(t)} < \mathrm{rank}_{\mathrm{tc}}(\mathcal{X}^{(t)} + \mathcal{N}_{\leq\ell^{(t)}}^{(t)}) \leq \mathbf{r}$$

成立. 因此我们可以沿着满足 $\langle \mathcal{N}_{\leq\ell^{(t)}}^{(t)}, -\nabla f(\mathcal{X}^{(t)})\rangle \geq 0$ 的 $\mathcal{N}_{\leq\ell^{(t)}}^{(t)}$ 进行线搜索, 以达到增加 $\mathcal{X}^{(t)}$ 的秩的同时, 降低函数值的目的. 特别地, 对任意满足 $(\mathbf{U}_{k,1}^{(t)})^{\top}\mathbf{U}_k^{(t)} = 0$ 的 $\mathbf{U}_{k,1}^{(t)} \in \mathrm{St}(\ell_k^{(t)}, n_k)$, 搜索方向

$$\mathcal{N}_{\leq\ell^{(t)}}^{(t)} := -\nabla f(\mathcal{X}^{(t)}) \times_{k=1}^{d} \mathrm{P}_{\mathbf{U}_{k,1}^{(t)}} \in \mathrm{N}_{\leq\ell^{(t)}}(\mathcal{X}^{(t)})$$

总是一个下降方向. 为保证收敛性, 我们采用 Armijo 回溯线搜索即式 (4-19). 于是, 我们得到一个新的张量

$$\tilde{\mathcal{X}}^{(t+1)} = \mathcal{X}^{(t)} + s\mathcal{N}_{\leq\ell^{(t)}}^{(t)} \in \mathcal{M}_{\leq\mathbf{r}} \text{ 满足 } \mathrm{rank}_{\mathrm{tc}}(\tilde{\mathcal{X}}^{(t+1)}) > \underline{\mathbf{r}}^{(t)}.$$

从张量空间的角度来讲, 秩增机制将一个张量 $\mathcal{X}^{(t)}$ 更新到一个在更大张量空间中的张量 $\tilde{\mathcal{X}}^{(t+1)}$, 即

$$\bigotimes_{k=1}^{d} \operatorname{span}(\mathbf{U}_k^{(t)}) \longrightarrow \bigotimes_{k=1}^{d} \left(\operatorname{span}(\mathbf{U}_k^{(t)}) + \operatorname{span}(\mathbf{U}_{k,1}^{(t)}) \right).$$

当 $d = 3$ 时, 图 4-6给出了秩增机制一个几何上的展示. 算法 16总结了上述秩增机制的完整框架.

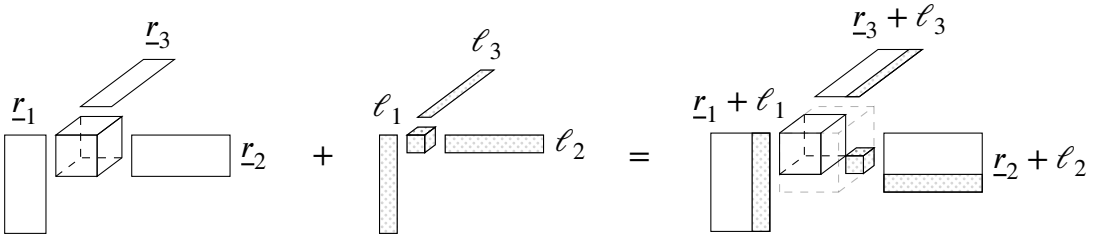


图 4-6 当 $d = 3$ 时秩增机制示意图

Figure 4-6 Illustration of rank-increasing procedure for $d = 3$.

算法 16 秩增机制

输入: $\mathcal{X}^{(t)} = \mathcal{G}^{(t)} \times_{k=1}^{d} \mathbf{U}_k^{(t)}, \ell^{(t)}$.

1: 随机选择满足 $\mathbf{U}_{k,1}^{(t)} \in \operatorname{St}(\ell_k^{(t)}, n_k)$ 和 $(\mathbf{U}_{k,1}^{(t)})^{\top}\mathbf{U}_k^{(t)} = 0$ 的一组基 $\mathbf{U}_{1,1}^{(t)}, \mathbf{U}_{2,1}^{(t)}, \ldots, \mathbf{U}_{d,1}^{(t)}$.

2: 计算 $\hat{\mathcal{G}}^{(t)} = -\nabla f(\mathcal{X}^{(t)}) \times_{k=1}^{d} (\mathbf{U}_{k,1}^{(t)})^{\top}$.

3: 通过 Armijo 回溯线搜索式 (4-19)选择步长 s.

4: 合并核张量 $\bar{\mathcal{G}}^{(t)} = \operatorname{diag}(\mathcal{G}^{(t)}, s\hat{\mathcal{G}}^{(t)})$, 以及因子矩阵 $\bar{\mathbf{U}}_k^{(t)} = \begin{bmatrix} \mathbf{U}_k^{(t)} & \mathbf{U}_{k,1}^{(t)} \end{bmatrix}$.

输出: 增秩后的张量 $\tilde{\mathcal{X}}^{(t+1)} = \bar{\mathcal{G}}^{(t)} \times_{k=1}^{d} \bar{\mathbf{U}}_k^{(t)}$, 此时 Tucker 秩为 $\underline{\mathbf{r}}^{(t)} + \ell^{(t)}$.

4.5.4 Tucker 秩自适应算法

我们结合上述的步骤, 针对问题 (4-1)提出 Tucker 秩自适应算法 (Tucker rank-adaptive method, TRAM), 算法的框架详见算法 17.

我们提出的秩自适应方法在每一步中首先以 $\tilde{\mathcal{X}}^{(t)}$ 作为初始点运行算法 14, 并返回结果 $\mathcal{X}^{(t)}$. 根据 $\mathcal{X}^{(t)}$ 的不同性质进行相应的秩调整. 若发现 $\mathcal{X}^{(t)}$ 是秩亏的, 则触发算法 15 中的秩减机制, 以避免潜在的秩退化问题. 否则, 可以考虑通过秩增机制来提升精度. 为此, 我们首先检验条件

$$\|\mathcal{N}_{\leq \ell^{(t)}}^{(t)}\|_{\mathrm{F}} \geq \varepsilon_1 \|\mathcal{T}^{(t)}\|_{\mathrm{F}} \quad \text{和} \quad \varepsilon_2 \|\nabla f(\mathcal{X}^{(t)})\|_{\mathrm{F}} \leq \|\mathcal{T}^{(t)}\|_{\mathrm{F}}, \tag{4-27}$$

其中 $\mathcal{T}^{(t)} := \operatorname{grad} f(\mathcal{X}^{(t)})$. 若迭代点满足该条件, 则表明此时进行秩增是有效的. 于是, 我们采用算法 16 实现秩增. 否则, 结合 (4-5) 及图 4-3, 可以观察到

$$\mathrm{T}_{\mathcal{X}^{(t)}} \mathcal{M}_{\underline{\mathbf{r}}^{(t)}} + \mathrm{N}_{\leq \ell^{(t)}}(\mathcal{X}^{(t)}) \subsetneq \mathrm{T}_{\mathcal{X}^{(t)}} \mathcal{M}_{\leq \mathbf{r}}.$$

算法 17 求解问题 (4-1) 的 Tucker 秩自适应算法 (TRAM)

输入: 初始值 $\mathcal{X}^{(0)} = \tilde{\mathcal{X}}^{(0)} \in \mathcal{M}_{\leq \mathbf{r}}$ 以及 $\mathrm{rank}_{\mathrm{tc}}(\mathcal{X}^{(0)}) = \mathrm{rank}_{\mathrm{tc}}(\tilde{\mathcal{X}}^{(0)}) = \underline{\mathbf{r}}^{(0)}$; 参数 $\varepsilon_R^{(0)} > 0$, $\rho_R \in (0,1)$; 秩减参数 $\Delta > 0$, $\rho_1 \in (0,1)$; 秩增参数 $\{\ell^{(t)}\}_{t\geq 0}$; 线搜索参数 $\rho, a \in (0,1)$, $s_{\min} > 0$.

while 停机准则未被满足 **do**
　输入迭代点 $\tilde{\mathcal{X}}^{(t)}$ 和阈值 $\varepsilon_R^{(t)}$ 执行算法 14 得到点 $\mathcal{X}^{(t)}$, 并得到黎曼梯度 $\mathcal{T}^{(t)} = \mathrm{grad} f(\mathcal{X}^{(t)})$.
　if 算法 14停止原因为检测到秩亏 **then**
　　触发秩减机制, 利用算法 15将点 $\mathcal{X}^{(t)}$ 更新为 $\tilde{\mathcal{X}}^{(t+1)}$.
　　if $\mathrm{rank}_{\mathrm{tc}}(\tilde{\mathcal{X}}^{(t+1)}) = \underline{\mathbf{r}}^{(t)}$ **then**
　　算法终止.
　　end if
　else ▷ 未检测到秩亏
　　if $\underline{\mathbf{r}}^{(t)} = \mathbf{r}$ **then** ▷ 当前迭代点满秩
　　令 $\tilde{\mathcal{X}}^{(t+1)} = \mathcal{X}^{(t)}$ 和 $\varepsilon_R^{(t+1)} = \rho_R \varepsilon_R^{(t)}$.
　　else ▷ 当前迭代点不满秩, 秩增可行
　　通过算法 16的第 1–2 行计算 $\mathcal{N}^{(t)}_{\leq \ell^{(t)}} = \hat{\mathcal{G}}^{(t)} \times_{k=1}^{d} \mathbf{U}^{(t)}_{k,1}$.
　　if $\|\mathcal{N}^{(t)}_{\leq \ell^{(t)}}\|_{\mathrm{F}} \geq \varepsilon_1 \|\mathcal{T}^{(t)}\|_{\mathrm{F}}$ and $\varepsilon_2 \|\nabla f(\mathcal{X}^{(t)})\|_{\mathrm{F}} \leq \|\mathcal{T}^{(t)}\|_{\mathrm{F}}$ **then**
　　应用秩增机制 (即算法 16) 并得到 $\tilde{\mathcal{X}}^{(t+1)}$.
　　else if $\varepsilon_2 \|\nabla f(\mathcal{X}^{(t)})\|_{\mathrm{F}} > \|\mathcal{T}^{(t)}\|_{\mathrm{F}}$ **then**
　　通过算法12的第2–4行更新 $\tilde{\mathcal{X}}^{(t+1)} = \mathrm{P}^{\mathrm{HO}}_{\leq \mathbf{r}}(\mathcal{X}^{(t)} + s^{(t)} \tilde{\mathrm{P}}_{\mathrm{T}_{\mathcal{X}^{(t)}}\mathcal{M}_{\leq \mathbf{r}}}(-\nabla f(\mathcal{X}^{(t)})))$. ▷ 重启动
　　else
　　令 $\tilde{\mathcal{X}}^{(t+1)} = \mathcal{X}^{(t)}$ 和 $\varepsilon_R^{(t+1)} = \rho_R \varepsilon_R^{(t)}$.
　　end if
　　end if
　end if
　$t = t + 1$.
end while

输出: $\mathcal{X}^{(t)}$

因此, 我们进一步检验如下重启判据:

$$\varepsilon_2\|\nabla f(\mathcal{X}^{(t)})\|_{\mathrm{F}} \geq \|\mathcal{T}^{(t)}\|_{\mathrm{F}}. \tag{4-28}$$

若该条件成立, 我们执行算法 12 中第 2–4 行, 沿方向 $\tilde{\mathrm{P}}_{\mathrm{T}_{\mathcal{X}^{(t)}}\mathcal{M}_{\leq \underline{\mathbf{r}}^{(t)}}}(-\nabla f(\mathcal{X}^{(t)}))$ 进行线搜索, 这相当于一步重启动.

注. 在实际应用中, 由于搜索空间的扩大, 增加秩总是能够改善近似误差 (参见 [60, §4.9]). 然而, 若在固定 (且较大) 的秩参数 $\mathbf{r}$ 下直接应用 GRAP、rfGRAP 或黎曼共轭梯度方法, 则可能导致严重的过拟合问题 (参见第 4.6.2 节及 [15, §4.3]). 所提出的秩增加策略仅在搜索方向 $\mathcal{N}^{(t)}_{\leq \ell^{(t)}}$ 在 (4-27) 的意义下占主导时才会增加秩. 此外, 该秩增加过程具备理论保证, 详见定理 4.13.

4.5.5 收敛性结果

设 $\{\mathcal{X}^{(t)}\}_{t\geq 0}$ 为算法 17 生成的无穷序列. 注意到从算法 14–15中我们可以得到点 $\mathcal{X}^{(t+1)}$ 满足

$$f(\mathcal{X}^{(t+1)}) \leq f(\tilde{\mathcal{X}}^{(t+1)}) \leq f(\mathcal{X}^{(t)}),$$

也就是说, $\{f(\mathcal{X}^{(t)})\}_{t\geq 0}$ 单调不增. 接下来, 我们证明 TRAM 方法的全局收敛性.

引理 4.12. *设 $\{\mathcal{X}^{(t)}\}_{t\geq 0}$ 为算法 17 生成的无穷序列. 假设函数 f 在下界 f^* 上有界, 则有*

$$\liminf_{t\to\infty} \|\mathrm{grad} f(\mathcal{X}^{(t)})\|_{\mathrm{F}} = \liminf_{t\to\infty} \|\mathrm{P}_{\mathrm{T}_{\mathcal{X}^{(t)}}\mathcal{M}_{\underline{\mathbf{r}}^{(t)}}}(\nabla f(\mathcal{X}^{(t)}))\|_{\mathrm{F}} = 0.$$

证明. 设 $\{\tilde{\mathcal{X}}^{(t)}\}_{t\geq 0}$ 是由算法 17生成的无限序列. 首先, 我们证明, 存在无限个 t 满足 $\|\mathcal{T}^{(t)}\|_{\mathrm{F}} \leq \varepsilon_R^{(t)}$, 这里我们回顾 $\mathcal{T}^{(t)} = \mathrm{grad} f(\mathcal{X}^{(t)})$. 否则, 我们根据算法 14 的停机准则, 对于充分大的 t, 算法总会产生一个秩亏点. 因此, 秩减机制在每一步都会被触发, 这与一个张量的 Tucker 秩有限产生矛盾. 此外, 我们假设第 10 和 18 行的参数更新被执行有限次. 否则, $\varepsilon_R^{(t+1)} = \rho_R \varepsilon_R^{(t)}$ 会被执行无限次 $\lim_{t\to\infty} \varepsilon_R^{(t)} = 0$, 这也就意味着 $\liminf_{t\to\infty} \|\mathrm{P}_{\mathrm{T}_{\mathcal{X}^{(t)}}\mathcal{M}_{\underline{\mathbf{r}}^{(t)}}}(-\nabla f(\mathcal{X}^{(t)}))\|_{\mathrm{F}} = 0$.

于是, 存在一个满足 $\|\mathcal{T}^{(t_j)}\|_{\mathrm{F}} \leq \varepsilon_R^{(t_j)}$ 的子列 $\{\mathcal{X}^{(t_j)}\}_{j\geq 0}$. 我们的目标是证明 $\|\mathcal{T}^{(t_j)}\|_{\mathrm{F}}$ 收敛到 0. 由于 f 是下有界的, 我们有

$$0 \leq \lim_{j\to\infty} f(\mathcal{X}^{(t_j)}) - f(\tilde{\mathcal{X}}^{(t_j+1)}) \leq \lim_{j\to\infty} f(\mathcal{X}^{(t_j)}) - f(\mathcal{X}^{(t_j+1)}) = 0.$$

接下来, 对于足够大的 j, 我们分成秩增机制与重启机制两类情况讨论点 $\mathcal{X}^{(t_j)}$ 的更新: 1) 如果第 14行的秩增机制被触发, 根据算法 16中的回溯线搜索以及式 (4-27) 得到

$$\begin{aligned} f(\mathcal{X}^{(t_j)}) - f(\tilde{\mathcal{X}}^{(t_j+1)}) &\geq s_{\min} a \left\langle \mathcal{N}^{(t_j)}_{\leq \ell^{(t_j)}}, -\nabla f(\mathcal{X}^{(t_j)}) \right\rangle \\ &= s_{\min} a \|\mathcal{N}^{(t_j)}_{\leq \ell^{(t_j)}}\|_{\mathrm{F}}^2 \\ &\geq s_{\min} a \varepsilon_1^2 \|\mathcal{T}^{(t_j)}\|_{\mathrm{F}}^2; \end{aligned}$$

2) 如果第 16行的重启机制被触发, 则搜索方向 $\tilde{\mathrm{P}}_{\mathrm{T}_{\mathcal{X}^{(t_j)}}\mathcal{M}_{\leq \mathbf{r}}}(-\nabla f(\mathcal{X}^{(t_j)}))$ 会被采用. 于是我们有

$$f(\mathcal{X}^{(t_j)}) - f(\tilde{\mathcal{X}}^{(t_j+1)}) \geq s_{\min} a \,\|\tilde{\mathrm{P}}_{\mathrm{T}_{\mathcal{X}^{(t_j)}}\mathcal{M}_{\leq \mathbf{r}}}(-\nabla f(\mathcal{X}^{(t_j)}))\|_{\mathrm{F}}^2 \geq s_{\min} a \,\|\mathcal{T}^{(t_j)}\|_{\mathrm{F}}^2.$$

这里最后一个不等号来源于 $\mathrm{T}_{\mathcal{X}^{(t_j)}}\mathcal{M}_{\underline{\mathbf{r}}^{(t_j)}} \subseteq \mathrm{T}_{\mathcal{X}^{(t_j)}}\mathcal{M}_{\leq \mathbf{r}}$.

总的来说, 对于充分大的 j 我们有

$$\|\mathcal{T}^{(t_j)}\|_{\mathrm{F}}^2 \leq \frac{f(\mathcal{X}^{(t_j)}) - f(\tilde{\mathcal{X}}^{(t_j+1)})}{s_{\min} a \min\{\varepsilon_1^2, 1\}}$$

于是它收敛到 0. □

通过使用引理 4.12, 我们可以证明如下的更强的结论.

定理 4.13. *设 $\{\mathcal{X}^{(t)}\}_{t\geq 0}$ 为算法 17 生成的无穷序列. 假设函数 f 在下界 f^* 上有界, 则有*

$$\liminf_{t\to\infty} \|\mathrm{P}_{\mathrm{T}_{\mathcal{X}^{(t)}}\mathcal{M}_{\leq \mathbf{r}}}(-\nabla f(\mathcal{X}^{(t)}))\|_{\mathrm{F}} = 0.$$

证明. 设 $\{\tilde{\mathcal{X}}^{(t)}\}_{t\geq 0}$ 是由算法 17 生成的一个无穷序列. 回顾 $\mathcal{T}^{(t)} = \mathrm{grad} f(\mathcal{X}^{(t)})$ 的定义. 由引理 4.12 可知, 存在一个子序列 $\{\mathcal{X}^{(t_j)}\}_{j\geq 0}$, 使得 $\|\mathcal{T}^{(t_j)}\|_{\mathrm{F}} \leq \varepsilon_R^{(t_j)}$ 且 $\lim_{j\to\infty} \|\mathcal{T}^{(t_j)}\|_{\mathrm{F}} = 0$. 进一步假设对于所有 $j \geq 0$ 及某个常数 $\varepsilon_0 > 0$, 均有 $\|\nabla f(\mathcal{X}^{(t_j)})\|_{\mathrm{F}} \geq \varepsilon_0$, 否则结论是显然成立的.

若算法 17 中第 10 行被执行无限次, 则存在 $\{\mathcal{X}^{(t_j)}\}_{j\geq 0}$ 的一个子列 $\{\mathcal{X}^{(t_{j_l})}\}_{l\geq 0}$, 使得 $\mathrm{rank}_{\mathrm{tc}}(\mathcal{X}^{(t_{j_l})}) = \mathbf{r}$. 因此,

$$\lim_{l\to\infty} \|\mathrm{P}_{\mathrm{T}_{\mathcal{X}^{(t_{j_l})}}\mathcal{M}_{\leq \mathbf{r}}}(-\nabla f(\mathcal{X}^{(t_{j_l})}))\|_{\mathrm{F}} = \lim_{l\to\infty} \|\mathcal{T}^{(t_{j_l})}\|_{\mathrm{F}} = 0.$$

否则, 由于 $\|\mathcal{T}^{(t_j)}\|_{\mathrm{F}}$ 收敛至 0 且 $\|\nabla f(\mathcal{X}^{(t_j)})\|_{\mathrm{F}} \geq \varepsilon_0$, 由条件 (4-28) 可知, 当 j 足够大时, 算法 17 中第 16 行的重启动将被持续执行. 进一步结合第 16 行中的回溯线搜索, 可得

$$\begin{aligned} f(\mathcal{X}^{(t_j)}) - f(\tilde{\mathcal{X}}^{(t_j+1)}) &\geq s_{\min} a \,\|\tilde{\mathrm{P}}_{\mathrm{T}_{\mathcal{X}^{(t_j)}}\mathcal{M}_{\leq \mathbf{r}}}(-\nabla f(\mathcal{X}^{(t_j)}))\|_{\mathrm{F}}^2 \\ &\geq s_{\min} a\, \omega^2 \|\mathrm{P}_{\mathrm{T}_{\mathcal{X}^{(t_j)}}\mathcal{M}_{\leq \mathbf{r}}}(-\nabla f(\mathcal{X}^{(t_j)}))\|_{\mathrm{F}}^2. \end{aligned}$$

最终我们得到,

$$\lim_{j\to\infty} \|\mathrm{P}_{\mathrm{T}_{\mathcal{X}^{(t_j)}}\mathcal{M}_{\leq \mathbf{r}}}(-\nabla f(\mathcal{X}^{(t_j)}))\|_{\mathrm{F}} = 0.$$

□

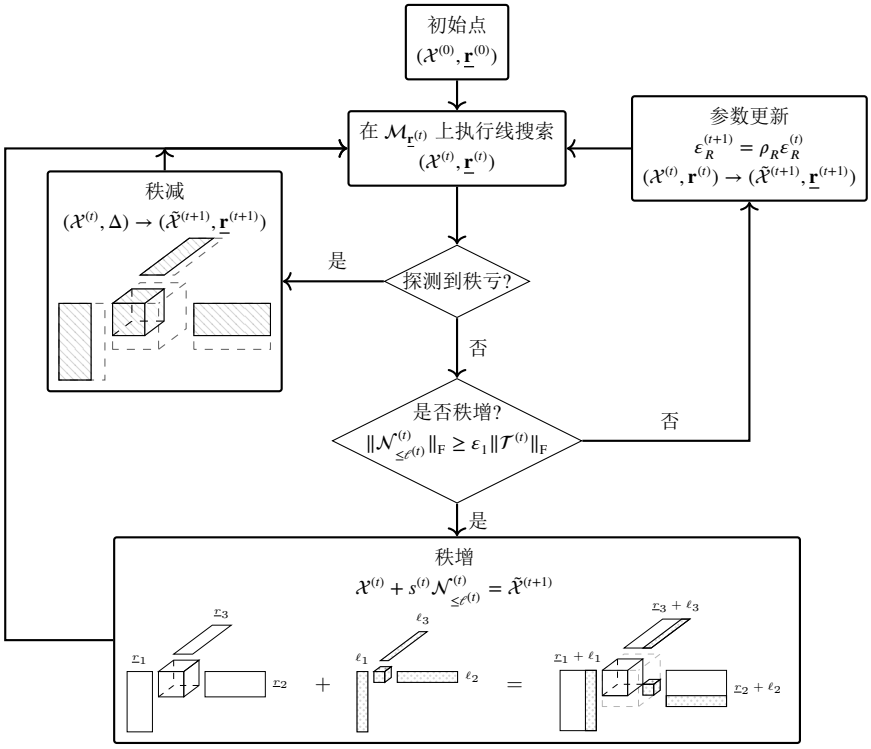


图 4-7 TRAM 方法在实际计算中的流程图.

Figure 4-7 A flowchart of the practical Tucker rank-adaptive method.

4.5.6 TRAM 方法的计算细节

在实际计算当中, TRAM 方法按照图 4-7的流程图执行. 对于秩增机制, 我们发现算法 17当中的重启动机制会带来较大的开销, 这是因为重启动机制会大概率将迭代点的秩直接增加到 $\mathbf{r}$. 当 $\mathcal{N}^{(t)}_{\leq \ell^{(t)}}$ 被判定为不满足条件 (4-27) 时, 上述情形就会出现. 此时, 我们通过设置 $\varepsilon_R^{(t+1)} = \rho_R \varepsilon_R^{(t)}$ 来收紧停机准则, 并再次运行 RGD 方法. 对于秩减机制, 设 $\{\mathcal{Y}^{(i)}\}_{i\geq 0}$ 为由 RGD 生成的序列, 且初始点为 $\mathcal{Y}^{(0)} = \tilde{\mathcal{X}}^{(t)}$. 在算法 14的实际代码实现中, 每一个点 $\mathcal{Y}^{(i)} = \mathcal{G}^{(i)} \times_{k=1}^{d} \mathbf{U}_k^{(i)} \in \mathcal{M}_{\underline{\mathbf{r}}^{(t)}}$ 都会计算比值 $\sigma_{\min}(\mathbf{Y}^{(i)}_{(k)})/\sigma_{\max}(\mathbf{Y}^{(i)}_{(k)})$ 以检测是否出现秩亏. 我们注意到

$$\mathbf{Y}^{(i)}_{(k)} = \mathbf{U}_k^{(i)} \mathbf{G}^{(i)}_{(k)} (\mathbf{V}_k^{(i)})^{\mathsf{T}} = \mathbf{U}_k^{(i)} \check{\mathbf{U}}_k^{(i)} \check{\Sigma} (\check{\mathbf{V}}_k^{(i)})^{\mathsf{T}} (\mathbf{V}_k^{(i)})^{\mathsf{T}}$$

构成了 $\mathbf{Y}^{(i)}_{(k)}$ 的一个奇异值分解, 其中 $\check{\mathbf{U}}_k^{(i)} \check{\Sigma} (\check{\mathbf{V}}_k^{(i)})^{\mathsf{T}}$ 是 $\mathbf{G}^{(i)}_{(k)}$ 的奇异值分解. 因此, 有

$$\frac{\sigma_{\min}(\mathbf{Y}^{(i)}_{(k)})}{\sigma_{\max}(\mathbf{Y}^{(i)}_{(k)})} = \frac{\sigma_{\min}(\mathbf{G}^{(i)}_{(k)})}{\sigma_{\max}(\mathbf{G}^{(i)}_{(k)})}.$$

利用这一性质, 我们可以通过仅使用尺寸较小的核张量 $\mathcal{G}^{(i)}$ 完成秩亏的检测, 从而避免显式构造大规模的 $\mathcal{Y}^{(i)}$. 此外, 在算法 14 中, 条件 $f(\mathcal{X}^{(t)}) \geq f(\mathrm{P}^{\mathrm{HO}}_{\leq \hat{\mathbf{r}}}(\mathcal{X}^{(t)}))$ 将不会被检查, 因此秩确实会被降低.

4.6　数值实验

在本节中, 我们测试所提出的 GRAP(算法 12)、rfGRAP(算法 13)、TRAM(算法 17) 以及其他已有方法在张量补全问题上的性能. 具体而言, 给定一个在索引集 $\Omega \subseteq [n_1] \times [n_2] \times \cdots \times [n_d]$ 上部分观测到的张量 $\mathcal{A} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$, Tucker 张量补全的目标是基于低秩 Tucker 分解, 从其在 Ω 上的观测条目中恢复完整张量 $\mathcal{A}$. 相应的优化问题可以在 Tucker 张量代数簇 $\mathcal{M}_{\leq \mathbf{r}}$ 上表述为

$$\begin{aligned} \min \quad & \frac{1}{2} \| \mathrm{P}_\Omega(\mathcal{X}) - \mathrm{P}_\Omega(\mathcal{A}) \|_{\mathrm{F}}^2 \\ \text{s.t.} \quad & \mathcal{X} \in \mathcal{M}_{\leq \mathbf{r}}, \end{aligned}$$

其中 P_Ω 表示投影算子到 Ω 上, 即对任意 $\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$, $\mathrm{P}_\Omega(\mathcal{X})(i_1, i_2, \ldots, i_d) = \mathcal{X}(i_1, i_2, \ldots, i_d)$ 当 $(i_1, i_2, \ldots, i_d) \in \Omega$ 时成立, 否则 $\mathrm{P}_\Omega(\mathcal{X})(i_1, i_2, \ldots, i_d) = 0$. 采样率定义为 $p := |\Omega|/(n_1 n_2 \cdots n_d)$.

4.6.1　算法实现细节

首先, 我们介绍所有的默认设置与实现细节. 总体而言, 所提出方法中与张量相关的实现基于 Tensor-Toolbox v3.4[1] 工具包. 所有实验均在一台工作站上完成, 该工作站配备两颗 Intel(R) Xeon(R) Gold 6330 处理器 (2.00GHz×28, 42M Cache), 512GB 内存, 操作系统为 Ubuntu 22.04.3, 运行 Matlab R2019b. 我们提出方法的代码可在 https://github.com/JimmyPeng1998/TRAM网站上获取.

计算投影　给定张量 $\mathcal{T} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$ 以及 $\mathcal{X} = \mathcal{G} \times_{k=1}^{d} \mathbf{U}_k$, 且 $\mathrm{rank}_{\mathrm{tc}}(\mathcal{X}) = \underline{\mathbf{r}}$, 所提出的方法涉及到对切锥 $\mathrm{T}_{\mathcal{X}} \mathcal{M}_{\leq \mathbf{r}}$ 以及 Tucker 张量簇 $\mathcal{M}_{\mathbf{r}}$ 的投影. 我们给出两种投影 $\tilde{\mathrm{P}}_{\mathrm{T}_{\mathcal{X}} \mathcal{M}_{\leq \mathbf{r}}}(\mathcal{T})$ 与 $\mathrm{P}^{\mathrm{HO}}_{\leq \mathbf{r}}(\mathcal{T})$ 的计算细节. 在实际实现中, 我们从不在 $\tilde{\mathrm{P}}_{\mathrm{T}_{\mathcal{X}} \mathcal{M}_{\leq \mathbf{r}}}(\mathcal{T})$ 和 $\mathrm{P}^{\mathrm{HO}}_{\leq \mathbf{r}}(\mathcal{T})$ 中直接操作具有 $n_1 n_2 \cdots n_d$ 个参数的大规模完整张量 $\mathcal{X}$, 而是仅操作核心张量及其展开矩阵.

在 (4-16) 中, 近似投影 $\tilde{\mathrm{P}}_{\mathrm{T}_{\mathcal{X}} \mathcal{M}_{\leq \mathbf{r}}}(\mathcal{T})$ 需要为 $k \in [d]$ 选择满足 $\tilde{\mathbf{U}}_{k,1}^{\mathsf{T}} \mathbf{U}_k = 0$ 的矩阵 $\tilde{\mathbf{U}}_{k,1} \in \mathrm{St}(r_k - \underline{r}_k, n_k)$. 我们生成一个随机矩阵 $\mathbf{M}_{k,1} \in \mathbb{R}^{n_k \times (r_k - \underline{r}_k)}$, 其元素服从正态分布 $N(0,1)$ 的独立同分布采样, 并将 $\tilde{\mathbf{U}}_{k,1}$ 选为矩阵 $[\mathbf{U}_k \ \mathbf{M}_{k,1}] \in \mathbb{R}^{n_k \times r_k}$ 的 Q 因子的最后 $(r_k - \underline{r}_k)$ 列. 随后, 可以通过算法 11 计算到 $\mathrm{T}_{\mathcal{X}} \mathcal{M}_{\leq \mathbf{r}}$ 的近似投影. 该选择 $\tilde{\mathbf{U}}_{k,1}$ 的方式同样被用于算法 16 中 $\{\mathbf{U}_{k,1}\}_{k=1}^{d}$ 的选取. 由于角度条件在计算上难以处理, 我们在实际中不对角度条件进行验证. 此外, 切空间上的正交投影由 GeomCG 工具箱[2] 计算. 需要注意的是, 当 GRAP 方法中 $\mathrm{rank}_{\mathrm{tc}}(\mathcal{X}^{(t)}) = \mathbf{r}$ 时, 为了公平比较, 切锥上的投影同样由 GeomCG 工具箱计算, 这是因为 $\mathrm{T}_{\mathcal{X}^{(t)}} \mathcal{M}_{\leq \mathbf{r}} = \mathrm{T}_{\mathcal{X}^{(t)}} \mathcal{M}_{\mathbf{r}}$.

[1] Tensor-Toolbox v3.4: http://www.tensortoolbox.org/

[2] GeomCG toolbox: https://www.epfl.ch/labs/anchp/index-html/software/geomcg/.

对于 (1-12) 中的投影 $\mathrm{P}^{\mathrm{HO}}_{\leq \mathbf{r}}(\mathcal{T})$, 结合算法 12, 我们设 $\mathcal{T}$ 具有形式 $\mathcal{T} = \mathcal{X} + \mathcal{V}$, 其中 $\mathcal{V} \in \mathrm{T}_{\mathcal{X}}\mathcal{M}_{\leq \mathbf{r}}$. 由 (4-5) 可知,

$$\mathcal{T} = \mathcal{X} + \mathcal{V} \in \bigotimes_{k=1}^{d} \left(\operatorname{span}(\mathbf{U}_k) + \operatorname{span}(\mathbf{U}_{k,1}) + \operatorname{span}(\mathbf{U}_{k,2}\mathbf{R}_{k,2})\right) \subseteq \mathcal{M}_{\leq(\mathbf{r}+\underline{\mathbf{r}})}.$$

因此, 张量 $\mathcal{T}$ 可以表示为 Tucker 分解 $\mathcal{T} = \tilde{\mathcal{G}} \times_{k=1}^{d} \tilde{\mathbf{U}}_k \in \mathcal{M}_{\tilde{\mathbf{r}}}$, 其中 $\tilde{\mathbf{r}} \leq \mathbf{r} + \underline{\mathbf{r}}$. 我们不直接对完整张量 $\mathcal{T} \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ 执行 HOSVD, 而是利用其低秩结构, 仅对 $\mathcal{T}$ 的核张量 $\tilde{\mathcal{G}} \in \mathbb{R}^{\tilde{r}_1 \times \cdots \times \tilde{r}_d}$ 执行 HOSVD, 其规模要小得多. 具体而言, 设 $\tilde{\mathcal{G}}$ 的秩-$\mathbf{r}$ HOSVD 为 $(\tilde{\mathcal{G}} \times_{k=1}^{d} \hat{\mathbf{U}}_k^{\top}) \times_{k=1}^{d} \hat{\mathbf{U}}_k$, 其中 $\hat{\mathbf{U}}_k \in \mathrm{St}(r_k, \tilde{r}_k)$ 为 $\tilde{\mathbf{G}}_{(k)}$ 的前 r_k 个奇异向量. 于是有

$$\mathrm{P}^{\mathrm{HO}}_{\leq \mathbf{r}}(\mathcal{T}) = ((\tilde{\mathcal{G}} \times_{k=1}^{d} \hat{\mathbf{U}}_k^{\top}) \times_{k=1}^{d} \hat{\mathbf{U}}_k) \times_{k=1}^{d} \tilde{\mathbf{U}}_k = (\tilde{\mathcal{G}} \times_{k=1}^{d} \hat{\mathbf{U}}_k^{\top}) \times_{k=1}^{d} (\tilde{\mathbf{U}}_k \hat{\mathbf{U}}_k).$$

注意到 $\tilde{\mathbf{U}}_k \hat{\mathbf{U}}_k \in \mathrm{St}(r_k, n_k)$, 因为 $(\tilde{\mathbf{U}}_k \hat{\mathbf{U}}_k)^{\top}(\tilde{\mathbf{U}}_k \hat{\mathbf{U}}_k) = \mathbf{I}_{r_k}$. 该技巧同样被用于秩减机制以及在 $\mathcal{M}_{\underline{\mathbf{r}}}$ 上的黎曼梯度法 (算法 14) 中的收缩映射.

在切锥上的精确线搜索 类似于在固定秩 Tucker 张量流形上的优化 [15], 给定一点 $\mathcal{X}^{(t)}$ 及下降方向 $\mathcal{V}^{(t)} \in \mathrm{T}_{\mathcal{X}^{(t)}}\mathcal{M}_{\leq \mathbf{r}}$, 优化问题

$$s_0^{(t)} = \arg\min_{s \geq 0} \| \mathrm{P}_{\Omega}(\mathcal{X}^{(t)} + s\mathcal{V}^{(t)}) - \mathrm{P}_{\Omega}\mathcal{A} \|_{\mathrm{F}}^2$$

的解具有显式表达

$$s_0^{(t)} = \frac{\langle \mathrm{P}_{\Omega}\mathcal{V}^{(t)}, \mathrm{P}_{\Omega}(\mathcal{A} - \mathcal{X}^{(t)}) \rangle}{\langle \mathrm{P}_{\Omega}\mathcal{V}^{(t)}, \mathrm{P}_{\Omega}\mathcal{V}^{(t)} \rangle} \geq 0.$$

$\mathrm{P}_{\Omega}\mathcal{V}^{(t)}$ 的计算通过 MEX 函数实现. 我们将 $s_0^{(t)}$ 作为 (4-19) 中 Armijo 回溯线搜索的初始步长.

对比的已有方法 对于基于 Tucker 分解的方法, 我们将所提出的方法与固定秩流形上的黎曼共轭梯度法 (GeomCG) [15] 以及在预条件度量下的商流形黎曼共轭梯度法 (Tucker-RCG) [96] 进行比较[3].

此外, 我们还将所提出的方法与基于其他张量分解的方法进行比较. 对于 CP 分解, 我们选择 Guan 等人 [156] 提出的基于图的交替最小化方法[4], 记为 CP-AltMin. 对于张量链分解补全问题, 我们考虑文献 [60] 中的黎曼共轭梯度法 (TT-RCG)[5]. 对于张量环分解下的张量补全, 我们采用在预条件度量下的黎曼梯度下降方法 (TR-RGD) [17][6].

[3]代码可见 https://bamdevmishra.in/codes/tensorcompletion/.

[4]代码可见: https://gitlab.com/ricky7guanyu/tensor-completion-with-regularization-term.

[5]TTeMPS 工具箱: https://www.epfl.ch/labs/anchp/index-html/software/ttemps/.

[6]LRTCTR 工具箱: https://github.com/JimmyPeng1998/LRTCTR

停机准则　所有方法的性能均通过训练误差与测试误差进行评估:

$$\varepsilon_{\Omega}(\mathcal{X}) := \frac{\|\mathrm{P}_{\Omega}(\mathcal{X}) - \mathrm{P}_{\Omega}(\mathcal{A})\|_{\mathrm{F}}}{\|\mathrm{P}_{\Omega}(\mathcal{A})\|_{\mathrm{F}}} \quad \text{和} \quad \varepsilon_{\Gamma}(\mathcal{X}) := \frac{\|\mathrm{P}_{\Gamma}(\mathcal{X}) - \mathrm{P}_{\Gamma}(\mathcal{A})\|_{\mathrm{F}}}{\|\mathrm{P}_{\Gamma}(\mathcal{A})\|_{\mathrm{F}}},$$

其中 Γ 为与训练集 Ω 不同的测试集. 当满足以下任一条件时, 算法终止: 1) 训练误差 $\varepsilon_{\Omega}(\mathcal{X}^{(t)}) < 10^{-12}$; 2) 训练误差的相对变化 $(\varepsilon_{\Omega}(\mathcal{X}^{(t)}) - \varepsilon_{\Omega}(\mathcal{X}^{(t-1)}))/\varepsilon_{\Omega}(\mathcal{X}^{(t-1)}) < 10^{-8}$; 3) 达到最大迭代次数; 4) 超出时间预算.

所提出方法参数的默认设置　所提出方法的默认参数设置如下: 在 TRAM 中, 取 $\rho_R = 0.5$、$\varepsilon_R^{(0)} = 0.1$, 秩减参数 $\Delta = 0.01$ 与 $\rho_1 = 0.5$, 以及秩增参数 $\ell = (1, 1, \ldots, 1)$ 与 $\varepsilon_1 = 0.01$. 回溯线搜索参数设置为 $\rho = 0.5$、$a = 10^{-4}$ 和 $s_{\min} = 10^{-10}$. 此外, TRAM 方法中固定秩线搜索的最大迭代次数为 5.

4.6.2　在人工数据集上的实验

我们在合成数据上测试基于 Tucker 的方法的恢复性能. 给定真实秩 $\mathbf{r}^* = (r_1^*, r_2^*, \ldots, r_d^*)$, 我们考虑如下方式生成的低秩张量 $\mathcal{A}$:

$$\mathcal{A} = \mathcal{G}^* \times_{k=1}^{d} \mathbf{U}_k^*,$$

其中, $\mathcal{G}^* \in \mathbb{R}^{r_1^* \times r_2^* \times \cdots \times r_d^*}$ 和 $\mathbf{U}_k^* \in \mathbb{R}^{n_k \times r_k^*}$ 的元素均独立采样自正态分布 $N(0, 1)$. 随后, 我们通过 QR 分解正交化 $\mathbf{U}_k^*$. 我们设 $d = 3$, $n_1 = n_2 = n_3 = 400$, 测试集大小为 $|\Gamma| = pn_1n_2n_3$, 并取 $r_1^* = r_2^* = r_3^* = 6$. 初始点 $\mathcal{X}^{(0)}$ 以相同方式生成, 但其秩为给定的 $\underline{\mathbf{r}}^{(0)}$. 当训练误差 $\varepsilon_{\Omega}(\mathcal{X}^{(t)}) \leq 10^{-12}$ 或运行时间超过 200s 时, 算法终止.

真实秩条件下的测试　首先, 我们考察在真实秩条件下 (即 $\mathbf{r} = \mathbf{r}^* = (6, 6, 6)$) 各类基于 Tucker 分解方法的性能. 为保证公平比较, 我们将所提出的方法与 GeomCG 以及 Tucker-RCG 进行对比, 并统一采用初始点 $\mathcal{X}^{(0)} \in \mathcal{M}_{\mathbf{r}}$. 图 4-8 给出了在采样率 $p = 0.01, 0.05$ 下, 各种基于 Tucker 分解方法的测试误差. 首先可以观察到, GRAP 与 TRAM 方法在性能上与 GeomCG 和 Tucker-RCG 相当. 其次, rfGRAP 方法所需的迭代次数多于其他方法, 这是因为其仅利用了切锥的部分信息以避免收缩映射.

在低秩初始点下的测试　不同于运行在 $\mathcal{M}_{\mathbf{r}}$ 上的黎曼方法, 所提出的方法可以接受任意初始点 $\mathcal{X}^{(0)} \in \mathcal{M}_{\leq \mathbf{r}}$. 因此, 我们在不同初始秩 $\mathbf{r}^{(0)} = (r^{(0)}, r^{(0)}, r^{(0)})$ 下比较所提出的方法, 其中 $r^{(0)} = 1$ 或 5. 采样率取 $p = 0.05$. 需要注意的是, GeomCG 和 Tucker-RCG 仍然在 $\mathcal{M}_{\mathbf{r}^{(0)}}$ 上运行. 测试误差如图 4-9 所示. 由图 4-9 可见, 所提出的 GRAP 和 rfGRAP 方法相较于 TRAM 具有更好的性能. 这是因为 TRAM 方法需要通过秩增机制来寻找真实秩 $\mathbf{r}^*$. 此外, 所提出的 TRAM 方法能够成功识别真实秩 $\mathbf{r}^*$. 然而, 由于 $\mathbf{r}^{(0)} < \mathbf{r}^*$, GeomCG 与 Tucker-RCG 只能获得对数据张量 $\mathcal{A}$ 的较差低秩近似. 因此, 秩增过程确实使我们能够在更大的搜索空间中进行优化, 从而获得更高的精度.

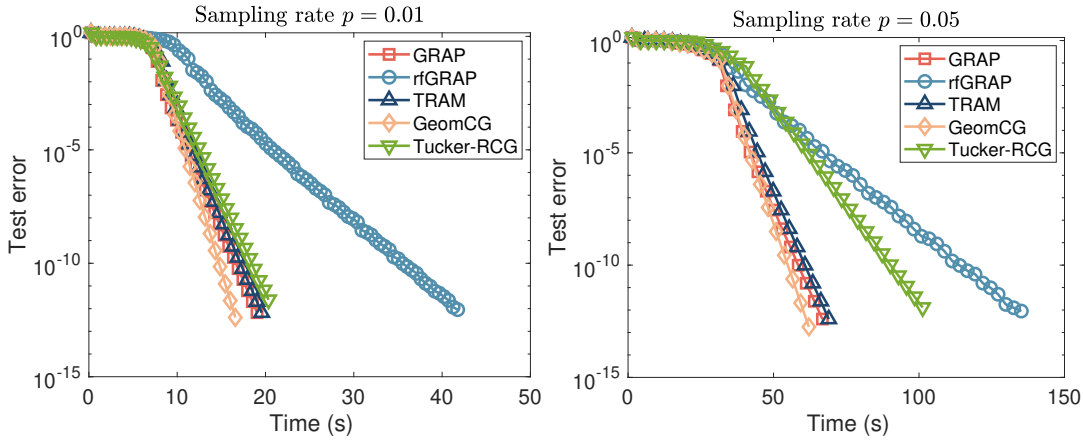


图 4-8 在采样率 $p = 0.01, 0.05$ 下, 各种方法的恢复性能.

Figure 4-8 The recovery performance under sampling rate $p = 0.01, 0.05$.

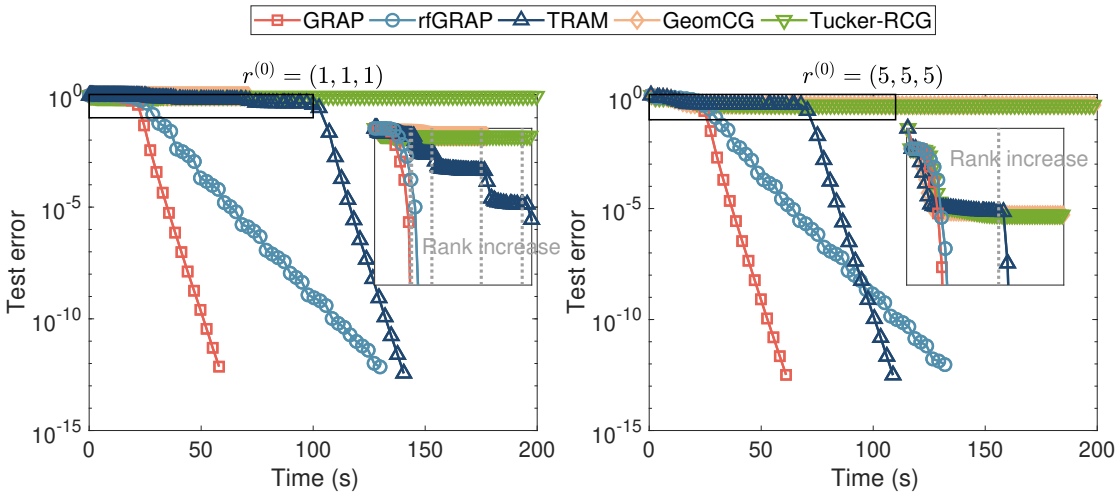


图 4-9 在不同初始秩 $\mathbf{r}^{(0)} = (1, 1, 1)$ 和 $\mathbf{r}^{(0)} = (5, 5, 5)$ 下的测试误差.

Figure 4-9 Test error under different initial ranks $\mathbf{r}^{(0)} = (1, 1, 1)$ and $\mathbf{r}^{(0)} = (5, 5, 5)$.

在秩参数被高估条件下的测试　我们在一组被高估的秩参数 $\mathbf{r} = (r, r, r)$ 下测试基于 Tucker 分解方法的性能, 其中 $r = 7, 8, 9, 10, 11, 12 > r^* = 6$. 采样率设置为 $p = 0.01$. 为保证与 GeomCG 和 Tucker-RCG(Q) 的公平比较, 初始点 $\mathcal{X}^{(0)}$ 从 $\mathcal{M}_{\mathbf{r}}$ 中生成. 数值结果如图 4-10 和图 4-11 所示. 首先, 从图 4-10 可以观察到, 当秩参数被高估时, 所提出的 TRAM 方法仍然能够收敛, 而其他方法由于秩参数设置过大而无法成功恢复数据张量. 其次, 图 4-10 (右) 表明, 在所有秩参数选择下, TRAM 都能够成功恢复数据张量 $\mathcal{A}$ 的真实 Tucker 秩. 因此, TRAM 的整体表现优于其他方法. 此外, 图 4-11 给出了在 $\mathbf{r} = (8, 8, 8)$ 情形下, TRAM 方法中展平矩阵 $\mathbf{X}_{(1)}^{(t)}$、$\mathbf{X}_{(2)}^{(t)}$ 和 $\mathbf{X}_{(3)}^{(t)}$ 的奇异值随迭代点的变化. 可以看到, 所提出的 TRAM 方法确实能够识别前六个主奇异值与其后两个奇异值之间的显著差异, 并触发秩减机制, 将秩参数降低至真实秩 $\mathbf{r}^*$.

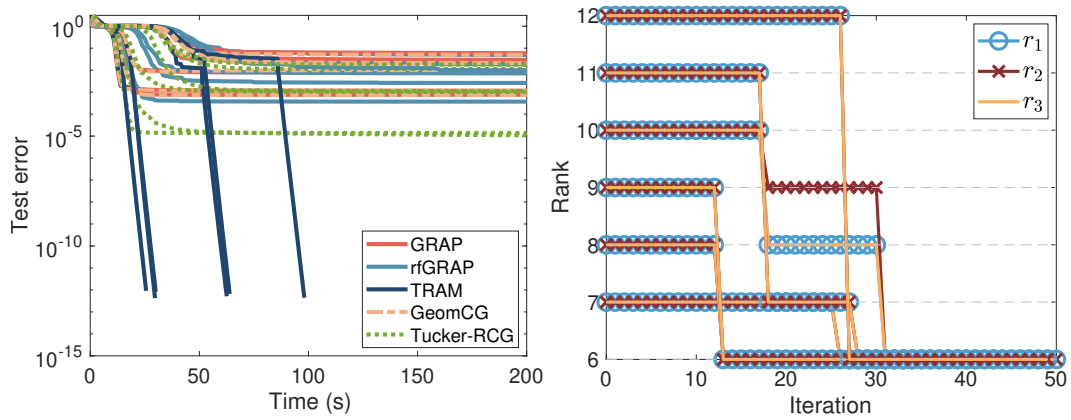


图 4-10 在合成数据集上秩参数被高估条件下, 基于 Tucker 分解方法的数值结果. 左: 测试误差. 右: TRAM 每步迭代点的秩.

Figure 4-10 Numerical results on synthetic dataset under over-estimated rank parameter of Tucker-based methods. Left: test error. Right: rank update of TRAM.

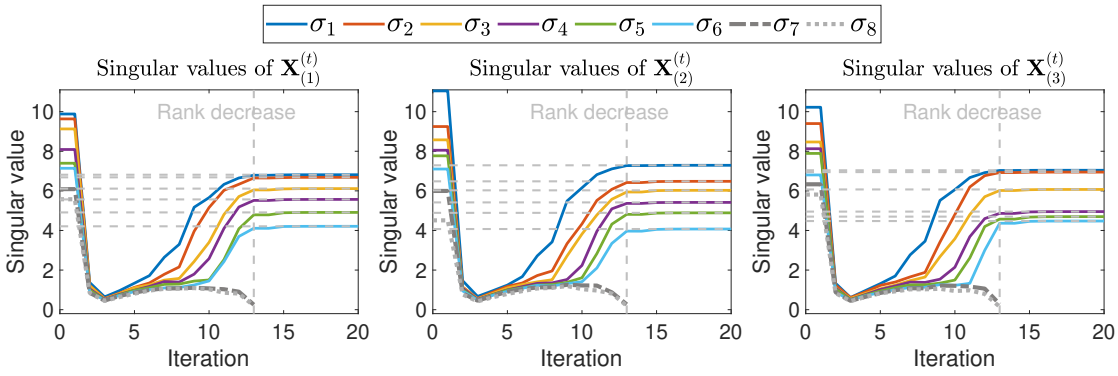


图 4-11 在秩参数 $\mathbf{r} = (8, 8, 8)$ 下, TRAM 方法中展平矩阵 $\mathbf{X}_{(1)}^{(t)}$、$\mathbf{X}_{(2)}^{(t)}$ 和 $\mathbf{X}_{(3)}^{(t)}$ 的奇异值随迭代点的变化.

Figure 4-11 The history of singular values of unfolding matrices $\mathbf{X}_{(1)}^{(t)}$, $\mathbf{X}_{(2)}^{(t)}$, and $\mathbf{X}_{(3)}^{(t)}$ for $\mathbf{r} = (8, 8, 8)$ in TRAM.

此外, 为了验证秩增机制的作用, 我们在初始秩 $\mathbf{r}^{(0)} = (1, 1, 1)$ 以及一组被高

估的秩参数 $\mathbf{r} = (r, r, r)$(其中 $r = 7, 8, 9, 10, 11, 12 > r^* = 6$) 下, 对所提出的方法进行比较. 图 4-12 给出了测试误差以及 TRAM 方法中秩更新的历史. 可以观察到, 由于引入了秩增机制, 只有 TRAM 方法能够成功识别并恢复真实秩.

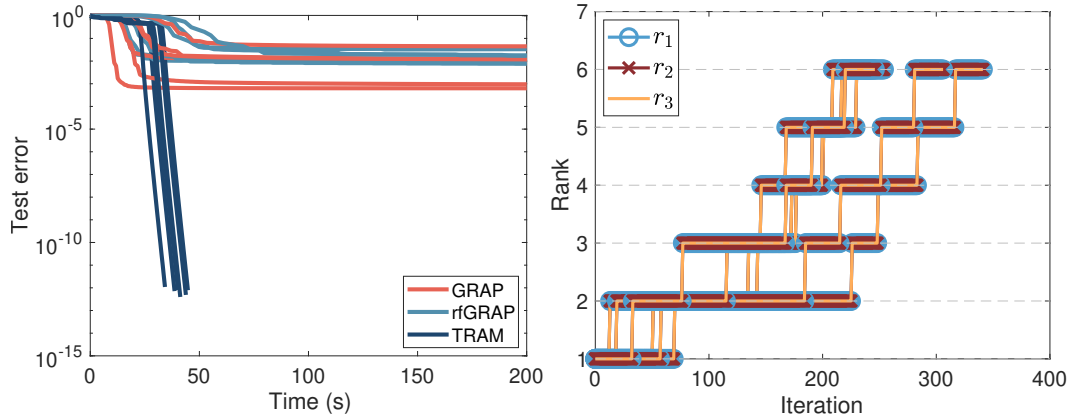


图 4-12 在合成数据集上、秩参数被高估且初始秩被低估为 $\mathbf{r}^{(0)} = (1, 1, 1)$ 的情况下, 基于 Tucker 分解方法的数值结果. 左: 测试误差. 右: TRAM 迭代点的秩更新过程.

Figure 4-12 Numerical results on synthetic dataset under over-estimated rank parameters and under-estimated initial rank $\mathbf{r}^{(0)} = (1, 1, 1)$. Left: test error. Right: rank update of TRAM.

总的来说, 如果真实秩未知, 秩增和秩减机制对于找到合适的秩参数都是至关重要的.

4.6.3 在高光谱图像上的数值实验

在本实验中, 我们考虑高光谱图像的张量补全问题, 并测试所提出的方法与其他方法的恢复性能. 高光谱图像可以表示为一个三阶张量 $\mathcal{A} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$. 其中, 第三个模态对应于光的 n_3 个波长维度, 而第一个模态和第二个模态分别表示在不同波长下的空间反射率分布. 我们选取了两幅高光谱图像: "Ribeira Hotel Image", 简称为 "Ribeira"[7], 其尺寸为 249×329×33; "220 Band AVIRIS Hyperspectral Image", 简称为 "AVIRIS"[8], 其尺寸为 $145 \times 145 \times 220$. 图 4-13 展示了这两幅高光谱图像的第 24 帧的灰度图.

为了评估图像补全的恢复性能, 我们采用峰值信噪比来衡量两幅图像之间的相似性, 其定义为

$$\mathrm{PSNR} := 10 \log_{10}\left(\frac{\max(\mathcal{A})^2}{\mathrm{MSE}}\right) = 10 \log_{10}\left(n_1 n_2 n_3 \frac{\max(\mathcal{A})^2}{\|\mathcal{X} - \mathcal{A}\|_{\mathrm{F}}^2}\right),$$

其中, $\max(\mathcal{A})$ 表示张量 $\mathcal{A}$ 中的最大像素值, MSE 为均方误差, 其定义为 MSE :=

[7]图片来源: 链接 https://figshare.manchester.ac.uk/articles/dataset/Fifty_hyperspectral_reflectance_images_of_outdoor_scenes/14877285中的 hsi_32.mat 文件.

[8]图片来源: https://purr.purdue.edu/publications/1947/1

图 4-13 两幅高光谱图像的第 24 帧的灰度图. 左图: “Ribeira”. 右图: “AVIRIS”.

Figure 4-13 The twenty-fourth frame of two images. Left: “Ribeira”. Right: “AVIRIS”.

$\|\mathcal{X}-\mathcal{A}\|_{\mathrm{F}}^2/(n_1n_2n_3)$. 此外, 我们还展示相对误差

$$\mathrm{relerr}(\mathcal{X}) := \frac{\|\mathcal{X}-\mathcal{A}\|_{\mathrm{F}}}{\|\mathcal{A}\|_{\mathrm{F}}}.$$

采样率取为 $p=0.1$. 我们在秩参数 $\mathbf{r}=(r,r,r)$(其中 $r=5,10,15,\ldots,30$) 下测试基于 Tucker 的方法. 为保证公平比较, 当算法达到最大迭代次数 250 时即终止, 该设置与文献 [96, §5] 是一致的.

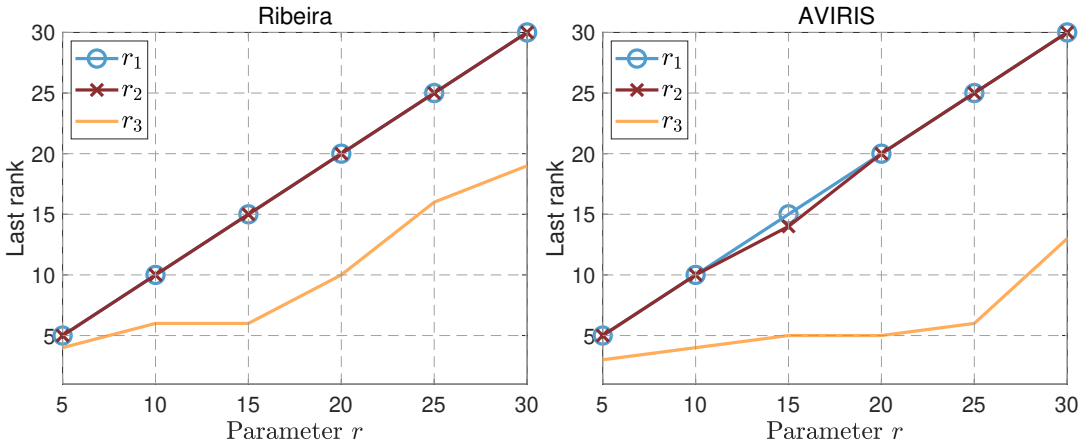


图 4-14 在不同秩参数 $\mathbf{r}=(r,r,r)$ 下, TRAM 在 “Ribeira” 与 “AVIRIS” 图像上得到的最终迭代点的秩.

Figure 4-14 The last rank obtained from TRAM for “Ribeira” and “AVIRIS” images under different parameters $\mathbf{r}=(r,r,r)$.

图 4-14 与表 4-1 展示了基于 Tucker 分解方法的恢复结果. 由图 4-14 可以观察到, TRAM 能够检测到沿第三模态的低秩结构, 即图像张量 $\mathcal{A}$ 在不同波长取值之间存在相似性. 值得注意的是, 在 “Ribeira” 图像中, 当 $\mathbf{r}=(15,15,15)$ 时, TRAM 得到的最终秩为 $(15,15,6)$, 这与 [15, §4.3.1] 中的秩参数选择是一致的. 此外, 表 4-1 给出了定量的恢复指标结果. 所提出的 GRAP 与 rfGRAP 方法在性能上与 GeomCG 和 Tucker-RCG 相当. 具体而言, 在大多数秩参数设置下, TRAM 方法取得了最高的 PSNR 以及最低的相对误差.

表 4-1 在 “Ribeira” 和 “AVIRIS” 图像上的相对误差与峰值信噪比.

Table 4-1 Relative error and PSNR on “Ribeira” and “AVIRIS” image.

Tucker 秩 **r** (r_1, r_2, r_3)	指标	GRAP	rfGRAP	TRAM	GeomCG	Tucker-RCG
		“Ribeira”				
(5, 5, 5)	PSNR	**24.9351**	24.9325	**24.9351**	**24.9351**	24.9350
	relerr	**0.2984**	0.2985	**0.2984**	**0.2984**	**0.2984**
(10, 10, 10)	PSNR	26.8481	26.8482	**26.8648**	26.8483	26.8482
	relerr	0.2394	0.2394	**0.2389**	0.2394	0.2394
(15, 15, 15)	PSNR	28.3451	28.3450	**28.4127**	28.3451	28.3451
	relerr	0.2015	0.2015	**0.1999**	0.2015	0.2015
(20, 20, 20)	PSNR	29.3908	29.3934	**29.5197**	29.3917	29.3924
	relerr	0.1786	0.1786	**0.1760**	0.1786	0.1786
(25, 25, 25)	PSNR	30.2324	30.1852	**30.3897**	30.2315	30.2332
	relerr	0.1621	0.1630	**0.1592**	0.1622	0.1621
(30, 30, 30)	PSNR	30.7088	30.7182	**30.9921**	30.7579	30.7566
	relerr	0.1535	0.1533	**0.1486**	0.1526	0.1527
		“AVIRIS”				
(5, 5, 5)	PSNR	**31.7181**	**31.7181**	31.6955	**31.7181**	**31.7181**
	relerr	**0.0835**	**0.0835**	0.0837	**0.0835**	**0.0835**
(10, 10, 10)	PSNR	33.7393	33.7393	**33.7517**	33.7393	33.7394
	relerr	0.0661	0.0661	**0.0660**	0.0661	0.0661
(15, 15, 15)	PSNR	35.1308	35.1157	**35.1427**	35.1144	35.1251
	relerr	0.0564	0.0564	**0.0563**	0.0565	0.0564
(20, 20, 20)	PSNR	36.1776	36.1777	**36.5438**	36.1781	36.1780
	relerr	0.0500	0.0500	**0.0479**	0.0500	0.0500
(25, 25, 25)	PSNR	36.6010	36.6430	**37.5433**	36.6142	36.6002
	relerr	0.0476	0.0473	**0.0427**	0.0475	0.0476
(30, 30, 30)	PSNR	36.3106	36.4263	**37.4879**	36.1278	36.1505
	relerr	0.0492	0.0485	**0.0430**	0.0502	0.0501

4.6.4　在 “MovieLens 1M” 数据集上的实验

我们进一步考虑在真实数据集 “MovieLens 1M”[9] 上进行张量补全测试. 该数据集包含从 1997 年 9 月 19 日至 1998 年 4 月 22 日期间, 6040 名用户对 3952 部电影给出的 1000209 条评分. 以一周为一个时间段, 这些评分可以构成一个大小为 $6040 \times 3952 \times 150$ 的三阶张量 $\mathcal{A}$. 我们随机选取 80% 的已知评分作为训练集 Ω, 其余 20% 作为测试集 Γ. 秩参数设置为 $\mathbf{r} = (r, r, r)$, 其中 $r = 1, 2, \ldots, 15$. 此外, 我们不仅将所提出的方法与其他基于 Tucker 的方法进行比较, 还与 CP-AltMin、TT-RCG 以及 TR-RGD 等方法进行对比. 为保证不同张量分解格式下参数数量具有可比性, 我们选取 CP 秩为 9, 张量链分解秩为 $(1, 4, 4, 1)$, 以及张量环补全中的秩为 $(3, 3, 3)$. 所提出方法的初始点 $\mathcal{X}^{(0)}$ 按照第 4.6.2 节中的 Tucker 分解方式生成; 随后, 其他方法的初始点分别通过 CP-ALS [38, Fig. 3.3]、TT-SVD [41, Theorem 2.1] 以及 TR-SVD [42, Algorithm 1] 从 $\mathcal{X}^{(0)}$ 转换得到. 需要注意的是, 不同张量格式下的初始点具有相近数量的参数. 当运行时间超过 3000s 时, 算法终止.

图 4-15 给出了 “MovieLens 1M” 数据集上的数值结果. 可以观察到: 1) 在不同秩参数 $\mathbf{r}$ 下, 所提出的方法在测试误差方面优于 GeomCG 与 Tucker-RCG; 2) 随着 $\mathbf{r}$ 的增大, TRAM 方法的测试误差对秩参数不敏感, 而其他方法的测试误差开始上升; 3) 图 4-15(右) 给出了 TRAM 得到的最终秩结果, 表明 TRAM 能够自适应地找到合适的秩参数 $\mathbf{r}^{(t)}$, 并揭示 “MovieLens 1M” 数据张量 $\mathcal{A}$ 在模态 2(电影类别) 上的低秩结构.

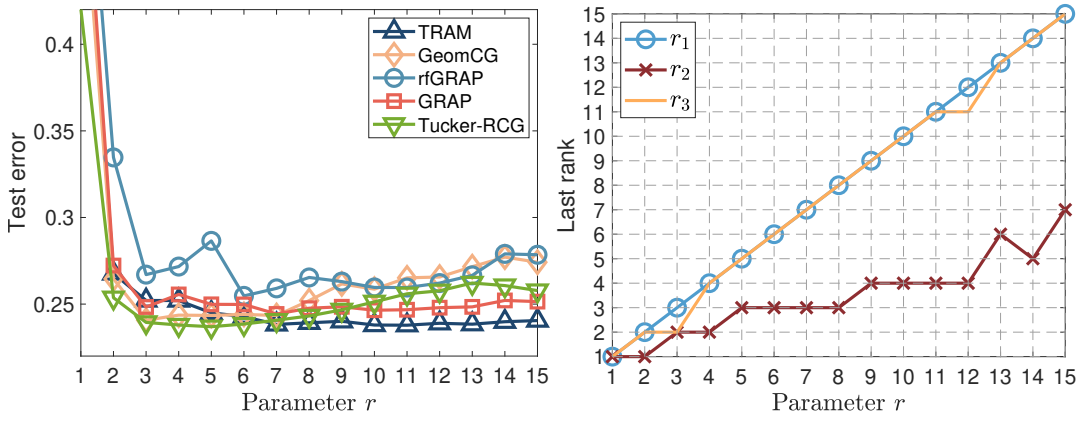


图 4-15 在不同秩参数 $\mathbf{r} = (r, r, r)$ 下的测试误差以及 TRAM 得到的最终秩参数. 左: 测试误差. 右: TRAM 的最终秩参数.

Figure 4-15 Test error and the last rank obtained from TRAM under different rank parameters $\mathbf{r} = (r, r, r)$. Left: test error. Right: last rank of TRAM.

此外, 我们在秩参数 $\mathbf{r} = (9, 9, 9)$ 下, 将基于 Tucker 的方法与其他方法进行比较. 由图 4-16(左) 可以看到, 所提出的方法与其他方法相比具有更优或可比的性能, 其中 TRAM 方法表现最佳. 图 4-16(右) 表明, 在 $\mathbf{r} = (9, 9, 9)$ 的设置下, 参数 $\mathbf{r}^{(t)}$ 被降低至 $(9, 4, 9)$. 这一结果意味着 TRAM 在 “MovieLens 1M” 数据中成功识

[9] 数据集可见 https://grouplens.org/datasets/movielens/1m/.

别出了四个不同的电影类别.

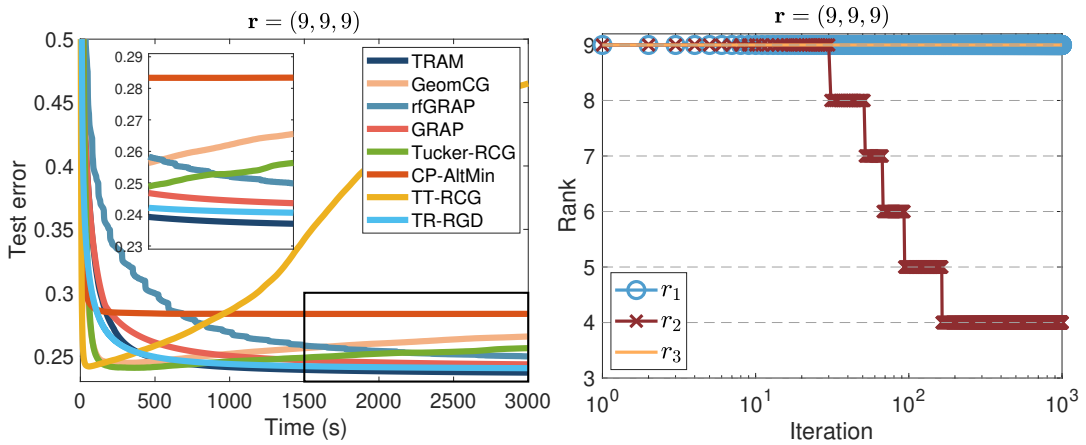


图 4-16 在秩参数 $\mathbf{r} = (9, 9, 9)$ 下, "MovieLens 1M" 数据集的数值结果. 左: 测试误差. 右: TRAM 迭代过程中秩的更新.

Figure 4-16 Numerical results on "MovieLens 1M" dataset under rank parameter $\mathbf{r} = (9, 9, 9)$. Left: test error. Right: rank update in iterations.

注. 在第3.5.2节中, 我们同样在电影评分数据集 "MovieLens 1M" 上测试了基于不同张量分解的张量补全算法. 因此在本节中, 我们选用第3章中的 TR-RGD 方法 (算法8) 作为 TR 类张量补全方法的代表进行数值比较. 我们发现 TRAM 方法相较于 TR-RGD 方法能达到更低的测试误差, 这是因为 TRAM 方法中迭代点的秩会被动态调整以选用合适的秩参数, 而 TR-RGD 方法不具有类似的机制. 但 TR-RGD 方法作为预条件方法, 相较于其他类型的方法能更快地达到更低的测试误差, 这也说明了 TR-RGD 方法的有效性. 总的来说, 我们认为, 在张量补全问题中, 如果我们事先预知合适的秩参数, 我们推荐 TR-RGD 方法以避免 TRAM 方法中的秩寻找过程, 如果秩参数未知, 我们建议选择 TRAM 方法并设置一个较大的秩参数, 通过 TRAM 方法中自带的秩自适应机制寻找合适的秩参数以达到更好的效果.

4.7 本章小结

在本章中, 我们对 Tucker 张量代数簇的几何结构进行了深入研究, 并提出了用于 Tucker 张量代数簇优化的新几何方法与秩自适应方法. 我们给出了 Tucker 张量代数簇每一点切锥的显式表达. 我们观察到, Tucker 张量代数簇的几何结构与矩阵代数簇密切相关, 但其远比矩阵代数簇复杂. 所有结果都可以通过几何示意图优雅地归纳到已知的矩阵代数簇结果上. 此外, Tucker 张量代数簇优化的一个核心在于度量投影. 基于建立的几何结构, 我们提出了近似投影方法, 以规避度量投影的显式计算. 令人惊讶的是, 我们发现仅利用切锥的部分信息即可获得无需收缩映射的搜索方向. 张量补全的数值实验表明, 所提出的方法在不同秩参数选择下均优于现有最先进方法.

第 5 章　归一化的张量链分解: 几何与应用

5.1　引言

给定一个张量 $\mathcal{A} \in \mathbb{C}^{n_1 \times n_2 \times \cdots \times n_d}$, $\mathcal{A}$ 的归一化的张量链分解 (normalized tensor train decomposition, NTT) 旨在用一个具有单位 Frobenius 范数的低秩张量 $[\![\mathcal{U}_1, \mathcal{U}_2, \ldots, \mathcal{U}_d]\!]$ 来近似 $\mathcal{A}$, 即

$$\mathcal{A}(i_1, i_2, \ldots, i_d) \approx \mathbf{U}_1(i_1)\mathbf{U}_2(i_2)\cdots\mathbf{U}_d(i_d) \quad \text{满足} \quad \left\|[\![\mathcal{U}_1, \mathcal{U}_2, \ldots, \mathcal{U}_d]\!]\right\|_{\mathrm{F}} = 1$$

其中 $i_k = 1, 2, \ldots, n_k$ 和 $k = 1, 2, \ldots, d$, 这里 $\mathcal{U}_k \in \mathbb{C}^{r_{k-1} \times n_k \times r_k}$ 是第 k 个核张量, $\mathbf{U}_k(i_k) = \mathcal{U}_k(:, i_k, :)$, 并且 $r_0, r_1, \ldots, r_d$ 是满足 $r_0 = r_d = 1$ 的正整数. 我们将 $[\![\mathcal{U}_1, \mathcal{U}_2, \ldots, \mathcal{U}_d]\!]$ 称为一个 NTT 张量. 图 5-1 描述了一个张量的 NTT 分解. 需要注意的是, NTT 分解同样可以定义在 $\mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$ 中的张量上.

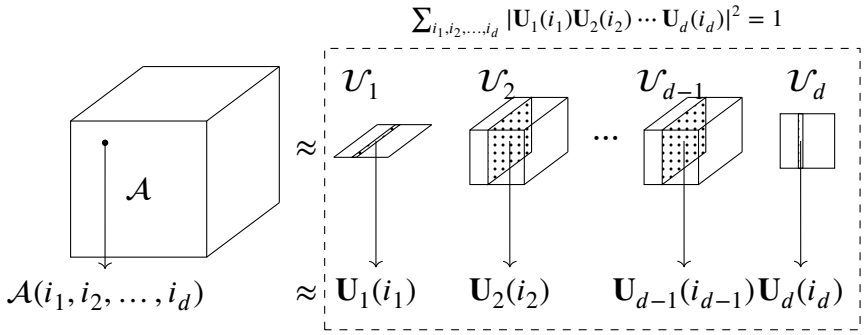


图 5-1 一个张量的归一化张量链分解

Figure 5-1 Normalized tensor train decomposition of a tensor.

5.1.1　NTT 分解的应用

我们展示 NTT 分解的应用, 第一类是在科学计算中的应用.

低秩张量恢复　给定一个在索引集 $\Omega \subset [n_1] \times [n_2] \times \cdots \times [n_d]$ 上部分观测到的、具有单位范数的 NTT 张量 $\mathcal{A} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$, 我们的目标是利用 $\mathcal{A}$ 的低秩结构来恢复完整张量. 该问题出现在诸多应用中, 例如统计学、机器学习以及高维函数的压缩；参见例如 [137, 157]. 具体而言, 低秩张量恢复可以通过求解如下定义在 NTT 张量集合上的优化问题来实现:

$$\begin{aligned} \min \quad & f(\mathcal{X}) = \frac{1}{2}\| \mathrm{P}_\Omega(\mathcal{X}) - \mathrm{P}_\Omega(\mathcal{A})\|_{\mathrm{F}}^2 \\ \text{s.t.} \quad & \mathcal{X} \text{ 是一个 NTT 张量;} \end{aligned}$$

详见第5.3.1节.

具有张量积结构的特征值问题 对于一个对称矩阵 $\mathbf{A} \in \mathbb{R}^{(n_1 n_2 \cdots n_d)\times(n_1 n_2 \cdots n_d)}$, 计算其最小 (最大) 特征值 $\lambda_{\min}(\lambda_{\max})$ 及对应的特征向量 $\mathbf{x} \in \mathbb{R}^{n_1 n_2 \cdots n_d}$ 是数值线性代数与计算物理中的核心问题之一 [158]. 其中, 空间 $\mathbb{R}^{n_1 n_2 \cdots n_d}$ 例如可以来源于在张量积空间 $\mathbb{R}^{n_1} \otimes \mathbb{R}^{n_2} \otimes \cdots \otimes \mathbb{R}^{n_d}$ 上对高维偏微分方程的离散化. 然而, 直接求解该问题将遭遇维数灾难. 类似的困难也出现在量子多体物理中, 其中量子系统的哈密顿量由一个 Hermitian 矩阵 $\mathbf{A} \in \mathbb{C}^{(n_1 n_2 \cdots n_d)\times(n_1 n_2 \cdots n_d)}$ 描述. 寻找哈密顿量的基态及其能量同样可以建模为一个特征值问题. 此外, 低秩 MPS 能够忠实地表示局部哈密顿量的基态 [159], 这催生了重要的数值算法, 例如用于求解局部哈密顿量的密度矩阵重整化群 (density matrix renormalization group, DMRG) 方法 [160, 161]. 基于这一观察, 我们的目标是在利用 NTT 形式寻找该特征值问题的低秩解, 即

$$\begin{aligned} \min_{\mathcal{X}}(\max_{\mathcal{X}}) \quad & \mathrm{vec}(\mathcal{X})^{\top}\mathbf{A}\mathrm{vec}(\mathcal{X}) \\ \text{s.t.} \quad & \mathcal{X} \text{ 是一个 NTT 张量;} \end{aligned}$$

详见第 5.3.2节.

第二类应用来源于量子信息理论, 其中 NTT 分解使得从数值上对两个基本概念的研究变得高效: 量子资源的量化以及信道容量的可加性.

稳定子秩的近似 我们考虑量子信息中的非稳定子性 (non-stabilizerness), 亦称为魔性 (magic) [162], 这是体现量子计算机优势的一个关键要素. 刻画纯态非稳定子性的一个核心度量是稳定子秩 (stabilizer rank) [163], 它的定义是使目标态能够表示为 R 个稳定子态的凸组合所需的最小整数 R. 在操作层面上, 它对应于量子线路的经典模拟代价. 然而, 计算一个给定 n-比特量子态的稳定子秩是不可行的, 因为稳定子态的数量以 $2^{\Theta(n^2)}$ 的速度超指数增长. 因此, 任何穷举类方法 (例如遍历所有可能的稳定子态组合以寻找分解) 在计算上都是不可行的, 即便对于较小的系统也是如此. 基于此, 我们引入纯态 $|\psi\rangle$ 的 (ϵ, δ)-近似稳定子秩的概念, 并提出通过在多个固定秩 NTT 张量集合的笛卡尔积上求解如下优化问题估计该近似秩:

$$\begin{aligned} \min_{\{c_j\}_j, \{|\phi_j\rangle\}_j} \quad & \frac{1}{2}\Big\| \sum_{j=1}^{R} c_j |\phi_j\rangle - |\psi\rangle \Big\|_{\mathrm{F}}^2 + \sum_{j=1}^{R} M_2(|\phi_j\rangle) \\ \text{s.t.} \quad & c_1, \ldots, c_R \in \mathbb{C}, \text{ 每一个 } |\phi_j\rangle \text{ 是 NTT 张量;} \end{aligned}$$

详见第 5.4.1节.

最小输出 Rényi p-熵 我们利用 NTT 分解研究量子信道 (完全正且保迹的线性映射) 的最小输出 Rényi p-熵的可加性. 当 p 趋于 1 时, 该量的非可加性已由 Hastings 在高维情形下证明 [164], 这一结果解决了量子信息理论中的一个重大公开问题: 量子信道经典容量的非可加性. 然而, 构造显式反例仍然极具挑战, 目前仅有少数例子被发现 [165–168]. 检验超可加性的主要瓶颈在于, 需要计算信道 n 次张量积的最小输出熵, 而该优化是在输入量子态空间上进行的, 其维数随 n 指数级增长.

通过将高维输入态表示为 NTT 形式, 我们将最小输出熵的计算转化为一个在固定秩 NTT 张量集合上的可处理优化问题:

$$\begin{aligned} \min_{|\psi\rangle} \quad & \frac{1}{1-p}\log \mathrm{tr}(N_{A\to B}^{\otimes n}(|\psi\rangle\langle\psi|_{A^n})^p) \\ \text{s.t.} \quad & |\psi\rangle \text{ 是一个 NTT 张量;} \end{aligned}$$

详见第 5.4.2 节.

由于张量结构本身的复杂性, 归一化张量链分解的性质与几何结构无法直接从已有结果中推广得到. 首先, 额外的单位范数约束从根本上改变了低秩逼近问题的性质. 对于标准的张量链分解, 可以通过顺序奇异值分解计算得到一种拟优的低秩逼近. 然而, 在引入单位范数约束之后, 如何在 NTT 形式下构造拟优逼近是未知的. 其次, 众所周知, 固定秩的 TT 张量构成的集合是一个光滑流形. 然而, 这一性质在 NTT 形式下是否仍然成立, 目前尚不清楚. 第三, 将一个张量投影到固定秩的 NTT 张量集合上, 需要额外的步骤以同时满足单位范数约束和低秩约束, 这不可避免地增加了计算成本, 从而我们需要针对 NTT 张量设计高效的基本运算方式.

5.1.2　本章主要内容

在本章中, 我们提出了归一化张量链分解, 并系统研究了 NTT 形式下张量的性质与几何结构. 首先, 我们证明了对于具有单位 Frobenius 范数的张量, NTT 分解是存在的, 且与标准的 TT 分解等价. 对于不具有单位范数的张量 $\mathcal{A}$, 我们通过 TT-SVD 以及向单位球的投影, 构造了一个秩为 $\mathbf{r}$ 的近似算子 $\mathrm{P}_{\mathbf{r}}^{\mathrm{NTTSVD}}$, 并证明其满足如下拟优性:

$$\|\mathrm{P}_{\mathbf{r}}^{\mathrm{NTTSVD}}(\mathcal{A})-\mathcal{A}\|_{\mathrm{F}} \le (2\sqrt{d-1}+1)\|\mathrm{P}_{\mathcal{N}_{\mathbf{r}}}(\mathcal{A})-\mathcal{A}\|_{\mathrm{F}},$$

其中 $\mathrm{P}_{\mathcal{N}_{\mathbf{r}}}(\mathcal{A})$ 表示 $\mathcal{A}$ 在 NTT 形式下的最佳秩-$\mathbf{r}$ 逼近.

随后, 我们考虑集合

$$\mathcal{N}_{\mathbf{r}} = \{\mathcal{X}\in\mathbb{C}^{n_1\times n_2\times\cdots\times n_d} : \mathrm{rank}_{\mathrm{TT}}(\mathcal{X})=\mathbf{r} \text{ 和 } \|\mathcal{X}\|_{\mathrm{F}}=1\},$$

即秩为 $\mathbf{r}$ 的 NTT 张量集合. 该集合可以看作是固定秩 TT 张量流形与单位球面的交集. 由于两个光滑流形是横截相交的, 我们由此推导出 $\mathcal{N}_{\mathbf{r}}$ 本身也是一个光滑流形. 在此基础上, 我们建立了 $\mathcal{N}_{\mathbf{r}}$ 上的黎曼几何结构, 从而为在 $\mathcal{N}_{\mathbf{r}}$ 上提出几何优化方法打下基础. $\mathcal{N}_{\mathbf{r}}$ 的低秩结构不仅显著降低了存储与计算成本, 同时也保留了张量的内在单位范数结构. 张量链分解与归一化张量链分解之间的差异总结于表 5-1 中.

基于所提出的 NTT 分解, 我们进一步研究了在流形 $\mathcal{N}_{\mathbf{r}}$ 上光滑函数的几何优化问题, 并提出了一种黎曼共轭梯度方法, 称为 NTT-RCG. 该方法被系统应用于多个典型问题中, 包括低秩张量恢复、特征值问题、量子物理中的稳定子秩计算以及最小输出 Rényi p-熵的数值计算.

表 5-1 张量链分解与归一化张量链分解之间的差异, 详见第 1.3 和 5.2节. 其中 $\mathbf{r} = (r_0, r_1, \dots, r_d)$. $\mathcal{B}_1 = \{\mathcal{X} \in \mathbb{C}^{n_1 \times n_2 \times \cdots \times n_d} : \|\mathcal{X}\|_{\mathrm{F}} = 1\}$.

Table 5-1 The differences between tensor train decomposition and normalized TT decomposition; see Sections 1.3 and 5.2 for details. $\mathbf{r} = (r_0, r_1, \dots, r_d)$. $\mathcal{B}_1 = \{\mathcal{X} \in \mathbb{C}^{n_1 \times n_2 \times \cdots \times n_d} : \|\mathcal{X}\|_{\mathrm{F}} = 1\}$.

性质	张量链分解	归一化张量链分解
秩-$\mathbf{r}$ 逼近	$\mathrm{P}_{\mathbf{r}}^{\mathrm{TT-SVD}}$	$\mathrm{P}_{\mathcal{B}_1} \circ \mathrm{P}_{\mathbf{r}}^{\mathrm{TT-SVD}}$
拟优性	$\sqrt{d-1}$ in (1-17)	$(2\sqrt{d-1}+1)$ in (5-2)
参数空间	$\mathbb{C}^{r_0 \times n_1 \times r_1} \times \mathbb{C}^{r_1 \times n_2 \times r_2} \times \cdots \times \mathbb{C}^{r_{d-1} \times n_d \times r_d}, \quad r_0 = r_d = 1$	
固定秩流形	$\mathcal{M}_{\mathbf{r}}$	$\mathcal{N}_{\mathbf{r}} = \mathcal{M}_{\mathbf{r}} \cap \mathcal{B}_1$
维数	$\sum_{k=1}^{d} r_{k-1} n_k r_k - \sum_{k=1}^{d-1} r_k^2$	$\sum_{k=1}^{d} r_{k-1} n_k r_k - \sum_{k=1}^{d-1} r_k^2 - 1$
切空间	$\sum_{k=1}^{d} [\![\mathcal{U}_1, \dots, \mathcal{U}_{k-1}, \dot{\mathcal{U}}_k, \mathcal{U}_{k+1}, \dots, \mathcal{U}_d]\!]$	
	$\mathrm{L}(\dot{\mathcal{U}}_k)^{\dagger}\mathrm{L}(\mathcal{U}_k) = 0, k = 1, 2, \dots, d-1$	$\mathrm{L}(\dot{\mathcal{U}}_k)^{\dagger}\mathrm{L}(\mathcal{U}_k) = 0, k = 1, 2, \dots, d$

在科学计算应用中, 数值结果表明, NTT-RCG 方法能够在无噪声和含噪声情形下有效恢复低秩张量. 在特征值问题中, 我们将所提出的方法与交替线性化方法 (即物理中常称的 DMRG 方法) 进行了比较, 结果显示 NTT-RCG 在收敛速度和最大 (最小) 特征值的计算精度方面均具有明显优势.

在量子信息的应用中, NTT 分解不仅显著降低了存储成本, 还使得目标函数具有高效的计算方法. 数值实验表明, NTT-RCG 方法能够有效地用若干非稳定性显著更低的状态来逼近非稳定子态. 此外, NTT-RCG 还提供了一种直接计算最小输出 Rényi p-熵的实用方法. 针对反对称信道和广义振幅阻尼信道的实验结果表明, 该方法的计算复杂度关于量子比特数和键维数呈多项式增长, 并在最多 12 个量子比特、NTT 秩不超过 $(1, 10, 10, \dots, 10, 1)$ 的情况下, 数值验证了不存在超可加性现象.

5.2 归一化的张量链分解

在本节中, 我们首先定义 NTT 分解. 随后, 我们给出一种将完整张量近似投影到低秩 NTT 张量的方法. 此外, 我们还将研究固定秩 NTT 张量集合的几何结构.

5.2.1 NTT 分解的定义

张量链分解能够将一个高维张量分解为一系列较小的核张量. 然而, 标准的 TT 分解并未考虑单位范数约束, 而该约束在许多应用中是自然出现的. 为此, 我们引入归一化张量链分解, 其目标是用一个 Frobenius 范数为 1 的 TT 张量来近似给定的完整张量.

定义 5.1 (归一化的张量链分解)**.** 给定张量 $\mathcal{A} \in \mathbb{C}^{n_1 \times n_2 \times \cdots \times n_d}$, $\mathcal{A}$ 的归一化张量列分解旨在用一个 TT 张量 $[\![\mathcal{U}_1, \mathcal{U}_2, \ldots, \mathcal{U}_d]\!]$ 近似 $\mathcal{A}$, 满足

$$\mathcal{A} \approx [\![\mathcal{U}_1, \mathcal{U}_2, \ldots, \mathcal{U}_d]\!] \quad \text{且} \quad \|[\![\mathcal{U}_1, \mathcal{U}_2, \ldots, \mathcal{U}_d]\!]\|_{\mathrm{F}} = 1,$$

其中 $\mathcal{U}_k \in \mathbb{C}^{r_{k-1} \times n_k \times r_k}$, $k \in [d]$, 并且 r_k 为正整数 ($k \in [d-1]$). 在物理学中, 正整数 $r_1, r_2, \ldots, r_{d-1}$ 通常被称为键维数 (bond dimensions).

需要注意的是, NTT 分解给出的是对任意张量的单位范数低秩近似, 而非严格意义上的精确分解. 若 $\|\mathcal{A}\|_{\mathrm{F}} = 1$, 则张量 $\mathcal{A}$ 存在如下形式的精确 NTT 分解:

$$\mathcal{A} = [\![\mathcal{U}_1, \mathcal{U}_2, \ldots, \mathcal{U}_d]\!] \quad \text{且} \quad \|[\![\mathcal{U}_1, \mathcal{U}_2, \ldots, \mathcal{U}_d]\!]\|_{\mathrm{F}} = 1,$$

其中 $r_k = \operatorname{rank}(\mathbf{A}_{\langle k \rangle})$, $k \in [d-1]$. 事实上, 在这种情况下, 所得的 NTT 分解与 $\mathcal{A}$ 的标准 TT 分解是等价的, 因为 TT 分解本身保张量的 Frobenius 范数. 若 $\|\mathcal{A}\|_{\mathrm{F}} \neq 1$, 则可以先对 $\mathcal{A}$ 执行标准 TT 分解, 再对所得的 TT 张量进行归一化, 从而得到 NTT 分解. 注意这两步操作是不可交换的. 此外, 该构造满足拟优性性质, 具体见命题 5.1. 最后需要指出的是, NTT 分解同样可以自然地定义在实张量空间 $\mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$ 上.

核张量的正交化　NTT 张量同样可以像 TT 那样, 通过 QR 分解进行左正交化或右正交化. 然而, 正交化之后核张量所满足的正交性质在 TT 与 NTT 两种分解形式下是不同的. 具体而言, 设 $\mathcal{X} = [\![\mathcal{U}_1, \mathcal{U}_2, \ldots, \mathcal{U}_d]\!] \in \mathcal{N}_{\mathbf{r}}$ 是一个左正交化的 NTT 张量, 即对于 $k \in [d-1]$, $\mathrm{L}(\mathcal{U}_k)^{\dagger}\mathrm{L}(\mathcal{U}_k) = \mathbf{I}_{r_k}$, 由 $\|\mathcal{X}\|_{\mathrm{F}} = 1$ 以及 $\mathrm{L}(\mathcal{U}_d) \in \mathbb{C}^{r_{d-1} n_d}$ 可得

$$\mathrm{L}(\mathcal{U}_d)^{\dagger}\mathrm{L}(\mathcal{U}_d) = \|\mathcal{U}_d\|_{\mathrm{F}}^2 = \|\mathbf{X}_{\leq k-1}^{\dagger}\mathrm{R}(\mathcal{U}_d)\|_{\mathrm{F}}^2 = \|\mathbf{X}_{\langle d-1 \rangle}\|_{\mathrm{F}}^2 = \|\mathcal{X}\|_{\mathrm{F}}^2 = 1, \tag{5-1}$$

其中我们使用了 $\mathbf{X}_{\leq d-1}^{\dagger}\mathbf{X}_{\leq d-1} = \mathbf{I}_{r_{d-1}}$. 因此, 与 TT 分解的不同点在于: 在 TT 分解中, 经过左正交化后, 最后一个核张量 $\mathcal{U}_d$ 并不一定满足左正交性; 而在 NTT 分解中, 所有核张量 $\mathcal{U}_k$(包括最后一个 $\mathcal{U}_d$) 在左正交化后均满足左正交性. 这一差异正是由 NTT 分解中引入的单位范数约束所导致的.

5.2.2　NTT-SVD 算法

给定一个完整张量 $\mathcal{A} \in \mathbb{C}^{n_1 \times n_2 \times \cdots \times n_d}$ 以及秩参数数组 $\mathbf{r} = (1, r_1, r_2, \ldots, r_{d-1}, 1)$, 一个自然的问题是: 如何计算 $\mathcal{A}$ 在 NTT 形式下的最优秩 $\mathbf{r}$ 近似, 即如何求解如下的度量投影问题

$$\mathrm{P}_{\mathcal{N}_{\mathbf{r}}}(\mathcal{A}) := \operatorname*{arg\,min}_{\mathcal{X} \in \mathcal{N}_{\mathbf{r}}} \|\mathcal{X} - \mathcal{A}\|_{\mathrm{F}}.$$

注意到集合

$$\mathcal{N}_{\mathbf{r}} = \mathcal{M}_{\mathbf{r}} \cap \mathcal{B}_1 = \{\mathcal{X} \in \mathbb{C}^{n_1 \times n_2 \times \cdots \times n_d} : \operatorname{rank}_{\mathrm{TT}}(\mathcal{X}) = \mathbf{r} \text{ 和 } \|\mathcal{X}\|_{\mathrm{F}} = 1\}$$

是固定 TT 秩张量集合 $\mathcal{M}_{\mathbf{r}}$(见式 (1-15)) 与单位球 $\mathcal{B}_1 = \{\mathcal{X} \in \mathbb{C}^{n_1\times n_2\times\cdots\times n_d} : \|\mathcal{X}\|_{\mathrm{F}} = 1\}$ 的交集. 因此, 原则上可以通过依次投影到 $\mathcal{M}_{\mathbf{r}}$ 和 $\mathcal{B}_1$ 来构造到 $\mathcal{N}_{\mathbf{r}}$ 的投影, 这两步分别对应于 TT 形式下的最优秩-$\mathbf{r}$ 近似以及归一化操作.

然而, 众所周知, TT 形式下的最优秩-$\mathbf{r}$ 近似并不存在显式表达式 [41], 这使得精确计算投影 $\mathrm{P}_{\mathcal{N}_{\mathbf{r}}}(\mathcal{A})$ 在实际中不可行. 为此, 我们采用一种近似投影策略: 首先对 $\mathcal{A}$ 执行 TT-SVD, 得到一个秩为 $\mathbf{r}$ 的近似 $\mathrm{P}_{\mathbf{r}}^{\mathrm{TTSVD}}(\mathcal{A}) \in \mathcal{M}_{\mathbf{r}}$, 然后再将其归一化到单位球 $\mathcal{B}_1$ 上, 即

$$\mathrm{P}_{\mathbf{r}}^{\mathrm{NTTSVD}}(\mathcal{A}) := \mathrm{P}_{\mathcal{B}_1}(\mathrm{P}_{\mathbf{r}}^{\mathrm{TTSVD}}(\mathcal{A}));$$

其流程示意见图 5-2. 我们将该近似投影算子 $\mathrm{P}_{\mathbf{r}}^{\mathrm{NTTSVD}}$ 称为 NTT-SVD 算法. 需要强调的是, 上述两步操作不能交换顺序. 在实际计算中, 我们可以高效完成投影到单位球 $\mathcal{B}_1$ 的归一化: 由于 TT-SVD 算法生成的核张量 $\mathcal{U}_1, \mathcal{U}_2, \ldots, \mathcal{U}_{d-1}$ 是左正交的, 并且由式 (5-1) 可知 $\|[\![\mathcal{U}_1, \mathcal{U}_2, \ldots, \mathcal{U}_d]\!]\|_{\mathrm{F}} = \|\mathcal{U}_d\|_{\mathrm{F}}$, 因此只需对最后一个核张量 $\mathcal{U}_d$ 进行归一化即可完成整体张量的单位范数约束.

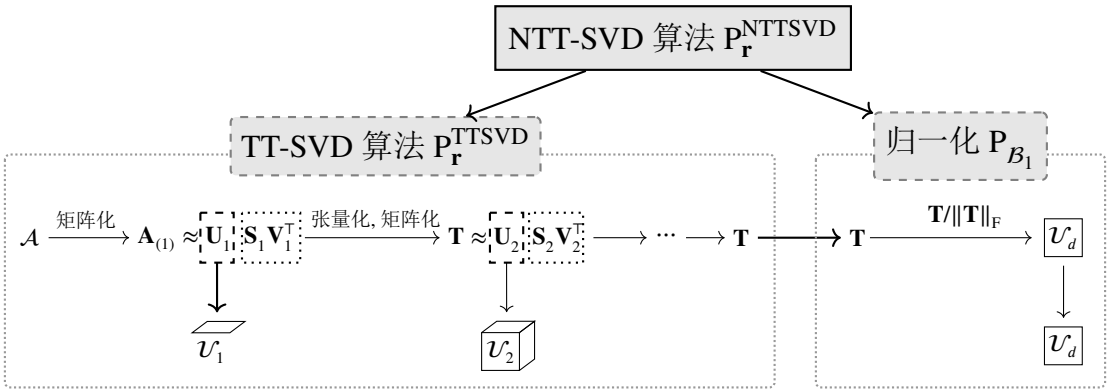


图 5-2 NTT-SVD 算法的流程图.

Figure 5-2 Flowchart of the NTT-SVD algorithm.

尽管所提出的近似投影 $\mathrm{P}_{\mathbf{r}}^{\mathrm{NTTSVD}}$ 并不一定是一个度量投影 $\mathrm{P}_{\mathcal{N}_{\mathbf{r}}}$, 下面的命题刻画了 $\mathrm{P}_{\mathbf{r}}^{\mathrm{NTTSVD}}$ 与 $\mathrm{P}_{\mathcal{N}_{\mathbf{r}}}$ 之间的关系, 该性质通常被称为拟优性 (quasi-optimality). 从几何角度来看，$\mathrm{P}_{\mathcal{N}_{\mathbf{r}}}(\mathcal{A})$ 是在低秩集合 $\mathcal{N}_{\mathbf{r}}$ 上距离 $\mathcal{A}$ 最近的点 (度量投影), 而 $\mathrm{P}_{\mathbf{r}}^{\mathrm{NTTSVD}}(\mathcal{A})$ 则提供了一个可计算的近似点. 命题 5.1 表明, 该近似点到 $\mathcal{A}$ 的距离在一个与阶数 d 相关的常数因子内逼近最优距离.

命题 5.1 (拟优性)**.** *近似投影对任意张量 $\mathcal{A} \in \mathbb{C}^{n_1\times n_2\times\cdots\times n_d}$ 以及秩参数 $\mathbf{r}$ 都满足*

$$\|\mathrm{P}_{\mathcal{N}_{\mathbf{r}}}(\mathcal{A}) - \mathcal{A}\|_{\mathrm{F}} \le \|\mathrm{P}_{\mathbf{r}}^{\mathrm{NTTSVD}}(\mathcal{A}) - \mathcal{A}\|_{\mathrm{F}} \le (2\sqrt{d-1}+1)\|\mathrm{P}_{\mathcal{N}_{\mathbf{r}}}(\mathcal{A}) - \mathcal{A}\|_{\mathrm{F}} \quad (5\text{-}2)$$

成立.

证明. 由于 $\mathrm{P}_{\mathcal{N}_{\mathbf{r}}}(\mathcal{A})$ 是 $\mathcal{A}$ 在 NTT 形式下的最优秩-$\mathbf{r}$ 近似, 显然有 $\|\mathrm{P}_{\mathcal{N}_{\mathbf{r}}}(\mathcal{A}) -$

$\mathcal{A}\|_{\mathrm{F}} \leq \|\mathrm{P}_{\mathbf{r}}^{\mathrm{NTTSVD}}(\mathcal{A}) - \mathcal{A}\|_{\mathrm{F}}$. 下面证明第二个不等式. 注意到

$$
\begin{aligned}
\|\mathrm{P}_{\mathbf{r}}^{\mathrm{NTTSVD}}(\mathcal{A}) - \mathcal{A}\|_{\mathrm{F}} &= \|\mathrm{P}_{\mathcal{B}_1}(\mathrm{P}_{\mathcal{M}_{\mathbf{r}}}^{\mathrm{TTSVD}}(\mathcal{A})) - \mathcal{A}\|_{\mathrm{F}} \\
&\leq \|\mathrm{P}_{\mathcal{B}_1}(\mathrm{P}_{\mathcal{M}_{\mathbf{r}}}^{\mathrm{TTSVD}}(\mathcal{A})) - \mathrm{P}_{\mathcal{M}_{\mathbf{r}}}^{\mathrm{TTSVD}}(\mathcal{A})\|_{\mathrm{F}} + \|\mathrm{P}_{\mathcal{M}_{\mathbf{r}}}^{\mathrm{TTSVD}}(\mathcal{A}) - \mathcal{A}\|_{\mathrm{F}} \\
&\leq \|\mathrm{P}_{\mathcal{N}_{\mathbf{r}}}(\mathcal{A}) - \mathrm{P}_{\mathcal{M}_{\mathbf{r}}}^{\mathrm{TTSVD}}(\mathcal{A})\|_{\mathrm{F}} + \|\mathrm{P}_{\mathcal{M}_{\mathbf{r}}}^{\mathrm{TTSVD}}(\mathcal{A}) - \mathcal{A}\|_{\mathrm{F}} \qquad (5\text{-}3) \\
&\leq \|\mathrm{P}_{\mathcal{N}_{\mathbf{r}}}(\mathcal{A}) - \mathcal{A}\|_{\mathrm{F}} + \|\mathcal{A} - \mathrm{P}_{\mathcal{M}_{\mathbf{r}}}^{\mathrm{TTSVD}}(\mathcal{A})\|_{\mathrm{F}} + \|\mathrm{P}_{\mathcal{M}_{\mathbf{r}}}^{\mathrm{TTSVD}}(\mathcal{A}) - \mathcal{A}\|_{\mathrm{F}} \\
&\leq \|\mathrm{P}_{\mathcal{N}_{\mathbf{r}}}(\mathcal{A}) - \mathcal{A}\|_{\mathrm{F}} + 2\sqrt{d-1}\|\mathrm{P}_{\mathcal{M}_{\mathbf{r}}}(\mathcal{A}) - \mathcal{A}\|_{\mathrm{F}} \qquad (5\text{-}4) \\
&\leq (2\sqrt{d-1}+1)\|\mathrm{P}_{\mathcal{N}_{\mathbf{r}}}(\mathcal{A}) - \mathcal{A}\|_{\mathrm{F}}, \qquad (5\text{-}5)
\end{aligned}
$$

其中, 不等式 (5-3) 利用了 $\mathrm{P}_{\mathcal{B}_1}$ 是到单位球的度量投影且 $\mathrm{P}_{\mathcal{N}_{\mathbf{r}}}(\mathcal{A}) \in \mathcal{B}_1$；不等式 (5-4) 源于 TT 形式下 TT-SVD 的拟优性估计 (1-17)；而不等式 (5-5) 则利用了包含关系 $\mathcal{N}_{\mathbf{r}} \subseteq \mathcal{M}_{\mathbf{r}}$. □

5.2.3　流形结构

事实上, NTT 分解生成了一类具有流形结构的低秩张量集合. 具体而言, 给定整数数组 $\mathbf{r} = (1, r_1, r_2, \ldots, r_{d-1}, 1)$, 秩-$\mathbf{r}$ 的 NTT 张量构成的集合

$$
\mathcal{N}_{\mathbf{r}} = \left\{ \mathcal{X} \in \mathbb{C}^{n_1 \times n_2 \times \cdots \times n_d} : \mathrm{rank}_{\mathrm{TT}}(\mathcal{X}) = \mathbf{r},\ \|\mathcal{X}\|_{\mathrm{F}} = 1 \right\}
$$

是 $\mathbb{C}^{n_1 \times n_2 \times \cdots \times n_d}$ 中的一个复子流形. 注意到

$$
\mathcal{N}_{\mathbf{r}} = \mathcal{M}_{\mathbf{r}} \cap \mathcal{B}_1
$$

是两个流形的交集: 即固定秩 TT 张量流形 $\mathcal{M}_{\mathbf{r}}$(见式 (1-15)) 与单位球面 $\mathcal{B}_1$ 的交. 我们观察到, 流形 $\mathcal{M}_{\mathbf{r}}$ 与 $\mathcal{B}_1$ 是横截相交的, 即对任意 $\mathcal{X} \in \mathcal{N}_{\mathbf{r}}$, 都有 $\mathrm{T}_{\mathcal{X}}\mathcal{M}_{\mathbf{r}} + \mathrm{T}_{\mathcal{X}}\mathcal{B}_1 = \mathbb{C}^{n_1 \times n_2 \times \cdots \times n_d}$. 因此, 由 [169, Theorem 6.30] 可知, $\mathcal{N}_{\mathbf{r}}$ 是一个光滑流形.

具体地, 回顾 $\mathcal{M}_{\mathbf{r}}$ 在点 $\mathcal{X} = [\![\mathcal{U}_1, \mathcal{U}_2, \ldots, \mathcal{U}_d]\!] \in \mathcal{M}_{\mathbf{r}}$ 处的切空间参数化 (见式 (1-16)):

$$
\mathrm{T}_{\mathcal{X}}\mathcal{M}_{\mathbf{r}} = \left\{ \begin{array}{l} [\![\dot{\mathcal{U}}_1, \mathcal{U}_2, \mathcal{U}_3, \ldots, \mathcal{U}_d]\!] \\ +\,[\![\mathcal{U}_1, \dot{\mathcal{U}}_2, \mathcal{U}_3, \ldots, \mathcal{U}_d]\!] \\ \vdots \\ +\,[\![\mathcal{U}_1, \mathcal{U}_2, \ldots, \mathcal{U}_{d-1}, \dot{\mathcal{U}}_d]\!] \end{array} : \begin{array}{c} \dot{\mathcal{U}}_k \in \mathbb{C}^{r_{k-1} \times n_k \times r_k}, k \in [d], \\ \mathrm{L}(\dot{\mathcal{U}}_k)^{\dagger}\mathrm{L}(\mathcal{U}_k) = 0, k \in [d-1] \end{array} \right\}.
$$

$\mathcal{B}_1$ 在 $\mathcal{X} \in \mathcal{B}_1$ 处的切空间可以表示为

$$
\mathrm{T}_{\mathcal{X}}\mathcal{B}_1 = \{\mathcal{V} \in \mathbb{C}^{n_1 \times n_2 \times \cdots \times n_d} : \langle \mathcal{V}, \mathcal{X} \rangle = 0\}. \qquad (5\text{-}6)
$$

由此可得 $(\mathrm{T}_{\mathcal{X}}\mathcal{B}_1)^{\perp} = \{t\mathcal{X} : t \in \mathbb{C}\}$. 由于在式 (1-16) 中我们并未对 $\mathrm{L}(\dot{\mathcal{U}}_d)$ 施加正交性约束, 因此, 令 $\dot{\mathcal{U}}_k = 0 (k = 1, 2, \ldots, d-1)$ 以及 $\dot{\mathcal{U}}_d = t\mathcal{U}_d$, 即可得到 $t\mathcal{X} \in \mathrm{T}_{\mathcal{X}}\mathcal{M}_{\mathbf{r}}$. 因此有

$$
\mathbb{C}^{n_1 \times n_2 \times \cdots \times n_d} = (\mathrm{T}_{\mathcal{X}}\mathcal{B}_1)^{\perp} + \mathrm{T}_{\mathcal{X}}\mathcal{B}_1 \subseteq \mathrm{T}_{\mathcal{X}}\mathcal{M}_{\mathbf{r}} + \mathrm{T}_{\mathcal{X}}\mathcal{B}_1 \subseteq \mathbb{C}^{n_1 \times n_2 \times \cdots \times n_d},
$$

从而 $\mathcal{N}_{\mathbf{r}}$ 是一个光滑流形.

5.2.4 NTT 张量的黎曼几何

在本节中我们将构建流形 $\mathcal{N}_{\mathbf{r}}$ 上的黎曼几何结构, 包括切空间、黎曼度量、到切空间的投影以及收缩映射；其整体流程示意见图 5-3. 具体而言, 在 $\mathcal{N}_{\mathbf{r}}$ 上进行搜索可以分为两个步骤. 1) 投影到切空间: 给定一点 $\mathcal{X} \in \mathcal{N}_{\mathbf{r}}$ 以及一个方向 $\mathcal{A} \in \mathbb{C}^{n_1 \times n_2 \times \cdots \times n_d}$, 我们将 $\mathcal{A}$ 投影到切空间 $\mathrm{T}_{\mathcal{X}}\mathcal{N}_{\mathbf{r}}$ 上, 并得到 $\mathcal{V}$. 该过程可以理解为先后对 $\mathrm{T}_{\mathcal{X}}\mathcal{M}_{\mathbf{r}}$ 和 $\mathrm{T}_{\mathcal{X}}\mathcal{B}_1$ 进行逐步投影; 2) 沿流形移动: 给定步长 $s > 0$, 我们将 $(\mathcal{X} + s\mathcal{V})$ 通过 NTT-SVD 算法投影到 $\mathcal{N}_{\mathbf{r}}$ 上, 其中我们依次执行 TT-SVD 算法以及归一化操作.

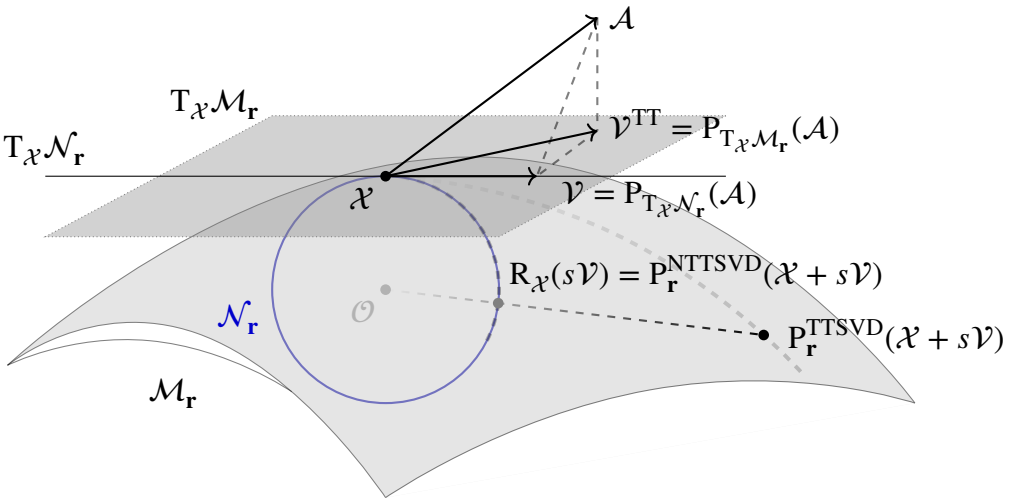


图 5-3 流形 $\mathcal{N}_{\mathbf{r}}$ 上几何的示意图. $\mathcal{O} \in \mathbb{C}^{n_1 \times n_2 \times \cdots \times n_d}$: 零张量.

Figure 5-3 An illustration of the geometry of $\mathcal{N}_{\mathbf{r}}$. $\mathcal{O} \in \mathbb{C}^{n_1 \times n_2 \times \cdots \times n_d}$: zero tensor.

切空间 由于 $\mathcal{M}_{\mathbf{r}}$ 与 $\mathcal{B}_1$ 横截相交 (intersect transversally), 由 [169] 可知, $\mathcal{N}_{\mathbf{r}}$ 在点 $\mathcal{X}$ 处的切空间可以表示为切空间 $\mathrm{T}_{\mathcal{X}}\mathcal{M}_{\mathbf{r}}$ 与 $\mathrm{T}_{\mathcal{X}}\mathcal{B}_1$ 的交, 即 $\mathrm{T}_{\mathcal{X}}\mathcal{N}_{\mathbf{r}} = \mathrm{T}_{\mathcal{X}}\mathcal{M}_{\mathbf{r}} \cap \mathrm{T}_{\mathcal{X}}\mathcal{B}_1$. 因此, 我们可以得到如下的切空间参数化表示.

命题 5.2 (切空间)**.** 给定左正交的 $\mathcal{X} = [\![\mathcal{U}_1, \mathcal{U}_2, \ldots, \mathcal{U}_d]\!] \in \mathcal{N}_{\mathbf{r}}$, $\mathcal{N}_{\mathbf{r}}$ 在 $\mathcal{X}$ 处的切空间可以参数化为

$$\mathrm{T}_{\mathcal{X}}\mathcal{N}_{\mathbf{r}} = \left\{ \begin{array}{l} [\![\dot{\mathcal{U}}_1, \mathcal{U}_2, \mathcal{U}_3, \ldots, \mathcal{U}_d]\!] \\ + [\![\mathcal{U}_1, \dot{\mathcal{U}}_2, \mathcal{U}_3, \ldots, \mathcal{U}_d]\!] \\ \vdots \\ + [\![\mathcal{U}_1, \mathcal{U}_2, \ldots, \mathcal{U}_{d-1}, \dot{\mathcal{U}}_d]\!] \end{array} \right. : \begin{array}{r} \dot{\mathcal{U}}_k \in \mathbb{C}^{r_{k-1} \times n_k \times r_k}, k \in [d], \\ \mathrm{L}(\dot{\mathcal{U}}_k)^{\dagger}\mathrm{L}(\mathcal{U}_k) = 0, k \in [d-1], \\ \langle \dot{\mathcal{U}}_d, \mathcal{U}_d \rangle = 0 \end{array} \left. \right\}. \quad (5\text{-}7)$$

证明. 将右端集合记为 T. 一方面, 对任意 $\mathcal{V} \in T$, 显然有 $\mathcal{V} \in \mathrm{T}_{\mathcal{X}}\mathcal{M}_{\mathbf{r}}$, 且

$$\begin{aligned}\langle \mathcal{V}, \mathcal{X} \rangle &= \langle [\![\dot{\mathcal{U}}_1, \mathcal{U}_2, \dots, \mathcal{U}_d]\!], [\![\mathcal{U}_1, \mathcal{U}_2, \dots, \mathcal{U}_d]\!] \rangle \\ &\quad + \langle [\![\mathcal{U}_1, \dot{\mathcal{U}}_2, \mathcal{U}_3, \dots, \mathcal{U}_d]\!], [\![\mathcal{U}_1, \mathcal{U}_2, \dots, \mathcal{U}_d]\!] \rangle \\ &\quad + \cdots + \langle [\![\mathcal{U}_1, \mathcal{U}_2, \dots, \mathcal{U}_{d-1}, \dot{\mathcal{U}}_d]\!], [\![\mathcal{U}_1, \mathcal{U}_2, \dots, \mathcal{U}_d]\!] \rangle \\ &= \langle [\![\mathcal{U}_1, \mathcal{U}_2, \dots, \mathcal{U}_{d-1}, \dot{\mathcal{U}}_d]\!], [\![\mathcal{U}_1, \mathcal{U}_2, \dots, \mathcal{U}_d]\!] \rangle \\ &= \langle \dot{\mathcal{U}}_d, \mathcal{U}_d \rangle = 0,\end{aligned}$$

其中等式成立是由于对所有 $k \in [d-1]$ 满足正交条件 $\mathrm{L}(\dot{\mathcal{U}}_k)^\dagger \mathrm{L}(\mathcal{U}_k) = 0$ 以及 $\langle \dot{\mathcal{U}}_d, \mathcal{U}_d \rangle = 0$. 因此, 由 (5-6) 可知 $\mathcal{V} \in \mathrm{T}_{\mathcal{X}}\mathcal{B}_1$, 从而 $T \subseteq \mathrm{T}_{\mathcal{X}}\mathcal{M}_{\mathbf{r}} \cap \mathrm{T}_{\mathcal{X}}\mathcal{B}_1 = \mathrm{T}_{\mathcal{X}}\mathcal{N}_{\mathbf{r}}$.

另一方面, 对任意切向量 $\mathcal{V} \in \mathrm{T}_{\mathcal{X}}\mathcal{N}_{\mathbf{r}} \subseteq \mathrm{T}_{\mathcal{X}}\mathcal{M}_{\mathbf{r}}$, 存在 $\dot{\mathcal{X}}_k \in \mathbb{C}^{r_{k-1} \times n_k \times r_k}$ 使得 $\mathcal{V} = [\![\dot{\mathcal{U}}_1, \mathcal{U}_2, \mathcal{U}_3, \dots, \mathcal{U}_d]\!] + [\![\mathcal{U}_1, \dot{\mathcal{U}}_2, \mathcal{U}_3, \dots, \mathcal{U}_d]\!] + \cdots + [\![\mathcal{U}_1, \mathcal{U}_2, \dots, \mathcal{U}_{d-1}, \dot{\mathcal{U}}_d]\!]$ 以及对 $k \in [d-1]$ 有 $\mathrm{L}(\dot{\mathcal{U}}_k)^\dagger \mathrm{L}(\mathcal{U}_k) = 0$. 又由于 $\mathcal{V} \in \mathrm{T}_{\mathcal{X}}\mathcal{N}_{\mathbf{r}} \subseteq \mathrm{T}_{\mathcal{X}}\mathcal{B}_1$, 可得 $\langle \mathcal{V}, \mathcal{X} \rangle = 0$, 即 $\langle \dot{\mathcal{U}}_d, \mathcal{U}_d \rangle = 0$. 因此, $\mathcal{V} \in T$.

综上可得 $T = \mathrm{T}_{\mathcal{X}}\mathcal{N}_{\mathbf{r}}$. □

注意到 $\langle \dot{\mathcal{U}}_d, \mathcal{U}_d \rangle = 0$ 等价于 $\mathrm{L}(\dot{\mathcal{U}}_d)^\dagger \mathrm{L}(\mathcal{U}_d) = 0$. 因此, 式 (5-7) 中的参数 $\dot{\mathcal{U}}_d$ 满足正交条件, 而式 (1-16) 中的 $\dot{\mathcal{U}}_d$ 是任意的. 在实际计算中, 对于一个切向量, 我们只需存储参数 $\dot{\mathcal{U}}_1, \dot{\mathcal{U}}_2, \dots, \dot{\mathcal{U}}_d$ 即可.

往切空间上的投影　接下来, 我们计算一个张量到切空间上的投影. 我们采用内积 $\langle \cdot, \cdot \rangle$ 作为 $\mathcal{N}_{\mathbf{r}}$ 的黎曼度量.

命题 5.3. *给定左正交的 $\mathcal{X} = [\![\mathcal{U}_1, \mathcal{U}_2, \dots, \mathcal{U}_d]\!] \in \mathcal{N}_{\mathbf{r}}$, $\mathcal{A} \in \mathbb{C}^{n_1 \times n_2 \times \cdots \times n_d}$ 到切空间 $\mathrm{T}_{\mathcal{X}}\mathcal{N}_{\mathbf{r}}$ 上的投影可以表示为一个切向量 $\mathrm{P}_{\mathrm{T}_{\mathcal{X}}\mathcal{N}_{\mathbf{r}}}(\mathcal{A}) \in \mathrm{T}_{\mathcal{X}}\mathcal{N}_{\mathbf{r}}$ 其参数 $\mathcal{W}_k \in \mathbb{C}^{r_{k-1} \times n_k \times r_k}$ 满足*

$$\begin{aligned}\mathrm{L}(\mathcal{W}_k) &= (\mathbf{I}_{r_{k-1}n_k} - \mathbf{P}_k)(\mathbf{I}_{n_k} \otimes \mathbf{X}_{\leq k-1})^\dagger \mathbf{A}_{\langle k \rangle} \operatorname{conj}(\mathbf{X}_{\geq k+1})(\mathbf{X}_{\geq k+1}^\top \operatorname{conj}(\mathbf{X}_{\geq k+1}))^{-1} \\ \operatorname{vec}(\mathcal{W}_d) &= (\mathbf{I}_{r_{d-1}n_d} - \mathbf{P}_d)(\mathbf{I}_{n_d} \otimes \mathbf{X}_{\leq d-1})^\dagger \operatorname{vec}(\mathbf{A})\end{aligned}$$

其中 $k \in [d-1]$, 并且 $\mathbf{P}_k = \mathrm{L}(\mathcal{U}_k)\mathrm{L}(\mathcal{U}_k)^\dagger$ 是到 $\mathrm{L}(\mathcal{U}_k)$ 的像空间上的正交投影算子.

证明. 由于 $\mathrm{P}_{\mathrm{T}_{\mathcal{X}}\mathcal{N}_{\mathbf{r}}}(\mathcal{A}) \in \mathrm{T}_{\mathcal{X}}\mathcal{N}_{\mathbf{r}}$, 它可以按照 (5-7) 的形式进行参数化, 其参数为 $\mathcal{W}_k \in \mathbb{C}^{r_{k-1} \times n_k \times r_k}$, $k = 1, 2, \dots, d$. 我们的目标是确定这些参数.

我们回顾参数化表示 (5-7), 并将每一项记为 $\mathcal{V}_k$, 即 $\mathcal{V} = \mathcal{V}_1 + \mathcal{V}_2 + \cdots + \mathcal{V}_d$. 我们注意到, 对任意切向量 $\mathcal{V} \in \mathrm{T}_{\mathcal{X}}\mathcal{N}_{\mathbf{r}}$ 都有 $\langle \mathcal{A} - \mathrm{P}_{\mathrm{T}_{\mathcal{X}}\mathcal{N}_{\mathbf{r}}}(\mathcal{A}), \mathcal{V} \rangle = 0$. 于是可得

$$\begin{aligned}&\langle \mathcal{A} - \mathrm{P}_{\mathrm{T}_{\mathcal{X}}\mathcal{N}_{\mathbf{r}}}(\mathcal{A}), \mathcal{V}_k \rangle = \langle \mathcal{A} - \mathrm{P}_{\mathrm{T}_{\mathcal{X}}\mathcal{N}_{\mathbf{r}}}(\mathcal{A}), [\![\mathcal{U}_1, \dots, \mathcal{U}_{k-1}, \dot{\mathcal{X}}_k, \mathcal{U}_{k+1}, \dots, \mathcal{U}_d]\!] \rangle \\ &= \langle \mathbf{A}_{\langle k \rangle} - (\mathbf{I}_{n_k} \otimes \mathbf{X}_{\leq k-1})\mathrm{L}(\mathcal{W}_k)\mathbf{X}_{\geq k+1}^\top, (\mathbf{I}_{n_k} \otimes \mathbf{X}_{\leq k-1})\mathrm{L}(\dot{\mathcal{X}}_k)\mathbf{X}_{\geq k+1}^\top \rangle \\ &= \langle (\mathbf{I}_{n_k} \otimes \mathbf{X}_{\leq k-1})^\dagger \mathbf{A}_{\langle k \rangle} \operatorname{conj}(\mathbf{X}_{\geq k+1}) - \mathrm{L}(\mathcal{W}_k)\mathbf{X}_{\geq k+1}^\top \operatorname{conj}(\mathbf{X}_{\geq k+1}), \mathrm{L}(\dot{\mathcal{X}}_k) \rangle \\ &= 0\end{aligned}$$

对任意满足 $\mathrm{L}(\dot{\mathcal{X}}_k)^\dagger \mathrm{L}(\mathcal{U}_k) = 0$ 的 $k \in [d]$ 成立. 这里我们使用了接口矩阵的正交性 (1-7), 以及 $\mathbf{X}_{\leq k-1} = (\mathcal{V}_k)_{\leq k-1}$ 和 $\mathbf{X}_{\geq k+1} = (\mathcal{V}_k)_{\geq k+1}$. 因此, 我们得到

$$\mathrm{L}(\mathcal{W}_k) = (\mathbf{I}_{r_{k-1}n_k} - \mathbf{P}_k)(\mathbf{I}_{n_k} \otimes \mathbf{X}_{\leq k-1})^\dagger \mathbf{A}_{\langle k \rangle} \operatorname{conj}(\mathbf{X}_{\geq k+1})(\mathbf{X}_{\geq k+1}^\top \operatorname{conj}(\mathbf{X}_{\geq k+1}))^{-1}$$

对 $k = 1, 2, \ldots, d-1$ 成立, 并且

$$\operatorname{vec}(\mathcal{W}_d) = (\mathbf{I}_{r_{d-1}n_d} - \mathbf{P}_d)(\mathbf{I}_{n_d} \otimes \mathbf{X}_{\leq d-1})^\dagger \operatorname{vec}(\mathbf{A}).$$

□

由命题 5.3 的证明可知, 到切空间 $\mathrm{T}_{\mathcal{X}}\mathcal{N}_{\mathbf{r}}$ 的投影可以表示为到切空间 $\mathrm{T}_{\mathcal{X}}\mathcal{M}_{\mathbf{r}}$ 以及单位球切空间 $\mathrm{P}_{\mathrm{T}_{\mathcal{X}}\mathcal{B}_1}$ 的投影的复合.

推论 5.4. 我们有

$$\mathrm{P}_{\mathrm{T}_{\mathcal{X}}\mathcal{N}_{\mathbf{r}}} = \mathrm{P}_{\mathrm{T}_{\mathcal{X}}\mathcal{B}_1} \circ \mathrm{P}_{\mathrm{T}_{\mathcal{X}}\mathcal{M}_{\mathbf{r}}} = \mathrm{P}_{\mathrm{T}_{\mathcal{X}}\mathcal{M}_{\mathbf{r}}} \circ \mathrm{P}_{\mathrm{T}_{\mathcal{X}}\mathcal{B}_1}.$$

在实际计算中, 我们注意到参数 $\mathcal{W}_k$ 中包含 $(\mathbf{X}_{\geq k+1}^\top \operatorname{conj}(\mathbf{X}_{\geq k+1}))^{-1}$, 该项可能是接近奇异的. 因此, 受 [60, §3.3] 启发, 我们考虑如下表示:

$$\mathcal{X} = [\![\mathcal{U}_1, \ldots, \mathcal{U}_{k-1}, \tilde{\mathcal{X}}_k, \mathcal{Y}_{k+1}, \ldots, \mathcal{Y}_d]\!],$$

其中参数是 k-正交的, 即 $\mathcal{U}_1, \ldots, \mathcal{U}_{k-1}$ 是左正交的, $\mathcal{Y}_{k+1}, \ldots, \mathcal{Y}_d$ 是右正交的, 而 $\tilde{\mathcal{X}}_k$ 不要求是左正交或右正交的. 随后, 每一个分量 $\mathcal{V}_k$ 都可以进行 k-正交化写成 $\mathcal{V}_k = [\![\mathcal{U}_1, \ldots, \mathcal{U}_{k-1}, \tilde{\mathcal{W}}_k, \mathcal{Y}_{k+1}, \ldots, \mathcal{Y}_d]\!]$. 于是, 我们得到 $\mathrm{P}_{\mathrm{T}_{\mathcal{X}}\mathcal{N}_{\mathbf{r}}}(\mathcal{A}) \in \mathrm{T}_{\mathcal{X}}\mathcal{N}_{\mathbf{r}}$ 的一个等价表示:

$$\mathrm{P}_{\mathrm{T}_{\mathcal{X}}\mathcal{N}_{\mathbf{r}}}(\mathcal{A}) = \sum_{k=1}^{d} [\![\mathcal{U}_1, \ldots, \mathcal{U}_{k-1}, \tilde{\mathcal{W}}_k, \mathcal{Y}_{k+1}, \ldots, \mathcal{Y}_d]\!] \tag{5-8}$$

其中 $\mathrm{L}(\tilde{\mathcal{W}}_k) = (\mathbf{I}_{r_{k-1}n_k} - \mathbf{P}_k)(\mathbf{I}_{n_k} \otimes \mathbf{X}_{\leq k-1})^\dagger \mathbf{A}_{\langle k \rangle} \operatorname{conj}(\mathbf{X}_{\geq k+1})$. 注意到由于 k-正交性, 有 $\mathbf{X}_{\geq k+1} \in \mathrm{St}_{\mathbb{C}}(r_k, n_{k+1}n_{k+1} \cdots n_d)$.

收缩映射 为了在流形 $\mathcal{N}_{\mathbf{r}}$ 上进行线搜索, 我们需要一个收缩映射. 回顾1.4节, 若映射 $\mathrm{R} : \mathrm{T}\mathcal{N}_{\mathbf{r}} \to \mathcal{N}_{\mathbf{r}}$ 在 $\mathcal{X} \in \mathcal{N}_{\mathbf{r}}$ 的邻域内满足以下条件, 则称其为 $\mathcal{N}_{\mathbf{r}}$ 上的一个收缩映射: 1) 存在 $(\mathcal{X}, 0) \in \mathrm{T}\mathcal{N}_{\mathbf{r}}$ 的一个邻域 U 使得 $U \subseteq \operatorname{dom}(\mathrm{R})$ 且 R 在 U 上光滑; 2) $\mathrm{R}_{\mathcal{X}}(0) = \mathcal{X}$ 对所有的 $\mathcal{X} \in \mathcal{N}_{\mathbf{r}}$ 成立; 3) $\mathrm{DR}_{\mathcal{X}}(\cdot)[0] = \mathrm{id}_{\mathrm{T}_{\mathcal{X}}\mathcal{N}_{\mathbf{r}}}$, 即等于切空间上的恒等映射.

命题 5.5. 给定 $\mathcal{X} \in \mathcal{N}_{\mathbf{r}}$ 以及切向量 $\mathcal{V} \in \mathrm{T}_{\mathcal{X}}\mathcal{N}_{\mathbf{r}}$, 映射 $\mathrm{R}_{\mathcal{X}}(\mathcal{V}) = \mathrm{P}_{\mathbf{r}}^{\mathrm{NTTSVD}}(\mathcal{X} + \mathcal{V})$ 是一个收缩映射.

证明. 我们只需验证上述三个性质. 对于第一条性质, 由 [60, Proposition 4] 可知, 存在 $\mathcal{X}$ 的一个邻域 $U \subseteq \mathbb{C}^{n_1 \times n_2 \times \cdots \times n_d}$ 使得 $0 \notin U$ 且 $\mathrm{P}_{\mathbf{r}}^{\mathrm{TTSVD}}$ 在该邻域内是光滑的. 由于 $\mathrm{P}_{\mathcal{B}_1}$ 在 U 上同样是光滑的, 并且 $\mathrm{R}_{\mathcal{X}}(\mathcal{V}) = \mathrm{P}_{\mathcal{B}_1}(\mathrm{P}_{\mathbf{r}}^{\mathrm{TTSVD}}(\mathcal{X}+\mathcal{V}))$, 因此映射 R 在 $(\mathcal{X}, 0) \in \mathcal{N}_{\mathbf{r}} \times \mathrm{T}_{\mathcal{X}}\mathcal{N}_{\mathbf{r}}$ 的邻域内是光滑的. 第二条性质是显然成立的.

我们通过拟优性 (5-2) 来证明第三条性质. 由于度量投影 $\mathrm{P}_{\mathcal{N}_{\mathbf{r}}}$ 是 $\mathcal{N}_{\mathbf{r}}$ 上的一个收缩映射, 我们有

$$\|(\mathcal{X}+t\mathcal{V}) - \mathrm{R}_{\mathcal{X}}(t\mathcal{V})\|_{\mathrm{F}} \leq (2\sqrt{d-1}+1)\|(\mathcal{X}+t\mathcal{V}) - \mathrm{P}_{\mathcal{N}_{\mathbf{r}}}(\mathcal{X}+t\mathcal{V})\|_{\mathrm{F}} = \mathcal{O}(t^2).$$

因此有 $\mathrm{R}_{\mathcal{X}}(t\mathcal{V}) = \mathcal{X} + t\mathcal{V} + \mathcal{O}(t^2)$, 即 $\mathrm{DR}_{\mathcal{X}}(\cdot)[0] = \mathrm{id}_{\mathrm{T}_{\mathcal{X}}\mathcal{N}_{\mathbf{r}}}$. 由此可知, R 在 $\mathcal{X} \in \mathcal{N}_{\mathbf{r}}$ 附近定义了一个收缩映射. □

5.2.5　几何方法

基于固定秩 NTT 张量所构成的集合 $\mathcal{N}_{\mathbf{r}}$ 的几何结构, 我们考虑如下在光滑流形 $\mathcal{N}_{\mathbf{r}}$ 上的优化问题:

$$\min\ f(\mathcal{X}), \quad \text{s.t. } \mathcal{X} \in \mathcal{N}_{\mathbf{r}} = \mathcal{M}_{\mathbf{r}} \cap \mathcal{B}_1, \tag{5-9}$$

其中 $f : \mathbb{C}^{n_1 \times n_2 \times \cdots \times n_d}$ 是光滑函数.

算法 18 针对 (5-9) 的黎曼共轭梯度方法 (NTT-RCG).

输入: 初始点 $\mathcal{X}^{(0)} \in \mathcal{N}_{\mathbf{r}}$, $t = 0$, $\beta^{(0)} = 0$.

1: **while** 停机准则未被满足 **do**
2: 通过 (5-8) 计算切向量 $\mathcal{V}^{(t)} = \mathrm{P}_{\mathrm{T}_{\mathcal{X}^{(t)}}\mathcal{M}}(-\nabla f(\mathcal{X}^{(t)})) + \beta^{(t)}\mathcal{T}_{\mathcal{X}^{(t)} \leftarrow \mathcal{X}^{(t-1)}}\mathcal{V}^{(t-1)}$ 的参数 $\tilde{\mathcal{W}}_k^{(t)}$.
3: 选择步长 $s^{(t)}$.
4: 通过图 5-2更新 $\mathcal{X}^{(t+1)} = \mathrm{P}_{\mathbf{r}}^{\mathrm{NTTSVD}}(\mathcal{X}^{(t)} + s^{(t)}\mathcal{V}^{(t)})$; $t = t+1$.
5: **end while**

输出: $\mathcal{X}^{(t)}$.

我们采用黎曼共轭梯度方法来求解 (5-9)；见算法 18. 我们将向量传输算子 $\mathcal{T}_{\mathcal{X}^{(t)} \leftarrow \mathcal{X}^{(t-1)}}\mathcal{V}^{(t-1)}$ 设置为正交投影 (5-8). 因此, $\mathcal{V}^{(t)}$ 的参数 $\tilde{\mathcal{W}}_k^{(t)}$ 可以通过将 $\mathrm{P}_{\mathrm{T}_{\mathcal{X}^{(t)}}\mathcal{M}}(-\nabla f(\mathcal{X}^{(t)}))$ 的参数与 $\beta^{(t)}\mathrm{P}_{\mathrm{T}_{\mathcal{X}^{(t)}}\mathcal{M}}(\mathcal{V}^{(t-1)})$ 的参数按照 (5-8) 计算出来以后相加得到. 给定 $\mathcal{V} = \mathrm{P}_{\mathrm{T}_{\mathcal{X}}\mathcal{N}_{\mathbf{r}}}(-\nabla f(\mathcal{X}))$ 在 (5-8) 中的参数化表达, 张量 $\mathcal{X}+\mathcal{V}$ 可以用 TT 形式的张量表示, 其中 $(i_1, i_2, \ldots, i_d)$ 元素为:

$$\begin{bmatrix} \tilde{\mathbf{W}}_1(i_1) & \mathbf{U}_1(i_1) \end{bmatrix} \begin{bmatrix} \mathbf{Y}_2(i_2) & 0 \\ \tilde{\mathbf{W}}_2(i_2) & \mathbf{U}_2(i_2) \end{bmatrix} \begin{bmatrix} \mathbf{Y}_3(i_3) & 0 \\ \tilde{\mathbf{W}}_3(i_3) & \mathbf{U}_3(i_3) \end{bmatrix} \cdots \begin{bmatrix} \mathbf{Y}_{d-1}(i_{d-1}) & 0 \\ \tilde{\mathbf{W}}_{d-1}(i_{d-1}) & \mathbf{U}_{d-1}(i_{d-1}) \end{bmatrix} \begin{bmatrix} \mathbf{Y}_d(i_d) \\ \mathbf{U}_d(i_d) + \tilde{\mathbf{W}}_d(i_d) \end{bmatrix},$$

其中 $\mathcal{X} = [\![\mathcal{U}_1, \mathcal{U}_2, \ldots, \mathcal{U}_d]\!]$ 和 $\mathcal{X} = [\![\mathcal{Y}_1, \mathcal{Y}_2, \ldots, \mathcal{Y}_d]\!]$ 是 $\mathcal{X}$ 的两个等价 NTT 分解, 分别具有左正交和右正交的核张量. 随后, NTT-SVD 算法可以被高效实现.

注. 在实际应用中, 如果 $\nabla f(\mathcal{X})$ 是稀疏张量或者可以用低秩 TT 张量表示, 则投影梯度 $\mathrm{P}_{\mathrm{T}_{\mathcal{X}}\mathcal{N}_{\mathbf{r}}}(\nabla f(\mathcal{X}))$ 可以像 [60, §4.2] 中的方法一样高效计算. 例如, 特征值问题中目标函数 f 的欧氏梯度可以用 TT 形式的张量表示, 因此投影梯度可以高效计算; 详见5.3.2节. 对于量子信息论中的应用, 目标函数 f 可以高效计算, 但欧氏梯度 ∇f 是完整张量. 因此, 我们采用有限差分的方法来近似投影梯度: 1) 生成 $\mathrm{T}_{\mathcal{X}}\mathcal{N}_{\mathbf{r}}$ 的正交基 $\mathcal{V}_1, \mathcal{V}_2, \ldots, \mathcal{V}_{\dim(\mathcal{N}_{\mathbf{r}})}$; 2) 通过下式近似投影梯度

$$\mathrm{P}_{\mathrm{T}_{\mathcal{X}}\mathcal{N}_{\mathbf{r}}}(\nabla f(\mathcal{X})) \approx \sum_{k=1}^{\dim(\mathcal{N}_{\mathbf{r}})} \alpha_k \mathcal{V}_k \quad \text{其中} \quad \alpha_k = \frac{f(\mathrm{P}_{\mathbf{r}}^{\mathrm{NTTSVD}}(\mathcal{X} + t\mathcal{V}_k)) - f(\mathcal{X})}{t},$$

这里我们选择足够小的 t.

5.3 在科学计算中的应用

在本节中, 我们将 NTT 分解应用于科学计算中的两个问题: 低秩张量的恢复以及高维特征值问题.

我们首先介绍实验中的默认设置. 张量链分解相关的计算基于 TTeMPS 工具箱[1]完成, 而提出的 NTT-RCG 方法在 Manopt 工具箱[2] v7.1.0 中实现, 这是一个用于几何方法的 Matlab 库. 所有实验均在配备两颗 Intel(R) Xeon(R) Gold 6330 处理器 (2.00GHz × 28 核心, 42M 缓存) 和 512GB 内存、运行 Ubuntu 22.04.3 的工作站上使用 Matlab R2019b 完成. 所提方法的代码可在 https://github.com/JimmyPeng1998/GeomNTT 获取.

5.3.1 低秩张量恢复

给定部分观测的张量 $A \in \mathcal{N}_{\mathbf{r}}$, 其观测索引集为 $\Omega \subseteq [n_1] \times [n_2] \times \cdots \times [n_d]$, 我们的目标是通过求解以下优化问题, 基于其在 Ω 上的元素恢复完整的张量 $\mathcal{A}$:

$$\begin{aligned} \min \quad & f(\mathcal{X}) = \frac{1}{2} \| \mathrm{P}_{\Omega}(\mathcal{X}) - \mathrm{P}_{\Omega}(\mathcal{A}) \|_{\mathrm{F}}^2 \\ \text{s.t.} \quad & \mathcal{X} \in \mathcal{N}_{\mathbf{r}}, \end{aligned} \tag{5-10}$$

其中 P_{Ω} 定义为: 若 $(i_1, i_2, \ldots, i_d) \in \Omega$, 则 $\mathrm{P}_{\Omega}(\mathcal{A})(i_1, i_2, \ldots, i_d) = \mathcal{A}(i_1, i_2, \ldots, i_d)$; 否则 $\mathrm{P}_{\Omega}(\mathcal{A})(i_1, i_2, \ldots, i_d) = 0$. NTT-RCG 的数值性能通过训练误差 $\| \mathrm{P}_{\Omega}(\mathcal{X}) - \mathrm{P}_{\Omega}(\mathcal{A}) \|_{\mathrm{F}} / \| \mathrm{P}_{\Omega}(\mathcal{A}) \|_{\mathrm{F}}$ 和测试误差 $\| \mathrm{P}_{\Gamma}(\mathcal{X}) - \mathrm{P}_{\Gamma}(\mathcal{A}) \|_{\mathrm{F}} / \| \mathrm{P}_{\Gamma}(\mathcal{A}) \|_{\mathrm{F}}$ 来衡量, 其中 $\Gamma \subseteq [n_1] \times [n_2] \times \cdots \times [n_d]$ 为另一验证集.

无噪声数据测试 我们考虑无噪声情况, 即 $\mathcal{A} \in \mathcal{N}_{\mathbf{r}}$ 是低秩张量. 我们的目标是展示 NTT-RCG 方法在不同张量大小 n 和样本量 $|\Omega|$ 下恢复低秩张量的能力. 按照 [60, §5.3] 的设置, 我们取 $d = 5$, $\mathbf{r} = (1, 3, 3, 3, 3, 1)$, 张量大小 $n =$

[1]可在https://www.epfl.ch/labs/anchp/index-html/software/ttemps/中获取.

[2]可在https://www.manopt.org/获取

$50, 100, \ldots, 400$, 样本量 $|\Omega| = 2000, 4000, \ldots, 60000$. 对于每一组 $(n, |\Omega|)$, 我们运行 NTT-RCG 方法五次. 若在 250 次迭代内测试误差小于 10^{-4}, 则认为 NTT-RCG 方法成功恢复张量. 图 5-4(左) 展示了 NTT-RCG 方法的相图, 其中白色块表示五次运行全部成功恢复, 黑色块表示五次运行全部恢复失败, 红色曲线表示 $\mathcal{O}(n\log(n))$. 相图显示了与已有结果类似的恢复能力；参见例如 [60, §5.3].

有噪声数据测试　我们考虑有噪声情况, 即 $\mathcal{A} = \hat{\mathcal{A}} + \lambda\mathcal{E}/\|\mathcal{E}\|_{\mathrm{F}}$ 由单位范数的低秩张量 $\hat{\mathcal{A}} \in \mathcal{N}_{\mathbf{r}}$ 和噪声水平为 λ 的噪声张量 $\mathcal{E} \in \mathbb{R}^{n_1\times n_2\times\cdots\times n_d}$ 相加得到. 张量 $\mathcal{E}$ 的每个元素均从标准高斯分布 $N(0,1)$ 独立同分布采样. 我们设置 $\lambda = 10^{-4}, 10^{-6}, \ldots, 10^{-12}, 0$, $d = 3$, $n = 100$, $\mathbf{r} = (1, r_1, r_2, 1) = (1, 3, 3, 1)$, 并且 $|\Omega| = 10dnr_1^2$. 图 5-4(右) 展示了 NTT-RCG 方法的收敛结果. 可以观察到, NTT-RCG 方法在不同噪声水平下均能成功恢复潜在的低秩张量.

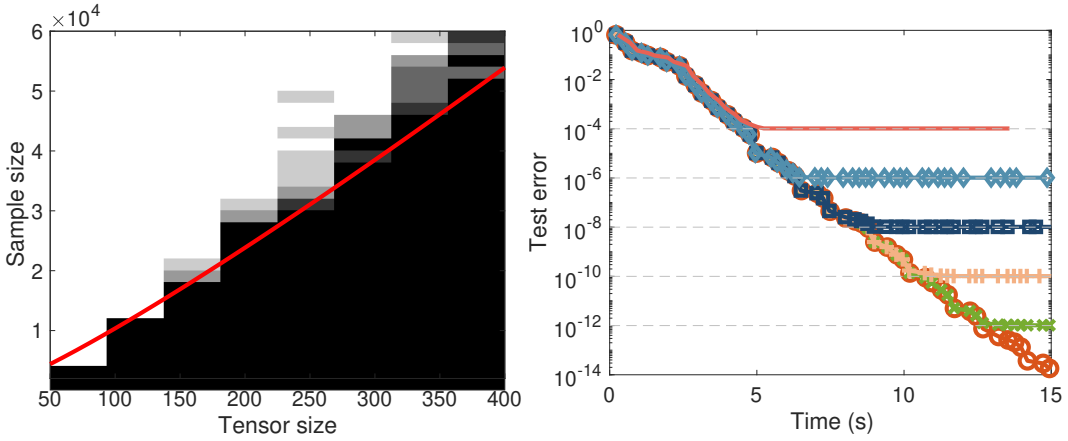


图 5-4 左图: 五次运行的恢复结果相位图. 白色方块表示五次运行均恢复成功, 黑色方块表示五次运行均恢复失败. 右图: 在噪声水平 $\lambda = 10^{-4}, 10^{-6}, \ldots, 10^{-12}, 0$ 下的测试误差.

Figure 5-4 Left: phase plot of recovery results for five runs. The white block indicates successful recovery in all five runs, while the black block indicates failure of recovery in all five runs. Right: test error under noise levels $\lambda = 10^{-4}, 10^{-6}, \ldots, 10^{-12}, 0$.

5.3.2　具有张量积结构的特征值问题

在本节中, 我们将 NTT 分解应用于具有张量积结构的特征值问题. 考虑如下张量积结构的最小 (或最大) 特征值问题:

$$\begin{aligned} \min_{\mathbf{x}}(\max_{\mathbf{x}}) \quad & f(\mathbf{x}) = \mathbf{x}^{\top}\mathbf{H}\mathbf{x} = \mathbf{x}^{\top}\left(\sum_{\ell=1}^{L}\mathbf{H}_{\ell,d}\otimes\mathbf{H}_{\ell,d-1}\otimes\cdots\otimes\mathbf{H}_{\ell,1}\right)\mathbf{x} \\ \text{s.t.} \quad & \mathbf{x}\in\mathbb{R}^{n_1n_2\cdots n_d},\ \|\mathbf{x}\|_2^2 = 1, \end{aligned} \tag{5-11}$$

其中 $\mathbf{H}_{\ell,k}\in\mathbb{R}^{n_k\times n_k}$, $\ell\in[L], k\in[d]$, 矩阵 $\mathbf{H}$ 可表示为 Kronecker 积的和. 一个典型例子是 d 维 Laplace 算子的离散化形式

$$\mathbf{H} = \mathbf{T}_{n_d}\otimes\mathbf{I}_{n_{d-1}}\otimes\cdots\otimes\mathbf{I}_{n_1} + \cdots + \mathbf{I}_{n_d}\otimes\mathbf{I}_{n_{d-1}}\otimes\cdots\otimes\mathbf{I}_{n_{d-1}}\otimes\mathbf{T}_{n_1}, \tag{5-12}$$

这里 $\mathbf{T}_n = \mathrm{tridiag}(-1, 2, -1) \in \mathbb{R}^{n\times n}$ 是一个三对角矩阵.

直接求解该问题是不可行的, 因为参数数量随 d 指数增长. 为此, 我们利用张量积结构, 采用归一化张量列 (NTT) 分解, 将优化问题限制在 $\mathcal{N}_{\mathbf{r}} \subseteq \mathbb{R}^{n_1\times n_2\times\cdots\times n_d}$ 上, 即

$$\begin{aligned} \min_{\mathcal{X}}(\max_{\mathcal{X}}) \quad & f(\mathcal{X}) = \mathrm{vec}(\mathcal{X})^\top \mathbf{H}\mathrm{vec}(\mathcal{X}) \\ \text{s.t.} \quad & \mathcal{X} = [\![\mathcal{U}_1, \mathcal{U}_2, \dots, \mathcal{U}_d]\!] \in \mathcal{N}_{\mathbf{r}}. \end{aligned} \tag{5-13}$$

为了高效计算 f, 可利用下述张量积性质.

命题 5.6. 给定 $\mathcal{X} = [\![\mathcal{U}_1, \dots, \mathcal{U}_d]\!]$ 与矩阵 $\mathbf{K}_k \in \mathbb{R}^{n_k\times n_k}$, 有

$$(\mathbf{K}_d \otimes \mathbf{K}_{d-1} \otimes \cdots \otimes \mathbf{K}_1)\mathrm{vec}(\mathcal{X}) = \mathrm{vec}([\![\mathcal{U}_1 \times_2 \mathbf{K}_1, \mathcal{U}_2 \times_2 \mathbf{K}_2, \dots, \mathcal{U}_d \times_2 \mathbf{K}_d]\!]).$$

证明. 我们首先从第一展平矩阵 $\mathbf{X}_{\langle 1\rangle}$ 开始推导, 有

$$\begin{aligned} (\mathbf{K}_d \otimes \cdots \otimes \mathbf{K}_1)\mathrm{vec}(\mathcal{X}) &= (\mathbf{K}_d \otimes \mathbf{K}_{d-1} \otimes \cdots \otimes \mathbf{K}_1)\mathrm{vec}(\mathrm{L}(\mathcal{U}_1)\mathbf{X}_{\geq 2}^\top) \\ &= (\mathbf{K}_d \otimes \mathbf{K}_{d-1} \otimes \cdots \otimes \mathbf{K}_1)(\mathbf{X}_{\geq 2} \otimes \mathbf{I}_{n_1})\mathrm{vec}(\mathcal{U}_1) \\ &= ((\mathbf{K}_d \otimes \mathbf{K}_{d-1} \otimes \cdots \otimes \mathbf{K}_2)\mathbf{X}_{\geq 2} \otimes \mathbf{K}_1)\mathrm{vec}(\mathcal{U}_1) \\ &= ((\mathbf{K}_d \otimes \mathbf{K}_{d-1} \otimes \cdots \otimes \mathbf{K}_2)\mathbf{X}_{\geq 2} \otimes \mathbf{I}_{n_1})(\mathbf{I}_{r_1} \otimes \mathbf{K}_1)\mathrm{vec}(\mathcal{U}_1) \\ &= (((\mathbf{K}_d \otimes \mathbf{K}_{d-1} \otimes \cdots \otimes \mathbf{K}_2)\mathbf{X}_{\geq 2}) \otimes \mathbf{I}_{n_1})\mathrm{vec}(\mathcal{U}_1 \times_2 \mathbf{K}_1), \end{aligned}$$

这里最后一个等号来自于 (1-3) 和 $(\mathbf{I}_{r_1} \otimes \mathbf{K}_1)\mathrm{vec}(\mathcal{U}_1) = (\mathbf{I}_{r_1} \otimes \mathbf{K}_1 \otimes \mathbf{I}_{r_0})\mathrm{vec}(\mathcal{U}_1) = \mathrm{vec}(\mathcal{U}_1 \times_2 \mathbf{K}_1)$. 接下来, 我们利用 (1-7) 和 (1-3), 可以得到

$$\begin{aligned} (\mathbf{K}_d \otimes \cdots \otimes \mathbf{K}_2)\mathbf{X}_{\geq 2} &= (\mathbf{K}_d \otimes \mathbf{K}_{d-1} \otimes \cdots \otimes \mathbf{K}_2)(\mathbf{X}_{\geq 3} \otimes \mathbf{I}_{n_2})\mathrm{R}(\mathcal{U}_2)^\top \\ &= (((\mathbf{K}_d \otimes \mathbf{K}_{d-1} \otimes \cdots \otimes \mathbf{K}_3)\mathbf{X}_{\geq 3}) \otimes \mathbf{I}_{n_3})(\mathbf{I}_{r_2} \otimes \mathbf{K}_2)\mathrm{R}(\mathcal{U}_2)^\top \\ &= (((\mathbf{K}_d \otimes \mathbf{K}_{d-1} \otimes \cdots \otimes \mathbf{K}_3)\mathbf{X}_{\geq 3}) \otimes \mathbf{I}_{n_3})\mathrm{R}(\mathcal{U}_2 \times_2 \mathbf{K}_2)^\top. \end{aligned}$$

我们发现对于 $k = 3, 4, \dots, d$, 张量积 $(\mathbf{K}_d \otimes \mathbf{K}_{d-1} \otimes \cdots \otimes \mathbf{K}_k)\mathbf{X}_{\geq k}$ 同样地可以被递归计算. 最终我们得到

$$\begin{aligned} & (\mathbf{K}_d \otimes \mathbf{K}_{d-1} \otimes \cdots \otimes \mathbf{K}_1)\mathrm{vec}(\mathcal{X}) \\ &= \left(((\mathrm{R}(\mathcal{U}_d \times_2 \mathbf{K}_d)^\top \otimes \mathbf{I}_{d-1})\mathrm{R}(\mathcal{U}_{d-1} \times_2 \mathbf{K}_{d-1})) \otimes \mathbf{I}_{d-2}) \cdots\right) \mathrm{vec}(\mathcal{U}_1) \\ &= \mathrm{vec}([\![\mathcal{U}_1 \times_2 \mathbf{K}_1, \mathcal{U}_2 \times_2 \mathbf{K}_2, \dots, \mathcal{U}_d \times_2 \mathbf{K}_d]\!]). \end{aligned}$$

□

因此, 式 (5-13) 中的目标函数 f 可以通过 $\mathcal{X}$ 的核张量 $\mathcal{U}_1, \mathcal{U}_2, \dots, \mathcal{U}_d$ 高效地计算, 即

$$f(\mathcal{X}) = \mathrm{vec}(\mathcal{X})^\top \mathbf{H}\mathrm{vec}(\mathcal{X}) = \sum_{\ell=1}^{L} \langle \mathcal{X}, [\![\mathcal{U}_1 \times_2 \mathbf{H}_{\ell,1}, \mathcal{U}_2 \times_2 \mathbf{H}_{\ell,2}, \dots, \mathcal{U}_d \times_2 \mathbf{H}_{\ell,d}]\!] \rangle.$$

相比于直接计算 $\mathrm{vec}(\mathcal{X})^\top \mathbf{H}\mathrm{vec}(\mathcal{X})$ 所需的 $\mathcal{O}(Ln^{2d})$ 次浮点运算, 新方法的计算复杂度为 $\mathcal{O}(Ldn^2r_{\max}^2)$, 其随 d 线性增长, 其中 $r_{\max} = \max r_1, r_2, \ldots, r_{d-1}$.

我们将 NTT-RCG 方法应用于问题 (5-13), 其性能通过以下指标进行衡量: 1) 最小 (最大) 特征值的相对误差 $|\lambda_{\min} - \lambda|/|\lambda_{\min}|(|\lambda_{\max} - \lambda|/|\lambda_{\max}|)$; 2) 若内存允许, 则计算子空间距离 $\mathrm{dist}(\mathcal{X}, \mathbf{x}^*) = \|\mathrm{vec}(\mathcal{X})\mathrm{vec}(\mathcal{X})^\top - \mathbf{x}^*(\mathbf{x}^*)^\top\|_{\mathrm{F}}$.

Laplace 算子上的测试　我们考虑对 d 维 Laplace 算子 (5-12) 的离散化. 其特征值 $\lambda_{i_d,i_{d-1},\ldots,i_1}$ 及对应的特征向量 $\mathbf{v}_{i_d,i_{d-1},\ldots,i_1}$ 具有如下的显式表达式:

$$\lambda_{i_d,\ldots,i_1} = 4\sum_{k=1}^{d} \sin^2\left(\frac{i_k\pi}{2(n_k+1)}\right) \quad \text{和} \quad \mathbf{v}_{i_d,\ldots,i_1}(j_d,\ldots,j_1) = \prod_{k=1}^{d} \sin\left(\frac{i_k j_k \pi}{n_k+1}\right) \tag{5-14}$$

其中 $i_k, j_k \in [n_k]$, $k \in [d]$. 由引理 5.6 可知, 目标函数 f 在 $\mathcal{X} = [\![\mathcal{U}_1, \mathcal{U}_2, \ldots, \mathcal{U}_d]\!]$ 处的欧氏梯度可以高效地计算为

$$\begin{aligned}
&\nabla f(\mathcal{X}) \\
&= [\![\mathcal{U}_1 \times_2 \mathbf{T}_{n_1}, \mathcal{U}_2, \ldots, \mathcal{U}_d]\!] + [\![\mathcal{U}_1, \mathcal{U}_2 \times_2 \mathbf{T}_{n_2}, \ldots, \mathcal{U}_d]\!] + \cdots + [\![\mathcal{U}_1, \mathcal{U}_2, \ldots, \mathcal{U}_d \times_d \mathbf{T}_{n_d}]\!] \\
&= [\![\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_d]\!],
\end{aligned}$$

这里

$$\mathbf{G}_1(i_1) = \begin{bmatrix} \tilde{\mathbf{U}}_1(i_1) & \mathbf{U}_1(i_1) \end{bmatrix}, \quad \mathbf{G}_k(i_k) = \begin{bmatrix} \mathbf{U}_k(i_k) & 0 \\ \tilde{\mathbf{U}}_k(i_k) & \mathbf{U}_k(i_k) \end{bmatrix}, \quad \mathbf{G}_d(i_d) = \begin{bmatrix} \mathbf{U}_d(i_d) \\ \tilde{\mathbf{U}}_d(i_d) \end{bmatrix},$$

$k = 2, 3, \ldots, d-1$ 以及 $\tilde{\mathcal{U}}_k = \mathcal{U}_k \times_2 \mathbf{T}_{n_k}$. 注意到 $\mathcal{G}_k$ 的大小与维数 d 无关, 从而保证了计算的可扩展性.

我们将所提出的 NTT-RCG 方法与交替线性方法 [170] 进行比较, 该方法也被称为单站点密度矩阵重整化群 (single-site DMRG) 方法 [160, 161]. 我们注意到, 式 (5-14) 中的任意特征向量 $\mathbf{v}_{i_d,i_{d-1},\ldots,i_1}$ 都可以重排为 $\mathbb{R}^{n_d \times n_{d-1} \times \cdots \times n_1}$ 中的一个秩为 1 的张量, 因此特征值问题 (5-11) 等价于 $\mathbf{r} = (1, 1, \ldots, 1)$ 情形下的 (5-13). 在数值实验中, 我们取 $n_1 = n_2 = \cdots = n_d = 10$, 并令 $d = 8, 16, 32, \ldots, 256$. 表 5-2 和图 5-5 给出了相应的数值结果. 可以观察到, 在不同的 d 取值下, 所有方法均收敛到最大特征值及其对应的特征向量. 值得注意的是, 所提出的方法相较于单站点 DMRG 表现出更快的收敛速度, 并且在最大特征值的计算精度方面具有更优的表现.

横场 Ising 哈密顿量的测试　考虑一个具有 d 个格点的 Ising 模型, 其哈密顿量为

$$\mathbf{H} = -\sum_{k=1}^{d-1} \sigma_k^z \sigma_{k+1}^z - t\sum_{k=1}^{d} \sigma_k^x,$$

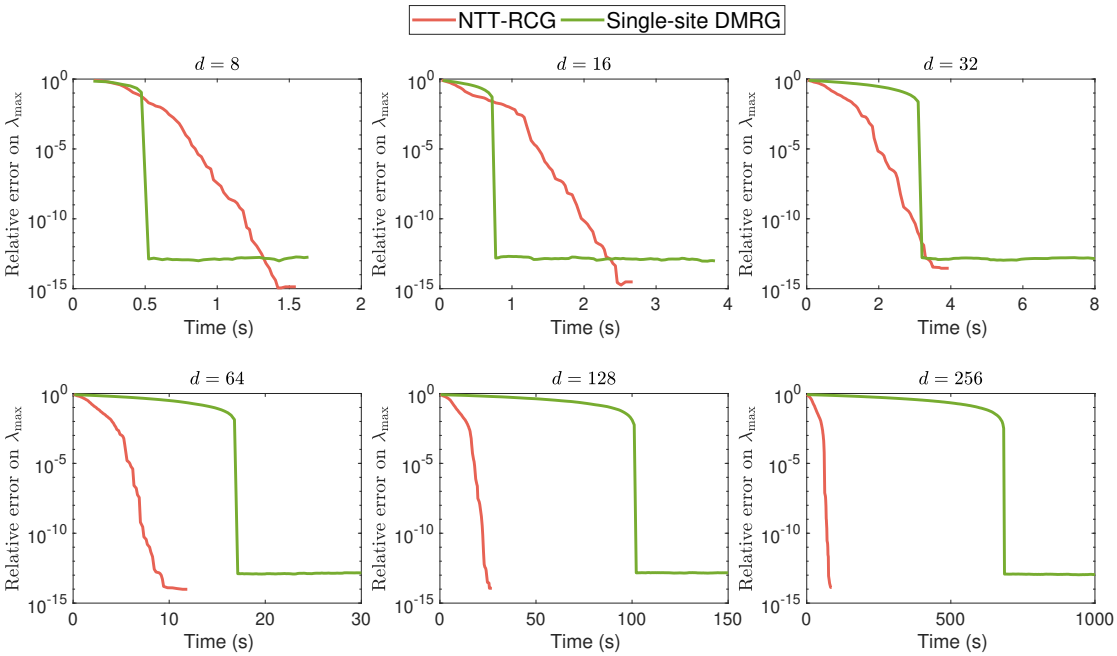


图 5-5 在 $d = 8, 16, 32, \ldots, 256$ 下两种方法的收敛性结果.

Figure 5-5 Convergence of two methods for $d = 8, 16, 32, \ldots, 256$.

表 5-2 Laplace 算子离散化问题的数值实验结果.

Table 5-2 Numerical results on the discretization of the Laplace operator.

d	NTT-RCG			单站点 DMRG		
	时间 (秒)	$\lambda_{\max}$ 的相对误差	$\mathrm{dist}(\mathcal{X}, \mathbf{x}^*)$	时间 (秒)	$\lambda_{\max}$ 的相对误差	$\mathrm{dist}(\mathcal{X}, \mathbf{x}^*)$
8	1.55	1.3598e-15	2.1491e-07	1.63	1.8085e-13	5.5952e-06
16	2.67	3.0596e-15	6.3571e-07	3.82	9.6773e-14	5.0423e-06
32	3.94	2.8896e-14	6.1342e-06	17.05	1.2057e-13	9.1063e-06
64	11.88	9.7453e-15	6.1905e-06	85.50	1.4165e-13	1.4724e-05
128	27.03	1.1558e-14	1.2112e-05	512.65	1.1332e-13	1.7337e-05
256	86.49	1.3258e-14	2.4897e-05	3438.71	1.1751e-13	2.6551e-05

其中 σ^x, σ^z 为在第 5.4 节中定义的 Pauli 矩阵, $\sigma_k^z = \mathbf{I}_{2^{k-1}} \otimes \sigma^z \otimes \mathbf{I}_{2^{d-k}}$, $\sigma_k^x = \mathbf{I}_{2^{k-1}} \otimes \sigma^x \otimes \mathbf{I}_{2^{d-k}}$, 且 $t \in \mathbb{R}$. 矩阵 $\mathbf{H}$ 的特征值可以通过 Jordan–Wigner 变换高效计算得到, 然而其特征向量并不存在显式表达. 为此, 我们通过问题 (5-13) 寻求对应于最小特征值的特征向量的低秩近似解.

函数 f 在 $\mathcal{X} \in \mathcal{N}_{\mathbf{r}}$ 处的欧氏梯度可以表示为张量 $\nabla f(\mathcal{X}) = [\![\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_d]\!]$, 其中

$$\mathbf{G}_1(i_1) = [\,\tilde{\mathbf{U}}_1(i_1)\ \check{\mathbf{U}}_1(i_1)\ \mathbf{U}_1(i_1)\,],\ \mathbf{G}_k(i_k) = \begin{bmatrix} \mathbf{U}_k(i_k) & 0 & 0 \\ \check{\mathbf{U}}_k(i_k) & 0 & 0 \\ \tilde{\mathbf{U}}_k(i_k) & \check{\mathbf{U}}_k(i_k) & \mathbf{U}_k(i_k) \end{bmatrix},\ \mathbf{G}_d(i_d) = \begin{bmatrix} \mathbf{U}_d(i_d) \\ \check{\mathbf{U}}_d(i_d) \\ \tilde{\mathbf{U}}_d(i_d) \end{bmatrix},$$

$k = 2, 3, \ldots, d-1$, $\check{\mathcal{U}}_k = \mathcal{U}_k \times_2 \mathbf{S}^{(z)}$ 以及 $\tilde{\mathcal{U}}_k = \mathcal{U}_k \times_2 \mathbf{S}^{(x)}$. 上述表达式是由引理 5.6 推导得到的.

在数值实验中, 我们取 $t = 1$, 并令 $d = 8, 16, 32, \ldots, 256$. 为获得更好的数值性能, 我们对 NTT-RCG 方法采用了秩递增策略, 参见例如 [60, §4.9]. 从初始秩 $\mathbf{r}^{(0)} = (1, 1, \ldots, 1)$ 开始, 我们在每一个秩下运行 NTT-RCG 方法 50 次迭代, 并将秩更新为 $\mathbf{r}^{(t+1)} = \min\{(1, 2, \ldots, 2^{\lfloor d/2 \rfloor}, 2^{\lfloor d/2 \rfloor - 1}, \ldots, 1), \mathbf{r}^{(t)} + 1\}$, 直至达到预设的最大秩 $\mathbf{r}$. 最大秩设定为 $\mathbf{r} = \min\{(1, 2, \ldots, 2^{\lfloor d/2 \rfloor}, 2^{\lfloor d/2 \rfloor - 1}, \ldots, 1), (1, r, r, \ldots, 1)\}$ 其中 $r = 1, 4, 6, 8, \ldots, 14$. 图 5-6和表 5-3 给出了 NTT-RCG 方法的数值表现. 首先, 对于较小的系统规模 ($d = 8, 16$), 可通过 MATLAB 函数 eigs 计算参考特征向量, 最小特征值 $\lambda_{\min}$ 的相对误差以及子空间距离均随着参数 r 的增大而减小, 即更高秩的解能够更精确地逼近真实特征向量. 其次, 在所有 d 的取值下, NTT-RCG 方法均能在较小秩参数的条件下获得较小的 $\lambda_{\min}$ 相对误差, 并保持可接受的计算时间. 最后, NTT 表示中的参数数量仅随 d 线性增长, 这与完整张量表示中参数数量的指数级增长形成了鲜明对比. 正是这种低秩结构, 使得我们能够对大规模自旋链进行计算 (在实验中可达 $d = 256$ 个格点).

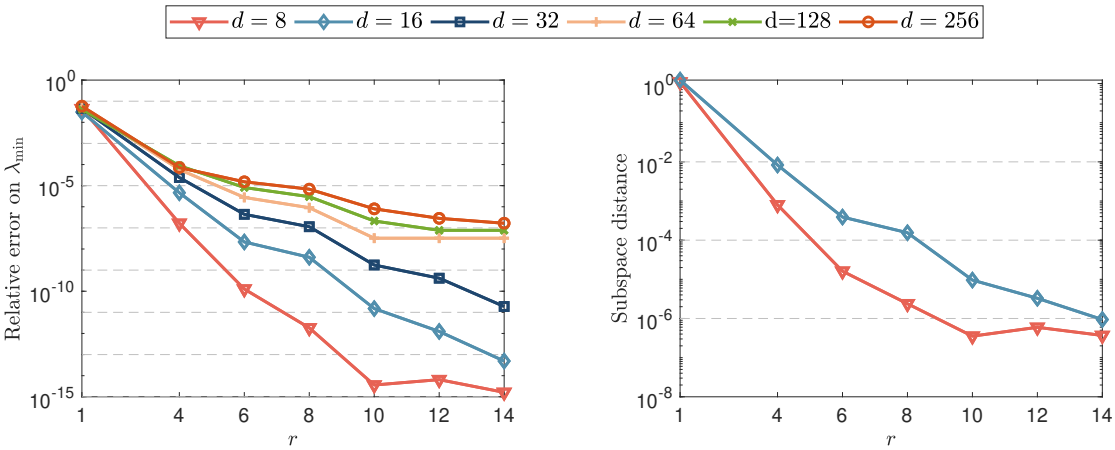


图 5-6 Ising 哈密顿量的数值结果. 左图: 在 $d = 8, 16, 32, \ldots, 256$ 情形下, 最小特征值 $\lambda_{\min}$ 的相对误差. 右图: 在 $d = 8, 16$ 情形下的子空间距离.

Figure 5-6 Numerical results on the Ising Hamiltonian. Left: relative error on $\lambda_{\min}$ with $d = 8, 16, 32, \ldots, 256$. Right: subspace distance with $d = 8, 16$.

表 5-3 NTT-RCG 方法在 Ising 哈密顿量问题中的数值性能.

Table 5-3 The performance of the proposed NTT-RCG method on the Ising Hamiltonian.

d	r	参数量	时间 (秒)	$\lambda_{\min}$ 上的相对误差	d	r	参数量	时间 (秒)	$\lambda_{\min}$ 上的相对误差
	1	16	0.66	4.4265e-02		1	128	30.10	4.2145e-02
	4	168	1.17	1.6557e-07		4	1960	24.74	1.3958e-04
	6	280	1.10	1.2708e-10		6	4312	40.19	2.7756e-06
8	8	424	1.25	1.8347e-12	64	8	7592	55.97	9.0321e-07
	10	488	1.39	1.0834e-15		10	11688	73.31	3.2456e-08
	12	552	1.49	8.3058e-15		12	16680	74.70	3.2456e-08
	14	616	1.55	3.4307e-15		14	22568	76.08	3.2456e-08
	1	32	2.24	7.3024e-02		1	256	70.28	7.8158e-02
	4	424	2.86	4.6116e-06		4	4008	73.67	1.6158e-04
	6	856	2.63	2.1532e-08		6	8920	111.12	1.1990e-05
16	8	1448	3.90	4.0551e-09	128	8	15784	148.26	3.0342e-06
	10	2088	4.01	1.5077e-11		10	24488	184.17	2.1188e-07
	12	2856	3.47	1.2472e-12		12	35112	187.36	2.1188e-07
	14	3752	5.06	6.5316e-14		14	47656	209.00	3.8778e-08
	1	64	8.92	4.5299e-02		1	512	187.28	8.2331e-02
	4	936	7.65	2.4219e-05		4	8104	194.31	1.8673e-03
	6	2008	8.99	4.4011e-07		6	18136	265.97	3.3424e-04
32	8	3496	9.23	1.1340e-07	256	8	32168	331.03	5.6170e-05
	10	5288	11.32	1.7451e-09		10	50088	434.01	1.2384e-06
	12	7464	13.79	4.1455e-10		12	71976	537.43	2.8209e-07
	14	10024	13.20	1.8888e-11		14	97832	594.19	1.6480e-07

5.4　量子信息理论中的应用

本节中, 我们展示 NTT 分解在稳定子秩 (stabilizer rank) 以及量子信道的最小输出熵 (minimum output entropy) 计算中的应用. 这两个量都是量子信息理论中的核心量.

我们首先引入量子信息理论中的一些记号. 由 n 个量子比特组成的量子系统 A 由希尔伯特空间 $H_A = (\mathbb{C}^2)^{\otimes n}$ 描述, 其维数为 2^n. 纯态是 H_A 中的一个单位范数向量 $|\psi\rangle$, 即满足 $\langle\psi|\psi\rangle = 1$. 混合态由密度矩阵 ρ_A 描述, 其中 ρ_A 是定义在 H_A 上的半正定算子, 并满足 $\mathrm{tr}\rho_A = 1$. 量子信道 $N_{A\to B} : L(H_A) \to L(H_B)$ 是作用在线性算子空间之间的线性映射, 且该映射是完全正的并保持迹不变. 量子信道的作用可以通过 Kraus 算子表示为 $N(\rho) = \sum_k \mathbf{K}_k \rho \mathbf{K}_k^\dagger$, 其中 Kraus 算子满足 $\sum_k \mathbf{K}_k^\dagger \mathbf{K}_k = \mathbf{I}$. n 比特 Pauli 群定义为 $\hat{P}_n \coloneqq \{i^k \sigma_{h_1} \otimes \sigma_{h_2} \otimes \cdots \otimes \sigma_{h_n} : k, h_j \in \{0,1,2,3\}\}$, 其中

$$\sigma_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \sigma_1 = \sigma^x = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \sigma_2 = \sigma^y = \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix}, \sigma_3 = \sigma^z = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$$

为 Pauli 矩阵. 模去相位因子后的 n 比特 Pauli 群记为 $P_n \coloneqq \hat{P}_n / \langle \pm i \mathbb{1}_{2^n} \rangle$. n 比特 Clifford 群是 n 比特 Pauli 群的正规化子 (normalizer), 定义为 $\mathrm{Cl}_n \coloneqq \left\{U : UPU^\dagger \in \hat{P}_n, \forall P \in \hat{P}_n\right\}$. n 比特纯稳定子态 (pure stabilizer states) 定义为 Clifford 群作用下的轨道: $\mathrm{STAB}_n \coloneqq \left\{U|0\rangle^{\otimes n} : U \in \mathrm{Cl}_n\right\}$. 最后, 我们设定秩参数为

$$\mathbf{r} = \min\{(1, 2, 4, \ldots, 2^{\lfloor d/2 \rfloor}, 2^{\lfloor d/2 \rfloor - 1}, \ldots, 2, 1), (1, r, r, \ldots, r, 1)\} \tag{5-15}$$

其中 r 为默认给定的整数参数.

5.4.1　稳定子秩的近似计算

非稳定性 (nonstabilizerness), 亦称为“魔性”(magic) [162], 是释放量子计算全部潜力、并实现超越经典算法的通用量子算法所必需的一种资源 [171, 172]. 其根源在于 Gottesman–Knill 定理 [173], 该定理表明, 任何稳定子态与 *Clifford* 操作都可以在经典计算机上被高效模拟. 稳定子秩 (stabilizer rank) 为刻画非稳定性提供了一种定量度量 [163, 174]. 对于一个 n 比特纯量子态, 其稳定子秩定义为

$$\chi(|\psi\rangle) \coloneqq \left\{ \min R \in \mathbb{N}_+ : \begin{array}{l} \exists c_1, \ldots, c_R \in \mathbb{C}, |s_1\rangle, \ldots, |s_R\rangle \in \mathrm{STAB}_n, \\ \text{s.t. } |\psi\rangle = \sum_{j=1}^R c_j |s_j\rangle \end{array} \right\}. \tag{5-16}$$

确定量子态的稳定子秩对于量化实现量子加速所需的魔性资源, 以及理解经典与量子计算能力之间的边界, 具有基础性意义 [163, 174, 175].

由稳定子秩的定义可知, 对于态的张量积 $|\psi\rangle^{\otimes n}$, 直接计算 $\chi(|\psi\rangle^{\otimes n})$ 是不可行的, 因为稳定子态的数量随 n 的增长呈 $2^{(1/2+o(1))n^2}$ 的规模增长; 参见 [176, Proposition 2]. 因此, 为了获得其估计值, 我们提出了一种高效的几何方法, 用以验证给定的 R 是否为 (5-16) 的一个可行解. 为此, 首先需要对稳定子态进行一种高

效刻画. 近期提出的一种用于刻画 n 比特纯量子态魔性的度量是 α-稳定子 Rényi 熵 (stabilizer Rényi entropy, SRE) [177], 其定义为

$$M_\alpha(|\psi\rangle) \coloneqq \frac{1}{1-\alpha}\log_2 \sum_{P\in P_n} \Xi_P^\alpha(|\psi\rangle) - n,$$

其中 $\Xi_P(|\psi\rangle) \coloneqq \langle\psi|P|\psi\rangle^2/2^n$, $P \in P_n$ 为一个 n 比特 Pauli 串. 当且仅当 $|\psi\rangle \in \mathrm{STAB}_n$ 时, 有 $M_\alpha(|\psi\rangle) = 0$. 更为关键的是, SRE 可以对矩阵乘积态 (MPS) 进行高效计算. Haug 和 Piroli [178] 提出了一种方法, 可以将具有 n 个键维数 $r, r, \ldots, r$ 的 MPS 的 SRE 计算, 转化为对一个键维数为 $r^{2\alpha}, r^{2\alpha}, \ldots, r^{2\alpha}$ 的 MPS 的范数计算.

这一结果启发我们针对给定的纯态 $|\psi\rangle$、秩参数 $\mathbf{r}$ 以及分解项数 $R \in \mathbb{N}_+$, 考虑如下的优化问题:

$$\begin{aligned} \min_{\{c_j\}_j, \{|\phi_j\rangle\}_j} \ & f(\{c_j\}_j, \{|\phi_j\rangle\}_j) = \frac{1}{2}\Big\| \sum_{j=1}^R c_j|\phi_j\rangle - |\psi\rangle\Big\|_{\mathrm{F}}^2 + \lambda \sum_{j=1}^R M_2(|\phi_j\rangle) \\ \text{s.t.} \ & c_1, c_2, \ldots, c_R \in \mathbb{C}, \text{ 每一个 } |\phi_j\rangle \in \mathcal{N}_{\mathbf{r}}, \end{aligned} \tag{5-17}$$

其中 $\lambda > 0$ 为罚参数. 目标函数由两部分组成: 第一项为保真度项 $\|\sum_{j=1}^R c_j|\phi_j\rangle - |\psi\rangle\|_{\mathrm{F}}^2/2$, 用于衡量重构误差; 第二项为 SRE 正则化项 $M_2(|\phi_j\rangle)$, 用于得到低魔性的解. 问题 (5-17) 可被视为定义在如下乘积流形上的优化问题:

$$\mathcal{M} = \mathbb{C}^n \times \mathcal{N}_{\mathbf{r}} \times \mathcal{N}_{\mathbf{r}} \times \cdots \times \mathcal{N}_{\mathbf{r}}.$$

关于乘积流形上的优化问题, 可参见本章第 2章的内容.

寻找一个目标函数值严格为零的精确分解可以给出 $\chi(|\psi\rangle)$ 的一个严格上界, 但在数值上具有很大挑战性. 因此, 我们转而引入 (ϵ, δ)-近似稳定子秩, 其定义为满足如下条件的最小 R: 存在一组 $\{c_j, |\phi_j\rangle\}_{j=1}^R\}$ 满足下面两个条件: i) 具有较小的不保真度, 即 $1 - |\sum_{j=1}^R c_j\langle\phi_j|\psi\rangle|^2 \le \epsilon$; ii) 具有较低的魔性, 即对所有 $j = 1, 2, \ldots, R$, 均有 $M_2(|\phi_j\rangle) \le \delta$. 当 R 较小时, 找到这样一种分解便给出了对魔性态的一种物理上有意义的近似, 其分解分量均接近稳定子态集合.

一种具有代表性的魔性态是单比特 H 态 $|H\rangle = \cos(\pi/8)|0\rangle + \sin(\pi/8)|1\rangle$. Bravyi 等人 [163] 证明了 $|H^{\otimes n}\rangle$ 的稳定子秩满足上界 $\chi(|H^{\otimes n}\rangle) \le 7^{n/6}$. 因此, 在数值实验中, 我们取 $\lambda = 1$, $n = 2, 3, 4, 5, 6$, 并对每个 n 令 $R = 1, 2, \ldots, \lceil 7^{n/6}\rceil$. 当 $n \le 4$ 时, 选取 (5-15) 中参数 $r = 1, 2$ 的秩参数 $\mathbf{r}$; 当 $n = 5, 6$ 时, 选取 $r = 1, 2, \ldots, 5$ 的 $\mathbf{r}$. 由于 $M_2(|\psi\rangle)$ 的欧几里得梯度包含 4^n 项, 其欧几里得梯度及投影梯度在计算上是不可行的. 因此, 我们采用有限差分方法 (见注释 5.2.5), 通过沿 $\mathcal{M}$ 的切空间基方向计算方向导数来近似投影梯度, 该过程需要 $\mathcal{O}(2Rnr_{\max}^2)$ 次目标函数 f 的计算, 其中 $r_{\max} = \max r_1, r_2, \ldots, r_{n-1}$.

数值结果见表 5-4 与图 5-7. 我们观察到, 随着分解数 R 的增加, 近似的不保真度不断减小, 同时, 各分量中的最大 SRE 始终保持在较低水平. 例如, 如表 5-4 所示, 在 $n = 4$ 比特的情形下, 当 $R = 3$ 时, 我们得到的不保真度低于

1.2×10^{-3}, 且每个分量的 SRE 均小于 8.3×10^{-3}. 这表明 $R=3$ 是 $|H^{\otimes4}\rangle$ 的一个 $(1.2\times10^{-3}, 8.3\times10^{-3})$-近似稳定子秩. 上述结果表明, 所提出的方法能够为基于“将量子态近似分解为低魔性分量”的经典模拟算法设计提供有效指导.

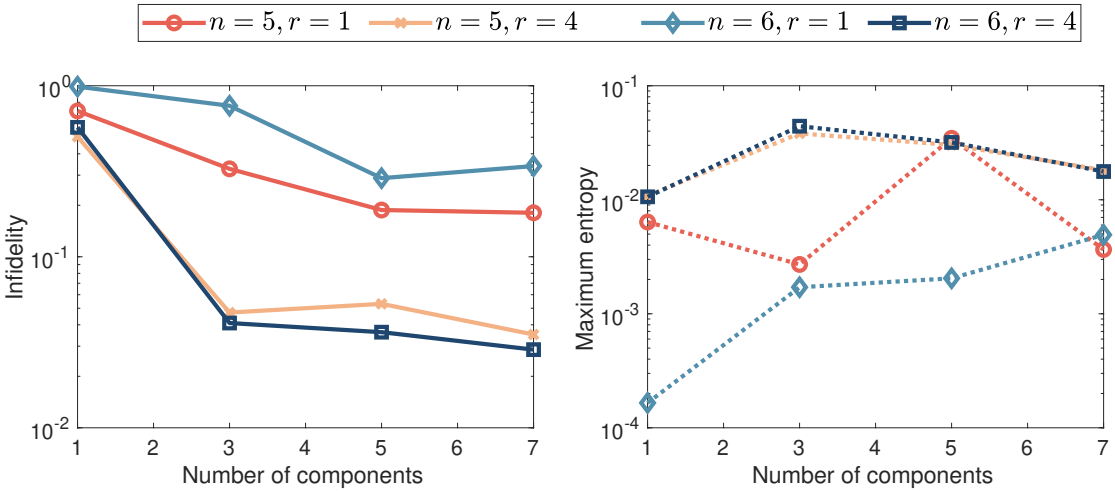


图 5-7 对 $n=5,6$ 比特的 $|H^{\otimes n}\rangle$ 近似稳定子秩的估计的数值结果. 左图: 不保真度. 右图: 各分量中的最大 SRE.

Figure 5-7 Numerical results on estimating approximate stabilizer rank of $|H^{\otimes n}\rangle$ for $n=5,6$ qubits. Left: infidelity. Right: the maximum 2-stabilizer Rényi entropy among each component.

5.4.2　最小输出 Rényi p-熵

给定一个量子通道 $\mathcal{N}_{A\to B}: L(H_A)\to L(H_B)$, 其最小输出 Rényi p-熵定义为

$$S_p^{\min}(\mathcal{N}_{A\to B}) := \min_{\rho_A} \frac{1}{1-p}\log \mathrm{tr}(\mathcal{N}_{A\to B}(\rho_A)^p), \quad p\in(0,1)\cup(1,+\infty), \tag{5-18}$$

其中最小化是对 $L(H_A)$ 中的所有密度矩阵 ρ_A 进行的. 当 $p\to1$ 时, 有 $S_1^{\min}(\mathcal{N}) = \min_{\rho_A} H(N(\rho_A))$, 这里 $H(\rho)=-\mathrm{tr}(\rho\log\rho)$ 为冯 · 诺依曼熵. 由 Rényi 熵的凹性可知, 最小值 (5-18) 可在纯态输入 $\rho_A=|\psi\rangle\langle\psi|$ 处取得. 因此, 对于给定量子通道计算 $S_p^{\min}(\mathcal{N}_{A\to B})$ 可理解为在单位球面上的优化问题. 关于给定量子通道 $N_{A\to B}$ 的最小输出 Rényi p-熵, 一个核心问题是其严格次可加性 (strict subadditivity), 即既然次可加性总是成立, 是否存在某个 n 使得

$$S_p^{\min}(\mathcal{N}_{A\to B}^{\otimes n}) < nS_p^{\min}(\mathcal{N}_{A\to B}).$$

解决这一问题的一种直接数值方法是对大 n 计算 $n^{-1}S_p^{\min}(\mathcal{N}_{A\to B}^{\otimes n})$, 并与一拷贝 (one-shot) 最小值 $S_p^{\min}(\mathcal{N}_{A\to B})$ 进行比较. 然而, $S_p^{\min}(\mathcal{N}_{A\to B}^{\otimes n})$ 的参数空间随 n 呈指数增长. 考虑到 $N_{A\to B}^{\otimes n}$ 的张量积结构, 我们采用低秩 NTT 张量对 Rényi p-熵进行最小化, 即

$$\begin{aligned} \min_{|\psi\rangle} \quad & f^{(n)}(|\psi\rangle) = \frac{1}{1-p}\log \mathrm{tr}(N_{A\to B}^{\otimes n}(|\psi\rangle\langle\psi|)^p) \\ \text{s.t.} \quad & |\psi\rangle = \mathrm{vec}(\llbracket \mathcal{U}_1, \mathcal{U}_2, \ldots, \mathcal{U}_n\rrbracket), \quad \llbracket \mathcal{U}_1, \mathcal{U}_2, \ldots, \mathcal{U}_n\rrbracket \in \mathcal{N}_{\mathbf{r}}. \end{aligned} \tag{5-19}$$

表 5-4 对 $n=2,3,4$ 比特的 $|H^{\otimes n}\rangle$ (ϵ,δ)-近似稳定子秩的估计的数值结果. 分解分量数 $R=1,2,\dots,\lceil 7^{n/6}\rceil$, 秩参数 $r=1,2$.

Table 5-4 Numerical results on estimating (ϵ,δ)-approximate stabilizer rank of $|H^{\otimes n}\rangle$ for $n=2,3,4$ qubits, number of components $R=1,2,\dots,\lceil 7^{n/6}\rceil$, and parameter $r=1,2$.

n	R	r	不保真度	2-稳定 Rényi 熵
2	1	1	2.4127e-01	7.4318e-03
		2	8.5398e-01	5.4383e-03
	2	1	1.1871e-01	1.5668e-04, 6.4077e-03
		2	6.9951e-05	1.2532e-03, 1.2405e-03
3	1	1	4.0758e-01	1.1000e-01
		2	3.3873e-01	9.7383e-03
	2	1	2.4934e-01	4.4777e-03, 9.9008e-04
		2	2.0903e-01	2.4321e-03, 1.2648e-02
	3	1	1.1337e-01	2.9432e-02, 1.6710e-02, 8.9772e-04
		2	2.8777e-03	5.6882e-03, 4.4175e-03, 3.8043e-03
4	1	1	4.3528e-01	8.5178e-03
		2	8.9291e-01	4.1529e-03
	2	1	4.2484e-01	9.0938e-05, 9.7877e-03
		2	1.0962e-03	6.5625e-03, 7.5606e-03
	3	1	1.0278e-01	2.5400e-04, 3.5386e-03, 7.1955e-03
		2	1.1394e-03	8.2699e-03, 5.7145e-04, 5.5453e-03
	4	1	2.6224e-01	6.2629e-05, 8.8638e-04, 2.1451e-03, 1.7844e-03
		2	2.4029e-04	1.5270e-03, 1.7929e-06, 1.6754e-03, 1.4706e-03

对于 $p=2$, 可得到

$$
\begin{aligned}
&\operatorname{tr}(\mathcal{N}_{A\to B}^{\otimes n}(|\psi\rangle\langle\psi|)^2)\\
&=\operatorname{tr}\left(\Big(\sum_{k_1,k_2,\dots,k_n=1}^{K}(\bigotimes_{j=1}^{n}\mathbf{K}_{k_j})|\psi\rangle\langle\psi|\bigotimes_{j=1}^{n}\mathbf{K}_{k_j}^{\dagger}\Big)^2\right)\\
&=\left\|\sum_{k_1,k_2,\dots,k_n=1}^{K}(\bigotimes_{j=1}^{n}\mathbf{K}_{k_j})|\psi\rangle\langle\psi|\bigotimes_{j=1}^{n}\mathbf{K}_{k_j}^{\dagger}\right\|_{\mathrm{F}}^2\\
&=\operatorname{tr}\Big(\sum_{k_1^{(1)},\dots,k_n^{(1)}=1}^{K}\sum_{k_1^{(2)},\dots,k_n^{(2)}=1}^{K}(\bigotimes_{j=1}^{n}\mathbf{K}_{k_j^{(1)}})|\psi\rangle\langle\psi|\bigotimes_{j=1}^{n}(\mathbf{K}_{k_j^{(1)}}^{\dagger}\mathbf{K}_{k_j^{(2)}})|\psi\rangle\langle\psi|\bigotimes_{j=1}^{n}\mathbf{K}_{k_j^{(2)}}^{\dagger}\Big)\\
&=\sum_{k_1^{(1)},k_2^{(1)},\dots,k_n^{(1)}=1}^{K}\sum_{k_1^{(2)},k_2^{(2)},\dots,k_n^{(2)}=1}^{K}\Big|\langle\psi|\bigotimes_{j=1}^{n}(\mathbf{K}_{k_j^{(1)}}^{\dagger}\mathbf{K}_{k_j^{(2)}})|\psi\rangle\Big|^2,
\end{aligned}
$$

其中 $\{\mathbf{K}_{k_j}\}_{k_j}$ 是 $\mathcal{N}_{A\to B}$ 的 Kraus 算子集合. 沿用 [178] 中的思路, 目标函数 $f(|\psi\rangle)$ 可以通过 TT 形式张量的 Frobenius 范数高效计算, 该张量大小为 $2^2\times 2^2\times\cdots\times 2^2$, 键维数为 $(1,r_1^2,r_2^2,\dots,r_{n-1}^2,1)$.

反对称通道　反对称子空间 asym_p^d 是 $(\mathbb{C}^d)^{\otimes p}$ 的子空间, 定义为 $\mathrm{asym}_d^p \coloneqq \{|\psi\rangle \in (\mathbb{C}^d)^{\otimes p} : |\psi\rangle = (-1)^{\mathrm{sgn}(\sigma)} P_\sigma |\psi\rangle, \text{ for all } \sigma \in S_p\}$, 其中 S_p 是对称群, $\mathrm{sgn}(\sigma)$ 是置换 σ 的奇偶性, P_σ 是将 p 个子系统按置换 σ 重新排列的酉算符, 即 $P_\sigma(|\psi_1\rangle \otimes \cdots \otimes |\psi_p\rangle) \coloneqq |\psi_{\sigma(1)}\rangle \otimes \cdots \otimes |\psi_{\sigma(p)}\rangle$. 当 $d = 3, p = 2$ 时, 反对称子空间的一组基为 $|\psi_1\rangle = (|01\rangle - |10\rangle)/\sqrt{2}$, $|\psi_2\rangle = (|02\rangle - |20\rangle)\sqrt{2}$, $|\psi_3\rangle = (|12\rangle - |21\rangle)\sqrt{2}$. 接下来考虑通道 $N_{\mathrm{as}}(\cdot)$, 其 Stinespring 等距映射为 $V : \mathbb{C}^3 \to \mathrm{asym}_3^2$, 矩阵形式为 $V = (|\psi_1\rangle, |\psi_2\rangle, |\psi_3\rangle)$. 该通道的 Kraus 算子为 $\mathbf{K}_1, \mathbf{K}_2, \mathbf{K}_3 \in \mathbb{C}^{3\times 3}$, 定义为

$$\mathbf{K}_1 = -\frac{1}{\sqrt{2}}(|1\rangle\langle 0| + |2\rangle\langle 1|),\ \mathbf{K}_2 = \frac{1}{\sqrt{2}}(|0\rangle\langle 0| - |2\rangle\langle 2|),\ \mathbf{K}_3 = \frac{1}{\sqrt{2}}(|0\rangle\langle 1| + |1\rangle\langle 2|).$$

我们设置 $n = 11, 12, \ldots, 15$, 并采用 (5-15) 中的秩参数 $\mathbf{r}$, $r = 1, 2, \ldots, 10$. 对 (5-19) 使用 NTT-RCG 方法进行求解, 当迭代次数达到 2000 时终止. 对于每个比特数和秩参数组合, NTT-RCG 方法重复运行五次. 反对称通道的数值结果见图 5-8. 从图 5-8(左) 和图 5-8(中) 可以观察到, 每次迭代的平均时间随着比特数和秩参数呈多项式增长. 同时可以得出, 对于 $10, 11, \ldots, 16$ 个比特以及秩参数不超过 10 的 NTT 张量, 可加性成立.

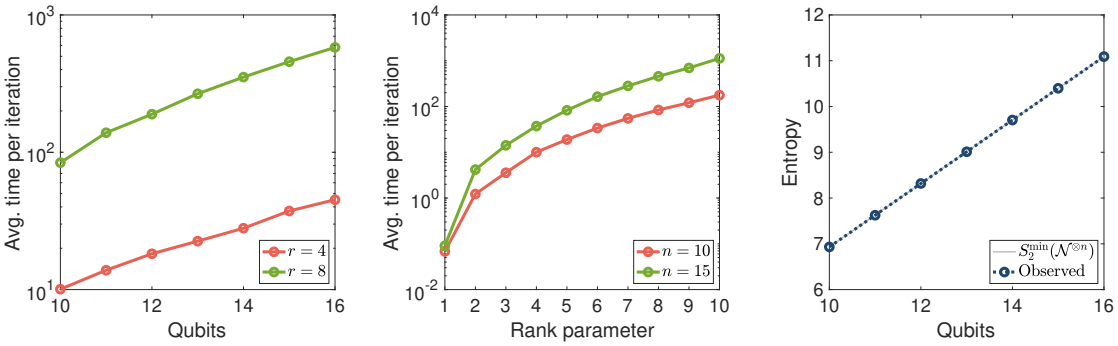


图 5-8 反对称通道的数值结果. 左图: 每次迭代的平均时间随比特数变化. 中图: 每次迭代的平均时间随秩参数变化. 右图: NTT-RCG 计算得到的最小熵.

Figure 5-8 Numerical results on the antisymmetric channel. Left: average time per iteration with respect to qubit. Middle: average time per iteration with respect to the rank parameter. Right: smallest entropy computed from NTT-RCG.

广义振幅阻尼通道 (GADC)　广义振幅阻尼通道定义为 $A_{\gamma,N} : \rho \mapsto \sum_{k=1}^4 \mathbf{A}_k \rho \mathbf{A}_k^\dagger$, 其中 Kraus 算子 $\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3, \mathbf{A}_4 \in \mathbb{C}^{2\times 2}$ 为

$$\mathbf{A}_1 = \sqrt{1-N}\big(|0\rangle\langle 0| + \sqrt{1-\gamma}|1\rangle\langle 1|\big), \qquad \mathbf{A}_2 = \sqrt{\gamma(1-N)}|0\rangle\langle 1|,$$
$$\mathbf{A}_3 = \sqrt{N}\big(\sqrt{1-\gamma}|0\rangle\langle 0| + |1\rangle\langle 1|\big), \qquad \mathbf{A}_4 = \sqrt{\gamma N}|1\rangle\langle 0|,$$

且 $\gamma, N \in [0, 1]$. 我们设置 $n = 2, 3, \ldots, 12$, 并采用 (5-15) 中的秩参数 $\mathbf{r}$, $r = 1, 2, \ldots, 10$. 广义振幅阻尼通道的数值结果见图 5-9. 从图 5-9(左) 和图 5-9(中) 可以看出, 每次迭代的平均时间随着比特数 n 和秩参数 r 呈多项式增长. 同时可以得出, 对于 $2, 3, \ldots, 12$ 个比特以及秩参数不超过 10 的 NTT 张量, 可加性成立.

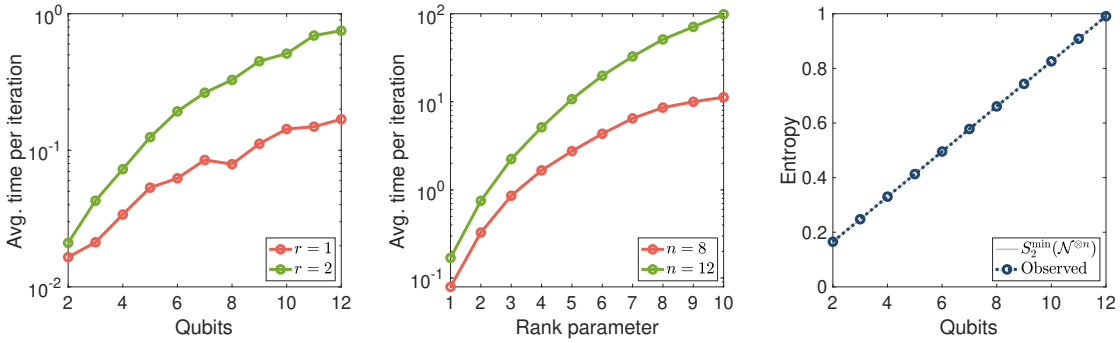


图 5-9 广义振幅阻尼通道的数值结果. 左图: 每次迭代的平均时间随比特数变化. 中图: 每次迭代的平均时间随秩参数变化. 右图: NTT-RCG 计算得到的最小熵.

Figure 5-9 Numerical results on generalized amplitude damping channel. Left: average time per iteration with respect to qubit. Middle: average time per iteration with respect to the rank parameter. Right: smallest entropy computed from NTT-RCG.

5.5 本章小结

我们引入并研究了归一化张量链分解, 用于表示低秩张量以及建模新的低秩张量优化问题. 我们证明了单位范数张量的 NTT 分解的存在性, 并构造了一个拟优的近似投影算子, 即 NTT-SVD 算法. 我们还证明了固定秩 NTT 张量构成的集合是一个光滑流形, 并推导出了相应的几何工具. 虽然 NTT 与 TT 分解在实践中有许多相似之处, 但它们的底层几何结构本质上是不同的. NTT 自然地将张量编码在单位球上, 因此特别适合那些归一化是内在需求的问题, 例如单位范数的低秩张量恢复、特征向量计算、稳定子秩的近似、最小输出 Rényi p-熵的计算, 以及其他受单位范数约束的应用. 我们展示了 NTT 分解在若干应用中的优势. 基于 NTT 的黎曼共轭梯度法在大规模局部哈密顿量的特征值与特征向量计算上, 相比单点 DMRG 方法具有更高的效率. 在量子信息理论中, 基于 NTT 分解的几何方法为估计魔态的稳定子秩以及研究量子通道熵的可加性提供了可行的数值工具.

第 6 章　总结与展望

低秩张量优化是处理高维结构化数据的核心工具, 在数据分析、量子计算及机器学习等领域中具有广泛的应用价值. 本文从低秩张量集合的几何特性出发, 从几何结构、算法构建到实际应用, 系统进行了研究, 形成了从基础几何理解到高效算法实现的完整工作链.

在第二章中, 我们通过理论分析与数值实验验证了精心构造的度量能够有效加速黎曼优化方法. 具体而言, 我们提出了一个利用预条件度量, 求解乘积流形上优化问题的一般框架, 该框架包含了三种具体方法来构造算子以近似黎曼 Hessian. 包括黎曼高斯牛顿方法以及数值线性代数中的块 Jacobi 预条件法在内的多种现有方法, 都可以通过选取特定度量的方式, 利用所提出的框架进行解释. 我们基于所提出的框架, 为典型相关分析和截断奇异值分解设计了新的预条件度量, 并通过计算局部极小点处黎曼 Hessian 的条件数验证了所提出度量的效果, 理论结果显示条件数确实得到了改善. 数值结果进一步验证, 精心设计的度量确实能够提升黎曼优化方法的性能.

在第三章中, 我们基于张量环分解提出了用于张量补全问题的黎曼预条件算法. 预条件效果来源于定义在由核张量的模态 2 展平矩阵所构成的矩阵乘积空间上的一种度量. 在预条件方法中, 直接计算黎曼梯度需要进行大规模矩阵乘法, 这在实际计算中代价过高. 为此, 我们采用了一种高效的计算策略, 在不显式构造大矩阵的情况下计算黎曼梯度. 所提出的算法具有全局收敛性保证, 并且在人工合成数据和真实数据集上的数值实验均展示出了良好的性能.

在第四章中, 我们对 Tucker 张量代数簇的几何结构进行了深入研究, 并提出了用于 Tucker 张量代数簇优化的新几何方法与秩自适应方法. 我们给出了 Tucker 张量代数簇每一点切锥的显式表达. 此外, Tucker 张量代数簇上优化的一个核心难点在于度量投影不具有显式解. 基于得到的几何结构, 我们提出了近似投影方法, 以规避度量投影的显式计算. 我们同时也发现, 仅利用切锥的部分信息即可获得无需收缩映射的搜索方向. 张量补全的数值实验表明, 所提出的方法在不同秩参数选择下均优于现有最好的方法. 总体而言, 当可靠的秩参数可用时, 我们推荐使用 GRAP 与 rfGRAP 方法, 以避免可能冗长的秩参数选择过程; 而在秩参数不确定的情形下, 则建议使用 TRAM 方法, 因为其能够自适应地识别合适的秩参数.

在第五章中, 我们引入并研究了归一化张量链分解, 用于表示低秩张量以及建模新的低秩张量优化问题. 我们证明了单位范数张量的 NTT 分解的存在性, 并提出 NTT-SVD 算法, 它可以得到拟优的近似投影. 我们还证明了固定秩 NTT 张量构成的集合是一个光滑流形, 并推导出了相应的几何工具. NTT 自然地将张量约束在单位球上, 因此特别适合归一化是内在需求的问题, 例如单位范数的低秩张量恢复、特征向量计算、稳定子秩的近似、最小输出 Rényi p-熵的计算, 以及其他受单位范数约束的应用. 我们在数值实验中展示了 NTT 分解在若干应用中

的优势.

实际上, 在博士阶段我们还有一些其他低秩张量相关研究. 例如, 我们针对 Tucker 张量代数簇与张量链分解代数簇, 提出了一种去奇异化方法, 通过引入松弛变量, 将非光滑的张量代数簇参数化为高维空间中的低维流形, 为低秩张量优化提供了一个新的角度. 我们通过在近似投影梯度法中引入秩减机制, 设计了一个 “无灾难点现象” 的一阶方法, 该方法具有理论收敛性保证, 即聚点均为稳定点. 此外, 我们在张量环分解参数空间上引入商几何结构并设计优化方法. 由于篇幅限制, 这些工作未在正文展开, 感兴趣的读者可参见预印本 [97, 179, 180].

本文的研究为低秩张量优化提供了新的视角, 未来工作可沿以下方向展开:

- 将所提出的乘积流形预条件框架应用于其他问题, 以及通过并行计算有望进一步加速黎曼优化方法的运算效率.
- 在特征值问题中, 我们主要关注单个特征向量的计算, 这在文献中已有较多研究. 将 NTT 框架扩展到高效计算多个特征向量或特征子空间仍是一个有趣的开放问题. 由于在张量环分解中取不太大的秩参数就能对局部哈密顿量的基态提供较好的近似, 如何针对张量环分解设计高效计算多个特征向量或特征子空间的算法仍是一个开放问题.
- 如何将低秩张量优化中的几何结构与优化方法引入到大规模神经网络训练与模型压缩中, 并提高训练效率, 是未来一个值得深入探索的研究方向.

参考文献

[1] 袁亚湘, 孙文瑜. 最优化理论与方法 [M]. 北京: 科学出版社, 1997.

[2] 袁亚湘. 非线性优化计算方法 [M]. 北京: 科学出版社, 2008.

[3] Boyd S, Vandenberghe L. Convex optimization [M/OL]. Cambridge university press, 2004. https://web.stanford.edu/~boyd/cvxbook/.

[4] Nocedal J, Wright S J. Numerical optimization (2nd edition) [M]. Springer, 2006.

[5] Sun W, Yuan Y X. Optimization theory and methods: nonlinear programming [M]. Springer, 2006.

[6] Acar E, Yener B. Unsupervised multiway data analysis: A literature survey [J/OL]. IEEE transactions on knowledge and data engineering, 2008, 21(1): 6-20. DOI: 10.1109/TKDE.2008.112.

[7] Vasilescu M A O, Terzopoulos D. Multilinear analysis of image ensembles: Tensorfaces [C]// European conference on computer vision. Springer, 2002: 447-460.

[8] Vasilescu M A O, Terzopoulos D. Multilinear subspace analysis of image ensembles [C/OL]// 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.: volume 2. IEEE, 2003: II-93. DOI: 10.1109/CVPR.2003.1211457.

[9] Grasedyck L, Kressner D, Tobler C. A literature survey of low-rank tensor approximation techniques [J/OL]. GAMM-Mitteilungen, 2013, 36(1): 53-78. DOI: 10.1002/gamm.201310004.

[10] Udell M, Townsend A. Why are big data matrices approximately low rank? [J/OL]. SIAM Journal on Mathematics of Data Science, 2019, 1(1): 144-160. DOI: 10.1137/18M1183480.

[11] De Lathauwer L, De Moor B, Vandewalle J. An introduction to independent component analysis [J/OL]. Journal of Chemometrics: A Journal of the Chemometrics Society, 2000, 14 (3): 123-149. DOI: 10.1002/1099-128X(200005/06)14:3<123::AID-CEM589>3.0.CO;2-1.

[12] Peng Y, Meng D, Xu Z, et al. Decomposable nonlocal tensor dictionary learning for multispectral image denoising [C]//2014 IEEE Conference on Computer Vision and Pattern Recognition. 2014: 2949-2956.

[13] Cichocki A, Mandic D, De Lathauwer L, et al. Tensor decompositions for signal processing applications: From two-way to multiway component analysis [J/OL]. IEEE signal processing magazine, 2015, 32(2): 145-163. DOI: 10.1109/MSP.2013.2297439.

[14] Vandereycken B. Low-rank matrix completion by Riemannian optimization [J/OL]. SIAM Journal on Optimization, 2013, 23(2): 1214-1236. DOI: 10.1137/110845768.

[15] Kressner D, Steinlechner M, Vandereycken B. Low-rank tensor completion by Riemannian optimization [J/OL]. BIT Numerical Mathematics, 2014, 54(2): 447-468. DOI: 10.1007/s10543-013-0455-z.

[16] Gao B, Peng R, Yuan Y x. Optimization on product manifolds under a preconditioned metric [J]. ArXiv preprint arXiv:2306.08873, 2023.

[17] Gao B, Peng R, Yuan Y x. Riemannian preconditioned algorithms for tensor completion via tensor ring decomposition [J/OL]. Computational Optimization and Applications, 2024, 88: 443-468. DOI: 10.1007/s10589-024-00559-7.

[18] Kressner D, Steinlechner M, Vandereycken B. Preconditioned low-rank Riemannian opti-

mization for linear systems with tensor product structure [J/OL]. SIAM Journal on Scientific Computing, 2016, 38(4): A2018-A2044. DOI: 10.1137/15M1032909.

[19] Glau K, Kressner D, Statti F. Low-rank tensor approximation for Chebyshev interpolation in parametric option pricing [J/OL]. SIAM Journal on Financial Mathematics, 2020, 11(3): 897-927. DOI: 10.1137/19M1244172.

[20] Verstraete F, Murg V, Cirac J I. Matrix product states, projected entangled pair states, and variational renormalization group methods for quantum spin systems [J/OL]. Advances in physics, 2008, 57(2): 143-224. DOI: 10.1080/14789940801912366.

[21] Bachmayr M, Eisenmann H, Uschmajew A. Dynamical low-rank tensor approximations to high-dimensional parabolic problems: existence and convergence of spatial discretizations [J]. ArXiv preprint arXiv:2308.16720, 2023.

[22] Wang Y, Lin Z, Liao Y, et al. Solving high dimensional partial differential equations using tensor neural network and a posteriori error estimators [J]. ArXiv preprint arXiv:2311.02732, 2023.

[23] Uschmajew A, Vandereycken B. Geometric methods on low-rank matrix and tensor manifolds [M]//Handbook of variational methods for nonlinear geometric data. Springer, 2020: 261-313.

[24] Koren Y, Bell R, Volinsky C. Matrix factorization techniques for recommender systems [J/OL]. Computer, 2009, 42(8): 30-37. DOI: 10.1109/MC.2009.263.

[25] Gupta V, Koren T, Singer Y. Shampoo: Preconditioned stochastic tensor optimization [C]// Dy J, Krause A. Proceedings of Machine Learning Research: volume 80 Proceedings of the 35th International Conference on Machine Learning. PMLR, 2018: 1842-1850.

[26] Hu E J, Shen Y, Wallis P, et al. Lora: Low-rank adaptation of large language models. [J]. Iclr, 2022, 1(2): 3.

[27] Gu Y, Zhou W, Iacovides G, et al. Tensorllm: Tensorising multi-head attention for enhanced reasoning and compression in llms [C]//2025 International Joint Conference on Neural Networks (IJCNN). IEEE, 2025: 1-8.

[28] Yang Y, Zhou J, Wong N, et al. Loretta: Low-rank economic tensor-train adaptation for ultra-low-parameter fine-tuning of large language models [C]//Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). 2024: 3161-3176.

[29] Pan Y, Yuan Y, Yin Y, et al. Reusing pretrained models by multi-linear operators for efficient training [J]. Advances in Neural Information Processing Systems, 2023, 36: 3248-3262.

[30] Wang W, Sun Y, Eriksson B, et al. Wide compression: Tensor ring nets [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 9329-9338.

[31] Hamilton W R. On some extensions of quaternions [J]. Philosophical Magazine (4th series), 1854, 7: 492-499.

[32] Voigt W. Die fundamentalen physikalischen eigenschaften der krystalle in elementarer darstellung [M]. de Gruyter, 1898.

[33] De Silva V, Lim L H. Tensor rank and the ill-posedness of the best low-rank approximation problem [J/OL]. SIAM Journal on Matrix Analysis and Applications, 2008, 30(3): 1084-1127. DOI: 10.1137/06066518X.

[34] Rota G C. Finite-dimensional multilinear algebra: M. marcus, dekker, part 1, 1973, 292 p.; part 2, 1975, 715 p. [M]. Academic Press, 1976.

[35] Adkins W A, Weintraub S H. Algebra: an approach via module theory: volume 136 [M]. Springer Science & Business Media, 2012.

[36] Comon P, Golub G, Lim L H, et al. Symmetric tensors and symmetric tensor rank [J/OL]. SIAM Journal on Matrix Analysis and Applications, 2008, 30(3): 1254-1279. DOI: 10.1137/060661569.

[37] 张贤达. 矩阵分析与应用 [M]. 清华大学出版社有限公司, 2004.

[38] Kolda T G, Bader B W. Tensor decompositions and applications [J/OL]. SIAM review, 2009, 51(3): 455-500. DOI: 10.1137/07070111X.

[39] Hitchcock F L. Multiple invariants and generalized rank of a p-way matrix or tensor [J]. Journal of Mathematics and Physics, 1928, 7(1-4): 39-79.

[40] Tucker L R, et al. The extension of factor analysis to three-dimensional matrices [J]. Contributions to mathematical psychology, 1964, 110119.

[41] Oseledets I V. Tensor-train decomposition [J/OL]. SIAM Journal on Scientific Computing, 2011, 33(5): 2295-2317. DOI: 10.1137/090752286.

[42] Zhao Q, Zhou G, Xie S, et al. Tensor ring decomposition [J]. ArXiv preprint arXiv:1606.05535, 2016.

[43] Hitchcock F L. The expression of a tensor or a polyadic as a sum of products [J/OL]. Journal of Mathematics and Physics, 1927, 6(1-4): 164-189. DOI: 10.1002/sapm192761164.

[44] Cattell R B. “parallel proportional profiles”and other principles for determining the choice of factors by rotation [J]. Psychometrika, 1944, 9(4): 267-283.

[45] Cattell R B. The three basic factor-analytic research designs—their interrelations and derivatives. [J]. Psychological bulletin, 1952, 49(5): 499.

[46] Carroll J D, Chang J J. Analysis of individual differences in multidimensional scaling via an N-way generalization of “Eckart-Young” decomposition [J/OL]. Psychometrika, 1970, 35 (3): 283-319. DOI: 10.1007/BF02310791.

[47] Harshman R A, et al. Foundations of the PARAFAC procedure: Models and conditions for an “explanatory” multimodal factor analysis [J]. UCLA Working Papers in Phonetics, 1970, 16: 1-84.

[48] Möcks J. Topographic components model for event-related potentials and some biophysical considerations [J]. IEEE transactions on biomedical engineering, 1988, 35(6): 482-484.

[49] Kruskal J B. Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics [J]. Linear algebra and its applications, 1977, 18(2): 95-138.

[50] Håstad J. Tensor rank is NP-complete [C]//International Colloquium on Automata, Languages, and Programming. Springer, 1989: 451-460.

[51] Hillar C J, Lim L H. Most tensor problems are NP-hard [J]. Journal of the ACM (JACM), 2013, 60(6): 1-39.

[52] Shitov Y. How hard is the tensor rank? [J]. ArXiv preprint arXiv:1611.01559, 2016.

[53] Tucker L R. Some mathematical notes on three-mode factor analysis [J/OL]. Psychometrika, 1966, 31(3): 279-311. DOI: 10.1007/BF02289464.

[54] Levin J. Three-mode factor analysis [J/OL]. Psychological Bulletin, 1965, 64(6): 442. DOI: 10.1037/h0022603.

[55] Kapteyn A, Neudecker H, Wansbeek T. An approach to n-mode components analysis [J/OL]. Psychometrika, 1986, 51(2): 269-275. DOI: 10.1007/BF02293984.

[56] Kruskal J B. Rank, decomposition, and uniqueness for 3-way and n-way arrays [J]. Multiway data analysis, 1989: 7-18.

[57] Uschmajew A, Vandereycken B. The geometry of algorithms using hierarchical tensors [J]. Linear Algebra and its Applications, 2013, 439(1): 133-166.

[58] De Lathauwer L, De Moor B, Vandewalle J. A multilinear singular value decomposition [J]. SIAM journal on Matrix Analysis and Applications, 2000, 21(4): 1253-1278.

[59] Schollwöck U. The density-matrix renormalization group in the age of matrix product states [J/OL]. Annals of physics, 2011, 326(1): 96-192. DOI: 10.1016/j.aop.2010.09.012.

[60] Steinlechner M. Riemannian optimization for high-dimensional tensor completion [J/OL]. SIAM Journal on Scientific Computing, 2016, 38(5): S461-S484. DOI: 10.1137/15M1010506.

[61] Helmke U, Shayman M A. Critical points of matrix least squares distance functions [J/OL]. Linear Algebra and its Applications, 1995, 215: 1-19. DOI: 10.1016/0024-3795(93)00070-G.

[62] Bruns W, Vetter U. Determinantal rings: volume 1327 [M]. Springer, 2006.

[63] Schneider R, Uschmajew A. Convergence results for projected line-search methods on varieties of low-rank matrices via Łojasiewicz inequality [J/OL]. SIAM Journal on Optimization, 2015, 25(1): 622-646. DOI: 10.1137/140957822.

[64] Koch O, Lubich C. Dynamical tensor approximation [J/OL]. SIAM Journal on Matrix Analysis and Applications, 2010, 31(5): 2360-2375. DOI: 10.1137/09076578X.

[65] Vannieuwenhoven N, Vandebril R, Meerbergen K. A new truncation strategy for the higher-order singular value decomposition [J/OL]. SIAM Journal on Scientific Computing, 2012, 34(2): A1027-A1052. DOI: 10.1137/110836067.

[66] Grasedyck L. Hierarchical singular value decomposition of tensors [J/OL]. SIAM journal on matrix analysis and applications, 2010, 31(4): 2029-2054. DOI: 10.1137/090764189.

[67] O'Shea D B, Wilson L C. Limits of tangent spaces to real surfaces [J/OL]. American journal of mathematics, 2004, 126(5): 951-980. DOI: 10.1353/ajm.2004.0040.

[68] Haegeman J, Mariën M, Osborne T J, et al. Geometry of matrix product states: Metric, parallel transport, and curvature [J/OL]. Journal of Mathematical Physics, 2014, 55(2). DOI: 10.1063/1.4862851.

[69] Holtz S, Rohwedder T, Schneider R. On manifolds of tensors of fixed TT-rank [J]. Numerische Mathematik, 2012, 120(4): 701-731.

[70] Boumal N. An introduction to optimization on smooth manifolds [M/OL]. Cambridge University Press, 2023. DOI: 10.1017/9781009166164.

[71] Absil P A, Mahony R, Sepulchre R. Optimization algorithms on matrix manifolds [M/OL]// Optimization Algorithms on Matrix Manifolds. Princeton University Press, 2009. DOI: 10.1515/9781400830244.

[72] Sato H. Riemannian conjugate gradient methods: General framework and specific algorithms with convergence analyses [J/OL]. SIAM Journal on Optimization, 2022, 32(4): 2690-2717. DOI: 10.1137/21M1464178.

[73] Shalit U, Weinshall D, Chechik G. Online learning in the embedded manifold of low-rank matrices [J]. Journal of Machine Learning Research, 2012, 13(2).

[74] Boumal N, Absil P A, Cartis C. Global rates of convergence for nonconvex optimization on manifolds [J/OL]. IMA Journal of Numerical Analysis, 2019, 39(1): 1-33. DOI: 10.1093/imanum/drx080.

[75] Da Silva C, Herrmann F J. Optimization on the hierarchical Tucker manifold–applications to tensor completion [J]. Linear Algebra and its Applications, 2015, 481: 131-173.

[76] Hosseini S, Uschmajew A. A gradient sampling method on algebraic varieties and application to nonsmooth low-rank optimization [J/OL]. SIAM Journal on Optimization, 2019, 29(4): 2853-2880. DOI: 10.1137/17M1153571.

[77] Jain P, Tewari A, Kar P. On iterative hard thresholding methods for high-dimensional m-estimation [C]//Ghahramani Z, Welling M, Cortes C, et al. Advances in Neural Information Processing Systems: volume 27. Curran Associates, Inc., 2014.

[78] Zhou G, Huang W, Gallivan K A, et al. A Riemannian rank-adaptive method for low-rank optimization [J/OL]. Neurocomputing, 2016, 192: 72-80. DOI: 10.1016/j.neucom.2016.02.030.

[79] Gao B, Absil P A. A Riemannian rank-adaptive method for low-rank matrix completion [J/OL]. Computational Optimization and Applications, 2022, 81: 67-90. DOI: 10.1007/s10589-021-00328-w.

[80] Tang T, Toh K C. Solving graph equipartition sdps on an algebraic variety [J/OL]. Mathematical Programming, 2023: 1-49. DOI: 10.1007/s10107-023-01952-6.

[81] Olikier G, Absil P A. An apocalypse-free first-order low-rank optimization algorithm with at most one rank reduction attempt per iteration [J/OL]. SIAM Journal on Matrix Analysis and Applications, 2023, 44(3): 1421-1435. DOI: 10.1137/22M1518256.

[82] Olikier G, Gallivan K A, Absil P A. First-order optimization on stratified sets [J]. ArXiv preprint arXiv:2303.16040, 2023.

[83] Kutschan B. Tangent cones to tensor train varieties [J/OL]. Linear Algebra and its Applications, 2018, 544: 370-390. DOI: 10.1016/j.laa.2018.01.012.

[84] Vermeylen C, Olikier G, Van Barel M. An approximate projection onto the tangent cone to the variety of third-order tensors of bounded tensor-train rank [J]. ArXiv preprint arXiv:2306.13360, 2023.

[85] Vermeylen C, Olikier G, Absil P A, et al. Rank estimation for third-order tensor completion in the tensor-train format [C]//2023 31st European Signal Processing Conference (EUSIPCO). IEEE, 2023: 965-969.

[86] Luo Z, Qi L. Optimality conditions for Tucker low-rank tensor optimization [J/OL]. Computational Optimization and Applications, 2023: 1-24. DOI: 10.1007/s10589-023-00465-4.

[87] Gao B, Peng R, Yuan Y x. Low-rank optimization on Tucker tensor varieties [J/OL]. Mathematical Programming, 2025, 214: 357-407. DOI: 10.1007/s10107-024-02186-w.

[88] Ha W, Liu H, Barber R F. An equivalence between critical points for rank constraints versus low-rank factorizations [J/OL]. SIAM Journal on Optimization, 2020, 30(4): 2927-2955. DOI: 10.1137/18M1231675.

[89] Levin E, Kileel J, Boumal N. Finding stationary points on bounded-rank matrices: a geometric hurdle and a smooth remedy [J/OL]. Mathematical Programming, 2023, 199(1-2): 831-864. DOI: 10.1007/s10107-022-01851-2.

[90] Naldi S. Exact algorithms for determinantal varieties and semidefinite programming [D]. INSA de Toulouse, 2015.

[91] Khrulkov V, Oseledets I. Desingularization of bounded-rank matrix sets [J/OL]. SIAM Journal on Matrix Analysis and Applications, 2018, 39(1): 451-471. DOI: 10.1137/16M1108194.

[92] Rebjock Q, Boumal N. Optimization over bounded-rank matrices through a desingularization enables joint global and local guarantees [J]. ArXiv preprint arXiv:2406.14211, 2024.

[93] Yang Y, Gao B, Yuan Y x. A space-decoupling framework for optimization on bounded-rank matrices with orthogonally invariant constraints [J]. ArXiv preprint arXiv:2501.13830, 2025.
[94] Ye K, Lim L H. Tensor network ranks [J]. ArXiv preprint arXiv:1801.02662, 2018.
[95] Swijsen L, Van der Veken J, Vannieuwenhoven N. Tensor completion using geodesics on Segre manifolds [J/OL]. Numerical Linear Algebra with Applications, 2022, 29(6): e2446. DOI: 10.1002/nla.2446.
[96] Kasai H, Mishra B. Low-rank tensor completion: a Riemannian manifold preconditioning approach [C]//Balcan M F, Weinberger K Q. Proceedings of Machine Learning Research: volume 48 Proceedings of The 33rd International Conference on Machine Learning. New York, New York, USA: PMLR, 2016: 1012-1021.
[97] Gao B, Peng R, Yuan Y x. Desingularization of bounded-rank tensor sets [J]. ArXiv preprint arXiv:2411.14093, 2024.
[98] Natarajan B K. Sparse approximate solutions to linear systems [J/OL]. SIAM journal on computing, 1995, 24(2): 227-234. DOI: 10.1137/S0097539792240406.
[99] Fazel M. Matrix rank minimization with applications [D]. PhD thesis, Stanford University, 2002.
[100] Recht B, Fazel M, Parrilo P A. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization [J/OL]. SIAM review, 2010, 52(3): 471-501. DOI: 10.1137/070697835.
[101] Candès E J, Tao T. Decoding by linear programming [J]. IEEE transactions on information theory, 2005, 51(12): 4203-4215.
[102] Candès E J, Tao T. The power of convex relaxation: Near-optimal matrix completion [J/OL]. IEEE Transactions on Information Theory, 2010, 56(5): 2053-2080. DOI: 10.1109/TIT.2010.2044061.
[103] Recht B. A simpler approach to matrix completion. [J]. Journal of Machine Learning Research, 2011, 12(12).
[104] Candès E J, Recht B. Exact matrix completion via convex optimization [J/OL]. Communications of the ACM, 2012, 55(6): 111-119. DOI: 10.1145/2184319.2184343.
[105] Ding L, Chen Y. Leave-one-out approach for matrix completion: Primal and dual analysis [J/OL]. IEEE Transactions on Information Theory, 2020, 66(11): 7274-7301. DOI: 10.1109/TIT.2020.2992769.
[106] Gandy S, Recht B, Yamada I. Tensor completion and low-n-rank tensor recovery via convex optimization [J/OL]. Inverse problems, 2011, 27(2): 025010. DOI: 10.1088/0266-5611/27/2/025010.
[107] Yuan M, Zhang C H. On tensor completion via nuclear norm minimization [J/OL]. Foundations of Computational Mathematics, 2016, 16(4): 1031-1068. DOI: 10.1007/s10208-015-9269-5.
[108] Barak B, Moitra A. Noisy tensor completion via the sum-of-squares hierarchy [C]// Conference on Learning Theory. PMLR, 2016: 417-445.
[109] Cason T P, Absil P A, Van Dooren P. Iterative methods for low rank approximation of graph similarity matrices [J/OL]. Linear Algebra and its Applications, 2013, 438(4): 1863-1882. DOI: 10.1016/j.laa.2011.12.004.
[110] Rakhuba M, Oseledets I V. Jacobi–Davidson method on low-rank matrix manifolds [J/OL]. SIAM Journal on Scientific Computing, 2018, 40(2): A1149-A1170. DOI: 10.1137/17M1123080.

[111] Krumnow C, Pfeffer M, Uschmajew A. Computing eigenspaces with low rank constraints [J/OL]. SIAM Journal on Scientific Computing, 2021, 43(1): A586-A608. DOI: 10.1137/19M1308384.

[112] Sato H, Iwai T. A Riemannian optimization approach to the matrix singular value decomposition [J/OL]. SIAM Journal on Optimization, 2013, 23(1): 188-212. DOI: 10.1137/120872887.

[113] Usevich K, Li J, Comon P. Approximate matrix and tensor diagonalization by unitary transformations: convergence of Jacobi-type algorithms [J/OL]. SIAM Journal on Optimization, 2020, 30(4): 2998-3028. DOI: 10.1137/19M125950X.

[114] Yamamoto M S, Yger F, Chevallier S. Subspace oddity-optimization on product of Stiefel manifolds for EEG data [C/OL]//ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021: 1080-1084. DOI: 10.1109/ICASSP39728.2021.9413730.

[115] Shustin B, Avron H. Riemannian optimization with a preconditioning scheme on the generalized Stiefel manifold [J/OL]. Journal of Computational and Applied Mathematics, 2023, 423: 114953. DOI: 10.1016/j.cam.2022.114953.

[116] Boumal N, Absil P A. Low-rank matrix completion via preconditioned optimization on the Grassmann manifold [J/OL]. Linear Algebra and its Applications, 2015, 475: 200-239. DOI: 10.1016/j.laa.2015.02.027.

[117] Dong S, Gao B, Guan Y, et al. New Riemannian preconditioned algorithms for tensor completion via polyadic decomposition [J/OL]. SIAM Journal on Matrix Analysis and Applications, 2022, 43(2): 840-866. DOI: 10.1137/21M1394734.

[118] Cai J F, Huang W, Wang H, et al. Tensor completion via tensor train based low-rank quotient geometry under a preconditioned metric [J]. ArXiv preprint arXiv:2209.04786, 2022.

[119] Udriste C. Convex functions and optimization methods on Riemannian manifolds: volume 297 [M/OL]. Springer Science & Business Media, 1994. DOI: 10.1007/978-94-015-8390-9.

[120] Mishra B, Sepulchre R. Riemannian preconditioning [J/OL]. SIAM Journal on Optimization, 2016, 26(1): 635-660. DOI: 10.1137/140970860.

[121] Mishra B, Apuroop K A, Sepulchre R. A Riemannian geometry for low-rank matrix completion [J]. ArXiv preprint arXiv:1211.1550, 2012.

[122] Shustin B, Avron H. Faster randomized methods for orthogonality constrained problems [J/OL]. Journal of Machine Learning Research, 2024, 25(257): 1-59. http://jmlr.org/papers/v25/21-1022.html.

[123] Tong T, Ma C, Chi Y. Accelerating ill-conditioned low-rank matrix estimation via scaled gradient descent [J/OL]. The Journal of Machine Learning Research, 2021, 22(1): 6639-6701. DOI: 10.5555/3546258.3546408.

[124] Bian F, Cai J F, Zhang R. A preconditioned Riemannian gradient descent algorithm for low-rank matrix recovery [J]. ArXiv preprint arXiv:2305.02543, 2023.

[125] Hamed M, Hosseini R. Riemannian preconditioned coordinate descent for low multilinear rank approximation [J/OL]. SIAM Journal on Matrix Analysis and Applications, 2024, 45 (2): 1054-1075. DOI: 10.1137/21M1463896.

[126] Yger F, Berar M, Gasso G, et al. Adaptive canonical correlation analysis based on matrix manifolds [C]//ICML'12: Proceedings of the 29th International Coference on International Conference on Machine Learning. Madison, WI, USA: Omnipress, 2012: 299-306.

[127] Demmel J. Nearly optimal block-Jacobi preconditioning [J/OL]. SIAM Journal on Matrix Analysis and Applications, 2023, 44(1): 408-413. DOI: 10.1137/22M1504901.

[128] Shima H, Yagi K. Geometry of Hessian manifolds [J/OL]. Differential geometry and its applications, 1997, 7(3): 277-290. DOI: 10.1016/S0926-2245(96)00057-5.

[129] Absil P, Trumpf J, Mahony R, et al. All roads lead to Newton: Feasible second-order methods for equality-constrained optimization [J]. Technical Report UCL-INMA-2009.024, 2009.

[130] yon Neumann J. Some matrix inequalities and metrization of matrix space [J]. Tomsk Univ. Rev, 1937, 1: 286-300.

[131] Shustin B, Avron H. Faster randomized methods for orthogonality constrained problems [J]. ArXiv preprint arXiv:2106.12060, 2021.

[132] Horn R A, Johnson C R. Matrix analysis [M]. Cambridge university press, 2012.

[133] Sato H, Aihara K. Cholesky QR-based retraction on the generalized Stiefel manifold [J/OL]. Computational Optimization and Applications, 2019, 72: 293-308. DOI: 10.1007/s10589-018-0046-7.

[134] Boumal N, Mishra B, Absil P A, et al. Manopt, a Matlab toolbox for optimization on manifolds [J]. The Journal of Machine Learning Research, 2014, 15(1): 1455-1459.

[135] Liu J, Musialski P, Wonka P, et al. Tensor completion for estimating missing values in visual data [J/OL]. IEEE transactions on pattern analysis and machine intelligence, 2012, 35(1): 208-220. DOI: 10.1109/TPAMI.2012.39.

[136] Breiding P, Vannieuwenhoven N. A Riemannian trust region method for the canonical tensor rank approximation problem [J/OL]. SIAM Journal on Optimization, 2018, 28(3): 2435-2465. DOI: 10.1137/17M114618X.

[137] Khoo Y, Lu J, Ying L. Efficient construction of tensor ring representations from sampling [J/OL]. Multiscale Modeling & Simulation, 2021, 19(3): 1261-1284. DOI: 10.1137/17M1154382.

[138] Zhao X, Bai M, Sun D, et al. Robust tensor completion: Equivalent surrogates, error bounds, and algorithms [J/OL]. SIAM Journal on Imaging Sciences, 2022, 15(2): 625-669. DOI: 10.1137/21M1429539.

[139] Jain P, Oh S. Provable tensor factorization with missing data [C]//Ghahramani Z, Welling M, Cortes C, et al. Advances in Neural Information Processing Systems: volume 27. Curran Associates, Inc., 2014.

[140] Andersson C A, Bro R. Improving the speed of multi-way algorithms: Part I. Tucker3 [J/OL]. Chemometrics and Intelligent Laboratory Systems, 1998, 42(1): 93-103. DOI: https://doi.org/10.1016/S0169-7439(98)00010-0.

[141] Oseledets I, Tyrtyshnikov E. TT-cross approximation for multidimensional arrays [J/OL]. Linear Algebra and its Applications, 2010, 432(1): 70-88. DOI: 10.1016/j.laa.2009.07.024.

[142] Grasedyck L, Kluge M, Kramer S. Variants of alternating least squares tensor completion in the tensor train format [J/OL]. SIAM Journal on Scientific Computing, 2015, 37(5): A2424-A2450. DOI: 10.1137/130942401.

[143] Wang W, Aggarwal V, Aeron S. Efficient low rank tensor ring completion [C/OL]// Proceedings of the IEEE International Conference on Computer Vision. 2017: 5697-5705. DOI: 10.1109/ICCV.2017.607.

[144] Chen Z, Li Y, Lu J. Tensor ring decomposition: optimization landscape and one-loop convergence of alternating least squares [J/OL]. SIAM Journal on Matrix Analysis and Applications, 2020, 41(3): 1416-1442. DOI: 10.1137/19M1270689.

[145] Acar E, Dunlavy D M, Kolda T G, et al. Scalable tensor factorizations for incomplete data [J/OL]. Chemometrics and Intelligent Laboratory Systems, 2011, 106(1): 41-56. DOI: 10.1016/j.chemolab.2010.08.004.

[146] Zhao Q, Sugiyama M, Yuan L, et al. Learning efficient tensor representations with ring-structured networks [C/OL]//ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2019: 8608-8612. DOI: 10.1109/ICASSP.2019.8682231.

[147] Srebro N, Rennie J, Jaakkola T. Maximum-margin matrix factorization [J]. Advances in neural information processing systems, 2004, 17.

[148] Yuan L, Cao J, Zhao X, et al. Higher-dimension tensor completion via low-rank tensor ring decomposition [C/OL]//2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). IEEE, 2018: 1071-1076. DOI: 10.23919/APSIPA.2018.8659708.

[149] Iannazzo B, Porcelli M. The Riemannian Barzilai–Borwein method with nonmonotone line search and the matrix geometric mean computation [J/OL]. IMA Journal of Numerical Analysis, 2018, 38(1): 495-517. DOI: 10.1093/imanum/drx015.

[150] Hestenes M R, Stiefel E. Methods of conjugate gradients for solving linear systems [J]. Journal of research of the National Bureau of Standards, 1952, 49(6): 409.

[151] Bader B W, Kolda T G. Efficient MATLAB computations with sparse and factored tensors [J/OL]. SIAM Journal on Scientific Computing, 2008, 30(1): 205-231. DOI: 10.1137/060676489.

[152] Sobral A, Zahzah E. Matrix and tensor completion algorithms for background model initialization: A comparative evaluation [J/OL]. Pattern Recognition Letters, 2016. DOI: 10.1016/j.patrec.2016.12.019.

[153] Barzilai J, Borwein J M. Two-point step size gradient methods [J/OL]. IMA J. Numer. Anal., 1988, 8(1): 141-148. DOI: 10.1093/imanum/8.1.141.

[154] Keshavan R, Montanari A, Oh S. Matrix completion from noisy entries [J]. Advances in neural information processing systems, 2009, 22.

[155] Foster D H, Reeves A. Colour constancy failures expected in colourful environments [J/OL]. Proceedings of the Royal Society B, 2022, 289(1967): 20212483. DOI: 10.1098/rspb.2021.2483.

[156] Guan Y, Dong S, Absil P A, et al. Alternating minimization algorithms for graph regularized tensor completion [J]. ArXiv preprint arXiv:2008.12876, 2020.

[157] Chu M, Del Buono N, Lopez L, et al. On the low-rank approximation of data on the unit sphere [J/OL]. SIAM Journal on Matrix Analysis and Applications, 2005, 27(1): 46-60. DOI: 10.1137/S0895479803433295.

[158] Jones R O. Density functional theory: Its origins, rise to prominence, and future [J/OL]. Reviews of modern physics, 2015, 87(3): 897-923. DOI: 10.1103/RevModPhys.87.897.

[159] Verstraete F, Cirac J I. Matrix product states represent ground states faithfully [J/OL]. Phys. Rev. B, 2006, 73: 094423. DOI: 10.1103/PhysRevB.73.094423.

[160] White S R. Density matrix renormalization group algorithms with a single center site [J/OL]. Phys. Rev. B, 2005, 72: 180403. DOI: 10.1103/PhysRevB.72.180403.

[161] Schollwöck U. The density-matrix renormalization group in the age of matrix product states [J/OL]. Annals of Physics, 2011, 326(1): 96-192. DOI: https://doi.org/10.1016/j.aop.2010.09.012.

[162] Bravyi S, Kitaev A. Universal quantum computation with ideal clifford gates and noisy ancillas [J/OL]. Physical Review A, 2005, 71(2). DOI: 10.1103/physreva.71.022316.

[163] Bravyi S, Smith G, Smolin J A. Trading classical and quantum computational resources [J/OL]. Phys. Rev. X, 2016, 6: 021043. DOI: 10.1103/PhysRevX.6.021043.

[164] Hastings M B. Superadditivity of communication capacity using entangled inputs [J/OL]. Nature Physics, 2009, 5(4): 255–257. DOI: 10.1038/nphys1224.

[165] Werner R F, Holevo A S. Counterexample to an additivity conjecture for output purity of quantum channels [J/OL]. Journal of Mathematical Physics, 2002, 43(9): 4353–4357. DOI: 10.1063/1.1498491.

[166] Cubitt T, Harrow A W, Leung D, et al. Counterexamples to additivity of minimum output p-Rényi entropy for p close to 0 [J/OL]. Communications in Mathematical Physics, 2008, 284(1): 281–290. DOI: 10.1007/s00220-008-0625-z.

[167] Grudka A, Horodecki M, Pankowski L. Constructive counterexamples to the additivity of the minimum output Rényi entropy of quantum channels for all $p \geq 2$ [J/OL]. Journal of Physics A: Mathematical and Theoretical, 2010, 43(42): 425304. DOI: 10.1088/1751-8113/43/42/425304.

[168] Szczygielski K, Studziński M. New constructive counterexamples to additivity of minimum output Rényi p-entropy of quantum channels [J/OL]. IEEE Transactions on Information Theory, 2024, 70(10): 7023–7035. DOI: 10.1109/tit.2024.3446191.

[169] Lee J M. Smooth manifolds [M/OL]. Springer, 2012. DOI: 10.1007/978-1-4419-9982-5_1.

[170] Holtz S, Rohwedder T, Schneider R. The alternating linear scheme for tensor optimization in the tensor train format [J]. SIAM Journal on Scientific Computing, 2012, 34(2): A683-A713.

[171] Shor P W. Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer [J/OL]. SIAM Journal on Computing, 1997, 26(5): 1484-1509. DOI: 10.1137/s0097539795293172.

[172] Grover L K. A fast quantum mechanical algorithm for database search [Z]. 1996.

[173] Gottesman D. Stabilizer codes and quantum error correction [M]. California Institute of Technology, 1997.

[174] Bravyi S, Gosset D. Improved classical simulation of quantum circuits dominated by clifford gates [J/OL]. Phys. Rev. Lett., 2016, 116: 250501. DOI: 10.1103/PhysRevLett.116.250501.

[175] Bravyi S, Browne D, Calpin P, et al. Simulation of quantum circuits by low-rank stabilizer decompositions [J/OL]. Quantum, 2019, 3: 181. DOI: 10.22331/q-2019-09-02-181.

[176] Aaronson S, Gottesman D. Improved simulation of stabilizer circuits [J/OL]. Phys. Rev. A, 2004, 70: 052328. DOI: 10.1103/PhysRevA.70.052328.

[177] Leone L, Oliviero S F E, Hamma A. Stabilizer Rényi entropy [J/OL]. Physical Review Letters, 2022, 128(5): 50402. DOI: 10.1103/PhysRevLett.128.050402.

[178] Haug T, Piroli L. Quantifying nonstabilizerness of matrix product states [J/OL]. Physical Review B, 2023, 107(3). DOI: 10.1103/physrevb.107.035148.

[179] Gao B, Peng R, Yuan Y x. First-order methods on bounded-rank tensors converging to stationary points [J]. ArXiv preprint arXiv:2503.04523, 2025.

[180] Gao B, Peng R, Yuan Y x. Quotient geometry of tensor ring decomposition [J]. ArXiv preprint arXiv:2601.21874, 2026.

致 谢

从高中毕业至今, 我已在数学的道路上行走了近十年. 一路从同济大学数学科学学院, 到中国科学院数学与系统科学研究院, 这段求学旅程既漫长又深刻. 回望来时路, 我从最初的青涩与犹疑, 逐渐走向更加自信、也更加坚定的自己. 数学不仅塑造了我的逻辑思维能力, 也悄然改变了我看待世界、理解问题的方式. 回首在科学院度过的 5 年时光, 恍若弹指一挥间. 如今离别在即, 心中既有对过往岁月的感激, 也生出了几分不舍, 这些情绪交织在此刻的心头.

感谢我的导师袁亚湘研究员. 我与袁老师结识于 2020 年的夏天, 那时还在疫情防控阶段, 第一次线上见到袁老师的场景仿佛还在昨日, 袁老师和蔼可亲与平易近人的形象给我留下了深刻的印象. 在科研上, 袁老师给予我充分的自由, 让我能在张量优化与流形优化的海洋里自由的探索. 每次和袁老师讨论后, 我总是茅塞顿开, 会有一些新的想法迫不及待地去尝试; 而当科研工作遭遇瓶颈时, 袁老师也常以自身的学术经历与人生体会鼓励我坚定信心、继续前行. 我最敬佩袁老师的是, 不管是谁的报告, 不管是什么方向的研究, 袁老师总能一针见血点出问题的本质. 张量优化涉及复杂的数学符号与运算规则, 要在繁复公式中清晰传达核心思想并非易事, 但袁老师总能在我 “词穷” 的时候, 帮助我给出最直观、最形象也最本质的解释. 耳濡目染下, 我也尝试学着袁老师严谨、多角度、更本质地思考问题, 而这个能力值得我一辈子去体会去学习. 同时, 感谢袁老师给我很多外出参加学术会议的机会, 让我开阔了学术视野, 结识了许多学术前辈、同行以及朋友. 在生活中, 袁老师亦如一位长者, 乃至父亲般关心着我和我的家人. 袁老师兴趣爱好广泛, 每次登山总是身先士卒、走在第一梯队, 激励着学生们不断前行. 一日为师终身为父, 能够成为袁老师的学生, 是我一生中最幸福、也最幸运的事情. 衷心感谢袁老师照亮了我前行的道路. 未来的路上, 我将更加努力, 继续以袁老师为榜样, 潜心学问, 不负所托.

感谢中国科学院数学与系统科学研究院的高斌副研究员在博士生涯中的合作中对我的帮助. 高老师在讨论学术时总能提出有意思的问题, 将我的科研想法变得更加成熟, 并针对我遇到的问题提出相应的建议. 高老师最令我佩服的是对待科研认真严谨的态度, 不放过任何一个小问题, 修改文章时每一句话都会斟酌多次, 力求达到完美. 感谢香港科技大学 (广州) 的王鑫教授以及朱成开博士在量子信息方面的指导与帮助.

感谢戴彧虹老师在讨论班上提出的一些富有启发性的建议. 这些建议使我的科研工作更加完善. 感谢课题组的刘歆老师、马俊杰老师、陈亮老师、陈圣杰老师、北京邮电大学刘亚锋老师、孙聪老师以及中国科学院大学王小玉老师对我的照顾与帮助. 感谢已在外工作曾给予我帮助的师兄师姐们: 王彦飞, 文再文, 夏勇, 马士谦, 张在坤师兄, 牛凌峰, 王晓等师姐.

感谢计算数学所刘颖、陈瑾、李雨霏、衡思宇、丁如娟、钱莹、研究生部刘

霞主任、尹永华、陆凤斌、卢佳艺、财务处袁晨处长、科研处王晓欢处长、王丽等老师们的帮助. 感谢你们为我们的科研生活保驾护航.

感谢和我同期的课题组的师兄弟姐妹们. 感谢黄磊、姜博鸥、陈圣杰、王磊、陈硕、谢鹏程、裴骞、王圣超、胡雨宽、武哲宇、张亦、章煜海、胡雨婷、刘上琳、王子岳等师兄师姐们对我的关心与照顾. 感谢一同入学的李博文、汤宇杨、苏昭纲、李冠达、姜林硕、郑浩然、范熙来同学, 我会记得我们一起在雁栖湖、中关村校区一同求学的时光. 感谢杨俨、胡威、刁若愉、张宇航、岳艺双、张思远、罗舟行、徐勐、李新鹏、李雨芯、刘亚琛、金泽龙、孙胜煜、王兆维、王宇杨、黄辰飞、张博洋、丁一翕、邹海军、杨茹赟、邓家懿、杨景添、苏园茗、阮博元、王拓、熊鑫辉等师弟师妹们的陪伴与帮助. 感谢张凯丽、章丽、张旭、张帆、魏奇远、彭真、李世茹等课题组博士后们对我的关爱与照顾. 感谢舍友李岩、薛舒晨同学这五年以来的关心和照顾, 有你们的相伴科研生活更加丰富.

感谢高斌副研究员、彭真博士、杨俨、熊鑫辉、郝鹏飞、胡同欣同学、香港科技大学 (广州) 朱成开博士审阅我的论文初稿并提出了许多宝贵的建议.

感谢我的朋友们官毅、郭纪新、胡煊赫、彭李轶哲、唐金龙、王浩羽、王一凡、杨杰翔、周正浩, 感谢你们在我感到彷徨的时候向我伸出援手. 感谢在机场跑道尽头、在阳光与暮色中与我一同等待航班的朋友们: 陈孟涛、陈睿轩、方圣超、甘鹏翼、龚翰威、胡雨廷、康雨新、李登安、李木子、梁裕、刘炳睿、刘家璇、刘义鹏、刘宗桓、陆芝演、孟远飞、田文雨、王博文、王绍恒、王文铉、王啸越、王泽睿、王智博、王子铭、王梓旭、吴庆彤、闫博瑞、颜绍铭、燕一晨、尹赫、尤政旭、张宇飞、赵乐、赵栩枫、赵中琦、郑佳鑫, 我们一起用相机记录下飞机—人类工业皇冠上的明珠—的优雅身姿, 你们让我在科研之外, 始终记得世界的辽阔, 愿你我的人生都能起落安妥.

最后, 感谢我的父母和家人们多年来对我一如既往的关爱、支持和照顾. 你们的爱一直陪伴着我, 一直鼓励着我, 让我有勇气面对一切困难与挫折. 愿我的努力能不负你们的期望.

2026 年 6 月

作者简历及攻读学位期间发表的学术论文与研究成果

作者简历:

彭任锋, 湖南省长沙市人, 出生于 1998 年 9 月.

2017 年 9 月—2021 年 7 月, 在同济大学数学科学学院获得学士学位.

2021 年 9 月—2026 年 7 月, 在中国科学院数学与系统科学研究院攻读博士学位.

邮箱: pengrenfeng@lsec.cc.ac.cn　个人主页: https://jimmypeng1998.github.io

获得奖励:

(1) 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
(2) 2022 年 6 月　中国科学院大学三好学生
(3) 2023 年 6 月　中国科学院数学与系统科学研究院优秀共青团员
(4) 2024 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) 2025 年 9 月　中国科学院数学与系统科学研究院院长奖学金特别奖

已发表（或正式接受）的学术论文:

(1) Bin Gao, **Renfeng Peng**, Ya-xiang Yuan, *Riemannian Preconditioned Algorithms for Tensor Completion via Tensor Ring Decomposition*, Computational Optimization and Applications, 2024, 88(2): 443–468. (作者署名按姓氏排序.)
(2) Bin Gao, **Renfeng Peng**, Ya-xiang Yuan, *Low-rank optimization on Tucker tensor varieties*, Mathematical Programming, 2025, 214: 357–407. (作者署名按姓氏排序.)
(3) Bin Gao, **Renfeng Peng**, Ya-xiang Yuan, *Optimization on product manifolds under a preconditioned metric*, SIAM Journal on Matrix Analysis and Applications, 2025, 46(3): 1816–1845. (作者署名按姓氏排序.)
(4) Chengkai Zhu, **Renfeng Peng**, Bin Gao, and Xin Wang. *Riemannian optimization for Holevo capacity*, 2025 IEEE International Symposium on Information Theory (ISIT), Ann Arbor, MI, USA, 2025, pp: 1–6.

已完成的学术论文:

(1) Bin Gao, **Renfeng Peng**, Ya-xiang Yuan, *Desingularization on bounded-rank tensor sets*, In: arXiv preprint arXiv:2411.14093 (2024). (作者署名按姓氏排序.)
(2) Bin Gao, **Renfeng Peng**, Ya-xiang Yuan. *First-order methods on bounded-rank*

tensors converging to stationary points. In: arXiv preprint arXiv:2503.04523 (2025). (作者署名按姓氏排序.)

(3) **Renfeng Peng**, Chengkai Zhu, Bin Gao, Xin Wang, Ya-xiang Yuan, *Normalized tensor train decomposition*. In: arXiv preprint arXiv:2511.0436 (2025).

(4) Bin Gao, **Renfeng Peng**, Ya-xiang Yuan, *Quotient geometry of tensor ring decomposition*. In: arXiv preprint arXiv:2601.21874 (2026). (作者署名按姓氏排序.)