arXiv:2601.20152v2 [math.ST] 9 Feb 2026

可交换张量与矩阵值数据的集中不等式

Chen Cheng* Rina Foygel Barber*

2026年2月11日

摘要

我们研究随机数据结构化加权和的集中不等式，包括 (i) 张量内积和 (ii) 序列矩阵和。我们关注可交换性条件下这些结构化加权和的尾界与集中不等式，将经典的独立项框架推广至此。

我们在结构依赖的可交换性条件下推导了 Hoeffding 界和 Bernstein 界。在此过程中，我们将可交换随机变量加权和以及独立同分布随机矩阵和的已知结果推广至最优常数。值得注意的是，相较于先前基于 Chatterjee 的可交换对方法推导的结果，我们为矩阵数组的组合和推导出了更紧的集中界。

在应用方面，更丰富的结构为我们提供了新颖的分析工具，可用于估计多因素响应模型的平均效应，并研究联邦平均中的固定设计草图方法。我们将结果应用于这些问题，并发现数值证据证实了我们的理论预测。

目录

1 引言 **2**1.1 示例 .31.2 相关工作 .5

2 模态可交换张量的集中不等式 **6**
2.1 霍夫丁界 .82.2 Bernstein 界 .82.2.1 比较可交换性与独立性：维度的作用 9

3 集中不等式 矩阵值数据的不等式 9
3.1 霍夫丁界（一般情形）.103.2 霍夫丁界（满足交换性条件时） .103.3 Bernstein 界（满足交换性条件时） .11

4 理论推论 **12**
4.1 可交换随机变量的加权和 .124.2 与独立同分布情形的联系 .144.2.1 模可交换张量与独立同分布情形的比较 .14

*统计学系，芝加哥大学

4.2.2 可交换性与独立同分布界之间的插值 .154.2.3 与独立同分布矩阵集中不等式的联系 .164.3 矩阵数组的组合和 .16

5 应用 **17**
5.1 估计多因子平均效应的实验设计 .175.2 基于离散正弦变换的固定设计草图化 .195.3 通信高效的岭调优联邦平均 .22

6 讨论 **23**

A 技术引理 **28**

B 第2节证明 **30**
B.1 定理1的证明 .30B.2 定理 2 的证明 .32B.3 引理B.1的证明 .32B.4 引理 B.3 的证明 .33

C 第 3 节的证明 **35**
C.1 定理 4 的证明 .35C.2 定理 5 的证明 .38C.3 引理 C.4 的证明 .40

D 理论推论与应用的证明 **41**
D.1 与独立同分布情形相关的证明 .41D.2 推论4.3 的证明 .42D.3 推论 5.1 的证明 .42D.4 推论 5.2 的证明 .43

1 引言

集中不等式的一个核心目标，是为中心化随机变量 $Z - \mathbb{E}Z$ 推导尾界。该变量等于随机数据的加权和，其最简单的形式为

$$Z = \langle w, X \rangle := w_1 X_1 + w_2 X_2 + \cdots + w_N X_N,$$

其中，$w = (w_1, \cdots, w_N)^\top$ 是 $\mathbb{R}$ 中的固定权重，而 $X = (X_1, \cdots, X_N)^\top$ 是 $[-1, 1]$ 中的独立同分布随机变量。在独立同分布情形下，关于界定 $\mathbb{P}(Z - \mathbb{E}Z \geq t)$ 的一些经典结果广为人知，例如 Hoeffding 界和 Bernstein 界，以及鞅情形下的 Azuma-Hoeffding 界 [45, 第 2] 章。近期的工作 [3]，中，Barber 将 Hoeffding 界和 Bernstein 界推广到了可交换情形：在任意置换 π 作用于 $[N]$ 时，随机数据 $(X_1, \cdots, X_N)$ 的联合分布与 $(X_{\pi(1)}, \cdots, X_{\pi(N)})$ 相同。

因此，一个自然的问题是（这在一定程度上由 Barber 提出）：这类尾界如何推广到具有更丰富结构的可交换性？我们考虑一个固定权重（或一组有序权重） $W \in \mathbb{T}$ 以及一个来自向量空间 $\mathbb{T}, \mathbb{T}'$ 的随机数据对象 $X \in \mathbb{T}'$，其中 X 在与 $\mathbb{T}'$ 相关联的对称群作用下具有分布不变性（即具有可交换性）。我们将感兴趣的量 Z 概括为双线性型 $Z := W, X \in \mathbb{T}''$，其中 $\mathbb{T}'' = \mathbb{R}$ 或 $\mathbb{R}^{n \times m}$。我们的目标是为 $\mathbb{P}(Z - \mathbb{E}Z > t)$ 或 $\mathbb{P}(\|Z - \mathbb{E}Z\| > t)$ 提供 Hoeffding 型和 Bernstein 型尾界。我们考虑两种情形：

1. 模态可交换张量： 第一种情形考虑一个固定的 K 模张量 W，以及一个满足沿每个模可交换（定义见下文）的随机张量 X。为了便于理解，我们先介绍一个具有 $K = 2$ 个模的特殊情形： W 和 X 都是矩阵， $W,\ X \in \mathbb{R}^{N\times M}$，且 X 是**按行和按列可交换的**，也就是说，对行和列进行置换后的矩阵 $(X_{\pi(i)\pi'(j)})_{i\in[N],\ j\in[M]}$ 与 $X = (X_{ij})_{i\in[N],\ j\in[M]}$ 具有相同的分布，其中 π 是 $[N]$ 上的任意置换， π' 是 $[M]$ 上的任意置换。随后，我们研究实值内积 $Z = \sum_{i=1}^{N}\sum_{j=1}^{M} W_{ij}X_{ij} \in \mathbb{R}$ 的集中（更一般地，Z 由张量内积给出）。

2. 矩阵值可交换数据： 令 $W_1,\cdots,W_N \in \mathbb{R}^{p\times q}$ 为固定值，并令 $X_1,\cdots,X_N \in \mathbb{R}^{q\times r}$ 为随机变量，使得在 $[N]$ 上进行任意置换 π 后，$(X_1,\cdots,X_N)$ 与 $(X_{\pi(1)},\cdots,X_{\pi(N)})$ 具有相同的联合分布。随后，我们建立关于矩阵值和 $Z = \sum_{k=1}^{N} W_kX_k \in \mathbb{R}^{p\times r}$ 的集中界。

我们在表 1 中展示了这些设定的一个示例。本文的主要理论贡献包括第 2 节，其中我们研究模可交换张量的集中不等式；以及第 3 节，其中我们转而讨论可交换矩阵值数据。本文其余部分安排如下。在第 4 节中，我们给出结果的理论推论，恢复 Barber 关于可交换加权和的结果，并将其应用于独立同分布的张量型数据和矩阵值数据。特别是，我们的工具可以推出关于 Chatterjee [10], Mackey 等人 [30] 研究的矩阵数组组合和的尾界。在第 5 节中，我们将工具应用于两个实际问题，并提出具体的理论预测，同时通过数值模拟提供实证证据。

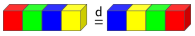
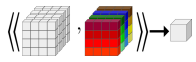
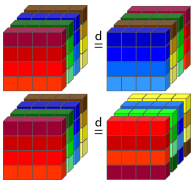
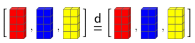
W 和 X	$\mathbb{T}$	$\mathbb{T}'$	$W, X \mapsto Z$	可交换性 X的假设
标量序列	$\mathbb{R}^N$	$\mathbb{R}^N$		
张量	$\mathbb{R}^{N_1\times\cdots\times N_K}$	$\mathbb{R}^{N_1\times\cdots\times N_K}$		
矩阵值数据	$\mathbb{R}^{N\times p\times q}$	$\mathbb{R}^{N\times q\times r}$		

表1. 本文所考虑情形的示意图：将 [3] 中研究的可交换标量加权和（上排）推广至模态可交换张量（中排），以及矩阵值数据的可交换序列（下排）。

1.1 示例

在形式化数学框架并引入集中界之前，我们先通过两个现实世界的例子进行说明：在这些例子中，传统的独立同分布或可交换假设并不成立，但随机性中更丰富的对称结构使我们能够运用新颖的分析工具。我们将在第5节中再次讨论这些例子。

下面的第一个例子涉及沿每个模都具有可交换性的张量值数据。

例1（多因素响应的平均效应）**：** 考虑一个多因子模型，其中因素 $k \in [K]$ 具有离散水平 $[N_k]$。多重索引 $(i_1,\ \cdots,\ i_K)$ $i_l \in [N_l]$ 对应于一个

实验条件，其中第 k个因素为 i_k。对于每种实验设置，进行测试会产生随机且可能被删失的响应 $Y_{i_1\cdots i_K}\in\mathbb{R}$。设 $\sum_{i_l\in[N_l],\ l\in[K]} p_{i_1\cdots i_K}=1$ 是定义在 $[N_1]\times\cdots\times[N_K]$上的概率分布 P，我们感兴趣的是估计多因素响应的平均效应

$$\mu=\mathbb{E}_{(i_1,\cdots,i_K)\sim P}[Y_{i_1\cdots i_K}]=\sum_{i_l\in[N_l],l\in[K]} p_{i_1\cdots i_K}Y_{i_1\cdots i_K}.$$

在实践中，观测以下内容可能并不可行： $Y_{i_1\cdots i_K}$ 对于所有 $N_1\cdot\cdots\cdot N_K$ 可能的因素水平组合。我们可以改为通过对子集的笛卡尔积进行子抽样，考虑以下平均效应估计程序。具体而言，我们无放回地选择$I_l\subset[N_l]$ 对于所有 $l\in[K]$ 满足 $|I_l|=n_l$的样本，并令 $X_{i_1\cdots i_K}=\prod_{l=1}^K \mathbb{1}_{i_l\in I_l}$ 为抽样实验的指示随机变量——显然， $X_{i_1\cdots i_K}$ 沿每个固定因素水平可交换（我们将在下文的定义2.1中正式定义， X 是模态可交换的）。在各种实际问题中，即使与样本数量相同的独立抽样相比，这种笛卡尔积子抽样方案也颇具吸引力，因为为单个实验更改因素设置可能代价高昂，而批量实验则能大幅降低成本。多个实际领域都促使人们采用此类降本方案，包括多种因素（温度、荷电状态窗口、充放电倍率和放电深度）作用下锂离子电池的老化 [36, 39]、癌症领域的高通量药物组合筛选（药物 A、药物 B、剂量水平和细胞系的联合效应） [20] ，以及半导体制造工艺优化（因素包括气体流量、腔室压力、射频功率和电极间距） [32]。

假设我们知道 $p_{i_1\cdots i_K}$ （或者我们有频率的经验估计）。在给定真实响应 $Y_{i_1\cdots i_K}$ 的条件下，令 $W_{i_1\cdots i_K}=p_{i_1\cdots i_K}Y_{i_1\cdots i_K}$，我们可以使用经典的霍维茨-汤普森估计量来估计平均效应：

$$\widehat{\mu}=\frac{N_1\cdots N_K}{n_1\cdots n_K}\cdot Z=\frac{N_1\cdots N_K}{n_1\cdots n_K}\cdot\langle\!\langle W,X\rangle\!\rangle:=\frac{N_1\cdots N_K}{n_1\cdots n_K}\sum_{i_l\in[N_l],l\in[K]} W_{i_1\cdots i_K}X_{i_1\cdots i_K}.$$

接下来要研究 $\widehat{\mu}-\mathbb{E}[\widehat{\mu}\,|\,\{Y_{i_1\cdots i_K}\}_{i_l\in[N_l],\ l\in[K]}]$的集中，尤其是在子抽样预算 $n_1\cdots n_K\le B$固定的情况下，如何将估计误差降至最低。◇

在下一个示例中，我们将讨论统计机器学习中广泛用作降维方法的一类草图技术，其中 W， X 以及感兴趣的量 $\langle\!\langle W,\ X\rangle\!\rangle$属于表 1中的矩阵值领域。在高斯草图和计数草图等经典设置中，草图矩阵彼此独立。我们考虑从固定设计草图矩阵池中进行子抽样，以此作为现有方法的替代方案。尽管这些矩阵仍是随机的，但这将独立性要求降低为可交换性。

例2 （固定设计草图矩阵）**：** 考虑一个由 N 可交换参数 $\theta_k\in\mathbb{R}^{q\times r}$构成的序列，我们关注参数的平均值 $\theta=\frac{1}{N}\sum_{k=1}^N\theta_k$。在分布式优化设置中， θ_k 是局部参数，例如联邦平均 [24, 13] 和协作学习 [12], ，计算密集地在本地代理上进行，而且中央服务器与本地代理之间的通信成本往往受到限制。因此，理想的做法是采用草图技术压缩本地数据 θ_k：我们将使用维度为 $q'\ll q$ 的草图矩阵 $U_k\in\mathbb{R}^{q'\times q}$，它们满足

$$\sum_{k=1}^N U_k^\top U_k\approx I_q.$$

随后将本地数据 $\theta_k \in \mathbb{R}^{q\times r}$ 压缩为 $z_k = U_k\theta_k \in \mathbb{R}^{q'\times r}$，本地代理将$z_k$ 传输给中央服务器，中央服务器再对压缩数据进行聚合：

$$\widehat{\theta} = \sum_{k=1}^{N} U_k^\top z_k = \sum_{k=1}^{N} U_k^\top U_k \theta_k.$$

由于 $\sum_{k=1}^{N} U_k^\top U_k \approx I_q$，因此我们可以预期 $\widehat{\theta} \approx \overline{\theta}$。

标准方法采用独立的草图矩阵 U_k，例如高斯随机矩阵、Rademacher 随机矩阵和子采样随机傅里叶变换（SRFT）矩阵 [18]。其他常用方法包括计数草图和行列子采样技术，它们可以降低计算成本，因为 U_k 是稀疏的 [8, 31]。

在我们的设定中，鉴于本地代理之间具有可交换性 $\theta_1, \cdots, \theta_N$，我们可以采用一组固定的草图矩阵 U_k，并满足精确等式 $\sum_{k=1}^{N} U_k^\top U_k = I_q$，这有助于减小 $\sum_{k=1}^{N} U_k^\top U_k \approx I_q$随机草图化设定中产生的近似误差。记 $W_k = U_k^\top U_k$和$X_k = \theta_k$，则有$\hat{}\theta = Z = \langle\!\langle W, X \rangle\!\rangle = \sum_{k=1}^{N} W_k X_k$，我们的结果进而确立了表1中矩阵值可交换性条件下的集中保证。◇

1.2 相关工作

针对独立随机变量之和及其加权变体的经典集中不等式，如今已是标准结果，最常用的形式是 Hoeffding [21] 界和 Bernstein 界 [6], 。 [7, 45]对于具有独立加项的矩阵值和，Tropp的一系列重要工作将随机变量之和推广到了矩阵 Hoeffding 不等式、Bernstein不等式及其他变体 [43, 44],，并借助矩阵拉普拉斯变换和算子不等式等工具。技术思路可参见Ahlswede 和 Winter的工作 [1]以及 Lieb 不等式 [29]。

关于独立性之外的类似紧界，已知结果较少，即使对于可交换有界随机变量也是如此 $X_1, \cdots$ 而可交换性 X_N 在保留强对称性的同时，为独立性提供了一种典型的放宽。德·费内蒂著名的表示定理 [14] 表明，可以得到一般而紧致的 Hoeffding 界和 Bernstein 界，并且在独立同分布情形下达到最优常数 $N \to \infty$。例如， [37], 作者证明了无限可交换性条件下具有最优常数的均匀 Hoeffding 不等式。Diaconis 对可交换序列的分布近似进行了定量刻画：当 N 有限时，可用独立同分布律的混合在全变差距离意义下进行近似 [15]。另见 [16]中关于库尔贝克–莱布勒散度的近期变体。在无权情形下 $w_1 = w_2 = \cdots = w_n = 1$， $w_{n+1} = \cdots = w_N = 0$通常将其视为无放回抽样，Serfling 不等式 [40] 给出了比 Hoeffding 界更强的精确刻画。近期的研究 [4, 17]针对这一特定设定推导出了更精细的界。

Barber近期的研究探讨了有限可交换性情形下的集中不等式，并为可交换序列的一般加权和推导出了精确的 Hoeffding 界和 Bernstein 界 [3]。受这项研究的直接启发，本文将相关结果推广到 (i) 模可交换张量，以及 (ii) 由可交换序列构成的矩阵值和；这也将结果大幅推广到更丰富的对称结构类别。在随机矩阵和的设定下，研究依赖情形下集中不等式的另一条相关路线是Chatterjee的可交换对方法 [9, 10, 30, 19]。

可交换性下的序列数据 $X_1, \cdots, X_N$ 在现代统计学中十分常见，例如置换检验和保形预测 [38, 41, 27, 2]。除了多因子模型和固定设计草图化示例之外，模可交换性和随机矩阵的可交换和在实际应用中也很有用，包括二部网络中的 U统计量 [26] 以及组合非参数回归 [19],，这些领域都可能用到我们提出的新工具。

记号。对于任意整数 $n \geq 1$，$[n]$ 表示集合 $\{1, \cdots, n\}$。对于向量 $x \in \mathbb{R}^n$ 和 $p \geq 1$，我们用 $\|x\|_p$ 表示其 ℓ_p 范数，并用 $\|x\|$ 表示 $p = 2$。对于方阵 $A \in \mathbb{R}^{n \times n}$，我们用$\|A\|$ 表示其算子范数，用 $\|A\|_F$ 表示其Frobenius范数。我们用 $\mathbb{S}^n$ 表示 $\mathbb{R}^{n \times n}$ 中所有对称实值矩阵构成的向量空间，用 $\mathbb{S}^n_+$ 表示 $\mathbb{S}^n$中所有对称半正定矩阵构成的凸锥。令 $\mathbb{S}^n_{++}$ 为对称正定矩阵锥。令 $\preceq$ 表示由锥 $\mathbb{S}^n_+$ 诱导的偏序，$\prec$ 表示由 $\mathbb{S}^n_{++}$诱导的偏序。对于满足$m \leq n$的矩阵 $A \in \mathbb{R}^{n \times m}$，其伪逆记为 $A^\dagger$。我们将投影到 A的列空间上的投影矩阵定义为 $\mathrm{P}_A := A(A^\top A)^\dagger A^\top$，其正交补的投影矩阵记为 $\mathrm{P}_A^\perp := I_n - A(A^\top A)^\dagger A^\top$。特别地，对于单位向量 $w \in \mathbb{R}^n$，$\mathrm{P}_w = ww^\top$ 和 $\mathrm{P}_w^\perp = I_n - ww^\top$。令 $\mathscr{S}_N$ 为 $[N]$中所有置换映射构成的集合，令 $\mathscr{P}_N$为 $\mathbb{R}^{N \times N}$中所有置换矩阵构成的集合。我们用 $\mathsf{d}=$表示分布相等。

接下来，我们定义一些用于处理张量的其他符号和术语。对于两个张量 $A \in \mathbb{R}^{N_1 \times \cdots \times N_K}$ 和 $B \in \mathbb{R}^{N'_1 \times \cdots \times N'_{K'}}$，它们的外积 $A \otimes B \in \mathbb{R}^{N_1 \times \cdots \times N_K \times N'_1 \times \cdots \times N'_{K'}}$ 定义为

$$(A \otimes B)_{i_1 \cdots i_K i'_1 \cdots i'_{K'}} = A_{i_1 \cdots i_K} B_{i'_1 \cdots i'_{K'}} \qquad \text{for all } \ i_l \in [N_l], i'_{l'} \in [N'_{l'}],\ l \in [K], l' \in [K'].$$

我们还会经常用到内积，

$$\langle A, B \rangle_{\mathbb{T}} = \sum_{i_1=1}^{N_1} \cdots \sum_{i_K=1}^{N_K} A_{i_1 \cdots i_K} B_{i_1 \cdots i_K} \in \mathbb{R}.$$

回顾一下 k-模张量-矩阵乘积（参见 [25, 第 2 节。5]），对于 $A \in \mathbb{T}$ 和一个矩阵 $B \in \mathbb{R}^{M \times N_k}$，$A \times_k B \in \mathbb{R}^{N_1 \times \cdots \times N_{k-1} \times M \times N_{k+1} \times \cdots \times N_K}$ 其元素为

$$(A \times_k B)_{i_1 \cdots i_{k-1} j i_k \cdots i_K} = \sum_{i_k=1}^{N_k} B_{j i_k} X_{i_1 \cdots i_K}.$$

我们用 $A := \frac{\top}{N_1 \cdots N_K} \sum_{i_1=1}^{N_1} \cdots \sum_{i_K=1}^{N_K} A_{i_1 \cdots i_K} = A, \mathbf{P}^{\mathbf{1}_{N_1}} \otimes \cdots \mathbf{P}^{\mathbf{1}_{N_K}}$ 表示张量所有元素的均值。类似于矩阵的 Frobenius范数，我们可以定义希尔伯特空间 $\mathbb{T}$中的张量 A的范数为 $\|A\|_{\mathbb{T}} := \sqrt{A, A_{\mathbb{T}}}$。我们还记$\|A\|_1 := \sum_{i_l \in [N_l], l \in [K]} |A_{i_1 \cdots i_K}|$和 $\|A\|_\infty := \max_{i_l \in [N_l], l \in [K]} |A_{i_1 \cdots i_K}|$。

2 模态可交换张量的集中不等式

本节将讨论张量结构化数据 $X \in \mathbb{R}^{N_1 \times \cdots \times N_K}$，其中 $X_{i_1 \cdots i_K} \in [-1, 1]$ 对所有 $i_k \in [N_k]$，$k \in [K]$均成立。这里 K 是张量的模数，特殊情形包括 $K = 1$ （向量）和 $K = 2$ （矩阵）。在为一般的 K情形定义问题之前，我们先分别讨论这两种特殊情形。对于向量情形（$K = 1$），正如 Barber [3], 所研究的，我们有

$$Z = \langle w, X \rangle := w^\top X = \sum_{i=1}^N w_i X_i.$$

在这个向量情形中，可交换性意味着 $(X_1, \cdots, X_N)$ 的联合分布在对索引进行任意置换后保持不变，即对于所有 $\pi \in \mathscr{S}_N$，$(X_1, \cdots, X_N) \mathsf{d} = (X_{\pi(1)}, \cdots, X_{\pi(N)})$。接下来，对于矩阵情形 $(K = 2)$，我们有

$$Z = \langle W, X \rangle := \mathsf{Tr}(W^\top X) = \sum_{i=1}^N \sum_{j=1}^M W_{ij} X_{ij},$$

此时，可交换性假设为行列可交换性：对于所有 $\pi \in \mathscr{S}_N, \pi' \in \mathscr{S}_M$，$(X_{ij})_{i\in[N],j\in[M]} \overset{\mathrm{d}}{=} (X_{\pi(i)\pi'(j)})_{i\in[N],j\in[M]}$。在这两个问题中，权重向量 $w \in \mathbb{R}^N$ 或权重矩阵 $W \in \mathbb{R}^{N\times M}$ 都是固定的。

现在，我们转向模态可交换 张量的一般形式。这一形式自然地推广了上述两种情形，并支持对更复杂数据结构进行细致分析。在本节中， W 和 X 都属于 $\mathbb{T} := \mathbb{R}^{N_1\times\cdots\times N_K}$。我们关注它们在张量空间中的内积的集中，

$$Z = \langle W, X\rangle_{\mathbb{T}} := \sum_{i_1=1}^{N_1} \cdots \sum_{i_K=1}^{N_K} W_{i_1\cdots i_K} X_{i_1\cdots i_K},$$

其中， W 是一个固定的 K模张量，而 X 是满足模可交换性的随机张量，定义如下。

定义 2.1（模可交换性）。对于随机张量 $X \in \mathbb{T}$，如果对任意 $l \in [K]$ 和置换矩阵 $\Pi_l \in \mathscr{P}_{N_l}$，

$$X \times_l \Pi_l \overset{\mathrm{d}}{=} X.$$

也就是说， X 的分布在沿任意一个模进行置换时保持不变。等价地，对于任意一组置换矩阵 $\Pi_k \in \mathscr{P}_{N_k}$、 $k = 1$、2、 $\cdots$、 K，

$$(X \times_1 \Pi_1 \times_2 \cdots \times_K \Pi_K) \overset{\mathrm{d}}{=} X,$$

即， 当我们沿每个模进行置换时，分布 X 保持不变。

在此设定下，我们关注 $W, X_{\mathbb{T}} - \mathbb{E}[W, X_{\mathbb{T}}]$ 的集中不等式，其中$\mathbb{E}[W, X_{\mathbb{T}}] = N_1\cdots N_k \cdot W \cdot X$。要得到单侧尾界 $\mathbb{P}\{W, X_{\mathbb{T}} - \mathbb{E}[W, X_{\mathbb{T}}] > t\}$，关键在于对其矩母函数 $\mathbb{E}[\exp\{\lambda(W, X_{\mathbb{T}} - \mathbb{E}[W, X_{\mathbb{T}}])\}]$ 在适当取值下进行上界估计 λ。本节接下来将给出两种此类界：首先是 Hoeffding 型界，然后是更精细的 Bernstein 型界。

在继续之前，我们需要先介绍一些额外的记号，用于计算张量内的逐模平均。对于任意张量 $A \in \mathbb{T}$，我们可以定义一个张量序列 $A_k \in \mathbb{R}^{N_1\times\cdots\times N_k} =: \mathbb{T}_k$ ，其定义如下

$$\overline{A}_k := A \times_{k+1} \frac{1}{N_{k+1}} \mathbf{1}_{N_{k+1}} \times_{k+2} \cdots \times_K \frac{1}{N_K} \mathbf{1}_{N_K}.$$

换言之， A_k是通过对各模 $k+1, .., K$取平均得到的，其元素为

$$(\overline{A}_k)_{i_1\cdots i_k} = \frac{1}{N_{k+1}\cdots N_K} \sum_{i_{k+1}=1}^{N_{k+1}} \cdots \sum_{i_K=1}^{N_K} A_{i_1\cdots i_k i_{k+1}\cdots i_K}.$$

我们还定义了一个序列 $A_k \in \mathbb{T}$ ，其定义如下

$$A_k := A \times_{k+1} \mathsf{P}_{\mathbf{1}_{N_{k+1}}} \times_{k+2} \cdots \times_K \mathsf{P}_{\mathbf{1}_{N_K}},$$

也就是说，对于任意 i_1、.....、 i_K，我们有 $(A_k)_{i_1\cdots i_k i_{k+1}\cdots i_K} = (A_k)_{i_1\cdots i_k}$。对于 $k = 0$，此外还有 $\overline{A}_0 = \underline{A} \in \mathbb{R}$ 是 A 所有元素的平均值，且 $A_0 = A_0 \cdot \mathbf{1}_{N_1} \otimes \cdots \otimes \mathbf{1}_{N_K}$。另一方面，$A_K = A_K = A$。

2.1 Hoeffding 界

如 [3], 我们定义一个由辅助常数组成的序列 ϵ_n ，满足 $\epsilon_1 = 0$，并且当 $n \geq 2$ 时，

$$\epsilon_n = \frac{H_n - 1}{n - H_n}, \qquad H_n := 1 + \frac{1}{2} + \cdots + \frac{1}{n}.$$

该常数与 [46], 中 WoR 均值估计器对应的优势项相关，其中在 $t = n$ 情形下， $A_n = n/(1+\epsilon_{n+1})$。以下结果给出了关于 $Z = W, X_{\mathbb{T}} - \mathbb{E}[W, X_{\mathbb{T}}]$ 的 Hoeffding 型矩母函数界。

定理 1 (**模态可交换性下的 Hoeffding 型矩母函数界**)。对于固定的 $W \in \mathbb{T}$和模态可交换的 $X \in \mathbb{T}$，对任意 $\lambda \in \mathbb{R}$ 均有

$$\mathbb{E}\left[\exp\{\lambda Z\}\right] \leq \exp\left\{\frac{\lambda^2}{2} \cdot \prod_{l=1}^{K} N_l \cdot \sum_{k=1}^{K} \frac{1+\epsilon_{N_k}}{N_k} \cdot \sigma_{W,k}^2\right\}, \qquad \text{where } \sigma_{W,k}^2 := \left\| W_k \times_k \mathsf{P}^{\perp}_{\mathbf{1}_{N_k}} \right\|_{\mathbb{T}}^2. \tag{1a}$$

因此，对所有 $\delta \in (0,1)$,

$$\mathbb{P}\left\{ Z \geq \sqrt{2 \log \frac{1}{\delta} \cdot \prod_{l=1}^{K} N_l \cdot \sum_{k=1}^{K} \frac{1+\epsilon_{N_k}}{N_k} \sigma_{W,k}^2} \right\} \leq \delta. \tag{1b}$$

该结果的证明见附录B.1。

2.2 Bernstein 界

我们为张量 X （ $k \in [K]$）定义一列方差参数 $\sigma_{X,\,k}^2$ 和 $\widetilde{\sigma}_{X,\,k}^2$，

$$\sigma_{X,k}^2 = \frac{1}{N_k} \left\| \overline{X}_k \times_k \mathsf{P}^{\perp}_{\mathbf{1}_{N_k}} \right\|_{\mathbb{T}_k}^2 = \frac{1}{\prod_{l=k}^{K} N_l} \cdot \left\| X_k \times_k \mathsf{P}^{\perp}_{\mathbf{1}_{N_k}} \right\|_{\mathbb{T}}^2,$$

$$\widetilde{\sigma}_{X,k}^2 = \sigma_{X,k}^2 + \epsilon_{N_k} \cdot \prod_{l=0}^{k-1} N_l \cdot \left\| \overline{X}_k \times_k \mathsf{P}^{\perp}_{\mathbf{1}_{N_k}} \right\|_{\infty}^2 = \sigma_{X,k}^2 + \epsilon_{N_k} \cdot \prod_{l=0}^{k-1} N_l \cdot \left\| X_k \times_k \mathsf{P}^{\perp}_{\mathbf{1}_{N_k}} \right\|_{\infty}^2,$$

其定义类似于上文对 $\sigma_{W,\,k}^2$ 的定义。定义向量 $w_k \in \mathbb{R}^{N_k}$ 为(对于 $k \in [K]$)，其 i 个坐标为 $\sum_{i_l \in [N_l],\, l \in [k-1]} \left| (W_k \times_k \mathsf{P}^{\perp}_{\mathbf{1}_{N_k}})_{i_1 \cdots i_{k-1} i} \right|$。现在我们可以给出 Bernstein 型矩母函数界了。证明见附录 B.2。

定理 2 (**模交换性下的 Bernstein 型矩母函数界**)。对于固定的 $W \in \mathbb{T}$和模态可交换的 $X \in \mathbb{T}$，令 $\widetilde{\mathcal{F}}_X = \sigma(\widetilde{\sigma}_{X,1}, \cdots, \widetilde{\sigma}_{X,K})$，则对于 $|\lambda| \leq$ 最小值 $_{k \in [K]} \frac{3}{2\|w_k\|_{\infty} (1+\epsilon_{N_k})}$，

$$\mathbb{E}\left[\exp\{\lambda Z\} \mid \widetilde{\mathcal{F}}_X\right] \leq \exp\left\{ \frac{\lambda^2}{2} \cdot \sum_{k=1}^{K} \frac{(1+\epsilon_{N_k}) \cdot \prod_{l=k+1}^{K} N_l \cdot \sigma_{W,k}^2 \cdot \widetilde{\sigma}_{X,k}^2}{1 - \frac{2|\lambda|}{3} \|\overline{w}_k\|_{\infty} (1+\epsilon_{N_k})} \right\} \tag{2a}$$

因此，对于所有 $\delta \in (0,1)$,

$$\mathbb{P}\left\{ Z \geq \sqrt{2 \log \frac{1}{\delta} \cdot \sum_{k=1}^{K} \prod_{l=k+1}^{K} N_l \cdot (1+\epsilon_{N_k})\, \sigma_{W,k}^2 \widetilde{\sigma}_{X,k}^2} + \frac{2 \max_{k \in [K]} (1+\epsilon_{N_k}) \|\overline{w}_k\|_{\infty}}{3} \cdot \log \frac{1}{\delta} \right\} \leq \delta. \tag{2b}$$

为了使界限更易于解释，我们可以进一步放宽式（2a）和（2b）来简化它们。注意 $\sigma_{W,k}^2 = \left\|W_k \times_k \mathbf{P}_{\mathbf{1}_{N_k}}^{\perp}\right\|_{\mathbb{T}}^2 = \|W_k\|_{\mathbb{T}}^2 - \left\|W_k \times_k \mathbf{P}_{\mathbf{1}_{N_k}}\right\|_{\mathbb{T}}^2 = \|W_k\|_{\mathbb{T}}^2 - \|W_{k-1}\|_{\mathbb{T}}^2$，这表明 $\sum_{k=1}^K \sigma_{W,k}^2 = \|W\|_{\mathbb{T}}^2 - \|W_0\|_{\mathbb{T}}^2 \le \|W\|_{\mathbb{T}}^2$。因此

$$\sum_{k=1}^{K} \prod_{l=k+1}^{K} N_l \cdot (1+\epsilon_{N_k})\, \sigma_{W,k}^2 \widetilde{\sigma}_{X,k}^2 \le \|W\|_{\mathbb{T}}^2 \cdot \max_{k\in[K]} \prod_{l=k+1}^{K} N_l \cdot (1+\epsilon_{N_k})\, \widetilde{\sigma}_{X,k}^2.$$

这立即导出以下推论：

推论 2.1。 在定理 2的设定下，也有

$$\mathbb{P}\left\{ Z \ge \sqrt{2\log\frac{1}{\delta} \cdot \|W\|_{\mathbb{T}}^2 \cdot \max_{k\in[K]} \prod_{l=k+1}^{K} N_l \cdot (1+\epsilon_{N_k})\, \widetilde{\sigma}_{X,k}^2} + \frac{2\max_{k\in[K]}(1+\epsilon_{N_k})\, \|\overline{w}_k\|_{\infty}}{3} \cdot \log\frac{1}{\delta} \right\} \le \delta. \quad (3)$$

2.2.1 比较可交换性与独立性：维度的作用

为了帮助理解这些结果，我们接下来考察维度 N_1、……、N_K 在这些界中的作用。作为基线，假设 X 的各个条目是独立同分布的无偏符号——也就是说，$X_{i_1\cdots i_K} \sim$ 均匀分布$(\{\pm 1\})$。此时，$Z = \langle X, W\rangle$ 的均值为零，方差为 $\|W\|_{\mathbb{T}}^2$，因此有$|Z| = O_P(\|W\|_{\mathbb{T}})$。另一方面，在推论 2.1中，如果只考虑首项，则有 $|Z| = O_P(\|W\|_{\mathbb{T}} \cdot \sqrt{\prod_{k=2}^K N_K})$。

然而，接下来我们将看到，由于当 $K > 1$ 时，可交换性明显弱于独立性，这是不可避免的。具体而言，考虑以下模可交换张量：令 $Y = (Y_1, \ldots, Y_{N_1})$，其中 $Y_1, \ldots, Y_{N_1} \sim$ 均匀分布$(\{\pm 1\})$且相互独立同分布，并令$X = Y \otimes \mathbf{1}_{N_2} \otimes \cdots \otimes \mathbf{1}_{N_K}$。因此，$X$ 由 N_1 个独立样本构造而成（而不是 $\prod_{k=1}^K N_k$）。特别地，$Z = X, W_{\mathbb{T}}$ 仍然具有零均值，但现在其方差为 $\|W_1\|_{\mathbb{T}}^2 \cdot \prod_{k=2}^K N_k$。这解释了为什么可交换情形与独立同分布情形之间的集中速率相差 $\sum_{k=2}^K N_k$ 倍。

3 矩阵值数据的集中不等式

在本节中，我们研究矩阵值条目序列数据的集中不等式。回顾一下，在这种设定中，我们使用一列权重矩阵 $W_1, \cdots, W_N \in \mathbb{R}^{p\times q}$和随机矩阵 $X_1, \cdots, X_N \in \mathbb{R}^{q\times r}$，并假设它们是可交换的——即，对于任意 $\pi \in \mathscr{S}_N$，都有 $(X_1, \cdots, X_N) \overset{\mathrm{d}}{=} (X_{\pi(1)}, \cdots, X_{\pi(N)})$。使用张量记号，我们将张量 $(W_1, \cdots, W_N)$ 和 $(X_1, \cdots, X_N)$ 分别记为 $W \in \mathbb{R}^{N\times p\times q}$和 $X \in \mathbb{R}^{N\times q\times r}$。接下来，我们考虑双线性映射 $\cdot, \cdot : \mathbb{R}^{N\times p\times q} \times \mathbb{R}^{N\times q\times r} \to \mathbb{R}^{p\times r}$，

$$Z = \langle\!\langle W, X \rangle\!\rangle := \sum_{k=1}^{N} W_k X_k.$$

对于任意 $A \in \mathbb{R}^{N\times N}$，该双线性映射满足以下性质（类似于 $\cdot, \cdot_{\mathbb{T}}$）：

$$\langle\!\langle W \times_1 A, X \rangle\!\rangle = \sum_{k=1}^{N} \left(\sum_{l=1}^{N} A_{kl} W_l\right) X_k = \sum_{k=1}^{N} W_k \left(\sum_{l=1}^{N} A_{lk} X_l\right) = \langle\!\langle W, X \times_1 A^{\top} \rangle\!\rangle.$$

令 $\overline{W} = \frac{1}{N}\sum_{k=1}^N W_k$ 和 $\overline{X} = \frac{1}{N}\sum_{k=1}^N X_k$。由于 Z 现在是矩阵值，在此情形下，我们可以通过考察 $\left\| W, X - N \cdot \overline{W}\, \overline{X} \right\|$ 在算子范数意义下的尾界，来研究 Z 的集中。由于

这是标量情形下矩母函数的自然推广，我们将看到，通过研究由迹指数函数定义的矩阵矩母函数，可以推导出切尔诺夫型集中界，包括 Hoeffding 界和 Bernstein 界（见下文第3.1节）。

3.1 Hoeffding 界（一般情形）

首先，我们要强调矩阵值情形与标量值情形之间的一个关键区别：对称性不再成立。即使 $p = r$，我们也不能普遍地认为 W， $X = X$， W 成立，除非 $p = q = r = 1$。这是由于矩阵乘法的非交换性。正如我们将在下文看到的，这种不对称性给推导集中界带来了一些挑战。尽管如此，我们仍然可以在不对 W 和 X 之间的可交换性作任何额外假设的情况下，得到如下通用的霍夫丁型尾界。

首先，我们回顾处理矩阵值数据时的一些记号。对于方阵 $A \in \mathbb{R}^{n\times n}$，回忆矩阵指数 $\exp A := \sum_{k=0}^{\infty} \frac{1}{k!} A^k$。我们将如下定义归一化指数算子 $\overline{\exp} : \mathbb{R}^{n\times m} \to \mathbb{R}$

$$\overline{\exp}(A) := \frac{1}{n+m} \operatorname{Tr} \left\{ \exp \begin{bmatrix} 0 & A \\ A^\top & 0 \end{bmatrix} \right\}.$$

$\mathbb{S}^{n+m}$ 右侧的扩大的对称化矩阵是 A的扩张矩阵。所谓的膨胀技巧 [35] 是算子理论中的一个强大思想。这些工具导出了以下界。

定理 3 (**可交换矩阵值数据的通用 Hoeffding 型界**)。 对于任意正整数 N, p, q, r， 设 $W_1, \cdots, W_N \in \mathbb{R}^{p\times q}$ 为一个固定的权重矩阵序列，且 $X_1, \cdots, X_N \in \mathbb{R}^{q\times r}$ 为一个随机数据的可交换序列，其中 $\|X_k\| \le 1$ 对所有 $k \in [N]$均成立。那么，对于任意 $\lambda \in \mathbb{R}$ ，有

$$\mathbb{E}\left[\overline{\exp}\left\{\lambda\left(\langle\!\langle W, X\rangle\!\rangle - N \cdot \overline{W}\ \overline{X}\right)\right\}\right] \le \exp\left\{8\lambda^2 (1+\epsilon_N)^2 \cdot \sum_{k=1}^{N} \|W_k\|^2\right\}, \tag{4a}$$

因此，对于所有 $\delta \in (0,1)$，

$$\mathbb{P}\left\{\left\|\langle\!\langle W, X\rangle\!\rangle - N \cdot \overline{W}\ \overline{X}\right\| \ge 4(1+\epsilon_N)\sqrt{2\log\frac{p+r}{\delta} \cdot \sum_{k=1}^{N} \|W_k\|^2}\right\} \le \delta. \tag{4b}$$

这一结果普遍成立，不需要任何类型的交换性条件。然而，正如我们接下来将看到的，在附加条件下可以推导出强得多的结果——尤其是通过改进项 $\sum_{k=1}^{N} \|W_k\|^2$。

3.2 Hoeffding 界（在交换性条件下）

尽管一般情况下交换性并不成立，但在可以假设权重矩阵 W_k 和随机矩阵 X_k 满足西尔维斯特方程等额外约束的情形下，我们可以改进上述推导出的尾界。我们将假设表述如下。

假设 1 （交换性条件）。 对于给定的权重 $W \in \mathbb{R}^{N\times p\times q}$，如果存在一个

正整数 $\widetilde{q}$、与之关联的权重 $V \in \mathbb{R}^{N\times\widetilde{q}\times r}$ 以及一个单射 $\iota : \mathbb{R}^{q\times r} \to \mathbb{R}^{p\times\widetilde{q}}$，使得以下条件在 X_1 的支撑集上成立：

$$W_k X_1 = \iota(X_1)V_k =: \widetilde{X}_1 V_k, \qquad \text{for all } k \in [N].$$

记为 $\widetilde{X} = (\widetilde{X}_1, \cdots, \widetilde{X}_N) \in \mathbb{R}^{N\times\widetilde{q}\times r}$，我们称$(\widetilde{X}, V)$为$(W, X)$的可交换对。

注 1. 显然，根据 X 经 ι 得到的推前分布，X 也是一个可交换矩阵序列。因此，研究量$\langle\!\langle W, X\rangle\!\rangle$ 的集中等价于研究$\langle\!\langle X, V\rangle\!\rangle$，从而消除了前文所述数据左乘的不对称性。举例来说，对于可交换标量序列$(x_1, \cdots, x_N)$，在 $q = r$ 且 $X_k = x_k I_q$ 的特殊情形下，对于任意权重 W，我们可以设定 $\widetilde{q} = p$，且 $X_k = x_k I_p, V_k = W_k$。

事实上，交换性条件的适用范围远不止上述注释中的标量随机序列情形——我们将在下文第 4 节和 5节中看到几个具体示例。

有了这一假设，我们现在可以给出适用于矩阵情形的、更强的 Hoeffding 型界。

定理 *4*（可交换矩阵值数据的 *Hoeffding* 型矩母函数界）。对于任意正整数 N, p, q, r，设 $W_1, \cdots, W_N \in \mathbb{R}^{p\times q}$ 是固定的权重矩阵序列，$X_1, \cdots, X_N \in \mathbb{R}^{q\times r}$ 是可交换的随机数据序列，其中对所有 $k \in [N]$，$\|X_k\| \le 1$。若对于可交换对 $(\widetilde{X}, V)$，假设 *1* 成立，其中对所有 $k \in [N]$，$\|\widetilde{X}_k\| \le 1$，

$$\mathbb{E}\left[\overline{\exp}\left\{\lambda\left(\langle\!\langle W, X\rangle\!\rangle - N\cdot\overline{W}\;\overline{X}\right)\right\}\right] \le \exp\left\{8\lambda^2(1+\epsilon_N)\cdot\max\left\{\left\|\sum_{k=1}^N W_k W_k^\top\right\|, \left\|\sum_{k=1}^N V_k^\top V_k\right\|\right\}\right\}. \tag{5a}$$

因此，对于所有 $\delta \in (0,1)$，

$$\mathbb{P}\left\{\left\|\langle\!\langle W, X\rangle\!\rangle - N\cdot\overline{W}\;\overline{X}\right\| \ge 4\sqrt{2\log\frac{p+r}{\delta}\cdot(1+\epsilon_N)\max\left\{\left\|\sum_{k=1}^N W_k W_k^\top\right\|, \left\|\sum_{k=1}^N V_k^\top V_k\right\|\right\}}\right\} \le \delta. \tag{5b}$$

将这些界与前一个定理的结果进行比较，可以看到，定理 4中得到的尾界依赖于 最大值$\{\|\sum_{k=1}^N W_k W_k^\top\|, \|\sum_{k=1}^N V_k^\top V_k\|\}$，而不是像定理 $\sum_{k=1}^N \|W_k\|^2$3中那样依赖于该项。这可能是一个紧得多的界：事实上，我们总有

$$\max\left\{\left\|\sum_{k=1}^N W_k W_k^\top\right\|, \left\|\sum_{k=1}^N V_k^\top V_k\right\|\right\} \le \max\left\{\sum_{k=1}^N \|W_k\|^2, \sum_{k=1}^N \|V_k\|^2\right\}$$

而该不等式中的差距可能高达 $O(N)$ 倍（例如，如果每个 W_k 和 V_k 都是秩一矩阵）。然而，定理4 所假设的交换性条件可能具有较强限制，而定理3始终适用，因而具有更大的灵活性。

3.3 伯恩斯坦界（交换性条件下）

接下来，我们将推导交换性条件成立时可交换矩阵值数据的伯恩斯坦型界。令

$$\Sigma = \frac{1}{N}\sum_{k=1}^N (X_k - \overline{X})(X_k - \overline{X})^\top, \qquad \widetilde{\Sigma} = \frac{1}{N}\sum_{k=1}^N (\widetilde{X}_k - \overline{\widetilde{X}})^\top(\widetilde{X}_k - \overline{\widetilde{X}}),$$

并且我们可以定义方差参数 $\sigma_X^2 = \|\Sigma\|$ 和 $\sigma_{\widetilde{X}}^2 = \|\widetilde{\Sigma}\|$，以及

$$\widetilde{\sigma}_X^2 = \sigma_X^2 + \epsilon_N \max_{k\in[N]} \left\|X_k - \overline{X}\right\|^2, \qquad \widetilde{\sigma}_{\widetilde{X}}^2 = \sigma_{\widetilde{X}}^2 + \epsilon_N \max_{k\in[N]} \left\|\widetilde{X}_k - \overline{\widetilde{X}}\right\|^2.$$

定理 5 (**可交换矩阵值数据的 Bernstein 型矩母函数界**)。 设 假设 *1* 对 $N \geq 2$成立。若 $\|X_k\| \leq 1$ 且 $\left\|\widetilde{X}_k\right\| \leq 1$ 对所有 $k \in [N]$均成立，则对于 $w = (\|W_1\|, \cdots, \|W_N\|)^\top$, $v = (\|V_1\|, \cdots, \|V_N\|)^\top$ 以及 $|\lambda| \leq \sqrt[3]{\min\{\|w\|_\infty, \|v\|_\infty\}(1+\epsilon_N)}$，有

$$\begin{aligned}
&\mathbb{E}\left[\overline{\exp}\left\{\lambda\left(\langle\!\langle W, X\rangle\!\rangle - N \cdot \overline{W}\ \overline{X}\right)\right\} \mid \widetilde{\sigma}_X^2, \widetilde{\sigma}_{\widetilde{X}}^2\right] \\
&\quad \leq \exp\left\{\frac{\lambda^2(1+\epsilon_N)^2}{2\left(1 - \frac{2|\lambda|(1+\epsilon_N)}{3}\min\{\|w\|_\infty, \|v\|_\infty\}\right)} \cdot \max\left\{\widetilde{\sigma}_X^2 \left\|\sum_{k=1}^N W_k W_k^\top\right\|, \widetilde{\sigma}_{\widetilde{X}}^2 \left\|\sum_{k=1}^N V_k^\top V_k\right\|\right\}\right\}. \qquad (6a)
\end{aligned}$$

因此，对于所有 $\delta \in (0,1)$，

$$\begin{aligned}
&\mathbb{P}\left\{\left\|\langle\!\langle W, X\rangle\!\rangle - N \cdot \overline{W}\ \overline{X}\right\| \geq (1+\epsilon_N)\sqrt{2\log\frac{p+r}{\delta} \cdot \max\left\{\widetilde{\sigma}_X^2 \left\|\sum_{k=1}^N W_k W_k^\top\right\|, \widetilde{\sigma}_{\widetilde{X}}^2 \left\|\sum_{k=1}^N V_k^\top V_k\right\|\right\}}\right. \\
&\qquad \left. + \frac{2\min\{\|w\|_\infty, \|v\|_\infty\}}{3}(1+\epsilon_N)\log\frac{p+r}{\delta}\right\} \leq \delta. \qquad (6b)
\end{aligned}$$

其证明推迟至附录C.2。

4 理论推论

本节考察我们一般性结果的若干推论，并为一系列不同的特殊情形推导界。

4.1 可交换随机变量的加权和

我们首先将第 2 节和 3 节的结果应用于随机项由可交换随机变量序列 ξ_1、$\cdots$、$\xi_N \in [-1,1]$组成的情形。具体而言，我们考虑以下一般形式的可交换随机变量加权和，其中权重可以是标量值或矩阵值：

$$\begin{aligned}
&\textbf{（标量权重）} && \langle w, \xi\rangle = \sum_{k=1}^N w_k \xi_k, \\
&\textbf{（矩阵值权重）} && \langle\!\langle W, \xi I_r\rangle\!\rangle = \sum_{k=1}^N W_k \xi_k.
\end{aligned}$$

为了将这些和与前文的一般记号联系起来，在第一种情形下，我们只需令 $X_k = \xi_k$；而在第二种情形下，我们处理的是矩阵值随机变量 $X_k = \xi_k I_r$。

令 $\bar{\xi} := \frac{1}{N}\sum_{k=1}^N \xi_k$、$\sigma_\xi^2 := \frac{1}{N}\sum_{k=1}^N (\xi - \bar{\xi})^2$和 $\widetilde{\sigma}_\xi^2 := \sigma_\xi^2 + \epsilon_N \left\|\xi - \bar{\xi}\right\|_\infty^2 \leq \sigma_\xi^2 + 4\epsilon_N$。对于标量权重的情形，我们应用定理 1 和 2，令 $K = 1$。请注意，这些界恰好恢复了 Barber [3,定理 3.1 和 3.3]中的主要结果，且常数相同。

推论 4.1 （**可交换随机变量的标量加权和**）。对于 $w = (w_1, \cdots, w_N)^\top \in \mathbb{R}^N$以及所有 $\delta \in (0, 1)$，假设$w = \frac{1}{N}\sum_{k=1}^N \overline{w}_k = 0$。

1. （***Hoeffding*** 界）对于任意 $\lambda \in \mathbb{R}$，有

$$\mathbb{E}\left[\exp\left\{\lambda\langle w,\xi\rangle\right\}\right] \le \exp\left\{\frac{\lambda^2(1+\epsilon_N)}{2}\cdot\|w\|^2\right\},$$

and consequently $\mathbb{P}\left\{\langle w,\xi\rangle \ge \sqrt{2(1+\epsilon_N)\log 1/\delta}\cdot\|w\|\right\} \le \delta.$

2. ***(Bernstein*** 界***)*** 此外，对于 $|\lambda| \le \frac{3}{2\|w\|_\infty(1+\epsilon_N)}$

$$\mathbb{E}\left[\exp\left\{\lambda\langle w,\xi\rangle\right\} \mid \widetilde{\sigma}_\xi^2\right] \le \exp\left\{\frac{\lambda^2(1+\epsilon_N)}{2\left(1-\frac{2|\lambda|}{3}\|w\|_\infty(1+\epsilon_N)\right)}\cdot\widetilde{\sigma}_\xi^2\|w\|^2\right\},$$

以及 *c* 因此 $\mathbb{P}\{w,\xi \ge \sqrt{2(1+\epsilon_N)\log 1/\delta}\cdot\widetilde{\sigma}_\xi\|w\| + \frac{2}{3}\|w\|_\infty(1+\epsilon_N)\log 1/\delta\} \le \delta$.

证明 我们在定理 $\mathbb{T} = \mathbb{R}^N, W = w - \overline{w}\mathbf{1}_N$和 $X = \xi$中取 1 和 2。 □

接下来，我们将定理 4 和 5 应用于一列固定矩阵 $W_1,\cdots,W_N$。由此得到的结果恢复了 Tropp [43, 推论 7。2] 中的 Hoeffding 界，以及 Tropp 等人 [44,定理 6.1。1] 中的 Bernstein 界，二者相差一个乘法因素 $(1+\epsilon_N)$；这是因为 Tropp's 的结果针对的是独立随机变量。（下一节将讨论如何将传输可交换结果还原为独立同分布结果，从而消去 $(1+\epsilon_N)$ 这一因素。）

然而，与标量权重的情形相比，只要满足 $\xi_k \overset{d}{=} -\xi_k$，霍夫丁界中的数值常数确实可以改进。关于进一步讨论，请读者参阅 [43, 备注 7。4] 此外，在 Mackey 等人的研究中，作者采用可交换对方法，在独立同分布数据下无需分布对称性，便建立了具有最优常数的 Hoeffding 型不等式。Bernstein 界中的数值常数确实是最优的。

推论 4.2 (**可交换随机变量的矩阵加权和**)。对于 $W_1,\ \cdots,\ W_N \in \mathbb{R}^{p\times r}$ 以及所有$\delta \in (0,\ 1)$，假设 $W = \frac{1}{N}\sum_{k=1}^N W_k = \overline{0}$并令 $\sigma_W^2 := \max\{\|\sum_{k=1}^N W_k W_k^\top\|,\ \|\sum_{k=1}^N W_k^\top W_k\|\}$。

1. （***Hoeffding*** 界）对于任意 $\lambda \in \mathbb{R}$，均有

$$\mathbb{E}\left[\overline{\exp}\left\{\lambda\sum_{k=1}^N W_k\xi_k\right\}\right] \le \exp\left\{8\lambda^2(1+\epsilon_N)\cdot\sigma_W^2\right\},$$

and consequently $\mathbb{P}\left\{\|\sum_{k=1}^N W_k\xi_k\| \ge 4\sqrt{2(1+\epsilon_N)\log\frac{p+r}{\delta}}\cdot\sigma_W\right\} \le \delta.$

2. （***Bernstein*** 界）此外，假设 $\|W_k\| \le L$ 对所有 $k \in [N]$，有 $|\lambda| \le \frac{3}{2L(1+\epsilon_N)}$

$$\mathbb{E}\left[\overline{\exp}\left\{\lambda\sum_{k=1}^N W_k\xi_k\right\} \mid \widetilde{\sigma}_\xi^2\right] \le \exp\left\{\frac{\lambda^2(1+\epsilon_N)}{2\left(1-\frac{2|\lambda|L}{3}(1+\epsilon_N)\right)}\cdot\widetilde{\sigma}_\xi^2\sigma_W^2\right\},$$

and consequently $\mathbb{P}\left\{\|\sum_{k=1}^N W_k\xi_k\| \ge \sqrt{2(1+\epsilon_N)\log\frac{p+r}{\delta}}\cdot\widetilde{\sigma}_\xi\sigma_W + \frac{2L}{3}(1+\epsilon_N)\log\frac{p+r}{\delta}\right\} \le \delta.$

证明 我们首先通过取 $X_k = \xi_k I_r,\quad X_k = \xi_k I_p$ 和 $V_k = W_k$，验证假设 1 。随后由定理 4 和 5可得矩母函数不等式。尾界则是引理 A.5的推论。 □

4.2 与独立同分布情形的联系

接下来，我们将探讨在可交换性假设下得到的结果与更常见的独立同分布数据情形之间的联系。

为引出这个问题，我们先回顾 [3]，中研究的标量情形，其目标是证明可交换随机变量X_1、...、$X_N \in \mathbb{R}$ 的加权和 $w_1X_1 + \cdots + w_NX_N$ 集中。文献 [3] 表明，如果 X_1、...、X_N 可以嵌入到更长的可交换$\widetilde{\text{}}$序列 X_1、...、$X_{\widetilde{N}}$ 中（对 $N \geq N$），那么推论 4.1 的集中界中可以用 $\epsilon_{\widetilde{N}}\widetilde{\text{}}$ 替代 ϵ_N。极端情况下，如果这些 X_i 是独立同分布的，我们可以令 N 任意大；由于$\sim\epsilon_{\widetilde{N}} \to 0$ 当 $N \to \infty$ 时，这意味着推论 4.1 的结果可以恢复独立随机变量之和的通常 Hoeffding 界和 Bernstein 界。本节将看到，本文所考虑的张量和矩阵情形也成立同类结论。

4.2.1 与模可交换张量的独立同分布情形进行比较

如第 2.2.1 节所述，对于模态可交换张量的情形，我们可以将其与“沿一个模独立同分布”的情形进行比较——也就是说，我们对独立同分布的 $(K-1)$ 模张量进行抽样。具体而言，假设 $X \in [-1,1]^{N_1 \times \cdots \times N_K}$ 沿第 K 模独立同分布：也就是说，$(K-1)$ 模张量

$$\mathscr{X}_l := X \times_K e_l = (X_{i_1 \ldots i_{K-1} l})_{i_1 \in [N_1], \ldots, i_{K-1} \in [N_{K-1}]} \in \mathbb{T}_{K-1}$$

对于每个 $l = 1, \ldots, N_K$，均为独立同分布，服从 $\mathbb{T}_{K-1}$ 上的概率分布 P，其取实数值的均值和方差参数为

$$\mu_P = \mathbb{E}_{\mathscr{X} \sim P}[\overline{\mathscr{X}}], \qquad \sigma_P^2 = \mathbb{E}_{\mathscr{X} \sim P}\left[\|\mathscr{X} - \mathbb{E}_{\mathscr{X} \sim P}[\mathscr{X}]\|_{\mathbb{T}_{K-1}}^2\right].$$

类似地，我们可以定义 N_K 权重张量 $\mathscr{W}_1$，$\cdots$，$\mathscr{W}_{N_K} \in \mathbb{T}_{K-1}$。在这种情况下，相关内积可以写成 N_K 个独立项之和：

$$Z = \langle W, X \rangle_{\mathbb{T}_K} = \sum_{l=1}^{N_K} \langle \mathscr{W}_l, \mathscr{X}_l \rangle_{\mathbb{T}_{K-1}}.$$

令 $w \in \mathbb{R}^{N_K}$ 为一个向量，其第 i个坐标为 $\sum_{i_l \in [N_l],\ l \in [K-1]} |W_{i_1 \cdots i_{K-1} i}| = \|\mathscr{W}_i\|_1$。为了应用关于独立随机变量之和的经典结果，我们注意到第 l个加数 $\mathscr{W}_l$，$\mathscr{X}_l\mathbb{T}_{K-1}$ 满足

$$|\langle \mathscr{W}_l, \mathscr{X}_l \rangle_{\mathbb{T}_{K-1}}| \leq \|\mathscr{W}_l\|_1 \cdot \|\mathscr{X}_l\|_\infty \leq w_l, \quad \mathsf{Var}(\langle \mathscr{W}_l, \mathscr{X}_l \rangle_{\mathbb{T}_{K-1}}) = \mathbb{E}\left[(\langle \mathscr{W}_l, \mathscr{X}_l - \mu_P \rangle_{\mathbb{T}_{K-1}})^2\right] \leq \|\mathscr{W}_l\|_{\mathbb{T}_{K-1}}^2 \cdot \sigma_P^2.$$

应用关于独立随机变量之和的经典 Hoeffding 不等式，可得对所有 $\lambda \in \mathbb{R}$,

$$\mathbb{E}[\exp\{\lambda(Z - \mathbb{E}[Z])\}] \leq \exp\left\{\frac{\lambda^2}{2} \cdot \sum_{l=1}^{N_K} \|\mathscr{W}_l\|_1^2\right\},$$

而经典的 Bernstein 界对所有 $|\lambda| \leq \frac{3}{2\|w\|_\infty}$ 给出如下矩母函数界，

$$\mathbb{E}[\exp\{\lambda(Z - \mathbb{E}[Z])\}] \leq \exp\left\{\frac{\lambda^2}{2\left(1 - \frac{2|\lambda|}{3}\|w\|_\infty\right)} \cdot \sum_{l=1}^{N_K} \|\mathscr{W}_l\|_{\mathbb{T}_{K-1}}^2 \cdot \sigma_P^2\right\}.$$

为更好地与我们的界建立联系，我们应用柯西-施瓦茨不等式 $\|\mathscr{W}_l\|_1^2 \le \prod_{l=1}^{K-1} N_l \cdot \|\mathscr{W}_l\|_{\mathbb{T}_{K-1}}^2$（当 $K=2$ 时取等号）以及恒等式 $\sum_{l=1}^{N_K} \|\mathscr{W}_l\|_{\mathbb{T}_{K-1}}^2 = \|W\|_{\mathbb{T}}^2$，得到如下结果。

$$\mathbb{E}[\exp\{\lambda(Z-\mathbb{E}[Z])\}] \le \exp\left\{\frac{\lambda^2}{2}\cdot\prod_{l=1}^{K-1} N_l \cdot \|W\|_{\mathbb{T}}^2\right\} \qquad \text{for all } \lambda\in\mathbb{R}, \tag{7a}$$

$$\mathbb{E}[\exp\{\lambda(Z-\mathbb{E}[Z])\}] \le \exp\left\{\frac{\lambda^2}{2\left(1-\frac{2|\lambda|}{3}\|w\|_\infty\right)}\cdot\|W\|_{\mathbb{T}}^2\,\sigma_P^2\right\} \qquad \text{for all } |\lambda| \le \frac{3}{2\|w\|_\infty}. \tag{7b}$$

4.2.2 可交换性界与独立同分布界之间的插值

接下来，我们将看到，在可交换性假设下，我们的结果在极限情形下如何与这些界相吻合。假设 $\widetilde{X}\in\mathbb{R}^{N_1\times\cdots\times N_{K-1}\times\widetilde{N}_K} =: \widetilde{\mathbb{T}}$ 对于某个 $\widetilde{N}_K \ge N_K$ 是模可交换的，且 $X\in\mathbb{R}^{N_1\times\cdots\times N_K}$ 是 $\widetilde{X}$ 的相应子张量。我们还需要一个额外假设：

定义 4.1 (沿模-K的平衡权重)。 如果 W 沿模-K 是平衡的，即 $W_{K-1} = W\mathbf{1}_{N_1}\otimes\cdots\otimes\mathbf{1}_{N_{K-1}}$，则称其满足该条件。

那么，假设 W 沿模-K是平衡的，定理 1 中的矩母函数界可通过以下方式

$$\sigma_{W,k} = \left\|W_k \times_k \mathsf{P}^{\perp}_{\mathbf{1}_{N_k}}\right\|_{\mathbb{T}}^2 = \left\|\overline{W}_{K-1}\times_k \mathsf{P}^{\perp}_{\mathbf{1}_{N_k}}\right\|_{\mathbb{T}_{K-1}}^2 = 0, \qquad \text{for all } k\in[K-1].$$

因此（从 $\mathbb{T}$ 到 $\widetilde{\mathbb{T}}$ 的化简细节见附录 D.1）

$$\mathbb{E}[\exp\{\lambda(Z-\mathbb{E}[Z])\}] \le \exp\left\{\frac{\lambda^2}{2}\cdot\prod_{l=1}^{K-1} N_l\cdot(1+\epsilon_{\widetilde{N}_K})\,\|W\|_{\mathbb{T}}^2\right\} \qquad \text{for all } \lambda\in\mathbb{R}.$$

特别地，取 $N_K\to\infty$（即，若 X 沿 K-模独立同分布），我们便看到，这恰好复现了经典计算（7a），且常数正确。类似地，对于 Bernstein 界，同样假设 W 沿模-K平衡，则定理 2 给出（详见附录 D.1）

$$\mathbb{E}[\exp\{\lambda(Z-\mathbb{E}[Z])\}] \le \exp\left\{\frac{\lambda^2\cdot(1+\epsilon_{\widetilde{N}_K})}{2\left(1-\frac{2|\lambda|}{3}\|w\|_\infty(1+\epsilon_{\widetilde{N}_K})\right.}\cdot\|W\|_{\mathbb{T}}^2\,\widetilde{\sigma}^2_{\widetilde{X},K}\right\} \qquad \text{for all } |\lambda|\le\frac{3}{2\|w\|_\infty},$$

若取$\widetilde{N}_K\to\infty$ 且样本方差 $\widetilde{\sigma}^2_{\widetilde{X},K}\to\sigma_P^2$ 几乎必然收敛，则这同样恰好复现了经典计算（7b），且常数正确。

换言之，我们可以将我们的结果理解为在可交换情形与独立同分布情形之间提供了一种自然的插值：若沿模 N_K -K 的维度 N_K仅等于（有限可交换情形），则我们得到较弱的集中界；而若 $N_K\to\infty$（对应于数据沿模-K“独立同分布”的情形），则可恢复独立同分布数据的标准界。值得注意的是，Ramdas 和 Manole 证明，在无限可交换性的假设下，可以针对任意给定的阈值 $N_K\ge M$，对所有 M得到一致界，正如 [37]中所示。这一观察表明，我们有望强化结果，从而在我们的框架内确保一致控制。

注 2。事实上，如果 W沿模-K是平衡的，令 $\mu_P := \mathbb{E}_{\mathscr{X}\sim P}[\mathscr{X}] \in [-1,1]$，$Z - \mathbb{E}Z$可得以下简化形式

$$Z - \mathbb{E}Z = \sum_{l=1}^{N_K} \langle \mathscr{W}_l, \mathscr{X}_l - \mathbb{E}_{\mathscr{X}\sim P}[\mathscr{X}] \rangle_{\mathbb{T}_{K-1}} = \langle W, X - \mu_P \mathbf{1}_{\mathbb{T}} \rangle_{\mathbb{T}}.$$

具体细节留待附录 *D.1*。

4.2.3 与独立同分布矩阵集中不等式的联系

类似地，我们可以通过将定理 4 和定理 5 嵌入由随机矩阵 $X_1, X_2, \cdots, X_{\widetilde{N}}$ 组成的更长可交换序列中来推广它们，无论是否满足交换性条件；并令 $N \to \infty$，从而逼近~独立同分布结果。由于一般情形的启示有限，且符号繁复，为简便起见，我们仅简要说明如何将上一节中的推论 4.2 推广到一般情形，其中我们有$W_1, \cdots, W_N \in \mathbb{R}^{p\times r}$和 $X_k = \xi_k I_r$。首先回顾独立同分布情形，其中 $\xi_1, \cdots, \xi_N \overset{\text{iid}}{\sim} P$于 $[-1,1]$。对于 $Z = W, X = \sum_{k=1}^N \xi_k W_k$，Tropp 等人已确立的独立同分布结果 [43, 44] 表明，对于参数

$$\sigma_W^2 = \max\left\{ \left\| \sum_{k=1}^N W_k W_k^\top \right\|, \left\| \sum_{k=1}^N W_k^\top W_k \right\| \right\}, \qquad L = \max_{k\in[N]} \|W_k\|,$$

我们有

$$\mathbb{E}[\overline{\exp}\{\lambda(Z - \mathbb{E}[Z])\}] \le \exp\left\{8\lambda^2 \cdot \sigma_W^2\right\} \qquad \text{for all } \lambda \in \mathbb{R}, \tag{8a}$$

$$\mathbb{E}[\overline{\exp}\{\lambda(Z - \mathbb{E}[Z])\}] \le \exp\left\{ \frac{\lambda^2}{2\left(1 - \frac{2|\lambda|L}{3}\right)} \cdot \sigma_W^2 \mathsf{Var}(\xi) \right\} \qquad \text{for all } |\lambda| \le \frac{3}{2L}. \tag{8b}$$

现在我们看到，仅有可交换性还不够；但如果我们能将 $\xi_1, \cdots, \xi_N$嵌入到一个更长的序列中，其~长度为 N，则推论 4.2 中的结果便立即推出

$$\mathbb{E}[\overline{\exp}\{\lambda(Z - \mathbb{E}[Z])\}] \le \exp\left\{8\lambda^2(1+\epsilon_{\widetilde{N}}) \cdot \sigma_W^2\right\} \qquad \text{对所有 } \lambda \in \mathbb{R},$$

$$\mathbb{E}[\overline{\exp}\{\lambda(Z - \mathbb{E}[Z])\}] \le \exp\left\{ \frac{\lambda^2(1+\epsilon_{\widetilde{N}})}{2\left(1 - \frac{2|\lambda|L}{3}(1+\epsilon_{\widetilde{N}})\right)} \cdot \sigma_W^2 \widetilde{\sigma}_\xi^2 \right\} \qquad \text{for all } |\lambda| \le \frac{3}{2L}.$$

取 $\widetilde{N} \to \infty$，我们有 $\epsilon_{\widetilde{N}} \to 0$，且 $\widetilde{\sigma}_\xi^2 \to \mathsf{Var}(\xi)$ 几乎必然成立。因此，我们重新得到式（8a）和（8b）中的标准独立同分布矩阵集中不等式，常数也同样正确。

4.3 矩阵数组的组合和

作为对我们工具的最后一次理论展示，我们重新考察矩阵的组合和 [30,19]。Mackey 等人 考虑了一个厄米矩阵的确定性数组 $\{A_{i,\ j}\}_{i,\ j\in[N]}$，并使用著名的 Chatterjee 的可交换对方法 [10] 研究随机矩阵 $\sum_{k=1}^N A_{k,\ \pi(k)}$的集中（该方法最早出现在Chatterjee的博士论文 [9]中）。我们将结果推广到矩形情形，并假设 $A_{i,\ j} \in \mathbb{R}^{m\times n}$。在前几节中，我们将张量积 $\otimes$推广为矩阵的克罗内克积。对于 $A = (a_{ij}) \in \mathbb{R}^{m\times n}$，$B = (b_{ij}) \in \mathbb{R}^{m'\times n'}$，令 $A \otimes B \in \mathbb{R}^{mm'\times nn'}$为

$$A \otimes B := \begin{bmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{bmatrix}.$$

~我们有一个关键观察：可以找到满足假设 $(W,\ X)$ 和 $(X,\ V)$ 1 的可交换对 1，使得 $\sum_{k=1}^{N} A_{k,\ \pi(k)} = \sum_{k=1}^{N} W_k X_k = \sum_{k=1}^{N} \widetilde{X}_k V_k$。具体而言，

$$W_k = \begin{bmatrix} A_{k,1} & \cdots & A_{k,N} \end{bmatrix} \in \mathbb{R}^{m\times nN}, \qquad X_k = e_{\pi(k)} \otimes I_n \in \mathbb{R}^{nN\times n};$$

$$\widetilde{X}_k = e_{\pi(k)}^{\top} \otimes I_m \in \mathbb{R}^{m\times mN}, \qquad V_k = \begin{bmatrix} A_{k,1} \\ \vdots \\ A_{k,N} \end{bmatrix} \in \mathbb{R}^{mN\times n}.$$

$X_1,\ \cdots,\ X_N$ 因此自然构成一个可交换矩阵序列。现在我们给出矩阵数组组合和的 Bernstein 型集中不等式。

推论4.3 (组合矩阵伯恩斯坦不等式)。 假设矩阵数组 $\{A_{i,\ j}\}_{i,\ j\in[N]}$ 满足 $\sum_{i,\ j\in[N]} A_{i,\ j} = 0$ 且 $\|A_{i,\ j}\| \le R$。则有

$$\sigma^2 = \max\left\{ \left\| \sum_{i,j\in[N]} A_{i,j} A_{i,j}^{\top} \right\|, \left\| \sum_{i,j\in[N]} A_{i,j}^{\top} A_{i,j} \right\| \right\}$$

当 $|\lambda| < {}_2 \quad {}^{3}_{(1+\epsilon_N)R}$ 时，

$$\mathbb{E}\left[\overline{\exp}\left\{ \lambda \sum_{k=1}^{N} A_{k,\pi(k)} \right\} \right] \le \exp\left\{ \frac{\lambda^2 (1+\epsilon_N)^2}{2\left(1 - \frac{2|\lambda|(1+\epsilon_N)R}{3}\right)} \cdot \left(\frac{1}{N} + \epsilon_N \cdot \left(1 - \frac{1}{N}\right)\right) \sigma^2 \right\},$$

因此，对于 $\delta \in (0,1)$，

$$\mathbb{P}\left\{ \left\| \sum_{k=1}^{N} A_{k,\pi(k)} \right\| \ge (1+\epsilon_N)\sigma \sqrt{2\left(\frac{1}{N} + \epsilon_N \cdot \left(1 - \frac{1}{N}\right)\right) \cdot \log\frac{m+n}{\delta}} + \frac{2(1+\epsilon_N)R}{3} \log\frac{m+n}{\delta} \right\} \le \delta.$$

证明推迟至附录 D.2。该结果以相同的速率重现了 Mackey 等人 [30, 推论10。3] 中的结论 N，同时得到了更紧的常数。

5 应用

接下来，我们回到第 1 节中介绍的两个示例应用，探讨我们的结果在这些具体情境中的含义。第 5.1 节重新讨论示例 1，而第 5.2 节和 5.3 节则进一步展开示例 2 中提出的问题。本节实验的代码见此处[1]。

5.1 用于估计多因子平均效应的实验设计

我们重新讨论示例 1，在该示例中，我们的目标是估计加权均值

$$\mu = \sum_{i_l\in[N_l], l\in[K]} p_{i_1\cdots i_K} Y_{i_1\cdots i_K}$$

同时只对张量 Y 的一个区块进行子抽样：我们只观测 $Y_{i_1\cdots i_K}$ 对应的索引 $(i_1,\ \ldots\ldots,\ i_K) \in I_1 \times \cdots \times I_K$。我们将使用霍维茨-汤普森估计量 $\hat{}\mu$，与示例 1 中相同。根据构造，$\widehat{\mu}$ 的均值为 μ，方差为

$$\sigma^2 = \prod_{l=1}^{K} N_l \cdot \sum_{i_l\in[N_l], l\in[K]} p_{i_1\cdots i_K}^2 Y_{i_1\cdots i_K}^2 - \mu^2.$$

[1] https://github.com/Moriartycc/exchangeable-tensors-matrix-valued-data

估计误差的尾界。接下来的估计误差界由定理2得出。证明见附录D.3。

推论 5.1。 对于 $k \in [K]$，令 $B_k > 0$ 表示对所有 $i_k \in [N_k]$、$\sum_{i_l \in [N_l],\ l \neq k} p_{i_1 \cdots i_K} |Y_{i_1 \cdots i_K}| \le B_k$ 以及 $\alpha_k \in (0, 1]$，满足 $\prod_{l=1}^k n_l = \alpha_k \prod_{l=1}^k N_l$ 的子抽样率。对于所有 $\delta \in (0, 1)$，

$$\mathbb{P}\left\{\widehat{\mu} - \mu \ge \sigma \sqrt{2 \log \frac{1}{\delta} \cdot \max_{k \in [K]} \frac{1 + \epsilon_{N_k}}{N_k \alpha_k} \cdot \left(1 - \frac{n_k}{N_k} + \frac{\epsilon_{N_k}}{\alpha_k} \cdot \mathbb{1}_{n_k < N_k}\right)} + \frac{2 \max_{k \in [K]} B_k (1 + \epsilon_{N_k})}{3 \alpha_K} \log \frac{1}{\delta}\right\} \le \delta. \tag{9a}$$

对于满足 $N_K \alpha \ge 1$ 为整数的 $\alpha \in (0, 1)$，在笛卡尔子抽样配置 $(n_1, \cdots, n_K)$ 下，我们有如下尾界 $= (N_1, \cdots, N_{K-1}, N_K \alpha)$：

$$\mathbb{P}\left\{\widehat{\mu} - \mu \ge \sigma \sqrt{2 \log \frac{1}{\delta} \cdot \frac{1 + \epsilon_{N_K}}{N_K \alpha} \cdot \left(1 - \alpha + \frac{\epsilon_{N_K}}{\alpha}\right)} + \frac{2 \max_{k \in [K]} B_k (1 + \epsilon_{N_k})}{3 \alpha} \log \frac{1}{\delta}\right\} \le \delta. \tag{9b}$$

一般而言（假设方差项主导前述尾界），我们可以从推论 5.1 得出，在子抽样率约束 $\alpha \ge 1/N_K$ 下，我们可以将偏差控制在 $1 - \delta$ 概率不超过

$$O\left(\sigma \sqrt{\log \frac{1}{\delta}} \cdot \frac{\sqrt{\log N_K}}{\alpha N_K} \vee \sqrt{\frac{1 - \alpha}{\alpha N_K}}\right),$$

恢复具有 $p = \alpha$ 的 N_K 独立伯努利随机变量的收敛速率。

作为比较，我们可以考虑另一种抽样策略：假设 $X_{i_1 \cdots i_K}$ 均为具有 $p = \alpha$ 的独立同分布伯努利随机变量。我们预期其尾界为

$$O\left(\sigma \sqrt{\log \frac{1}{\delta}} \cdot \sqrt{\frac{1 - \alpha}{\alpha N_1 \cdots N_K}}\right)$$

根据标准集中不等式。我们建议主要在某一维度远大于其他所有维度之和的情况下使用笛卡尔积子抽样策略，即当$N_K \gg N_1 N_2 \cdots N_{K-1}$时。更一般地，正如第 1 节所述，即使笛卡尔积子抽样无法带来更好的集中效果，出于实际实现方面的考虑，也可能需要采用该方法。（当然，由于我们可以自由地对 Cor. $(N_1, \cdots, N_K)$ 5.1 中的量重新排序，根据式（9b），可以看出，对于任意 i、$j \in [K]$，在相同的速率 α 下，若 $N_i > N_j$，沿模 i 进行子抽样的尾界比沿模 j 进行子抽样更紧。）

数值实验。为验证我们的理论预测，我们进行了数值模拟。对于 $K = 3$，我们构造权重张量 $W_{ijk} = (ijk)^2/(N_1 N_2 N_3)^2 \in [0, 1]$。考虑以下三种笛卡尔子抽样程序

$$I \times J \times K \subset [N_1] \times [N_2] \times [N_3]$$

使得 (i) $|I| = 0.4 N_1$，$|J| = N_2$，$|K| = N_3$; (ii) $|I| = N_1$，$|J| = 0.4 N_2$，$|K| = N_3$; (iii) $|I| = N_1$，$|J| = N_2$，$|K| = 0.4 N_3$; 以及 (iv) 采用 Bernoulli(0.4) 进行独立同分布子抽样，即每个条目以 0.4 的概率被观测到。这四种方案对应相同的 40% 抽样率。

我们的推论 5.1 理论预测，对于抽样策略 (i)、(ii)、(iii)，估计误差分别为 $O(1/\sqrt{N_1})$、$O(1/\sqrt{N_2})$、$O(1/\sqrt{N_3})$ 而独立同分布抽样 (iv) 的估计误差为

is $O(1/\sqrt{N_1N_2N_3})$。我们报告模拟结果，并在两种设置下于图 1 中验证这些预测：(A) $(N_1,N_2,N_3)=(5,10,N)$ 和 (B) $(N_1,N_2,N_3)=(0.1N,0.2N,0.5N)$，其中 N 取值于 $\{10,20,\cdots,500\}$。正如预测所示，独立同分布抽样 (iv) 的误差最低；但是，当 N_1,N_2 较小而 N_3 较大时（如设置 (A)），沿模态 3 子抽样 (iii) 的误差远低于沿另外两个模进行子抽样的策略 (i) 和 (ii)，这与预测一致；另一方面，当 N_1,N_2,N_3 相近时（如设置 (B)），抽样策略 (i)、(ii)、(iii) 的表现相似，不过其相对排序再次与维度的排序 $N_1<N_2<N_3$一致。

5.2 通过离散正弦变换进行固定设计草图化

接下来，我们将数学工具应用于示例2。考虑 $X_k=\theta_k\in\mathbb{R}^{q\times r}$ 一列可交换矩阵，我们的目标是在 $\bar{\theta}=\frac{1}{N}\sum_{k=1}^N\theta_k$ 的同时，通过草图化降低通信成本。

令 $U\in\mathbb{R}^{q\times q}$为一个固定的正交矩阵。令 $D_1,\dots,D_N$为非负对角矩阵，每个矩阵至多有 q'个非零项，且满足 $\sum_{k=1}^N D_k=I_q$。对于每个 k，我们都可以构造一个矩阵$U_k\in\mathbb{R}^{q'\times q}$，使得 $U_k^\top U_k=U^\top D_kU$——也就是说，U_k的各行是通过选取（并重新缩放）U中与 D_k的非零项相对应的行构造而成的。随后，这些固定矩阵 $U_1,\dots,U_N$将用作草图矩阵。

推论 5.2。设 $\theta_1,\cdots,\theta_N$ 是 $\mathbb{R}^{q\times r}$ 中由数据矩阵组成的可交换序列。设 $U_1,\dots,U_N\in\mathbb{R}^{q'\times q}$ 为按上述方式构造的固定矩阵。假设存在常数 $L,M>0$，使得$\|\theta_k\|_F\le 1$,$\|U_k^\top U_k\|=\|D_k\|\le L$ 和 $\|\sum_{k=1}^N U_k^\top U_kU_k^\top U_k\|=\|\sum_{k=1}^N D_k^2\|\le M$，并定义方差参数

$$\widetilde{\sigma}_\theta^2:=\frac{1}{N}\sum_{k=1}^N\left\|\theta_k-\bar{\theta}\right\|_F^2+\epsilon_N\max_{k\in[N]}\left\|\theta_k-\bar{\theta}\right\|_F^2.$$

对于所有 $\delta\in(0,1)$，均有

$$\mathbb{P}\left\{\left\|\sum_{k=1}^N U_k^\top U_k\theta_k-\bar{\theta}\right\|\ge\widetilde{\sigma}_\theta\sqrt{2(1+\epsilon_N)q'M\log\frac{q+r}{\delta}}+\frac{2L}{3}(1+\epsilon_N)\log\frac{q+r}{\delta}\right\}\le\delta. \tag{10}$$

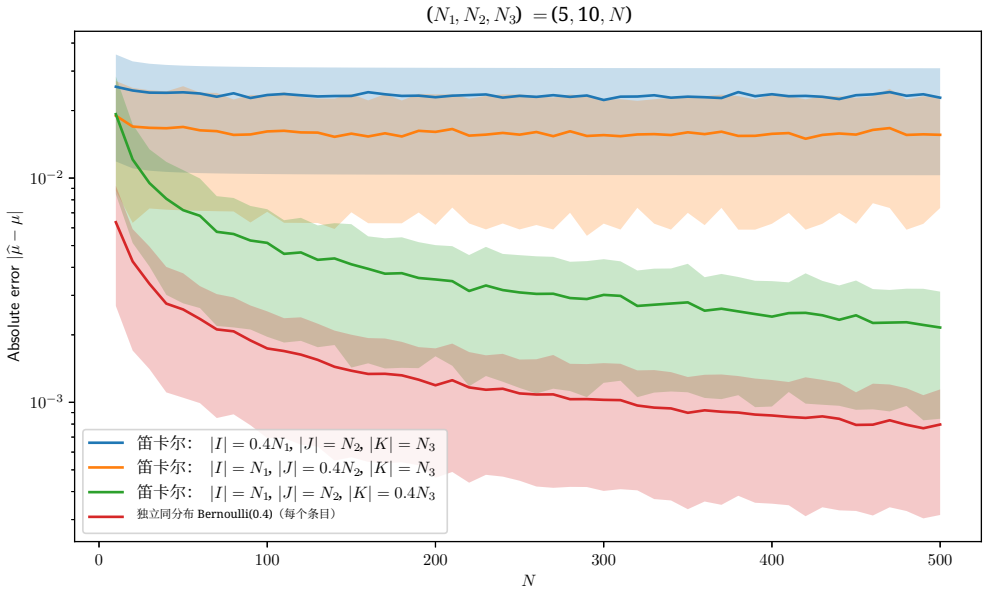
其证明留待附录 D.4。我们指出，这里给出结果时采用 q'，以直观的方式展示通过草图化实现降维的优势；不过，推论5.2 在 N中仍能给出紧确的界，只要 $q'\asymp q$。例如，当 $D_k=I_q/N$（即 $q'=q$，且 $M=1/N$）时，尾界（10）中的主导项按 $q'M=q/N$ 的尺度变化。

为了了解我们的结果如何应用于具体情形，考虑正交离散正弦变换（DST）矩阵 $U\in\mathbb{R}^{q\times q}$，其元素为

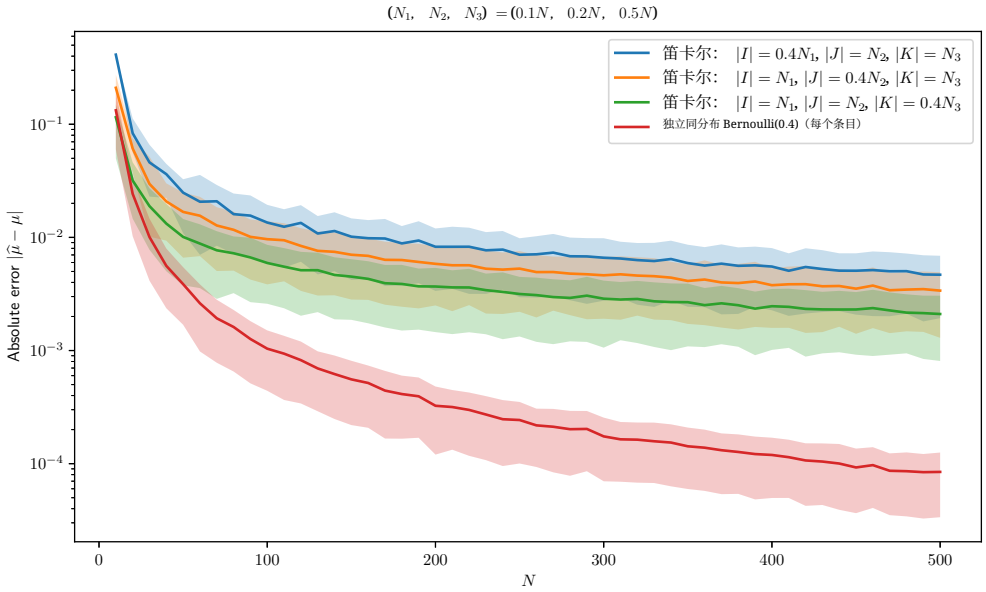
$$U_{ij}=\sqrt{\frac{2}{q+1}}\sin\left(\frac{\pi ij}{q+1}\right).$$

令 $U=\begin{bmatrix}u_1&\cdots&u_q\end{bmatrix}^\top$，我们采用周期索引约定定义 $U_k=\sqrt{q/(q'N)}\cdot\begin{bmatrix}u_{(k-1)q'+1}&\cdots&u_{(k-1)q'+q'}\end{bmatrix}^\top$，即对所有 $i'\in\mathbb{N}_+$，以及满足 $i'\equiv i$ (模 q) 的 $i\in[q]$，定义 $u_{i'}=u_i$。当 $q'N$ 是 q 的倍数时，容易验证 $\sum_{k=1}^N U_k^\top U_k=U^\top U=I_q$。若进一步有 $q'\le q$，则推论 5.2 适用于 D_k，其第 i 个对角元素在 $(k-1)q'<i\le(k-1)q'+q'$ 时为 $q/(q'N)$，否则为零。因此，应用推论 5.2 时的参数为 $L=M=q/(q'N)$，这意味着

$$\mathbb{P}\left\{\left\|\sum_{k=1}^N U_k^\top U_k\theta_k-\bar{\theta}\right\|\ge\widetilde{\sigma}_\theta\sqrt{\frac{2q(1+\epsilon_N)}{N}\cdot\log\frac{q+r}{\delta}}+\frac{2q(1+\epsilon_N)}{3q'N}\log\frac{q+r}{\delta}\right\}\le\delta.$$



设置 (A) 下的实验



设置 (B) 下的实验

图1. 子抽样程序的数值模拟。在两种（N_1, N_2, N_3）配置下，我们模拟笛卡尔积子抽样：(i) $|I| = 0.4N_1$, $|J| = N_2$, $|K| = N_3$；(ii) $|I| = N_1$, $|J| = 0.4N_2$, $|K| = N_3$；(iii) $|I| = N_1$, $|J| = N_2$, $|K| = 0.4N_3$；并纳入 (iv) 使用 Bernoulli(0.4) 的独立同分布子抽样。对于每个 N，我们对每种程序运行 $T = 1,000$ 次独立试验，并报告霍维茨-汤普森估计量$|\hat{\mu} - \mu|$的绝对估计误差均值，以及第 25 至第 75 分位数之间的误差带。

从上述界可以看出一些值得注意的结论：(i) 对独立同分布随机变量 U_k 进行采样不会产生额外的波动，因为主导随机项 $O(\bar{\sigma}_\theta/\sqrt{N})$ 仅取决于平均数据矩阵；(ii) 主导噪声项与压缩空间的维度 q' 无关；(iii) 随着 q' 增大，阶为$O(1/N)$的高阶噪声项对应的界会变得更紧。

数值实验。 令 $q = 1000$， $r = 1$， $N = 50$。我们构造一组可交换向量θ_1， $\cdots$， θ_N独立同分布~ $\mathrm{N}(\mu,\ I_q/q)$，条件是给定独立采样的 $\mu \sim \mathrm{N}(0,\ \Sigma/q)$，其中 Σ 是 AR(1) 协方差矩阵 $\Sigma_{ij} = \rho^{|i-j|}$。在同一数据集上，并取一个满足 $q'N$ 是 q 的倍数的 q'，我们测试四种草图方案：

(A) 上述固定设计草图矩阵；(B) 无放回随机子采样 DST 草图矩阵 $U_k^\top U_k = q/(q'N) \cdot U D_k U$，其中矩阵 D_k 的对角元素是相互独立的随机变量，服从 Bernoulli(q'/q) 分布；(C) 有放回随机子采样 DST 矩阵。与 (B) 不同，对每个 D_k，独立生成 q 个随机数对 (i_l, ξ_l) $l \in [q]$，其中 i_l 在 $[q]$ 上服从均匀分布， ξ_l 服从 Bernoulli(q'/q) 分布。然后构造 $\mathrm{diag}(\widetilde{D}_k)_i = \sum_{l=1}^{q} \mathbb{1}_{\xi_l=1, i_l=i}$；(D) 随机高斯草图矩阵，其 U_k的各行独立地从 $\mathrm{N}(0, I_q/(q'N))$ 中抽取；

我们在图中展示了数值模拟结果。2，从中可以清楚地看到方法 (A) 中固定设计草图化的优势，尤其是与方法 (C) 中对 DST 的行进行有放回子抽样，以及方法 (D) 中完全随机化的高斯草图化相比。

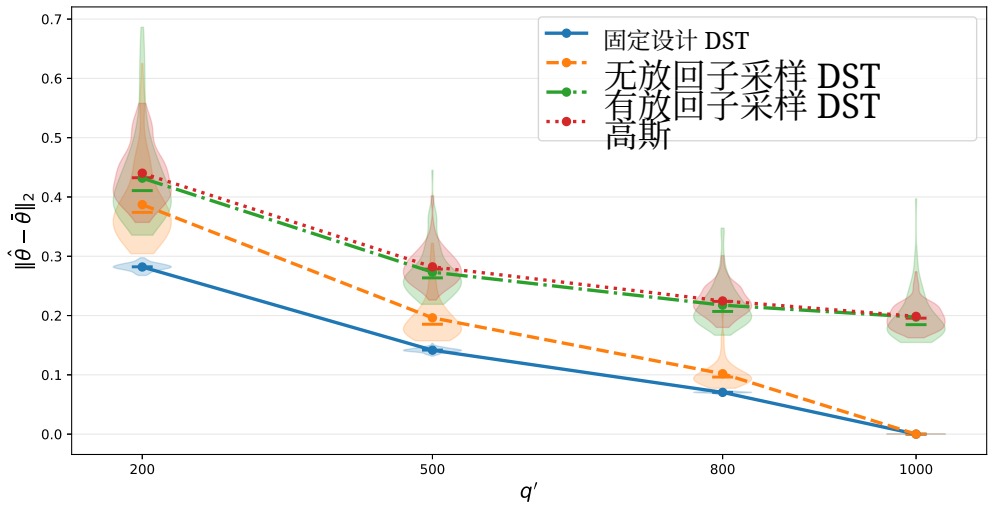


图 2。 当 $\rho = 0.99$时误差统计量 $\|\widehat{\theta} - \theta\|$ 的小提琴图，对比以下草图方案：(A) 固定设计 DST，(B) 子采样 DST（无放回），(C) 子采样 DST（有放回）和 (D) 高斯草图。我们对 $q' \in \{200, 500, 800, 1000\}$进行了 100 次试验。短横线"–"表示中位数，圆点"·"表示均值。曲线连接了 q'不同取值对应的均值。

5.3 通信高效的岭调优联邦平均

在固定设计草图化的模拟示例基础上，我们现在转向一个具体场景：对相互依赖但可交换的参数进行联邦平均 $\theta_1, \cdots, \theta_N$ 在该场景中具有实际意义。正如示例2中简要提到的，联邦学习（FL）涉及一组 N 客户端或设备，通过本地计算并与中央服务器协调，协作求解机器学习模型 [24]。

我们研究标准联邦平均的一种推广 [34] ，其中 [28, 13], 利用中央服务器的广播，通过相对于全局参数的岭调优惩罚项，在本地客户端上微调个性化模型。有关通过与中央服务器通信并与本地模型交互以实现个性化的类似思想，请参见 [42] 。不同于第 5.2 节中的静态平均设置，在每一轮 $t = 0$、1、2、$\cdots$、T中，我们考虑一个全局参数 $\theta_t \in \mathbb{R}^{q\times r}$，在 θ_0 中初始化；以及局部参数θ_k、$_t \in \mathbb{R}^{q\times r}$，其初始化在 $t = 0$ 中独立同分布地进行，适用于 $k \in [N]$。为此，我们引入一系列公开已知的固定设计草图矩阵 $U_k \in \mathbb{R}^{q'\times q}$，其中 $q' \ll q$，以及满足 $\sum_{k=1}^N U_k^\top U_k = I_q$的 $k \in [N]$。随后，算法按照算法1进行。

算法1 固定设计草图岭调优联邦平均

1: **给定：** 公开已知的草图矩阵 $U_1, \cdots, U_N \in \mathbb{R}^{q'\times q}$.2: **结果：** 全局参数 θ_T 和局部参数 $\theta_{k,\ T}$.3: **对于** $t \leftarrow 0$ **到** $T-1$ **执行**4: 服务器向本地代理广播 θ_t .5: 服务器在 $[N]$ 上生成一个新的随机排列 π，并广播给各代理。6: **并行** $k \in [N]$ **执行**7: 令 $\theta_k^{(0)}$, $_t = \theta_{k,\ t}$.8: 在本地代理上以步长 η 执行 m_t 步本地梯度下降，使得

$$\theta_{k,t}^{(i+1)} \leftarrow \theta_{k,t}^{(i)} - \eta\left(\nabla L_k(\theta_{k,t}^{(i)}) + \lambda(\theta_{k,t}^{(i)} - \bar{\theta}_t)\right), \qquad i = 0, 1, \cdots, m_t - 1.$$

9: $\theta_{k,\ t+1} \leftarrow \theta^{(m_t)}{}_{k,\ t}$。上传草图参数 $\phi_{k,\ t+1} \to U_{\pi(k)}\theta_{k,\ t+1} \in \mathbb{R}^{p\times r}$。10: **循环结束**11: 中央服务器聚合 $\theta_{t+1} = \sum_{k=1}^N U_{\pi\ (k)}^\top \phi_{k,\ t+1} = \sum_{k=1}^N U_k^\top U_k \theta_{\pi^{-1}(k)\ t+1}$。12: **循环结束**

我们指出，在每一轮 t 中，$(\theta_{\pi^{-1}\ (1)\ t}, \cdots, \theta_{\pi^{-1}\ (N)\ t})$ 都是可交换的，但并非相互独立，因为它们共享来自先前聚合估计的互信息，θ_t。固定的草图矩阵序列使我们能够考虑更广泛的结构化草图设计，例如傅里叶基，而不必在每一步都生成新的高斯随机矩阵 [5]或计数草图 [23] 来确保独立性，同时还能大幅降低上传通信成本。

数值实验。令 $q = 1000, r = 1, n = 200, N = 50$ 和 $\theta^\star = e_1 = (1, 0, \cdots, 0)^\top \in \mathbb{R}^q$。假设对于每个本地代理，其观测到的数据 X_k 的各行均独立同分布于 $\mathsf{N}(0, I_q/q)$，且 $y_k = X_k\theta^\star + \epsilon_k \in \mathbb{R}^n$。设 $\epsilon_1, \cdots, \epsilon_N$ 独立同分布于 $\mathsf{N}(0, \sigma^2 I_n)$，本地损失函数为 $L_k(\theta) = \|X_k\theta - y_k\|_2^2$。为简化实验，我们在算法1中设定 $m_t = \infty$，即在每一轮都通过闭式更新精确求解优化问题

$$\theta_{k,t} = (X_k^\top X_k + \lambda I_q)^{-1}(X_k^\top y_k + \lambda\bar{\theta}_{t-1}), \qquad \text{for all } k \in [N].$$

在没有通信约束时，当 $t \to \infty$ 对于 θ_t i $\quad -$ s

$$\bar{\theta}^\star = \frac{1}{N}\sum_{k=1}^{N}(X_k^\top X_k + \lambda I_q)^{-1}(X_k^\top y_k + \lambda\bar{\theta}^\star),$$

因此

$$\bar{\theta}^\star = \left\{\sum_{k=1}^{N} X_k^\top X_k (X_k^\top X_k + \lambda I_q)^{-1}\right\}^{-1}\left\{\sum_{k=1}^{N}(X_k^\top X_k + \lambda I_q)^{-1} X_k^\top y_k\right\}.$$

考虑两种草图设置：(i) 第 5.2 节介绍的固定设计 DST 草图矩阵；(ii) 随机高斯草图矩阵。我们以 (ii) 为基线，其中 $U_k \in \mathbb{R}^{q'\times q}$ 在每一轮随机生成 t，其各行均为来自 $\mathsf{N}(0, I_q/(q'N))$ 的独立同分布向量，此时 $\sum_{k=1}^{N}\mathbb{E}[U_k^\top U_k] = I_{q\circ-}$ 从零初始化 $\theta_0 = 0$开始，我们运行算法 1，令 $m_t = \infty$，并绘制两种草图方案$\|\theta_t - \theta^\star\|$ 在每次迭代时的误差曲线。模拟结果见图 3，结果表明，采用固定设计草图可降低高斯草图矩阵固有的方差，从而显著提高联邦平均估计精度。

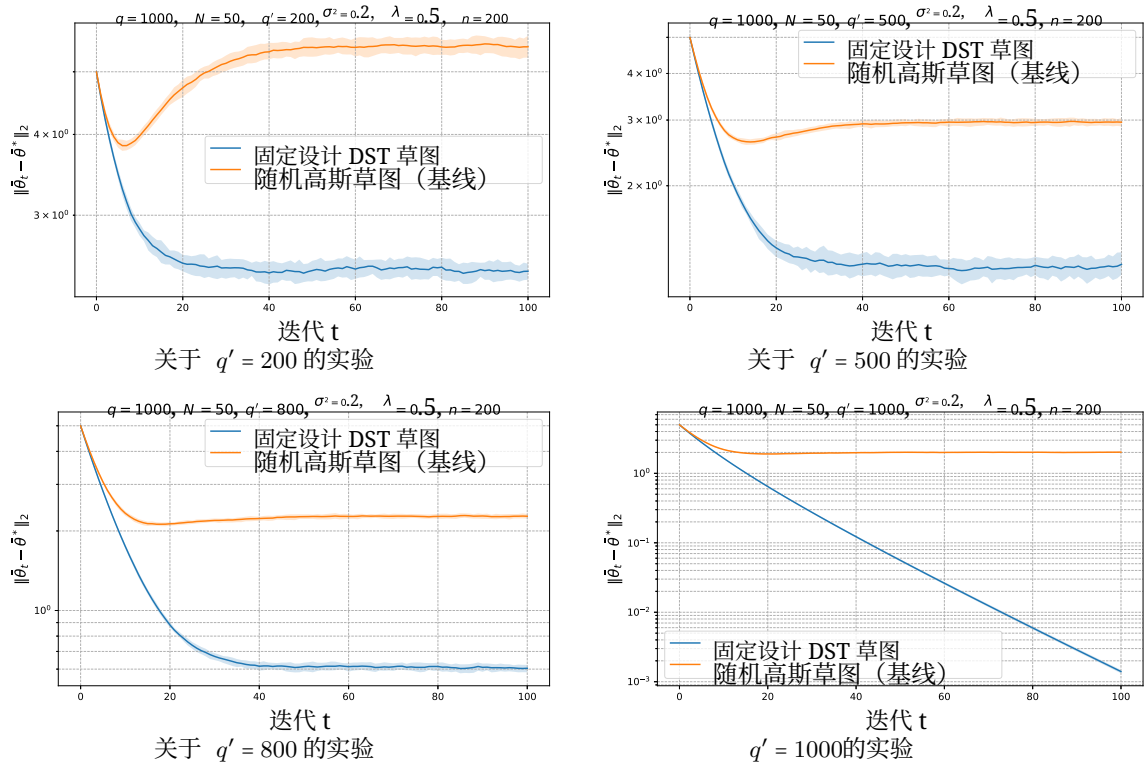


图 3. 针对 (i) 的草图 RTFA 数值模拟：蓝色：固定设计 DST 草图 与 (ii)橙色：高斯草图。我们针对 $\sigma^2 = 0.2$、$\lambda = 0.5$ 和 $q' \in \{200、500、800、1000\}$进行实验，共 100 次试验。我们绘制了 $\|\theta_t - \theta^\star\|$ 在每次迭代 t 时的均值误差曲线，误差带表示第 25 至第 75 分位数。

6 讨论

我们为模可交换张量和可交换矩阵值数据的加权和提供了 Hoeffding 型和 Bernstein 型集中界。我们得到的界恢复并

改进了此前已知的结果，尤其是在独立同分布情形、随机标量的加权和以及矩阵数组的组合和方面。我们能够利用本文研究的更丰富的分布对称性类别，分析更广泛的问题，包括估计多因子模型的平均效应，以及应用于联邦平均的固定设计草图化。尽管如此，仍有一些值得探讨的问题可作为未来的研究方向。

非线性交换和。 本文的结果聚焦于双线性映射 $Z = \langle\!\langle W, X \rangle\!\rangle$ 的集中不等式。一个有趣的问题是，在张量或矩阵值数据具有相同可交换性假设的条件下，一般非线性映射 $Z = f(W, X)$ 是否会出现集中。在无权且独立的情形下，McDiarmid 不等式 [33] 成立，并表明在诸如 f 的利普希茨连续性等温和假设下，加权且可交换的情形也应有类似结果。作为一个具体示例，对于矩阵值数据，为非线性求和 $Z = \sum_{k=1}^{N} A_k W_k B_k$（其中 $X = (A_k, B_k)$）建立集中界，在盲反卷积 [11]等问题的分析中至关重要。

非渐近最优性。 正如先前在 [3], 中所提出的，一个有趣的问题是我们理论结果中乘法项 $(1+\epsilon_{N_k})$ 的非渐近最优性。在随机变量无放回抽样的无权情形下，Serfling 不等式 [40]表明此类放大常数并非必要（事实上，无放回抽样甚至可以导致比独立同分布界更强的集中）；另一方面， [3]表明对于一般可交换情形，某种放大是不可避免的。在本文研究的张量数据与矩阵数据情形中，是否存在这种放大并非必要的情形（类似于无放回抽样）？又有哪些具体示例会使放大的尾部出现? 另一个问题是张量结构的作用：我们对多路张量的界（参见定理 4 与 5）通常依赖于各模的顺序 $(N_1, \cdots, N_K)$。当然可以通过选择最优顺序来优化这些界，但这也表明，有可能改进为与模态顺序无关的结果。

矩阵值数据更丰富的对称结构。我们的论文将Barber的结果推广到了表1所示的两个独立分支。一个自然的未来方向是发展关于如下求和的一般结果： $Z = \sum_{i=1}^{N_1} \cdots \sum_{i=1}^{N_K} W_{i_1 \cdots i_K} X_{i_1 \cdots i_K}$， 其中 $W_{i_1 \cdots i_K}$和 $X_{i_1 \cdots i_K}$均为矩阵值，且 $X_{i_1 \cdots i_K}$ 为模态可交换，从而提供一个允许矩阵值数据具有更丰富对称结构的统一框架。这有助于简化第4.3节中矩阵数组组合和的证明，因为我们可以自然地取 $K = 2$, $X_{ij} = A_{\pi(i),\ \pi'(j)}$和 $W_{ij} = \mathbb{1}_{i=j}$，从而避免像本文中那样仔细构造可交换对。

其他对称群与分布不变性。 从更高层次看，可交换性是将独立性放宽为分布对称性。按模以及按行列的可交换性是群作用下不变性和较弱对称性的特殊情形，相较于序贯可交换性而言如此。然而，群作用下分布对称性的概念可以更为一般，例如循环群对称性，其中 $(X_1, \cdots, X_N) \overset{d}{=} (X_2, \cdots, X_N, X_1) \overset{d}{=} \cdots \overset{d}{=} (X_N, X_1, \cdots, X_{N-1})$。对于 $(X_1, \cdots, X_N) \sim \mathrm{N}(0, \Sigma)$，其中 Σ 是一般的半正定托普利兹矩阵，此时通常的可交换性不再成立。未来的研究方向是，在给定一般群作用集合的情况下推导集中界。

致谢

R.F.B. 获美国国家科学基金会资助，资助项目编号为 DMS-2023109。C.C. 和 R.F.B. 还获得美国海军研究办公室资助，资助项目编号为 N00014-24-1-2544。Chen Cheng 也要感谢 Gary Cheng 提供的宝贵讨论。

参考文献

[1] R. Ahlswede 和 A. Winter. 通过量子信道进行识别的强逆定理. IEEE信息论汇刊, 48(3):569–579, 2002.[2] A. N. Angelopoulos 和 S. Bates. 保形预测：简明介绍. 机器学习基础与趋势, 16(4):494–591, 2023.[3]R. F. Barber. 可交换随机变量加权和的 Hoeffding 不等式和 Bernstein不等式. Electronic Communications in Probability, 29(ECP620), 2024. doi: 10.1214/24-ECP620. 发表于 2024 年 4 月；亦可见于 arXiv.[4] R. Bardenet 和 O.-A. Maillard. 无放回抽样的集中不等式. Bernoulli, 21(3):1361–1385, 2015.[5] B. Bartan 和 M. Pilanci. 随机优化的分布式草图：精确刻画、集中与下界. IEEE信息论汇刊, 69(6): 3850–3879, 2023.[6] S. Bernstein. 概率论, 1946.[7] S. Boucheron、G. Lugosi 和 P. Massart. 集中不等式：独立性的非渐近理论. 牛津大学出版社, 02 2013. ISBN 9780199535255. doi: 10.1093/acprof:oso/9780199535255.001.0001. URL https://doi.org/10.1093/acprof:oso/9780199535255.001.0001.[8] M. Charikar、K. Chen 和 M. Farach-Colton. 在数据流中查找频繁项. 载于自动机、语言与编程国际学术讨论会, 第 693–703 页. 施普林格, 2002.[9] S. Chatterjee. 可交换对的集中不等式. 斯坦福大学, 2005.[10] S. Chatterjee. 用于集中不等式的 Stein方法. 概率论及相关领域, 138, 2007.[11] Y. Chen、Y. Chi、J. Fan、C. Ma 和 Y. Yan. 含噪矩阵补全：通过非凸优化理解凸松弛的统计保证. SIAM优化杂志, 30(4):3098–3121, 2020.[12] C. Cheng、G. Cheng 和 J. C. Duchi. 利用结构化缺失数据协同学习线性模型. 神经信息处理系统进展, 36:7529–7540, 2023.[13] G. Cheng、K. Chadha 和 J. Duchi. 联邦学习中的微调效果不错. arXiv 预印本 arXiv:2108.07313, 3, 2021.[14] B. De Finetti. 随机现象的特征函数. 载于《国际数学家大会会议录：1928 年 9 月 3 日至 10 日，博洛尼亚》，第 179–190 页, 1929.

[15] P. Diaconis. 德·菲内蒂可交换性定理的有限形式. *Synthese*, 36(2):271–281, 1977.[16] L. Gavalakis 和 I. Kontoyiannis. 有限德·菲内蒂定理的信息论证明.*Electronic Communications in Probability*, 26:1–5, 2021.[17] E. Greene 和 J. A. Wellner. 超几何分布的指数界. *Bernoulli:* 伯努利数学统计与概率学会官方期刊, 23(3):1911, 2017.[18] N. Halko、P.-G. Martinsson 和 J. A. Tropp. 借助随机性发现结构：用于构造近似矩阵分解的概率算法. *SIAM Review*, 53(2):217–288, 2011.[19]F. Han. 置换过程导论（版本 0.5）. *arXiv* 预印本 *arXiv:2407.09664*, 2024.[20] L. He、E. Kulesskiy、J. Saarela、L. Turunen、K. Wennerberg、T. Aittokallio 和 J. Tang. 高通量药物组合筛选与协同作用评分方法. 载于癌症系统生物学：方法与规程, 第 351–398 页. 施普林格, 2018.[21] W. Hoeffding. 有界随机变量之和的概率不等式. 美国统计学会会刊, 58(301):13–30, 1963.[22] R. A. Horn 和 C. R. Johnson. 矩阵分析专题. 剑桥大学出版社, 1994.[23] N. Ivkin、D. Rothchild、E. Ullah、I. Stoica、R. Arora 等. 基于草图化的通信高效分布式随机梯度下降. 神经信息处理系统进展, 32, 2019.[24] P. Kairouz、H. B. McMahan、B. Avent、A. Bellet、M. Bennis、A. N. Bhagoji、K. Bonawitz、Z. Charles、G. Cormode、R. Cummings 等. 联邦学习的进展与未决问题. 机器学习基础与趋势® 机器学习, 14(1–2):1–210, 2021.[25] T. G. Kolda 和 B. W. Bader. 张量分解及其应用. *SIAM Review*, 51(3): 455–500, 2009.[26] T. Le Minh. 二部可交换网络上的 U统计量. *ESAIM*：概率与统计, 27:576–620, 2023.[27] J. Lei、M. G'Sell、A. Rinaldo、R. J. Tibshirani 和 L. Wasserman. 回归的无分布预测推断. 美国统计学会会刊, 113(523):1094–1111, 2018.[28] T. Li、S. Hu、A. Beirami 和 V. Smith. Ditto：通过个性化实现公平、稳健的联邦学习. 载于国际机器学习大会, 第 6357–6368 页. PMLR, 2021.[29] E. H. Lieb. 凸迹函数与维格纳–柳泽–戴森猜想. 斯特拉斯堡物理学家与数学家会议-*RCP25*, 19:0–35, 1973.[30] L. Mackey、M. I. Jordan、R. Y. Chen、B. Farrell 和 J. A. Tropp. 通过可交换对方法得到的矩阵集中不等式. 概率年刊, 42(3):906–945, 2014.

[31] M. W. Mahoney 等人. 矩阵与数据的随机化算法. 基础与趋势®机器学习, 3(2):123–224, 2011.[32] G. S. May, J. Huang, 和 C. J. Spanos. 等离子体蚀刻建模中的统计实验设计.*IEEE*半导体制造汇刊, 4(2):83–98, 1991.[33] C. McDiarmid 等人. 关于有界差分法. 组合学综述, 141(1): 148–188, 1989.[34] B. McMahan, E. Moore, D. Ramage, S. Hampson, 和 B. A. y Arcas. 基于去中心化数据的深度网络通信高效学习. 载于人工智能与统计学, 第1273–1282页. PMLR, 2017.[35] V. Paulsen. 完全有界映射与算子代数, 第78卷. 剑桥大学出版社, 2002.[36] S. Pelletier, O. Jabali, G. Laporte, 和 M. Veneroni. 电动汽车的电池衰减与性能：若干模型的综述与数值分析. 交通研究*B*辑：方法论, 103:158–187, 2017.[37] A. Ramdas 和 T. Manole. 马尔可夫、切比雪夫和切尔诺夫不等式的随机化与可交换改进. 统计科学, 41(1):121–142, 2026.[38] A. Ramdas, R. F. Barber, E. J. Candès, 和 R. J. Tibshirani. 使用任意置换分布的置换检验. $Sankhya\ A$, 85(2):1156–1177, 2023.[39] L. A. Román-Ramírez 和 J. Marco. 应用于锂离子电池的实验设计：文献综述. 应用能源, 320:119305, 2022.[40] R. J. Serfling. 无放回抽样中样本和的概率不等式. 统计年鉴, 第39–48页, 1974.[41] G. Shafer 和 V. Vovk. 保形预测教程. 机器学习研究杂志, 9(3), 2008.[42] C. T Dinh, N. Tran, 和 J. Nguyen. 基于莫罗包络的个性化联邦学习.神经信息处理系统进展, 33:21394–21405, 2020.[43] J. A. Tropp. 易于使用的随机矩阵和尾界. 计算数学基础, 12(4):389–434, 2012.[44] J. A. Tropp 等人. 矩阵集中不等式导论. 基础与趋势® 机器学习, 8(1-2):1–230, 2015.[45] M. J. Wainwright. 高维统计学：非渐近视角, 第48卷. 剑桥大学出版社, 2019.[46] I. Waudby-Smith 和 A. Ramdas. 无放回抽样的置信序列.神经信息处理系统进展, 33:20204–20214, 2020.

A 技术引理

本节列出了附录中主要证明部分通篇使用的技术引理。为求完整，尽管在使用这些引理时我们可能不会明确提及，仍将它们列于此处。

引理 A.1. 对于两个张量 A、$B \in \mathbb{T} := \mathbb{R}^{N_1 \times \cdots \times N_K}$、一个整数 $l \in [K]$ 和两个矩阵C、$D \in \mathbb{R}^{N_l \times N_l}$，有

$$A \times_l C \times_l D = A \times_l DC,$$
$$\langle A \times_l C, B \rangle_{\mathbb{T}} = \langle A, B \times_l C^\top \rangle_{\mathbb{T}}.$$

证明 这是一个初等线性代数练习。 □

引理 A.2. 对于任意满足 $A\mathbf{1}_n = 0$ 的矩阵 $A \in \mathbb{R}^{n \times n}$，有

$$\frac{1}{n!} \sum_{\Pi_n \in \mathscr{P}_n} \Pi_n^\top A \Pi_n = \frac{\mathsf{Tr}(A)}{n-1} \cdot \mathsf{P}_{\mathbf{1}_n}^\perp.$$

证明 令 A' 为左侧的量。由 $\Pi_n^\top A' \Pi_n = A'$ 可知，对于任意 $\Pi_n \in \mathscr{P}_n$，A' 的非对角元素和对角元素必须分别相同，即 $A' = aI_n + b\mathbf{1}_n\mathbf{1}_n^\top$。另一方面，

$$A'\mathbf{1}_n = \frac{1}{n!} \sum_{\Pi_n \in \mathscr{P}_n} \Pi_n^\top A \Pi_n \mathbf{1}_n = A'\mathbf{1}_n = \frac{1}{n!} \sum_{\Pi_n \in \mathscr{P}_n} \Pi_n^\top A \mathbf{1}_n = 0$$

强制 $A' = c\mathbf{P}_{\mathbf{1}_n}^\perp$。为求常数 c，我们对等式两边取迹，这意味着

$$(n-1)c = c\,\mathsf{Tr}(\mathsf{P}_{\mathbf{1}_n}^\perp) = \frac{1}{n!} \sum_{\Pi_n \in \mathscr{P}_n} \mathsf{Tr}(\Pi_n^\top A \Pi_n) = \frac{1}{n!} \sum_{\Pi_n \in \mathscr{P}_n} \mathsf{Tr}(A \Pi_n \Pi_n^\top) = \mathsf{Tr}(A).$$

□

下一个引理涉及归一化指数算子exp （即迹指数函数）的凸性和凹性。值得注意的是，第三个结论源自关于凸迹函数的基本 Lieb 定理 [29, 定理 6] 。我们引用其一个推论，见 [43, 推论 3。3] 和 [44, 推论 3.4。2]。前两个结果见 [43, 第 2] 节，最后一个是 [43, 引理 7。6]。

引理 A.3。 设 A 和 B 是 $\mathbb{S}^n \subset \mathbb{R}^{n \times n}$中的对称矩阵。

1. 映射 $A \mapsto \mathsf{Tr} \exp A$ 关于 $\mathbb{S}^n$是凸的，并且相对于由 $\mathbb{S}^n_{++}$诱导的偏序 $\prec$单调递增。

2. （戈登–汤普森不等式）$\mathsf{Tr}\exp$（$A + B$） $\leq \mathsf{Tr}$（$\exp A \exp B$）。 3.（Lieb 不等式）设 为固定的，为随机的，则 $\mathsf{Tr}\exp$（$A + B$） $\leq \mathsf{Tr}\exp$（$A + \log \mathbb{E}\exp B$）。其中，log 定义在正定锥 $\mathbb{S}^n_{++}$ 上，使得对于所有 $B \in \mathbb{S}^n$都有 $\log \exp B = B$。

4.（对称化）假设 A 是固定的， B 是满足 $\mathbb{E}B = 0$ 的随机变量。令 ϵ 为一个与 B 独立的 *Rademacher* 随机变量，则 $\mathbb{E}$迹 指数$(A + B) \leq \mathbb{E}$ 迹 指数$(A + 2\epsilon B)$。

对于标量随机变量 Z，下面的引理将矩母函数界与单侧尾事件界联系起来。关于第一个结果，请参见 [7,第2节。3] 关于第二个结果，请参见 [7, 第2节。4]。

引理 A.4。 对于任意随机变量 $Z \in \mathbb{R}$，若存在常数 $a, b > 0$ 使得：

1. $\mathbb{E}[\exp\{\lambda Z\}] \leq \exp\{\lambda^2 a^2/2\}$ 对所有 $\lambda \in \mathbb{R}$ 均成立，则有

$$\mathbb{P}\left\{Z \geq a\sqrt{2\log\frac{1}{\delta}}\right\} \leq \delta, \qquad \textit{for all } \delta \in (0,1);$$

2. $\mathbb{E}[\exp\{\lambda Z\}] \leq \exp\{\lambda^2 a^2/(2-2b|\lambda|)\}$ 对所有 $|\lambda| < 1/b$ 均成立，则有

$$\mathbb{P}\left\{Z \geq a\sqrt{2\log\frac{1}{\delta}} + b\log\frac{1}{\delta}\right\} \leq \delta, \qquad \textit{for all } \delta \in (0,1).$$

当 $Z \in \mathbb{R}^{n\times m}$ 时，Ahlswede 和 Winter [1] 提出使用迹指数函数 exp 以及矩阵矩母函数界来推导算子范数界。事实上，我们可以证明以下引理。我们指出，当 A.4 满足 $n = m = 1$ 时，它恢复了前述引理的双侧版本。

引理 A.5。 对于任意随机矩阵 $Z \in \mathbb{R}^{n\times m}$，若存在常数 $a, b > 0$ 使得：

1. $\mathbb{E}[\overline{\exp}\{\lambda Z\}] \leq \exp\{\lambda^2 a^2/2\}$ 对所有 $\lambda \in \mathbb{R}$ 均成立，则有

$$\mathbb{P}\left\{\|Z\| \geq a\sqrt{2\log\frac{n+m}{\delta}}\right\} \leq \delta, \qquad \textit{for all } \delta \in (0,1);$$

2. $\mathbb{E}[\overline{\exp}\{\lambda Z\}] \leq \exp\{\lambda^2 a^2/(2-2b|\lambda|)\}$ 对所有 $|\lambda| < 1/b$，均成立

$$\mathbb{P}\left\{\|Z\| \geq a\sqrt{2\log\frac{n+m}{\delta}} + b\log\frac{n+m}{\delta}\right\} \leq \delta, \qquad \textit{for all } \delta \in (0,1).$$

证明 我们给出该引理的一个简短证明，该证明实质上也证明了引理 A.4 。记 Z的奇异值为 σ_1、…、$\sigma_{n\wedge m} \geq 0$，容易验证扩张矩阵满足

$$\operatorname{spec}\left(\begin{bmatrix} 0 & A \\ A^\top & 0 \end{bmatrix}\right) \subset \{0, \sigma_1, \cdots, \sigma_{n\wedge m}, -\sigma_1, \cdots, -\sigma_{n\wedge m}\}.$$

因此 $\overline{\exp}\{\lambda Z\} \geq$ —— 当 $\lambda > 0$ 时，$\frac{1}{n+m}\exp(\lambda\|Z\|)$。因此，应用 Chernoff界，

$$\mathbb{P}\{\|Z\| \geq t\} \leq \inf_{\lambda>0} \exp\{-\lambda t + (n+m)\mathbb{E}[\overline{\exp}\{\lambda Z\}]\}.$$

在第一种情况下代入 ~~$\lambda =$~~ $\frac{t}{(n+m)a^2}$，并且

$$\lambda = \frac{2}{\sqrt{(n+m)^2a^4 + 2(n+m)a^2bt} + (n+m)a^2 + 2bt}$$

后者完成了证明。 □

第 B 节证明：2

B.1 定理 1的证明

设 $\mathrm{F}_k = \sigma(X_k)$，对于 $k = 0, \cdots, K$，由于

$$(\overline{X}_k)_{i_1\cdots i_k} = \frac{1}{N_{k+1}} \sum_{i_{k+1}=1}^{N_{k+1}} (\overline{X}_{k+1})_{i_1\cdots i_k i_{k+1}},$$

我们有滤过 $\sigma(X) = \mathrm{F}_0 \subset \cdots \subset \mathrm{F}_K = \sigma(X)$。定义矩阵 $\mathrm{A}_n \in \mathbb{R}^{n\times n}$，其第 i 行为 $(\mathbf{0}_{i-1}, \mathbf{1}, -\frac{1}{n-i}\cdot\mathbf{1}_{n-i})$，其中 $i \in [n-1]$，其第 n 行为 $\mathbf{0}_n$。作为一个特例，$\mathrm{A}_1 = 0$。我们还使用辅助矩阵 $\mathrm{B}_n \in \mathbb{R}^{n\times n}$，其第 i 行为 $(\mathbf{0}_{i-1}, \frac{1}{n+1-i}\cdot\mathbf{1}_{n+1-i})$，并满足性质 $\mathrm{A}_n^\dagger = I_n - \mathrm{B}_n^\top$ 和 $\mathrm{A}_n^\top(I_n - \mathrm{B}_n) = \mathrm{P}^\perp_{\mathbf{1}_n}$。两个矩阵都出现在 [3] 中。对于任意固定的 $U \in \mathbb{T}$ 及其相关序列 $U_k \in \mathbb{T}_k$，我们有以下引理。

引理B.1。对于任意 $k = 0, 1, \cdots K-1$ 以及上文定义的滤过 $\{\mathcal{F}_k\}_{k=0}^K$，令 $N_0 = 1$，则有

$$\mathbb{E}\left[\exp\left\{\lambda\left(\langle\overline{U}_{k+1}, \overline{X}_{k+1}\rangle_{\mathbb{T}_{k+1}} - N_{k+1}\langle\overline{U}_k, \overline{X}_k\rangle_{\mathbb{T}_k}\right)\right\} \mid \mathcal{F}_k\right] \le \exp\left\{\frac{\lambda^2}{2}\cdot\prod_{l=0}^{k} N_l \cdot \left\|\overline{U}_{k+1} \times_{k+1} \mathrm{A}_{N_{k+1}}\right\|^2_{\mathbb{T}_{k+1}}\right\}.$$

其证明留待附录 B.3。建立上述引理后，我们可以将其改进为下一个结果。

引理 B.2。对于任意 $k = 0, 1, \cdots K-1$ 以及上文定义的滤过 $\{\mathcal{F}_k\}_{k=0}^K$，令 $N_0 = 1$，并令$W \in \mathbb{T}$ 为权重张量。我们有

$$\begin{aligned}&\mathbb{E}\left[\exp\left\{\lambda\left(\langle\overline{W}_{k+1}, \overline{X}_{k+1}\rangle_{\mathbb{T}_{k+1}} - N_{k+1}\langle\overline{W}_k, \overline{X}_k\rangle_{\mathbb{T}_k}\right)\right\} \mid \mathcal{F}_k\right]\\ &\le \exp\left\{\frac{\lambda^2(1+\epsilon_{N_{k+1}})}{2}\cdot\prod_{l=0}^{k} N_l \cdot \left\|\overline{W}_{k+1} \times_{k+1} \mathrm{P}^\perp_{\mathbf{1}_{N_{k+1}}}\right\|^2_{\mathbb{T}_{k+1}}\right\}.\end{aligned}$$

证明 由 $\mathrm{A}_n\mathbf{1}_n = 0$，我们有 $\mathrm{A}_n\mathrm{P}^\perp_{\mathbf{1}_n} = \mathrm{A}_n$ 以及 $(\mathrm{A}_n^\top\mathrm{A}_n)^\dagger\mathrm{P}^\perp_{\mathbf{1}_n} = (\mathrm{A}_n^\top\mathrm{A}_n)^\dagger$。事实上，我们可以从 $I_n - \mathrm{B}_n = \mathrm{A}_n(\mathrm{A}_n^\top\mathrm{A}_n)^\dagger$ 明确推导出

$$(\mathrm{A}_n^\top\mathrm{A}_n)^\dagger = (I_n - \mathrm{B}_n^\top)(I_n - \mathrm{B}_n).$$

对于任意 $\Pi_{k+1} \in \mathscr{P}_{N_{k+1}}$，由于 $X \times_{k+1} \Pi_{k+1} \overset{\mathrm{d}}{=} X$，我们可以在引理 B.1中代入 $U_{k+1} = W_{k+1}\times_{k+1}(\mathrm{A}_{N_{k+1}}^\top\mathrm{A}_{N_{k+1}})^\dagger\Pi_{k+1}$并将 X 替换为 $X \times_{k+1} \Pi_{k+1}$。通过以下计算化简可得，

$$\begin{aligned}&\langle\overline{U}_{k+1}, \overline{(X\times_{k+1}\Pi_{k+1})}_{k+1}\rangle_{\mathbb{T}_{k+1}} - \langle\overline{U}_k, \overline{(X\times_{k+1}\Pi_{k+1})}_k\rangle_{\mathbb{T}_k} = \langle\overline{U}_{k+1}, \overline{(X\times_{k+1}\Pi_{k+1})}_{k+1} \times_{k+1} \mathrm{P}_{\mathbf{1}^\perp_{N_{k+1}}}\rangle_{\mathbb{T}_{k+1}}\\ &= \langle\overline{W}_{k+1}\times_{k+1}(\mathrm{A}_{N_{k+1}}^\top\mathrm{A}_{N_{k+1}})^\dagger\Pi_{k+1}, \overline{X}_{k+1}\times_{k+1}\Pi_{k+1}\times_{k+1}\mathrm{P}_{\mathbf{1}^\perp_{N_{k+1}}}\rangle_{\mathbb{T}_{k+1}}\\ &\overset{\text{(i)}}{=} \langle\overline{W}_{k+1}\times_{k+1}\Pi_{k+1}^\top\mathrm{P}_{\mathbf{1}^\perp_{N_{k+1}}}(\mathrm{A}_{N_{k+1}}^\top\mathrm{A}_{N_{k+1}})^\dagger\Pi_{k+1}, \overline{X}_{k+1}\rangle_{\mathbb{T}_{k+1}}\\ &\overset{\text{(ii)}}{=} \langle\overline{W}_{k+1}\times_{k+1}\Pi_{k+1}^\top(\mathrm{A}_{N_{k+1}}^\top\mathrm{A}_{N_{k+1}})^\dagger\Pi_{k+1}, \overline{X}_{k+1}\rangle_{\mathbb{T}_{k+1}},\\ &\left\|\overline{U}_{k+1}\times_{k+1}\mathrm{A}_{N_{k+1}}\right\|^2_{\mathbb{T}_{k+1}} = \left\|\overline{W}_{k+1}\times_{k+1}(\mathrm{A}_{N_{k+1}}^\top\mathrm{A}_{N_{k+1}})^\dagger\Pi_{k+1}\times_{k+1}\mathrm{A}_{N_{k+1}}\right\|^2_{\mathbb{T}_{k+1}}\\ &= \langle\overline{W}_{k+1}\times_{k+1}\mathrm{A}_{N_{k+1}}(\mathrm{A}_{N_{k+1}}^\top\mathrm{A}_{N_{k+1}})^\dagger\Pi_{k+1}, \overline{W}_{k+1}\times_{k+1}\mathrm{A}_{N_{k+1}}(\mathrm{A}_{N_{k+1}}^\top\mathrm{A}_{N_{k+1}})^\dagger\Pi_{k+1}\rangle_{\mathbb{T}_{k+1}}\\ &\overset{\text{(i)}}{=} \langle\overline{W}_{k+1}\times_{k+1}\Pi_{k+1}^\top(\mathrm{A}_{N_{k+1}}^\top\mathrm{A}_{N_{k+1}})^\dagger\Pi_{k+1}, \overline{W}_{k+1}\rangle_{\mathbb{T}_{k+1}},\end{aligned}$$

其中，在 (i) 中我们应用引理 A.1 ，在 (ii) 中我们使用 $\mathsf{A}_n\mathbf{1}_n = 0$，因此可以得到

$$\mathbb{E}\left[\exp\left\{\lambda\langle\overline{W}_{k+1}\times_{k+1}\Pi_{k+1}^{\top}(\mathsf{A}_{N_{k+1}}^{\top}\mathsf{A}_{N_{k+1}})^{\dagger}\Pi_{k+1},\overline{X}_{k+1}\rangle_{\mathbb{T}_{k+1}}\right\}\mid\mathcal{F}_k\right]$$
$$\leq\exp\left\{\frac{\lambda^2}{2}\cdot\prod_{l=0}^{k}N_l\cdot\langle\overline{W}_{k+1}\times_{k+1}\Pi_{k+1}^{\top}(\mathsf{A}_{N_{k+1}}^{\top}\mathsf{A}_{N_{k+1}})^{\dagger}\Pi_{k+1},\overline{W}_{k+1}\rangle_{\mathbb{T}_{k+1}}\right\}.$$

接下来我们应用詹森不等式，对 $\Pi_{k+1}\in\mathscr{P}_{N_{k+1}}$ 中的所有置换矩阵取平均，可得

$$\mathbb{E}\left[\exp\left\{\lambda\langle\overline{W}_{k+1}\times_{k+1}\frac{1}{N_{k+1}!}\sum_{\Pi_{k+1}\in\mathscr{P}_{N_{k+1}}}\Pi_{k+1}^{\top}(\mathsf{A}_{N_{k+1}}^{\top}\mathsf{A}_{N_{k+1}})^{\dagger}\Pi_{k+1},\overline{X}_{k+1}\rangle_{\mathbb{T}_{k+1}}\right\}\mid\mathcal{F}_k\right]$$
$$\leq\exp\left\{\frac{\lambda^2}{2}\cdot\prod_{l=0}^{k}N_l\cdot\langle\overline{W}_{k+1}\times_{k+1}\frac{1}{N_{k+1}!}\sum_{\Pi_{k+1}\in\mathscr{P}_{N_{k+1}}}\Pi_{k+1}^{\top}(\mathsf{A}_{N_{k+1}}^{\top}\mathsf{A}_{N_{k+1}})^{\dagger}\Pi_{k+1},\overline{W}_{k+1}\rangle_{\mathbb{T}_{k+1}}\right\}.$$

应用引理A.2 可进一步化简

$$\frac{1}{N_{k+1}!}\sum_{\Pi_{k+1}\in\mathscr{P}_{N_{k+1}}}\Pi_{k+1}^{\top}(\mathsf{A}_{N_{k+1}}^{\top}\mathsf{A}_{N_{k+1}})^{\dagger}\Pi_{k+1}=\frac{\mathsf{Tr}((\mathsf{A}_{N_{k+1}}^{\top}\mathsf{A}_{N_{k+1}})^{\dagger})}{N_{k+1}-1}\cdot\mathsf{P}_{\mathbf{1}_{N_{k+1}}}^{\perp}.$$

通过以下步骤即可完成证明

$$\mathsf{Tr}((\mathsf{A}_{N_{k+1}}^{\top}\mathsf{A}_{N_{k+1}})^{\dagger})=\mathsf{Tr}((I_{N_{k+1}}-\mathsf{B}_{N_{k+1}}^{\top})(I_{N_{k+1}}-\mathsf{B}_{N_{k+1}}))=\left\|I_{N_{k+1}}-\mathsf{B}_{N_{k+1}}^{\top}\right\|_F^2$$
$$=\sum_{i=1}^{N_{k+1}}\left\{\left(\frac{N_{k+1}-i}{N_{k+1}+1-i}\right)^2+(N_{k+1}-i)\cdot\frac{1}{(N_{k+1}+1-i)^2}\right\}=N_{k+1}-H_{N_{k+1}},$$

并将 λ 替换为 $\lambda(1+\epsilon_{N_{k+1}})$。 □

回到主证明，结论可通过应用 K 次塔式法则和引理B.2得到。事实上，通过代入

$$\langle W_k,X_k\rangle_{\mathbb{T}}=\langle\overline{W}_k\otimes\mathbf{1}_{N_{k+1}}\cdots\otimes\mathbf{1}_{N_K},\overline{X}_k\otimes\mathbf{1}_{N_{k+1}}\cdots\otimes\mathbf{1}_{N_K}\rangle_{\mathbb{T}}=\prod_{l=k+1}^{K}N_l\cdot\langle\overline{W}_k,\overline{X}_k\rangle_{\mathbb{T}_k},$$
$$\left\|\overline{W}_k\times_k\mathsf{P}_{\mathbf{1}_{N_k}}^{\perp}\right\|_{\mathbb{T}_k}^2=\frac{1}{\prod_{l=k+1}^{K}N_l}\left\|W_k\times_k\mathsf{P}_{\mathbf{1}_{N_k}}^{\perp}\right\|_{\mathbb{T}}^2,$$

并将 λ 替换为 $(\prod_{l=k+2}^{K}N_l)\cdot\lambda$ 到引理 B.2中，得到

$$\mathbb{E}\left[\exp\left\{\lambda\left(\prod_{l=k+2}^{K}N_l\cdot\langle\overline{W}_{k+1},\overline{X}_{k+1}\rangle_{\mathbb{T}_{k+1}}-\prod_{l=k+1}^{K}N_l\cdot\langle\overline{W}_k,\overline{X}_k\rangle_{\mathbb{T}_k}\right)\right\}\mid\mathcal{F}_k\right]$$
$$\leq\exp\left\{\frac{\lambda^2(1+\epsilon_{N_{k+1}})}{2}\cdot\left(\prod_{l=k+2}^{K}N_l\right)^2\cdot\prod_{l=0}^{k}N_l\cdot\frac{1}{\prod_{l=k+2}^{K}N_l}\cdot\left\|W_{k+1}\times_{k+1}\mathsf{P}_{\mathbf{1}_{N_{k+1}}}^{\perp}\right\|_{\mathbb{T}}^2\right\}$$
$$=\exp\left\{\frac{\lambda^2}{2}\cdot\prod_{l=0}^{K}N_l\cdot\frac{1+\epsilon_{N_{k+1}}}{N_{k+1}}\cdot\left\|W_{k+1}\times_{k+1}\mathsf{P}_{\mathbf{1}_{N_{k+1}}}^{\perp}\right\|_{\mathbb{T}}^2\right\}.$$

记为 $M_k:=\prod_{l=k+1}^{K}N_l\cdot W_k$，$X_{\overline{k^{\mathbb{T}}k}}$，然后我们通过 $N_0=1$ and 完成证明

$$\mathbb{E}\left[\exp\left\{\lambda\left(\langle W,X\rangle_{\mathbb{T}}-\prod_{l=1}^{K}N_l\cdot\overline{W}\,\overline{X}\right)\right\}\right]=\mathbb{E}\left[\exp\left\{\lambda\left(\langle\overline{W}_K,\overline{X}_K\rangle_{\mathbb{T}_K}-\prod_{l=1}^{K}N_l\cdot\langle\overline{W}_0,\overline{X}_0\rangle_{\mathbb{T}_0}\right)\right\}\right]$$

$$= \mathbb{E}\left[\mathbb{E}\left[\exp\left\{\lambda\left(M_K - M_{K-1}\right)\right\} \mid \mathcal{F}_{K-1}\right] \cdot \exp\left\{\lambda\left(M_{K-1} - M_0\right)\right\}\right]$$
$$\leq \exp\left\{\frac{\lambda^2}{2} \cdot \prod_{l=0}^{K} N_l \cdot \frac{1+\epsilon_{N_K}}{N_K} \cdot \left\| W_K \times_K \mathsf{P}^{\perp}_{\mathbf{1}_{N_K}} \right\|_{\mathbb{T}}^2 \right\} \cdot \mathbb{E}\left[\exp\left\{\lambda\left(M_{K-1} - M_0\right)\right\}\right]$$
$$\leq \cdots \leq \exp\left\{\frac{\lambda^2}{2} \cdot \prod_{l=0}^{K} N_l \cdot \sum_{k=1}^{K} \frac{1+\epsilon_{N_k}}{N_k} \cdot \left\| W_k \times_k \mathsf{P}^{\perp}_{\mathbf{1}_{N_k}} \right\|_{\mathbb{T}}^2 \right\}.$$

然后，我们通过确定 $\prod_{l=1}^{K} N_l \cdot W\ X = \mathbb{E}[\langle W,\ X\rangle_{\mathbb{T}} \mid X]$，并应用詹森不等式对随机变量 X取平均，得到矩母函数界。尾界则由应用引理 A.4 得出。

定理 2 的证明

与附录B.1中的证明类似，我们首先为此证明构造滤过 $\{\mathrm{F}_k\}_{k=0}^{K}$，其中当 $k \leq K-1$ 时，$\mathrm{F}_k = \sigma(X_k,\ \sigma^2_{X,\ k+1},\ \widetilde{\sigma}^2_{X,\ k+1})$，而 $\mathrm{F}_K = \sigma(X)$。由于 $\sigma^2_{X,\ k+1},\ \widetilde{\sigma}^2_{X,\ k+1} \in \sigma(X_{k+1}) \subset \mathrm{F}_{k+1}$，因此确有 $\mathrm{F}_0 \subset \cdots \subset \mathrm{F}_K$。于是，在给定 F_k 的条件下，我们得到伯恩斯坦型引理。

引理 B.3。对于任意 $k = 0$、1、$\cdots$、$K-1$ 以及上述定义的滤过$\{\mathcal{F}_k\}_{k=0}^{K}$，设 $N_0 = 1$，则对于 $|\lambda| \leq \frac{3}{2\|w_{k+1}\|_{\infty}(1+\epsilon_{N_{k+1}})}$，有

$$\mathbb{E}\left[\exp\left\{\lambda\left(\langle \overline{W}_{k+1}, \overline{X}_{k+1}\rangle_{\mathbb{T}_{k+1}} - N_{k+1}\langle \overline{W}_k, \overline{X}_k\rangle_{\mathbb{T}_k}\right)\right\} \mid \mathcal{F}_k\right]$$
$$\leq \exp\left\{\frac{\lambda^2(1+\epsilon_{N_{k+1}})}{2\left(1 - \frac{2|\lambda|}{3}\|\overline{w}_{k+1}\|_{\infty}(1+\epsilon_{N_{k+1}})\right)} \cdot \left\| \overline{W}_{k+1} \times_{k+1} \mathsf{P}^{\perp}_{\mathbf{1}_{N_{k+1}}} \right\|^2_{\mathbb{T}_{k+1}} \cdot \widetilde{\sigma}^2_{X,k+1}\right\}.$$

证明见附录B.4。应用前面的引理 K 次可得

$$\mathbb{E}\left[\exp\left\{\lambda\left(\langle W, X\rangle_{\mathbb{T}} - \prod_{l=1}^{K} N_l \cdot \overline{W}\ \overline{X}\right)\right\} \mid \mathcal{F}_0\right] \leq \exp\left\{\frac{\lambda^2}{2} \cdot \sum_{k=1}^{K} \frac{(1+\epsilon_{N_k}) \cdot \prod_{l=k+1}^{K} N_l \cdot \left\| W_k \times_k \mathsf{P}^{\perp}_{\mathbf{1}_{N_k}} \right\|^2_{\mathbb{T}} \cdot \widetilde{\sigma}^2_{X,k}}{1 - \frac{2|\lambda|}{3}\|\overline{w}_k\|_{\infty}(1+\epsilon_{N_k})}\right\}.$$

矩母函数界的证明是对两边关于 F_X 取条件期望，并使用$\widetilde{\sigma}^2_{X,1}, \cdots, \widetilde{\sigma}^2_{X,K} \in \mathrm{F}_X \subset \mathrm{F}_0$。进一步，由于

$$\frac{\lambda^2}{2} \cdot \sum_{k=1}^{K} \frac{(1+\epsilon_{N_k}) \cdot \prod_{l=k+1}^{K} N_l \cdot \left\| W_k \times_k \mathsf{P}^{\perp}_{\mathbf{1}_{N_k}} \right\|^2_{\mathbb{T}} \cdot \widetilde{\sigma}^2_{X,k}}{1 - \frac{2|\lambda|}{3}\|\overline{w}_k\|_{\infty}(1+\epsilon_{N_k})} \leq \frac{\lambda^2 \sum_{k=1}^{K}(1+\epsilon_{N_k}) \cdot \prod_{l=k+1}^{K} N_l \cdot \left\| W_k \times_k \mathsf{P}^{\perp}_{\mathbf{1}_{N_k}} \right\|^2_{\mathbb{T}} \cdot \widetilde{\sigma}^2_{X,k}}{2\left(1 - \frac{2}{3}\max_{k\in[K]}\|\overline{w}_k\|_{\infty}(1+\epsilon_{N_k}) \cdot |\lambda|\right)},$$

通过应用引理A.4即可得到尾界。

B.3 引理B.1的证明

我们首先构造一个更细的滤过 $\{\mathrm{F}_{k,i}\}_{i=0}^{N_{k+1}}$，它连接 $\mathrm{F}_{k,0} = \sigma(X_k) = \mathrm{F}_k$ 与 $\mathrm{F}_{k,N_{k+1}} = \sigma(X_{k+1}) = \mathrm{F}_{k+1}$；其中，对于 $i = 1, 2, \cdots, N_{k+1}$,

$$\mathcal{F}_{k,i} = \sigma\left(\mathcal{F}_{k,i-1}, \left\{(\overline{X}_{k+1})_{i_1\cdots i_k i}\right\}_{i_l\in[N_l], l\in[k]}\right) = \sigma\left(\overline{X}_k, \left\{(\overline{X}_{k+1})_{i_1\cdots i_k i_{k+1}}\right\}_{i_l\in[N_l], l\in[k], i_{k+1}\in[i]}\right).$$

对于任意固定的 $i_l \in [N_l], l \in [k]$，我们用简写 i_1^k 表示索引组 $i_1\cdots i_k$，并将 $(X_{k+1})_{i_1^k, \geq i} = \frac{1}{N_{k+1}+1-i}\sum_{i_{k+1}=i}^{N_{k+1}+1}(X_{k+1})_{i_1^k i_{k+1}}$记作 It is clear that $(X_{k+1})_{i_1^k,\geq i} = \frac{1}{N_{k+1}+1-i}(N_{k+1}(X_k)_{i_1^k} - \sum_{i_{k+1}=1}^{i-1}(X_{k+1})_{i_1^k i_{k+1}} \in \mathrm{F}_{k,i-1}$。由沿 $(k+1)$-模的模交换性，

$$\mathbb{E}\left[(\overline{X}_{k+1})_{i_1^k i} - (\overline{X}_{k+1})_{i_1^k,\geq i} \mid \mathcal{F}_{k,i-1}\right] = 0.$$

这使我们能够对固定的 [45, 命题 25] 应用标准的霍夫丁引理（例如参见 $V \in \mathbb{T}_{k+1}$,

$$\mathbb{E}\left[\exp\left\{\lambda \sum_{i_1^k \in [N_1]\times\cdots\times[N_k]} \overline{V}_{i_1^k i} \cdot \left((\overline{X}_{k+1})_{i_1^k i} - (\overline{X}_{k+1})_{i_1^k, \geq i}\right)\right\} \mid \mathcal{F}_{k,i-1}\right]$$
$$\leq \exp\left\{\frac{\lambda^2}{2} \cdot \left(\sum_{i_1^k \in [N_1]\times\cdots\times[N_k]} \left|\overline{V}_{i_1^k i}\right|\right)^2\right\} \leq \exp\left\{\frac{\lambda^2}{2} \cdot \prod_{l=0}^{k} N_l \cdot \sum_{i_1^k \in [N_1]\times\cdots\times[N_k]} \overline{V}_{i_1^k i}^2\right\},$$

其中，最后一个不等式由柯西-施瓦茨不等式得出。通过望远镜乘积并依次对$\mathbf{F}_{k,i-1}$取条件期望，便可得到

$$\mathbb{E}\left[\exp\left\{\lambda \sum_{i_1^k \in [N_1]\times\cdots\times[N_k]} \sum_{i=1}^{N_{k+1}} \overline{V}_{i_1^k i} \cdot \left((\overline{X}_{k+1})_{i_1^k i} - (\overline{X}_{k+1})_{i_1^k, \geq i}\right)\right\} \mid \mathcal{F}_k\right] \leq \exp\left\{\frac{\lambda^2}{2} \cdot \prod_{l=0}^{k} N_l \cdot \left\|\overline{V}\right\|_{\mathbb{T}_{k+1}}^2\right\}.$$

对于一个张量 $A \in \mathbb{T}_{k+1}$ 及其前 k 个指标 i_1^k，我们将$A_{i_1^k} := (A_{i_1^k 1}, \cdots, A_{i_1^k N_{k+1}})^\top \in \mathbb{R}^{N_{k+1}}$表示为，这使我们可以写出

$$\sum_{i_1^k \in [N_1]\times\cdots\times[N_k]} \sum_{i=1}^{N_{k+1}} \overline{V}_{i_1^k i} \cdot \left((\overline{X}_{k+1})_{i_1^k i} - (\overline{X}_{k+1})_{i_1^k, \geq i}\right) = \sum_{i_1^k \in [N_1]\times\cdots\times[N_k]} \overline{V}_{i_1^k}^\top (I_{N_{k+1}} - \mathsf{B}_{N_{k+1}})(\overline{X}_{k+1})_{i_1^k}$$
$$= \langle \overline{V}, \overline{X}_{k+1} \times_{k+1} (I_{N_{k+1}} - \mathsf{B}_{N_{k+1}})\rangle_{\mathbb{T}_{k+1}} = \langle \overline{V} \times_{k+1} (I_{N_{k+1}} - \mathsf{B}_{N_{k+1}})^\top, \overline{X}_{k+1}\rangle_{\mathbb{T}_{k+1}}. \tag{11}$$

最后，代入 $V = U_{k+1} \overline{\times}_{k+1} \overline{\mathsf{A}}^{N_{k+1}}$并利用 $U_{k+1} \times_{k+1} \overline{\mathsf{A}_{N_{k+1}}} \times_{k+1} (I_{N_{k+1}} - \mathsf{B}^{N_{k+1}})^\top = U_{k+1} \times_{k+1} (I_{\overline{N_{k+1}}} - \mathsf{B}^{N_{k+1}})^\top \mathsf{A}_{N_{k+1}} = U_{k+1} \times_{k+1} \overline{\mathsf{P}^\perp_{\mathbf{1}_{N_{k+1}}}}$，我们得到

$$\langle \overline{V} \times_{k+1} (I_{N_{k+1}} - \mathsf{B}_{N_{k+1}})^\top, \overline{X}_{k+1}\rangle_{\mathbb{T}_{k+1}} = \langle \overline{U}_{k+1} \times_{k+1} \mathsf{P}^\perp_{\mathbf{1}_{N_{k+1}}}, \overline{X}_{k+1}\rangle_{\mathbb{T}_{k+1}}$$
$$= \langle \overline{U}_{k+1}, \overline{X}_{k+1} \times_{k+1} \mathsf{P}^\perp_{\mathbf{1}_{N_{k+1}}}\rangle_{\mathbb{T}_{k+1}} = \langle \overline{U}_{k+1}, \overline{X}_{k+1} - \overline{X}_k \otimes \frac{1}{N_{k+1}} \mathbf{1}_{N_{k+1}}\rangle_{\mathbb{T}_{k+1}}$$
$$= \langle \overline{U}_{k+1}, \overline{X}_{k+1}\rangle_{\mathbb{T}_{k+1}} - N_{k+1} \langle \overline{U}_k \otimes \frac{1}{N_{k+1}} \mathbf{1}_{N_{k+1}}, \overline{X}_k \otimes \frac{1}{N_{k+1}} \mathbf{1}_{N_{k+1}}\rangle_{\mathbb{T}_{k+1}}$$
$$= \langle \overline{U}_{k+1}, \overline{X}_{k+1}\rangle_{\mathbb{T}_{k+1}} - N_{k+1} \langle \overline{U}_k, \overline{X}_k\rangle_{\mathbb{T}_k}.$$

由以上各式即可完成证明

$$\mathbb{E}\left[\exp\left\{\lambda\left(\langle \overline{U}_{k+1}, \overline{X}_{k+1}\rangle_{\mathbb{T}_{k+1}} - N_{k+1}\langle \overline{U}_k, \overline{X}_k\rangle_{\mathbb{T}_k}\right)\right\} \mid \mathcal{F}_k\right]$$
$$= \mathbb{E}\left[\exp\left\{\lambda \sum_{i_1^k \in [N_1]\times\cdots\times[N_k]} \sum_{i=1}^{N_{k+1}} \overline{V}_{i_1^k i} \cdot \left((\overline{X}_{k+1})_{i_1^k i} - (\overline{X}_{k+1})_{i_1^k, \geq i}\right)\right\} \mid \mathcal{F}_k\right] \leq \exp\left\{\frac{\lambda^2}{2} \cdot \prod_{l=0}^{k} N_l \cdot \left\|\overline{V}\right\|_{\mathbb{T}_{k+1}}^2\right\}$$
$$= \exp\left\{\frac{\lambda^2}{2} \cdot \prod_{l=0}^{k} N_l \cdot \left\|\overline{U}_{k+1} \times_{k+1} \mathsf{A}_{N_{k+1}}\right\|_{\mathbb{T}_{k+1}}^2\right\}$$

B.4 引理B.3 的证明

与附录中 B.3 的证明类似，对于 $i = 1, 2, \cdots, N_{k+1}$，我们令

$$\mathcal{F}_{k,i} = \sigma(\mathcal{F}_{k,i-1}, \{(\overline{X}_{k+1})_{i_1 \cdots i_k i}\}_{i_l \in [N_l], l \in [k]})$$
$$= \sigma(\overline{X}_k, \{\sigma_{X,l}^2, \widetilde{\sigma}_{X,l}^2\}_{l \in [k+1]}, \{(\overline{X}_{k+1})_{i_1 \cdots i_k i_{k+1}}\}_{i_l \in [N_l], l \in [k], i_{k+1} \in [i]}).$$

对于任意置换 $\pi_{k+1} \in \mathscr{P}_{N_{k+1}}$ 及其对应的置换矩阵 $\Pi_{k+1} \in \mathscr{S}_{N_{k+1}}$，我们定义

$$\overline{V}_{k+1}^{\pi_{k+1}} := \overline{W}_{k+1} \times_{k+1} \mathsf{P}^{\perp}_{\mathbf{1}_{N_{k+1}}} \times_{k+1} \Pi_{k+1},$$

并考虑随机变量 $Z_i^{\pi_{k+1}} := \sum_{i_1^k \in [N_1]\times\cdots\times[N_k]} V_{i_1^k i} \cdot ((X_{k+1})_{i_1^k i} - (\overline{X_{k+1}})_{i_1^k, \geq i})$。根据以下界

$$\left|Z_i^{\pi_{k+1}}\right| \leq \sum_{i_1^k \in [N_1]\times\cdots\times[N_k]} \left|\overline{V}_{i_1^k i}\right| = \sum_{i_1^k \in [N_1]\times\cdots\times[N_k]} \left|\left(\overline{W}_{k+1} \times \mathsf{P}^{\perp}_{\mathbf{1}_{N_{k+1}}}\right)_{i_1^k \pi_{k+1}(i)}\right| \leq \|\overline{w}_{k+1}\|_{\infty},$$

以及给定 $\mathrm{F}_{k,\ i-1}$ 时由模交换性可知，

$$\mathbb{E}\left[Z_i^{\pi_{k+1}} \mid \mathcal{F}_{k,i-1}\right] = 0,$$

$$\begin{aligned}
\mathsf{Var}\left(Z_i^{\pi_{k+1}} \mid \mathcal{F}_{k,i-1}\right) &\leq \left(\sum_{i_1^k \in [N_1]\times\cdots\times[N_k]} \overline{V}_{i_1^k i}^2\right) \cdot \frac{1}{N_{k+1}+1-i} \sum_{l=i}^{N_{k+1}} \left(\sum_{i_1^k \in [N_1]\times\cdots\times[N_k]} \left((\overline{X}_{k+1})_{i_1^k l} - (\overline{X}_{k+1})_{i_1^k, \geq i}\right)^2\right) \\
&\leq \left(\sum_{i_1^k \in [N_1]\times\cdots\times[N_k]} \overline{V}_{i_1^k i}^2\right) \cdot \frac{1}{N_{k+1}+1-i} \sum_{l=i}^{N_{k+1}} \underbrace{\left(\sum_{i_1^k \in [N_1]\times\cdots\times[N_k]} \left((\overline{X}_{k+1})_{i_1^k l} - (\overline{X}_k)_{i_1^k}\right)^2\right)}_{=:\widetilde{X}_{k+1,l}} \\
&= \underbrace{\left(\sum_{i_1^k \in [N_1]\times\cdots\times[N_k]} \left(\overline{W}_{k+1} \times \mathsf{P}^{\perp}_{\mathbf{1}_{N_{k+1}}}\right)^2_{i_1^k \pi_{k+1}(i)}\right)}_{=:\widetilde{W}_{k+1,\pi_{k+1}(i)}} \cdot \frac{1}{N_{k+1}+1-i} \sum_{l=i}^{N_{k+1}} \widetilde{X}_{k+1,l},
\end{aligned}$$

由伯恩斯坦引理（例如，参见 [45, 命题 2。10]）可得，对于 $|\lambda| \leq \frac{3}{2\|w_{k+1}\|_{\infty}}$，

$$\begin{aligned}
&\mathbb{E}\left[\exp\left\{\lambda \sum_{i_1^k \in [N_1]\times\cdots\times[N_k]} \overline{V}_{i_1^k i} \cdot \left((\overline{X}_{k+1})_{i_1^k i} - (\overline{X}_{k+1})_{i_1^k, \geq i}\right)\right\} \mid \mathcal{F}_{k,i-1}\right] \\
&\leq \exp\left\{\frac{\lambda^2}{2\left(1 - \frac{2|\lambda|}{3}\|\overline{w}_{k+1}\|_{\infty}\right)} \cdot \widetilde{W}_{k+1,\pi_{k+1}(i)} \cdot \frac{1}{N_{k+1}+1-i} \sum_{l=i}^{N_{k+1}} \widetilde{X}_{k+1,l}\right\}.
\end{aligned}$$

记 $\widetilde{W}_{k+1}$ 和 $\widetilde{X}_{k+1}$ 分别为$\mathbb{R}^{N_{k+1}}$中的向量 $(\widetilde{W}_{k+1,\ 1},\ \cdots,\ \widetilde{W}_{k+1,\ N_{k+1}})^{\top}$ 和 $(\widetilde{X}_{k+1,\ 1},\ \cdots,\ \widetilde{X}_{k+1,\ N_{k+1}})^{\top}$，则上述公式可简洁地写为

$$\begin{aligned}
&\mathbb{E}\left[\exp\left\{\lambda \sum_{i_1^k \in [N_1]\times\cdots\times[N_k]} \overline{V}_{i_1^k i} \cdot \left((\overline{X}_{k+1})_{i_1^k i} - (\overline{X}_{k+1})_{i_1^k, \geq i}\right)\right\} \mid \mathcal{F}_{k,i-1}\right] \\
&\leq \exp\left\{\frac{\lambda^2}{2\left(1 - \frac{2|\lambda|}{3}\|\overline{w}_{k+1}\|_{\infty}\right)} \cdot (\Pi_{k+1}\widetilde{W}_{k+1})^{\top} e_i e_i^{\top} \mathsf{B}_{N_{k+1}} \widetilde{X}_{k+1}\right\}.
\end{aligned}$$

通过望远镜乘积并识别式（11），可得

$$\mathbb{E}\left[\exp\left\{\lambda \langle \overline{W}_{k+1} \times_{k+1} \mathsf{P}^{\perp}_{\mathbf{1}_{N_{k+1}}} \times_{k+1} (I_{N_{k+1}} - \mathsf{B}_{N_{k+1}})^{\top} \Pi_{k+1}, \overline{X}_{k+1} \rangle_{\mathbb{T}_{k+1}}\right\} \mid \mathcal{F}_k\right]$$

$$\leq \exp\left\{\frac{\lambda^2}{2\left(1-\frac{2|\lambda|}{3}\left\|\overline{w}_{k+1}\right\|_\infty\right)}\cdot\sum_{i\in[N_{k+1}]}(\Pi_{k+1}\widetilde{W}_{k+1})^\top e_i e_i^\top \mathsf{B}_{N_{k+1}}\widetilde{X}_{k+1}\right\}$$
$$=\exp\left\{\frac{\lambda^2}{2\left(1-\frac{2|\lambda|}{3}\left\|\overline{w}_{k+1}\right\|_\infty\right)}\cdot\widetilde{W}_{k+1}^\top\Pi_{k+1}^\top\mathsf{B}_{N_{k+1}}\widetilde{X}_{k+1}\right\}.$$

我们可以再次应用根据詹森不等式，从$(X_{k+1},\widetilde{X}_{k+1})\,\mathsf{d}=(X_{k+1}\,\overline{\times}_{k+1}\,\Pi_{k+1},\Pi_{k+1}\widetilde{X}_{k+1})$得到

$$\mathbb{E}\left[\exp\left\{\lambda\langle\overline{W}_{k+1}\times_{k+1}\mathsf{P}^\perp_{\mathbf{1}_{N_{k+1}}}\times_{k+1}\frac{1}{N_{k+1}!}\sum_{\Pi_{k+1}\in\mathscr{P}_{N_{k+1}}}\Pi_{k+1}^\top(I_{N_{k+1}}-\mathsf{B}_{N_{k+1}})^\top\Pi_{k+1},\overline{X}_{k+1}\rangle_{\mathbb{T}_{k+1}}\right\}\mid\mathcal{F}_k\right]$$
$$=\exp\left\{\frac{\lambda^2}{2\left(1-\frac{2|\lambda|}{3}\left\|\overline{w}_{k+1}\right\|_\infty\right)}\cdot\widetilde{W}_{k+1}^\top\frac{1}{N_{k+1}!}\sum_{\Pi_{k+1}\in\mathscr{P}_{N_{k+1}}}\Pi_{k+1}^\top\mathsf{B}_{N_{k+1}}\Pi_{k+1}\widetilde{X}_{k+1}\right\}.$$

给定$(I_{N_{k+1}}-\mathsf{B}^{N_{k+1}})\mathbf{1}_{N_{k+1}}=0$，我们参照引理 A.2 并计算

$$\frac{1}{N_{k+1}!}\sum_{\Pi_{k+1}\in\mathscr{P}_{N_{k+1}}}\Pi_{k+1}^\top(I_{N_{k+1}}-\mathsf{B}_{N_{k+1}})^\top\Pi_{k+1}=\frac{N_{k+1}-\mathsf{Tr}(B_{N_{k+1}})}{N_{k+1}-1}\cdot\mathsf{P}^\perp_{\mathbf{1}_{N_{k+1}}}=\frac{1}{1+\epsilon_{N_{k+1}}}\cdot\mathsf{P}^\perp_{\mathbf{1}_{N_{k+1}}},$$
$$\frac{1}{N_{k+1}!}\sum_{\Pi_{k+1}\in\mathscr{P}_{N_{k+1}}}\Pi_{k+1}^\top\mathsf{B}_{N_{k+1}}\Pi_{k+1}=I-\frac{1}{1+\epsilon_{N_{k+1}}}\cdot\mathsf{P}^\perp_{\mathbf{1}_{N_{k+1}}}=\frac{1}{1+\epsilon_{N_{k+1}}}\cdot\mathsf{P}_{\mathbf{1}_{N_{k+1}}}+\frac{\epsilon_{N_{k+1}}}{1+\epsilon_{N_{k+1}}}\cdot I_{N_{k+1}}.$$

因此，右侧的上界为

$$\begin{aligned}
&\widetilde{W}_{k+1}^\top\frac{1}{N_{k+1}!}\sum_{\Pi_{k+1}\in\mathscr{P}_{N_{k+1}}}\Pi_{k+1}^\top\mathsf{B}_{N_{k+1}}\Pi_{k+1}\widetilde{X}_{k+1}\\
&\leq\frac{1}{1+\epsilon_{N_{k+1}}}\cdot\widetilde{W}_{k+1}^\top\mathbf{1}_{N_{k+1}}\cdot\frac{1}{N_{k+1}}\mathbf{1}_{N_{k+1}}^\top\widetilde{X}_{k+1}+\frac{\epsilon_{N_{k+1}}}{1+\epsilon_{N_{k+1}}}\widetilde{W}_{k+1}^\top\widetilde{X}_{k+1}\\
&\leq\frac{1}{1+\epsilon_{N_{k+1}}}\cdot\left\|\widetilde{W}_{k+1}\right\|_1\cdot\left(\frac{1}{N_{k+1}}\mathbf{1}_{N_{k+1}}^\top\widetilde{X}_{k+1}+\epsilon_{N_{k+1}}\left\|\widetilde{X}_{k+1}\right\|_\infty\right)\\
&\leq\frac{1}{1+\epsilon_{N_{k+1}}}\cdot\left\|\overline{W}_{k+1}\times_{k+1}\mathsf{P}^\perp_{\mathbf{1}_{N_{k+1}}}\right\|^2_{\mathbb{T}_{k+1}}\cdot\widetilde{\sigma}^2_{X,k+1},
\end{aligned}$$

其中，在最后一行中，我们使用了$\left\|\widetilde{W}_{k+1}\right\|_1=\left\|W_{k+1}\times_{k+1}\mathsf{P}^\perp_{\mathbf{1}_{N_{k+1}}}\right\|^2_{\mathbb{T}_{k+1}}$和柯西-施瓦茨不等式$\left\|\widetilde{X}_{k+1}\right\|_\infty\leq\prod_{l=0}^k N_l\cdot\left\|X_{k+1}\times_{k+1}\mathsf{P}^\perp_{\mathbf{1}_{N_{k+1}}}\right\|^2_\infty$。通过对 λ 重新参数化为$\lambda\cdot(1+\epsilon_{N_{k+1}})$，证明即告完成。

C 第 3 节的证明

C.1 定理 4 的证明

在此证明中，我们使用滤过$\{\mathsf{F}_k\}_{k=0}^N$，其中$\mathsf{F}_0=\sigma(X)$且$\overline{\mathsf{F}}_k=\sigma(\mathsf{F}_{k-1},X_k)$。若假设 1 成立，我们还有$\mathsf{F}_k=\sigma(X,\widetilde{X},\{X_i,\widetilde{X}_i\}_{i=1}^k)$。首先，我们引入一个关于 Rademacher 随机变量的对称化矩阵累积量生成函数（CGF）的引理。其证明见 [43,引理 4.3]。

引理 C.1 (Hoeffding 累积生成函数界)。 对于固定的对称矩阵 $A \in \mathbb{S}^n$，令 ϵ 为一个 *Rademacher* 随机变量。我们有

$$\log \mathbb{E} \exp(\epsilon A) \preceq \frac{A^2}{2}.$$

这里，log 是正定锥 $\mathbb{S}^n_{++}$ 上exp 在 $\mathbb{S}^n$上的函数逆。

步骤 1：通过鞅论证得到$\{X_k - X_{\geq k}\}_{k=1}^N$的矩母函数界。 对于任意 $U \in \mathbb{R}^{N\times p\times q}$, $k \in [N]$ $H_{k-1} \in \mathbb{S}^{p+r}$ $H_{k-1} \in$ $_{k-1}$ —— 并且若 $\mathcal{F}$ 是可适应的，令 $X_{\geq k} := \frac{1}{N+1-k}\sum_{i=k}^N X_i$。 根据$\mathbb{E}[X_k - X_{\geq k} \mid \mathcal{F}_{k-1}] = 0$和对称化引理（参见引理 A.3），

$$\begin{aligned}
\mathsf{M}_k(H_{k-1}) &:= \mathbb{E}\left[\mathsf{Tr}\exp\left\{H_{k-1} + \lambda\begin{bmatrix} 0 & U_k(X_k - X_{\geq k}) \\ (X_k - X_{\geq k})^\top U_k^\top & 0 \end{bmatrix}\right\} \mid \mathcal{F}_{k-1}\right] \\
&\leq \mathbb{E}\left[\mathsf{Tr}\exp\left\{H_{k-1} + \epsilon\cdot 2\lambda\begin{bmatrix} 0 & U_k(X_k - X_{\geq k}) \\ (X_k - X_{\geq k})^\top U_k^\top & 0 \end{bmatrix}\right\} \mid \mathcal{F}_{k-1}\right].
\end{aligned}$$

利用 Lieb 不等式和引理C.1，还可以得到

$$\begin{aligned}
&\mathbb{E}\left[\mathsf{Tr}\exp\left\{H_{k-1} + \lambda\begin{bmatrix} 0 & U_k(X_k - X_{\geq k}) \\ (X_k - X_{\geq k})^\top U_k^\top & 0 \end{bmatrix}\right\} \mid \mathcal{F}_{k-1}\right] \\
&\leq \mathbb{E}\left[\mathbb{E}\left[\mathsf{Tr}\exp\left\{H_{k-1} + \epsilon\cdot 2\lambda\begin{bmatrix} 0 & U_k(X_k - X_{\geq k}) \\ (X_k - X_{\geq k})^\top U_k^\top & 0 \end{bmatrix}\right\} \mid \mathcal{F}_N\right] \mid \mathcal{F}_{k-1}\right] \\
&\leq \mathbb{E}\left[\mathsf{Tr}\exp\left\{H_{k-1} + \log\mathbb{E}\left[\exp\left\{\epsilon\cdot 2\lambda\begin{bmatrix} 0 & U_k(X_k - X_{\geq k}) \\ (X_k - X_{\geq k})^\top U_k^\top & 0 \end{bmatrix}\right\} \mid \mathcal{F}_N\right]\right\} \mid \mathcal{F}_{k-1}\right] \\
&\leq \mathbb{E}\left[\mathsf{Tr}\exp\left\{H_{k-1} + 2\lambda^2\begin{bmatrix} U_k(X_k - X_{\geq k})(X_k - X_{\geq k})^\top U_k^\top & 0 \\ 0 & (X_k - X_{\geq k})^\top U_k^\top U_k (X_k - X_{\geq k}) \end{bmatrix}\right\} \mid \mathcal{F}_{k-1}\right],
\end{aligned}$$

其中，在最后一个不等式中，我们使用了引理 C.1 以及 Trexp的单调性。由$\|X_k\| \leq 1$，可得

$$\begin{aligned}
&\begin{bmatrix} U_k(X_k - X_{\geq k})(X_k - X_{\geq k})^\top U_k^\top & 0 \\ 0 & (X_k - X_{\geq k})^\top U_k^\top U_k (X_k - X_{\geq k}) \end{bmatrix} \\
&\preceq \begin{bmatrix} 4U_k U_k^\top & 0 \\ 0 & (X_k - X_{\geq k})^\top \cdot \left\|U_k^\top U_k\right\| I_q \cdot (X_k - X_{\geq k}) \end{bmatrix} \preceq \begin{bmatrix} 4U_k U_k^\top & 0 \\ 0 & 4\left\|U_k^\top U_k\right\| I_r \end{bmatrix} \preceq 4\|U_k\|^2 I_{p+r},
\end{aligned}$$

~并且还给定假设 1对 (U, X) 及其可交换对 (X, Z)成立，

$$\begin{aligned}
&\begin{bmatrix} U_k(X_k - X_{\geq k})(X_k - X_{\geq k})^\top U_k^\top & 0 \\ 0 & (X_k - X_{\geq k})^\top U_k^\top U_k (X_k - X_{\geq k}) \end{bmatrix} \\
&= \begin{bmatrix} U_k(X_k - X_{\geq k})(X_k - X_{\geq k})^\top U_k^\top & 0 \\ 0 & Z_k^\top(\widetilde{X}_k - \widetilde{X}_{\geq k})^\top(\widetilde{X}_k - \widetilde{X}_{\geq k}) Z_k \end{bmatrix} \preceq \begin{bmatrix} 4U_k U_k^\top & 0 \\ 0 & 4Z_k^\top Z_k \end{bmatrix}.
\end{aligned}$$

综合以上结果，并采用简写 $\mathbb{E}_k[\cdot] := \mathbb{E}[\cdot \mid \mathrm{F}_k]$，我们可以将其总结为以下引理。

引理 C.2. 对于任意 $k = 0, 1, \cdots, N-1$ 以及本节中定义的滤过 $\{\mathcal{F}_k\}_{k=0}^N$。我们有

$$\mathsf{M}_{k+1}(H_k) \leq \begin{cases} \mathbb{E}_k\left[\mathsf{Tr}\exp(H_k + 8\lambda^2\|U_{k+1}\|^2 I_{p+r})\right], & \textit{in general;} \\ \mathbb{E}_k\left[\mathsf{Tr}\exp\left(H_k + 8\lambda^2\begin{bmatrix} U_k U_k^\top & 0 \\ 0 & Z_k^\top Z_k \end{bmatrix}\right)\right], & \textit{If Assumption 1 holds.} \end{cases}$$

我们对某个 $H_0 \in \mathbb{S}^{p+r}$ 应用前述引理 N 次，并且

$$H_k = H_0 + \lambda \begin{bmatrix} 0 & \sum_{i=1}^{k} U_i (X_i - X_{\geq i}) \\ \sum_{i=1}^{k} (X_i - X_{\geq i})^\top U_i^\top & 0 \end{bmatrix}.$$

于是我们得到

$$\begin{aligned}
&\mathbb{E}\left[\mathsf{Tr}\exp\left\{H_0 + \lambda \begin{bmatrix} 0 & \sum_{k=1}^{N} U_k (X_k - X_{\geq k}) \\ \sum_{k=1}^{N} (X_k - X_{\geq k})^\top U_k^\top & 0 \end{bmatrix}\right\}\right] = \mathbb{E}\left[\mathsf{M}_N(H_{N-1})\right] \\
&\leq \mathbb{E}\left[\mathbb{E}_N\left[\mathsf{Tr}\exp(H_{N-1} + 8\lambda^2 \|U_N\|^2 I_{p+r})\right]\right] = \mathbb{E}\left[\mathbb{E}_{N-1}\left[\mathsf{Tr}\exp(H_{N-1} + 8\lambda^2 \|U_N\|^2 I_{p+r})\right]\right] \\
&= \mathbb{E}\left[\mathsf{M}_{N-1}\left(H_{N-2} + 8\lambda^2 \|U_N\|^2 I_{p+r}\right)\right] \leq \mathbb{E}\left[\mathsf{M}_{N-2}\left(H_{N-3} + 8\lambda^2 \|U_{N-1}\|^2 I_{p+r} + 8\lambda^2 \|U_N\|^2 I_{p+r}\right)\right] \\
&\leq \cdots \leq \mathbb{E}\left[\mathsf{Tr}\exp\left(H_0 + 8\lambda^2 \sum_{k=1}^{N} \|U_k\|^2 I_{p+r}\right)\right].
\end{aligned}$$

令 $H_0 = -8\lambda^2 \sum_{k=1}^{N} \|U_k\|^2 I_{p+r}$，得到

$$\frac{1}{p+r}\mathbb{E}\left[\mathsf{Tr}\exp\left\{\lambda \begin{bmatrix} 0 & \sum_{k=1}^{N} U_k (X_k - X_{\geq k}) \\ \sum_{k=1}^{N} (X_k - X_{\geq k})^\top U_k^\top & 0 \end{bmatrix} - 8\lambda^2 \sum_{k=1}^{N} \|U_k\|^2 I_{p+r}\right\}\right] \leq 1.$$

若假设 1成立，我们同样有

$$\frac{1}{p+r}\mathbb{E}\left[\mathsf{Tr}\exp\left\{\lambda \begin{bmatrix} 0 & \sum_{k=1}^{N} U_k (X_k - X_{\geq k}) \\ \sum_{k=1}^{N} (X_k - X_{\geq k})^\top U_k^\top & 0 \end{bmatrix} - 8\lambda^2 \begin{bmatrix} \sum_{k=1}^{N} U_k U_k^\top & 0 \\ 0 & \sum_{k=1}^{N} Z_k^\top Z_k \end{bmatrix}\right\}\right] \leq 1.$$

步骤 2：一般情形下 $\{X_k - X\}_{k=1}^{N}$ 的矩母函数界。 令 $u = (\|U_1\|, \cdots, \|U_N\|) \in \mathbb{R}^N, U^\top \in \mathbb{R}^{N \times q \times p}$，$Z^\top \in \mathbb{R}^{N \times q' \times r}$，类似地 $X^\top$，$\widetilde{X}^\top$，$W^\top$，$V^\top$。我们发现 $\sum_{k=1}^{N} U_k (X_k - X_{\geq k}) = U$，$X \times_1 (I - \mathsf{B}_N)$，且 $\sum_{k=1}^{N} \|U_k\|^2 = \|u\|^2$。令 $U = W \times_1 \Pi$ 和 $X = X \times_1 \Pi$，得到

$$\langle\!\langle U, X \times_1 (I - \mathsf{B}_N)\Pi \rangle\!\rangle = \langle\!\langle W \times_1 \Pi^\top (I - \mathsf{B}_N)^\top \Pi, X \rangle\!\rangle, \tag{12a}$$

$$\|u\|^2 = \sum_{k=1}^{N} \|W_k\|^2. \tag{12b}$$

利用 Trexp 的凸性并代入方程（12a）和（12b），然后对所有 $\Pi \in \mathscr{P}_N$ 取平均（计算细节见附录 B.1 和 B.2），从而得到

$$\frac{1}{p+r}\mathbb{E}\left[\mathsf{Tr}\exp\left\{\frac{\lambda}{1+\epsilon_N} \begin{bmatrix} 0 & \langle\!\langle W \times_1 \mathsf{P}_{\mathbf{1}_N}^{\perp}, X \rangle\!\rangle \\ \langle\!\langle W \times_1 \mathsf{P}_{\mathbf{1}_N}^{\perp}, X \rangle\!\rangle^\top & 0 \end{bmatrix} - 8\lambda^2 \cdot \sum_{k=1}^{N} \|W_k\|^2 I_{p+r}\right\}\right] \leq 1,$$

步骤 3：给定交换性条件时，$\{X_k - X\}_{k=1}^{N}$ 的矩母函数界。注意，$\sum_{k=1}^{N} U_k U_k^\top = U, U^\top$ 和 $\sum_{k=1}^{N} Z_k^\top Z_k = Z^\top, Z$。对于任意 $\Pi \in \mathscr{P}_N$，我们代入 $U = W \times_1 \mathsf{A}_N (\mathsf{A}_N^\top \mathsf{A}_N)^\dagger \Pi$，并将 X 替换为 $X \times_1 \Pi$。如果假设 1 对可交换对 (W, X) 和 $(\widetilde{X}, V)$ 成立，则容易验证 $Z = V \times_1 (\mathsf{A}_N^\top \mathsf{A}_N)^\dagger \Pi$。进行这些替换后，以下恒等式成立

$$U, X \times_1 (I - \mathsf{B}_N)\Pi = W \times_1 \Pi^\top (I - \mathsf{B}_N)^\top \mathsf{A}_N (\mathsf{A}_N^\top \mathsf{A}_N)^\dagger \Pi, X = W \times_1 \Pi^\top (\mathsf{A}_N^\top \mathsf{A}_N)^\dagger \Pi, X, \tag{13a}$$

$$\langle\!\langle U, U^\top \rangle\!\rangle = \langle\!\langle W \times_1 \Pi^\top (\mathsf{A}_N^\top \mathsf{A}_N)^\dagger \Pi, W^\top \rangle\!\rangle, \tag{13b}$$

$$\langle\!\langle Z^\top, Z \rangle\!\rangle = \langle\!\langle V^\top \times_1 \Pi^\top (\mathsf{A}_N^\top \mathsf{A}_N)^\dagger \Pi, V \rangle\!\rangle. \tag{13c}$$

再次利用 Trexp 的凸性，并代入方程（13a）至（13c，然后在所有 $\Pi \in \mathscr{P}_N$ 满足交换性条件时取平均，

$$\frac{1}{p+r}\mathbb{E}\left[\mathsf{Tr}\exp\left\{\frac{\lambda}{1+\epsilon_N} \begin{bmatrix} 0 & \langle\!\langle W \times_1 \mathsf{P}_{\mathbf{1}_N}^{\perp}, X \rangle\!\rangle \\ \langle\!\langle W \times_1 \mathsf{P}_{\mathbf{1}_N}^{\perp}, X \rangle\!\rangle^\top & 0 \end{bmatrix} - \frac{8\lambda^2}{1+\epsilon_N} \begin{bmatrix} \sum_{k=1}^{N} W_k W_k^\top & 0 \\ 0 & \sum_{k=1}^{N} V_k^\top V_k \end{bmatrix}\right\}\right] \leq 1.$$

步骤 4：完成证明。通过应用 Golden-Thompson 不等式，并将 λ 替换为 $\lambda\cdot(1+\epsilon_N)$ 即可完成证明。例如，利用最后一个不等式，并将迹指数算子中的矩阵记为$\Omega(\lambda)$，当假设 1成立时，我们有：

$$
\begin{aligned}
&\mathbb{E}\left[\overline{\exp}\left\{\lambda\left(\langle\!\langle W,X\rangle\!\rangle - N\cdot\overline{W}\,\overline{X}\right)\right\}\right]\\
&=\frac{1}{p+r}\mathbb{E}\left[\mathsf{Tr}\exp\left\{\lambda\begin{bmatrix}0 & \langle\!\langle W\times_1\mathsf{P}^{\perp}_{\mathbf{1}_N},X\rangle\!\rangle\\ \langle\!\langle W\times_1\mathsf{P}^{\perp}_{\mathbf{1}_N},X\rangle\!\rangle^{\top} & 0\end{bmatrix}\right\}\right]\\
&=\frac{1}{p+r}\mathbb{E}\left[\mathsf{Tr}\exp\left\{\Omega(\lambda(1+\epsilon_N))+8\lambda^2(1+\epsilon_N)\cdot\begin{bmatrix}\sum_{k=1}^N W_kW_k^{\top} & 0\\ 0 & \sum_{k=1}^N V_k^{\top}V_k\end{bmatrix}\right\}\right]\\
&\overset{\text{(i)}}{\le}\frac{1}{p+r}\mathbb{E}\left[\mathsf{Tr}\exp\left\{\Omega(\lambda(1+\epsilon_N))\right\}\cdot\exp\left\{8\lambda^2(1+\epsilon_N)\cdot\begin{bmatrix}\sum_{k=1}^N W_kW_k^{\top} & 0\\ 0 & \sum_{k=1}^N V_k^{\top}V_k\end{bmatrix}\right\}\right]\\
&\overset{\text{(ii)}}{\le}\frac{1}{p+r}\mathbb{E}\left[\mathsf{Tr}\exp\left\{\Omega(\lambda(1+\epsilon_N))\right\}\right]\cdot\exp\left\{8\lambda^2(1+\epsilon_N)\cdot\max\left\{\left\|\sum_{k=1}^N W_kW_k^{\top}\right\|,\left\|\sum_{k=1}^N V_k^{\top}V_k\right\|\right\}\right\}\\
&\le\exp\left\{8\lambda^2(1+\epsilon_N)\cdot\max\left\{\left\|\sum_{k=1}^N W_kW_k^{\top}\right\|,\left\|\sum_{k=1}^N V_k^{\top}V_k\right\|\right\}\right\},
\end{aligned}
$$

其中 (i) 中使用戈登–汤普森不等式，(ii) 中使用 Tr exp 的单调性。

C.2 定理 5

首先定义本证明中用于鞅论证的滤过。设

$$\mathcal{F}_0=\sigma\left(\overline{X},\overline{\widetilde{X}},\Sigma,\widetilde{\Sigma},\max_{k\in[N]}\left\|X_k-\overline{X}\right\|^2,\max_{k\in[N]}\left\|\widetilde{X}_k-\overline{\widetilde{X}}\right\|^2\right).$$

然后，我们可以对 $k\in[N]$ 归纳定义 $\mathcal{F}_k=\sigma(\mathcal{F}_{k-1},\ X_k)$。我们将用到 [44,引理 6.6.2]中的以下累积生成函数界。

引理 C.3 (Bernstein 累积生成函数界)。 对于满足 $\mathbb{E}A=0$ 和 $\|A\|\le L$ 的随机对称矩阵 $A\in\mathbb{S}^n$，对于 $|\lambda|<3/L$，有

$$\log\mathbb{E}\exp(\lambda A)\le\frac{\lambda^2}{2\left(1-\frac{|\lambda|L}{3}\right)}\cdot\mathbb{E}A^2.$$

下面给出定理 5 的一个稍强版本，这有助于推导组合矩阵求和不等式（参见推论 4.3）。

定理 5'。 定理 5成立，且 $w=$本质上确界$_X$(最大值$_{k\in[N]}\|W_1X_k\|,\cdots,$最大值$_{k\in[N]}\|W_NX_k\|)^{\top}$且$v=$本质上确界 $_{\widetilde{X}}$(最大值$_{k\in[N]}\left\|\widetilde{X}_kV_1\right\|,\cdots,$最大值$_{k\in[N]}\left\|\widetilde{X}_kV_N\right\|)^{\top}$。

显然，定理 5 可直接由定理 5'推出；证明的其余部分将专门用于定理 5'。

步骤 1：通过鞅论证得到 $\{X_k-X_{\ge k}\}_{k=1}^N$ 的矩母函数界。 对于任意满足假设且对应的可交换对为 $(U,\widetilde{X})$ 和 $(Z,\widetilde{X})$ 的，让 $u=(\max_{k\in[N]}\|U_1X_k\|,\cdots,\max_{k\in[N]}\|U_NX_k\|)^{\top}$ 和 $z=(\max_{k\in[N]}\left\|\widetilde{X}_kZ_1\right\|,\cdots,\max_{k\in[N]}\left\|\widetilde{X}_kZ_N\right\|)^{\top}$ 属于$\mathbb{R}^N$。定义以下 $\mathcal{F}_{k-1}$ 中的半正定矩阵

$$\Sigma_{\ge k}=\frac{1}{N+1-k}\sum_{i=k}^N(X_i-\overline{X})(X_i-\overline{X})^{\top},\qquad\widetilde{\Sigma}_{\ge k}=\frac{1}{N+1-k}\sum_{i=k}^N(\widetilde{X}_i-\overline{\widetilde{X}})^{\top}(\widetilde{X}_i-\overline{\widetilde{X}}).$$

显然， $\Sigma_{\geq 1} = \Sigma$ 和 $\widetilde{\Sigma}_{\geq 1} = \widetilde{\Sigma}$。考虑 $H_k = \sum_{i=1}^k D_i$，其中

$$D_k = \lambda \begin{bmatrix} 0 & U_k(X_k - X_{\geq k}) \\ (X_k - X_{\geq k})^\top U_k^\top & 0 \end{bmatrix} - \underbrace{\frac{\lambda^2}{2\left(1 - \frac{2|\lambda|}{3}\min\{\|u\|_\infty, \|z\|_\infty\}\right)}}_{=:c_\lambda} \cdot \underbrace{\begin{bmatrix} U_k \Sigma_{\geq k} U_k^\top & 0 \\ 0 & Z_k^\top \widetilde{\Sigma}_{\geq k} Z_k \end{bmatrix}}_{=:\Delta_k}.$$

采用记号 $\mathbb{E}_k[\cdot] := \mathbb{E}[\cdot \,|\, \mathcal{F}_k]$，我们注意到 $\mathbb{E}_{k-1}[U_k(X_k - X_{\geq k})] = 0$ 以及 $\|U_k(X_k - X_{\geq k})\| = \|(\widetilde{X}_k - \widetilde{X}_{\geq k}) Z_k\| \leq 2\min\{\|u\|_\infty, \|z\|_\infty\}$。接着，我们可以对C.3应用引理，取 $|\lambda| \leq \frac{3}{2\min\{\|u\|_\infty, \|z\|_\infty\}}$

$$\begin{aligned}
&\log \mathbb{E}_{k-1} \exp\left\{\lambda \begin{bmatrix} 0 & U_k(X_k - X_{\geq k}) \\ (X_k - X_{\geq k})^\top U_k^\top & 0 \end{bmatrix}\right\} \\
&\leq c_\lambda \begin{bmatrix} U_k \cdot \mathbb{E}_{k-1}\left[(X_k - X_{\geq k})(X_k - X_{\geq k})^\top\right] \cdot U_k^\top & 0 \\ 0 & Z_k^\top \cdot \mathbb{E}_{k-1}\left[(\widetilde{X}_k - \widetilde{X}_{\geq k})^\top (\widetilde{X}_k - \widetilde{X}_{\geq k})\right] \cdot Z_k \end{bmatrix} \\
&\leq c_\lambda \begin{bmatrix} U_k \cdot \mathbb{E}_{k-1}\left[(X_k - \overline{X})(X_k - \overline{X})^\top\right] \cdot U_k^\top & 0 \\ 0 & Z_k^\top \cdot \mathbb{E}_{k-1}\left[(\widetilde{X}_k - \overline{\widetilde{X}})^\top (\widetilde{X}_k - \overline{\widetilde{X}})\right] \cdot Z_k \end{bmatrix} \\
&= c_\lambda \Delta_k.
\end{aligned}$$

因此，根据 Lieb 不等式，对所有 $k \in [N]$，可得

$$\begin{aligned}
\mathbb{E}\left[\mathsf{Tr}\exp(H_k)\right] &\leq \mathbb{E}\mathbb{E}_{k-1}\left[\mathsf{Tr}\exp\left\{H_{k-1} - c_\lambda V_k + \log \mathbb{E}_{k-1}\exp\left\{\lambda \begin{bmatrix} 0 & U_k(X_k - X_{\geq k}) \\ (X_k - X_{\geq k})^\top U_k^\top & 0 \end{bmatrix}\right\}\right\}\right] \\
&\leq \mathbb{E}\mathbb{E}_{k-1}\left[\mathsf{Tr}\exp(H_{k-1})\right] = \mathbb{E}\left[\mathsf{Tr}\exp(H_{k-1})\right].
\end{aligned}$$

由于 $H_0 = 0$，于是有

$$\begin{aligned}
&\frac{1}{p+r}\mathbb{E}\left[\mathsf{Tr}\exp\left\{\lambda \begin{bmatrix} 0 & \sum_{k=1}^N U_k(X_k - X_{\geq k}) \\ \sum_{k=1}^N (X_k - X_{\geq k})^\top U_k^\top & 0 \end{bmatrix} - c_\lambda \begin{bmatrix} \sum_{k=1}^N U_k \Sigma_{\geq k} U_k^\top & 0 \\ 0 & \sum_{k=1}^N Z_k^\top \widetilde{\Sigma}_{\geq k} Z_k \end{bmatrix}\right\}\right] \\
&= \frac{1}{p+r}\mathbb{E}\left[\mathsf{Tr}\exp(H_N)\right] \leq \frac{1}{p+r}\mathbb{E}\left[\mathsf{Tr}\exp(H_0)\right] = 1
\end{aligned}$$

步骤 2：利用詹森不等式，给出 $\{X_k - \overline{X}\}_{k=1}^N$ 的矩母函数界。 对所有 $\Pi \in \mathscr{P}_N$，令 $U = W \times_1 \Pi$ 并且 $X = X \times_1 \Pi$。我们仍有公式（12a），且有 $\|u\|_\infty = \|w\|_\infty$，$\|z\|_\infty = \|v\|_\infty$。正如附录C.1中所证明的，对所有置换取平均可得，

$$\frac{1}{N!}\sum_{\Pi \in \mathscr{P}_N} \langle\!\langle W \times_1 \Pi, X \times_1 (I - \mathsf{B}_N)\Pi \rangle\!\rangle = \frac{1}{N!}\sum_{\pi \in \mathscr{S}_N}\sum_{k=1}^N W_{\pi(k)}(X_{\pi(k)} - X_{\geq k}^\pi) = \frac{1}{1+\epsilon_N}\langle\!\langle W \times_1 \mathsf{P}_{\mathbf{1}_N}^\perp, X \rangle\!\rangle,$$

其中 $X_{\geq k}^\pi := \frac{1}{N+1-k}\sum_{i=k}^N X_{\pi(i)}$。问题归结为控制

$$\begin{aligned}
\frac{1}{N!}\sum_{\pi \in \mathscr{S}_N}\sum_{k=1}^N W_{\pi(k)}\Sigma_{\geq k}^\pi W_{\pi(k)}^\top &:= \frac{1}{N!}\sum_{\pi \in \mathscr{S}_N}\sum_{k=1}^N W_{\pi(k)} \cdot \left\{\frac{1}{N+1-k}\sum_{i=k}^N (X_{\pi(i)} - \overline{X})(X_{\pi(i)} - \overline{X})^\top\right\} \cdot W_{\pi(k)}^\top, \\
\frac{1}{N!}\sum_{\pi \in \mathscr{S}_N}\sum_{k=1}^N V_{\pi(k)}^\top \widetilde{\Sigma}_{\geq k}^\pi V_{\pi(k)} &:= \frac{1}{N!}\sum_{\pi \in \mathscr{S}_N}\sum_{k=1}^N V_{\pi(k)}^\top \cdot \left\{\frac{1}{N+1-k}\sum_{i=k}^N (\widetilde{X}_{\pi(i)} - \overline{\widetilde{X}})^\top (\widetilde{X}_{\pi(i)} - \overline{\widetilde{X}})\right\} \cdot V_{\pi(k)}^\top.
\end{aligned}$$

下面的引理随之成立。

引理 C.4. 对于所有 $N \geq 2$,

$$\frac{1}{N!}\sum_{\pi\in\mathscr{S}_N}\sum_{k=1}^{N} W_{\pi(k)}\Sigma^{\pi}_{\geq k}W^{\top}_{\pi(k)} \preceq \left(\frac{1}{1+\epsilon_N}\|\Sigma\| + \frac{\epsilon_N}{1+\epsilon_N}\max_{l\in[N]}\left\|X_l-\overline{X}\right\|^2\right)\cdot\sum_{k=1}^{N} W_kW_k^{\top} = \frac{\widetilde{\sigma}_X^2}{1+\epsilon_N}\sum_{k=1}^{N} W_kW_k^{\top},$$

$$\frac{1}{N!}\sum_{\pi\in\mathscr{S}_N}\sum_{k=1}^{N} V^{\top}_{\pi(k)}\widetilde{\Sigma}^{\pi}_{\geq k}V_{\pi(k)} \preceq \left(\frac{1}{1+\epsilon_N}\|\widetilde{\Sigma}\| + \frac{\epsilon_N}{1+\epsilon_N}\max_{l\in[N]}\left\|\widetilde{X}_l-\overline{\widetilde{X}}\right\|^2\right)\cdot\sum_{k=1}^{N} V_k^{\top}V_k = \frac{\widetilde{\sigma}_{\widetilde{X}}^2}{1+\epsilon_N}\sum_{k=1}^{N} V_k^{\top}V_k.$$

根据詹森不等式和上述引理，我们有

$$\frac{1}{p+r}\mathbb{E}\left[\mathsf{Tr}\exp\left\{\frac{\lambda}{1+\epsilon_N}\begin{bmatrix}0 & \langle\!\langle W\times_1\mathsf{P}^{\perp}_{\mathbf{1}_N},X\rangle\!\rangle\\ \langle\!\langle W\times_1\mathsf{P}^{\perp}_{\mathbf{1}_N},X\rangle\!\rangle^{\top} & 0\end{bmatrix} - \frac{c_\lambda}{1+\epsilon_N}\begin{bmatrix}\widetilde{\sigma}_X^2\sum_{k=1}^{N}W_kW_k^{\top} & 0\\ 0 & \widetilde{\sigma}_{\widetilde{X}}^2\sum_{k=1}^{N}V_k^{\top}V_k\end{bmatrix}\right\}\right] \leq 1.$$

步骤 3：完成证明。通过应用戈登–汤普森不等式，将λ 替换为 $\lambda\cdot(1+\epsilon_N)$，并利用以下不等式即可完成证明

$$\begin{bmatrix}\sum_{k=1}^{N}W_kW_k^{\top} & 0\\ 0 & \sum_{k=1}^{N}V_k^{\top}V_k\end{bmatrix} \leq \max\left\{\widetilde{\sigma}_X^2\left\|\sum_{k=1}^{N}W_kW_k^{\top}\right\|, \widetilde{\sigma}_{\widetilde{X}}^2\left\|\sum_{k=1}^{N}V_k^{\top}V_k\right\|\right\} I_{p+r}.$$

C.3 引理C.4 的证明

我们详细证明

$$\frac{1}{N!}\sum_{\pi\in\mathscr{S}_N}\sum_{k=1}^{N} W_{\pi(k)}\Sigma^{\pi}_{\geq k}W^{\top}_{\pi(k)} \preceq \left(\frac{1}{1+\epsilon_N}\|\Sigma\| + \frac{\epsilon_N}{1+\epsilon_N}\max_{l\in[N]}\left\|X_l-\overline{X}\right\|^2\right)\cdot\sum_{k=1}^{N} W_kW_k^{\top}$$

第二个不等式可由相同的论证得到。设 $\mathsf{Uni}(\mathscr{S}_N)$ 为 $\mathscr{S}_N$ 上的均匀分布，我们可以写成

$$\frac{1}{N!}\sum_{\pi\in\mathscr{S}_N}\sum_{k=1}^{N} W_{\pi(k)}\Sigma^{\pi}_{\geq k}W^{\top}_{\pi(k)} = \mathbb{E}_{\pi\sim\mathsf{Uni}(\mathscr{S}_N)}\left[\sum_{k=1}^{N} W_{\pi(k)}\cdot\left\{\frac{1}{N+1-k}\sum_{i=k}^{N}(X_{\pi(i)}-\overline{X})(X_{\pi(i)}-\overline{X})^{\top}\right\}\cdot W^{\top}_{\pi(k)}\right]$$

$$= \sum_{k=1}^{N}\frac{1}{N+1-k}\mathbb{E}_{\pi\sim\mathsf{Uni}(\mathscr{S}_N)}\left[W_{\pi(k)}(X_{\pi(k)}-\overline{X})(X_{\pi(k)}-\overline{X})^{\top}W^{\top}_{\pi(k)}\right]$$

$$+ \sum_{k=1}^{N}\sum_{i=k+1}^{N}\frac{1}{N+1-k}\mathbb{E}_{\pi\sim\mathsf{Uni}(\mathscr{S}_N)}\left[W_{\pi(k)}(X_{\pi(i)}-\overline{X})(X_{\pi(i)}-\overline{X})^{\top}W^{\top}_{\pi(k)}\right].$$

由于对任意固定的 $k\neq i$，$\pi(k)$ 在 $[N]$ 上服从均匀分布，而 $(\pi(k)\ \pi(i))$ 在 $\{(j,\ l):j\neq l,\ j,\ l\in[N]\}$ 上服从均匀分布，因此我们可以继续计算

$$\frac{1}{N!}\sum_{\pi\in\mathscr{S}_N}\sum_{k=1}^{N} W_{\pi(k)}\Sigma^{\pi}_{\geq k}W^{\top}_{\pi(k)}$$

$$= \sum_{k=1}^{N}\frac{1}{N+1-k}\cdot\left\{\frac{1}{N}\sum_{j=1}^{N}W_j(X_j-\overline{X})(X_j-\overline{X})^{\top}W_j^{\top} + \frac{N-k}{N(N-1)}\sum_{j,l\in[N],j\neq l}W_j(X_l-\overline{X})(X_l-\overline{X})^{\top}W_j^{\top}\right\}$$

$$= \sum_{k=1}^{N}\frac{1}{N+1-k}\cdot\left\{\frac{k-1}{N(N-1)}\sum_{j=1}^{N}W_j(X_j-\overline{X})(X_j-\overline{X})^{\top}W_j^{\top} + \frac{N-k}{N(N-1)}\sum_{j=1}^{N}\sum_{l=1}^{N}W_j(X_l-\overline{X})(X_l-\overline{X})^{\top}W_j^{\top}\right\}.$$

利用以下不等式

$$\sum_{j=1}^{N} W_j(X_j-\overline{X})(X_j-\overline{X})^\top W_j^\top \preceq \max_{l\in[N]}\left\|X_l-\overline{X}\right\|^2\cdot\sum_{j=1}^{N} W_jW_j^\top,$$
$$\sum_{j=1}^{N}\sum_{l=1}^{N} W_j(X_l-\overline{X})(X_l-\overline{X})^\top W_j^\top = N\sum_{j=1}^{N} W_j\Sigma W_j^\top \preceq N\|\Sigma\|\cdot\sum_{j=1}^{N} W_jW_j^\top,$$

我们可以通过以下方式完成证明

$$\begin{aligned}
&\frac{1}{N!}\sum_{\pi\in\mathscr{S}_N}\sum_{k=1}^{N} W_{\pi(k)}\Sigma_{\geq k}^{\pi}W_{\pi(k)}^\top\\
&\preceq\left\{\frac{1}{N-1}\sum_{k=1}^{N}\frac{N-k}{N+1-k}\|\Sigma\|+\frac{1}{N(N-1)}\sum_{k=1}^{N}\frac{k-1}{N+1-k}\max_{l\in[N]}\left\|X_l-\overline{X}\right\|^2\right\}\cdot\sum_{j=1}^{N} W_jW_j^\top\\
&=\left\{\frac{N-H_N}{N-1}\|\Sigma\|+\frac{N(H_N-1)}{N(N-1)}\max_{l\in[N]}\left\|X_l-\overline{X}\right\|^2\right\}\cdot\sum_{j=1}^{N} W_jW_j^\top\\
&=\left\{\frac{1}{1+\epsilon_N}\|\Sigma\|+\frac{\epsilon_N}{1+\epsilon_N}\max_{l\in[N]}\left\|X_l-\overline{X}\right\|^2\right\}\cdot\sum_{j=1}^{N} W_jW_j^\top.
\end{aligned}$$

D 理论推论与应用的证明

D.1 与独立同分布情形相关的证明

首先，说明 $\sum_{l=1}^{N_K}\mathscr{W}_l, \mathbb{E}_{\mathscr{X}\sim P}[\mathscr{X}]\rangle_{\mathbb{T}_{K-1}} = W, \mu_P\mathbf{1}_{\mathbb{T}}\rangle_{\mathbb{T}}$ 是有益的。事实上，利用 W 沿模-K 平衡这一点，我们验证

$$\begin{aligned}
\sum_{l=1}^{N_K}\langle\mathscr{W}_l,\mathbb{E}_{\mathscr{X}\sim P}[\mathscr{X}]\rangle_{\mathbb{T}_{K-1}} &= \left\langle\sum_{l=1}^{N_K}\mathscr{W}_l,\mathbb{E}_{\mathscr{X}\sim P}[\mathscr{X}]\right\rangle_{\mathbb{T}_{K-1}} = N_K\left\langle\overline{W}_{K-1},\mathbb{E}_{\mathscr{X}\sim P}[\mathscr{X}]\right\rangle_{\mathbb{T}_{K-1}}\\
&=\langle\overline{W}\mathbf{1}_{\mathbb{T}},\mathbb{E}_{\mathscr{X}\sim P}[\mathscr{X}]\rangle_{\mathbb{T}} = \langle W\times_1\mathsf{P}_{\mathbf{1}_{N_1}}\times_2\cdots\times_K\mathsf{P}_{\mathbf{1}_{N_K}},\mathbb{E}_{\mathscr{X}\sim P}[\mathscr{X}]\rangle_{\mathbb{T}}\\
&=\langle W,\mathbb{E}_{\mathscr{X}\sim P}[\mathscr{X}]\times_1\mathsf{P}_{\mathbf{1}_{N_1}}\times_2\cdots\times_K\mathsf{P}_{\mathbf{1}_{N_K}}\rangle_{\mathbb{T}} = \langle W,\mu_P\mathbf{1}_{\mathbb{T}}\rangle_{\mathbb{T}}.
\end{aligned}$$

回到主要证明。对于任意 $\widetilde{N}_K\geq N_K$，令 $\widetilde{W}\in\mathbb{R}^{N_1\times N_2\cdots\times N_{K-1}\times\widetilde{N}_K}=:\widetilde{\mathbb{T}}$ 为

$$\widetilde{W}_{i_1\cdots i_{K-1}i_K}=\begin{cases}W_{i_1\cdots i_{K-1}i_K}, & i_K\leq N_K,\\ 0, & i_K>N_K,\end{cases}$$

$\widetilde{\sim}$ $\widetilde{\sim}$ $\widetilde{\sim}$，并令随机数据 $X\in\mathbb{T}$ 沿模-K 无限可交换，使得当 $i_K\in[N_K]$ 时，$X_{i_1\cdots i_{K-1}i_K}=\widetilde{\sim}X_{i_1\cdots i_{K-1}i_K}$。我们可以验证 W 沿模-K 平衡，并且

$$\lambda\left(\langle\widetilde{W},\widetilde{X}\rangle_{\widetilde{\mathbb{T}}}-\prod_{l=1}^{K-1}N_lN\cdot\overline{\widetilde{W}}\overline{\widetilde{X}}\right)=\lambda\left(\langle W,X\rangle_{\mathbb{T}}-\prod_{l=1}^{K}N_l\cdot\overline{W}\overline{\widetilde{X}}\right)\overset{a.s.}{\to}\lambda\left(\langle W,X\rangle_{\mathbb{T}}-\prod_{l=1}^{K}N_l\cdot\overline{W}\mu_P\right).$$

取 $\widetilde{N}_K\to\infty$，令 $\widetilde{W}_k\times_k\mathsf{P}^{\perp}_{\mathbf{1}_{N_k}}=0$ 和 $\left\|\widetilde{W}_K\times_K\mathsf{P}^{\perp}_{\mathbf{1}_{\widetilde{N}_K}}\right\|_{\widetilde{\mathbb{T}}}^2\to\|W\|_{\mathbb{T}}^2$，我们将 $\epsilon_{\widetilde{N}_K}\to0$ 代入定理 1，即可得到独立 X 沿模-K的 Hoeffding 界。对于 Bernstein 界，还需利用 $\|w_K\|_\infty\to\|w\|_\infty$ 以及

$$\frac{1}{\widetilde{N}_K}\left\|\widetilde{X}\times_K\mathsf{P}^{\perp}_{\mathbf{1}_{\widetilde{N}_K}}\right\|_{\mathbb{T}}^2\to\sigma_P^2$$

即可完成证明。矩阵值结果的证明完全相同。重复性的细节从略。

D.2 推论4.3 的证明

由于假设1 成立，我们可以应用定理 5 。利用矩阵克罗内克积的性质（见 [22, 第 4]章），我们验证得

$$\|X_k\| = \left\|e_{\pi(k)} \otimes I_n\right\| = \left\|e_{\pi(k)}\right\| \cdot \|I_n\| = 1,$$

以及 $\left\|\widetilde{X}_k\right\| = 1$。现在我们根据以下内容计算方差参数

$$\begin{aligned}
\Sigma &= \frac{1}{N}\sum_{k=1}^{N}(X_k - \overline{X})(X_k - \overline{X})^\top = \frac{1}{N}\sum_{k=1}^{N}(e_k \otimes I_n)(e_k \otimes I_n)^\top - \left(\frac{1}{N}\sum_{k=1}^{N} e_k \otimes I_n\right)\left(\frac{1}{N}\sum_{k=1}^{N} e_k \otimes I_n\right)^\top \\
&= \frac{1}{N}\sum_{k=1}^{N} e_k e_k^\top \otimes I_n - \left(\frac{1}{N}\mathbf{1}_N \otimes I_n\right)\left(\frac{1}{N}\mathbf{1}_N \otimes I_n\right)^\top = \frac{1}{N}\cdot \mathsf{P}^{\perp}_{\mathbf{1}_N} \otimes I_n, \\
\widetilde{\Sigma} &= \frac{1}{N}\sum_{k=1}^{N}(\widetilde{X}_k - \overline{\widetilde{X}})^\top(\widetilde{X}_k - \overline{\widetilde{X}}) = \frac{1}{N}\sum_{k=1}^{N}(e_k \otimes I_m)(e_k \otimes I_m)^\top - \left(\frac{1}{N}\sum_{k=1}^{N} e_k \otimes I_m\right)\left(\frac{1}{N}\sum_{k=1}^{N} e_k \otimes I_m\right)^\top \\
&= \frac{1}{N}\sum_{k=1}^{N} e_k e_k^\top \otimes I_m - \left(\frac{1}{N}\mathbf{1}_N \otimes I_m\right)\left(\frac{1}{N}\mathbf{1}_N \otimes I_m\right)^\top = \frac{1}{N}\cdot \mathsf{P}^{\perp}_{\mathbf{1}_N} \otimes I_m,
\end{aligned}$$

以及 $\max_{k\in[N]}\left\|X_k - \overline{X}\right\|^2 = \max_{k\in[N]}\left\|\widetilde{X}_k - \overline{\widetilde{X}}\right\|^2 = (1-1/N)^2 + (N-1)\cdot 1/N^2 = 1 - 1/N$，由此可得

$$\widetilde{\sigma}_X^2 = \widetilde{\sigma}_{\widetilde{X}}^2 = \frac{1}{N} + \epsilon_N \cdot \left(1 - \frac{1}{N}\right).$$

另一方面，由于

$$\sum_{k=1}^{N} W_k W_k^\top = \sum_{i,j\in[N]} A_{i,j}A_{i,j}^\top, \qquad \text{and} \qquad \sum_{k=1}^{N} V_k^\top V_k = \sum_{i,j\in[N]} A_{i,j}^\top A_{i,j}.$$

我们将调用略作修改的定理 5' ，其中 $W\ X = 0$ 以及前面计算出的参数。接下来需要界定 $\|w\|_\infty$ 和 $\|v\|_\infty$ ，它们在定理 5'中定义。由于以下内容成立，这些量很容易得到

$$\max_{k\in[N]}\|W_i X_k\| = \max_{k\in[N]}\left\|\widetilde{X}_k V_i\right\| = \max_{j\in[N]}\|A_{i,j}\| \le R.$$

最后，我们应用引理A.5 来得到如下尾界。

D.3 推论 5.1 的证明

为了应用定理 2，我们将使用推论 2.1 中的界，并计算参数$\widetilde{\sigma}^2_{X,\ k}$。由

$$\left(\overline{X}_k \times_k \mathsf{P}^{\perp}_{\mathbf{1}_{N_k}}\right)_{i_1\cdots i_k} = \mathbf{1}_{i_l\in I_l, l\in[k-1]}\cdot\left\{\mathbf{1}_{i_k\in I_k}\frac{n_{k+1}\cdots n_K}{N_{k+1}\cdots N_K}\left(1 - \frac{n_k}{N_k}\right) - \mathbf{1}_{i_k\notin I_k}\frac{n_k\cdots n_K}{N_k\cdots N_K}\right\},$$

我们有

$$\left\|\overline{X}_k \times_k \mathsf{P}^{\perp}_{\mathbf{1}_{N_k}}\right\|^2_{\mathbb{T}_k} = \prod_{l=1}^{k-1} n_l \cdot \left\{\left(\frac{n_{k+1}\cdots n_K}{N_{k+1}\cdots N_K}\right)^2\left(1 - \frac{n_k}{N_k}\right)^2 \cdot n_k + \left(\frac{n_k\cdots n_K}{N_k\cdots N_K}\right)^2 \cdot (N_k - n_k)\right\}$$

$$= \prod_{l=1}^{k} n_l \cdot \left(\frac{n_{k+1}\cdots n_K}{N_{k+1}\cdots N_K}\right)^2 \left(1-\frac{n_k}{N_k}\right)$$

$$\left\|\overline{X}_k \times_k \mathsf{P}^{\perp}_{\mathbf{1}_{N_k}}\right\|_\infty^2 = \mathbb{1}_{n_k<N_k} \cdot \left(\max\left\{1-\frac{n_k}{N_k}, \frac{n_k}{N_k}\right\} \cdot \frac{n_{k+1}\cdots n_K}{N_{k+1}\cdots N_K}\right)^2.$$

因此

$$\begin{aligned}
\widetilde{\sigma}^2_{X,k} &= \frac{1}{N_k} \cdot \left\|\overline{X}_k \times_k \mathsf{P}^{\perp}_{\mathbf{1}_{N_k}}\right\|_{\mathbb{T}_k}^2 + \epsilon_{N_k} \cdot \prod_{l=0}^{k-1} N_l \cdot \left\|\overline{X}_k \times_k \mathsf{P}^{\perp}_{\mathbf{1}_{N_k}}\right\|_\infty^2 \\
&= \prod_{l=1}^{k-1} n_l \cdot \left(\frac{n_{k+1}\cdots n_K}{N_{k+1}\cdots N_K}\right)^2 \left(1-\frac{n_k}{N_k}\right) \cdot \frac{n_k}{N_k} + \epsilon_{N_k} \cdot \prod_{l=1}^{k-1} N_l \cdot \mathbb{1}_{n_k<N_k} \cdot \left(\max\left\{1-\frac{n_k}{N_k}, \frac{n_k}{N_k}\right\} \cdot \frac{n_{k+1}\cdots n_K}{N_{k+1}\cdots N_K}\right)^2 \\
&= \left\{\frac{\prod_{l=1}^{k} n_l}{\prod_{l=1}^{k} N_l} \cdot \left(1-\frac{n_k}{N_k}\right) + \epsilon_{N_k} \cdot \mathbb{1}_{n_k<N_k} \max\left\{1-\frac{n_k}{N_k}, \frac{n_k}{N_k}\right\}^2\right\} \cdot \prod_{l=1}^{k-1} N_l \cdot \left(\frac{n_{k+1}\cdots n_K}{N_{k+1}\cdots N_K}\right)^2 \\
&\le \left\{\frac{\prod_{l=1}^{k} n_l}{\prod_{l=1}^{k} N_l} \cdot \left(1-\frac{n_k}{N_k}\right) + \epsilon_{N_k} \cdot \mathbb{1}_{n_k<N_k}\right\} \cdot \prod_{l=1}^{k-1} N_l \cdot \left(\frac{n_{k+1}\cdots n_K}{N_{k+1}\cdots N_K}\right)^2,
\end{aligned}$$

使用 $\|w_{\overline{k}}\|_\infty \le B_k$，对于 $\widehat{\mu}-\mu = \frac{\prod_{l=1}^{K} N_l}{\prod_{l=1}^{K} n_l} \cdot \langle W,\ X\rangle_{\mathbb{T}} - \prod_{l=1}^{K} N_l \cdot \overline{W}\ \overline{X}$），我们可以调用定理2（特别是，公式（2b）和推论2.1），从而

$$\mathbb{P}\left\{\widehat{\mu}-\mu \ge \sigma\sqrt{2\log\frac{1}{\delta} \cdot \max_{k\in[K]} \Delta_k} + \frac{2\max_{k\in[K]} B_k(1+\epsilon_{N_k})}{3\alpha_K}\log\frac{1}{\delta}\right\} \le \delta,$$

其中 $\sigma^2 = \prod_{l=1}^{K} N_l \cdot (\|W\|_{\mathbb{T}}^2 - \|W_0\|_{\mathbb{T}}^2)$，并且

$$\begin{aligned}
\Delta_k &= \frac{1}{\prod_{l=1}^{K} N_l} \cdot \frac{\prod_{l=1}^{K} N_l^2}{\prod_{l=1}^{K} n_l^2} \cdot \prod_{l=k+1}^{K} N_l \cdot (1+\epsilon_{N_k})\,\widetilde{\sigma}^2_{X,k} \\
&\le \frac{\prod_{l=1}^{K} N_l}{\prod_{l=1}^{K} n_l^2} \cdot \prod_{l=k+1}^{K} N_l \cdot (1+\epsilon_{N_k}) \cdot \left\{\frac{\prod_{l=1}^{k} n_l}{\prod_{l=1}^{k} N_l} \cdot \left(1-\frac{n_k}{N_k}\right) + \epsilon_{N_k} \cdot \mathbb{1}_{n_k<N_k}\right\} \cdot \prod_{l=1}^{k-1} N_l \cdot \left(\frac{n_{k+1}\cdots n_K}{N_{k+1}\cdots N_K}\right)^2 \\
&= \frac{\prod_{l=1}^{k} N_l}{\prod_{l=1}^{k} n_l^2} \cdot (1+\epsilon_{N_k}) \cdot \left\{\frac{\prod_{l=1}^{k} n_l}{\prod_{l=1}^{k} N_l} \cdot \left(1-\frac{n_k}{N_k}\right) + \epsilon_{N_k} \cdot \mathbb{1}_{n_k<N_k}\right\} \cdot \frac{\prod_{l=1}^{k} N_l}{N_k} \\
&= \frac{1+\epsilon_{N_k}}{N_k} \cdot \left\{\frac{\prod_{l=1}^{k} N_l}{\prod_{l=1}^{k} n_l} \cdot \left(1-\frac{n_k}{N_k}\right) + \epsilon_{N_k} \cdot \mathbb{1}_{n_k<N_k} \left(\frac{\prod_{l=1}^{k} N_l}{\prod_{l=1}^{k} n_l}\right)^2\right\} \\
&= \frac{1+\epsilon_{N_k}}{N_k\alpha_k} \cdot \left(1-\frac{n_k}{N_k} + \frac{\epsilon_{N_k}}{\alpha_k} \cdot \mathbb{1}_{n_k<N_k}\right).
\end{aligned}$$

这就完成了证明的第一部分。对于后半部分，由于 $N_K\alpha \ge 1$，我们可以将 $(n_1,\ \cdots,\ n_k) = (N_1,\ \cdots,\ N_{K-1},\ \alpha N_K)$ 代入前面的界。由于在这一选择下

$$\frac{1+\epsilon_{N_k}}{N_k\alpha_k} \cdot \left(1-\frac{n_k}{N_k} + \frac{\epsilon_{N_k}}{\alpha_k} \cdot \mathbb{1}_{n_k<N_k}\right) = \mathbb{1}_{k=K} \cdot \frac{1+\epsilon_{N_K}}{N_K\alpha} \cdot \left(1-\alpha+\frac{\epsilon_{N_K}}{\alpha}\right).$$

D.4 推论5.2 的证明

令 $U = \begin{bmatrix} u_1 & \cdots & u_q \end{bmatrix}^\top = \begin{bmatrix} \widetilde{u}_1 & \cdots & \widetilde{u}_q \end{bmatrix}$和 $\theta_k = \begin{bmatrix} \theta_{k,1} & \cdots \theta_{k,r} \end{bmatrix}$。回顾 $U_k^\top U_k = U^\top D_k U$，其中 $D_k = \mathrm{diag}(d_{k,\ 1},\ \cdots,\ d_{k,\ q}) \ge 0$ 是一个至多含有 q' 个非零元素的对角矩阵。

~**步骤 1：通过定义来检验交换性约束** $(X,\ V)$。 交换性条件 1 对 $W_k = U_k U_k^\top$和 $X_k = \theta_k$成立。事实上

$$W_k X_k = U^\top D_k U \theta_k = \sum_{l=1}^{q} d_l u_l u_l^\top \theta_k = \underbrace{\begin{bmatrix} u_1^\top\theta_k \cdots u_q^\top\theta_k & & & \\ & u_1^\top\theta_k \cdots u_q^\top\theta_k & & \\ & & \ddots & \\ & & & u_1^\top\theta_k \cdots u_q^\top\theta_k \end{bmatrix}}_{q\times q^2 r} \underbrace{\begin{bmatrix} d_{k,1}u_{l,1}I_r \\ \vdots \\ d_{k,q}u_{q,1}I_r \\ d_{k,1}u_{l,2}I_r \\ \vdots \\ d_{k,q}u_{q,2}I_r \\ \vdots \\ d_{k,q}u_{q,q}I_r \end{bmatrix}}_{q^2 r\times r}$$

$$= \underbrace{\left(I_q \otimes \begin{bmatrix} u_1^\top\theta_k & \cdots & u_q^\top\theta_k \end{bmatrix}\right)}_{:=\widetilde{X}_k} \underbrace{\begin{bmatrix} D_k\widetilde{u}_1 \otimes I_r \\ \vdots \\ D_k\widetilde{u}_q \otimes I_r \end{bmatrix}}_{=:V_k} = \widetilde{X}_k V_k.$$

步骤 2：检查有界性条件和权重参数。 我们验证条件~对于定理 5。对于数据 X_k 和 X_k，我们有

$$\begin{aligned} \|X_k\| &= \|\theta_k\| \le \|\theta_k\|_F \le 1, \\ \left\|\widetilde{X}_k\right\| &= \sqrt{\left\|\widetilde{X}_k\widetilde{X}_k^\top\right\|} = \sqrt{\left\|\left(I_q \otimes \begin{bmatrix} u_1^\top\theta_k & \cdots & u_q^\top\theta_k \end{bmatrix}\right)\left(I_q \otimes \begin{bmatrix} u_1^\top\theta_k & \cdots & u_q^\top\theta_k \end{bmatrix}^\top\right)\right\|} \\ &= \sqrt{\left\|I_q \otimes \begin{bmatrix} u_1^\top\theta_k & \cdots & u_q^\top\theta_k \end{bmatrix}\begin{bmatrix} u_1^\top\theta_k & \cdots & u_q^\top\theta_k \end{bmatrix}^\top\right\|} \\ &= \sqrt{\left\|I_q \otimes \sum_{l=1}^{q} u_l^\top\theta_k\theta_k^\top u_l\right\|} = \sqrt{\mathsf{Tr}\left(U\theta_k\theta_k^\top U\right)} = \|\theta_k\|_F \le 1. \end{aligned}$$

对于权重矩阵，我们有 $\min\{\|w\|_\infty,\ \|v\|_\infty\} \le \|w\|_\infty = \max_{k\in[N]}\|W_k\| = \max_{k\in[N]}\|D_k\| = L$，并且

$$\begin{aligned} \left\|\sum_{k=1}^{N} W_k W_k^\top\right\| &= \left\|\sum_{k=1}^{N} D_k^2\right\| \le M, \\ \left\|\sum_{k=1}^{N} V_k^\top V_k\right\| &= \left\|\sum_{k=1}^{N}\sum_{l=1}^{q} (D_k\widetilde{u}_l \otimes I_r)^\top (D_k\widetilde{u}_l \otimes I_r)\right\| = \left\|\sum_{k=1}^{N}\sum_{l=1}^{q} \widetilde{u}_l^\top D_k^2 \widetilde{u}_l \otimes I_r\right\| \\ &= \sum_{l=1}^{q} \widetilde{u}_l^\top\left(\sum_{k=1}^{N} D_k^2\right)\widetilde{u}_l = \mathsf{Tr}\left(U^\top \sum_{k=1}^{N} D_k^2 U\right) = \mathsf{Tr}\left(\sum_{k=1}^{N} D_k^2\right) \\ &\le q'\left\|\sum_{k=1}^{N} D_k^2\right\| = q' M. \end{aligned}$$

在最后一个不等式中，我们利用了这样一个事实：每个 D_k 都是对角的，且至多有 q' 个非零项。

步骤 3：界定数据方差参数。 接下来，我们计算方差参数$\widetilde{\sigma}_X^2$ 和 $\widetilde{\sigma}_{\widetilde{X}}^2$。显然，如 $X_k = \theta_k$ 所示，

$$\widetilde{\sigma}_X^2 = \left\|\frac{1}{N}\sum_{k=1}^{N}(\theta_k - \overline{\theta})(\overline{\theta} - \theta_k)^\top\right\| + \epsilon_N \max_{k\in[N]}\left\|\theta_k - \overline{\theta}\right\|^2.$$

~而对于 X，由于

$$\begin{aligned}\widetilde{\Sigma} &= \frac{1}{N}\sum_{k=1}^{N}(\widetilde{X}_k-\overline{\widetilde{X}})^\top(\widetilde{X}_k-\overline{\widetilde{X}})\\ &= \frac{1}{N}\sum_{k=1}^{N}\left(I_q\otimes\begin{bmatrix}u_1^\top\theta_k & \cdots & u_q^\top\theta_k\end{bmatrix}\right)^\top\left(I_q\otimes\begin{bmatrix}u_1^\top\theta_k & \cdots & u_q^\top\theta_k\end{bmatrix}\right)-\left(I_q\otimes\begin{bmatrix}u_1^\top\overline{\theta} & \cdots & u_q^\top\overline{\theta}\end{bmatrix}\right)^\top\left(I_q\otimes\begin{bmatrix}u_1^\top\overline{\theta} & \cdots & u_q^\top\overline{\theta}\end{bmatrix}\right)\\ &= \frac{1}{N}\sum_{k=1}^{N}I_q\otimes\left[\theta_k^\top u_i u_j^\top\theta_k\right]_{i,j\in[q]}-I_q\otimes\left[\overline{\theta}u_i u_j^\top\overline{\theta}\right]_{i,j\in[q]},\end{aligned}$$

我们有

$$\sigma_{\widetilde{X}}^2=\left\|\frac{1}{N}\sum_{k=1}^{N}I_q\otimes\left[\theta_k^\top u_i u_j^\top\theta_k\right]_{i,j\in[q]}-I_q\otimes\left[\overline{\theta}u_i u_j^\top\overline{\theta}\right]_{i,j\in[q]}\right\|=\left\|\frac{1}{N}\sum_{k=1}^{N}\left[(\theta_k-\overline{\theta})^\top u_i u_j^\top(\theta_k-\overline{\theta})\right]_{i,j\in[q]}\right\|.$$

因此

$$\begin{aligned}\sigma_{\widetilde{X}}^2 &= \left\|\frac{1}{N}\sum_{k=1}^{N}\left[(\theta_k-\overline{\theta})^\top u_i u_j^\top(\theta_k-\overline{\theta})\right]_{i,j\in[q]}\right\|=\left\|\frac{1}{N}\begin{bmatrix}\sum_{k=1}^N(\theta_k-\overline{\theta})^\top u_1\\ \vdots\\ \sum_{k=1}^N(\theta_k-\overline{\theta})^\top u_q\end{bmatrix}\begin{bmatrix}\sum_{k=1}^N(\theta_k-\overline{\theta})^\top u_1\\ \vdots\\ \sum_{k=1}^N(\theta_k-\overline{\theta})^\top u_q\end{bmatrix}^\top\right\|\\ &= \left\|\frac{1}{N}\begin{bmatrix}\sum_{k=1}^N(\theta_k-\overline{\theta})^\top u_1\\ \vdots\\ \sum_{k=1}^N(\theta_k-\overline{\theta})^\top u_q\end{bmatrix}^\top\begin{bmatrix}\sum_{k=1}^N(\theta_k-\overline{\theta})^\top u_1\\ \vdots\\ \sum_{k=1}^N(\theta_k-\overline{\theta})^\top u_q\end{bmatrix}\right\|=\left\|\sum_{l=1}^{q}u_l^\top\left(\frac{1}{N}\sum_{k=1}^{N}(\theta_k-\overline{\theta})(\theta_k-\overline{\theta})^\top\right)u_l\right\|\\ &= \mathsf{Tr}\left(U\left(\frac{1}{N}\sum_{k=1}^{N}(\theta_k-\overline{\theta})(\theta_k-\overline{\theta})^\top\right)U^\top\right)=\frac{1}{N}\sum_{k=1}^{N}\left\|\theta_k-\overline{\theta}\right\|_F^2.\end{aligned}$$

结合

$$\begin{aligned}\left\|\widetilde{X}_k-\overline{\widetilde{X}}\right\|^2 &= \left\|I_q\otimes\begin{bmatrix}u_1^\top(\theta_k-\overline{\theta}) & \cdots & u_q^\top(\theta_k-\overline{\theta})\end{bmatrix}\right\|^2\\ &= \left\|\begin{bmatrix}u_1^\top(\theta_k-\overline{\theta}) & \cdots & u_q^\top(\theta_k-\overline{\theta})\end{bmatrix}\right\|^2=\sum_{l=1}^{q}\left\|u_l^\top(\theta_k-\overline{\theta})\right\|^2\\ &= \mathsf{Tr}\left(U(\theta_k-\overline{\theta})(\theta_k-\overline{\theta})U^\top\right)=\left\|\theta_k-\overline{\theta}\right\|_F^2,\end{aligned}$$

可得

$$\widetilde{\sigma}_X^2\le\widetilde{\sigma}_{\widetilde{X}}^2=\frac{1}{N}\sum_{k=1}^{N}\left\|\theta_k-\overline{\theta}\right\|_F^2+\epsilon_N\max_{k\in[N]}\left\|\theta_k-\overline{\theta}\right\|_F^2=\widetilde{\sigma}_\theta^2.$$

步骤 4：完成证明。 现在，我们可以将上述计算结果代入定理5。由于

$$\begin{aligned}&\min\{\|w\|_\infty,\|v\|_\infty\}\le L,\\ &\max\left\{\widetilde{\sigma}_X^2\left\|\sum_{k=1}^{N}W_kW_k^\top\right\|,\widetilde{\sigma}_{\widetilde{X}}^2\left\|\sum_{k=1}^{N}V_k^\top V_k\right\|\right\}\le q'\widetilde{\sigma}_\theta^2M,\end{aligned}$$

应用引理 A.5可得

$$\mathbb{P}\left\{\left\|\sum_{k=1}^{N}U_k^\top U_k\theta_k-\overline{\theta}\right\|\ge\widetilde{\sigma}_\theta\sqrt{2(1+\epsilon_N)q'M\log\frac{q+r}{\delta}}+\frac{2L}{3}(1+\epsilon_N)\log\frac{q+r}{\delta}\right\}\le\delta.$$

证明至此完成。