arXiv:2601.01371v1 [math.ST] 4 Jan 2026

依赖数据下的随机梯度下降：最优估计、遗憾与推断

Yinan Shen[1], Yichen Zhang[2], Wen-Xin Zhou[3]

[1]南加州大学数学系[2]普渡大学丹尼尔斯商学院[3]伊利诺伊大学芝加哥分校信息与决策科学系

摘要

本文研究时间依赖数据下随机梯度下降（SGD）所产生的最终迭代点的表现。我们考虑两种互补的依赖来源：(*i*) 协变量过程和噪声过程中的鞅型依赖，这种依赖形式适用于非平稳和非混合时间序列数据；以及 (*ii*) 序贯决策所引起的依赖关系。我们的表述与经典的（局部）平稳性和强混合概念相呼应，但两种框架都无法完全涵盖彼此。值得注意的是，研究表明，在一大类步长方案和探索率方案下，随机梯度下降能够自动适应独立信息和相关信息。

在非渐近情形下，我们证明随机梯度下降能够同时实现统计最优的估计误差和遗憾，从而拓展并改进了现有结果。特别是，即使时域 $T = +\infty$ 可能无限，我们的尾界仍然是紧的。在渐近情形下，随机梯度下降迭代序列收敛到高斯分布，且余项仅为 $O_{\mathbb{P}}(1/\sqrt{t})$，这表明先前工作所声称的估计—遗憾权衡实际上可以避免。我们还提出了一种新的决策区域“锥形”近似，使协变量可以具有无界支撑。针对在线稀疏回归，我们开发了一种新的基于随机梯度下降的算法，该算法仅需 d 单位的存储，每次迭代需要 $O(d)$ 次浮点运算，从而实现长期统计最优性。直观而言，每个新到来的观测都有助于提高估计精度，而汇总统计量则为支撑集恢复提供指导。

关键词：渐近分布、条件 Orlicz 范数、决策、鞅、非渐近分析、随机梯度下降、稀疏回归

1 引言

随机梯度下降（SGD）是在线学习和序贯决策的基础算法。尽管其形式简单，但当输入数据存在依赖关系时，理解其统计行为会变得更加困难。这种依赖关系可能源于协变量或噪声构成时间序列，也可能源于数据通过学习者的决策以自适应方式收集。此类依赖关系广泛存在于现代应用中，包括上下文老虎机、金融与经济预测（Black，1986年；Ubukata and Oya，2009年；Jacod et al.，2017），以及流式预测问题（Robbins和Monro，1951；Lai，1987年；Wuet al.，2006；Wu，2007；Wu andZhao，2007）。因此，一个核心的方法论与理论问题是：如何设计并分析能够自动适应潜在混合来源的独立与相关观测的随机梯度下降过程，而无需明确了解其潜在依赖结构。

关于分析相关过程的工具，已有大量研究成果，涵盖（局部）平稳性、物理或函数依赖度量（Dedecker and Rio, 2000; Wu, 2007; Xiao and Wu, 2012; Nagaraj et al., 2020; Srikant, 2024）、 ρ-混合及其他相关弱依赖条件（Masry and Fan, 1997; Doukhan et al., 1994; Merlevède et al., 2009; Lu et al., 2022），以及对条件分布的基于密度的刻画（Tartakovsky, 2024）。尽管这些框架提供了强有力的理论保证，但它们通常依赖于对时间依赖的精确刻画，例如平稳性、混合速率或条件密度。然而，现代在线环境往往不满足这些假设：协变量可能既非平稳，也不满足混合条件；噪声分布可能随时间变化；自适应数据收集还会引入额外的依赖渠道，而经典随机过程理论无法充分刻画这些依赖（Azuma, 1967; Koltchinskii, 1994; van de Geer, 2002; Lauer, 2023）。

本文针对随机梯度下降建立了一个估计与推断框架，该框架能够容纳独立观测与相关观测同时存在的情形，而无需试图区分或剥离它们的来源。我们的表述与经典依赖概念相呼应或对其进行了扩展，但具有更大的灵活性。具体而言，协变量可以按照既不满足时间上的独立性、也不局部平稳或强混合的过程演化，但仍属于我们理论的适用范围。此外，在上下文老虎机设定中，自适应收集的奖励引入了第二种依赖关系：不断演化的决策规则与数据生成机制相互作用，共同塑造观测序列。我们的框架在统一的分析中纳入了这两类依赖关系。为说明我们的分析所涵盖的广泛依赖结构，考虑一列按以下任一动态演化的 d 维协变量 $\{\mathbf{X}_t\}$：（i）$\mathbf{X}_t = \nu_t \mathbf{X}_{t-1}/\|\mathbf{X}_{t-1}\| + \mathbf{E}_t$，或（ii）$\mathbf{X}_0 \sim$均匀分布($\mathbb{S}^{d-1}$)，且$\mathbf{X}_t = \nu_{t,0}\mathbf{X}_0 + \cdots +$

$\nu_{t,t-1}\mathbf{X}_{t-1} + \mathbf{E}_t$ 对于 $t \geq 1$，其中 $\{\nu_t\}$、$\{\nu_{t,i}\}$ 和 $\{\mathbf{E}_t\}$ 是随机的，但不一定平稳或相互独立，并且满足 $\sum_{i=0}^{t} |\nu_{t,i}| \leq 1/2$。在这类设定下，经典的平稳或混合分析通常不适用。然而，我们证明，在协变量与噪声满足适当正则性条件的情况下，具有这些相关协变量的随机梯度下降的非渐近性能可以达到与理想化独立同分布高斯设定相当的水平。此外，在温和的收敛条件和有限矩条件下，我们推导出随机梯度下降迭代序列的渐近分布。

噪声过程中也可能出现同类型的依赖关系，而我们的框架同样能够对此加以处理。更广泛地说，本文所考虑的依赖关系具有双重性：除了协变量和噪声固有的时间依赖之外，我们还纳入了由决策引起的依赖关系。这后一种依赖是上下文老虎机问题所固有的（Goldenshluger and Zeevi，2013 年；Hao 等人，2020年；Bastani and Bayati，2020年；Bastani 等人，2021 年；Chen 等人，2020、2021b、a；Han 等人，2025；Duan 等人，2024年；Han 等人，2024a），在这类问题中，学习者观测协变量、选择一个动作，并据此获得奖励。探索（随机抽取）与利用（抽取经验最优臂）之间的平衡至关重要。现有的探索策略可分为两类。第一类依赖于预先设定的探索时刻网格（Goldenshluger and Zeevi，2013 年；Hao 等人，2020年；Bastani and Bayati，2020年；Bastani 等人，2021 年；Ren and Zhou，2024年），这使得优雅的离线分析成为可能，但需要存储并反复使用历史观测，同时还需要预先知道总时域。第二类由 ε-贪心型方案构成（Chen 等人，2020、2021a、b；Han 等人，2025；Chen 等人，2022；Duan 等人，2024年），这些方案在不知道时域的情况下指定探索概率 π_t，并且不存储完整的数据流。虽然这带来了实践上颇具吸引力的完全在线算法，但也引入了重大的理论挑战。特别是，在独立数据下，现有的在线分析通常需要相对较高的探索率，往往假设 $\lim_{t\to\infty} \pi_t > 0$ 或 $\pi_t = t^{-\alpha}$ 且 $\alpha \in \in (0, 1)$，从而得到阶为 $O(T)$ 或 $O(T^{2/3})$ 的遗憾率，这些遗憾率严格劣于离线保证。

本研究旨在弥合这一差距。我们允许采用广泛的探索计划，包括以下形式的选择：$\pi_t = f(t)$ 其中函数 f 在定义 3中给出。尽管具有这种灵活性，我们的算法仍能同时实现最优的估计误差和遗憾。为处理依赖型流数据而无需存储过去的观测，我们通过随机梯度下降更新参数，并提出一类新的步长方案。如第 4节所示，当 $f(t)$ 满足定义 3 以及增长条件 $\int_0^T f(t)\mathrm{d}t \leq O(\sqrt{T})$时，即使对于离散协变量，遗憾和估计误差也都达到极小极大最优速率。尽管协变量与噪声可能存在依赖关系，我们的非渐近保证仍与假设数据独立同分布的最优离线速率相匹配。

在不保留历史数据的完全在线分析中，另一项挑战是尾概率的累积

（Jin 等人，2016；Chen 等人，2020，2021a；Li 等人，2023；Duan 等人，2024；Shen 等人，2025）。这一问题迫使先前研究将时域限制为 $T = O(d^C)$ 或 $T \leq \exp(cd)$。我们的分析避免了这一限制：在我们的随机梯度下降算法下，所有非渐近保证对 $T = +\infty$ 仍然成立。在第2节中，借助 Ma 等（2024）的下界结果，我们进一步表明，对于允许 $t = 1$、……、$+\infty$ 的算法而言，我们的结果在统计意义上是最优的。渐近地，我们得到一种 Bahadur 型表示

$$\sqrt{t}\,(\boldsymbol{\beta}_t - \boldsymbol{\beta}^*) = \text{a term that converges to Gaussian in distribution} + O_{\mathbb{P}}\left(\frac{1}{\sqrt{t}}\right),$$

对于在线程序而言，这是紧的。现有在线方法即使在独立同分布数据下且具有较弱的遗憾保证，也会由于依赖效应而产生阶为 $t^{1/5}$ 的残差。本文所实现的多重最优性源于算法与分析的协同设计。

在线稀疏线性回归长期以来一直是一个具有挑战性的问题。如何以在线方式估计稀疏参数，在很大程度上仍是一个开放问题。从概念上讲，在线稀疏线性回归将欧几里得范数下的估计（Blumensath and Davies，2009年；Bogdan 等人，2015年；Bellec 等人，2018年；Sun 等人，2020年）与支撑集恢复（Fan and Lv，2008年；Candes and Plan，2009年；Fan 等人，2014年；Zhang，2010）结合起来。随着流式数据不断到达，目标是在最终恢复整个支撑集的同时，实现单调递减的估计误差。然而，现有的离线理论与假设无法直接迁移到在线场景中。例如，离线分析通常要求最小信号强度具有下界（Zhang，2010），例如对于固定样本量 n，要求 $\min_{j:\beta_j \neq 0} |\beta_j| \geq O(1/\sqrt{n})$。相比之下，在线学习可能涉及未知或无界的时域。为规避这一困难，Fan 等人（2018）假设初始化步骤已经识别出真实支撑集，但他们仍然要求满足最小信号强度条件。Han 等人（2024b）和 Yang 等人（2023 年）提出了在线最小绝对收缩和选择算子算法，但每当有新的观测到达时，他们都会重新计算一个完整的最小绝对收缩和选择算子估计量。这三项研究均聚焦于独立同分布观测下的稀疏估计。

随机梯度下降（SGD）算法，也称随机逼近，可追溯至Robbins和Monro (1951年）；Kiefer和Wolfowitz(1952年）；Polyak和Juditsky (1992年）。尽管随机梯度下降的最终迭代点在实践中得到广泛应用，但其统计性能，尤其是它在估计误差和非渐近保证方面能否达到离线估计量的水平，仍不甚明了。本文研究了随机梯度下降的最终迭代点在不存储历史数据且不对过去的迭代量进行平均时的表现。下文将概述我们的主要贡献，第5节通过数值实验验证我们的理论结果。

1. 相关与独立数据：我们的依赖结构与经典的平稳或混合框架并行，且这两类框架互不包含。具体而言，我们将独立同分布高斯设定推广到条件 Orlicz 范数有界的随机变量。

有趣的是，即使协变量和噪声存在依赖关系，随机梯度下降的非渐近性能本质上仍与独立同分布高斯情形相同。在较弱的收敛条件和矩条件下，我们建立了渐近结果；马尔可夫结构不会降低随机梯度下降的渐近性能。由此得到的巴哈杜尔型余项小至 $1/\sqrt{t}$。

–

2. 线性回归中随机梯度下降的性能： 即使在线性回归中，在线随机梯度下降能否产生在统计上可与离线方法相媲美的估计量，仍然未知。在第 2 节中，我们不涉及决策，研究了在协变量和噪声可能相关的情况下，其非渐近和渐近性能。从非渐近角度来看，我们证明尾概率

$$\max_{\boldsymbol{\beta}_t(\{\mathbf{X}_l,Y_l\}_{l=1}^t)} \mathbb{P}\left(\cup_{t=0}^{T} \left\{ \|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|^2 \le \frac{Cd}{t} \right\} \right)$$

至少为 $1 - T\exp(-cd)$，与最佳离线结果相匹配。此外，我们证明轨迹

$$\cup_{t=0}^{+\infty} \left\{ \|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|^2 \le C\frac{d+\log t}{t} \right\}$$

以至少 $1-\exp(-cd)$ 的概率成立。借鉴 Ma 等人(2024年) 的结果，我们进一步说明，估计误差和尾概率在常数因子范围内是最优的。现有研究聚焦于独立同分布观测，并要求 $T \le d^C$ 或 $T \le \exp(cd)$ (Jin 等人，2016年；Chen 等人，2020年，2021a；Li 等人，2023 年；Duan 等人，2024年；Shen 等人，2025年）。

3. 决策与线性赌博机：即使在协变量和噪声独立同分布的情况下，随机梯度下降能否在线同时实现统计最优的估计和遗憾，一直是个未解问题。本文填补了这一空白，并在协变量和噪声相关的情况下确立了相应结果。对于广泛的探索率和步长，包括由定义 3 中一般函数生成的探索率和步长，我们在最小假设下给出了保证。我们进一步证明，采用本文步长方案的随机梯度下降迭代序列收敛到一个多元高斯分布。此外，通过充分纳入相关数据，我们首次针对递减探索率和零探索率建立了推断结果。本文提出的步长方案也是新的。

4. 在线稀疏学习： 如何设计一种稀疏线性回归在线算法，使其既不存储全部数据，也不必在每个新样本到来时都求解一个新的优化问题，这一问题一直悬而未决。本文提出一种新的基于随机梯度下降的稀疏线性回归算法，在推动最先进水平方面取得进展，其存储成本为 d，每次迭代计算量为 $O(d)$。直观而言，每个观测都能助力提升估计精度，而

汇总统计量则用于变量选择。从长远来看，我们的算法能够实现统计意义上最优的迭代量。

5. 锥形决策区域：本文提出了对真实决策区域的一种新型锥形近似，该近似允许协变量具有可能无界的支撑集，并且能够充分覆盖原点周围的一个邻域。此外，我们的渐近结果基于锥形测度，从而消除了现有研究中对分布连续性的条件（Chen 等人，2022，2020年，2021a）。当协变量服从离散分布时，我们的刻画与 Bastaniand Bayati （2020年）和Bastani 等人（2021 年）所采用的刻画一致。

此外，本文发展的统计理论本身也可能具有独立的研究价值。本节最后介绍全文中使用的符号。

符号说明：我们用粗体符号表示向量和矩阵（例如，g、X、Y），用花体字表示集合、算子或 σ-代数（例如， $\mathcal{A}$、$\mathcal{H}$、$\mathcal{F}$）。全文中， $[K]$ 表示集合$\{1, 2, \ldots, K\}$。对于两个对称矩阵 A 和 B，A $\succeq$ B 表示 A $-$ B 的特征值均非负。算子 $\|\cdot\|$ 表示向量的 L_2 范数和矩阵的算子范数，而 $\|\cdot\|_\infty$ 表示向量各分量绝对值的最大值。对于对称矩阵 A $=\mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top \in \mathbb{R}^{d\times d}$，其中 $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$，以及函数 $f(\cdot)$，我们定义 $f(\mathrm{A}) = \mathrm{U}f(\mathbf{\Lambda})\mathrm{U}^\top$，并且$f(\mathbf{\Lambda}) = \mathrm{diag}(f(\lambda_1)\ldots, f(\lambda_d))$。我们用 $\rightsquigarrow$表示依分布收敛。

对于向量 $\mathbf{X} \in \mathbb{R}^d$， $[\mathbf{X}]_i$ 表示其第 i 个元素；对于集合 $\mathcal{S} \subseteq [d]$， $[\mathbf{X}]_\mathcal{S}$ 表示$\mathbf{X}$ 中索引属于 $\mathcal{S}$ 的子向量。在稀疏线性回归中，算子 $\mathcal{H}_\mathcal{S}(\cdot): \mathbb{R}^d \to \mathbb{R}^d$仅保留 $\mathcal{S}$ 中的元素，并将 $\mathcal{S}^c$ 中的元素设为零；也就是说，对于 $i \in \mathcal{S}$， $[\mathcal{H}_\mathcal{S}(\mathbf{X})]_i = [\mathbf{X}]_i$，对于 $i \in \mathcal{S}^c$ 则为 0。

2 含相关数据的线性回归

我们首先介绍条件 Orlicz 范数，它在我们的分析中发挥着核心作用。这一概念并非新提出的，而是对若干情形的推广。Koltchinskii（1994）、van de Geer（1995）、van de Geer（2002）、Shamir（2011）和 Lauer（2023）等人的奠基性工作，以及经典的 Azuma 不等式（Azuma，1967），研究了有界或满足有界条件次高斯尾部的相依随机变量的集中不等式。近年来，随着随机梯度下降（SGD）估计误差的非渐近界日益受到关注，Han 等人（2025）和 Shen 等人（2025）等研究采用了基于条件 Orlicz 范数的方法。然而，这些研究聚焦于相互独立的协变量和噪声。我们的工作大幅推广了这一框架，允许存在更

一般的依赖结构。具体而言，我们证明，在适当条件下，相关数据下的随机梯度下降表现可与理想化的独立同分布高斯设定相媲美。

2.1 预备知识：Orlicz范数与条件 Orlicz 范数

回顾一下，单变量随机变量的 Orlicz范数定义如下:

$$\|X\|_{\Psi_\alpha} := \inf\left\{u > 0 : \mathbb{E}\exp(|X/u|^\alpha) \le 2\right\}, \quad \alpha \ge 1.$$

Let $\mathcal{F}$是一个 σ域。若随机变量 $K_{\mathcal{F}}$满足 $K_{\mathcal{F}} \in \mathcal{F}$，则称其为 $\mathcal{F}$可测的。

定义 1 (条件 Orlicz 范数)。 设 $\mathcal{F}$ 是一个 σ域，且 $K_{\mathcal{F}} > 0$ 是 $\mathcal{F}$可测的。当且仅当 X 满足 $\|X|\mathcal{F}\|_{\Psi_\alpha} \le K_{\mathcal{F}}$ 时，

$$\mathbb{E}\left\{\exp(|X/K_{\mathcal{F}}|^\alpha)|\mathcal{F}\right\} \le 2 \quad a.s. \tag{1}$$

注意，上述条件期望本身是 $\mathcal{F}$可测的随机变量。下面通过一些例子来说明条件 Orlicz 范数的概念。

例 1。 若 X是 $\mathcal{F}$可测的，则 $\|X|\mathcal{F}\|_{\Psi_\alpha} \le (\log 2)^{-1/\alpha} \cdot |X|$。若 X与 $\mathcal{F}$独立，则$\|X|\mathcal{F}\|_{\Psi_\alpha} \le \|X\|_{\Psi_2}$。

对于随机向量 $\mathbf{X} \in \mathbb{R}^d$，定义

$$\|\mathbf{X}\|_{\Psi_\alpha} := \sup_{\mathbf{u}\in\mathbb{S}^{d-1}} \|\mathbf{u}^\top \mathbf{X}\|_{\Psi_\alpha}.$$

条件版本的定义类似如下。

定义 2 (随机向量的条件 Orlicz 范数)。 假设 $\mathcal{F}$ 是一个 σ-域，且 $K_{\mathcal{F}} > 0$是 $\mathcal{F}$-可测的。当且仅当随机向量 $\mathbf{X}$满足 $\|\mathbf{X}|\mathcal{F}\|_{\Psi_\alpha} \le K_{\mathcal{F}}$时，

$$\sup_{\mathbf{u}\in\mathbb{S}^{d-1},\mathbf{u}\in\mathcal{F}} \mathbb{E}\left\{\exp(|\mathbf{u}^\top \mathbf{X}/K_{\mathcal{F}}|^\alpha)|\mathcal{F}\right\} \le 2 \quad a.s. \tag{2}$$

此外，对于随机矩阵 $\mathbf{X} \in \mathbb{R}^{d_1\times d_2}$，定义

$$\|\mathbf{X}\|_{\Psi_\alpha} := \sup_{\mathbf{u}\in\mathbb{S}^{d_1-1}} \sup_{\mathbf{v}\in\mathbb{S}^{d_2-1}} \|\mathbf{u}^\top \mathbf{X}\mathbf{v}\|_{\Psi_\alpha},$$

条件范数的定义类似。这些范数定义良好，并满足通常的范数性质；相关细节将在附录 A中给出。最后，我们给出条件 Orlicz 范数的另一个说明性示例，可直接验证。

例2. 设 $\mathcal{F} = \sigma(X)$。定义 $Y_1 = X + \xi$ 和 $Y_2 = \xi X$，其中 ξ 与 X 和$\|\xi\|_{\Psi_\alpha} < \infty$相互独立。则

$$\|Y_1|\mathcal{F}\|_{\Psi_\alpha} \le |X|/\log 2 + \|\xi\|_{\Psi_\alpha}, \quad \|Y_2|\mathcal{F}\|_{\Psi_\alpha} \le |X| \cdot \|\xi\|_{\Psi_\alpha}.$$

若 $|X| \le K$几乎必然， $\|Y_1|\mathcal{F}\|_{\Psi_\alpha} \le K/$对数 $2 + \|\xi\|_{\Psi_\alpha}$和 $\|Y_2|\mathcal{F}\|_{\Psi_\alpha} \le K \cdot \|\xi\|_{\Psi_\alpha}$。

2.2 相关协变量与噪声下的线性回归

为将依赖关系的影响与决策所导致的影响区分开来，我们首先通过随机梯度下降研究依赖数据下的线性回归。在每个时刻 t，我们观测到 $(\mathbf{X}_t,\ Y_t) \in \mathbb{R}^d \times \mathbb{R}$，这些数据由以下线性模型生成

$$Y_t = \mathbf{X}_t^\top \boldsymbol{\beta}^* + \xi_t,$$

其中 $\mathbf{X}_t$ 和 ξ_t 可能依赖于过去的观测。目标是以在线方式估计 $\boldsymbol{\beta}^* \in \mathbb{R}^d$，随着每个数据点到来逐一处理，而不存储全部历史数据。我们允许任意初始化 $\boldsymbol{\beta}_0 \in \mathbb{R}^d$。在时刻 t，假设当前迭代值为 $\boldsymbol{\beta}_{t-1}$，且一组新的数据对（$\mathbf{X}_t,\ Y_t$）可用。随后，我们使用标准的随机梯度下降规则更新参数:

$$\boldsymbol{\beta}_{t+1} = \boldsymbol{\beta}_t - \eta_t \cdot (\mathbf{X}_t^\top \boldsymbol{\beta}_t - Y_t)\mathbf{X}_t, \tag{3}$$

其中 η_t 是稍后指定的步长。令 $\mathcal{F}_t = \sigma(Y_{t-1}, \mathrm{X}_{t-1}, \ldots, Y_0, \mathrm{X}_0)$ 表示由截至时刻 $t-1$ 的所有随机性生成的 σ-域，并定义 $\mathcal{F}_t^+ = \sigma(\mathrm{X}_t, Y_{t-1}, \mathrm{X}_{t-1}, \ldots, Y_0, \mathrm{X}_0)$。(3) 中的更新确保 $\boldsymbol{\beta}_t$ 是 $\mathcal{F}_t$-可测的，且 $\mathcal{F}_t \subseteq \mathcal{F}_t^+ \subseteq \mathcal{F}_{t+1}$。接下来，我们将在以下假设下证明 (3) 的收敛性。

假设1（协变量：具有条件次高斯范数的鞅差）。协变量序列 $\mathbf{X}_t \in \mathbb{R}^d$ 满足 $\mathbb{E}\{\mathbf{X}_t|\mathcal{F}_t\} = 0$ 几乎必然成立，且 $\|\mathbf{X}_t|\mathcal{F}_t\|_{\Psi_2}^2 \le \lambda_{\max}$。此外，存在 $\lambda_{\min} > 0$ 使得 $\mathbb{E}\{\mathbf{X}_t\mathbf{X}_t^\top|\mathcal{F}_t\} \succeq \lambda_{\min} \cdot \mathbf{I}_d$ 几乎必然成立。这里，$\lambda_{\min}$ 和 $\lambda_{\max}$ 为常数。

假设 2（噪声：具有条件次高斯范数的鞅差）。噪声变量 $\xi_t \in \mathbb{R}$ 满足 $\mathbb{E}\{\xi_t|\mathcal{F}_t^+\} = 0$ 几乎必然成立，且 $\|\xi_t|\mathcal{F}_t^+\|_{\Psi_2} \le \sigma$，其中 σ 为常数。

假设1和假设 2 允许 X_t 和 ξ_t 依赖于历史数据。这推广了常见的独立同分布或协变量一致有界假设，例如 Perchet 和 Rigollet（2013 年）、Fang 等人（2018 年）、Chen 等人（2020年）、Han 等（2025）、Shao 和 Zhang（2022）、Li 等人（2023 年）、Agrawalla 等（2025）、Wei 等人（2025年）、Han 等人（2024b）以及 Shen 等人（2025年）的研究。我们强调，我们的框架既不包含于经典的平稳或强混合情形，也不包含经典的平稳或强混合情形；这两种情形从根本上不可比较（Wu，2007；Xiao 和 Wu，2012）。

例*3.* 令 X_0 为任意满足 $\|X_0\|_{\Psi_2} < \infty$ 且 $\lambda_{\min}(\mathbb{E}\{X_0X_0^\top\}) > 0$ 的中心化随机向量。对于 $t \ge 1$，考虑过程 $X_t = a_t \cdot X_{t-1} + E_t$，其中 $\mathbb{E}\{a_t|\mathcal{F}_t\} = 0$, $|a_t| \le 1$, $\|E_t|\mathcal{F}_t\|_{\Psi_2} \le \lambda$，且 $E_t \perp X_{t-1}|\mathcal{F}_t$，满足 $\lambda_{\min}(\mathbb{E}\{E_tE_t^\top\}) > 0$。在这一构造中，$\{X_t\}$ 是相关的，可能不满足平稳或局部平稳性。尽管如此，它仍满足假设*1*和假设*2*。

下面的定理给出了两种步长方案下随机梯度下降的非渐近性能。有趣的是，随机梯度下降能够很好地处理相关数据，效果如同数据是独立同分布高斯，这一点也通过数值实验（第 5节）得到了验证。

定理 1（非渐近性能）。 假设 *1* 和 *2* 成立。由更新 （3） 生成的迭代量满足：

1. 常数步长。 采用常数步长 $\eta_t := \eta \le \lambda_{\min}/(d\lambda_{\max}^2)$ 时，以至少 $1-(t+1)\exp(-c_1 d)$ 的概率，有如下结论：

$$\|\boldsymbol{\beta}_{t+1}-\boldsymbol{\beta}^*\|^2 \le 5(1-\eta\lambda_{\min}/2)^{t+1}\|\boldsymbol{\beta}_0-\boldsymbol{\beta}^*\|^2 + 5\frac{\lambda_{\max}}{\lambda_{\min}}\eta\sigma^2 d.$$

特别地，经过 $t_1 \ge \log(\|\boldsymbol{\beta}_0-\boldsymbol{\beta}^*\|/(\sqrt{\eta d}\sigma))$ 个后，我们有 $\|\boldsymbol{\beta}_{t_1}-\boldsymbol{\beta}^*\|^2 \le C\frac{\lambda_{\max}}{\lambda_{\min}}\eta\sigma^2 d \le C\frac{\sigma^2}{\lambda_{\max}}$。

2. 衰减步长。 假设某次迭代 t_1 满足 $\|\boldsymbol{\beta}_{t_1}-\boldsymbol{\beta}^*\|^2 \le \sigma^2/\lambda_{\max}$。[1] 取 $\eta_t = \frac{C_a/\lambda_{\min}}{t-t_1+C_b d}$，其中 $C_a \ge 2$ 且 $C_b \ge 3C_a^2(\lambda_{\max}^2/\lambda_{\min}^2)$。对于满足 (i) $0 \le \delta_t \le \delta_{t+1}$、$(ii)$ $\delta_t \le Ct/C_b$ 的任意非递减序列 $\{\delta_t\}$，以至少 $1-C\exp(-cd) - \sum_{l=t_1}^{t}\exp(-c(d+\delta_l))$ 的概率，对所有 $t \in [t_1, +\infty]$，有

$$\|\boldsymbol{\beta}_{t+1}-\boldsymbol{\beta}^*\|^2 \le C^*\frac{1}{\lambda_{\min}}\frac{\max\{d,\delta_{t+1}\}}{t+1+C_b d}\sigma^2, \quad where\ C^* = 2CC_a\frac{\lambda_{\max}}{\lambda_{\min}}.$$

第（1）部分的证明见附录 A。第（2）部分是定理 4 的一个特例，因此略去其证明。定理 1 提供了一种两阶段在线估计策略。第一阶段采用常数步长，确保向 $\boldsymbol{\beta}^*$ 的邻域快速线性收缩；该邻域的半径与指定的常数步长 η 成正比。第二阶段采用衰减步长，并达到统计最优性能，误差量级为 $O((d+\delta_t)/t)$。取 $\delta_t = c\log t$ 后，以至少 $1-C\exp(-cd)$ 的概率，对所有 $t = t_1, \ldots, +\infty$，均有：

$$\|\boldsymbol{\beta}_t-\boldsymbol{\beta}^*\|^2 \le C^*\frac{1}{\lambda_{\min}}\frac{\max\{d,\log(t)\}}{t+C_b d}\sigma^2. \tag{4}$$

与离线线性回归相比，在线学习中额外的因子 $\log(t)/t$ 是不可避免的，这是为了允许 $t=+\infty$。事实上，Ma 等人 (2024年)表明，如果 $[\mathbf{X}_1, \ldots, \mathbf{X}_t]^\top$ 满足适当的

[1] Th为使表述清晰，施加了这一条件。证明展示了更一般的动态过程，

$$\|\boldsymbol{\beta}_{t+1}-\boldsymbol{\beta}^*\|^2 \le C\left(\frac{d}{t+1+C_b d}\right)^{C_a-2}\|\boldsymbol{\beta}_0-\boldsymbol{\beta}^*\|^2 + \frac{C}{\lambda_{\min}}\frac{\max\{d,\delta_{t+1}\}}{t+1+C_b d}\sigma^2,$$

其中，当 C_a 较大时，初始误差的影响会迅速减小。由于常数步长会使初始误差呈线性衰减，为便于说明，我们假设满足简化条件 $\|\boldsymbol{\beta}_{t_1}-\boldsymbol{\beta}^*\|^2 \le \sigma^2/\lambda_{\max}$。

条件，则

$$\mathbb{P}\left\{\|\widehat{\boldsymbol{\beta}}_t(\{\mathbf{X}_l, Y_l\}_{l=1}^t) - \boldsymbol{\beta}^*\| \geq c\sigma^2 \cdot \frac{d + \log(\delta)}{t}\right\} \geq \frac{1}{\delta}.$$

令 $\delta = t$，可得

$$\sum_{t=1}^{+\infty} \mathbb{P}\left\{\|\widehat{\boldsymbol{\beta}}_t(\{\mathbf{X}_l, Y_l\}_{l=1}^t) - \boldsymbol{\beta}^*\| \geq c\frac{\sigma^2}{\lambda_{\min}} \cdot \frac{d + \log(t)}{t}\right\} = +\infty,$$

因此对数$(t)/t$ 项在定理 1 中是最优的，误差仅差一个常数因子。即使 $\boldsymbol{\beta}_t$ 相互独立，第二博雷尔-坎特利引理也意味着

$$\mathbb{P}\left\{\|\widehat{\boldsymbol{\beta}}_t(\{\mathbf{X}_l, Y_l\}_{l=1}^t) - \boldsymbol{\beta}^*\| \geq c\frac{\sigma^2}{\lambda_{\min}} \cdot \frac{d + \log(t)}{t}, \text{ i.o.}\right\} = 1.$$

注 1 (尾概率累积**)**。 在线估计和随机梯度下降的非渐近分析通常会在每个时间步应用集中界，因为数据是按顺序处理且不进行存储。这必然会导致尾概率累积（*Jin* 等人，*2016*年；*Han* 等人，*2025*年；*Li* 等人，*2023* 年；*Shen* 等人，*2025*年；*Liu* 和 *Zhou*，*2024*年）。定理 *1* 通过适用于无限时间范围，推动了最先进水平 $T \leq +\infty$，而早期结果仅适用于时域长度不超过 $T \leq C\exp(d)$（例如，*Shen* 等人，*2025*年）。能够对可能无界的时域进行非渐近分析，是一项重大且不平凡的改进。

备注 2 （遗憾）。 在线线性回归中的遗憾有多种定义。通常，遗憾衡量的是累积预测性能，而非某一时刻的估计质量。由于定理 *1* 保证对于所有足够大的 t，估计误差均达到最优，因此相应的遗憾在统计意义上也达到最优。为避免与老虎机式遗憾定义混淆，此处不再进一步讨论。

现在我们转而研究在具有收敛二阶矩的马尔可夫协变量和噪声条件下，$\boldsymbol{\beta}_t$ 的渐近行为。结果表明，随机梯度下降迭代序列依分布收敛于多元高斯分布。我们用 $O_{\mathbb{P}}(\cdot)$ 表示随机有界性。

定理 2 （渐近性能）。 假设 $\mathbb{E}\{\mathbf{X}_t|\mathcal{F}_t\} = 0$、$\mathbb{E}\{\xi_t|\mathcal{F}_t^+\} = 0$ 和 $\xi_t \perp \mathbf{X}_t|\mathcal{F}_t$。设存在 $\boldsymbol{\Sigma}^*$ 和 σ^*，使得

$$\lim_{t\to+\infty} \mathbb{E}\left\{\left\|\mathbb{E}\{\mathbf{X}_t^\top \mathbf{X}_t|\mathcal{F}_t\} - \boldsymbol{\Sigma}^*\right\|\right\} = 0, \quad and \quad \lim_{t\to+\infty} \mathbb{E}\left\{\left|\mathbb{E}\{\xi_t^2|\mathcal{F}_t\} - \sigma_*^2\right|\right\} = 0.$$

此外，假设存在 $\lambda_{\max}, \lambda_{\min}, \sigma > 0$，使得

$$\lambda_{\min}\mathbf{I} \preceq \mathbb{E}\{\mathbf{X}_t\mathbf{X}_t^\top|\mathcal{F}_t\}, \quad \sup_{\mathbf{V}\in\mathbb{S}^{d-1}:\,\mathbf{V}\in\mathcal{F}_t} \mathbb{E}\{(\mathbf{X}_t^\top\mathbf{V})^4|\mathcal{F}_t\} \leq \lambda_{\max}^2, \quad \mathbb{E}\{\xi_t^4|\mathcal{F}_t^+\} \leq \sigma^4 \quad a.s.$$

取步长 $\eta_t = \frac{C_a/\lambda_{\min}}{t-t_1+C_b d}$，其中 $C_a \geq 2$和 $C_b \geq 3C_a^2(\lambda_{\max}^2/\lambda_{\min}^2)$，我们可以将其分解为

$$\boldsymbol{\beta}_t - \boldsymbol{\beta}^* = \sum_{j=0}^{t} \prod_{l=j+1}^{t} (\mathbf{I} - \eta_l \boldsymbol{\Sigma}^*)\, \eta_j \cdot \xi_j \cdot \mathbf{X}_j + \mathbf{R}_t,$$

其中 $R_t = O_{\mathbb{P}}(1/t)$。此外，

$$\sqrt{t}\sum_{j=0}^{t} \prod_{l=j+1}^{t} (\mathbf{I} - \eta_l \boldsymbol{\Sigma}^*)\, \eta_j \cdot \xi_j \cdot \mathbf{X}_j \rightsquigarrow N(\mathbf{0}, \mathbf{U}^* \boldsymbol{\Lambda}^*_{C_a} \mathbf{U}^{*\top}),$$

其中 $\boldsymbol{\Sigma}^* = \mathbf{U}^* \boldsymbol{\Lambda}^* \mathbf{U}^{*\top}$ 是特征分解，$\boldsymbol{\Lambda}^* = \mathrm{diag}\{\lambda_1, \ldots, \lambda_d\}$，并且

$$\boldsymbol{\Lambda}^*_{C_a} = \mathrm{diag}\left\{ \frac{\sigma_*^2}{\lambda_i} \cdot \frac{(C_a \lambda_i/\lambda_{\min})^2}{2C_a\lambda_i/\lambda_{\min} - 1} \right\}.$$

定理 2 刻画了相关协变量和噪声下随机梯度下降迭代序列的渐近分布，其中包括独立同分布高斯情形。极限协方差反映了 λ_i 和 $\lambda_{\min}$之间的相互作用。另请注意，极限分布取决于 C_a，但不取决于 C_b：渐近地，$t + C_b d = O(t)$，而定理 1 则要求 C_b 足够大，以确保获得紧确的非渐近界和尾概率。多臂老虎机设定（第 4.3 节）扩展了定理 2；因此省略其证明，并将与现有文献的比较留待第 4.4 节讨论。关于独立同分布数据下、步长按 $1/t^\alpha$ 衰减且 $\alpha \in (1/2, 1)$的相关高维推断结果，请参阅 Agrawalla 等（2025）以及 Shao 和 Zhang（2022）。

3 相关数据下的稀疏线性回归

在本节中，我们研究相关协变量和噪声存在时的稀疏线性回归。协变量和响应变量满足与第2节相同的线性模型。与该情形不同的是，现在假设未知参数向量 $\boldsymbol{\beta}^*$ 位于由稀疏性约束所诱导的低维子空间中；具体而言，

$$Y_t = \mathbf{X}_t^\top \boldsymbol{\beta}^* + \xi_t, \quad \text{with} \quad |\mathrm{supp}(\boldsymbol{\beta}^*)| \leq s.$$

我们对协变量作出如下假设。由于 $\boldsymbol{\beta}^*$具有稀疏性，该假设比假设2 更弱。

假设 3. 协变量 $\mathbf{X}_t$ 满足 $\mathbb{E}\{\mathbf{X}_t | \mathcal{F}_t\} = 0$ 几乎必然。对于某个 $s \leq d$，存在 $\lambda_{\min} > 0$ 使得

$$\min_{\mathcal{S} \subseteq [d]:\, 1 \leq |\mathcal{S}| \leq s} \lambda_{\min}\Big(\mathbb{E}\big\{ [\mathbf{X}_t]_{\mathcal{S}} [\mathbf{X}_t]_{\mathcal{S}}^\top \big| \mathcal{F}_t \big\} \Big) \geq \lambda_{\min}, \quad \text{a.s.},$$

并且存在 $\lambda_{\max} > 0$ 使得

$$\max_{\mathcal{S}\subseteq[d]:\,|\mathcal{S}|\le s} \left\|[\mathbf{X}_t]_{\mathcal{S}}\middle|\mathcal{F}_t\right\|_{\Psi_2}^2 \le \lambda_{\max}, \quad \text{a.s.}.$$

其中 $\lambda_{\min}$、$\lambda_{\max}$ 是常数。

此假设不同于假设 1：此处仅对 s 维子向量施加受限特征值条件，因此条件严格较弱。此外，假设 3 保证存在有限常数 $\lambda^{\mathsf{Off}}_{\max,\,s}$，$\lambda^{\mathsf{Off}}_{\max,\,1} \ge 0$，使得

$$\max_{\mathcal{S}'\subseteq[d]:|\mathcal{S}'|\le s}\ \max_{\mathcal{S}\subseteq[d]:|\mathcal{S}|\le s} \|[\mathbf{X}_t]_{\mathcal{S}'}[\mathbf{X}_t]_{\mathcal{S}\setminus\mathcal{S}'}^\top|\mathcal{F}_t\|_{\Psi_1} \le \lambda^{\mathsf{Off}}_{\max,s}, \quad \max_{i\notin\mathcal{S}^*}\ \max_{\mathcal{S}\subseteq[d]:|\mathcal{S}|\le s} \|[\mathbf{X}_t]_i[\mathbf{X}_t]_{\mathcal{S}\setminus i}^\top|\mathcal{F}_t\|_{\Psi_1} \le \lambda^{\mathsf{Off}}_{\max,1}.$$

这些量衡量协变量子集之间的相关性。例如，当向量$\mathbf{X}_t$ 的任意两个分量具有不相交的支撑集时，这些量为零。为简便起见，我们假设

$$\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|^2 \le C\frac{\sigma^2}{\lambda_{\max}},$$

这可以通过离线初始化来保证，也可以通过运行一个较短的预热随机梯度下降阶段来实现，步长采用常数步长（如定理 1中所述）或采用脚注 $\frac{C_a}{t+C_b d}$ 1 中描述的步长方案1。

我们的分析分为三个步骤。第 3.1 节研究固定支撑集下随机梯度下降的行为，并重点分析其对 $\lambda^{\mathsf{Off}}_{\max,\,s}$ 的依赖，即使初始支撑集不包含真实支撑集也是如此。第 3.2 节介绍一种用于支撑集恢复的统计量，并建立检测缺失支持元素的条件。第 3.3 节将估计和支撑集恢复过程整合为一个统一的算法。关键在于，尽管单个数据点足以进行估计更新，但准确的支撑集恢复从根本上需要汇总历史信息。

3.1 稀疏随机梯度下降：固定支集

本节分析在整个迭代过程中支撑集保持固定时，稀疏随机梯度下降的行为。令 $\mathcal{S}_0$ 表示初始支撑集，并令 $\mathcal{S}^* := \mathrm{supp}(\boldsymbol{\beta}^*)$ 表示真实支撑集。需要强调的是，我们不对 $\mathcal{S}_0$ 和 $\mathcal{S}^*$之间施加任何结构关系。特别是，不要求 $|\mathcal{S}^*| \nmid |\mathcal{S}_0|$，也不要求包含关系 $\mathcal{S}^* \subseteq \mathcal{S}_0$ 成立。这与Fan 等人(2018)等情形形成鲜明对比；在那些情形中，通常假设初始支撑集包含真实支撑集。

更新规则如下：在每次迭代中，算法首先利用新收到的数据点计算随机梯度，据此更新当前估计值，然后将迭代结果投影回固定支集 $\mathcal{S}_0$，即将所有其他坐标置零。形式上，更新写作

$$\boldsymbol{\beta}_{t+1} = \mathcal{H}_{\mathcal{S}_0}\big(\boldsymbol{\beta}_t - \eta_t \cdot \mathbf{g}_t\big), \tag{5}$$

其中 $\mathbf{g}_t := (\mathbf{X}_t^\top \boldsymbol{\beta}_t - Y_t)\mathbf{X}_t$ 是瞬时梯度，而 $\mathcal{H}_{\mathcal{S}_0}$ 表示硬阈值算子，它仅保留 $\mathcal{S}_0$中的坐标。尽管支撑集保持固定，但该方案与第 2 节讨论的稠密设定有很大不同。即使 $\mathcal{S}_0$ 设定有误， $\mathcal{S}_0$ 上的估计误差也会因子向量 $\mathcal{S}_0 \setminus \mathcal{S}^*$ 之间的相关性而受到 $[\mathbf{X}_t]_{\mathcal{S}_0}$、 $[\mathbf{X}_t]_{\mathcal{S}^*\setminus\mathcal{S}_0}$ 和 $[\mathbf{X}_t]_{\mathcal{S}^*}$中分量的影响。下一个引理将明确说明这种依赖关系，并展示支撑集设定有误会如何影响收敛。

引理 1. 假设假设 *2* 和 *3* 成立，且 s ，其中 $|\mathcal{S}_0| \le s$ 和 $|\mathcal{S}^*| \le s$。考虑步长 $\eta_t = \frac{1}{\lambda_{\min}} \frac{t+C_b s}{\log(2d/s)}$ ，其中 ~~$C_a \ge 2$ 且 C_b~~ $\ge C(C_a\lambda_{\max}/\lambda_{\min})^2$。则，

$$\|[\boldsymbol{\beta}_t - \boldsymbol{\beta}^*]_{\mathcal{S}_0^c}\|^2 = \|[\boldsymbol{\beta}^*]_{\mathcal{S}^*\setminus\mathcal{S}_0}\|^2,$$

对于满足定理 $\{\delta_t\}$ 所要求条件的任意尾概率序列 *1*，以至少 $1 - \sum_{l=0}^{t} \exp(-c\{s\log(2d/s) + l/C_a\}) - \sum_{l=0}^{t} \exp(-c\{s\log(2d/s) + \delta_l\})$ 的概率，有

$$\|[\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*]_{\mathcal{S}_0}\|^2 \le \frac{C^*}{\lambda_{\min}} \frac{s\log(2d/s) + \delta_{t+1}}{t+1+C_b s\log(2d/s)}\sigma^2 + C\left(\frac{\lambda^{\mathit{Off}}_{\max,s}}{\lambda_{\min}}\right)^2 \cdot \|[\boldsymbol{\beta}^*]_{\mathcal{S}^*\setminus\mathcal{S}_0}\|^2.$$

其中 $C^* = CC_a^2\lambda_{\max}/\lambda_{\min} + C_b\lambda_{\min}/\lambda_{\max}$。

该结果揭示了两个关键现象。首先，如果 $\mathcal{S}_0$ ，则其上的估计误差以与稠密情形相同的极小极大速率 $\mathcal{S}^* \subseteq \mathcal{S}_0$衰减。其次，更重要的是，第二项揭示了遗漏相关协变量所带来的持续影响。当 $\mathcal{S}_0$ 未能包含完整的真实支撑集时，由 $\lambda^{\mathsf{Off}}_{\max,s}$ 所捕获的相关性会引入一个无法通过额外迭代消除的非消失偏差项。这凸显了支撑集恢复机制的必要性，该机制将在第3.2节中提出。

3.2 稀疏随机梯度下降：支撑集恢复

如引理 1 所示，固定但设定错误的支撑集必然会导致持续偏差。因此，必须自适应地恢复 $\boldsymbol{\beta}^*$ 的支撑集。本节将介绍一种基于统计量的支撑集更新机制，该统计量汇总给定索引集上协变量与残差之间的相互作用。对于任意索引集 $\mathcal{T}$（例如 $\mathcal{T} = \{1, \dots, T\}$） ，定义

$$\mathbf{G}(\{\boldsymbol{\beta}_t, \mathbf{X}_t, Y_t\}_{t\in\mathcal{T}}) := \sum_{t\in\mathcal{T}} (\mathbf{X}_t\boldsymbol{\beta}_t - Y_t)\mathbf{X}_t.$$

向量**G统计量(·)**编码了残差与协变量之间的对齐关系，并且等于平方损失下随机梯度的总和。重要的是，即使采用其他损失函数，这一量仍然具有信息价值。我们将证明，**G统计量** 在识别方面特别有效

$\mathcal{S}^* \setminus \mathcal{S}_0$中的坐标，从而揭示真实支撑集中缺失的分量。为明确这一作用，我们根据模型结构展开**G统计量**：

$$\mathbf{G}(\{\boldsymbol{\beta}_t, \mathbf{X}_t, Y_t\}_{t\in\mathcal{T}}) = \sum_{t\in\mathcal{T}} \mathbf{X}_t\mathbf{X}_t^\top(\boldsymbol{\beta}_t - \boldsymbol{\beta}^*) - \sum_{t\in\mathcal{T}} \xi_t \cdot \mathbf{X}_t.$$

我们称序列 $\{\boldsymbol{\beta}_t\}_{t\in\mathcal{T}}$具有共同支撑集 $\mathcal{S}_\mathcal{T}$，如果对于所有 $t \in \mathcal{T}$， $[\boldsymbol{\beta}_t]_{\mathcal{S}_\mathcal{T}^c} = \mathbf{0}$。对于这样的序列，我们引入事件

$$\mathcal{E}_{\{\mathcal{S}_\mathcal{T}, V_\mathcal{T}, W_\mathcal{T}\}} := \big\{\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\| \le V_t,\ \|[\boldsymbol{\beta}_t - \boldsymbol{\beta}^*]_{\mathcal{S}_\mathcal{T}}\| \le W_t,\ \text{for all } t \in \mathcal{T}\big\},$$

其中， V_t 和 W_t 是确定性界（稍后给出），用于控制轨迹上的全局估计误差和受限估计误差。下一个引理给出了$\mathbf{G}$ 在不同坐标子集上各项大小的下界和上界。这些界形成了间隔，从而能够可靠地进行支撑集检测。

引理 2. 假设 *2* 和 *3* 成立。设 $\{\boldsymbol{\beta}_t\}_{t\in\mathcal{T}}$ 是具有共同支撑集$\mathcal{S}_\mathcal{T}$的序列，其中 $|\mathcal{T}| \ge Cs$ 且 $|\mathcal{S}_\mathcal{T}| \le s$。在事件 $\mathcal{E}_{\{\mathcal{S}_\mathcal{T}, V_\mathcal{T}, W_\mathcal{T}\}}$下，对于任意 u， $\delta > 0$，以下结论成立：

- （未覆盖支撑集坐标的下界）对于 $\mathcal{S}^* \setminus \mathcal{S}_\mathcal{T}$ 中的坐标，

$$\|[\mathbf{G}(\{\boldsymbol{\beta}_t, \mathbf{X}_t, Y_t\}_{t\in\mathcal{T}})]_{\mathcal{S}^*\setminus\mathcal{S}_\mathcal{T}}\| \ge |\mathcal{T}| \cdot \lambda_{\min}\|[\boldsymbol{\beta}^*]_{\mathcal{S}^*\setminus\mathcal{S}_\mathcal{T}}\| - \lambda^{\mathit{Off}}_{\max,s}\sum_{t\in\mathcal{T}} W_t$$
$$- u - C\sigma\sqrt{|\mathcal{T}|}\sqrt{\lambda_{\max}}\sqrt{|\mathcal{S}^* \setminus \mathcal{S}_\mathcal{T}|\log(2d/|\mathcal{S}^* \setminus \mathcal{S}_\mathcal{T}|) + \delta}$$

  其概率至少为

$$1 - \exp\left(Cs - \min\left\{\frac{u^2}{(\lambda^{\mathit{Off}}_{\max,s})^2\sum_{t\in\mathcal{T}} W_t^2 + |\mathcal{T}|\lambda^2_{\max}\|[\boldsymbol{\beta}^*]_{\mathcal{S}^*\setminus\mathcal{S}_\mathcal{T}}\|^2}, \right.\right.$$
$$\left.\left.\frac{u}{\lambda^{\mathit{Off}}_{\max,s}\max_{t\in\mathcal{T}} W_t + \lambda_{\max}\|[\boldsymbol{\beta}^*]_{\mathcal{S}^*\setminus\mathcal{S}_\mathcal{T}}\|}\right\}\right)$$
$$- \exp\left(-c\min\left\{\delta, \sqrt{\frac{\delta|\mathcal{T}|}{|\mathcal{S}^* \setminus \mathcal{S}_\mathcal{T}|\log(2d/|\mathcal{S}^* \setminus \mathcal{S}_\mathcal{T}|)}}\right\}\right).$$

- （无关坐标的上界）对于 $\mathcal{S}^{*c} \cap \mathcal{S}_\mathcal{T}^c$ 中的坐标，

$$\|[\mathbf{G}(\{\boldsymbol{\beta}_t, \mathbf{X}_t, Y_t\}_{t\in\mathcal{T}})]_{\mathcal{S}^{*c}\cap\mathcal{S}_\mathcal{T}^c}\|_\infty \le \lambda^{\mathit{Off}}_{\max,1}\sum_{t\in\mathcal{T}} V_t + u + C\sigma\sqrt{\lambda_{\max}}\sqrt{\delta|\mathcal{T}|}$$

  概率至少为

$$1 - d\exp\left(-\min\left\{\frac{u^2}{(\lambda^{\mathit{Off}}_{\max,1})^2\sum_{t\in\mathcal{T}} V_t^2}, \frac{u}{2\lambda_{\max,1}\max_t V_t}\right\}\right) - d\exp\Big(-\min\big\{\delta, \sqrt{|\mathcal{T}|\delta}\big\}\Big).$$

引理 2 表明，**G统计量** 对应于缺失的真实支持坐标（$\mathcal{S}^* \backslash \mathcal{S}_{\mathcal{T}}$）以 $\Theta(|\mathcal{T}|)$ 的阶增长，而非活跃坐标（$\mathcal{S}^{*c} \cap \mathcal{S}_{\mathcal{T}}^c$）上的值则要小得多。这种分离对于正确识别缺失的支持元素至关重要。为保证这些界之间存在非平凡的差距，从而确保可检测的分离，我们要求协变量满足以下条件。

假设 4. 协变量$\mathbf{X}_t$ 满足

$$\lambda_{\min} \geq C\sqrt{s}\lambda_{\max,1}^{Off} \quad and \quad \lambda_{\min} \geq C\lambda_{\max,s}^{Off},$$

其中，常数 $(\lambda_{\min}, \lambda_{\max,1}^{Off}, \lambda_{\max,s}^{Off})$ 满足 $\min_{\mathcal{S}\subseteq[d],\, 1\leq|\mathcal{S}|\leq s} \lambda_{\min}(\mathbb{E}\{[X_t]_{\mathcal{S}}[X_t]_{\mathcal{S}}^\top|\mathcal{F}_t\}) \geq \lambda_{\min}$, $\max_{\mathcal{S}'\subseteq[d]:\, |\mathcal{S}'|\leq s} \max_{\mathcal{S}\subseteq[d]:|\mathcal{S}|\leq s} \|[X_t]_{\mathcal{S}'}[X_t]_{\mathcal{S}\backslash\mathcal{S}'}^\top|\mathcal{F}_t\|_{\Psi_1} \leq \lambda_{\max,s}^{Off}$并且 $\max_{i\notin\mathcal{S}^*} \max_{\mathcal{S}\subseteq[d]:|\mathcal{S}|\leq s} \|[X_t]_i[X_t]_{\mathcal{S}\backslash i}^\top|\mathcal{F}_t\|_{\Psi_1} \leq \lambda_{\max,1}^{Off}$。

假设 4 要求 $\mathbf{X}_t$ 的不同条目之间的相关性上界为 $\lambda_{\min}$；事实上，这等价于经典离线稀疏线性回归中的条件（Zhang, 2010）。引理 1为更新规则（$\{V_t\}$ 和 $\{W_t\}$）下的序列提供了有效的取值。具体而言，

$$W_t := \frac{\sqrt{C^*}}{\sqrt{\lambda_{\min}}}\sqrt{\frac{s\log(2d/s)+\delta_t}{t+C_b s\log(2d/s)}}\sigma + C\left(\frac{\lambda_{\max,s}^{\text{Off}}}{\lambda_{\min}}\right)\cdot\|[\boldsymbol{\beta}^*]_{\mathcal{S}^*\backslash\mathcal{S}_0}\|$$

and

$$V_t := \frac{\sqrt{C^*}}{\sqrt{\lambda_{\min}}}\sqrt{\frac{s\log(2d/s)+\delta_t}{t+C_b s\log(2d/s)}}\sigma + C\left(\frac{\lambda_{\max,s}^{\text{Off}}}{\lambda_{\min}}\right)\cdot\|[\boldsymbol{\beta}^*]_{\mathcal{S}^*\backslash\mathcal{S}_0}\| + \|[\boldsymbol{\beta}^*]_{\mathcal{S}^*\backslash\mathcal{S}_0}\|.$$

将 $\mathcal{S}_{\mathcal{T}} = \mathcal{S}_0$、 $\mathcal{T} = \{0, 1, \ldots, t\}$以及上述界代入引理 2 即可得到以下推论。

推论 1. 假设 2、3和4 均成立。取 $\delta_l = C\log(l) + \log(d/s)$，并假设定理 1 中的估计误差动态成立。则对于任意满足以下条件的 t

$$\frac{t}{\log(t)} \geq C|\mathcal{S}^* \setminus \mathcal{S}_0|\log\left(\frac{2d}{|\mathcal{S}^* \setminus \mathcal{S}_0|}\right)\cdot\frac{\sigma^2}{\lambda_{\min}\|[\boldsymbol{\beta}^*]_{\mathcal{S}^*\backslash\mathcal{S}_0}\|^2},$$

以至少 $1 - C(td/s)^{-100}$ 的概率，

$$\left\|[\mathbf{G}(\{\boldsymbol{\beta}_t, \mathbf{X}_t, Y_t\}_{t\in\mathcal{T}})]_{\mathcal{S}^*\backslash\mathcal{S}_0}\right\| > \sqrt{|\mathcal{S}^* \setminus \mathcal{S}_0|}\cdot\left\|[\mathbf{G}(\{\boldsymbol{\beta}_t, \mathbf{X}_t, Y_t\}_{t\in\mathcal{T}})]_{\mathcal{S}^{*c}\cap\mathcal{S}_0^c}\right\|_\infty,$$

这进而意味着 $\operatorname{argmax}_{i\in\mathcal{S}_{\mathcal{T}}^c}\left|[\textbf{G统计量}(\{\boldsymbol{\beta}_t, \mathbf{X}_t,\ Y_t\}_{t\in\mathcal{T}})]_i\right| \in \mathcal{S}^* \setminus \mathcal{S}_0$。

推论1 表明，一旦 t 足够大，统计量**G统计量**就能可靠地区分真实支撑集中缺失的坐标，从而实现准确的支撑集恢复。接下来，我们将这一选择机制与估计更新相结合。

3.3 稀疏随机梯度下降：估计与支撑集恢复

直观而言，每个新到的观测都有助于降低估计误差，但单个观测不足以可靠地进行变量选择。相比之下，通过汇总统计量**G统计量**聚合历史数据中的信息，我们便能在样本量足够大时识别缺失的支持元素。这启发我们采用一种迭代过程，在以下两步之间交替进行：(i) 在固定支集上进行估计；(ii) 使用**G统计量**扩展支撑集。下面我们将介绍一种以在线方式整合估计与支撑集恢复的算法。

算法。在时刻 $t \in \{0, 1, \dots, T\}$，其中 T 可以是 $+\infty$，选择更新时间 $\tau_0 = 0$、$0 < \tau_1 < \cdots < \tau_\alpha$。在每个区间 $[\tau_{l-1}, \tau_l - 1]$ 内，迭代量 $\{\boldsymbol{\beta}_t\}$ 共享共同支撑集 $\mathcal{S}_l$，并按照稀疏更新规则演化

$$\boldsymbol{\beta}_{t+1} = \mathcal{H}_{\mathcal{S}_l}(\boldsymbol{\beta}_t - \eta_t \cdot \mathbf{g}_t),$$

其中 $\mathbf{g}_t = (\mathbf{X}_t^\top \boldsymbol{\beta}_t - Y_t)\mathbf{X}_t$ 表示随机梯度，而 $\mathcal{H}_{\mathcal{S}_l}$ 是将向量投影到支撑集 $\mathcal{S}_l$ 上的硬阈值算子。支撑集序列满足 $\mathcal{S}_0 \subseteq \mathcal{S}_1 \subseteq \cdots \subseteq \mathcal{S}_\alpha$，在阶段 l 的更新通过以下方式执行

$$\mathcal{S}_l = \mathcal{S}_{l-1} \cup \left\{ \operatorname{argmax}_{i \in \mathcal{S}_{l-1}^c} \left| [\mathbf{G}(\{\boldsymbol{\beta}_t, \mathbf{X}_t, Y_t\}_{t \in \mathcal{T}_l})]_i \right| \right\},$$

其中 $\mathcal{T}_l \subseteq [t]$ 表示用于在步骤 l 计算统计量 $\mathbf{G}$ 的时间索引集。$\{\tau_l\}$ 和 $\{\mathcal{T}_l\}$ 的选择并不唯一，下面将对此进行详细讨论。对于索引集 $\mathcal{T}_l$，我们考虑两种自然的选择：（1）局部窗口，$\mathcal{T}_l := \{\tau_{l-1}, \dots, \tau_l - 1\}$；（2）累积窗口，$\mathcal{T}_l := \{0, 1, \dots, \tau_l - 1\}$。从另一个角度来看，如果每次更新都使 $\mathcal{S}^* \setminus \mathcal{S}_0$ 中的支撑集恰好扩大一个元素，那么完全恢复真实支撑集最多需要 $|\mathcal{S}^* \setminus \mathcal{S}_0|$ 次。以下命题刻画了这一集成更新方案下的估计误差动态。

命题 1. 假设 *2*、*3* 和 *4* 成立。设步长方案为 $\eta_t = \frac{C_a/\lambda_{\min}}{t + C_b s}$ *log(*$2d/s$*)*，其中常数满足 ~~$C_a \geq 5$~~ *and* $C_b \geq C(C_a \lambda_{\max}/\lambda_{\min})^2$。另外，假设对所有 $l \in \{1, 2, \dots, \alpha\}$，$\mathcal{S}_l \setminus \mathcal{S}_{l-1} \subseteq \mathcal{S}^*$ 均成立，且尾概率序列 $\{\delta_l\}$ 满足定理 *1* 中的条件。那么，对于任意 $t \in [\tau_l, \tau_{l+1} - 1]$，以至少 $1 - \sum_{l=0}^{t} \exp(-c\{s \log(2d/s) + l/C_a\}) - \sum_{l=0}^{t} \exp(-c\{s \log(2d/s) + \delta_l\})$ 的概率，以下界成立：

$$\begin{aligned} \|[\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*]_{\mathcal{S}_l}\|^2 \leq C \frac{s \log(2d/s) + \delta_{t+1}}{t + 1 + C_b s \log(2d/s)} \frac{\sigma^2}{\lambda_{\min}} + C \left(\frac{\lambda_{\max,s}^{Off}}{\lambda_{\min}} \right)^2 \|[\boldsymbol{\beta}^*]_{\mathcal{S}^* \setminus \mathcal{S}_l}\|^2 \\ + C \sum_{i=1}^{l} \left(\frac{\tau_i + C_b s \log(2d/s)}{t + 1 + C_b s \log(2d/s)} \right)^{C_a - 2} \cdot \|[\boldsymbol{\beta}^*]_{\mathcal{S}_i \setminus \mathcal{S}_{i-1}}\|^2. \end{aligned}$$

命题 1 刻画了支撑集恢复与估计精度之间的相互关系。如果$\lambda_{\max,\ s}^{\mathsf{Off}} = 0$，或者当前支撑集已包含真实支撑集（即 $\mathcal{S}^* \subseteq \mathcal{S}_t$），则界中的第二项消失。第三项揭示了一项关键设计原则：及早且准确地纳入缺失的支撑集坐标，对于将整体估计误差降至最低至关重要。然而，推论 1 表明，准确的支撑集检测要求窗口 $\mathcal{T}_l$内有足够大的样本量。因此，支撑集更新的时机必须兼顾两个相互制约的因素。更新过晚会减缓收敛，因为未选中的真实坐标得不到校正；而更新过早则可能在统计量**G统计量** 实现可靠区分之前选中错误变量。结合命题 1 与引理 2，我们现在得到同时估计与支撑集恢复的综合保证。

定理 3. 假设2、3和4 成立。令步长方案为 $\eta_t = \frac{C_a/\lambda_{\min}}{t+C_b s}\log(2d/s)$，其中 $C_a \geq 5$ 且 ~~$C_b \geq C(C_a\lambda_{\max}/\lambda_{\min})^2$~~。假设初始估计满足 $\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|^2 \leq \sigma^2/\lambda_{\min}$，并令 $s_+ := |\mathcal{S}^* \setminus \mathcal{S}_0|$。假设支撑集更新时间 $\{\tau_l\}$ 和索引集 $\{\mathcal{T}_l\}$ 满足：

1. $\tau_1/\log(\tau_1) \geq Cs_+\log(d/s_+) \cdot \frac{\sigma^2}{\lambda_{\min}\|[\boldsymbol{\beta}^*]_{\mathcal{S}^*\setminus\mathcal{S}_0}\|^2}$;
2. 对于每个 $l \in \{2, \ldots, s_+\}$，令 $s_l := |\mathcal{S}^* \setminus \mathcal{S}_l|$。

   (a) 若 $\mathcal{T}_l = \{\tau_{l-1}, \ldots, \tau_l - 1\}$（局部窗口），则要求

$$\tau_l \geq \max\left\{\left(\frac{C}{\tau_{l-1}^{C_a-4}\|[\boldsymbol{\beta}^*]_{\mathcal{S}^*\setminus\mathcal{S}_l}\|^2}\sum_{i=1}^{l-1}\tau_i^{C_a-2}\|[\boldsymbol{\beta}^*]_{\mathcal{S}_i\setminus\mathcal{S}_{i-1}}\|^2\right)^{\frac{1}{2}}, \tau_{l-1} + \frac{Cs_l\log(2d/s_l)\cdot\sigma^2}{\lambda_{\min}\|[\boldsymbol{\beta}^*]_{\mathcal{S}^*\setminus\mathcal{S}_l}\|^2}\right\},$$

   (b) 如果 $\mathcal{T}_l = \{0, 1, \ldots, \tau_l - 1\}$（累积窗口），则要求

$$\tau_l \geq \max\left\{\frac{C}{\|[\boldsymbol{\beta}^*]_{\mathcal{S}^*\setminus\mathcal{S}_l}\|}\sum_{i=1}^{l-1}\tau_i\|[\boldsymbol{\beta}^*]_{\mathcal{S}_i\setminus\mathcal{S}_{i-1}}\|, \frac{Cs_l\log(2d/s_l)\cdot\sigma^2}{\lambda_{\min}\|[\boldsymbol{\beta}^*]_{\mathcal{S}^*\setminus\mathcal{S}_l}\|^2}\right\}.$$

那么，以至少 $1 - Cd^{-100}$ 的概率，对于所有 $t \in [\tau_{s_0}, +\infty]$，我们有 $\mathcal{S}^* \subseteq \mathcal{S}_t$，且估计误差满足

$$\begin{aligned}\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\|^2 \leq\ & C^*\frac{s\log(2d/s)}{t+1+C_b s\log(2d/s)}\frac{\sigma^2}{\lambda_{\min}} \\ & + C\sum_{i=1}^{s_0}\left(\frac{\tau_i + C_b s\log(2d/s)}{t+1+C_b s\log(2d/s)}\right)^{C_a-2}\|[\boldsymbol{\beta}^*]_{\mathcal{S}_i\setminus\mathcal{S}_{i-1}}\|^2.\end{aligned}$$

定理3 表明，一旦支撑集被完全恢复，估计误差便以统计最优速率衰减。由于指数 $C_a - 2$的作用，早期支撑集不匹配阶段的贡献会迅速减小，长期性能则达到相关数据下稀疏线性回归的极小极大最优速率。在实践中，每当

统计量**G 统计量**在当前支撑集之外的坐标上呈现明显的 $O(t)$ 增长，这与推论1和定理3 所示一致。定理3在第5节中得到了数值验证。为此，我们的方法为现有方法（如在线 LASSO 算法（Yang 等人，2023 年；Han 等人，2024b））提供了一种计算效率高的替代方案；这些方法需要求解正则化优化问题，并在每一步维护 $O(d^2)$ 汇总统计量。

4 相关数据下的上下文线性赌博机

现在我们转向在线上下文线性老虎机问题。与回归情形相比，这个问题引入了一个额外且根本性的依赖来源：除了数据流$(\mathbf{X}_t,\ \xi_t)$中固有的依赖关系外，观测序列现在还通过决策策略自适应地生成。在时间 t 采取的每个动作都依赖于当前估计值，而当前估计值又取决于所有过去的结果。这一反馈回路形成了双重依赖结构：协变量和噪声中的时间依赖，以及探索-利用决策所导致的自适应依赖性。本节的目标是说明，所提出的随机梯度下降框架足够稳健，能够处理这种完整的依赖结构。具体而言，将依赖数据线性回归中开发的相同技术与 ε-贪心探索策略相结合，可以同时实现 (i) 每个臂的统计最优估计误差，以及 (ii) 统计最优遗憾。这将针对独立同分布协变量和噪声的经典分析（Gold- enshluger 和 Zeevi，2013 年；Gao 等人，2019年；Bastaniand Bayati，2020年；Bastani 等人，2021 年；Chen 等人，2021a，2022；Duan 等人，2024年）拓展到了广泛得多的相关数据情形。

在每个时间点 t，决策者都会观测到一个协变量向量$\mathbf{X}_t$，并必须从K 个臂中选择一个。每个臂 $i \in [K]$ 都由一个未知向量 $\boldsymbol{\beta}_i^*$进行参数化。选择臂 i 会获得奖励 $Y_t = \mathbf{X}_t^\top \boldsymbol{\beta}_i^* + \xi_t$，其中 ξ_t 表示噪声项，且允许其具有依赖性。令 a_t表示在时间点 t选择的臂。在一个（可能是无限的）时域 T内，目标是最大化累积奖励，或等价地，最小化遗憾

$$\mathsf{Regret}(T) := \mathbb{E}\left\{\sum_{t=1}^{T}\left(\max_{i\in[K]} \mathbf{X}_t^\top \boldsymbol{\beta}_i^* - \mathbf{X}_t^\top \boldsymbol{\beta}_{a_t}^*\right)\right\}. \tag{6}$$

等价地，目标是学习一个针对 a_t 的决策规则，使遗憾最小化；或者确定预言机决策区域，该区域将 $\mathbf{X}_t$ 映射到最优臂。尽管这一目标不同于最小化单个时间点的估计误差（如第 2 节所述），我们将证明，我们的方法能够同时实现这两个目标：对每个臂都能获得统计最优的估计误差和统计最优的遗憾。为此，我们考虑将随机梯度下降与 ε-贪心探索机制自然地结合起来。算法如下。

算法 令 $\{\boldsymbol{\beta}_i^{(0)}\}$ 为任意初始估计值，并令 $\{\eta_t\}$ 和 $\{\pi_t\}$ 分别表示步长序列和探索概率序列。在时刻 t，我们观测到协变量 $\mathbf{X}_t$，并抽取 $\alpha_t \sim$ 《伯努利》(π_t)。若 $\alpha_t = 0$（利用步骤），则拉取使估计奖励最大的臂，

$$a_t \in \arg\max_{i\in[K]} \mathbf{X}_t^\top \boldsymbol{\beta}_i^{(t)}.$$

如果 $\alpha_t = 1$（探索步骤），我们就均匀随机地选择一条臂：

$$a_t \sim \mathrm{Unif}\{1, 2, \ldots, K\}.$$

拉取臂 后，我们观测到奖励 Y_t，并使用随机梯度下降仅更新所选臂：

$$\boldsymbol{\beta}_i^{(t+1)} = \boldsymbol{\beta}_i^{(t)} - \eta_t \, \mathbb{I}_{\{a_t=i\}} \cdot (\mathbf{X}_t^\top \boldsymbol{\beta}_i^{(t)} - Y_t) \, \mathbf{X}_t, \quad i \in [K]. \tag{7}$$

注意，$\{\pi_t\}$ 和 $\{\eta_t\}$ 尚待指定。

记为 $\mathcal{F}_t = \sigma$（Y_{t-1}，α_t，$\mathbf{X}_{t-1}$，$\ldots$，Y_0，α_0，$\mathbf{X}_0$）。显然，$\boldsymbol{\beta}_i^{(t)} \in \mathcal{F}_t$。定义增广的 σ-域 $\mathcal{F}_t^+ = \sigma$（$\mathbf{X}_t$，Y_{t-1}，α_t，$\mathbf{X}_{t-1}$，$\ldots$，Y_0，α_0，$\mathbf{X}_0$），使得 $\mathcal{F}_t \subseteq \mathcal{F}_t^+ \subseteq \mathcal{F}_{t+1}$。在时刻 t，算法以概率π_t从 $[K]$中均匀随机地选择一条臂；这构成探索步骤，用于获取所有臂的信息。算法以概率 $1 - \pi_t$选择基于过去数据的估计奖励最大的臂；这就是利用步骤。探索率 π_t 决定了基本的权衡：充分探索会带来最小的估计误差，但遗憾较大；而纯粹利用则有可能导致算法选错臂。因此，π_t 会同时影响每条 $\boldsymbol{\beta}_i^*$ 的估计精度和累积遗憾。在这项工作中，我们允许采用广泛的探索计划，包括通过函数定义 π_t 的选择，具体形式见定义3。步长序列 $\{\eta_t\}$ 是另一个尚待指定的算法参数。

注释 *3*（$K = 2$ 与 $K \geq 2$）。本文始终考虑 $K \geq 2$。重要的是，多臂情形 $K > 2$ 并非双臂设置的简单扩展。当 $K = 2$时，每个臂都在协变量的某个非空区域内是最优臂。相反，当 $K \geq 3$时，某些臂可能整体上并非最优臂，即对于任何协变量都无法获得最高奖励。例如，令 $\boldsymbol{\beta}_1^* = (1, 0)^\top$，$\boldsymbol{\beta}_2^* = (0.1, 0)^\top$，$\boldsymbol{\beta}_3^* = (-1, 0)^\top$，$\boldsymbol{\beta}_4^* = (0, 1)^\top$，$\boldsymbol{\beta}_5^* = (0, -1)^\top$。那么，对于每个 $\boldsymbol{X} \in \mathbb{R}^2$，臂 *2* 都不是最优臂。这类非最优臂会带来额外的复杂性，而在 $K = 2$ 时则不会出现这种情况（*Chen* 等人，*2022*，*2021a*）。

4.1 决策区域刻画

本节介绍臂特定决策区域的锥形近似。对于臂 i，预言机决策区域定义为

$$\mathcal{U}_i^* = \left\{ \mathbf{X} \in \mathbb{R}^d : \ \mathbf{X}^\top \boldsymbol{\beta}_i^* > \max_{j \neq i} \mathbf{X}^\top \boldsymbol{\beta}_j^* \right\}.$$

不难验证，$\mathcal{U}_i^*$要么是一个锥，要么是空集。事实上，如果$\mathbf{x} \in \mathcal{U}_i^*$，那么对于任意正标量 a，都有 $a\mathbf{x} \in \mathcal{U}_i^*$。为了最小化遗憾，我们需要估计并近似预言机区域 $\mathcal{U}_i^*$。一系列具有影响力的研究（Goldenshluger and Zeevi，2013 年；Wang 等人，2018；Bastani and Bayati，2020年；Hao 等人，2020年；Chen 等人，2020年；Bastani 等人，2021 年）假设协变量位于有界域内，并使用由分离超平面构造的有界多面体区域来近似 $\mathcal{U}_i^*$。具体而言，他们考虑

$$\widetilde{\mathcal{U}}_i(h; D) := \left\{ \|\mathbf{X}\| \leq D : \ \mathbf{X}^\top \boldsymbol{\beta}_i^* - \max_{j \neq i} \mathbf{X}^\top \boldsymbol{\beta}_j^* \geq h \right\}. \tag{8}$$

~集合 $\mathcal{U}_i(h; D)$ 排除了原点的整个邻域。为解决这一局限性，我们提出了另一种刻画方式，将其称为决策区域的锥形近似：

$$\mathcal{U}_i(h) := \left\{ \mathbf{X} \in \mathbb{R}^d : \ \frac{\mathbf{X}^\top \boldsymbol{\beta}_i^* - \max_{j \neq i} \mathbf{X}^\top \boldsymbol{\beta}_j^*}{\|\mathbf{X}\|} \geq h \right\}. \tag{9}$$

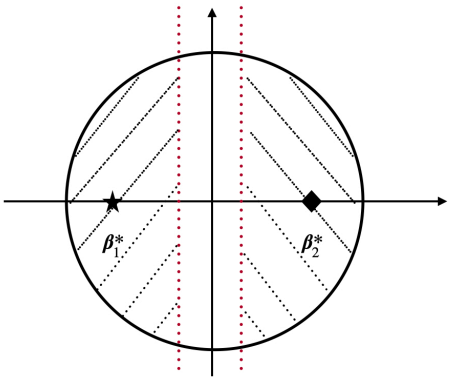
区域 $\mathcal{U}_i(h)$ 同样是锥形区域或空集，体现了 $\mathcal{U}_i^*$ 的几何结构。此外，族 $\mathcal{U}_i(h)$ 自然适用于无界的协变量支撑集，并保留了原点附近邻域的覆盖范围。这一特性在协变量分布高度集中于原点附近的情形中尤其有利；在这种情况下，$\mathcal{U}_i(h; D)$ 可能会因移除整个原点而舍弃大量概率质量。相比之下，当协变量取值来自离散集时，这两种构造 $\mathcal{U}_i(h)$ 和 $\mathcal{U}_i(h; D)$ 是一致的。为展示几何差异，图1展示了有界近似 $\mathcal{U}_i(h; D)$ 和锥形近似 $\mathcal{U}_i(h)$。下一个引理将正式刻画 $\mathcal{U}_i^*$、$\mathcal{U}_i(h)$ 和 $\mathcal{U}_i(h; D)$ 之间的关系，并确立所提出的锥形区域的若干关键几何性质。

引理3. 对于任意 $h > 0$ 和 $D > 0$，以下集合关系成立：

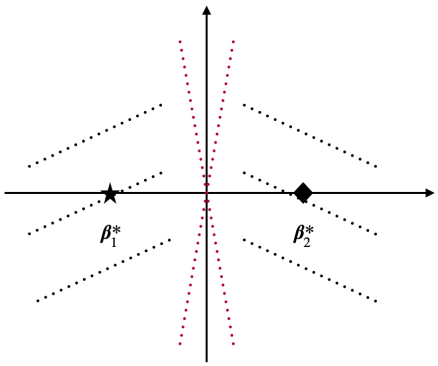
~ ~*(i)* 对于 $\mathcal{U}_i(h; D)$，它具有 $\mathcal{U}_i(h; D) \subsetneq \mathcal{U}_i(h/D) \cap \{\mathbf{X} : \|\mathbf{X}\| \leq D\} \subsetneq \mathcal{U}_i^* \cap \{\mathbf{X} : \|\mathbf{X}\| \leq D\}$。

~*(ii)* 反之，它具有 $\mathcal{U}_i(h) \cap \{\mathbf{X} : \|\mathbf{X}\| \leq D\} \not\subseteq \mathcal{U}_i(hD; D)$。

(iii) 此外，$\mathcal{U}_i(h) \subseteq \mathcal{U}_i^*$，特别是 $\mathcal{U}_i(0)$ 是 $\mathcal{U}_i^*$ 的闭包。对于任意 $h_1 > h_2$，它有 $\mathcal{U}_i(h_1) \subseteq \mathcal{U}_i(h_2)$。



(a) 区域 $\widetilde{\mathcal{U}}_i$



(b) $\mathcal{U}_i$的区域

图1：预言机决策区域 $\mathcal{U}_i^*$的两种近似。左图展示了~有界多面体近似$\mathcal{U}_i$，即由若干线段围成的阴影区域。右图展示了锥形近似 $\mathcal{U}_i$，它构成一个无界的尺度不变区域。

此外，如果 $\mathcal{U}_i^* \neq \emptyset$，那么对于任意

$$h \in \left(0, \min_{i\in\mathcal{A}} \sup_{\|\mathbf{z}\|=1} (\mathbf{z}^\top \boldsymbol{\beta}_i^* - \max_{j\neq i} \mathbf{z}^\top \boldsymbol{\beta}_j^*)\right),$$

我们有 $\mathcal{U}_i(h) \neq \emptyset$。此外，若 $\mathcal{U}_j^* = \emptyset$，则

$$\mathbf{X}^\top \boldsymbol{\beta}_j^* \leq \max_{i\neq j} \mathbf{X}^\top \boldsymbol{\beta}_i^* \quad \textit{for all} \quad \mathbf{X} \in \mathbb{R}^d.$$

~证明。我们仅证明(i)部分。为简洁起见，记 $\mathcal{X}$:= $\{\mathbf{X}: \|\mathbf{X}\| \leq D\}$。显然， $\mathcal{U}_i(h;\ D) \subsetneq \mathcal{X}$。我们证明 $\widetilde{\mathcal{U}}_i(h;\ D) \subsetneq \mathcal{U}_i(h/D)$。对于任意$\mathbf{x} \in \widetilde{\mathcal{U}}_i(h;\ D)$， $\mathbf{x}^\top \boldsymbol{\beta}_i^* - \max_{j\neq i} \mathbf{x}^\top \boldsymbol{\beta}_j^* \geq h \geq \frac{h}{D}\|\mathbf{x}\|$，这意味着$\widetilde{\mathcal{U}}_i(h;\ D) \subset \mathcal{U}_i(h/D)$。反过来，向量$\frac{h}{2}(\mathbf{x}^\top \boldsymbol{\beta}_i^* - \max_{j\neq i} \mathbf{x}^\top \boldsymbol{\beta}_j^*) \cdot \mathbf{x}$属于 $\mathcal{U}_i(h/D)$，但不属于~$\mathcal{U}_i(h;\ D)$，由此完成证明： $\mathcal{U}_i(h;\ \mathcal{X})$严格包含于 $\mathcal{X} \cap \mathcal{U}_i(h)$。□

接下来，令 $\boldsymbol{\beta}_i$ 表示 $\boldsymbol{\beta}_i^*$ 的估计量。相应的经验决策区域在（7）中为

$$\mathcal{X}\left(i, \{\boldsymbol{\beta}_j\}_{j\in[K]}\right) := \left\{\mathbf{X} \in \mathbb{R}^d : \ \mathbf{X}^\top \boldsymbol{\beta}_i > \max_{j\neq i} \mathbf{X}^\top \boldsymbol{\beta}_j\right\}. \tag{10}$$

在利用阶段，只要 $\mathbf{X}_t \in \mathcal{X}(i,\ \{\boldsymbol{\beta}_j^{(t)}\}_{j\in[K]})$，算法就会选择臂 i。重要的是， $\mathcal{X}(i,\ \{\boldsymbol{\beta}_j\}_{j\in[K]})$ 是一个锥形区域：如果 $\mathbf{X}$ 属于该集合，那么对于任意 $a > 0$， $a\mathbf{X}$ 也属于该集合。下一个结果将形式化描述经验区域与预言机近似 $\mathcal{U}_i$之间的关系。

引理 4。 如果 $\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^*\| \le h_0$ 对所有 $i \in [K]$ 都成立，那么有

$$\mathcal{U}_i(2h_0) \subseteq \mathcal{X}\Big(i, \{\boldsymbol{\beta}_j\}_{j\in[K]}\Big) \subseteq \mathcal{U}_i(-2h_0).$$

引理4 将更新规则（7）中使用的经验决策区域与预言机区域的锥形近似 $\mathcal{U}_i^*$联系起来，并明确揭示了其对估计误差 $\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^*\|$的可解释依赖关系。这一结果是我们有限样本分析和渐近分析中的关键结构性组成部分。它确保足够准确的参数估计能够产生正确近似真实最优区域的决策区域，从而在统一框架内同时控制遗憾和估计误差。

上下文老虎机中最优臂与次优臂之间的经典可分离性概念（BastaniandBayati，2020年；Bastani 等人，2021 年）通过有界域近似来表述，例如~，如 $\mathcal{U}_i$所示。接下来的假设将这一可分离性条件推广到尺度不变区域 $\mathcal{U}_i$（h）。

假设5（臂最优性）。存在 $h > 0$ 和一个划分 $\mathcal{A} \cup \mathcal{A}^c = [K]$，使得对于所有 t 和 $i \in \mathcal{A}$， $\mathbb{E}$（$\mathbf{X}_t\mathbf{X}_t^\top \cdot \mathbb{I}\{\mathbf{X}_t \in \mathcal{U}_i$（$h$）$\} \,|\, \mathcal{F}_t) \succeq \lambda_{\min} \cdot \mathbf{I}_d$，且对于每个 $j \in \mathcal{A}^c$，

$$\mathbb{P}\left(\mathbf{X}_t : \frac{\max_{i\in[K]} \mathbf{X}_t^\top \boldsymbol{\beta}_i^* - \mathbf{X}_t^\top \boldsymbol{\beta}_j^*}{\|\mathbf{X}_t\|} \ge h \middle| \mathcal{F}_t\right) = 1.$$

注释 *4*（关于最优区域的刻画）。假设*5*采用我们对最优决策区域的尺度不变定义，即 *(9)* 中的 $\mathcal{U}_i$ *(h)*。这与上下文老虎机文献中的传统表述存在细微但重要的差异。例如，*Bastani and Bayati*（*2020*）和 *Bastani* 等人（*2021*）等有影响力的研究采用有界域刻画，这对应于我们在 *(8)* 中构造的 $\mathcal{U}_i(h; D)$。这类有界域近似适用于协变量位于紧集内的情形，但也有两个局限。第一，它们要求预先知道有界支撑集，而在许多现代应用中，协变量自然无界，因此这一假设可能不成立。第二，它们会排除原点周围的整个邻域。对于表现良好的设计而言，这种排除可能会带来问题，例如多元高斯协变量的概率质量高度集中在原点附近。在这些情形下，移除零点附近的区域可能会排除信息性设计空间中的相当大一部分，从而降低适用性或统计效率。我们的 $\mathcal{U}_i(h)$ 定义避免了这些问题。按照构造， $\mathcal{U}_i(h)$ 是尺度不变的（即一个锥），因此非常适合无界协变量分布，同时保留预言机区域 $\mathcal{U}_i^*$ 的基本几何性质。值得注意的是， $\mathcal{U}_i(h)$ 包含原点，而非将其排除，从而确保不会丢弃高概率协变量质量中过大的区域

。因此，假设5 为上下文赌博机分析提供了更一般且适用范围更广的基础，使我们的理论保证能够适用于比有界域假设所涵盖的机制更一般的数据生成机制。

此外，若假设 5 对某个 h成立，则它也自动对任意 $h' \in (0, h)$成立。特别地，在 $K = 2$的情形下，该假设总是成立，因此无需额外条件。现在，我们将着手建立该问题的正则性性质，这些性质是分析所提出方法收敛动态的关键组成部分。这些性质背后的直觉部分受到 Bastani and Bayati (2020年) 的启发。值得注意的是，我们的框架既不要求 $\mathbf{X}$ 有界，也不要求参数 $\boldsymbol{\beta}^*$有界，同时也不依赖于 $\{\mathbf{X}_t\}$的独立同分布假设。

引理5 (正则性性质)。 假设假设 5 在 $h > 0$下成立。那么以下性质成立。

1. 若估计值满足 $\{\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^*\| \le \frac{h}{2},\ i \in [K]\}$，则以概率*1*，对任意 $j \in \mathcal{A}^c$,$\mathcal{X}(j, \{\boldsymbol{\beta}_i\}_{i\in[K]}) = \emptyset$, 即,

$$\operatorname{argmax}_{i\in[K]} \mathbf{X}_t^\top \boldsymbol{\beta}_i \subseteq \mathcal{A}, \quad \text{a.s.}$$

此外，对于每个 $i \in \mathcal{A}$,

$$\mathcal{U}_i\Big(2\max_{j\in\mathcal{A}} \|\boldsymbol{\beta}_j - \boldsymbol{\beta}_j^*\|\Big) \subseteq \mathcal{X}\Big(i, \{\boldsymbol{\beta}_j\}_{j\in[K]}\Big) \subseteq \mathcal{U}_i\Big(-2\max_{j\in\mathcal{A}} \|\boldsymbol{\beta}_j - \boldsymbol{\beta}_j^*\|\Big).$$

2. 对于任意 $i \in \mathcal{A}$ 和任意 $\boldsymbol{\beta}_i \in \mathcal{F}_t$,

$$\mathbb{E}\Big\{\mathbb{I}\Big\{\mathbf{X}_t^\top \boldsymbol{\beta}_i \ge \max_j \mathbf{X}_t^\top \boldsymbol{\beta}_j\Big\} \cdot \Big(\mathbf{X}_t^\top \boldsymbol{\beta}_i - \mathbf{X}_t^\top \boldsymbol{\beta}_i^*\Big)^2 \,\Big|\, \mathcal{F}_t\Big\}$$
$$\ge \mathbb{I}\Big\{\max_j \|\boldsymbol{\beta}_j - \boldsymbol{\beta}_j^*\| \le \frac{h}{2}\Big\} \cdot \lambda_{\min} \|\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^*\|^2. \quad (11)$$

引理5表明，当估计误差满足 $\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^*\| \le \frac{h}{2}$ 时，(11) 中的量在 $\|\boldsymbol{\beta}_i = \boldsymbol{\beta}_i^*\|$上具有一个紧致的二次下界。此外，与引理3不同，在假设5和相同误差条件下，经验决策区域的近似仅取决于 $\mathcal{A}$ 中各臂的估计精度。从技术上讲，引理5对于使利用阶段的观测能够有效地用于参数细化起着核心作用。当 $h' < h$ 时，与 $\mathcal{U}_i(h)$ 相比，区域 $\mathcal{U}_i(h')$ 对预言机区域 $\mathcal{U}_i^*$ 的近似更精确，并且并集 $\cup_{i\in[K]}\mathcal{U}_i(h')$ 覆盖了 $\mathbb{R}^d$ 的更大子集。然而，该引理也表明，要在 (11) 中获得严格为正的二次下界，就必须将估计误差控制在与 h' 成比例的尺度上。因此，减小 h 会扩大决策区域，但同时也要求参数估计更加精确，才能维持 (11) 中的曲率下界。这体现了选择 h 时固有的权衡：增加概率质量 $\mathbb{P}(\mathcal{U}_i)$，代价是削弱所保证的二次下界。

4.2 非渐近性能

本节中，我们研究所提出方法的非渐近统计性能。我们方法的一个显著特点是，与 $\mathcal{A}$ 中各臂相关的参数估计误差和累积遗憾均能达到统计最优速率，同时只需极少的存储，具体而言，只需存储最新的估计值以及迄今收集到的观测总数。下面我们将详细介绍这一结果背后的收敛动态。

为清晰起见，我们在全文中假设初始估计满足 $\|\boldsymbol{\beta}_i^{(0)} - \boldsymbol{\beta}_i^*\|^2 \leq C\sigma^2/\lambda_{\max}$。通过离线过程或在线方法均可实现这样的初始化。例如，如定理 1所示，采用足够小的常数步长进行随机梯度下降，可使 $O(1)$估计误差线性收敛，从而满足上述要求。接下来，我们给出探索率 π_t 及其对应尾概率序列$\{\delta_t\}$的一般设定方法。下面的定理确立了由此得到的非渐近估计误差动态。

定理 4 (尾部水平 δ_t下的非渐近误差动态**)**。 假设 *1*、*2*和*5*均成立。考虑步长 $\eta_t = \frac{1}{\lambda_{\min}} \frac{C_a}{t+C_b d}$，并设 $\pi \in (0,1]$ 为任意常数。对于满足$\{\delta_t\}$ 的任意序列 (i) $0 \leq \delta_t \leq \delta_{t+1}$ 和 (ii) $\delta_t \leq Ct/C_b$，以下结论成立：

(1) 如果 $\pi_t \geq \pi$、 $C_a \geq 2K/\pi$ 和 $C_b \geq 3C_a^2$（$\lambda_{\max}^2/\lambda_{\min}^2$），那么以至少 $1 - K\sum_{l=0}^{t} \exp(-c\ (d + l/C_a)\ - K\sum_{l=0}^{t}\exp(-c\ (d+\delta_l)$ 的概率，对于所有 $i \in [K]$，我们都有：

$$\|\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i^*\|^2 \leq \frac{C^*}{\lambda_{\min}} \frac{d+\delta_t}{t+C_b d}\sigma^2 \quad with \quad C^* = CC_a^2\frac{\lambda_{\max}}{\lambda_{\min}} + C_b\frac{\lambda_{\min}}{\lambda_{\max}}. \tag{12}$$

(2) 对于任意 $t \geq t_1 :\ = 16C^* d\sigma^2/\ (\lambda_{\min} h^2)\ - C_b d$ 以及任意探索计划 $\pi_t \in [0,1]$，我们以至少 $1 - |\mathcal{A}|\sum_{l=t_1}^{t+1} \exp(-c\ \ (\ d + l/C_a\ )\ \ - |\mathcal{A}|\sum_{l=t_1}^{t+1} \exp(-c\ \ (\ d+\delta_l\ )$ 的概率，对于所有 $i \in \mathcal{A}$，都有：

$$\|\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i^*\|^2 \leq \frac{C^{**}}{\lambda_{\min}} \frac{d+\delta_t}{t+C_b d}\sigma^2 \quad with \quad C^{**} = C^* + 2CC_a\frac{\lambda_{\max}}{\lambda_{\min}}. \tag{13}$$

此外，以至少 $1 - |\mathcal{A}^c|\exp(-cd)\ \cdot\sum_{l=t_1}^{t+1} \mathbb{I}\{\pi_l \neq 0\}$ 的概率，对于所有 $i \in \mathcal{A}^c$，我们都有：

$$\|\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i^*\|^2 \leq \frac{h^2}{4}. \tag{14}$$

备注 5 （关于衰减探索及其与无探索老虎机的关系）。 定理 *4* （*2*）的一个关键含义是，即使探索率 $\mathcal{A}$衰减至零，最优臂（即 π_t 中的臂）的统计最优估计速率仍然可以达到。这为自适应或逐渐减小的探索调度提供了严格的理论依据，并将我们的框架与日益发展的无探索 老虎机算法文献联系起来。特别是，*Bastani*

等人 （*2021* 年）提出了一类“基本无探索”的算法，在协变量分布满足一种称为协变量多样性的几何条件下实现最优遗憾。该条件确保协变量中的随机性提供足够的内在探索，从而无需显式强制探索。我们的结果提供了一个互补且更一般的观点：我们不依赖关于协变量的几何假设，而是表明，一旦估计误差满足$\max_i \|\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}^*\| \le h/2$，探索便不再必要。换言之，只有在估计误差降至临界阈值以下之前才需要显式探索；此后，算法可以在不损害长期性能的情况下采用近贪心策略运行。这一洞见凸显了一种能够大幅减少或消除探索的算法机制，即使在协变量多样性等几何条件不成立或难以验证的情形下也是如此。

定理4 刻画了存在尾概率时的估计误差率。第（1）部分描述了探索率 π_t 足够大时的收敛行为，从而保证统计最优估计。第（2）部分将分析扩展到探索率较小、可能单调递减甚至趋于零的情形，并证明对于 $\mathcal{A}$中的臂，长期仍可达到统计最优速率，而次优臂的估计误差则保持一致有界。该结果拓展了关于估计误差和探索计划的现有文献；第4.4节提供了详细比较。下面的推论将在 $\delta_t = C\log t$条件下给出更精确的刻画。

推论2（误差动态— $\delta_t = C\log t$）。在与定理*4*相同的条件以及相同的 t_1 和 $\{\eta_t\}$ 下，我们有：

(1) 若 $\pi_t \ge \pi$且 $C_a \ge 2K/\pi$， $C_b \ge CC_a^2\lambda_{\max}^2/\lambda_{\min}^2$，则以至少 $1 - CKC_a e^{-cd}$的概率，$\log_{t+C_b d}^{t}\sigma^2$ 对所有 t均成立。——

(2) 对任意 $\pi_t \in [0,1]$和 $t \ge t_1$，以至少 $1 - |\mathcal{A}|e^{-cd} - |\mathcal{A}^c|e^{-cd}\cdot\sum_{l=t_1}^{t+1}\mathbb{I}\{\pi_l \ne 0\}$ 的概率，对于所有 $i \in \mathcal{A}$, $\|\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i^*\|^2 \le \frac{C^{**}}{\lambda_{\min}}\,{}^{d+}_{\text{对数}}{}_{t+C_b d}^{t}\sigma^2$，以及所有 $i \in \mathcal{A}^c$, $\|\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i^*\|^2 \le \frac{h^2}{4}$，均有。—

推论 2(1) 表明，若 $\pi_t \ge \pi$，则对于所有 $T \le +\infty$，该估计量以至少 $1 - CKC_b\exp(-cd)$的概率满足，

$$\|\boldsymbol{\beta}_i^{(T)} - \boldsymbol{\beta}_i^*\|^2 \le C^*\frac{\max\{d, \log T\}}{\lambda_{\min}(T + C_b d)}\sigma^2, \quad T = 0, 1, \cdots, +\infty. \tag{15}$$

如果 $T \ge \exp(d)$，额外的 $\log T$ 项将不可避免，如第 2 节所示。推论 2(2)进一步涵盖了探索率可能任意小的情形。在适当条件下，例如当 $\pi_t = 0$ 对所有 $t \ge t_2$ 成立，或 $\mathcal{A}^c = \emptyset$成立时，臂 $i \in \mathcal{A}$ 的估计量对所有 $T \le \infty$都满足相同的界。

在给出遗憾界之前，我们先介绍以下模型复杂度度量：

$$\mathsf{CoM}_2 := \sum_{j=1}^{K}\sum_{i=1}^{K}\|\boldsymbol{\beta}_i^* - \boldsymbol{\beta}_j^*\|, \quad \mathsf{CoM}_1 := \sum_{j=1}^{K}\max_i \|\boldsymbol{\beta}_i^* - \boldsymbol{\beta}_j^*\|, \quad \mathsf{CoM}_\infty := \max_{i,j}\|\boldsymbol{\beta}_i^* - \boldsymbol{\beta}_j^*\|.$$

当参数之间的两两距离 $\{\boldsymbol{\beta}_i^*\}$ 较大时，假设 h 5 中的参数也可以取较大的值，从而可能简化最优臂的识别。然而，在这种情况下做出错误决策可能会导致更大的遗憾。下一个定理将形式化说明这些复杂度度量如何影响遗憾。定义累积尾概率

$$\mathsf{Tail}(T, d, \{\delta_l\}) := e^{-cd}\left(c'C_b + K\sum_{l=0}^{t_1} e^{-c\delta_l} + |\mathcal{A}|\sum_{l=t_1+1}^{T} e^{-c\delta_l} + |\mathcal{A}^c|\sum_{l=t_1+1}^{T}\mathbb{I}\{\pi_l \neq 0\}\right).$$

定理 5 （一般学习率下的遗憾界）。 在与定理 t_1 *4* 相同的假设、步长选择以及参数条件下，设 $\pi_t \geq 1/3$ 对 $t \leq t_1$成立，且 $\pi_t \in [0,1]$ 对 $t \geq t_1$成立。那么

$$\begin{aligned}\mathit{Regret}(T) \leq & \underbrace{\sqrt{\lambda_{\max}}\frac{\sum_{t=0}^{T}\pi_t}{K}\min\{\mathit{CoM}_2, \sqrt{d}\mathit{CoM}_1\}}_{R_1} \\ & + \underbrace{\sqrt{\lambda_{\max}}\,\mathit{Tail}(T, d, \{\delta_t\})\min\{\mathit{CoM}_2, \sqrt{d}\mathit{CoM}_\infty\}}_{R_2} \\ & + \underbrace{C\sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}}\sigma\min\{K, \sqrt{d}\}\sum_{t=0}^{T}\sqrt{\frac{d+\delta_t}{t+C_b d}}}_{R_3}.\end{aligned}$$

定理 5 将累积遗憾分解为三个部分。第一项 R_1表示探索导致的遗憾，其大小随 $\sum_{t=0}^{T}\pi_t$变化，后者表示近似的探索迭代次数。第二项 R_2源于定理 4中的失败概率，通常可以忽略不计。第三项 R_3是主要贡献项，其大小按 $\sqrt{T}$的量级增长。因此，为了最小化总体遗憾，应在满足必要条件的前提下，尽可能减小 $\sum_{t=0}^{T}\pi_t$。如定理 4所述，一个方便的选择是令$\pi_t = f(t)$ 其中 f属于类 $\Pi(\tau, \pi)$。 −

定义3。函数 $f(\cdot)$： $\mathbb{R}^+ \to \mathbb{R}^+$属于 $\Pi(\tau, \pi)$，如果 (i) f单调递减；(ii) $0 \leq f(x) \leq 1$；并且 (iii) $f(x) \geq \pi$对所有 $x \leq \tau$ 均成立。

由于 f单调递减，因此有 $\sum_{t=1}^{T}\pi_t \leq \int_0^T f(x)\mathrm{d}x$。以下是满足定义3的探索率 π_t。

例4 (满足定义的示例函数 3)。 以下函数属于 $\Pi(\tau,\ 1/3)$。

1. 令 $f(x) = \frac{C_\pi}{x+2C_\pi}$，其中 $C_\pi \geq \tau$。那么 $f \in \Pi(\tau, 1/3)$ 且 $\sum_{t=1}^{T} \pi_t \leq C_\pi \ln\left(\frac{T+2C_\pi}{T}\right)$;

2. 令 $f(x) = \left(\frac{C_\pi}{x+2^{1/p}C_\pi}\right)^p$，其中 $0 < p < 1$ 且 $C_\pi \geq 3^{1-1/p}\tau$。则 $f \in \Pi(\tau, 1/3)$ 且 $\sum_{t=1}^{T} \pi_t \leq C_\pi^p (T + 2^{1/p}C_\pi)^{1-p}$。

将 $\delta_t = 0$ 或 $\delta_t = \log t$ 代入定理 5 即可得到以下推论。

推论3 (遗憾：不知晓 h)。 假设与定理 5中的假设和参数相同。令 $\eta_t = \frac{1}{\lambda_{\min}} \frac{C_a}{t+C_b d}$ 且 $\pi_t = f(t) \in \Pi(t_1, 1/3)$。若 ~~$\mathcal{A}^c \neq \emptyset$~~，则对所有 $T \leq \exp(Cd)$,

$$\mathsf{Regret}(T) \leq \sqrt{\lambda_{\max}} \frac{\int_0^T f(t)\mathrm{d}t}{K} \min\{\mathsf{CoM}_2, \sqrt{d}\mathsf{CoM}_1\} + C\sqrt{C^*}\sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}} \frac{\sigma \min\{K, \sqrt{d}\} T\sqrt{d}}{\sqrt{T + C_b d} + \sqrt{C_b d}}.$$

若 $\mathcal{A}^c = \emptyset$，则对所有 $T \leq +\infty$,

$$\mathsf{Regret}(T) \leq \sqrt{\lambda_{\max}} \frac{\int_0^T f(t)\mathrm{d}t}{K} \min\{\mathsf{CoM}_2, \sqrt{d}\mathsf{CoM}_1\} + C\sqrt{C^*}\sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}} \frac{\sigma \min\{K, \sqrt{d}\} T\sqrt{d + \log T}}{\sqrt{T + C_b d} + \sqrt{C_b d}}.$$

累积遗憾界由两个主要部分组成。第一部分通过积分 T 依赖于时域 $\int_0^T f(s)\mathrm{d}s$，由探索计划决定，且其增长速度可能慢于 $\sqrt{T}$。第二项按 $\sqrt{T}$ 的量级增长。例如，选择 $\pi_t = \left(\frac{C_\pi}{t+2^{1/p}C_\pi}\right)^p$ 且满足 $1/2 \leq p \leq 1$ (例 4) 可得

$$\mathsf{Regret}(T) \leq O(\sqrt{T}).$$

因此，定理 4 和 5 共同表明，在广泛的一类探索计划下，可以同时实现最优遗憾和估计误差。此外，对 T 和 d 的依赖关系在统计意义上是最优的。在某些情形下，例如 Bastaniand Bayati (2020年中所讨论的情形，参数 h 可以视为已知。当 h 已知时，定理 4 和 5 表明，一旦估计误差满足 $\max_{i\in[K]} \|\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i^*\| \leq h/2$，就不再需要探索。因此，此后可以将探索率设为 $\pi_t = 0$，具体如下所述。

推论4 (遗憾：已知 h)。 在与定理 5相同的假设和条件下，取 $\eta_t = \frac{1}{\lambda_{\min}} \frac{C_a}{t+C_b d}$ 并令 $\pi_t \geq 1/2$ 当 $t < t_1$ 时，~~以及 $\pi_t = 0$~~ 当 $t \geq t_1$时。那么对所有 $T \leq \infty$,

$$\mathsf{Regret}(T) \leq \sqrt{\lambda_{\max}} \frac{t_1}{K} \min\{\mathsf{CoM}_2, \sqrt{d}\mathsf{CoM}_1\} + C\sqrt{C^*}\sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}} \frac{\sigma \min\{K, \sqrt{d}\} T\sqrt{d + \log T}}{\sqrt{T + C_b d} + \sqrt{C_b d}}.$$

设置 $\pi_t = 0$ 可确保次优臂在 $\mathcal{A}^c$ 中永远不会被选中。与推论3相比，该结果移除了限制 $T \leq C\exp(cd)$，即使 $\mathcal{A}^c \neq \emptyset$也不例外。这一改进源于：在 $\pi_t = 0$的情况下，探索引起的误差不再累积，因此非渐近保证得以扩展到无限时间范围。

4.3 渐近分析与推断

本节中，我们研究迭代量 $\boldsymbol{\beta}_i^{(t)}$ 的渐近行为，并据此推导出 $\boldsymbol{\beta}_i^*$的置信区间。关于上下文线性赌博机中随机梯度下降迭代序列推断的研究日益增多；例如，Chen 等人（2021b）、Han 等人（2025）、Chen 等人（2022）以及 Duan 等人（2024），他们研究了在独立同分布的协变量和噪声条件下，采用 $1/t^\alpha$ 步长方案、恒定探索率及相关设置时的推断。第 4.4 节详细比较了这些方法。

全文中，我们始终假设 $\mathbb{E}\{\xi_t|\mathcal{F}_t^+\}=0$ and $\mathbb{E}\{\mathbf{X}_t|\mathcal{F}_t\}=\mathbf{0}$ 几乎必然成立。为确保随机梯度下降迭代序列渐近收敛，我们要求协变量和噪声的二阶矩阵均存在极限。

假设6 （渐近收敛的协变量与噪声）。 存在矩阵 $\boldsymbol{\Sigma}^*$ 和$\boldsymbol{\Sigma}_i^*$ 使得

$$\lim_{t\to+\infty}\mathbb{E}\big\{\|\mathbb{E}\{\mathbf{X}_t^\top\mathbf{X}_t|\mathcal{F}_t\}-\boldsymbol{\Sigma}^*\|\big\}=0 \quad and \quad \lim_{t\to+\infty}\mathbb{E}\big\{\|\mathbb{E}\{\mathbf{X}_t\mathbf{X}_t^\top\cdot\mathbb{I}\{\mathbf{X}_t\in\mathcal{U}_i^*\}|\mathcal{F}_t\}-\boldsymbol{\Sigma}_i^*\|\big\}=0.$$

此外，存在常数 $\lambda_{\max},\lambda_{\min}>0$，使得以下关系几乎必然成立：

$$\lambda_{\min}\mathbf{I}\preceq\mathbb{E}\{\mathbf{X}_t\mathbf{X}_t^\top|\mathcal{F}_t\} \quad and \quad \sup_{\mathbf{V}\in\mathbb{S}^{d-1}:\,\mathbf{V}\in\mathcal{F}_t}\mathbb{E}\{(\mathbf{X}_t^\top\mathbf{V})^4|\mathcal{F}_t\}\le\lambda_{\max}^2.$$

假设 $\xi_t\perp\mathbf{X}_t|\mathcal{F}_t$。存在 $\sigma,\sigma_*>0$，使得 $\lim_{t\to+\infty}\mathbb{E}\{|\mathbb{E}\{\xi_t^2|\mathcal{F}_t\}-\sigma_*^2|\}=0$ 且 $\mathbb{E}\{\xi_t^4|\mathcal{F}_t^+\}\le\sigma^4$ 几乎必然成立。

假设6 是建立 $\boldsymbol{\beta}_i^{(t)}$ 的渐近分布所必需的。然而，在边界附近 $\partial\mathcal{U}_i^*$，$\mathbf{X}_t$ 的行为仍可能造成困难。为说明这一点，考虑 $\boldsymbol{\beta}_1^*=(3,\ 0)^\top$ 和 $\boldsymbol{\beta}_2^*=(-1,\ 0)^\top$ 的情形，并假设

$$\mathbb{P}\{\mathbf{X}=(0,1)^\top\}=\mathbb{P}\{\mathbf{X}=(0,-1)^\top\}=\frac{1}{2}\mathbb{P}\{\mathbf{X}=(2,0)^\top\}=\mathbb{P}\{\mathbf{X}=(-1,0)^\top\}=\frac{1}{5}.$$

这里，$\partial\mathcal{U}_1^*=\{(0,\ y)^\top:\ \ y\in\mathbb{R}\}$是 y轴，但 $\mathbb{P}(\mathbf{X}\in\partial\mathcal{U}_1^*)=2/5$。尽管定理4 和5 中的非渐近保证仍然成立，但渐近行为取决于 $\mathbf{X}_t$ 接近决策边界的频率所体现的敏感性。特别是，仅有假设6 是不够的。为此，我们引入球面测度，用以量化$\mathbf{X}_t$ 在单位球面上的分布情况。对于任意可测的 $A\in\sigma(\mathbb{S}^{d-1})$，定义

$$\nu_X(A;\ t):=\mathbb{E}\big(\|\mathbf{X}_t\|\cdot\mathbb{I}\{\mathbf{X}_t/\|\mathbf{X}_t\|\in A\}\big|\mathcal{F}_t\big),\quad \kappa_X(A;t):=\mathbb{E}\big(\|\mathbf{X}_t\|^2\cdot\mathbb{I}\{\mathbf{X}_t/\|\mathbf{X}_t\|\in A\}\big|\mathcal{F}_t\big).$$

我们在 $\partial\mathcal{U}_i^*$ 附近施加以下温和的正则性条件。

假设 7. 除假设5外，还存在 $h_0 \in (0, h]$ 和常数 $\kappa_0, \nu_0 > 0$，使得对于任意 $h_1, h_2 \in [-h_0, h_0]$，$\kappa_X(\mathcal{U}_i(h_1) \backslash \mathcal{U}_i(h_2)) \leq \kappa_0 |h_1 - h_2|$ 以及 $\nu_X(\mathcal{U}_i(h_1) \backslash \mathcal{U}_i(h_2)) \leq \nu_0 |h_1 - h_2|$。

假设 7 包含了 Chen 等人 (2021a) 中关于协变量的条件；Han 等人 (2025年) 中的条件。下面的注释给出了假设 7 的一个具体示例。

注释6。 假设 7涵盖了广泛的分布。为便于说明，我们令 $\mathbf{X}_t$ 为独立同分布，尽管相关情形下的论证同样成立。在本注释中，我们将 $\kappa_X(\cdot)$ 和 $\nu_X(\cdot)$ 中的索引 t省略。令 μ 表示球面 $\mathbb{S}^{d-1}$上的勒贝格测度。根据勒贝格分解定理，测度 ν_X 可分解为

$$\nu_X = \nu_X^{(1)} + \nu_X^{(2)},$$

其中 $\nu_X^{(1)} \ll \mu$ 和 $\nu_X^{(2)} \perp \mu$。拉东–尼科迪姆定理随后保证导数$\mathrm{d}\nu_X^{(1)} / \mathrm{d}\mu$ 的存在，并且对于任意 $A \in \sigma(\mathbb{S}^{d-1})$，

$$\nu_X^{(1)}(A) = \int_A \frac{\mathrm{d}\nu_X^{(1)}}{\mathrm{d}\mu} \mathrm{d}\mu.$$

如果 $\nu_X^{(2)}\big(\mathcal{U}_i(-h_0) \setminus \mathcal{U}_i(h_0)\big) = 0$ 且 $\frac{d\nu_X^{(1)}}{d\mu} \leq \kappa_0$，则假设 7对 ν_X成立。对 κ_X也可采用相同的论证。多元高斯分布和多元 *Student* t分布等常见分布均属于假设7的适用范围。

如果假设 5 和假设 7 分别在参数 h 和 h_0下成立，那么对于任何更小的 $h' < \min\{h, h_0\}$，它们也成立。因此，为简单起见，我们在全文中假设假设 5 和 7 在共同参数 h下成立。定义

$$\mathbf{\Sigma}_i(\pi^*) := \frac{\pi^*}{K} \mathbf{\Sigma}^* + (1 - \pi^*) \mathbf{\Sigma}_i^*,$$

并将其奇异值分解写作 $\mathbf{\Sigma}_i(\pi^*) = \mathbf{U}_i(\pi^*) \mathbf{\Lambda}_i(\pi^*) \mathbf{U}_i(\pi^*)^\top$，其中 $\mathbf{U}_i(\pi^*)$ 为正交矩阵，且 $\mathbf{\Lambda}_i(\pi^*) = \mathrm{diag}\{\lambda_j(\pi^*)\}$。令 $C_a' = C_a / \lambda_{\min}$，则步长为 $\eta_t = C_a' / (t + C_b d)$。进一步定义

$$\mathbf{\Lambda}_i(\pi^*, C_a', \sigma_*) := \mathrm{diag} \left\{ \frac{\sigma_*^2}{\lambda_j(\pi^*)} \cdot \frac{(C_a' \lambda_j(\pi^*))^2}{2 C_a' \lambda_j(\pi^*) - 1} \right\}.$$

现在，我们给出随机梯度下降迭代序列的 Bahadur 型表示。注意，$\{X_t\}$ 和 $\{\xi_t\}$构成的是马尔可夫过程，而非独立同分布序列。全文中，$O_{\mathbb{P}}(\cdot)$表示随机有界性。

命题 2（Bahadur 型表示）。假设 5、6和7 对某个共同的 h成立。 [2] 在与定理 5中相同的 $\{\eta_t\}$ 和 $\{\pi_t\}$ 的条件下，若 $\lim_{t\to+\infty}\pi_t=\pi^*$，则

$$\boldsymbol{\beta}_i^{(t)}-\boldsymbol{\beta}_i^*=\sum_{j=0}^{t}\prod_{l=j+1}^{t}\left\{\mathbf{I}-\eta_l\cdot\left(\frac{\pi_l}{K}\boldsymbol{\Sigma}+(1-\pi_l)\boldsymbol{\Sigma}_i^*\right)\right\}\eta_j\cdot\xi_j\cdot\mathbf{X}_j\cdot\mathbb{I}\{a_j=i\}+\mathbf{R}_i^{(t)}.$$

此外，在以下任一条件下：

(1) $\pi^*=0$, $i\in\mathcal{A}$ 且在事件 $\cup_{t=t_1}^{+\infty}\{$最大值 $\|\boldsymbol{\beta}_i^{(t)}-\boldsymbol{\beta}_i^*\|\le\frac{h}{2}\}$ 上；

(2) $\pi^*>0$, $C_a\ge 4K/\pi^*$, *and* $i\in[K]$,

我们有 $\mathbf{R}_i^{(t)}=O_{\mathbb{P}}(1/t)$ 且

$$\sqrt{t}\sum_{j=0}^{t}\prod_{l=j+1}^{t}\left\{\mathbf{I}-\eta_l\cdot\left(\frac{\pi_l}{K}\boldsymbol{\Sigma}+(1-\pi_l)\boldsymbol{\Sigma}_i(\pi^*)\right)\right\}\eta_j\xi_j\mathbf{X}_j\cdot\mathbb{I}\{a_j=i\}$$
$$\rightsquigarrow N(0,\mathbf{U}_i(\pi^*)\boldsymbol{\Lambda}_i(\pi^*,C_a',\sigma_*)\mathbf{U}_i(\pi^*)^\top).$$

渐近分布值得注意的是不依赖于 C_b。命题 2 使得能够为 $\boldsymbol{\beta}_i^*$构造置信区间。由于 π^* 和 C_a' 已知，剩下的任务是估计$\boldsymbol{\Sigma}$ 和 $\boldsymbol{\Sigma}_i^*$。

引理 6 ($\boldsymbol{\Sigma}$ 和 $\boldsymbol{\Sigma}_i^*$的估计)。在假设 6下，估计量

$$\widehat{\boldsymbol{\Sigma}}^*(t):=\frac{1}{t}\sum_{l=1}^{t}\mathbf{X}_l\mathbf{X}_l^\top$$

依概率收敛到 $\boldsymbol{\Sigma}^*$ （按 *Frobenius* 范数）。[3] 此外，若 $\boldsymbol{\beta}_i^{(t)}\to\boldsymbol{\beta}_i^*$ 依概率成立，则

$$\widehat{\boldsymbol{\Sigma}}_i(t):=\frac{1}{(1-\pi^*)t}\sum_{l=1}^{t}\mathbf{X}_l\mathbf{X}_l^\top\cdot\mathbb{I}\{a_l=i,\alpha_l=0\}\to\boldsymbol{\Sigma}_i^*,\quad\widehat{\sigma}_*^2(t):=\frac{1}{t}\sum_{l=1}^{t}(Y_l-\mathbf{X}_l^\top\boldsymbol{\beta}_{a_l})^2\to\sigma_*^2,$$

两者都依概率收敛。

由于已知 π^* 和 C_a'，我们构造

$$\widehat{\boldsymbol{\Sigma}}_i(\pi^*,t):=\frac{\pi^*}{K}\widehat{\boldsymbol{\Sigma}}^*(t)+(1-\pi^*)\widehat{\boldsymbol{\Sigma}}_i(t),$$

并令其奇异值分解为 $\widehat{\mathbf{U}}_i(\pi^*,t)\widehat{\boldsymbol{\Lambda}}_i(\pi^*,t)\ \widehat{\mathbf{U}}_i(\pi^*,t)^\top$。根据引理 6 和 Davis–Kahan 定理（Davis 和 Kahan，1970；Yu 等，2015），

$$\widehat{\mathbf{U}}_i(\pi^*,t)\to\mathbf{U}_i(\pi^*),\quad\widehat{\boldsymbol{\Lambda}}_i(\pi^*,t)\to\boldsymbol{\Lambda}_{\boldsymbol{i}}(\pi^*),$$

[2]为简洁起见，我们假设一个共同的 h；如果假设 5 对 h成立，那么它对任何更小的 $h'<h$也自动成立。[3]可以使用任何与 Frobenius 范数等价的范数。

以概率收敛。因此，

$$\widehat{\mathbf{\Lambda}}_i(t) := \operatorname{diag}\left\{ \frac{\widehat{\sigma}_*^2(t)}{\widehat{\lambda}_j(\pi^*, t)} \cdot \frac{(C_a' \widehat{\lambda}_j(\pi^*, t))^2}{2C_a' \widehat{\lambda}_j(\pi^*, t) - 1} \right\} \to \mathbf{\Lambda}_i(\pi^*, C_a', \sigma_*) \quad \text{in probability.}$$

对于任意集合 $\mathcal{C} \subseteq \mathbb{R}^d$，若其在 $N(\mathbf{0}, \mathbf{I})$ 下的概率质量为 $1-\alpha$，

$$\lim_{t\to\infty} \mathbb{P}\big\{ \sqrt{t}\widehat{\mathbf{\Lambda}}_i(t)^{-1/2}(\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i^*) \in \mathcal{C} \big\} = 1-\alpha,$$

由此可为 $\boldsymbol{\beta}_i^*$ 得到渐近有效的置信区间。

4.4 相关文献

本节回顾线性上下文老虎机及相关在线学习方法的文献。Goldenshluger and Zeevi（2013 年）、Bastani 和 Bayati（2020）以及 Bastani 等人（2021 年）的开创性研究，使用基于已存储历史观测的普通最小二乘（OLS）或最小绝对收缩和选择算子（LASSO），估计多臂上下文老虎机的底层参数。他们的分析是非渐近的，并依赖有界独立同分布协变量及独立同分布噪声。这些研究设定了一个间隔条件，在该条件下，最优遗憾率为 log(T)。相比之下，我们的框架不要求这一条件；相反，在我们的设定中，极小极大最优遗憾率的阶为 $\sqrt{T}$。其他离线上下文老虎机算法见 Dimakopoulou 等人（2017 年）、Kang 和 Kim（2023 年）以及 Rusmevichientong 和 Tsitsiklis（2010 年）。Khamaru 等人（2025年）和 Chen 等（2021a）研究了马尔可夫线性回归的渐近性质，但未涉及遗憾。这两项研究都依赖普通最小二乘估计量。具体而言，Chen 等（2021a）关注双臂这一特殊情形（$K = 2$）下的有界独立同分布协变量和独立同分布噪声。迄今为止，对于协变量和噪声均服从马尔可夫过程的情形，尚无已知离线算法能够实现最优非渐近性能。我们的工作通过在在线方式下针对这一设定提供非渐近和渐近分析，且无需存储历史观测，弥补了这一空白。

最近，Chen等（2021b）和Chen等（2022）通过采用随机梯度下降，将计算和内存约束纳入考量。他们的分析也聚焦于双臂设置（$K = 2$），其中次优臂集合 $\mathcal{A}^c$ 为空。他们假设协变量和噪声独立同分布。Chen等（2022）进一步研究了采用恒定探索率的情形，这会带来 O（T）遗憾，但能对估计量进行渐近高斯推断。需要强调的是，具有 $K > 2$ 的一般多臂设定并非双臂情形的简单延伸，无论从渐近角度还是非渐近角度皆是如此；例如，参见命题 2。通过将定理 4 和定理 5 应用于协变量和噪声独立同分布的双臂情形，我们的方法改进了现有的在线上下文老虎机结果，将遗憾从 Regret（T） $= O$（$T^{2/3}$）进一步收紧至极小极大最优的 $\sqrt{T}$ 率。此外，我们通过获得以下结果，强化了 Bahadur 型表示：

—

一个随机意义下为 $O(1/t)$ 的残差项。我们的框架通过充分利用相关数据，消除了经典的探索—利用权衡，使 $\mathcal{A}$ 中的臂的估计误差达到完全探索设计的表现。与要求探索率不能衰减过快的以往研究不同，我们的结果适用于广泛的探索方案，甚至允许 $\pi_t = 0$ 在足够大的 t时成立；参见定理 5 、命题 2 以及推论 3 和 4。

另一类研究关注具有低秩矩阵结构的在线赌博机（Han 等人，2025 年；Duan 等人，2024年；Li 等人，2023 年）。这些方法同样依赖 SGD 更新，但采用大小为 $1/t^{\alpha}$ 且 $\alpha \in (0,1)$ 的步长策略，与我们的方法不同。它们的遗憾界通常为 $T^{2/3}$，高于极小极大下界 $\sqrt{T}$。低秩结构是否会带来额外的技术挑战仍是一个开放问题，也是未来值得研究的方向。

5 数值实验

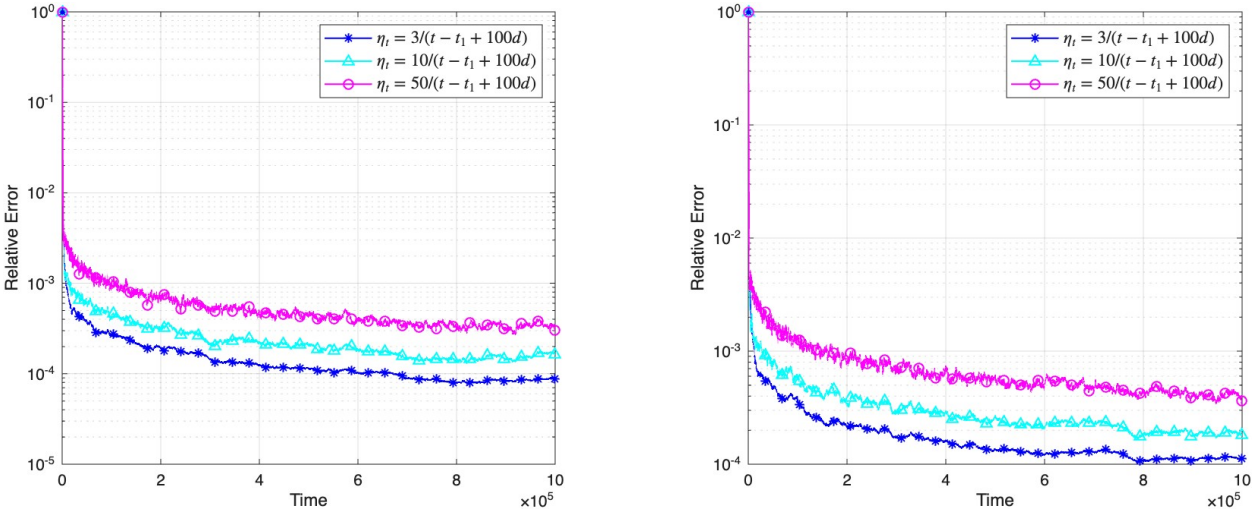
本节介绍一系列数值实验，用于评估所提算法的性能。在所有设置中，实验结果均与理论预测高度吻合。除非另有说明，我们将环境维度设为 $d = 99$ ，并将时间范围设为$T = 999$，999。

线性回归。我们首先进行在线线性回归实验（定理 1，第 2 节）。该过程先采用初始常数步长以稳定残差大小，随后采用形式为 $\frac{C_a}{t-t_1+C_b d}$ 的衰减步长。图 2 展示了在若干种 $C_a \in \{3, 10, 50\}$ 选择下、采用相同 C_b 时的收敛动态。不同步长大小下的估计轨迹非常相似。图 2a 和 2b 随后比较了两种数据生成过程：(i) 独立同分布的协变量和噪声；(ii) 按照以下方式生成的相关协变量和噪声：

$$\mathbf{X}_t = \text{Independent Rad}(1/2) \times \frac{1}{\|\mathbf{X}_{t-1}\|} \times \mathbf{X}_{t-1} + N(\mathbf{0}, \mathbf{I}_d), \tag{16}$$

$$\xi_t = \text{Independent Rad}(1/2) \times \frac{\text{sign}(\xi_{t-1}) \min\{|\xi_{t-1}|, 1\}}{|\mathbf{X}_{t-1}(1)|} \times \mathbf{X}_{t-1}(1) + N(\mathbf{0}, \mathbf{I}_d), \tag{17}$$

其中，Rad(1/2) 表示拉德马赫随机变量，而 $N(\mathbf{0}, \mathbf{I}_d)$ 表示标准多元高斯分布。两种设置使用相同的信噪比。值得注意的是，数据中的依赖关系并不会妨碍估计精度，这与我们的理论保证一致。



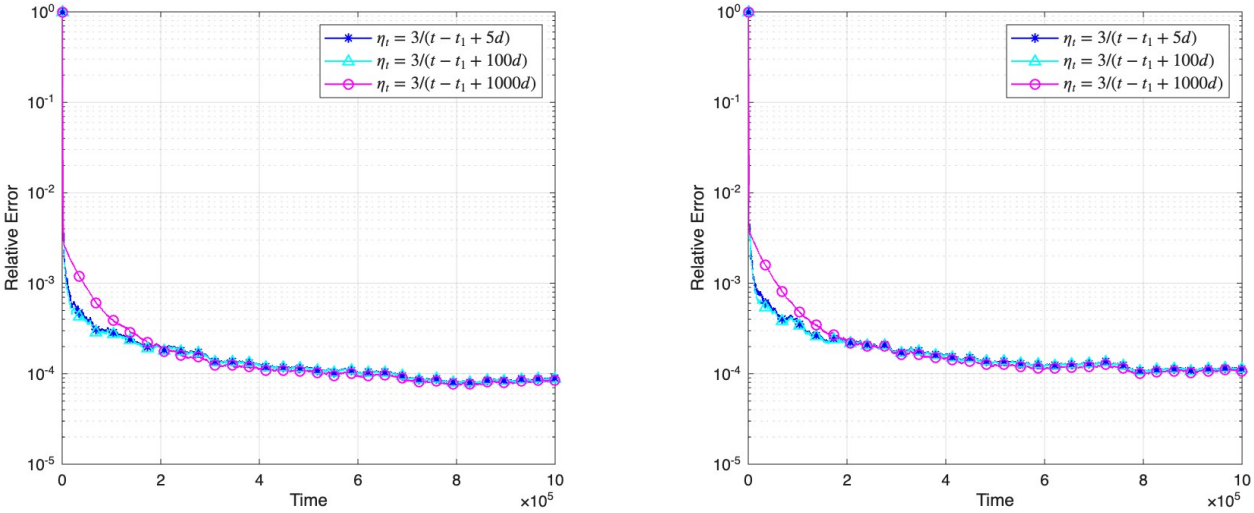
(a) 独立观测下的随机梯度下降　　(b) 相关观测下的随机梯度下降

图 2：关于 C_a 的敏感性：在独立同分布观测和相关观测下，在线性回归中的随机梯度下降。每条曲线描绘了 $C_a \in \{3, 10, 50\}$ 的收敛动态。

图3 研究了在固定 C_a的情况下，对 C_b 的敏感性。即使 C_b存在显著变化，长期协方差的表现仍几乎完全一致。图3b 显示，当协变量和噪声遵循相关过程（16）–（17）时，这种稳健性依然存在，再次与图3a中的独立同分布结果相呼应。

稀疏线性回归。 本部分介绍在线稀疏线性回归的数值实验（定理 3，第 3节）。当时间范围足够大时， $\boldsymbol{\beta}^*$ 的支撑集可以被完全恢复。我们设置 $T = 9,\ 999,\ 999$ 以及 $|\mathrm{supp}(\boldsymbol{\beta}^*)| = 4$。此外，协变量和噪声按照方程 16 和 17生成，并使其具有相关性。采用带常数步长方案的稠密随机梯度下降算法（定理 1(1)）进行初始化。在此初始化阶段之后，我们执行一次硬阈值化，保留 $\boldsymbol{\beta}_t$ 中绝对值最大的六个条目。我们将稀疏随机梯度下降方法（定理 3）的性能与稠密参数随机梯度下降算法（定理 1）进行比较。为确保公平比较，两种算法使用相同的调节参数 $C_a = 3$ 和 $C_b = 100$。

图 4a 对应于弱信噪比设置， $\|\boldsymbol{\beta}^*\|/\mathbb{E}|\xi| = 5$，其中最大的信号等于 $4\mathbb{E}|\xi|$ ，最小的信号低至 $0.004\mathbb{E}|\xi|$。初始化后，仅恢复出两个非零分量。我们在定理 $\alpha = 7$ 中设定3，这意味着稀疏算法最多更新 $\boldsymbol{\beta}_t$ 的支撑集七次。尽管剩余两个信号相对于噪声的幅度极其微弱，稀疏算法仍成功识别出全部四个支撑集条目



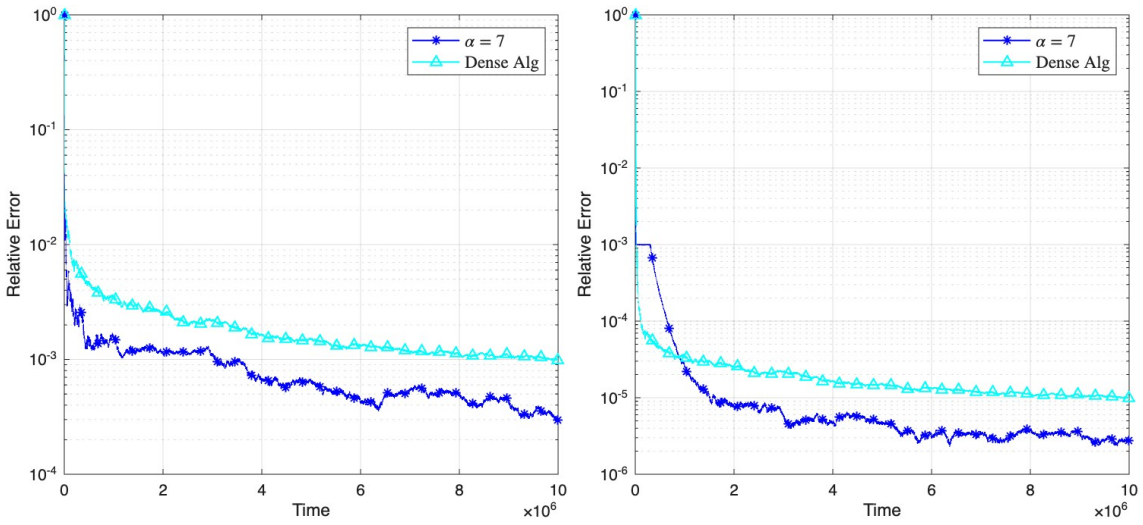
(a) 独立观测下的随机梯度下降　　(b) 相关观测下的随机梯度下降

图3：对 C_b 的敏感性：在独立同分布和相关观测下，在线线性回归中的随机梯度下降。每条曲线展示了 $C_b \in \{5, 100, 1000\}$ 的收敛动态。

在时间范围结束时。如图4a所示，稀疏估计量的估计误差显著小于稠密随机梯度下降估计量。图4b 报告了高信噪比条件下的结果。在恒定步长阶段和进行一次硬阈值化后，四个信号中的三个得到恢复。随后，稀疏算法识别出最后一个缺失信号，其估计误差再次显著小于稠密随机梯度下降算法。

线性赌博机。接下来，我们研究上下文线性赌博机设定（定理 4 和5，第 4 节）。根据线性回归实验的观察，在相关性下准确度损失可以忽略不计，因此我们在此重点研究相关数据生成过程（16）和（17）。我们设置 $K = 5$ 个臂，其中恰有一个是次优臂，并选择满足假设 5 的参数。步长计划与回归实验一致：先进行一个较短的热启动阶段，采用常数步长，随后采用 $\eta_t = \frac{20}{t - t_1 + 50d^\circ}$

图 5 比较了多种探索方案下的估计精度和遗憾。图 5a报告了最优臂的估计误差。值得注意的是，在递减探索率 $\pi_t = \frac{5}{\sqrt{t - t_1 + 50}}$ 和 $\pi_t = \frac{5}{t - t_1 + 50}$ 下的估计轨迹，与恒定探索率下得到的轨迹几乎无法区分。这表明，决策过程所诱导的依赖结构提供了足够的隐式探索，从而能够保持较高的估计精度，这与定理 4的保证一致。图 5b 展示了不同探索策略下的累积遗憾。当探索率保持恒定时，遗憾呈



(a) 低信噪比：$\frac{\|\boldsymbol{\beta}^*\|}{\mathbb{E}|\xi|} = 5$　　(b) 高信噪比：$\frac{\|\boldsymbol{\beta}^*\|}{\mathbb{E}|\xi|} = 500$

图4：稀疏线性回归：真实参数 $\boldsymbol{\beta}^* \in \mathbb{R}^{99}$ 的支撑集大小为 $|\mathrm{supp}(\boldsymbol{\beta}^*)| = 4$。“Dense Alg”曲线对应定理 1(2) 中的估计量；而“$\alpha = 7$”曲线展示了定理3中的算法所得到的估计误差。

线性增长，并且明显大于递减探索计划 $\pi_t = \frac{5}{\sqrt{t-t_1+50}}$ 或$\pi_t = \frac{5}{t-t_1+50}$下的遗憾。递减探索规则的优异表现与定理 5所确立的理论保证一致。

6 致谢

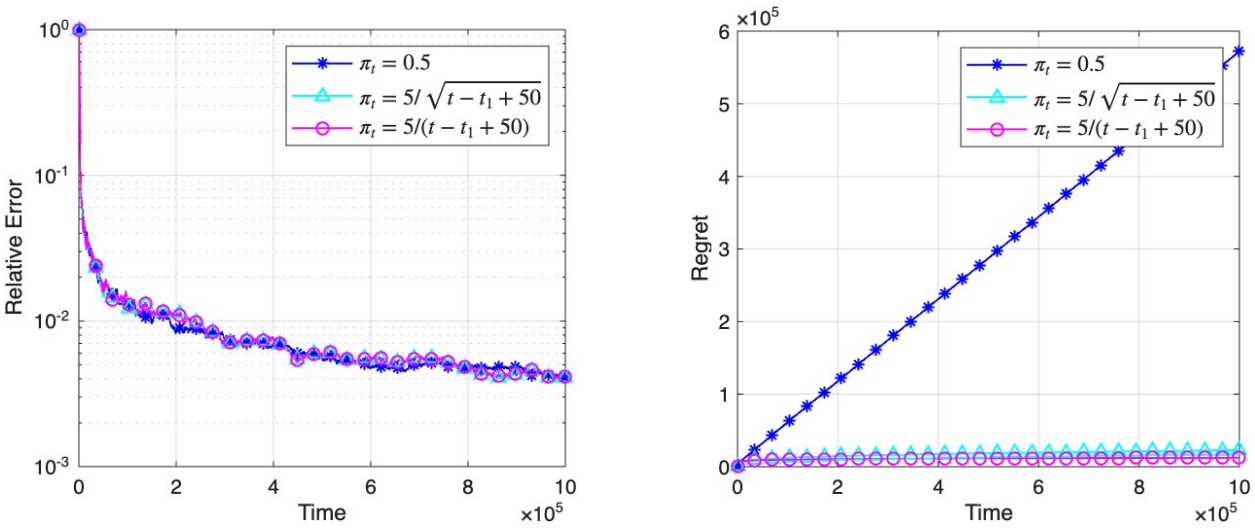
Y. Shen 感谢美国国家科学基金会项目 DMS CAREER-2045068 的资助。

参考文献

Bhavya Agrawalla、Krishnakumar Balasubramanian 和 Promit Ghosal。当 $t \gtrsim d^{1+\delta}$ 时在线最小二乘随机梯度下降的线性泛函的统计推断。*arXiv* 预印本 *arXiv:2510.19734*，2025年。

Kazuoki Azuma。某些相依随机变量的加权和。东北数学杂志，第二辑，19(3):357–367，1967年。

Emmanuel N Barron、Pierre Cardaliaguet 和 Reed Jensen。条件本质上确界及其应用。应用数学与优化，48(3):229–253，2003年。



(a) 估计误差　　(b) 遗憾

图5：探索率方案：三种探索方案下的估计误差和遗憾，$\pi_t = \frac{1}{2}$，$\frac{5}{\sqrt{t-t_1+50}}$，以及 $\frac{5}{t-t_1+50}$，均在同一个固定的随机梯度下降步长方案下计算。

Hamsa Bastani 和 Mohsen Bayati。基于高维协变量的在线决策。《运筹学》，68(1):276–294，2020年。

Hamsa Bastani、Mohsen Bayati 和 Khashayar Khosravi。上下文老虎机的免探索算法。《管理科学》，67(3):1329–1349，2021 年。

Pierre C Bellec、Guillaume Lecué 和 Alexandre B Tsybakov。Slope方法与最小绝对收缩和选择算子：改进的预言界与最优性。《统计年刊》，46(6B):3603–3642，2018 年。

Fischer Black。噪声。《金融学杂志》，41(3):528–543，1986 年。

Thomas Blumensath 和 Mike E Davies。用于压缩感知的迭代硬阈值。《应用与计算调和分析》，27(3):265–274，2009年。

Malgorzata Bogdan、Ewout Van Den Berg、Chiara Sabatti、Weijie Su 和 Emmanuel Candès。通过凸优化实现 Slope方法–自适应变量选择。《应用统计年鉴》，9 (65):1103–1140，2015。

B. M. Brown。鞅中心极限定理。《数学统计年刊》，42(1): 59–66，1971年。

Emmanuel J Candes 和 Yaniv Plan。通过 ℓ_1 最小化实现近乎理想的模型选择。《统计年刊》，37(5A)，2009年。

Haoyu Chen、Wenbin Lu 和 Rui Song。在线决策的统计推断：上下文老虎机设定下的研究。《美国统计学会杂志》，116(533):240–255，2021a。Haoyu Chen、Wenbin Lu 和 Rui Song。通过随机梯度下降进行在线决策的统计推断。《美国统计学会杂志》，116(534):708–719，2021b。Xi Chen、Jason D Lee、Xin T Tong 和 Yichen Zhang。随机梯度下降中模型参数的统计推断。《统计年刊》，48(1):251–273，2020年。Xi Chen、Zehua Lai、He Li 和 Yichen Zhang。通过随机梯度下降实现上下文老虎机的在线统计推断。*arXiv*预印本 *arXiv:2212.14883*，2022。H Cramér 和 H Wold。关于分布函数的一些定理。《伦敦数学学会杂志》，1(4):290–294，1936年。Chandler Davis 和 William Morton Kahan。扰动对特征向量的旋转。III。《*SIAM*数值分析杂志》，7(1):1–46，1970年。Jérôme Dedecker 和 Emmanuel Rio。关于平稳过程的函数中心极限定理。《亨利·庞加莱研究所年刊（*B*辑）：概率与统计》，36(1):1–34，2000年。Maria Dimakopoulou、Zhengyuan Zhou、Susan Athey 和 Guido Imbens。上下文老虎机中的估计考量。*arXiv*预印本 *arXiv:1711.07077*，2017年。Paul Doukhan、Pascal Massart 和 Emmanuel Rio。强混合过程的函数中心极限定理。《*IHP*概率与统计年刊》，30(1):63–82，1994年。Congyuan Duan、Jingyang Li 和 Dong Xia。通过矩阵补全进行在线策略学习与推断。*arXiv*预印本 *arXiv:2404.17398*，2024年。Rick Durrett。《概率论：理论与例题》，第49卷。剑桥大学出版社，2019年。Jianqing Fan 和 Jinchi Lv。超高维特征空间的确定性独立筛选。《皇家统计学会志*B*辑：统计方法学》，70(5):849–911，2008年。Jianqing Fan、Yingying Fan 和 Emre Barut。自适应稳健变量选择。《统计年刊》，42(1):324，2014年。Jianqing Fan、Wenyan Gong、Chris Junchi Li 和 Qiang Sun。统计稀疏在线回归：扩散近似视角。载于人工智能与统计国际会议，第1017–1026页。PMLR，2018年。

Yixin Fang、Jinfeng Xu 和 Lei Yang。随机梯度下降估计量的在线自助法置信区间。《机器学习研究杂志》，19(78):1–21，2018年。Hans Föllmer 和 Alexander Schied。《随机金融：离散时间导论》。载于《随机金融》。De Gruyter，2016年。Zijun Gao、Yanjun Han、Zhimei Ren 和 Zhengqing Zhou。批处理多臂老虎机问题。神经信息处理系统进展，32，2019年。Alexander Goldenshluger 和 Assaf Zeevi。线性响应老虎机问题。《随机系统》，3(1):230–261，2013 年。Qiyang Han、Koulik Khamaru 和 Cun-Hui Zhang。多臂老虎机的 UCB 算法：精确遗憾与自适应推断。*arXiv* 预印本 *arXiv:2412.06126*，2024a。Qiyu Han、Will Wei Sun 和 Yichen Zhang。矩阵上下文下决策中的在线统计推断。《统计年刊》，53(5):1963–1986，2025年。Ruijian Han、Lan Luo、Yuanyuan Lin 和 Jian Huang。采用去偏随机梯度下降进行在线推断。*Biometrika*，111(1):93–108，2024b。Botao Hao、Tor Lattimore 和 Mengdi Wang。高维稀疏线性老虎机。神经信息处理系统进展，33:10753–10763，2020年。Jean J acod、Yingying Li 和 Xinghua Zheng。微观结构噪声的统计性质。*Econo-metrica*，85(4):1133–1174，2017年。Chi Jin、Sham M Kakade 和 Praneeth Netrapalli。通过非凸随机梯度下降实现可证明高效的在线矩阵补全。神经信息处理系统进展，29，2016年。Minhyun Kang 和 Gi-Soo Kim。采用 Huber 回归的重尾线性老虎机。载于人工智能中的不确定性，第 1027–1036 页。PMLR，2023 年。Koulik Khamaru、Yash Deshpande、Tor Lattimore、Lester Mackey 和 Martin J Wainwright。自适应线性回归中的近最优推断。《统计年刊》，53(6)，2025年。Jack Kiefer 和 Jacob Wolfowitz。回归函数最大值的随机估计。《数学统计年刊》，第 462–466 页，1952年。Vladimir I Koltchinskii。一般经验过程的 Komlós-Major-Tusnády逼近以及函数类的 Haar展开。《理论概率杂志》，7:73–118，1994年。

Tze Leung Lai。自适应治疗分配与多臂老虎机问题。《统计年刊》，第 1091–1114 页，1987年。Fabien Lauer。基于相依数据序列进行学习的一致风险界。*arXiv* 预印本 *arXiv:2303.11650*，2023 年。Emmanuel Lépinette 和 Ilya Molchanov。随机集的条件核与条件凸包。数学分析与应用杂志，478(2):368–392，2019年。Jingyang Li、Jian-Feng Cai、Yang Chen 和 Dong Xia。在线张量学习：计算与统计权衡、自适应性及最优遗憾。*arXiv* 预印本 *arXiv:2306.03372*，2023 年。Zijian Liu 和 Zhengyuan Zhou。重新审视随机梯度方法的最后迭代收敛。载于第十二届国际学习表征会议，2024年。Jianya Lu、Wei Biao Wu、Zhijie Xiao 和 Lihu Xu。希尔伯特空间中 β-混合时间序列的几乎必然不变性原理。*arXiv* 预印本 *arXiv:2209.12535*，2022。

Tianyi Ma、Kabir A Verchand 和 Richard J Samworth。《高概率极小极大下界》。*arXiv* 预印本 *arXiv:2406.13447*，2024年。

Elias Masry 和 Jianqing Fan。混合过程回归函数的局部多项式估计。斯堪的纳维亚统计学杂志，24(2):165–179，1997年。

Florence Merlevède、Magda Peligrad 和 Emmanuel Rio。强混合条件下的伯恩斯坦不等式与中等偏差。载于高维概率 *V*：卢米尼卷，第 5 卷，第 273–293 页。数理统计学会，2009年。

Dheeraj Nagaraj、Xian Wu、Guy Bresler、Prateek Jain 和 Praneeth Netrapalli。马尔可夫数据上的最小二乘回归：基本极限与算法。神经信息处理系统进展，33：16666–16676，2020年。Vianney Perchet 和 Philippe Rigollet。带协变量的多臂老虎机问题。《统计年刊》，41(2)：693–721，2013 年。Boris T Polyak 和 Anatoli B Juditsky。通过平均加速随机逼近。*SIAM* 控制与优化杂志，30(4)：838–855，1992年。Zhimei Ren 和 Zhengyuan Zhou。高维稀疏线性上下文多臂老虎机中的动态批次学习。《管理科学》，70(2)：1315–1342，2024年。

Herbert Robbins 和 Sutton Monro。一种随机逼近方法。《数学统计年刊》，第 400–407 页，1951 年。

Paat Rusmevichientong 和 John N Tsitsiklis。线性参数化赌博机。《运筹学数学》，35(2):395–411，2010年。

Ohad Shamir。针对具有次高斯尾部的鞅的 Azuma 不等式变体。*arXiv* 预印本 *arXiv:1110.2392*，2011年。

Qi-Man Shao 和 Zhuo-Song Zhang。多元非线性统计量的 Berry–Esseen界及其在 M估计量和随机梯度下降算法中的应用。《伯努利》，28(3): 1548–1576，2022。

Yinan Shen、Dong Xia 和 Wen-Xin Zhou。在线分位数回归。《机器学习研究杂志》，26(231):1–55，2025年。

R Srikant。马尔可夫链中心极限定理中的收敛速率及其在时序差分学习中的应用。*arXiv* 预印本 *arXiv:2401.15719*，2024年。

Qiang Sun、Wen-XinZhou 和 Jianqing Fan。自适应Huber回归。《美国统计学会杂志》，115(529):254–265，2020年。

AlexanderG Tartakovsky。适用于一般非独立同分布随机模型的某些多重决策序贯规则的近最优性质。*arXiv* 预印本 *arXiv:2405.00928*，2024年。

Masato Ubukata 和 KosukeOya。市场微观结构噪声中依赖关系的估计与检验。《金融计量经济学杂志》，7(2):106–151，2009年。

Sara van de Geer。鞅的指数不等式及其在计数过程最大似然估计中的应用。《统计年刊》，第1779–1801页，1995年。

Sara A van de Geer。关于相依随机变量的Hoeffding不等式。载于《相依数据的经验过程技术》，第161–169页。施普林格，2002年。

Xue Wang、Mingcheng Wei 和 Tao Yao。《具有高维协变量的极小极大凹惩罚多臂赌博机模型》。发表于国际机器学习大会，第 5200–5208 页。PMLR，2018 年。

Ziyang Wei、Jiaqi Li、Likai Chen 和 Wei Biao Wu。通过恒定学习率随机梯度下降进行分位数在线推断。*arXiv* 预印本 *arXiv:2503.02178*，2025年。

Wei Biao Wu。相依随机变量的强不变性原理。《概率年刊》，35(6):2294–2320，2007年。

Wei Biao Wu 和 Zhibiao Zhao。时间序列趋势推断。《皇家统计学会志*B*辑：统计方法学》，69(3):391–410，2007 年。

Wei Biao Wu、Zhiwei Xu 和 Yu Wang。使用分位数回归模型稳健预测网络流量。载于*2006* 年*IEEE*信息复用与集成国际会议，第 220–225 页。IEEE，2006 年。

Han Xiao 和 Wei Biao Wu。平稳时间序列的协方差矩阵估计。《统计年刊》，第 466–493 页，2012 年。

Shuoguang Yang、Yuhao Yan、Xiuneng Zhu 和 Qiang Sun。在线线性化最小绝对收缩和选择算子。载于人工智能与统计国际会议，第 7594–7610 页。PMLR，2023 年。

Yi Yu、Tengyao Wang 和 Richard J Samworth。适用于统计学家的 Davis–Kahan 定理的一个实用变体。*Biometrika*，102(2):315–323，2015 年。

Cun-Hui Zhang. 极小极大凹惩罚下近乎无偏的变量选择。《统计年刊》，38(2):894–942，2010年。

“依赖数据下的随机梯度下降：最优估计、遗憾与推断”的补充材料

A 节中的证明 2

A.1 条件 Orlicz范数

本节将正式介绍条件 Orlicz 范数的定义。我们需要以下本质上确界的定义；有兴趣的读者可参阅 Barron等人(2003年),Föllmer和Schied(2016年)以及Lépinette和Molchanov(2019年，了解更详细的讨论。

定义 4 （本质上确界／下确界的定义）。设 Φ 是 $(\Omega\ \mathcal{F}, \mathbb{P})$ 上的任意一组随机变量。

1. 存在一个满足以下两个性质的随机变量 φ^* 。

(a) 对所有 $\varphi \in \Phi$， $\varphi^* \geq \varphi$ *P*-几乎处处；

(b) 对每个 $\psi \geq \varphi$， $\varphi \in \Phi$ 均满足 $\psi \geq \varphi$ *P-a.s.* 时，它具有 $\psi \geq \varphi^*$；

那么 φ^* 是 Φ 的本质上确界。 Φ 的本质下确界是 $-\Phi$ 的本质上确界。

2. 此外，假设 Φ 向上有向，即对每个 φ、$\widetilde{\varphi} \in \Phi$，都存在 $\psi \in \Phi$ 使得 $\psi \geq \varphi \vee \widetilde{\varphi}$。那么存在一个递增序列 $\varphi_1 \leq \varphi_2 \leq \cdots$，属于 Φ，使得 $\varphi^* =$ 极限$_{n\to+\infty}\varphi_n$ *P*-几乎必然。

定理6 (条件 Orlicz 范数). 假设 X 是定义在 $(\Omega\ \mathcal{F}_X, \mathbb{P})$ 上的实值随机变量。那么对任意 $\mathcal{F} \subseteq \mathcal{F}_X$，令 $\Phi := \{\varphi \in \mathcal{F}\colon\ \varphi \geq 0,\ \mathbb{E}\{\exp(|X/\varphi|^\alpha)|\mathcal{F}\} \leq 2\}$。条件 *Orlicz* 范数定义如下，

$$\|X|\mathcal{F}\|_{\Psi_\alpha} := \operatorname{ess\,inf} \Phi.$$

那么，对于 $\alpha \geq 1$， $\|\cdot|\mathcal{F}\|_{\Psi_\alpha}$作为范数定义良好，并满足 $\mathbb{E}\{\exp(|X/\|X|\mathcal{F}\|_{\Psi_\alpha}|^\alpha)|\mathcal{F}\} \leq 2$。

证明改编自 Orlicz范数 $\|\cdot\|_{\Psi_\alpha}$ 的证明。

证明。 我们首先证明 $\|\cdot|\mathcal{F}\|_{\Psi_\alpha}$是一个范数。显然，对于任意标量 a，它满足 $\|aX|\mathcal{F}\|_{\Psi_\alpha} = |a|\cdot\|X|\mathcal{F}\|_{\Psi_\alpha}$。另一方面，如果 $X = 0$几乎必然成立，那么 $\|X|\mathcal{F}\|_{\Psi_\alpha} = 0$几乎必然成立。反过来，如果 $\|X|\mathcal{F}\|_{\Psi_\alpha} = 0$几乎必然成立，那么令 $A := \{w\colon\ X(w) \neq 0\}$。根据詹森不等式，我们有 $\exp((\mathbb{E}\{|X|/K_{\mathcal{F}}|\mathcal{F}\})^\alpha) \leq \mathbb{E}\{\exp((|X|/K_{\mathcal{F}})^\alpha)|\mathcal{F}\} \leq 2$，由此可得 $\mathbb{E}|X| = 0$，这表明 $X = 0$几乎必然成立。我们只

需要验证三角不等式。对于任意 $\varepsilon > 0$，令 $u = \|X|\mathcal{F}\|_\alpha + \varepsilon$ 和 $v = \|Y|\mathcal{F}\|_{\Psi_\alpha} + \varepsilon$，则有

$$\begin{aligned}
\mathbb{E}\left\{\exp\left(\left(\frac{|X+Y|}{u+v}\right)^\alpha\right)\middle|\mathcal{F}\right\} &\leq \mathbb{E}\left\{\exp\left(\left(\frac{|X|+|Y|}{u+v}\right)^\alpha\right)\middle|\mathcal{F}\right\} \\
&= \mathbb{E}\left\{\exp\left(\left(\frac{u}{u+v}\frac{|X|}{u}+\frac{v}{u+v}\frac{|X|}{v}\right)^\alpha\right)\middle|\mathcal{F}\right\} \\
&\leq \frac{u}{u+v}\mathbb{E}\left\{\exp\left(\left(\frac{|X|}{u}\right)^\alpha\right)\middle|\mathcal{F}\right\}+\frac{v}{u+v}\mathbb{E}\left\{\exp\left(\left(\frac{|X|}{v}\right)^\alpha\right)\middle|\mathcal{F}\right\} \\
&\leq 2,
\end{aligned}$$

这证明 $\|\cdot|\mathcal{F}\|_{\Psi_\alpha}$ 是一个范数。此外，$-\Phi$ 向上有向，因此根据定义4，存在一列单调递减的正随机变量 $\varphi_1 \geq \varphi_2 \geq \cdots$ 属于 Φ，使得极限 $\lim_{n\to+\infty}\varphi_n = \varphi^* := \|X|\mathcal{F}\|_{\Psi_\alpha}$。因此，根据 Fatou 引理，我们有

$$\mathbb{E}\left\{\exp\left|\frac{X}{\|X|\mathcal{F}\|_{\Psi_\alpha}}\right|^\alpha|\mathcal{F}\right\} = \lim_{n\to+\infty}\mathbb{E}\left\{\exp\left|\frac{X}{\varphi_n}\right|^\alpha|\mathcal{F}\right\} \leq 2.$$

□

引理 7（预备）。设**X**、**Y** 为随机向量，且 $\mathcal{F}$ 是一个 σ-域，则有

$$\|\mathbf{X}\mathbf{Y}^\top|\mathcal{F}\|_{\Psi_1} \leq \|\mathbf{X}|\mathcal{F}\|_{\Psi_2}\cdot\|\mathbf{Y}|\mathcal{F}\|_{\Psi_2}.$$

若 $X \perp Y|\mathcal{F}$，则 $\mathbb{E}[XY|\mathcal{F}] = \mathbb{E}[X|\mathcal{F}]\cdot\mathbb{E}[Y|\mathcal{F}]$和 $\mathbb{E}[X|Y,\mathcal{F}] = \mathbb{E}[X|\mathcal{F}]$, $a.s.$。若 $\|X|\mathcal{F}\|_{\Psi_\alpha} \leq K$ 成立 $a.s.$，且 K是一个常数，则 $\|X\|_{\Psi_\alpha} \leq K$。若 X与 $\mathcal{F}$独立，则 $\|X|\mathcal{F}\|_{\Psi_1} =\|X\|_{\Psi_1}$, $a.s.$。若 X, Y是非负的非负随机变量，且 $|Y| \leq K$ 几乎处处成立，其中 K 是一个常数，则 $\|XY|\mathcal{F}\|_{\Psi_\alpha} \leq K\|X|\mathcal{F}\|$。

引理 8. 若 $\mathcal{F}_1 \subseteq \mathcal{F}_2$ 为σ域且 $\|\mathbf{X}|\mathcal{F}_2\|_{\Psi_2} \leq K$ 对某个常数 K成立，则 $\|\mathbf{X}|\mathcal{F}_1\|_{\Psi_2} \leq K$。此外，对于 $\mathbf{X} \in \mathcal{F}_2$，若 $\|\mathbf{X}|\mathcal{F}_1\|_{\Psi_2} \leq \lambda$ 且 $\|\xi|\mathcal{F}_2\|_{\Psi_2} \leq K$，则有 $\|\xi\mathbf{X}|\mathcal{F}_1\|_{\Psi_1} \leq K\lambda$。

接下来，我们将条件 Orlicz 范数的性质作为已知知识使用。

A.2 定理1的证明

本节仅证明定理 $\eta_t := \eta$ 1 的常数步长方案。其余部分的证明留待第 C 节，其中将证明一个更一般的收敛结果，并将定理 1 的第二部分作为特例纳入其中。

定理 *1* （*i*）的证明。我们用归纳法证明。记事件

$$\mathcal{E}_t := \left\{ \|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|^2 \leq C^*(1-c)^{2t}\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|^2 + 2\frac{\lambda_{\max}}{\lambda_{\min}}\eta\sigma^2 d \right\},$$

其中 c、C^* 的取值将在后文给出。显然，该事件以 $t = 0$ 成立。接下来，我们将在 $\cup_{l=0}^{t}\mathcal{E}_l$ 下证明 $\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*$ 的收敛动态。令 $\mathcal{E}_t^X$: $= \left\{\|\mathbf{X}_t\|^2 \leq 2\lambda_{\max}d\right\}$。假设 1 意味着 $\mathbb{P}(\mathcal{E}_t^X|\mathcal{F}_t) \geq 1 - \exp(-Cd)$。根据更新方程（3），我们可以将估计误差写为:

$$\left\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\right\|^2 = \|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|^2 - 2\eta\left(\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\right)^\top \mathbf{X}_t\left(\mathbf{X}_t^\top\boldsymbol{\beta}_t - Y_t\right) + \eta^2\left(\mathbf{X}_t^\top\boldsymbol{\beta}_t - Y_t\right)^2\|\mathbf{X}_t\|^2.$$

我们首先考虑在给定 $\mathcal{F}_t^+$ 条件下的期望。假设 2意味着 $\mathbb{E}\left\{\xi_t|\mathcal{F}_t^+\right\} = 0$。因此，我们有

$$\begin{aligned}
&\mathbb{E}\left\{\left\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\right\|^2|\mathcal{F}_t^+\right\} \\
&\quad\leq \|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|^2 - 2\eta\left(\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\right)^\top\mathbf{X}_t\mathbf{X}_t^\top\left(\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\right) + \eta^2\left(\mathbf{X}_t^\top\boldsymbol{\beta}_t - \mathbf{X}_t^\top\boldsymbol{\beta}^*\right)^2\|\mathbf{X}_t\|^2 + \eta^2\sigma^2\|\mathbf{X}_t\|^2.
\end{aligned}$$

We 进一步考虑在给定 $\mathcal{F}_t$ 条件下的期望。根据假设 1，我们有

$$\mathbb{E}\left\{\left\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\right\|^2|\mathcal{F}_t\right\} \leq \left(1 - 2\eta\lambda_{\min} + \eta^2\lambda_{\max}^2 d\right)\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|^2 + \eta^2\sigma^2\lambda_{\max}d.$$

接着，我们考虑 $\left\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\right\|^2$ 与 $\mathbb{E}\left\{\left\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\right\|^2|\mathcal{F}_t\right\}$之间的差:

$$\begin{aligned}
\left\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\right\|^2 - \mathbb{E}\left\{\left\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\right\|^2|\mathcal{F}_t\right\} &= \underbrace{2\eta\cdot\xi_t\left(\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\right)^\top\mathbf{X}_t}_{I_1^{(t)}} \\
&\underbrace{-2\eta\left(\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\right)^\top\left\{\mathbf{X}_t\mathbf{X}_t^\top - \mathbb{E}\left\{\mathbf{X}_t\mathbf{X}_t^\top|\mathcal{F}_t\right\}\right\}\left(\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\right)}_{I_2^{(t)}} + \underbrace{\eta^2\left\{\xi_t^2\|\mathbf{X}_t\|^2 - \mathbb{E}\left\{\xi_t^2\|\mathbf{X}_t\|^2|\mathcal{F}_t\right\}\right\}}_{I_3^{(t)}} \\
&\underbrace{+\eta^2\left\{\left(\mathbf{X}_t^\top\boldsymbol{\beta}_t - \mathbf{X}_t^\top\boldsymbol{\beta}^*\right)^2\|\mathbf{X}_t\|^2 - \mathbb{E}\left\{\left(\mathbf{X}_t^\top\boldsymbol{\beta}_t - \mathbf{X}_t^\top\boldsymbol{\beta}^*\right)^2\|\mathbf{X}_t\|^2|\mathcal{F}_t\right\}\right\}}_{I_4^{(t)}}.
\end{aligned}$$

接下来，我们将估计误差分解如下:

$$\begin{aligned}
\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\|^2 &= \mathbb{E}\left\{\left\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\right\|^2|\mathcal{F}_t\right\} + \left\{\left\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\right\|^2 - \mathbb{E}\left\{\left\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\right\|^2|\mathcal{F}_t\right\}\right\} \\
&\leq \left(1 - 2\eta\lambda_{\min} + \eta^2\lambda_{\max}^2 d\right)\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|^2 + \eta^2\sigma^2\lambda_{\max}d + I_1^{(t)} + I_2^{(t)} + I_3^{(t)} + I_4^{(t)}.
\end{aligned}$$

对于 $\eta \leq \lambda_{\min}/(d\lambda_{\max}^2)$，我们有

$$\|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*\|^2 \leq \left(1 - \eta\lambda_{\min}\right)\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|^2 + \eta^2\sigma^2\lambda_{\max}d + I_1^{(t)} + I_2^{(t)} + I_3^{(t)} + I_4^{(t)}.$$

然后，我们将上述等式累加至 $t=0$,

$$\begin{aligned}\|\boldsymbol{\beta}_{t+1}-\boldsymbol{\beta}^*\|^2 \leq (1-\eta\lambda_{\min})^{t+1}\|\boldsymbol{\beta}_0-\boldsymbol{\beta}^*\|^2+\sum_{l=0}^{t}(1-\eta\lambda_{\min})^{t-l}I_1^{(l)}+\sum_{l=0}^{t}(1-\eta\lambda_{\min})^{t-l}I_2^{(l)}\\ +\sum_{l=0}^{t}(1-\eta\lambda_{\min})^{t-l}I_3^{(l)}+\sum_{l=0}^{t}(1-\eta\lambda_{\min})^{t-l}I_4^{(l)}+\eta^2\sigma^2\lambda_{\max}d\sum_{l=0}^{t}(1-\eta\lambda_{\min})^{l}.\end{aligned} \tag{18}$$

然后，我们对方程 18 右侧的每一项进行界定。值得注意的是

$$(1-\eta\lambda_{\min})^{t+1}\leq(1-\eta\lambda_{\min}/2)^{2(t+1)},\quad \eta^2\sigma^2\lambda_{\max}d\sum_{l=0}^{t}(1-\eta\lambda_{\min})^{l}\leq\frac{\lambda_{\max}}{\lambda_{\min}}\eta\sigma^2 d.$$

然后，我们只需对累加项 I_1, I_2, I_3, I_4 进行界定。首先考虑 I_1 项，因为 $\boldsymbol{\beta}_t\in\mathcal{F}_t$，我们有

$$\mathbb{E}\left\{\xi_l\left(\boldsymbol{\beta}_l-\boldsymbol{\beta}^*\right)^\top\mathbf{X}_l\middle|\mathcal{F}_l\right\}=0,\quad \left\|\xi_l\left(\boldsymbol{\beta}_l-\boldsymbol{\beta}^*\right)^\top\mathbf{X}_l\middle|\mathcal{F}_l\right\|_{\Psi_1}\leq\sqrt{\lambda_{\max}}\sigma\|\boldsymbol{\beta}_l-\boldsymbol{\beta}^*\|.$$

此外，在事件 $\mathcal{E}_l$, $\|\boldsymbol{\beta}_l-\boldsymbol{\beta}^*\|^2\leq C^*(1-c)^{2l}\|\boldsymbol{\beta}_l-\boldsymbol{\beta}^*\|^2+2\frac{\lambda_{\max}}{\lambda_{\min}}\eta\sigma^2 d$ 下，我们使用技术性引理 9。因此，在事件 $\cup_{l=0}^{t}\mathcal{E}_l$ 下，我们有

$$\begin{aligned}\sum_{l=0}^{t}\left\|(1-\eta\lambda_{\min})^{t-l}I_1^{(l)}\middle|\mathcal{F}_l\right\|_{\Psi_1}^2 &\leq C\eta^2\sigma^2\lambda_{\max}\sum_{l=0}^{t}(1-\eta\lambda_{\min})^{2t-2l}((1-c)^{2l}\|\boldsymbol{\beta}_0-\boldsymbol{\beta}^*\|^2+\frac{\lambda_{\max}}{\lambda_{\min}}d\eta\sigma^2)\\ &\leq C\eta^2(1-c)^{2t}\frac{1}{1-\frac{1-\eta\lambda_{\min}}{1-c}}\|\boldsymbol{\beta}_0-\boldsymbol{\beta}^*\|^2+d\eta^3\sigma^4\frac{\lambda_{\max}^2}{\lambda_{\min}(1-(1-\eta\lambda_{\min})^2)}\\ &\leq C\frac{\eta\lambda_{\max}\sigma^2}{\lambda_{\min}}(1-c)^{2t}\|\boldsymbol{\beta}_0-\boldsymbol{\beta}^*\|^2+d\frac{\lambda_{\max}^2\eta^2\sigma^4}{\lambda_{\min}^2},\end{aligned}$$

其中最后一行由 $4c\leq\eta\lambda_{\min}$ 得出。此外，我们有

$$\max_l\left\|(1-\eta\lambda_{\min})^{t-l}I_1^{(l)}\middle|\mathcal{F}_l\right\|_{\Psi_1}\leq C\eta\sigma\sqrt{\lambda_{\max}}\left((1-c)^t\|\boldsymbol{\beta}_0-\boldsymbol{\beta}^*\|+\sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}}\sqrt{\eta}\sqrt{d}\sigma\right).$$

因此，根据引理 9，我们得到

$$\mathbb{P}\left(\left|\sum_{l=0}^{t}(1-\eta\lambda_{\min})^{t-l}I_1^{(l)}\right|\geq s\right)\leq\exp\left(-\min\left\{\frac{s^2}{C\left(\frac{\sigma^2\eta\lambda_{\max}}{\lambda_{\min}}(1-c)^{2t}\|\boldsymbol{\beta}_0-\boldsymbol{\beta}^*\|^2+d\frac{\eta^2\sigma^4\lambda_{\max}^2}{\lambda_{\min}^2}\right)},\right.\right.$$
$$\left.\left.\frac{s}{\eta\sigma\sqrt{\lambda_{\max}}\left((1-c)^t\|\boldsymbol{\beta}_0-\boldsymbol{\beta}^*\|+\sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}}\sqrt{\eta}\sqrt{d}\sigma\right)}\right\}\right).$$

将 $s=(1-c)^{2t+2}\|\boldsymbol{\beta}_0-\boldsymbol{\beta}^*\|^2+\frac{\lambda_{\max}}{\lambda_{\min}}\eta\sigma^2 d \geq 2(1-c)^{t+1}\sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}\sqrt{\eta d}\|\boldsymbol{\beta}_0-\boldsymbol{\beta}^*\|}\sigma$ 代入上述方程，得到

$$\mathbb{P}\left(\left|\sum_{l=0}^{t}(1-\eta\lambda_{\min})^{t-l}I_1^{(l)}\right|\geq(1-c)^{2t+2}\|\boldsymbol{\beta}_0-\boldsymbol{\beta}^*\|^2+\frac{\lambda_{\max}}{\lambda_{\min}}\eta\sigma^2 d\right)$$
$$\leq\exp\left(-c_1\min\left\{d,\sqrt{\frac{1}{\lambda_{\min}}\frac{d}{\eta}}\right\}\right)$$

接下来考虑项 $I_{2\circ}$ 假设 1 保证了，

$$\left\|(\boldsymbol{\beta}_l-\boldsymbol{\beta}^*)^\top\left\{\mathbf{X}_l\mathbf{X}_l^\top-\mathbb{E}\left\{\mathbf{X}_l\mathbf{X}_l^\top|\mathcal{F}_l\right\}\right\}(\boldsymbol{\beta}_l-\boldsymbol{\beta}^*)\,|\mathcal{F}_l\right\|_{\Psi_1}\leq\lambda_{\max}\|\boldsymbol{\beta}_l-\boldsymbol{\beta}^*\|^2.$$

在事件 $\cup_{l=0}^t\mathcal{E}_l$下，我们有

$$\sum_{l=0}^{t}\left\|(1-\eta\lambda_{\min})^{t-l}I_2^{(l)}|\mathcal{F}_l\right\|_{\Psi_1}^2\leq C\eta^2\lambda_{\max}^2\sum_{l=0}^{t}(1-\eta\lambda_{\min})^{2t-2l}\left((1-c)^{4l}\|\boldsymbol{\beta}_0-\boldsymbol{\beta}^*\|^4+\frac{\lambda_{\max}^2}{\lambda_{\min}^2}\eta^2\sigma^4d^2\right)$$
$$\leq C\eta\frac{\lambda_{\max}^2}{\lambda_{\min}}(1-c)^{4t}\|\boldsymbol{\beta}_0-\boldsymbol{\beta}^*\|^4+C\frac{\lambda_{\max}^4}{\lambda_{\min}^3}\eta^3\sigma^4d^2,$$

and

$$\max_l\left\|(1-\eta\lambda_{\min})^{t-l}I_2^{(l)}|\mathcal{F}_l\right\|_{\Psi_1}\leq C\eta\lambda_{\max}(1-c)^{2t}\|\boldsymbol{\beta}_0-\boldsymbol{\beta}^*\|^2+\eta^2\frac{\lambda_{\max}^2}{\lambda_{\min}}\sigma^2d.$$

接下来根据引理 9 并选取 $s=(1-c)^{2t+2}\|\boldsymbol{\beta}_0-\boldsymbol{\beta}^*\|^2+\frac{\lambda_{\max}}{\lambda_{\min}}\eta\sigma^2 d,$ 我们有

$$\mathbb{P}\left(\left|\sum_{l=0}^{t}(1-\eta\lambda_{\min})^{t-l}I_1^{(l)}\right|\geq(1-c)^{2t+2}\|\boldsymbol{\beta}_0-\boldsymbol{\beta}^*\|^2+\frac{\lambda_{\max}}{\lambda_{\min}}\eta\sigma^2d\right)\leq\exp\left(-c_1\frac{\lambda_{\min}}{\lambda_{\max}^2}\frac{1}{\eta}\right).$$

接下来考虑 $I_{3\circ}$ 在 $\mathcal{E}_l^X$下，根据假设 2，我们有

$$\left\|\left\{\xi_l^2\|\mathbf{X}_l\|^2-\mathbb{E}\left\{\xi_l^2\|\mathbf{X}_l\|^2|\mathcal{F}_l\right\}\right\}|\mathcal{F}_l\right\|_{\Psi_1}\leq C\lambda_{\max}\sigma^2d.$$

在事件 $\cup_{l=0}^t\mathcal{E}_l^X$下，我们有

$$\sum_{l=0}^{t}\left\|(1-\eta\lambda_{\min})^{t-l}I_3^{(l)}|\mathcal{F}_l\right\|_{\Psi_1}^2\leq\sum_{l=0}^{t}(1-\eta\lambda_{\min})^{2t-2l}\eta^4\lambda_{\max}^2\sigma^4d^2\leq\frac{\lambda_{\max}^2}{\lambda_{\min}}\eta^3d^2\sigma^4,$$

and

$$\max_l\left\|(1-\eta\lambda_{\min})^{t-l}I_3^{(l)}|\mathcal{F}_l\right\|_{\Psi_1}\leq\eta^2\lambda_{\max}\sigma^2d.$$

根据引理9，然后我们插入 $s=(1-c)^{2t+2}\|\boldsymbol{\beta}_0-\boldsymbol{\beta}^*\|^2+\frac{\lambda_{\max}}{\lambda_{\min}}\eta\sigma^2 d,$

$$\mathbb{P}\left(\left|\sum_{l=0}^{t}(1-\eta\lambda_{\min})^{t-l}I_3^{(l)}\right|\geq(1-c)^{2t+2}\|\boldsymbol{\beta}_0-\boldsymbol{\beta}^*\|^2+\frac{\lambda_{\max}}{\lambda_{\min}}\eta\sigma^2d\right)\leq\exp\left(-c_1\frac{1}{\lambda_{\max}}\frac{1}{\eta}\right).$$

最后，我们考虑 I_4。在事件 $\mathcal{E}_t^X$ 下，我们有

$$\left\|\left\{\left(\mathbf{X}_l^\top\boldsymbol{\beta}_l-\mathbf{X}_t^\top\boldsymbol{\beta}^*\right)^2\|\mathbf{X}_l\|^2-\mathbb{E}\left\{\left(\mathbf{X}_l^\top\boldsymbol{\beta}_l-\mathbf{X}_l^\top\boldsymbol{\beta}^*\right)^2\|\mathbf{X}_l\|^2|\mathcal{F}_l\right\}\right\}|\mathcal{F}_l\right\|_{\Psi_1}\leq C\lambda_{\max}^2 d\|\boldsymbol{\beta}_l-\boldsymbol{\beta}^*\|^2.$$

在事件 $\cup_{l=0}^t\mathcal{E}_l^X$下，我们有

$$\begin{aligned}\sum_{l=0}^t\left\|(1-\eta\lambda_{\min})^{t-l}I_4^{(l)}|\mathcal{F}_l\right\|_{\Psi_1}^2&\leq C\eta^4\lambda_{\max}^4 d^2\sum_{l=0}^t(1-\eta\lambda_{\min})^{2t-2l}\left((1-c)^{4l}\|\boldsymbol{\beta}_0-\boldsymbol{\beta}^*\|^4+\frac{\lambda_{\max}^2}{\lambda_{\min}^2}\eta^2\sigma^4d^2\right)\\&\leq C\frac{\lambda_{\max}^4}{\lambda_{\min}}\eta^3d^2\|\boldsymbol{\beta}_0-\boldsymbol{\beta}^*\|^4+C\frac{\lambda_{\max}^6}{\lambda_{\min}^3}\eta^5\sigma^4d^4,\end{aligned}$$

and

$$\max_l\left\|(1-\eta\lambda_{\min})^{t-l}I_4^{(l)}|\mathcal{F}_l\right\|_{\Psi_1}\leq C\eta^2\lambda_{\max}^2d(1-c)^{2t}\|\boldsymbol{\beta}_0-\boldsymbol{\beta}^*\|^2+C\eta^3(1-c)^{2t}\frac{\lambda_{\max}^3}{\lambda_{\min}}\sigma^2d^2.$$

将 $s=(1-c)^{2t+2}\|\boldsymbol{\beta}_0-\boldsymbol{\beta}^*\|^2+\frac{\lambda_{\max}}{\lambda_{\min}}\eta\sigma^2d$ 代入引理 9，得到

$$\begin{aligned}&\mathbb{P}\left(\left|\sum_{l=0}^t(1-\eta\lambda_{\min})^{t-l}I_4^{(l)}\right|\geq(1-c)^{2t+2}\|\boldsymbol{\beta}_0-\boldsymbol{\beta}^*\|^2+\frac{\lambda_{\max}}{\lambda_{\min}}\eta\sigma^2d\right)\\&\qquad\leq\exp\left(-\min\left\{\frac{\lambda_{\min}}{\lambda_{\max}^4\eta^3d^2},\frac{1}{\lambda_{\max}^2\eta^2d}\right\}\right).\end{aligned}$$

因此，得到方程（18）：

$$\|\boldsymbol{\beta}_{t+1}-\boldsymbol{\beta}^*\|^2\leq5(1-c)^{2t+2}\|\boldsymbol{\beta}_0-\boldsymbol{\beta}^*\|^2+5\frac{\lambda_{\max}}{\lambda_{\min}}\eta\sigma^2d,$$

with 概率大于 $1-5\exp(-c_1d)$。且 c 是满足 $c\leq\eta\lambda_{\min}/2$的任意常数 . □

B 第 3 节中的证明3

本节给出稀疏回归证明，具体包括引理 1、引理 2和命题 1。

B.1 引理 1

首先，具有固定支集的 SGD 更新 $\mathcal{S}_0$ 可写为：

$$\boldsymbol{\beta}_{t+1}-\boldsymbol{\beta}^*=\mathcal{H}_{\mathcal{S}_0}\left(\boldsymbol{\beta}_t-\eta_t\cdot\left(\mathbf{X}_t^\top\boldsymbol{\beta}_t-Y_t\right)\mathbf{X}_t\right)-\boldsymbol{\beta}^*=\boldsymbol{\beta}_t-\boldsymbol{\beta}^*-\eta_t\cdot\left(\mathbf{X}_t^\top\boldsymbol{\beta}_t-Y_t\right)\mathcal{H}_{\mathcal{S}_0}(\mathbf{X}_t).$$

值得注意的是， $\boldsymbol{\beta}_t-\boldsymbol{\beta}^*$ 的支撑集为 $\mathcal{S}^*\cup\mathcal{S}_0$，其中 $\mathcal{S}^*\cup\mathcal{S}_0=(\mathcal{S}^*\setminus\mathcal{S}_0)\cup\mathcal{S}_0$。一方面，限制在 $\mathcal{S}^*\setminus\mathcal{S}_0$ 上的迭代量满足

$$[\boldsymbol{\beta}_{t+1}-\boldsymbol{\beta}^*]_{\mathcal{S}^*\setminus\mathcal{S}_0}=[\boldsymbol{\beta}_t-\boldsymbol{\beta}^*]_{\mathcal{S}^*\setminus\mathcal{S}_0}=[-\boldsymbol{\beta}^*]_{\mathcal{S}^*\setminus\mathcal{S}_0},$$

而对于 $\mathcal{S}_0$上的元素，则有

$$\left[\boldsymbol{\beta}_{t+1}-\boldsymbol{\beta}^*\right]_{\mathcal{S}_0}=[\boldsymbol{\beta}_t-\boldsymbol{\beta}^*]_{\mathcal{S}_0}-\eta_t[\mathbf{X}_t]_{\mathcal{S}_0}(\mathbf{X}_t^\top\boldsymbol{\beta}_t-Y_t).$$

我们将通过归纳法证明引理 1 。定义事件

$$\mathcal{E}_t:=\left\{\|[\boldsymbol{\beta}_t-\boldsymbol{\beta}^*]_{\mathcal{S}_0}\|^2\le C^*\frac{s\log(2d/s)+\delta_t}{t+C_bs\log(2d/s)}\frac{\sigma^2}{\lambda_{\min}}+C\left(\frac{\lambda_{\max,s}^{\mathsf{Off}}}{\lambda_{\min}}\right)^2\cdot\|[\boldsymbol{\beta}^*]_{\mathcal{S}^*\setminus\mathcal{S}_0}\|^2\right\},$$

其中 C^* 是一个不依赖于 t、d、s的常数，其闭式见引理 1。显然， $\mathcal{E}_0$ 成立。我们将在 $\cup_{l=0}^t\mathcal{E}_l$ 下证明， $\mathcal{E}_{t+1}$ 以高概率成立。首先，支撑集为 $\mathcal{S}_0$ 的平方估计误差具有

$$\begin{aligned}\|[\boldsymbol{\beta}_{t+1}-\boldsymbol{\beta}^*]_{\mathcal{S}_0}\|^2=\|[\boldsymbol{\beta}_t-\boldsymbol{\beta}^*]_{\mathcal{S}_0}\|^2-2\eta_t(\mathbf{X}_t^\top\boldsymbol{\beta}_t-Y_t)[\boldsymbol{\beta}_t-\boldsymbol{\beta}^*]_{\mathcal{S}_0}^\top[\mathbf{X}_t]_{\mathcal{S}_0}\\+\eta_t^2(\mathbf{X}_t^\top\boldsymbol{\beta}_t-Y_t)^2\|[\mathbf{X}_t]_{\mathcal{S}_0}\|^2.\end{aligned}$$

我们注意到，$\mathbf{X}_t$ 可能依赖于 $\mathcal{S}_0$。根据假设 3 和引理 13，我们有$\mathbb{E}\{\|[\mathbf{X}_t]_{\mathcal{S}_0}\|^4|\mathcal{F}_t\}\le s^2\log^2(2d/s)\lambda_{\max}^2$ 此外，我们还有

$$\begin{aligned}&\mathbb{E}\left\{(\mathbf{X}_t^\top\boldsymbol{\beta}_t-Y_t)^2\|[\mathbf{X}_t]_{\mathcal{S}_0}\|^2\big|\mathcal{F}_t\right\}=\mathbb{E}\left\{(\mathbf{X}_t^\top\boldsymbol{\beta}_t-\mathbf{X}_t^\top\boldsymbol{\beta}^*)^2\|[\mathbf{X}_t]_{\mathcal{S}_0}\|^2\big|\mathcal{F}_t\right\}+\mathbb{E}\left\{\xi_t^2\|[\mathbf{X}_t]_{\mathcal{S}_0}\|^2\big|\mathcal{F}_t\right\}\\&\quad\le\frac{s\log(2d/s)}{2\|\boldsymbol{\beta}_t-\boldsymbol{\beta}^*\|^2}\mathbb{E}\left\{(\mathbf{X}_t\boldsymbol{\beta}_t-\mathbf{X}_t^\top\boldsymbol{\beta}^*)^4\big|\mathcal{F}_t\right\}+\frac{\|\boldsymbol{\beta}_t-\boldsymbol{\beta}^*\|^2}{2s\log(2d/s)}\mathbb{E}\left\{\|[\mathbf{X}_t]_{\mathcal{S}_0}\|^4\big|\mathcal{F}_t\right\}+\sigma^2s\log(2d/s)\lambda_{\max}\\&\quad\le\lambda_{\max}^2s\log(2d/s)\|\boldsymbol{\beta}_t-\boldsymbol{\beta}^*\|^2+2\sigma^2s\log(2d/s)\lambda_{\max}\le3\sigma^2s\log(2d/s)\lambda_{\max},\end{aligned}$$

其中第二行使用了不等式 $a^2+b^2\ge2ab$。因此，$\|\boldsymbol{\beta}_{t+1}-\boldsymbol{\beta}^*\|^2$ 的条件期望满足以下方程，

$$\begin{aligned}\mathbb{E}\left\{\left\|[\boldsymbol{\beta}_{t+1}-\boldsymbol{\beta}^*]_{\mathcal{S}_0}\right\|^2\big|\mathcal{F}_t\right\}\le(1-2\eta_t\lambda_{\min})\left\|[\boldsymbol{\beta}_t-\boldsymbol{\beta}^*]_{\mathcal{S}_0}\right\|^2+3\eta_t^2\lambda_{\max}\cdot s\log(2d/s)\sigma^2\\+2\eta_t\lambda_{\max,s}^{\mathsf{Off}}\cdot\|\left[\boldsymbol{\beta}_t-\boldsymbol{\beta}^*\right]_{\mathcal{S}_0}\|\cdot\|[\boldsymbol{\beta}^*]_{\mathcal{S}^*\setminus\mathcal{S}_0}\|.\end{aligned}$$

我们用 $2\lambda_{\max,s}^{\mathsf{Off}}\cdot\|\left[\boldsymbol{\beta}_t-\boldsymbol{\beta}^*\right]_{\mathcal{S}_0}\|\cdot\|[\boldsymbol{\beta}^*]_{\mathcal{S}^*\setminus\mathcal{S}_0}\|\le\lambda_{\min}\|[\boldsymbol{\beta}_t-\boldsymbol{\beta}^*]_{\mathcal{S}_0}\|^2+(\lambda_{\max,s}^{\mathsf{Off}})^2\|[\boldsymbol{\beta}^*]_{\mathcal{S}^*\setminus\mathcal{S}_0}\|^2/\lambda_{\min}$ 给出右侧第三项的界。化简后，得到

$$\begin{aligned}\mathbb{E}\left\{\left\|[\boldsymbol{\beta}_{t+1}-\boldsymbol{\beta}^*]_{\mathcal{S}_0}\right\|^2\big|\mathcal{F}_t\right\}\le(1-\eta_t\lambda_{\min})\left\|[\boldsymbol{\beta}_t-\boldsymbol{\beta}^*]_{\mathcal{S}_0}\right\|^2\\+3\eta_t^2\lambda_{\max}s\log(2d/s)\sigma^2+\eta_t\|[\boldsymbol{\beta}^*]_{\mathcal{S}^*\setminus\mathcal{S}_0}\|^2\cdot\frac{(\lambda_{\max,s}^{\mathsf{Off}})^2}{\lambda_{\min}}.\end{aligned}$$

接着，我们将平方估计误差与条件期望之间的差进行分解，

$$
\begin{aligned}
&\|[\boldsymbol{\beta}_{t+1}-\boldsymbol{\beta}^*]_{\mathcal{S}_0}\|^2-\mathbb{E}\left\{\left\|[\boldsymbol{\beta}_{t+1}-\boldsymbol{\beta}^*]_{\mathcal{S}_0}\right\|^2|\mathcal{F}_t\right\}\\
&\quad=-2\eta_t\underbrace{\left\{[\boldsymbol{\beta}_t-\boldsymbol{\beta}^*]_{\mathcal{S}_0}^\top[\mathbf{X}_t]_{\mathcal{S}_0}\mathbf{X}_t^\top(\boldsymbol{\beta}_t-\boldsymbol{\beta}^*)-\mathbb{E}\left\{[\boldsymbol{\beta}_t-\boldsymbol{\beta}^*]_{\mathcal{S}_0}^\top[\mathbf{X}_t]_{\mathcal{S}_0}\mathbf{X}_t^\top(\boldsymbol{\beta}_t-\boldsymbol{\beta}^*)|\mathcal{F}_t\right\}\right\}}_{\boldsymbol{\Xi}_1^{(t)}}\\
&\quad+2\eta_t\underbrace{\left\{\xi_t[\boldsymbol{\beta}_t-\boldsymbol{\beta}^*]_{\mathcal{S}_0}^\top[\mathbf{X}_t]_{\mathcal{S}_0}-\mathbb{E}\left\{\xi_t[\boldsymbol{\beta}_t-\boldsymbol{\beta}^*]_{\mathcal{S}_0}^\top[\mathbf{X}_t]_{\mathcal{S}_0}|\mathcal{F}_t\right\}\right\}}_{\boldsymbol{\Xi}_2^{(t)}}\\
&\quad+\eta_t^2\underbrace{\left\{(\mathbf{X}_t^\top\boldsymbol{\beta}_t-Y_t)^2\|[\mathbf{X}_t]_{\mathcal{S}_0}\|^2-\mathbb{E}\left\{(\mathbf{X}_t^\top\boldsymbol{\beta}_t-Y_t)^2\|[\mathbf{X}_t]_{\mathcal{S}_0}\|^2|\mathcal{F}_t\right\}\right\}}_{\boldsymbol{\Xi}_3^{(t)}}.
\end{aligned}
$$

显然，这三个项 $\boldsymbol{\Xi}_i^{(t)}$ 都是鞅差，其条件 Orlicz 范数有界，且

$$
\|\boldsymbol{\Xi}_1^{(t)}|\mathcal{F}_t\|_{\Psi_1}\le\lambda_{\max}\|[\boldsymbol{\beta}_t-\boldsymbol{\beta}^*]_{\mathcal{S}_0}\|^2+\lambda_{\max,s}^{\mathsf{Off}}\cdot\|[\boldsymbol{\beta}_t-\boldsymbol{\beta}^*]_{\mathcal{S}_0}\|\cdot\|[\boldsymbol{\beta}^*]_{\mathcal{S}^*\setminus\mathcal{S}_0}\|,
$$

and

$$
\|\boldsymbol{\Xi}_2^{(t)}|\mathcal{F}_t\|_{\Psi_1}\le\sqrt{\lambda_{\max}}\sigma\|[\boldsymbol{\beta}_t-\boldsymbol{\beta}^*]_{\mathcal{S}_0}\|.
$$

记作 $\mathcal{E}_t^X:=\{\|[\mathbf{X}_t]_{\mathcal{S}_0}\|^2\le C\lambda_{\max}(s\log(2d/s)+\delta_t)\}$，由引理 13，可得 $\mathbb{P}(\mathcal{E}_t^X)\ge1-\exp(-c_2(s\log(2d/s)+\delta_t))$。因此，在 $\mathcal{E}_t^X$ 条件下，有

$$
\|\boldsymbol{\Xi}_3^{(t)}|\mathcal{F}_t,\mathcal{E}_t^X\|_{\Psi_1}\le C(s\log(2d/s)+\delta_t)\lambda_{\max}\sigma^2.
$$

因此，将条件期望与差值结合起来，可得

$$
\begin{aligned}
\|[\boldsymbol{\beta}_{t+1}-\boldsymbol{\beta}^*]_{\mathcal{S}_0}\|^2&=\mathbb{E}\left\{\|[\boldsymbol{\beta}_{t+1}-\boldsymbol{\beta}^*]_{\mathcal{S}_0}\|^2|\mathcal{F}_t\right\}+\left\{\|[\boldsymbol{\beta}_{t+1}-\boldsymbol{\beta}^*]_{\mathcal{S}_0}\|^2-\mathbb{E}\left\{\|[\boldsymbol{\beta}_{t+1}-\boldsymbol{\beta}^*]_{\mathcal{S}_0}\|^2|\mathcal{F}_t\right\}\right\}\\
&\le(1-\eta_t\lambda_{\min})\left\|[\boldsymbol{\beta}_t-\boldsymbol{\beta}^*]_{\mathcal{S}_0}\right\|^2+3\eta_t^2\lambda_{\max}s\log(2d/s)\sigma^2\\
&\quad+\eta_t\|[\boldsymbol{\beta}^*]_{\mathcal{S}^*\setminus\mathcal{S}_0}\|^2\cdot\frac{(\lambda_{\max,s}^{\mathsf{Off}})^2}{\lambda_{\min}}-2\eta_t\boldsymbol{\Xi}_1^{(t)}+2\eta_t\boldsymbol{\Xi}_2^{(t)}+\eta_t^2\boldsymbol{\Xi}_3^{(t)}.
\end{aligned}
$$

将步长 $\eta_t = \frac{C_a}{\lambda_{\min}\, t+C_b s}^{1}$ ~~$\log(2d/s)$~~ 代入上述方程，可得

$$
\begin{aligned}
\|[\boldsymbol{\beta}_{t+1}-\boldsymbol{\beta}^*]_{\mathcal{S}_0}\|^2 \le & \underbrace{\prod_{j=0}^{t}\left(1-\frac{C_a}{j+C_b s\log(2d/s)}\right)\left\|[\boldsymbol{\beta}_0-\boldsymbol{\beta}^*]_{\mathcal{S}_0}\right\|^2}_{\boldsymbol{A}_1} \\
& \underbrace{+3s\log(2d/s)C_a^2\frac{\lambda_{\max}}{\lambda_{\min}^2}\sum_{l=0}^{t}\prod_{j=l}^{t}\left(1-\frac{C_a}{j+C_b s\log(2d/s)}\right)\left(\frac{1}{l+C_b s\log(2d/s)}\right)^2\sigma^2}_{\boldsymbol{A}_2} \\
& \underbrace{+C_a\sum_{l=0}^{t}\prod_{j=l}^{t}\left(1-\frac{C_a}{j+C_b s\log(2d/s)}\right)\frac{1}{l+C_b s\log(2d/s)}\|[\boldsymbol{\beta}^*]_{\mathcal{S}^*\setminus\mathcal{S}_0}\|^2\cdot\frac{(\lambda^{\mathsf{Off}}_{\max,s})^2}{\lambda_{\min}^2}}_{\boldsymbol{A}_3} \\
& \underbrace{-2\frac{C_a}{\lambda_{\min}}\sum_{l=0}^{t}\prod_{j=l}^{t}\left(1-\frac{C_a}{j+C_b s\log(2d/s)}\right)\frac{1}{l+C_b s\log(2d/s)}\boldsymbol{\Xi}_1^{(l)}}_{\boldsymbol{A}_4} \\
& \underbrace{+2\frac{C_a}{\lambda_{\min}}\sum_{l=0}^{t}\prod_{j=l}^{t}\left(1-\frac{C_a}{j+C_b s\log(2d/s)}\right)\frac{1}{l+C_b s\log(2d/s)}\boldsymbol{\Xi}_2^{(l)}}_{\boldsymbol{A}_5} \\
& \underbrace{+\frac{C_a^2}{\lambda_{\min}^2}\sum_{l=0}^{t}\prod_{j=l}^{t}\left(1-\frac{C_a}{j+C_b s\log(2d/s)}\right)\frac{1}{(l+C_b s\log(2d/s))^2}\boldsymbol{\Xi}_3^{(l)}}_{\boldsymbol{A}_6}.
\end{aligned}
$$

接下来分别对每一项进行界定。首先，考虑 $\mathbf{A}_1$。根据引理 16，我们有

$$
\prod_{j=0}^{t}\left(1-\frac{C_a}{j+C_b s\log(2d/s)}\right)\le\left(\frac{C_b s\log(2d/s)}{t+1+C_b s\log(2d/s)}\right)^{C_a},
$$

这意味着 $\mathrm{A}_1 \le {}^{C_b s}\log(2d/s)\, t+C_b s\log(2d/s)\|[\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*]_{\mathcal{S}_0}\|^2$。接下来考虑项 A_2。根据引理 16，我们有 $\prod_{j=l}^{t}\left(1-{}_{j+C_b s}{}^{C_a}\log(2d/s)\right)\left(l+C_b s\right)^{1}\log(2d/s)\le(l+C_b s\log(2d/s))^{C_a-2}(t+1+C_b s\log(2d/s))^{C_a}$，而引理17进一步推出 $\sum_{l=0}^{t}\prod_{j=l}^{t}\left(1-{}_{j+C_b s}{}^{C_a}\log(2d/s)\right)\,{}_{l+C_b s}{}^{1}(\log(2d/s))^2 \le {}_{t+1+C_b s}{}^{2/C_a}\log(2d/s)$。因此，我们有

$$
\boldsymbol{A}_2 \le C\frac{\lambda_{\max}}{\lambda_{\min}^2}\cdot\frac{C_a s\log(2d/s)}{t+1+C_b s\log(2d/s)}\sigma^2.
$$

对于项 $\boldsymbol{A}_3$，类似地，我们有

$$
\boldsymbol{A}_3 \le C\frac{(\lambda^{\mathsf{Off}}_{\max,s})^2}{\lambda_{\min}^2}\cdot\|[\boldsymbol{\beta}^*]_{\mathcal{S}^*\setminus\mathcal{S}_0}\|^2.
$$

接下来我们考虑 $\boldsymbol{A}_4$。我们注意到

$$\left\|\prod_{j=l}^{t}\left(1-\frac{C_a}{j+C_bs\log(2d/s)}\right)\frac{1}{l+C_bs\log(2d/s)}\boldsymbol{\Xi}_1^{(l)}\middle|\mathcal{F}_l\right\|_{\Psi_1}$$
$$\leq\frac{(l+C_bs\log(2d/s))^{C_a-1}}{(t+1+C_bs\log(2d/s))^{C_a}}\left(\lambda_{\max}\|[\boldsymbol{\beta}_l-\boldsymbol{\beta}^*]_{\mathcal{S}_0}\|^2+\frac{(\lambda_{\max,s}^{\mathsf{Off}})^2}{\lambda_{\max}}\|[\boldsymbol{\beta}^*]_{\mathcal{S}^*\setminus\mathcal{S}_0}\|^2\right).$$

然后根据引理 9，在事件 $\cup_{l=0}^{t}\mathcal{E}_l$下，我们有

$$\mathbb{P}\left(\left|\sum_{l=0}^{t}\prod_{j=l}^{t}\left(1-\frac{C_a}{j+C_bs\log(2d/s)}\right)\frac{1}{l+C_bs\log(2d/s)}\boldsymbol{\Xi}_1^{(l)}\right|\geq c_1\frac{C^*}{C_a}\frac{s\log(2d/s)+\delta_{t+1}}{t+1+C_bs}\sigma^2\right.$$
$$\left.+c_1\frac{(\lambda_{\max,s}^{\mathsf{Off}})^2}{C_a\lambda_{\max}}\|[\boldsymbol{\beta}^*]_{\mathcal{S}^*\setminus\mathcal{S}_0}\|^2\right)\leq\exp\left(-c_2\frac{C_bs\log(2d/s)+t}{C_a}\frac{\lambda_{\min}^2}{\lambda_{\max}^2}\right),$$

其中 $c_1<0.1$ 是一个足够小的常数，进而可得

$$|\boldsymbol{A}_4|\leq c_1C^*\frac{s\log(2d/s)+\delta_t}{t+1+C_bs\log(2d/s)}\frac{\sigma^2}{\lambda_{\min}}+c_1\left(\frac{\lambda_{\max,s}^{\mathsf{Off}}}{\lambda_{\min}}\right)^2\|[\boldsymbol{\beta}^*]_{\mathcal{S}\setminus\mathcal{S}_0}\|^2.$$

关于项 $\boldsymbol{A}_5$, ，我们有

$$\left\|\prod_{j=l}^{t}\left(1-\frac{C_a}{j+C_bs\log(2d/s)}\right)\frac{1}{l+C_bs\log(2d/s)}\boldsymbol{\Xi}_2^{(l)}|\mathcal{F}_l\right\|_{\Psi_1}$$
$$\leq C\sqrt{\lambda_{\max}}\sigma\frac{(l+C_bs\log(2d/s))^{C_a-1}}{(t+1+C_bs\log(2d/s))^{C_a}}\cdot\|[\boldsymbol{\beta}_l-\boldsymbol{\beta}^*]_{\mathcal{S}_0}\|.$$

那么，在事件 $\cup_{l=0}^{t}\{\mathcal{E}_l\}$ 下，根据引理 9，我们有

$$\mathbb{P}\left(\left|\sum_{l=0}^{t}\prod_{j=l}^{t}\left(1-\frac{C_a}{j+C_bs\log(2d/s)}\right)\frac{1}{l+C_bs\log(2d/s)}\boldsymbol{\Xi}_2^{(l)}\right|\right.$$
$$\left.\geq c_1\frac{C^*}{C_a}\frac{(s\log(2d/s)+\delta_{t+1})\sigma^2}{t+1+C_bs\log(2d/s)}+c_1\frac{1}{C_a}\frac{(\lambda_{\max,s}^{\mathsf{Off}})^2}{\lambda_{\min}}\|[\boldsymbol{\beta}^*]_{\mathcal{S}^*\setminus\mathcal{S}_0}\|^2\right)$$
$$\leq\exp\left(-c_2\min\left\{(s\log(2d/s)+\delta_{t+1})\frac{C^*}{C_a}\frac{\lambda_{\min}}{\lambda_{\max}},\sqrt{s\log(2d/s)+\delta_{t+1}}\sqrt{C_bs\log(2d/s)+t}\frac{\sqrt{C^*}}{C_a}\frac{\sqrt{\lambda_{\min}}}{\sqrt{\lambda_{\max}}}\right\}\right),$$

在此条件下，我们有，

$$|\boldsymbol{A}_5|\leq c_1\frac{C^*}{\lambda_{\min}}\frac{s\log(2d/s)+\delta_{t+1}}{t+1+C_bs\log(2d/s)}\sigma^2+c_1\left(\frac{\lambda_{\max,s}^{\mathsf{Off}}}{\lambda_{\min}}\right)^2\|[\boldsymbol{\beta}^*]_{\mathcal{S}^*\setminus\mathcal{S}_0}\|^2.$$

接下来，我们只需考虑项 $\boldsymbol{A}_6$,；在 $\mathcal{E}_l^X$ 下，它满足

$$\left\|\prod_{j=l}^{t}\left(1-\frac{C_a}{j+C_bs\log(2d/s)}\right)\frac{1}{(l+C_bs\log(2d/s))^2}\boldsymbol{\Xi}_3^{(l)}\middle|\mathcal{F}_l,\mathcal{E}_l^X\right\|_{\Psi_1}$$
$$\leq C\frac{(l+C_bs\log(2d/s))^{C_a-2}}{(t+1+C_bs\log(2d/s))^{C_a}}(s\log(2d/s)+\delta_l)\lambda_{\max}\sigma^2.$$

因此，根据引理9，在 $\cup_{l=0}^t\mathcal{E}_l^X$下，它满足

$$\mathbb{P}\left(\left|\prod_{j=l}^{t}\left(1-\frac{C_a}{j+C_bs\log(2d/s)}\right)\frac{1}{(l+C_bs\log(2d/s))^2}\boldsymbol{\Xi}_3^{(l)}\right|\geq c_1C^*\frac{\lambda_{\min}}{C_a^2}\frac{(s\log(2d/s)+\delta_{t+1})\sigma^2}{t+1+C_bs\log(2d/s)}\right)$$
$$\leq\exp\left(-c_2\left(s\log(2d/s)+\frac{t}{C_b}\right)\min\left\{\frac{\lambda_{\min}^2}{\lambda_{\max}^2}\frac{C_b}{C_a}\left(\frac{C^*}{C_a}\right)^2,\frac{\lambda_{\min}}{\lambda_{\max}}\frac{C^*C_b}{C_a^2}\right\}\right),$$

这意味着

$$|\boldsymbol{A}_6|\leq c_1C^*\frac{1}{\lambda_{\min}}\frac{s\log(2d/s)+\delta_{t+1}}{t+1+C_bs\log(2d/s)}\sigma^2.$$

因此，总体而言，在事件 $\cup_{l=0}^t\{\mathcal{E}_l,\mathcal{E}_l^X\}$ 发生且根据 $C_b\geq C_a$ 和 C^*的取值时，以至少 $1-\exp(-c(C_bs\log(2d/s)+t)/C_a)-\exp(-c(s\log(2d/s)+\delta_{t+1}))$的概率，我们有

$$\|[\boldsymbol{\beta}_{t+1}-\boldsymbol{\beta}^*]_{\mathcal{S}_0}\|^2\leq C^*\frac{1}{\lambda_{\min}}\frac{s\log(2d/s)+\delta_{t+1}}{t+1+C_bs\log(2d/s)}\sigma^2+c\frac{(\lambda_{\max,s}^{\mathsf{Off}})^2}{\lambda_{\min}^2}\cdot\|[\boldsymbol{\beta}^*]_{\mathcal{S}^*\setminus\mathcal{S}_0}\|^2,$$

由此完成证明。

B.2 引理2的证明

向量**G统计量**可以写成两个项之和

$$\mathbf{G}(\{\boldsymbol{\beta}_t,\mathbf{X}_t,Y_t\}_{t\in\mathcal{T}})=\sum_{t\in\mathcal{T}}\mathbf{X}_t\mathbf{X}_t^\top(\boldsymbol{\beta}_t-\boldsymbol{\beta}^*)-\sum_{t\in\mathcal{T}}\xi_t\cdot\mathbf{X}_t.\tag{19}$$

我们首先分析第二项，其中包含噪声。值得注意的是，$\xi_t\mathbf{X}_t$ 相对于 $\mathcal{F}_t$是鞅差，并且它具有 $\|\xi_t\mathbf{X}_t|\mathcal{F}_t\|_{\Psi_1}\leq\sigma\sqrt{\lambda_{\max}}$。引理15 证明，对于任意$\delta>0$，以超过$1-d\exp(-c\min\{\delta,\sqrt{\delta|\mathcal{T}|}\})$的概率，我们有

$$\left\|\sum_{t\in\mathcal{T}}\xi_t\mathbf{X}_t\right\|_\infty\leq C\sigma\sqrt{\lambda_{\max}}\sqrt{\delta|\mathcal{T}|}.\tag{20}$$

同样值得注意的是，$\mathbf{X}_t$ 与 $\mathcal{S}_0$之间可能存在依赖关系。根据引理12，对于任意 $\mathcal{S} \subseteq [d]$ 且大小为 $|\mathcal{S}| \leq s$的 $\mathcal{S} \subseteq [d]$，我们有

$$\begin{aligned}\mathbb{P}\left(\left\|\mathcal{H}_{\mathcal{S}}\left(\sum_{t\in\mathcal{T}}\xi_t\mathbf{X}_t\right)\right\| \geq C\sigma\sqrt{\lambda_{\max}}\sqrt{|\mathcal{T}|}\sqrt{s\log(2d/s)+\delta}\right) \\ \leq \exp\left(-c\min\left\{\delta, \sqrt{|\mathcal{T}|/(s\log(2d/s))}\sqrt{\delta}\right\}\right).\end{aligned} \tag{21}$$

那么，我们只需分析（19）中的第一项，即 $\sum_{t\in\mathcal{T}}\mathbf{X}_t\mathbf{X}_t^\top(\boldsymbol{\beta}_t-\boldsymbol{\beta}^*)$。具体而言，我们旨在为位于 $\mathcal{S}_{\mathcal{T}}^c \cap \mathcal{S}^{*c}$ 上的元素提供一个上界，这些元素未被 $\mathcal{S}^*$ 包含；并为支撑集位于 $\mathcal{S}^* \setminus \mathcal{S}_{\mathcal{T}}$ 上的元素提供一个下界。首先，对于 $i \in \mathcal{S}_{\mathcal{T}}^c \cap \mathcal{S}^{*c}$，我们将其分解如下：

$$\begin{aligned}&[\sum_{t\in\mathcal{T}}\mathbf{X}_t\mathbf{X}_t^\top(\boldsymbol{\beta}_t-\boldsymbol{\beta}^*)]_i = \sum_{t\in\mathcal{T}}[\mathbf{X}_t]_i\mathbf{X}_t^\top(\boldsymbol{\beta}_t-\boldsymbol{\beta}^*) \\ &\quad= \underbrace{\sum_{t\in\mathcal{T}}\mathbb{E}\left\{[\mathbf{X}_t]_i\mathbf{X}_t^\top(\boldsymbol{\beta}_t-\boldsymbol{\beta}^*)\,\middle|\,\mathcal{F}_t\right\}}_{\mathbf{D}_1} + \underbrace{\sum_{t\in\mathcal{T}}\left\{[\mathbf{X}_t]_i\mathbf{X}_t^\top(\boldsymbol{\beta}_t-\boldsymbol{\beta}^*) - \mathbb{E}\left\{[\mathbf{X}_t]_i\mathbf{X}_t^\top(\boldsymbol{\beta}_t-\boldsymbol{\beta}^*)\,\middle|\,\mathcal{F}_t\right\}\right\}}_{\mathbf{D}_2}.\end{aligned}$$

我们首先对项 $\mathbf{D}_1$ 作界。回顾 $\lambda_{\max,\,1}^{\mathsf{Off}}$ 的定义。在事件 $\mathcal{E}_{\{\mathcal{S}_{\mathcal{T}},V_{\mathcal{T}},W_{\mathcal{T}}\}}$ 下，我们有

$$\begin{aligned}&\left|\sum_{t\in\mathcal{T}}\mathbb{E}\left\{[\mathbf{X}_t]_i\mathbf{X}_t^\top(\boldsymbol{\beta}_t-\boldsymbol{\beta}^*)\,\middle|\,\mathcal{F}_t\right\}\right| \\ &\qquad\leq \sum_{t\in\mathcal{T}}\left\|\mathbb{E}\left\{[\mathbf{X}_t]_i[\mathbf{X}_t]_{\mathcal{S}^*\cup\mathcal{S}_{\mathcal{T}}}\right\}|\mathcal{F}_t\right\|\cdot\|\boldsymbol{\beta}_t-\boldsymbol{\beta}^*\| \leq \lambda_{\max,1}^{\mathsf{Off}}\sum_{t\in\mathcal{T}}\|\boldsymbol{\beta}_t-\boldsymbol{\beta}^*\| \leq \lambda_{\max,1}^{\mathsf{Off}}\sum_{t\in\mathcal{T}}V_t.\end{aligned}$$

接下来考虑项 $\mathbf{D}_2$。值得注意的是，$\mathbf{D}_2$ 是关于 $\{\mathcal{F}_t\}$ 的 $|\mathcal{T}|$ 个鞅差之和，并且在事件 $\mathcal{E}_{\{\mathcal{S}_{\mathcal{T}},D_{\mathcal{T}},U_{\mathcal{T}}\}}$ 下，$\mathbf{D}_2$ 的每一项都满足

$$\left\|\left\{[\mathbf{X}_t]_i\mathbf{X}_t^\top(\boldsymbol{\beta}_t-\boldsymbol{\beta}^*) - \mathbb{E}\left\{[\mathbf{X}_t]_i\mathbf{X}_t^\top(\boldsymbol{\beta}_t-\boldsymbol{\beta}^*)\,\middle|\,\mathcal{F}_t\right\}\right\}|\mathcal{F}_t\right\|_{\Psi_1} \leq \lambda_{\max,1}^{\mathsf{Off}}V_t.$$

因此，根据引理9，我们有

$$\begin{aligned}\mathbb{P}\left(\left|\sum_{t\in\mathcal{T}}\left\{[\mathbf{X}_t]_i\mathbf{X}_t^\top(\boldsymbol{\beta}_t-\boldsymbol{\beta}^*) - \mathbb{E}\left\{[\mathbf{X}_t]_i\mathbf{X}_t^\top(\boldsymbol{\beta}_t-\boldsymbol{\beta}^*)\,\middle|\,\mathcal{F}_t\right\}\right\}\right| \geq u\right) \\ \leq \exp\left(-\min\left\{\frac{u^2}{(\lambda_{\max,1}^{\mathsf{Off}})^2\sum_{t\in\mathcal{T}}V_t^2}, \frac{u}{2\lambda_{\max,1}^{\mathsf{Off}}\max_t V_t}\right\}\right).\end{aligned}$$

将上述两个方程结合起来，并对 $\mathcal{S}_{\mathcal{T}}^c \cap \mathcal{S}^{*c}$ 取均匀分布，我们有

$$\mathbb{P}\left(\left\|[\sum_{t\in\mathcal{T}}\mathbf{X}_t\mathbf{X}_t^\top(\boldsymbol{\beta}_t-\boldsymbol{\beta}^*)]_{\mathcal{S}^{*c}\cap\mathcal{S}_{\mathcal{T}}^c}\right\|_\infty \geq \lambda_{\max,1}^{\mathsf{Off}}\sum_{t\in\mathcal{T}}V_t+u\right)$$
$$\leq d\exp\left(-\min\left\{\frac{u^2}{(\lambda_{\max,1}^{\mathsf{Off}})^2\sum_{t\in\mathcal{T}}V_t^2},\frac{u}{2\lambda_{\max,1}^{\mathsf{Off}}\max_t V_t}\right\}\right).$$

因此，结合公式（20），我们得到向量 **G统计量**$(\{\boldsymbol{\beta}_t,\mathbf{X}_t,Y_t\}_{t\in\mathcal{T}})$的逐项范数，

$$\mathbb{P}\left(\left\|[\mathbf{G}(\{\boldsymbol{\beta}_t,\mathbf{X}_t,Y_t\}_{t\in\mathcal{T}})]_{\mathcal{S}^{*c}\cap\mathcal{S}_{\mathcal{T}}^c}\right\|_\infty \geq \lambda_{\max,1}^{\mathsf{Off}}\sum_{t\in\mathcal{T}}V_t+u+C\sigma\sqrt{\lambda_{\max}}\sqrt{\delta|\mathcal{T}|}\right)$$
$$\leq d\exp\left(-\min\left\{\frac{u^2}{(\lambda_{\max,1}^{\mathsf{Off}})^2\sum_{t\in\mathcal{T}}V_t^2},\frac{u}{2\lambda_{\max,1}^{\mathsf{Off}}\max_t V_t}\right\}\right)+d\exp(-c\min\{\delta,\sqrt{\delta|\mathcal{T}|}\}).$$

接下来，我们考虑支撑集为 $\mathcal{S}^*\setminus\mathcal{S}_{\mathcal{T}}$ 的向量的下界。我们仍将其分解为条件期望项和鞅差项，

$$[\sum_{t\in\mathcal{T}}\mathbf{X}_t\mathbf{X}_t^\top(\boldsymbol{\beta}_t-\boldsymbol{\beta}^*)]_{\mathcal{S}^*\setminus\mathcal{S}_{\mathcal{T}}}=\underbrace{\sum_{t\in\mathcal{T}}\mathbb{E}\left\{[\mathbf{X}_t]_{\mathcal{S}^*\setminus\mathcal{S}_{\mathcal{T}}}\mathbf{X}_t^\top(\boldsymbol{\beta}_t-\boldsymbol{\beta}^*)\,\middle|\mathcal{F}_t\right\}}_{\mathbf{E}_1}$$
$$+\underbrace{\sum_{t\in\mathcal{T}}\left\{[\mathbf{X}_t]_{\mathcal{S}^*\setminus\mathcal{S}_{\mathcal{T}}}\mathbf{X}_t^\top(\boldsymbol{\beta}_t-\boldsymbol{\beta}^*)-\mathbb{E}\left\{[\mathbf{X}_t]_{\mathcal{S}^*\setminus\mathcal{S}_{\mathcal{T}}}\mathbf{X}_t^\top(\boldsymbol{\beta}_t-\boldsymbol{\beta}^*)\,\middle|\mathcal{F}_t\right\}\right\}}_{\mathbf{E}_2}.$$

我们首先考虑项 $\mathbf{E}_1$。注意到 $\boldsymbol{\beta}_t-\boldsymbol{\beta}^*$ 的支撑集为 $\mathcal{S}_{\mathcal{T}}\cup\mathcal{S}^*$，我们进一步将其分解如下，

$$\left\|\sum_{t\in\mathcal{T}}\mathbb{E}\left\{[\mathbf{X}_t]_{\mathcal{S}^*\setminus\mathcal{S}_{\mathcal{T}}}\mathbf{X}_t^\top(\boldsymbol{\beta}_t-\boldsymbol{\beta}^*)\,\middle|\mathcal{F}_t\right\}\right\|=\left\|\sum_{t\in\mathcal{T}}\mathbb{E}\left\{[\mathbf{X}_t]_{\mathcal{S}^*\setminus\mathcal{S}_{\mathcal{T}}}[\mathbf{X}_t]_{\mathcal{S}^*\cup\mathcal{S}_{\mathcal{T}}}^\top\middle|\mathcal{F}_t\right\}(\boldsymbol{\beta}_t-\boldsymbol{\beta}^*)\right\|$$
$$\geq\left\|\sum_{t\in\mathcal{T}}\mathbb{E}\left\{[\mathbf{X}_t]_{\mathcal{S}^*\setminus\mathcal{S}_{\mathcal{T}}}[\mathbf{X}_t]_{\mathcal{S}^*\setminus\mathcal{S}_{\mathcal{T}}}^\top\middle|\mathcal{F}_t\right\}[\boldsymbol{\beta}^*]_{\mathcal{S}^*\setminus\mathcal{S}_{\mathcal{T}}}\right\|-\left\|\sum_{t\in\mathcal{T}}\mathbb{E}\left\{[\mathbf{X}_t]_{\mathcal{S}^*\setminus\mathcal{S}_{\mathcal{T}}}[\mathbf{X}_t]_{\mathcal{S}_{\mathcal{T}}}^\top\middle|\mathcal{F}_t\right\}[\boldsymbol{\beta}_t-\boldsymbol{\beta}^*]_{\mathcal{S}_{\mathcal{T}}}\right\|.$$

根据假设 3 和引理 14，我们有

$$\left\|\sum_{t\in\mathcal{T}}\mathbb{E}\left\{[\mathbf{X}_t]_{\mathcal{S}^*\setminus\mathcal{S}_{\mathcal{T}}}[\mathbf{X}_t]_{\mathcal{S}^*\setminus\mathcal{S}_{\mathcal{T}}}^\top\middle|\mathcal{F}_t\right\}[\boldsymbol{\beta}^*]_{\mathcal{S}^*\setminus\mathcal{S}_{\mathcal{T}}}\right\|\geq|\mathcal{T}|\cdot\lambda_{\min}\|[\boldsymbol{\beta}^*]_{\mathcal{S}^*\setminus\mathcal{S}_{\mathcal{T}}}\|,$$

另一方面，在事件 $\mathcal{E}_{\{\mathcal{S}_{\mathcal{T}},D_{\mathcal{T}},U_{\mathcal{T}}\}}$ 下，我们有

$$\left\|\sum_{t\in\mathcal{T}}\mathbb{E}\left\{[\mathbf{X}_t]_{\mathcal{S}^*\setminus\mathcal{S}_{\mathcal{T}}}[\mathbf{X}_t]_{\mathcal{S}_{\mathcal{T}}}^\top\middle|\mathcal{F}_t\right\}[\boldsymbol{\beta}_t-\boldsymbol{\beta}^*]_{\mathcal{S}_{\mathcal{T}}}\right\|\leq\lambda_{\max,s}^{\mathsf{Off}}\sum_{t\in\mathcal{T}}W_t.$$

因此综上，我们得到 $\|\mathbf{E}_1\|$ 的下界，

$$\|\mathbf{E}_1\| \geq |\mathcal{T}| \cdot \lambda_{\min} \|[\boldsymbol{\beta}^*]_{\mathcal{S}^* \setminus \mathcal{S}_\mathcal{T}}\| - \lambda_{\max,s}^{\mathsf{Off}} \sum_{t \in \mathcal{T}} W_t.$$

接下来，我们将求出$\mathbf{E}_2$的范数上界。注意到

$$\left\| \left\{ [\mathbf{X}_t]_{\mathcal{S}^* \setminus \mathcal{S}_\mathcal{T}} \mathbf{X}_t^\top (\boldsymbol{\beta}_t - \boldsymbol{\beta}^*) - \mathbb{E}\left\{ [\mathbf{X}_t]_{\mathcal{S}^* \setminus \mathcal{S}_\mathcal{T}} \mathbf{X}_t^\top (\boldsymbol{\beta}_t - \boldsymbol{\beta}^*) \,|\mathcal{F}_t \right\} \right\} |\mathcal{F}_t \right\|_{\Psi_1}$$
$$\leq \lambda_{\max,s}^{\mathsf{Off}} W_t + \lambda_{\max} \|[\boldsymbol{\beta}^*]_{\mathcal{S}^* \setminus \mathcal{S}_\mathcal{T}}\|.$$

因此，根据引理 11，我们有

$$\mathbb{P}\left(\left\| \sum_{t \in \mathcal{T}} \left\{ [\mathbf{X}_t]_{\mathcal{S}^* \setminus \mathcal{S}_\mathcal{T}} \mathbf{X}_t^\top (\boldsymbol{\beta}_t - \boldsymbol{\beta}^*) - \mathbb{E}\left\{ [\mathbf{X}_t]_{\mathcal{S}^* \setminus \mathcal{S}_\mathcal{T}} \mathbf{X}_t^\top (\boldsymbol{\beta}_t - \boldsymbol{\beta}^*) \,|\mathcal{F}_t \right\} \right\} \right\| \geq 2u \right) \leq \exp(Cs)$$
$$\times \exp\left(- \min\left\{ \frac{u^2}{(\lambda_{\max,s}^{\mathsf{Off}})^2 \sum_{t \in \mathcal{T}} W_t^2 + |\mathcal{T}| \lambda_{\max}^2 \|[\boldsymbol{\beta}^*]_{\mathcal{S}^* \setminus \mathcal{S}_\mathcal{T}}\|^2}, \frac{u}{\lambda_{\max,s}^{\mathsf{Off}} \max_{t \in \mathcal{T}} W_t + \lambda_{\max} \|[\boldsymbol{\beta}^*]_{\mathcal{S}^* \setminus \mathcal{S}_\mathcal{T}}\|} \right\} \right).$$

因此，结合这些不等式，我们有

$$\mathbb{P}\left(\left\| [\sum_{t \in \mathcal{T}} \mathbf{X}_t \mathbf{X}_t^\top (\boldsymbol{\beta}_t - \boldsymbol{\beta}^*)]_{\mathcal{S}^* \setminus \mathcal{S}_\mathcal{T}} \right\| \geq |\mathcal{T}| \cdot \lambda_{\min} \|[\boldsymbol{\beta}^*]_{\mathcal{S}^* \setminus \mathcal{S}_\mathcal{T}}\| - \lambda_{\max,s}^{\mathsf{Off}} \sum_{t \in \mathcal{T}} W_t - u \right) \geq 1 -$$
$$\exp\left(Cs - \min\left\{ \frac{u^2}{(\lambda_{\max,s}^{\mathsf{Off}})^2 \sum_{t \in \mathcal{T}} W_t^2 + |\mathcal{T}| \lambda_{\max}^2 \|[\boldsymbol{\beta}^*]_{\mathcal{S}^* \setminus \mathcal{S}_\mathcal{T}}\|^2}, \frac{u}{\lambda_{\max,s}^{\mathsf{Off}} \max_{t \in \mathcal{T}} W_t + \lambda_{\max} \|[\boldsymbol{\beta}^*]_{\mathcal{S}^* \setminus \mathcal{S}_\mathcal{T}}\|} \right\} \right).$$

结合公式 21，我们有

$$\mathbb{P}\Bigg(\left\| [\mathbf{G}(\{\boldsymbol{\beta}_t, \mathbf{X}_t, Y_t\}_{t \in \mathcal{T}})]_{\mathcal{S}^* \setminus \mathcal{S}_\mathcal{T}} \right\| \geq |\mathcal{T}| \lambda_{\min} \|[\boldsymbol{\beta}^*]_{\mathcal{S}^* \setminus \mathcal{S}_\mathcal{T}}\| - \lambda_{\max,s}^{\mathsf{Off}} \sum_{t \in \mathcal{T}} W_t$$
$$-u - C\sigma \sqrt{\lambda_{\max}} \sqrt{|\mathcal{T}|} \sqrt{|\mathcal{S}^* \setminus \mathcal{S}_\mathcal{T}| \log(2d/|\mathcal{S}^* \setminus \mathcal{S}_\mathcal{T}|) + \delta} \Bigg) \geq 1 - \exp(Cs)$$
$$\times \exp\left(- \min\left\{ \frac{u^2}{(\lambda_{\max,s}^{\mathsf{Off}})^2 \sum_{t \in \mathcal{T}} W_t^2 + |\mathcal{T}| \lambda_{\max}^2 \|[\boldsymbol{\beta}^*]_{\mathcal{S}^* \setminus \mathcal{S}_\mathcal{T}}\|^2}, \frac{u}{\lambda_{\max,s}^{\mathsf{Off}} \max_{t \in \mathcal{T}} W_t + \lambda_{\max} \|[\boldsymbol{\beta}^*]_{\mathcal{S}^* \setminus \mathcal{S}_\mathcal{T}}\|} \right\} \right)$$
$$- \exp\left(-c \min\left\{ \delta, \sqrt{\delta} \sqrt{|\mathcal{T}|/(|\mathcal{S}^* \setminus \mathcal{S}_\mathcal{T}| \log(2d/|\mathcal{S}^* \setminus \mathcal{S}_\mathcal{T}|))} \right\} \right),$$

从而完成证明。

B.3 命题1的证明

我们通过归纳法证明命题 1 。引理 1 给出了 $t \in [0, \tau_1 - 1]$ 的迭代序列动态。这里，我们只证明区间 $[\tau_1, \tau_2 - 1]$ 内的迭代量；对于一般的 $[\tau_l, \tau_l - 1]$，只需增加一些记号，便可用同样的方法证明。在 $t \in [\tau_1, \tau_2 - 1]$ 期间，更新为

$$\boldsymbol{\beta}_{t+1} = \mathcal{H}_{\mathcal{S}_1}(\boldsymbol{\beta}_t - \eta_t \cdot \mathbf{g}_t), \quad t \in [\tau_1, \tau_2 - 1].$$

在这一过程中，我们有

$$\|[\boldsymbol{\beta}_t - \boldsymbol{\beta}^*]_{\mathcal{S}^*\setminus\mathcal{S}_1}\| = \|[\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*]_{\mathcal{S}^*\setminus\mathcal{S}_1}\|.$$

对于 $\mathcal{S}_1 \setminus \mathcal{S}^*$中的项，采用与引理 1证明类似的分析，可得

$$\begin{aligned}\|[\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*]_{\mathcal{S}_1}\|^2 \le (1-\eta_t\lambda_{\min})\left\|[\boldsymbol{\beta}_t - \boldsymbol{\beta}^*]_{\mathcal{S}_1}\right\|^2 + 2\eta_t^2\lambda_{\max}s\sigma^2 \\ + \eta_t\|[\boldsymbol{\beta}^*]_{\mathcal{S}^*\setminus\mathcal{S}_1}\|^2\cdot\frac{(\lambda^{\mathsf{Off}}_{\max,s})^2}{\lambda_{\min}} - 2\eta_t\boldsymbol{\Xi}_1^{(t)} + 2\eta_t\boldsymbol{\Xi}_2^{(t)} + \eta_t^2\boldsymbol{\Xi}_3^{(t)},\end{aligned}$$

其中 $\boldsymbol{\Xi}_1^{(t)}$、$\boldsymbol{\Xi}_2^{(t)}$、$\boldsymbol{\Xi}_3^{(t)}$ 均沿用引理 1 证明中的结果，并相应地将 $\mathcal{S}_1$ 代入$\mathcal{S}_0$。将上述等式从 $t=\tau_1$ 累加，可得

$$\begin{aligned}\|[\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}^*]_{\mathcal{S}_1}\|^2 \le{}& \prod_{j=\tau_1}^{t}\left(1-\frac{C_a}{j+C_bs\log(2d/s)}\right)\left\|\left[\boldsymbol{\beta}_{\tau_1} - \boldsymbol{\beta}^*\right]_{\mathcal{S}_1}\right\|^2 \\ &+ 2s\sigma^2C_a^2\frac{\lambda_{\max}}{\lambda_{\min}^2}\sum_{l=\tau_1}^{t}\prod_{j=l}^{t}\left(1-\frac{C_a}{j+C_bs\log(2d/s)}\right)\left(\frac{1}{l+C_bs\log(2d/s)}\right)^2 \\ &+ C_a\sum_{l=\tau_1}^{t}\prod_{j=l}^{t}\left(1-\frac{C_a}{j+C_bs\log(2d/s)}\right)\frac{1}{l+C_bs\log(2d/s)}\|[\boldsymbol{\beta}^*]_{\mathcal{S}^*\setminus\mathcal{S}_1}\|^2\cdot\frac{(\lambda^{\mathsf{Off}}_{\max,s})^2}{\lambda_{\min}^2} \\ &- 2\frac{C_a}{\lambda_{\min}}\sum_{l=\tau_1}^{t}\prod_{j=l}^{t}\left(1-\frac{C_a}{j+C_bs\log(2d/s)}\right)\frac{1}{l+C_bs\log(2d/s)}\boldsymbol{\Xi}_1^{(l)} \\ &+ 2\frac{C_a}{\lambda_{\min}}\sum_{l=\tau_1}^{t}\prod_{j=l}^{t}\left(1-\frac{C_a}{j+C_bs\log(2d/s)}\right)\frac{1}{l+C_bs\log(2d/s)}\boldsymbol{\Xi}_2^{(l)} \\ &+ \frac{C_a^2}{\lambda_{\min}^2}\sum_{l=\tau_1}^{t}\prod_{j=l}^{t}\left(1-\frac{C_a}{j+C_bs\log(2d/s)}\right)\frac{1}{(l+C_bs\log(2d/s))^2}\boldsymbol{\Xi}_3^{(l)}.\end{aligned} \tag{22}$$

值得注意的是，在 τ_1 处，它有

$$\left\|\left[\boldsymbol{\beta}_{\tau_1} - \boldsymbol{\beta}^*\right]_{\mathcal{S}_1}\right\|^2 = \left\|\left[\boldsymbol{\beta}_{\tau_1} - \boldsymbol{\beta}^*\right]_{\mathcal{S}_0}\right\|^2 + \left\|[\boldsymbol{\beta}^*]_{\mathcal{S}_1\setminus\mathcal{S}_0}\right\|^2.$$

然后，我们将上述 $\left\|[\boldsymbol{\beta}_{\tau_1}-\boldsymbol{\beta}^*]_{\mathcal{S}_0}\right\|^2$ 方程代入方程22，并累积更新，直到 $t=0$，得到

$$
\begin{aligned}
\|[\boldsymbol{\beta}_{t+1}-\boldsymbol{\beta}^*]_{\mathcal{S}_1}\|^2 \leq & \underbrace{\prod_{j=0}^{t}\left(1-\frac{C_a}{j+C_b s\log(2d/s)}\right)\|[\boldsymbol{\beta}_0-\boldsymbol{\beta}^*]_{\mathcal{S}_0}\|^2}_{\boldsymbol{B}_1} \\
& + \underbrace{\prod_{j=\tau_1}^{t}\left(1-\frac{C_a}{j+C_b s\log(2d/s)}\right)\left\|[\boldsymbol{\beta}^*]_{\mathcal{S}_1\setminus\mathcal{S}_0}\right\|^2}_{\boldsymbol{B}_2} \\
& + \underbrace{2s\sigma^2 C_a^2\frac{\lambda_{\max}}{\lambda_{\min}^2}\sum_{l=0}^{t}\prod_{j=l}^{t}\left(1-\frac{C_a}{j+C_b s\log(2d/s)}\right)\left(\frac{1}{l+C_b s\log(2d/s)}\right)^2}_{\boldsymbol{B}_3} \\
& + \underbrace{C_a\sum_{l=0}^{t}\prod_{j=l}^{t}\left(1-\frac{C_a}{j+C_b s\log(2d/s)}\right)\frac{1}{l+C_b s\log(2d/s)}\|[\boldsymbol{\beta}^*]_{\mathcal{S}^*\setminus\mathcal{S}_l}\|^2\cdot\frac{(\lambda_{\max,s}^{\mathsf{Off}})^2}{\lambda_{\min}^2}}_{\boldsymbol{B}_4} \\
& - \underbrace{2\frac{C_a}{\lambda_{\min}}\sum_{l=0}^{t}\prod_{j=l}^{t}\left(1-\frac{C_a}{j+C_b s\log(2d/s)}\right)\frac{1}{l+C_b s\log(2d/s)}\boldsymbol{\Xi}_1^{(l)}}_{\boldsymbol{B}_5} \\
& + \underbrace{2\frac{C_a}{\lambda_{\min}}\sum_{l=0}^{t}\prod_{j=l}^{t}\left(1-\frac{C_a}{j+C_b s\log(2d/s)}\right)\frac{1}{l+C_b s\log(2d/s)}\boldsymbol{\Xi}_2^{(l)}}_{\boldsymbol{B}_6} \\
& + \underbrace{\frac{C_a^2}{\lambda_{\min}^2}\sum_{l=0}^{t}\prod_{j=l}^{t}\left(1-\frac{C_a}{j+C_b s\log(2d/s)}\right)\frac{1}{(l+C_b s\log(2d/s))^2}\boldsymbol{\Xi}_3^{(l)}}_{\boldsymbol{B}_7}.
\end{aligned}
$$

首先，我们通过引理 $\boldsymbol{B}_1$ 16对该项进行界定，

$$
\boldsymbol{B}_1 \leq \frac{C_b s\log(2d/s)}{t+1+C_b s\log(2d/s)}\|[\boldsymbol{\beta}_0-\boldsymbol{\beta}^*]_{\mathcal{S}_0}\|^2 \leq \frac{C_b s\log(2d/s)}{t+1+C_b s\log(2d/s)}\frac{\sigma^2}{\lambda_{\max}}.
$$

关于项 $\boldsymbol{B}_2$, 我们仍然应用引理 16 并且它有

$$
\boldsymbol{B}_2 \leq \left(\frac{\tau_1+C_b s\log(2d/s)}{t+C_b s\log(2d/s)}\right)^{C_a}\|[\boldsymbol{\beta}^*]_{\mathcal{S}_1\setminus\mathcal{S}_0}\|^2.
$$

至于项 $\mathbf{B}_3$,它有

$$\prod_{j=l}^{t}\left(1-\frac{C_a}{j+C_bs\log(2d/s)}\right)\left(\frac{1}{l+C_bs\log(2d/s)}\right)^2\sigma^2 \\ \leq\left(\frac{1}{t+1+C_bs\log(2d/s)}\right)^{C_a}(l+C_bs\log(2d/s))^{C_a-2}\sigma^2.$$

然后，根据引理 17，可将其界定为

$$\boldsymbol{B}_3\leq C_a\frac{\lambda_{\max}}{\lambda_{\min}^2}\frac{2s}{t+1+C_bs\log(2d/s)}\sigma^2.$$

类似地，对于 $\boldsymbol{B}_4$,，我们有如下界

$$\boldsymbol{B}_4\leq\left(\frac{\lambda_{\max,s}^{\mathsf{Off}}}{\lambda_{\min}}\right)^2\cdot\|[\boldsymbol{\beta}^*]_{\mathcal{S}^*\setminus\mathcal{S}_1}\|^2+\left(\frac{\tau_1+C_bs\log(2d/s)}{t+C_bs\log(2d/s)}\right)^{C_a}\left(\frac{\lambda_{\max,s}^{\mathsf{Off}}}{\lambda_{\min}}\right)^2\cdot\|[\boldsymbol{\beta}^*]_{\mathcal{S}_1\setminus\mathcal{S}_0}\|^2.$$

接下来考虑项 $\boldsymbol{B}_5$。具体而言，其条件 Orlicz 范数可界定为

$$\begin{aligned}\|\boldsymbol{\Xi}_1^{(l)}|\mathcal{F}_l\|_{\Psi_1}&\leq\lambda_{\max}\|[\boldsymbol{\beta}_l-\boldsymbol{\beta}^*]_{\mathcal{S}_l}\|^2+\lambda_{\max,s}^{\mathsf{Off}}\cdot\|[\boldsymbol{\beta}_l-\boldsymbol{\beta}^*]_{\mathcal{S}_l}\|\cdot\|[\boldsymbol{\beta}^*]_{\mathcal{S}^*\setminus\mathcal{S}_l}\|\\&\leq 2\lambda_{\max}\|[\boldsymbol{\beta}_l-\boldsymbol{\beta}^*]_{\mathcal{S}_l}\|^2+\frac{(\lambda_{\max,s}^{\mathsf{Off}})^2}{\lambda_{\max}}\|[\boldsymbol{\beta}^*]_{\mathcal{S}^*\setminus\mathcal{S}_l}\|^2.\end{aligned}$$

根据引理 9，在事件 $\cup_{l=0}^{t}\{\mathcal{E}_l\}$下，我们有

$$\mathbb{P}\left(\left|\sum_{l=0}^{t}\prod_{j=l}^{t}\left(1-\frac{C_a}{j+C_bs\log(2d/s)}\right)\frac{1}{l+C_bs\log(2d/s)}\boldsymbol{\Xi}_1^{(l)}\right|\geq c_1\frac{C^*}{C_a}\frac{(s\log(2d/s)+\delta_t)\sigma^2}{t+1+C_bs\log(2d/s)}\right. \\ \left.+\frac{c_1}{C_a}\frac{(\lambda_{\max,s}^{\mathsf{Off}})^2}{\lambda_{\min}}\|[\boldsymbol{\beta}^*]_{\mathcal{S}^*\setminus\mathcal{S}_1}\|^2+\lambda_{\min}\frac{c_1}{C_a}\left(\frac{\tau_1+C_bs\log(2d/s)}{t+1+C_bs\log(2d/s)}\right)^{C_a-2}\|[\boldsymbol{\beta}^*]_{\mathcal{S}_1\setminus\mathcal{S}_0}\|^2\right) \\ \leq\exp\left(-c_2\frac{C_bs\log(2d/s)+t}{C_a}\frac{\lambda_{\min}^2}{\lambda_{\max}^2}\right),$$

在该条件下，我们有

$$|\boldsymbol{B}_5|\leq c_1\frac{C^*}{\lambda_{\min}}\frac{(s\log(2d/s)+\delta_t)\sigma^2}{t+1+C_bs\log(2d/s)}+c_1\left(\frac{\lambda_{\max,s}^{\mathsf{Off}}}{\lambda_{\min}}\right)^2\|[\boldsymbol{\beta}^*]_{\mathcal{S}^*\setminus\mathcal{S}_1}\|^2 \\ +c_1\left(\frac{\tau_1+C_bs\log(2d/s)}{t+1+C_bs\log(2d/s)}\right)^{C_a-2}\cdot\|[\boldsymbol{\beta}^*]_{\mathcal{S}_1\setminus\mathcal{S}_0}\|^2.$$

至于术语 $\boldsymbol{B}_6$,，我们有

$$\|\boldsymbol{\Xi}_2^{(t)}|\mathcal{F}_t\|_{\Psi_1}\leq\sqrt{\lambda_{\max}}\sigma\|[\boldsymbol{\beta}_t-\boldsymbol{\beta}^*]_{\mathcal{S}_t}\|,$$

由此可得

$$\left\|\prod_{j=l}^{t}\left(1-\frac{C_a}{j+C_bs\log(2d/s)}\right)\frac{1}{l+C_bs\log(2d/s)}\boldsymbol{\Xi}_2^{(l)}|\mathcal{F}_l\right\|_{\Psi_1}$$
$$\leq C\sqrt{\lambda_{\max}}\sigma\frac{(l+C_bs\log(2d/s))^{C_a-1}}{(t+1+C_bs\log(2d/s))^{C_a}}\cdot\|[\boldsymbol{\beta}_l-\boldsymbol{\beta}^*]_{\mathcal{S}_l}\|.$$

那么在事件 $\cup_{l=0}^{t}\{\mathcal{E}_l\}$下，根据引理 9，我们有

$$\mathbb{P}\left(\left|\sum_{l=0}^{t}\prod_{j=l}^{t}\left(1-\frac{C_a}{j+C_bs\log(2d/s)}\right)\frac{1}{l+C_bs\log(2d/s)}\boldsymbol{\Xi}_2^{(l)}\right|\geq c_1\frac{C^*}{C_a}\frac{(s\log(2d/s)+\delta_{t+1})\sigma^2}{t+1+C_bs\log(2d/s)}\right.$$
$$\left.+\frac{c_1}{C_a}\frac{(\lambda^{\mathsf{Off}}_{\max,s})^2}{\lambda_{\min}}\|[\boldsymbol{\beta}^*]_{\mathcal{S}^*\setminus\mathcal{S}_1}\|^2+\lambda_{\min}\frac{c_1}{C_a}\left(\frac{\tau_1+C_bs\log(2d/s)}{t+1+C_bs\log(2d/s)}\right)^{C_a-2}\|[\boldsymbol{\beta}^*]_{\mathcal{S}_1\setminus\mathcal{S}_0}\|^2\right)$$
$$\leq\exp\left(-c_2\min\left\{(s\log(2d/s)+\delta_{t+1})\frac{C^*}{C_a}\frac{\lambda_{\min}}{\lambda_{\max}},\sqrt{s\log(2d/s)+\delta_{t+1}}\sqrt{C_bs\log(2d/s)+t}\frac{\sqrt{C^*}}{C_a}\frac{\sqrt{\lambda_{\min}}}{\sqrt{\lambda_{\max}}}\right\}\right),$$

这意味着以下界

$$|\boldsymbol{B}_6|\leq c_1\frac{C^*}{\lambda_{\min}}\frac{(s\log(2d/s)+\delta_{t+1})\sigma^2}{t+1+C_bs\log(2d/s)}$$
$$+c_1\frac{(\lambda^{\mathsf{Off}}_{\max,s})^2}{\lambda^2_{\min}}\|[\boldsymbol{\beta}^*]_{\mathcal{S}^*\setminus\mathcal{S}_1}\|^2+c_1\left(\frac{\tau_1+C_bs\log(2d/s)}{t+1+C_bs\log(2d/s)}\right)^{C_a-2}\cdot\|[\boldsymbol{\beta}^*]_{\mathcal{S}_1\setminus\mathcal{S}_0}\|^2.$$

回顾定义 $\mathcal{E}_t^X:=\{\max_{|\mathcal{S}|\leq s}\|[\mathbf{X}_t]_{\mathcal{S}}\|^2\leq C\lambda_{\max}(s\log(2d/s)+\delta_t)\}$（在第 B.1 节中定义）。因此，在 $\mathcal{E}_t^X$下，我们有

$$\|\boldsymbol{\Xi}_3^{(t)}|\mathcal{F}_t,\mathcal{E}_t^X\|_{\Psi_1}\leq C(s\log(2d/s)+\delta_t)\lambda_{\max}\sigma^2.$$

通过与第 B.1 节类似的分析，我们以 $1-\exp(-c_2(s\log(2d/s)+t/C_b))$ 的概率有，

$$|\boldsymbol{B}_7|\leq c_1C^*\frac{1}{\lambda_{\min}}\frac{s\log(2d/s)+\delta_{t+1}}{t+1+C_bs\log(2d/s)}\sigma^2.$$

因此，综上所述，我们完成了对 $t\in[\tau_1,\tau_2-1]$的收敛动态的证明，

$$\|[\boldsymbol{\beta}_{t+1}-\boldsymbol{\beta}^*]_{\mathcal{S}_1}\|^2\leq C^*\frac{s\log(2d/s)}{t+1+C_bs\log(2d/s)}\sigma^2+C\left(\frac{\lambda^{\mathsf{Off}}_{\max,s}}{\lambda_{\min}}\right)^2\|[\boldsymbol{\beta}^*]_{\mathcal{S}^*\setminus\mathcal{S}_1}\|^2$$
$$+C\left(\frac{\tau_1+C_bs\log(2d/s)}{t+1+C_bs\log(2d/s)}\right)^{C_a-2}\|[\boldsymbol{\beta}^*]_{\mathcal{S}_1\setminus\mathcal{S}_0}\|^2.$$

C 第4节中的证明

C.1 引理4的证明

证明。回忆 $\max_{i\in[K]}\|\boldsymbol{\beta}_i-\boldsymbol{\beta}_i^*\|\le h_0$。由三角不等式可得

$$\left|\frac{\mathbf{X}^\top\boldsymbol{\beta}_i^*}{\|\mathbf{X}\|}-\frac{\mathbf{X}^\top\boldsymbol{\beta}_i}{\|\mathbf{X}\|}\right|\le\|\boldsymbol{\beta}_i-\boldsymbol{\beta}_i^*\|\le h_0,\quad \left|\max_{j\neq i}\frac{\mathbf{X}^\top\boldsymbol{\beta}_j^*}{\|\mathbf{X}\|}-\max_{j\neq i}\frac{\mathbf{X}^\top\boldsymbol{\beta}_j}{\|\mathbf{X}\|}\right|\le\max_{j\neq i}\|\boldsymbol{\beta}_j-\boldsymbol{\beta}_j^*\|\le h_0.$$

因此，对于 $\mathbf{X}\in\mathcal{X}\left(i,\{\boldsymbol{\beta}_j\}\right)$，有

$$\frac{\mathbf{X}^\top\boldsymbol{\beta}_i^*}{\|\mathbf{X}\|}=\frac{\mathbf{X}^\top\boldsymbol{\beta}_i}{\|\mathbf{X}\|}+\left(\frac{\mathbf{X}^\top\boldsymbol{\beta}_i^*}{\|\mathbf{X}\|}-\frac{\mathbf{X}^\top\boldsymbol{\beta}_i}{\|\mathbf{X}\|}\right)\ge\max_{j\neq i}\frac{\mathbf{X}^\top\boldsymbol{\beta}_j}{\|\mathbf{X}\|}-h_0\ge\max_{j\neq i}\frac{\mathbf{X}^\top\boldsymbol{\beta}_j^*}{\|\mathbf{X}\|}-2h_0,$$

这意味着 $\mathbf{X}\in\mathcal{U}_i(-2h_0)$。因此，我们完成了对 $\mathcal{X}\left(i,\{\boldsymbol{\beta}_j\}\right)\subseteq\mathcal{U}_i(-2h_0)$ 的证明。另一方面，如果 $\mathbf{X}\in\mathcal{U}_i(2h_0)$，那么我们有

$$\frac{\mathbf{X}^\top\boldsymbol{\beta}_i}{\|\mathbf{X}\|}=\frac{\mathbf{X}^\top\boldsymbol{\beta}_i^*}{\|\mathbf{X}\|}+\left(\frac{\mathbf{X}^\top\boldsymbol{\beta}_i}{\|\mathbf{X}\|}-\frac{\mathbf{X}^\top\boldsymbol{\beta}_i^*}{\|\mathbf{X}\|}\right)\ge\max_{j\neq i}\frac{\mathbf{X}^\top\boldsymbol{\beta}_j^*}{\|\mathbf{X}\|}+2h_0-h_0\ge\max_{j\neq i}\frac{\mathbf{X}^\top\boldsymbol{\beta}_j}{\|\mathbf{X}\|},$$

这表明 $\mathbf{X}\in\mathcal{X}\left(i,\{\boldsymbol{\beta}_j\}\right)$。因此，我们完成了证明。 □

C.2 引理5的证明

引理 5 的证明。我们首先证明引理 (1) 中的结论 5。对于 $j\in\mathcal{A}^c$，我们有

$$\frac{\mathbf{X}_t^\top\boldsymbol{\beta}_j}{\|\mathbf{X}_t\|}=\frac{\mathbf{X}_t^\top\boldsymbol{\beta}_j^*}{\|\mathbf{X}_t\|}+\frac{\mathbf{X}_t^\top\left(\boldsymbol{\beta}_j-\boldsymbol{\beta}_j^*\right)}{\|\mathbf{X}_t\|}=\frac{\max_{i\in[K]}\mathbf{X}_t^\top\boldsymbol{\beta}_i^*}{\|\mathbf{X}_t\|}-\frac{\max_{i\in[K]}\mathbf{X}_t^\top\boldsymbol{\beta}_i^*-\mathbf{X}_t^\top\boldsymbol{\beta}_j^*}{\|\mathbf{X}_t\|}+\frac{\mathbf{X}_t^\top\left(\boldsymbol{\beta}_j-\boldsymbol{\beta}_j^*\right)}{\|\mathbf{X}_t\|}.$$

假设 5 保证了 $\max^{i\in[K]}\mathbf{X}\ {}_t^\top\boldsymbol{\beta}_i^*-\ {}_t^\top\boldsymbol{\beta}_j^*\mathbf{X}\mathbf{X}_t\|\ \ge h$，几乎处处成立。因此，结合 $\mathbf{X}$ ${}_t^\top\left(\boldsymbol{\beta}_j-\boldsymbol{\beta}_j^*\right)\le\|\mathbf{X}_t\|\cdot\|\boldsymbol{\beta}_i-\boldsymbol{\beta}_i^*\|$，并在事件 $\{\|\boldsymbol{\beta}_i-\boldsymbol{\beta}_i^*\|\le\frac{h}{2},\ i\in[K]\}$ 下，可得

$$\frac{\mathbf{X}_t^\top\boldsymbol{\beta}_j}{\|\mathbf{X}_t\|}\le\frac{\max_{i\in[K]}\mathbf{X}_t^\top\boldsymbol{\beta}_i^*}{\|\mathbf{X}_t\|}-\frac{h}{2}<\frac{\max_{i\in[K]}\mathbf{X}_t^\top\boldsymbol{\beta}_i^*}{\|\mathbf{X}_t\|},$$

这意味着 $\mathcal{X}(j,\ \{\boldsymbol{\beta}_i\})=\emptyset$ 对 $j\in\mathcal{A}^c$成立。

记 $h_{\mathcal{A}}:=\max_{i\in\mathcal{A}}\|\boldsymbol{\beta}_i-\boldsymbol{\beta}_i^*\|$。此外，在事件 $\{\|\boldsymbol{\beta}_i-\boldsymbol{\beta}_i^*\|\le\frac{h}{2},\ i\in[K]\}$ 下，对任意$i\in\mathcal{A}$，若 $\mathbf{X}\in\mathcal{X}\left(i,\ \{\boldsymbol{\beta}_j\}\right)$，则有

$$\frac{\mathbf{X}^\top\boldsymbol{\beta}_i^*}{\|\mathbf{X}\|}=\frac{\mathbf{X}^\top\boldsymbol{\beta}_i}{\|\mathbf{X}\|}+\left(\frac{\mathbf{X}^\top\boldsymbol{\beta}_i^*}{\|\mathbf{X}\|}-\frac{\mathbf{X}^\top\boldsymbol{\beta}_i}{\|\mathbf{X}\|}\right)\ge\max_{j\neq i,j\in\mathcal{A}}\frac{\mathbf{X}^\top\boldsymbol{\beta}_j}{\|\mathbf{X}\|}-h_{\mathcal{A}}\ge\max_{j\neq i,j\in\mathcal{A}}\frac{\mathbf{X}^\top\boldsymbol{\beta}_j^*}{\|\mathbf{X}\|}-2h_{\mathcal{A}},$$

并且根据假设 5，有

$$\max_{j\in\mathcal{A}}\frac{\mathbf{X}^\top\boldsymbol{\beta}_j^*}{\|\mathbf{X}\|}\ge\max_{j\in\mathcal{A}^c}\frac{\mathbf{X}^\top\boldsymbol{\beta}_j^*}{\|\mathbf{X}\|}+h.$$

上述两个方程共同推出 $\mathbf{X} \in \mathcal{U}_i(-2h_{\mathcal{A}})$。因此，我们完成了对 $\mathcal{X}\left(i, \left\{\boldsymbol{\beta}_j\right\}\right) \subseteq \mathcal{U}_i(-2h_{\mathcal{A}})$ 的证明。另一方面，对于任意 $\mathbf{X} \in \mathcal{U}_i(2h_{\mathcal{A}})$，有

$$\frac{\mathbf{X}^\top \boldsymbol{\beta}_i}{\|\mathbf{X}\|} = \frac{\mathbf{X}^\top \boldsymbol{\beta}_i^*}{\|\mathbf{X}\|} + \left(\frac{\mathbf{X}^\top \boldsymbol{\beta}_i}{\|\mathbf{X}\|} - \frac{\mathbf{X}^\top \boldsymbol{\beta}_i^*}{\|\mathbf{X}\|}\right) \geq \max_{j \neq i, j \in \mathcal{A}} \frac{\mathbf{X}^\top \boldsymbol{\beta}_j^*}{\|\mathbf{X}\|} + 2h_{\mathcal{A}} - h_{\mathcal{A}}.$$

然后，根据三角不等式 $\max_{j \neq i,\ j \in \mathcal{A}} \mathbf{x}\|^{\top \boldsymbol{\beta}_j^*}\|\mathbf{x}\| \geq \max_{j \neq i,\ j \in \mathcal{A}} \mathbf{x}\|^{\top \boldsymbol{\beta}_j}\|\mathbf{x}\| - h_{\mathcal{A}}$，我们有 $\mathbf{x}\|^{\top \boldsymbol{\beta}_i}\|\mathbf{x}\| \geq \max_{j \neq i,\ j \in \mathcal{A}} \mathbf{x}\|^{\top \boldsymbol{\beta}_j}\|\mathbf{x}$。此外，在事件 $\{\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^*\| \leq \frac{h}{2},\ i \in [K]\}$ 发生时，对于所有 $k \in \mathcal{A}^c$，我们有 $\mathbf{X} \notin \mathcal{X}\left(k, \left\{\boldsymbol{\beta}_j\right\}\right)$。因此，我们完成了对 $\mathbf{X} \in \mathcal{X}\left(i, \left\{\boldsymbol{\beta}_j\right\}\right)$ 的证明。

接下来，我们只需证明引理 (2) 5 中的断言。根据三角不等式，对于所有 $\mathbf{X} \in \mathbb{R}^d$，有

$$\begin{aligned}\mathbf{X}^\top \boldsymbol{\beta}_i &\geq \mathbf{X}^\top \boldsymbol{\beta}_i^* - \left|\mathbf{X}^\top \boldsymbol{\beta}_i - \mathbf{X}^\top \boldsymbol{\beta}_i^*\right| \\ &\geq \max_{j \neq i} \mathbf{X}^\top \boldsymbol{\beta}_j - 2\max_j \left|\mathbf{X}^\top \boldsymbol{\beta}_j - \mathbf{X}^\top \boldsymbol{\beta}_j^*\right| + \left(\mathbf{X}^\top \boldsymbol{\beta}_i^* - \max_{j \neq i} \mathbf{X}^\top \boldsymbol{\beta}_j^*\right).\end{aligned}$$

进一步可知

$$\begin{aligned}&\mathbb{I}\left\{\mathbf{X}^\top \boldsymbol{\beta}_i \geq \max_{j \neq i} \mathbf{X}^\top \boldsymbol{\beta}_j\right\} \\ &\qquad \geq \mathbb{I}\left\{2\max_j \left|\mathbf{X}^\top \boldsymbol{\beta}_j - \mathbf{X}^\top \boldsymbol{\beta}_j^*\right| \cdot \|\mathbf{X}\|^{-1} \leq \left(\mathbf{X}^\top \boldsymbol{\beta}_i^* - \max_{j \neq i} \mathbf{X}^\top \boldsymbol{\beta}_j^*\right) \cdot \|\mathbf{X}\|^{-1}\right\}.\end{aligned}$$

因此，利用上述等式，我们得到以下界

$$\begin{aligned}&\mathbb{E}\left\{\mathbb{I}\{\mathbf{X}_t^\top \boldsymbol{\beta}_i \geq \max_j \mathbf{X}_t^\top \boldsymbol{\beta}_j\} \cdot \mathbf{X}_t \mathbf{X}_t^\top \middle| \mathcal{F}_t\right\} \\ \succeq\ &\mathbb{E}\left\{\mathbb{I}\left\{2\max_j \left|\mathbf{X}_t^\top \boldsymbol{\beta}_j - \mathbf{X}_t^\top \boldsymbol{\beta}_j^*\right| \cdot \|\mathbf{X}_t\|^{-1} \leq \left(\mathbf{X}_t^\top \boldsymbol{\beta}_i^* - \max_{j \neq i} \mathbf{X}_t^\top \boldsymbol{\beta}_j^*\right) \cdot \|\mathbf{X}_t\|^{-1}\right\} \cdot \mathbf{X}_t \mathbf{X}_t^\top \middle| \mathcal{F}_t\right\}.\end{aligned}$$

回顾一下，在假设 5 下，$\mathbf{X}^\top \boldsymbol{\beta}_i^* - \max_{j \neq i} \mathbf{X}^\top \boldsymbol{\beta}_j^* \geq h\|\mathbf{X}\|$ 对所有 $\mathbf{X} \in \mathcal{U}_i(h)$均成立。因此，我们将右侧限制在区域 $\mathcal{U}_i(h)$，于是得到

$$\begin{aligned}&\mathbb{E}\left\{\mathbb{I}\{\mathbf{X}_t^\top \boldsymbol{\beta}_i \geq \max_j \mathbf{X}_t^\top \boldsymbol{\beta}_j\} \cdot \mathbf{X}_t \mathbf{X}_t^\top \middle| \mathcal{F}_t\right\} \\ &\qquad \succeq \mathbb{E}\left\{\mathbb{I}\left\{2\max_j \left|\mathbf{X}_t^\top \boldsymbol{\beta}_j - \mathbf{X}_t^\top \boldsymbol{\beta}_j^*\right| \cdot \|\mathbf{X}_t\|^{-1} \leq h\right\} \cdot \mathbf{X}_t \mathbf{X}_t^\top \cdot \mathbb{I}\left\{\mathbf{X}_t \in \mathcal{U}_i(h)\right\} \middle| \mathcal{F}_t\right\}.\end{aligned}$$

另一方面，显然 $\max_j |\mathbf{X}_t^\top \boldsymbol{\beta}_j - \mathbf{X}_t^\top \boldsymbol{\beta}_j^*| \cdot \|\mathbf{X}_t\|^{-1} \leq \max_j \|\boldsymbol{\beta}_j - \boldsymbol{\beta}_j^*\|$。于是我们有

$$\mathbb{E}\left\{\mathbb{I}\{\mathbf{X}_t^\top \boldsymbol{\beta}_i \geq \max_j \mathbf{X}_t^\top \boldsymbol{\beta}_j\} \cdot \mathbf{X}_t \mathbf{X}_t^\top \middle| \mathcal{F}_t\right\} \succeq \mathbb{I}\left\{\max_j \|\boldsymbol{\beta}_j - \boldsymbol{\beta}_j^*\| \leq \frac{h}{2}\right\} \cdot \mathbb{E}\left\{\mathbf{X}_t \mathbf{X}_t^\top \cdot \mathbb{I}\left\{\mathbf{X}_t \in \mathcal{U}_i\right\} \middle| \mathcal{F}_t\right\},$$

这证明了引理 5 的第一条结论。下面讨论 $\mathcal{A}^c$。至此，证明完成。

□

定理4的 C.3 证明

回顾 $\mathcal{F}_t = \sigma(Y_{t-1},\ \mathbf{X}_{t-1},\ \alpha_{t-1},\ \ldots\ldots)$。下面的命题为估计误差的条件期望提供了依据。

命题 3。 在与定理 4相同的条件下，对于 $i \in \mathcal{A}$，有

$$\begin{aligned}\mathbb{E}\left\{\|\boldsymbol{\beta}_i^{(t+1)} - \boldsymbol{\beta}_i^*\|^2 \middle| \mathcal{F}_t\right\} \leq \|\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i^*\|^2 - 2\eta_t \frac{\pi_t}{K}\lambda_{\min}\|\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i^*\|^2 + 2\eta_t^2 d\lambda_{\max}^2\|\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i^*\|^2 \\ - 2\eta_t(1-\pi_t)\lambda_{\min}\|\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i^*\|^2 \cdot \mathbb{I}\left\{\max_j \|\boldsymbol{\beta}_j^{(t)} - \boldsymbol{\beta}_i^*\| \leq \frac{h}{2}\right\} + 2\eta_t^2 d\lambda_{\max}\sigma^2,\end{aligned}$$

对于 $i \in \mathcal{A}^c$，我们有

$$\begin{aligned}\mathbb{E}\left\{\|\boldsymbol{\beta}_i^{(t+1)} - \boldsymbol{\beta}_i^*\|^2 \middle| \mathcal{F}_t\right\} \leq \|\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i^*\|^2 - 2\eta_t \frac{\pi_t}{K}\lambda_{\min}\|\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i^*\|^2 \\ + 2\eta_t^2 d\lambda_{\max}\sigma^2 \cdot \left(\frac{\pi_t}{K} + (1-\pi_t)\cdot \mathbb{I}\left\{\max_{i\in[K]} \|\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i^*\| \geq \frac{h}{2}\right\}\right) \\ + 2\eta_t^2 d\lambda_{\max}^2\|\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i^*\|^2 \cdot \left(\frac{\pi_t}{K} + (1-\pi_t)\cdot \mathbb{I}\left\{\max_{i\in[K]} \|\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i^*\| \geq \frac{h}{2}\right\}\right),\end{aligned}$$

其中 $C > 0$ 是某个常数。

命题 3的证明。 事实上，更新可表征如下：

$$\boldsymbol{\beta}_i^{(t+1)} - \boldsymbol{\beta}_i^* = \boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i^* - \eta_t\ \mathbb{I}\{a_t = i\} \cdot (\mathbf{X}_t^\top \boldsymbol{\beta}_i^{(t)} - Y_t)\ \mathbf{X}_t.$$

那么， L_2 范数下的估计误差率为：

$$\begin{aligned}\|\boldsymbol{\beta}_i^{(t+1)} - \boldsymbol{\beta}_i^*\|^2 = \|\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i^*\|^2 - 2\eta_t\ \mathbb{I}\{a_t = i\} \cdot (\mathbf{X}_t^\top \boldsymbol{\beta}_i^{(t)} - \mathbf{X}_t^\top \boldsymbol{\beta}_i^*)^2 \\ - 2\eta_t\ \mathbb{I}\{a_t = i\}\ \xi_t\ (\mathbf{X}_t^\top \boldsymbol{\beta}_i^{(t)} - \mathbf{X}_t^\top \boldsymbol{\beta}_i^*) + \eta_t^2\ \mathbb{I}\{a_t = i\} \cdot (\mathbf{X}_t^\top \boldsymbol{\beta}_i^{(t)} - Y_t)^2\ \|\mathbf{X}_t\|^2.\end{aligned}$$

回顾 $\mathcal{F}_t^+ := \sigma\left(\mathbf{X}_t, Y_{t-1}, \mathbf{X}_{t-1}, \alpha_{t-1}, \ldots\right)$，我们首先考虑上述等式的条件期望。首先，示性函数具有

$$\begin{aligned}\mathbb{E}\left\{\mathbb{I}\{a_t = i\}\middle|\mathcal{F}_t^+\right\} &= \mathbb{E}\left\{\mathbb{I}\{a_t = i\}\cdot\mathbb{I}\left\{\alpha_t = 1\right\}\middle|\mathcal{F}_t^+\right\} + \mathbb{E}\left\{\mathbb{I}\{a_t = i\}\cdot\mathbb{I}\left\{\alpha_t = 0\right\}\middle|\mathcal{F}_t^+\right\} \\ &= \frac{\pi_t}{K} + (1-\pi_t)\cdot\mathbb{I}\left\{i \in \arg\max_j \mathbf{X}_t^\top \boldsymbol{\beta}_j^{(t)}\right\}.\end{aligned}$$

其次，$(\mathbf{X}_t^\top \boldsymbol{\beta}_i^{(t)} - Y_t)^2$ 的条件期望为

$$\mathbb{E}\left\{\left(\mathbf{X}_t^\top \boldsymbol{\beta}_i^{(t)} - Y_t\right)^2 \middle| \mathcal{F}_t^+\right\} \leq \left(\mathbf{X}_t^\top \boldsymbol{\beta}_i^{(t)} - \mathbf{X}_t^\top \boldsymbol{\beta}_i^*\right)^2 + \sigma^2.$$

因此，综上所述，我们有

$$\begin{aligned}
&\mathbb{E}\left\{\left\|\boldsymbol{\beta}_i^{(t+1)}-\boldsymbol{\beta}_i^*\right\|^2\Big|\mathcal{F}_t^+\right\}\\
&\quad=\left\|\boldsymbol{\beta}_i^{(t)}-\boldsymbol{\beta}_i^*\right\|^2-2\eta_t\ \left(\frac{\pi_t}{K}+(1-\pi_t)\cdot\mathbb{I}\left\{i\in\arg\max_j \mathbf{X}_t^\top\boldsymbol{\beta}_j^{(t)}\right\}\right)\cdot\left(\mathbf{X}_t^\top\boldsymbol{\beta}_i^{(t)}-\mathbf{X}_t^\top\boldsymbol{\beta}_i^*\right)^2\\
&\qquad+\eta_t^2\ \left(\frac{\pi_t}{K}+(1-\pi_t)\cdot\mathbb{I}\left\{i=\arg\max_j \mathbf{X}_t^\top\boldsymbol{\beta}_j^{(t)}\right\}\right)\cdot\left((\mathbf{X}_t^\top\boldsymbol{\beta}_i^{(t)}-\mathbf{X}_t^\top\boldsymbol{\beta}_i^*)^2+\sigma^2\right)\ \|\mathbf{X}_t\|^2.
\end{aligned}$$

然后，给定 $\mathcal{F}_t$ 时的期望为，

$$\begin{aligned}
\mathbb{E}\left\{\left\|\boldsymbol{\beta}_i^{(t+1)}-\boldsymbol{\beta}_i^*\right\|^2\Big|\mathcal{F}_t\right\}&\le\|\boldsymbol{\beta}_i^{(t)}-\boldsymbol{\beta}_i^*\|^2-2\eta_t\frac{\pi_t}{K}\lambda_{\min}\|\boldsymbol{\beta}_i^{(t)}-\boldsymbol{\beta}_i^*\|^2\\
&\quad-2\eta_t(1-\pi_t)\cdot\underbrace{\mathbb{E}\left\{\mathbb{I}\{i=\arg\max_j \mathbf{X}_t^\top\boldsymbol{\beta}_j^{(t)}\}\cdot(\mathbf{X}_t^\top\boldsymbol{\beta}_i^{(t)}-\mathbf{X}_t^\top\boldsymbol{\beta}_i^*)^2\big|\mathcal{F}_t\right\}}_{\mathbf{A}_1}\\
&\quad+\eta_t^2\cdot\frac{\pi_t}{K}\ \sigma^2\cdot\mathbb{E}\|\mathbf{X}_t\|^2+\eta_t^2\cdot\frac{\pi_t}{K}\cdot\underbrace{\mathbb{E}\left\{(\mathbf{X}_t^\top\boldsymbol{\beta}_i^{(t)}-\mathbf{X}_t^\top\boldsymbol{\beta}_i^*)^2\cdot\|\mathbf{X}_t\|^2\,\big|\mathcal{F}_t\right\}}_{\mathbf{A}_2}\\
&\quad+\eta_t^2\cdot(1-\pi_t)\cdot\underbrace{\mathbb{E}\left\{\mathbb{I}\{i=\arg\max_j \mathbf{X}_t^\top\boldsymbol{\beta}_j^{(t)}\}\cdot(\mathbf{X}_t^\top\boldsymbol{\beta}_i^{(t)}-\mathbf{X}_t^\top\boldsymbol{\beta}_i^*)^2\cdot\|\mathbf{X}_t\|^2\big|\mathcal{F}_t\right\}}_{\mathbf{A}_3}\\
&\quad+\eta_t^2\cdot(1-\pi_t)\cdot\sigma^2\cdot\underbrace{\mathbb{E}\left\{\mathbb{I}\{i=\arg\max_j \mathbf{X}_t^\top\boldsymbol{\beta}_j^{(t)}\}\cdot\|\mathbf{X}_t\|^2\big|\mathcal{F}_t\right\}}_{\mathbf{A}_4}.
\end{aligned}$$

我们将讨论　　根据 $i\in\mathcal{A}$ 或 $i\in\mathcal{A}$ 是否成立，分别讨论$\mathbf{A}_1$,$\mathbf{A}_2,\mathbf{A}_3,\mathbf{A}_4$的界　　^{c}.

1 对于 $i\in\mathcal{A}$。

界$\mathbf{A}_1$ 引理5 给出

$$\mathbf{A}_1\ge\lambda_{\min}\|\boldsymbol{\beta}_i^{(t)}-\boldsymbol{\beta}_i\|^2\cdot\mathbb{I}\left\{\max_i\|\boldsymbol{\beta}_i^{(t)}-\boldsymbol{\beta}_i^*\|\le\frac{h}{2}\right\}.$$

界$\mathbf{A}_2$ 为了界定 $\mathbf{A}_2$,，我们知道 $\boldsymbol{\beta}_t\in\mathcal{F}_t$。那么在假设 **??** 下，我们有

$$\mathbf{A}_2\le\|\boldsymbol{\beta}_i^{(t)}-\boldsymbol{\beta}_i^*\|^2\cdot\left\|\mathbb{E}\left\{\mathbf{X}_t\mathbf{X}_t^\top\cdot\|\mathbf{X}_t\|^2\right\}\big|\mathcal{F}_t\right\|\le 2d\lambda_{\max}^2\|\boldsymbol{\beta}_i^{(t)}-\boldsymbol{\beta}_i^*\|^2,$$

其中 C 是某个常数。

界 $\mathbf{A}_3$ 类似于$\mathbf{A}_2$, 的界，我们有

$$\mathbf{A}_3\le 2d\lambda_{\max}^2\|\boldsymbol{\beta}_i^{(t)}-\boldsymbol{\beta}^*\|^2.$$

界 $\mathbf{A}_4$ 同样地，我们有

$$\mathbf{A}_4 \leq 2d\lambda_{\max}.$$

因此，总而言之，条件期望估计误差满足，

$$\begin{aligned}\mathbb{E}\left\{\|\boldsymbol{\beta}_i^{(t+1)} - \boldsymbol{\beta}_i^*\|^2 \middle| \mathcal{F}_t\right\} \leq \|\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i^*\|^2 - 2\eta_t \frac{\pi_t}{K}\lambda_{\min}\|\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i^*\|^2 + 2\eta_t^2 d\lambda_{\max}\sigma^2 \\ - 2\eta_t(1-\pi_t)\lambda_{\min}\|\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i^*\|^2 \cdot \mathbb{I}\left\{\max_i \|\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i^*\| \leq \frac{h}{2}\right\} + 2\eta_t^2 d\lambda_{\max}^2\|\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i^*\|^2.\end{aligned}$$

2 对于 $i \in \mathcal{A}^c$。

A 的界$_1$ 我们使用 $\mathbf{A}_1 \geq 0$。

A 的界 $\mathbf{A}_2$ 其界与 $i \in \mathcal{A}$ 时的情形相同。

A 的界 $\mathbf{A}_3$ 引理 5 表明，当 $\max_{i\in[K]} \|\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i^*\| \leq h/2$,

$$\arg\max_{i\in[K]} \mathbf{X}^\top \boldsymbol{\beta}_i^{(t)} \subseteq \mathcal{A}.$$

因此，我们有 $\mathbf{A}_3 \leq Cd\lambda_{\max}^2\|\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}^*\|^2 \cdot \mathbb{I}\left\{\max_{i\in[K]} \|\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i^*\| \geq h/2\right\}$。

的界 $\mathbf{A}_4$ 同理，我们有

$$\mathbf{A}_4 \leq 2d\lambda_{\max} \cdot \mathbb{I}\left\{\max_{i\in[K]} \|\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i^*\| \geq h/2\right\}.$$

因此，总体而言，估计误差率为

$$\begin{aligned}\mathbb{E}\left\{\|\boldsymbol{\beta}_i^{(t+1)} - \boldsymbol{\beta}_i^*\|^2 \middle| \mathcal{F}_t\right\} \leq \|\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i^*\|^2 - 2\eta_t \frac{\pi_t}{K}\lambda_{\min}\|\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i^*\|^2 \\ + \eta_t^2 d\lambda_{\max}\sigma^2 \cdot \left(\frac{\pi_t}{K} + (1-\pi_t) \cdot \mathbb{I}\left\{\max_{i\in[K]} \|\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i^*\| \geq \frac{h}{2}\right\}\right) \\ + 2\eta_t^2 d\lambda_{\max}^2\|\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i^*\|^2 \cdot \left(\frac{\pi_t}{K} + (1-\pi_t) \cdot \mathbb{I}\left\{\max_{i\in[K]} \|\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i^*\| \geq \frac{h}{2}\right\}\right),\end{aligned}$$

至此，证明完成。

□

接下来，我们准备证明定理 4。

定理 4的证明。 在讨论收敛之前，我们先引入

$$\begin{aligned}
&\|\boldsymbol{\beta}_i^{(t+1)} - \boldsymbol{\beta}_i^*\|^2 - \mathbb{E}\left\{\|\boldsymbol{\beta}_i^{(t+1)} - \boldsymbol{\beta}_i^*\|^2 \middle| \mathcal{F}_t\right\} \\
&\quad = -2\eta_t \underbrace{\left\{\mathbb{I}\{a_t = i\} \cdot (\mathbf{X}_t^\top \boldsymbol{\beta}_i^{(t)} - \mathbf{X}_t^\top \boldsymbol{\beta}_i^*)^2 - \mathbb{E}\left\{\mathbb{I}\{a_t = i\} \cdot (\mathbf{X}_t^\top \boldsymbol{\beta}_i^{(t)} - \mathbf{X}_t^\top \boldsymbol{\beta}_i^*)^2 \middle| \mathcal{F}_t\right\}\right\}}_{=:\boldsymbol{\Xi}_1^{(t)}} \\
&\quad - 2\eta_t \underbrace{\left\{\mathbb{I}\{a_t = i\}\ \xi_t\ (\mathbf{X}_t^\top \boldsymbol{\beta}_i^{(t)} - \mathbf{X}_t^\top \boldsymbol{\beta}_i^*) - \mathbb{E}\left\{\mathbb{I}\{a_t = i\}\ \xi_t\ (\mathbf{X}_t^\top \boldsymbol{\beta}_i^{(t)} - \mathbf{X}_t^\top \boldsymbol{\beta}_i^*) \middle| \mathcal{F}_t\right\}\right\}}_{=:\boldsymbol{\Xi}_2^{(t)}} \\
&\quad + \eta_t^2 \underbrace{\left\{\mathbb{I}\{a_t = i\} \cdot (\mathbf{X}_t^\top \boldsymbol{\beta}_i^{(t)} - Y_t)^2\ \|\mathbf{X}_t\|^2 - \mathbb{E}\left\{\mathbb{I}\{a_t = i\} \cdot (\mathbf{X}_t^\top \boldsymbol{\beta}_i^{(t)} - Y_t)^2\ \|\mathbf{X}_t\|^2\right\}\right\}}_{=:\boldsymbol{\Xi}_3^{(t)}}.
\end{aligned}$$

在给定 $\mathcal{F}_t$ 的条件下，它满足

$$\|\boldsymbol{\Xi}_1^{(t)} | \mathcal{F}_t\|_{\Psi_1} \le \lambda_{\max} \|\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i^*\|^2, \quad \|\boldsymbol{\Xi}_2^{(t)} | \mathcal{F}_t\|_{\Psi_1} \le \sigma \sqrt{\lambda_{\max}} \|\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i^*\|,$$

而在 $\mathcal{E}_t^X$ 下: $= \{\|\mathbf{X}_t\|^2 \le C\lambda_{\max}(d + \delta_t)\}$,

$$\|\boldsymbol{\Xi}_3^{(t)} | \mathcal{F}_t,\ \mathcal{E}_t\|_{\Psi_1} \le C(d + \delta_t)\lambda_{\max}\sigma^2 + (d + \delta_t)\lambda_{\max}^2 \|\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i^*\|^2.$$

具体而言， $\mathbb{P}(\mathcal{E}_t^X | \mathcal{F}_t) \ge 1 - \exp(-c_2(d + \delta_t))$。

第一阶段 定义事件

$$\mathcal{E}_t := \left\{\|\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i^*\|^2 \le \frac{C^*\ (d + \delta_t)}{t + C_b d} \frac{\sigma^2}{\lambda_{\min}}, \quad i \in [K]\right\},$$

其中 C^* 是一个与 t、d、n 不相关的常数，稍后将具体给出。我们将通过归纳法证明。显然， $\mathcal{E}_0$ 成立。接下来，我们将证明在 $\cup_{s=0}^t \mathcal{E}_s$ 下事件 $\mathcal{E}_{t+1}$ 成立。将探索率 $\pi_t \ge \pi$ 代入命题 3 ，即可得到

$$\mathbb{E}\left\{\|\boldsymbol{\beta}_i^{(t+1)} - \boldsymbol{\beta}_i^*\|^2 \middle| \mathcal{F}_t\right\} \le \|\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i^*\|^2 - 2\eta_t \frac{\pi}{K} \lambda_{\min} \|\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i^*\|^2 + C\eta_t^2 d\lambda_{\max}\sigma^2 + C\eta_t^2 d\lambda_{\max}^2 \|\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i^*\|^2.$$

然后，估计误差动态满足

$$\begin{aligned}
\|\boldsymbol{\beta}_i^{(t+1)} - \boldsymbol{\beta}_i^*\|^2 &= \mathbb{E}\left\{\|\boldsymbol{\beta}_i^{(t+1)} - \boldsymbol{\beta}_i^*\|^2 \middle| \mathcal{F}_t\right\} + \|\boldsymbol{\beta}_i^{(t+1)} - \boldsymbol{\beta}_i^*\|^2 - \mathbb{E}\left\{\|\boldsymbol{\beta}_i^{(t+1)} - \boldsymbol{\beta}_i^*\|^2 \middle| \mathcal{F}_t\right\} \\
&\le \|\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i^*\|^2 - 2\eta_t \frac{\pi}{K} \lambda_{\min} \|\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i^*\|^2 + 2\eta_t^2 d\lambda_{\max}\sigma^2 \\
&\qquad + 2\eta_t^2 d\lambda_{\max}^2 \|\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i\|^2 + 2\eta_t \boldsymbol{\Xi}_1^{(t)} + 2\eta_t \boldsymbol{\Xi}_2^{(t)} + \eta_t^2 \boldsymbol{\Xi}_3^{(t)}.
\end{aligned}$$

将步长 $\eta_t = \frac{1}{\lambda_{\min}} \frac{C_a}{t + C_b d}$ 代入上述方程，

$$\begin{aligned}
\|\boldsymbol{\beta}_i^{(t+1)} - \boldsymbol{\beta}_i^*\|^2 \le \left(1 - \frac{\pi}{K} \frac{C_a}{t + C_b d}\right) \|\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i^*\|^2 + \frac{2\,C_a^2}{(t + C_b d)^2} \frac{\lambda_{\max}}{\lambda_{\min}^2}\, d\,\sigma^2 \\
- \frac{2\,C_a}{\lambda_{\min}} \frac{1}{t + C_b d}\, \boldsymbol{\Xi}_1^{(t)} - \frac{2\,C_a}{\lambda_{\min}} \frac{1}{t + C_b d}\, \boldsymbol{\Xi}_2^{(t)} + \frac{C_a^2}{\lambda_{\min}^2} \frac{1}{(t + C_b d)^2}\, \boldsymbol{\Xi}_3^{(t)},
\end{aligned}$$

其中使用了 $2\eta_t d \le \lambda_{\min}/\lambda_{\max}^2$。我们将上界累积到 $t=0$,

$$\|\boldsymbol{\beta}_i^{(t+1)} - \boldsymbol{\beta}_i^*\|^2 \le \underbrace{\prod_{s=0}^{t}\left(1-\frac{\pi}{K}\frac{C_a}{s+C_b d}\right)\|\boldsymbol{\beta}_i^{(0)}-\boldsymbol{\beta}_i^*\|^2}_{\boldsymbol{\Delta}_1} + \underbrace{2\frac{\lambda_{\max}}{\lambda_{\min}^2}\sigma^2\sum_{s=0}^{t}\prod_{l=s+1}^{t}\left(1-\frac{\pi}{K}\frac{C_a}{l+C_b d}\right)\frac{C_a^2 d}{(s+C_b d)^2}}_{\boldsymbol{\Delta}_2}$$
$$\underbrace{-\frac{2\,C_a}{\lambda_{\min}}\sum_{s=0}^{t}\prod_{l=s+1}^{t}\left(1-\frac{\pi}{K}\frac{C_a}{l+C_b d}\right)\frac{1}{s+C_b d}\boldsymbol{\Xi}_1^{(s)}}_{\boldsymbol{\Delta}_3}$$
$$\underbrace{-\frac{2\,C_a}{\lambda_{\min}}\sum_{s=0}^{t}\prod_{l=s+1}^{t}\left(1-\frac{\pi}{K}\frac{C_a}{l+C_b d}\right)\frac{1}{s+C_b d}\boldsymbol{\Xi}_2^{(s)}}_{\boldsymbol{\Delta}_4}$$
$$\underbrace{-\frac{C_a^2}{\lambda_{\min}^2}\sum_{s=0}^{t}\prod_{l=s+1}^{t}\left(1-\frac{\pi}{K}\frac{C_a}{l+C_b d}\right)\left(\frac{1}{s+C_b d}\right)^2\boldsymbol{\Xi}_3^{(s)}}_{\boldsymbol{\Delta}_5}.$$

回顾一下，我们要求 $C_a \ge 2K/\pi$。然后，根据引理 16，我们有

$$\prod_{s=0}^{t}\left(1-\frac{\pi}{K}\frac{C_a}{s+C_b d}\right) \le \left(\frac{C_b\,d}{t+1+C_b d}\right)^{\frac{C_a\pi}{K}} \le \frac{C_b\,d}{t+1+C_b d},$$

and

$$\left(1-\frac{\pi}{K}\frac{C_a}{s+1+C_b d}\right)\cdots\left(1-\frac{\pi}{K}\frac{C_a}{t+C_b d}\right)\left(\frac{1}{s+C_b d}\right)^2 \le \frac{(s+C_b d)^{\pi\cdot C_a/K-2}}{(t+1+C_b d)^{\pi\cdot C_a/K}}.$$

然后，对上述等式关于 s 求和，并按照引理 17 进行计算，我们得到:

$$\sum_{s=0}^{t}\left(1-\frac{\pi}{K}\frac{C_a}{s+1+C_b d}\right)\cdots\left(1-\frac{\pi}{K}\frac{C_a}{t+C_b d}\right)\left(\frac{1}{s+C_b d}\right)^2 \le \frac{2K}{\pi\cdot C_a}\frac{1}{t+1+C_b d}.$$

因此，结合初始化，我们得到 $\boldsymbol{\Delta}_1$ 和 $\boldsymbol{\Delta}_2$的界,

$$\boldsymbol{\Delta}_1 \le \frac{C_b\,d}{t+1+C_b d}\frac{\sigma^2}{\lambda_{\max}}, \quad \boldsymbol{\Delta}_2 \le 4C_a\frac{K}{\pi}\frac{\lambda_{\max}}{\lambda_{\min}^2}\frac{d\,\sigma^2}{t+1+C_b d}.$$

然后考虑 $\boldsymbol{\Delta}_3$。条件 Orlicz 范数可由以下内容界定

$$\left\|\prod_{l=s+1}^{t}\left(1-\frac{\pi}{K}\frac{C_a}{l+C_b d}\right)\frac{1}{s+C_b d}\boldsymbol{\Xi}_1^{(s)}\middle|\mathcal{F}_s\right\|_{\Psi_1} \le C\lambda_{\max}\frac{(s+C_b d)^{C_a\cdot\pi/K-1}}{(t+1+C_b d)^{C_a\cdot\pi/K}}\cdot\|\boldsymbol{\beta}_i^{(s)}-\boldsymbol{\beta}_i^*\|^2,$$

并且在事件 $\cup_{s=0}^{t}\mathcal{E}_s$下，引理 9 推出

$$\mathbb{P}\left(\left|\sum_{s=0}^{t}\prod_{l=s+1}^{t}\left(1-\frac{\pi}{K}\frac{C_a}{l+C_b d}\right)\frac{1}{s+C_b d}\boldsymbol{\Xi}_1^{(s)}\right| \ge c_1\frac{C^*}{C_a}\frac{(d+\delta_{t+1})\,\sigma^2}{t+1+C_b d}\right)$$
$$\le \exp\left(-c_2\frac{\pi}{K}\frac{C_b d+t}{C_a}\frac{\lambda_{\min}^2}{\lambda_{\max}^2}\right),$$

其中 $c_1 < 0.1$ 是某个常数，并且在该条件下我们有

$$\left|\boldsymbol{\Delta}_3\right| \leq c_1 C^* \frac{1}{\lambda_{\min}} \frac{(d+\delta_{t+1})\,\sigma^2}{t+1+C_b d}.$$

对于项 $\boldsymbol{\Delta}_4$，我们有

$$\left\| \prod_{l=s+1}^{t} \left(1 - \frac{\pi}{K}\frac{C_a}{l+C_b d}\right) \frac{1}{s+C_b d} \boldsymbol{\Xi}_2^{(s)} \middle| \mathcal{F}_s \right\|_{\Psi_1} \leq C\sqrt{\lambda_{\max}}\sigma \frac{(s+C_b d)^{C_a\cdot\pi/K-1}}{(t+1+C_b d)^{C_a\cdot\pi/K}} \cdot \|\boldsymbol{\beta}_i^{(s)} - \boldsymbol{\beta}_i^*\|,$$

然后在事件 $\cup_{s=0}^{t}\mathcal{E}_s$下，引理9 保证

$$\begin{aligned}
&\mathbb{P}\left(\left|\sum_{s=0}^{t}\prod_{l=s+1}^{t}\left(1-\frac{\pi}{K}\frac{C_a}{l+C_b d}\right)\frac{1}{s+C_b d}\boldsymbol{\Xi}_2^{(s)}\right| \geq c_1\frac{C^*}{C_a}\frac{(d+\delta_{t+1})\,\sigma^2}{t+1+C_b d}\right) \\
&\qquad \leq \exp\left(-c_2\min\left\{(d+\delta_{t+1})\frac{C^*}{C_a}\frac{\pi}{K}\frac{\lambda_{\min}}{\lambda_{\max}}, \sqrt{d+\delta_{t+1}}\sqrt{C_b d + t}\frac{\sqrt{C^*}}{C_a}\frac{\sqrt{\lambda_{\min}}}{\sqrt{\lambda_{\max}}}\right\}\right).
\end{aligned}$$

因此，我们有 $\boldsymbol{\Delta}_4$ 的界，

$$\left|\boldsymbol{\Delta}_4\right| \leq c_1\frac{C^*}{\lambda_{\min}}\frac{(d+\delta_{t+1})\,\sigma^2}{t+1+C_b d}.$$

最后，在 $\mathcal{E}_s^X$ 和 $\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\| \leq 3\sigma^2/\lambda_{\max}$ 下， $\boldsymbol{\Delta}_5$ 项满足

$$\begin{aligned}
&\left\| \prod_{l=s+1}^{t}\left(1-\frac{\pi}{K}\frac{C_a}{l+C_b d}\right)\left(\frac{1}{s+C_b d}\right)^2 \boldsymbol{\Xi}_3^{(s)} \middle| \mathcal{F}_s, \mathcal{E}_s^X \right\|_{\Psi_1} \\
&\qquad \leq C\lambda_{\max}(d+\delta_s)\sigma^2\frac{(s+C_b d)^{C_a\cdot\pi/K-2}}{(t+1+C_b d)^{C_a\cdot\pi/K}} + C(d+\delta_s)^2\lambda_{\max}^2\frac{(s+C_b s)^{C_a\cdot\pi/K-3}}{(t+1+C_b d)^{C_a\cdot\pi/K}}.
\end{aligned}$$

这意味着在事件 $\cup\mathcal{E}_s^X$下，我们有

$$\begin{aligned}
&\mathbb{P}\left(\left|\sum_{s=0}^{t}\prod_{l=s+1}^{t}\left(1-\frac{\pi}{K}\frac{C_a}{l+C_b d}\right)\left(\frac{1}{s+C_b d}\right)^2\boldsymbol{\Xi}_3^{(s)}\right| \geq c_1 C^*\frac{\lambda_{\min}}{C_a^2}\frac{(d+\delta_{t+1})\,\sigma^2}{t+1+C_b d}\right) \\
&\qquad \leq \exp\left(-c_2\left(d+\frac{t}{C_b}\right)\min\left\{\frac{\lambda_{\min}^2}{\lambda_{\max}^2}\frac{\pi}{K}\frac{C_b}{C_a}\left(\frac{C^*}{C_a}\right)^2, \frac{\lambda_{\min}}{\lambda_{\max}}\frac{C^* C_b}{C_a^2}\right\}\right),
\end{aligned}$$

在此条件下，我们有

$$\left|\boldsymbol{\Delta}_5\right| \leq c_1\frac{C^*}{\lambda_{\min}}\frac{(d+\delta_{t+1})\,\sigma^2}{t+1+C_b d}.$$

因此，在上述事件下，我们有

$$\|\boldsymbol{\beta}_i^{(t+1)} - \boldsymbol{\beta}_i^*\|^2 \leq \frac{C_b\,d}{t+1+C_b d}\frac{\sigma^2}{\lambda_{\max}} + 4C_a\frac{K}{\pi}\frac{\lambda_{\max}}{\lambda_{\min}^2}\frac{\mathrm{d}\sigma^2}{t+1+C_b d} + 0.3\frac{C^*}{\lambda_{\min}}\frac{(d+\delta_{t+1})\,\sigma^2}{t+1+C_b d}.$$

有了 $C^* = 8C_a\frac{K}{\pi}\frac{\lambda_{\max}}{\lambda_{\min}} + 2C_b\frac{\lambda_{\min}}{\lambda_{\max}}$和 $C_b \geq 3C_a\frac{K}{\pi}\frac{\lambda_{\max}^2}{\lambda_{\min}^2}$，我们有

$$\|\boldsymbol{\beta}_i^{(t+1)} - \boldsymbol{\beta}_i^*\|^2 \leq \frac{C^*\,(d+\delta_{t+1})}{t+1+C_b d}\frac{\sigma^2}{\lambda_{\min}}, \quad \text{for all } i \in [K],$$

以超过 $1 - 2K\exp\left(-cd - c\frac{t}{C_a}\right) - K\exp\left(-c(d+\delta_{t+1})\right)$ 的概率

第二阶段 我们首先考虑集合 $\mathcal{A}$ 的收敛动态，然后讨论 $i \in \mathcal{A}^c$。在此阶段，$\pi_t \in [0,1]$ 是任意的。我们将事件记为

$$\mathcal{E}_t := \left\{\|\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i^*\|^2 \leq \frac{C^{**}\,(d+\delta_t)}{t+C_b d}\frac{\sigma^2}{\lambda_{\min}},\ i \in \mathcal{A}, \quad \|\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i^*\| \leq \frac{h}{2},\ i \in \mathcal{A}^c\right\},$$

这与第一阶段中的事件不同。我们通过归纳法证明。值得注意的是$\mathcal{E}_{t_1}$成立。接下来，我们将在 $\cup_{s=t_1}^{t}\mathcal{E}_s$, $\mathcal{E}_{t+1}$成立的条件下证明。然后给定 {最大值$_{i\in[K]}\|\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i^*\| \leq \frac{h}{2}$}，命题 3 证明，对所有 $i \in \mathcal{A}$,

$$\mathbb{E}\left\{\|\boldsymbol{\beta}_i^{(t+1)} - \boldsymbol{\beta}_i^*\|^2\middle|\mathcal{F}_t\right\} \leq \|\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i^*\|^2 - \eta_t\lambda_{\min}\frac{1}{K}\|\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i^*\|^2 + C\eta_t^2 d\lambda_{\max}\sigma^2,$$

其中 $\frac{\pi_t}{K} + (1-\pi_t) \geq \frac{1}{K}$。因此，估计误差率可以界定为

$$\begin{aligned}\|\boldsymbol{\beta}_i^{(t+1)} - \boldsymbol{\beta}_i^*\|^2 &= \mathbb{E}\left\{\|\boldsymbol{\beta}_i^{(t+1)} - \boldsymbol{\beta}_i^*\|^2\middle|\mathcal{F}_t\right\} + \left\{\|\boldsymbol{\beta}_i^{(t+1)} - \boldsymbol{\beta}_i^*\|^2 - \mathbb{E}\left\{\|\boldsymbol{\beta}_i^{(t+1)} - \boldsymbol{\beta}_i^*\|^2\middle|\mathcal{F}_t\right\}\right\} \\ &\leq \|\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i^*\|^2 - \eta_t\lambda_{\min}\frac{1}{K}\|\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i^*\|^2 + C\eta_t^2 d\lambda_{\max}\sigma^2 - 2\eta_t\boldsymbol{\Xi}_1^{(t)} - 2\eta_t\boldsymbol{\Xi}_2^{(t)} + \eta_t^2\boldsymbol{\Xi}_3^{(t)}.\end{aligned}$$

将步长 $\eta_t = \frac{C_a}{\lambda_{\min}}\frac{1}{t+C_b d}$ 代入上述方程，得到

$$\begin{aligned}\|\boldsymbol{\beta}_i^{(t+1)} - \boldsymbol{\beta}_i^*\|^2 \leq{}& \left(1 - \frac{C_a/K}{t+C_b d}\right)\|\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i^*\|^2 + \frac{C\,C_a^2}{(t+C_b d)^2}\frac{\lambda_{\max}}{\lambda_{\min}^2}\,d\,\sigma^2 \\ &- \frac{2\,C_a}{\lambda_{\min}}\frac{1}{t+C_b d}\boldsymbol{\Xi}_1^{(t)} - \frac{2\,C_a}{\lambda_{\min}}\frac{1}{t+C_b d}\boldsymbol{\Xi}_2^{(t)} + \frac{C_a^2}{\lambda_{\min}^2}\left(\frac{1}{t+C_b d}\right)^2\boldsymbol{\Xi}_3^{(t)}.\end{aligned}$$

由此可得

$$\|\boldsymbol{\beta}_i^{(t+1)}-\boldsymbol{\beta}_i^*\|^2 \le \underbrace{\prod_{s=t_1}^{t}\left(1-\frac{C_a/K}{s+C_bd}\right)\|\boldsymbol{\beta}_i^{(t_1)}-\boldsymbol{\beta}_i^*\|^2}_{\boldsymbol{\Theta}_1} + \underbrace{C\frac{\lambda_{\max}}{\lambda_{\min}^2}\sigma^2\sum_{s=t_1}^{t}\prod_{l=s+1}^{t}\left(1-\frac{C_a/K}{l+C_bd}\right)\frac{C_a^2d}{(s+C_bd)^2}}_{\boldsymbol{\Theta}_2}$$
$$\underbrace{-\frac{2\,C_a}{\lambda_{\min}}\sum_{s=t_1}^{t}\prod_{l=s+1}^{t}\left(1-\frac{C_a/K}{l+C_bd}\right)\frac{1}{s+C_bd}\boldsymbol{\Xi}_1^{(s)}}_{\boldsymbol{\Theta}_3}$$
$$\underbrace{-\frac{2\,C_a}{\lambda_{\min}}\sum_{s=t_1}^{t}\prod_{l=s+1}^{t}\left(1-\frac{C_a/K}{l+C_bd}\right)\frac{1}{s+C_bd}\boldsymbol{\Xi}_2^{(s)}}_{\boldsymbol{\Theta}_4}$$
$$\underbrace{-\frac{C_a^2}{\lambda_{\min}^2}\sum_{s=t_1}^{t}\prod_{l=s+1}^{t}\left(1-\frac{C_a/K}{l+C_bd}\right)\left(\frac{1}{s+C_bd}\right)^2\boldsymbol{\Xi}_3^{(s)}}_{\boldsymbol{\Theta}_5}.$$

根据引理 16 以及 $C_a > K$，我们有

$$\prod_{s=t_1}^{t}\left(1-\frac{C_a/K}{s+C_bd}\right) \le \frac{t_1+C_bd}{t+1+C_bd},$$

这 证明了 $\boldsymbol{\Theta}_1 \le \frac{t_1+C_bd}{t+1+C_bd}\|\boldsymbol{\beta}^{(t_1)}{}_i - \boldsymbol{\beta}_i^*\|^2 \le \frac{Cd}{t+1+C_bd}\sigma^2/\lambda_{\min}$。此外，引理 16 证明 s

$$\left(1-\frac{C_a/K}{s+1+C_bd}\right)\cdots\left(1-\frac{C_a/K}{t+C_bd}\right)\left(\frac{1}{s+C_bd}\right)^2 \le \left(\frac{s+C_bd}{t+1+C_bd}\right)^{C_a/K}\left(\frac{1}{s+C_bd}\right)^2.$$

接着，对 s 求和，并且引理 17 表明

$$\sum_{s=t_1}^{t}\left(1-\frac{C_a/K}{s+1+C_bd}\right)\cdots\left(1-\frac{C_a/K}{t+C_bd}\right)\left(\frac{1}{s+C_bd}\right)^2 \le \frac{2K}{C_a}\frac{1}{t+1+C_bd},$$

这证明了。

$$\boldsymbol{\Theta}_2 \le C\frac{C_a}{\lambda_{\min}}\frac{\lambda_{\max}}{\lambda_{\min}}\frac{\mathrm{d}\sigma^2}{t+1+C_bd}.$$

接下来考虑项 $\boldsymbol{\Theta}_3$。它的单个分量具有

$$\left\|\prod_{l=s+1}^{t}\left(1-\frac{C_a/K}{l+C_bd}\right)\frac{1}{s+C_bd}\boldsymbol{\Xi}_1^{(s)}\middle|\mathcal{F}_s\right\|_{\Psi_1} \le C\lambda_{\max}\frac{(s+C_bd)^{C_a/K-1}}{(t+1+C_bd)^{C_a/K}}\|\boldsymbol{\beta}_i^{(s)}-\boldsymbol{\beta}_i^*\|^2.$$

在事件 $\cup_{s=t_1}^{t}\mathcal{E}_s$下，引理 5 推得

$$\mathbb{P}\left(\left|\sum_{s=t_1}^{t}\prod_{l=s+1}^{t}\left(1-\frac{C_a/K}{l+C_bd}\right)\frac{1}{s+C_bd}\boldsymbol{\Xi}_1^{(s)}\right| \ge c_1\frac{C^*}{C_a}\frac{(d+\delta_{t+1})\,\sigma^2}{t+1+C_bd}\right) \le \exp\left(-c_2(C_bd+t)\frac{1}{C_a}\frac{\lambda_{\min}^2}{\lambda_{\max}^2}\right),$$

其中 $c_1 < 0.1$ 是某个小常数，于是我们有

$$|\mathbf{\Theta}_3| \leq c_1 C^{**} \frac{1}{\lambda_{\min}} \frac{(d+\delta_{t+1})\,\sigma^2}{t+1+C_b d}.$$

至于 $\mathbf{\Theta}_4$ 术语，我们有

$$\left\| \prod_{l=s+1}^{t} \left(1 - \frac{C_a/K}{l+C_b d}\right) \frac{1}{s+C_b d} \mathbf{\Xi}_2^{(s)} \middle| \mathcal{F}_s \right\|_{\Psi_1} \leq C\sqrt{\lambda_{\max}}\sigma \frac{(s+C_b d)^{C_a/K-1}}{(t+1+C_b d)^{C_a/K}} \|\boldsymbol{\beta}_i^{(s)} - \boldsymbol{\beta}_i^*\|.$$

然后在事件 $\cup_{s=t_1}^{t} \mathcal{E}_s$下，我们有

$$\begin{aligned}
&\mathbb{P}\left(\left| \sum_{s=t_1}^{t} \prod_{l=s+1}^{t} \left(1 - \frac{C_a/K}{l+C_b d}\right) \frac{1}{s+C_b d} \mathbf{\Xi}_2^{(s)} \right| \geq c_1 \frac{C^*}{C_a} \frac{(d+\delta_{t+1})\,\sigma^2}{t+1+C_b d} \right) \\
&\qquad \leq \exp\left(-c_2 \min\left\{ (d+\delta_{t+1}) \frac{C^*}{C_a} \frac{\lambda_{\min}}{\lambda_{\max}}, \sqrt{d+\delta_{t+1}} \sqrt{d+t/C_b} \frac{\sqrt{C^* C_b}}{C_a} \frac{\sqrt{\lambda_{\min}}}{\sqrt{\lambda_{\max}}} \right\} \right),
\end{aligned}$$

这将导致

$$|\mathbf{\Theta}_4| \leq c_1 \frac{C^{**}}{\lambda_{\min}} \frac{(d+\delta_{t+1})\,\sigma^2}{t+1+C_b d}.$$

最后，考虑术语 $\mathbf{\Theta}_{5\circ}$ 在 $\mathcal{F}_s$ 以及事件 $\mathcal{E}_s$， $\mathcal{E}_s^X$下，我们有

$$\left\| \prod_{l=s+1}^{t} \left(1 - \frac{C_a/K}{l+C_b d}\right) \left(\frac{1}{s+C_b d}\right)^2 \mathbf{\Xi}_3^{(s)} \middle| \mathcal{F}_s, \mathcal{E}_s^X \right\|_{\Psi_1} \leq C\lambda_{\max}(d+\delta_s)\sigma^2 \frac{(s+C_b d)^{C_a-2}}{(t+1+C_b d)^{C_a}}.$$

然后在事件 $\cup_{s=t_1}^{t} \{\mathcal{E}_s \cup \mathcal{E}_s^X\}$下，我们有

$$\begin{aligned}
&\mathbb{P}\left(\left| \sum_{s=t_1}^{t} \prod_{l=s+1}^{t} \left(1 - \frac{C_a/K}{l+C_b d}\right) \left(\frac{1}{s+C_b d}\right)^2 \mathbf{\Xi}_3^{(s)} \right| \geq c_1 C^{**} \frac{\lambda_{\min}}{C_a^2} \frac{(d+\delta_{t+1})\,\sigma^2}{t+1+C_b d} \right) \\
&\qquad \leq \exp\left(-c_2 \left(d + \frac{t}{C_b}\right) \min\left\{ \frac{\lambda_{\min}^2}{\lambda_{\max}^2} \frac{C_b}{C_a} \left(\frac{C^{**}}{C_a}\right)^2, \frac{\lambda_{\min}}{\lambda_{\max}} \frac{C^{**} C_b}{C_a^2} \right\} \right),
\end{aligned}$$

这表明

$$|\mathbf{\Theta}_5| \leq c_1 \frac{C^{**}}{\lambda_{\min}} \frac{(d+\delta_{t+1})\,\sigma^2}{t+1+C_b d}.$$

因此，综上所述，对于 $C^{**} = C^* + 2CC_a \frac{\lambda_{\max}}{\lambda_{\min}}$，我们有

$$\begin{aligned}
\|\boldsymbol{\beta}_i^{(t+1)} - \boldsymbol{\beta}_i^*\|^2 &\leq \frac{C^*}{\lambda_{\min}} \frac{d\,\sigma^2}{t+1+C_b d} + C \frac{C_a}{\lambda_{\min}} \frac{\lambda_{\max}}{\lambda_{\min}} \frac{\mathrm{d}\sigma^2}{t+1+C_b d} + 3c_1 \frac{C^{**}}{\lambda_{\min}} \frac{(d+\delta_{t+1})\,\sigma^2}{t+1+C_b d} \\
&\leq \frac{C^{**}}{\lambda_{\min}} \frac{(d+\delta_{t+1})\,\sigma^2}{t+1+C_b d},
\end{aligned}$$

对所有 $i \in \mathcal{A}$，以大于 $1 - |\mathcal{A}|\exp(-c(d + t/C_a)) - |\mathcal{A}|\exp(-c(d + \delta_{t+1}))$ 的概率成立。此外，对于 $C_b \geq C_a \frac{K}{\pi} \frac{\lambda_{\max}^2}{\lambda_{\min}^2}$，我们有 $\|\boldsymbol{\beta}_i^{(t+1)} - \boldsymbol{\beta}_i^*\|^2 \leq C \frac{\sigma^2}{\lambda_{\max}}$；给定 $\frac{C^*}{\lambda_{\min}} \frac{(d+\delta_{t_1}}{t_1 + C_b d} \frac{\sigma^2}{\leq} \frac{h^2)}{16}$，我们有 $\|\boldsymbol{\beta}_i^{(t+1)} - \boldsymbol{\beta}_i^*\|^2 \leq \frac{h^2}{4}$。

接下来考虑 $i \in \mathcal{A}^c$。命题3 证明，在 $\mathcal{E}_t$下，我们有 e

$$\mathbb{E}\left\{\|\boldsymbol{\beta}_i^{(t+1)} - \boldsymbol{\beta}_i^*\|^2 \middle| \mathcal{F}_t\right\} \leq \|\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i^*\|^2 - 2\eta_t \pi_t \lambda_{\min} \|\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i^*\|^2 + C\eta_t^2 \frac{\pi_t}{K} \lambda_{\max}\, d\sigma^2.$$

将 $\eta_t = \frac{C_a}{\lambda_{\min}} \frac{1}{t + C_b d}$ 代入上述方程，可得

$$\mathbb{E}\left\{\|\boldsymbol{\beta}_i^{(t+1)} - \boldsymbol{\beta}_i^*\|^2 \middle| \mathcal{F}_t\right\} \leq \left(1 - \frac{C_a\, \pi_t}{t + C_b d}\right) \|\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i^*\|^2 + CC_a^2 \frac{\pi_t}{K} \frac{\lambda_{\max}}{\lambda_{\min}^2} \left(\frac{1}{t + C_b d}\right)^2.$$

因此，我们有

$$\begin{aligned}
\|\boldsymbol{\beta}_i^{(t+1)} - \boldsymbol{\beta}_i^*\|^2 \leq\ & \underbrace{\prod_{s=t_1}^{t} \left(1 - \frac{C_a \pi_s}{s + C_b d}\right) \|\boldsymbol{\beta}_i^{(t_1)} - \boldsymbol{\beta}_i^*\|^2}_{\boldsymbol{\Gamma}_1} + \underbrace{CC_a^2 d\sigma^2 \frac{1}{K} \frac{\lambda_{\max}}{\lambda_{\min}^2} \sum_{s=t_1}^{t} \prod_{l=s+1}^{t} \left(1 - \frac{C_a \pi_l}{l + C_b d}\right) \frac{\pi_s}{(s + C_b d)^2}}_{\boldsymbol{\Gamma}_2} \\
& \underbrace{- \frac{2\, C_a}{\lambda_{\min}} \sum_{s=t_1}^{t} \prod_{l=s+1}^{t} \left(1 - \frac{C_a \pi_l}{l + C_b d}\right) \frac{1}{s + C_b d} \boldsymbol{\Xi}_1^{(s)}}_{\boldsymbol{\Gamma}_3} \\
& \underbrace{- \frac{2\, C_a}{\lambda_{\min}} \sum_{s=t_1}^{t} \prod_{l=s+1}^{t} \left(1 - \frac{C_a \pi_l}{l + C_b d}\right) \frac{1}{s + C_b d} \boldsymbol{\Xi}_2^{(s)}}_{\boldsymbol{\Gamma}_4} \\
& \underbrace{+ \frac{C_a^2}{\lambda_{\min}^2} \sum_{s=t_1}^{t} \prod_{l=s+1}^{t} \left(1 - \frac{C_a \pi_l}{l + C_b d}\right) \left(\frac{1}{s + C_b d}\right)^2 \boldsymbol{\Xi}_3^{(s)}}_{\boldsymbol{\Gamma}_5}.
\end{aligned}$$

对于 $\boldsymbol{\Gamma}_1$，其仅有 $\boldsymbol{\Gamma}_1 \leq \|\boldsymbol{\beta}^{(t_1)}{}_i - \boldsymbol{\beta}_i^*\|^2 \leq \frac{h^2}{16}$。至于 $\boldsymbol{\Gamma}_2$，我们有

$$\boldsymbol{\Gamma}_2 \leq CC_a^2 d\sigma^2 \frac{1}{K} \frac{\lambda_{\max}}{\lambda_{\min}^2} \sum_{s=t_1}^{t} \frac{1}{(s + C_b d)^2} \leq C \frac{C_a^2}{K} \frac{\lambda_{\max}}{\lambda_{\min}^2} \frac{d\sigma^2}{t_1 + C_b d}.$$

值得注意的是，在事件 $\{\max_j \|\boldsymbol{\beta}_j^{(s)} - \boldsymbol{\beta}_j^*\| \leq \frac{h}{2}\}$, $a_t = i$ 中，只有当 $\alpha_t = 1$ 且 $\mathbb{I}\{a_t = i\}$ 与 $\mathbf{X}_t$, $\boldsymbol{\beta}_i^{(t)}$ 独立时，才符合 $\mathbb{I}\{a_t = i\} \sim$《伯努利》$\left(\frac{\pi_t}{K}\right)$。于是，$\boldsymbol{\Gamma}_3$ 的单个分量具有

$$\left\| \prod_{l=s+1}^{t} \left(1 - \frac{C_a \pi_l}{l + C_b d}\right) \frac{1}{s + C_b d} \boldsymbol{\Xi}_1^{(s)} \middle| \mathcal{F}_s \right\|_{\Psi_1} \leq \mathbb{I}\{\pi_s \neq 0\} \cdot \lambda_{\max} \frac{h^2}{s + C_b d}.$$

根据引理 9，我们有

$$\mathbb{P}\left(\left|\sum_{s=t_1}^{t}\prod_{l=s+1}^{t}\left(1-\frac{C_a\pi_l}{l+C_bd}\right)\frac{1}{s+C_bd}\boldsymbol{\Xi}_1^{(s)}\right|>c_1\mathbb{I}\left\{\pi_s\neq 0\right\}h^2\frac{\lambda_{\min}}{C_a}\right)$$
$$\leq \mathbb{I}\left\{\pi_s\neq 0\right\}\exp\left(-c_2d\min\left\{\frac{C_b}{C_a^2}\frac{\lambda_{\min}^2}{\lambda_{\max}^2},\frac{C_b}{C_a}\frac{\lambda_{\min}}{\lambda_{\max}}\right\}\right),$$

在这种情况下，它具有

$$\left|\boldsymbol{\Gamma}_3\right|\leq c_1\mathbb{I}\left\{\pi_s\neq 0\right\}\cdot h^2.$$

同样，它具有

$$\left\|\prod_{l=s+1}^{t}\left(1-\frac{C_a\pi_l}{l+C_bd}\right)\frac{1}{s+C_bd}\boldsymbol{\Xi}_2^{(s)}\middle|\mathcal{F}_s\right\|_{\Psi_1}\leq \mathbb{I}\left\{\pi_s\neq 0\right\}\sqrt{\lambda_{\max}}\sigma\frac{h}{s+C_bd}.$$

于是，引理9 保证

$$\mathbb{P}\left(\left|\sum_{s=t_1}^{t}\prod_{l=s+1}^{t}\left(1-\frac{C_a\pi_l}{l+C_bd}\right)\frac{1}{s+C_bd}\boldsymbol{\Xi}_2^{(s)}\right|>c_1\mathbb{I}\left\{\pi_s\neq 0\right\}h^2\frac{\lambda_{\min}}{C_a}\right)$$
$$\leq \mathbb{I}\left\{\pi_s\neq 0\right\}\exp\left(-cd\min\left\{\frac{C^*}{C_a^2}\frac{\lambda_{\min}}{\lambda_{\max}},\frac{\sqrt{\lambda_{\min}}}{\sqrt{\lambda_{\max}}}\frac{\sqrt{C^*C_b}}{C_a}\right\}\right),$$

这意味着 $\left|\boldsymbol{\Gamma}_4\right|\leq c_1h^2$。然后考虑项 $\boldsymbol{\Gamma}_5$,

$$\left\|\prod_{l=s+1}^{t}\left(1-\frac{C_a\pi_l}{l+C_bd}\right)\left(\frac{1}{s+C_bd}\right)^2\boldsymbol{\Xi}_3^{(s)}\middle|\mathcal{F}_s\right\|_{\Psi_1}\leq \mathbb{I}\left\{\pi_s\neq 0\right\}\left(\frac{1}{s+C_bd}\right)^2 d\lambda_{\max}\sigma^2.$$

根据引理 9,

$$\mathbb{P}\left(\left|\sum_{s=t_1}^{t}\prod_{l=s+1}^{t}\left(1-\frac{C_a\pi_l}{l+C_bd}\right)\left(\frac{1}{s+C_bd}\right)^2\boldsymbol{\Xi}_3^{(s)}\right|>c_1\mathbb{I}\left\{\pi_s\neq 0\right\}\frac{\lambda_{\min}^2}{C_a^2}h^2\right)$$
$$\leq \mathbb{I}\left\{\pi_s\neq 0\right\}\exp\left(-cd\min\left\{\frac{\lambda_{\min}^2}{\lambda_{\max}^2}\frac{C^{*2}C_b}{C_a^4},\frac{\lambda_{\min}}{\lambda_{\max}}\frac{C^*C_b}{C_a^2}\right\}\right),$$

由此我们得到 $\left|\boldsymbol{\Gamma}_5\right|\leq c_1h^2$。因此，综上所述，给定 $C_b\geq CC_a^2\lambda_{\max}^2/\lambda_{\min}^2$，以下结论对所有 $i\in\mathcal{A}^c$ 均成立，其关于 $1-\mathbb{I}\left\{\pi_s\neq 0\right\}|\mathcal{A}^c|$ 的概率为 $\exp(-cd)$

$$\|\boldsymbol{\beta}_i^{(t+1)}-\boldsymbol{\beta}_i^*\|^2\leq\frac{h^2}{4}.$$

□

C.4 定理5的证明

证明。 首先，考虑时刻 t的遗憾，它可以分解为若干项，

$$\begin{aligned}
&\mathbb{E}\left\{\max_{i\in[K]}\mathbf{X}_t^\top\boldsymbol{\beta}_i^*-\mathbf{X}_t^\top\boldsymbol{\beta}_{a_t}^*\right\}\\
&\quad=\underbrace{\mathbb{E}\left\{\mathbb{I}\{\alpha_t=1\}\cdot\left(\max_{i\in[K]}\mathbf{X}_t^\top\boldsymbol{\beta}_i^*-\mathbf{X}_t^\top\boldsymbol{\beta}_{a_t}^*\right)\right\}}_{\mathbf{R}_1}+\underbrace{\mathbb{E}\left\{\mathbb{I}\{\alpha_t=0\}\cdot\left(\max_{i\in[K]}\mathbf{X}_t^\top\boldsymbol{\beta}_i^*-\mathbf{X}_t^\top\boldsymbol{\beta}_{a_t}^*\right)\right\}}_{\mathbf{R}_2}.
\end{aligned}$$

我们将分别分析 $t\le t_1$ 和 $t>t_1$。当 $t<t_1$时，

$$\begin{aligned}
\mathbf{R}_1&=\frac{\pi_t}{K}\cdot\sum_{j=1}^K\mathbb{E}\left\{\max_i\mathbf{X}_t^\top\boldsymbol{\beta}_i^*-\mathbf{X}_t^\top\boldsymbol{\beta}_j^*\right\}\\
&\le\frac{\pi_t}{K}\cdot\min\left\{\sum_{j=1}^K\sum_{i=1}^K\mathbb{E}\left\{\left|\mathbf{X}_t^\top\boldsymbol{\beta}_i^*-\mathbf{X}_t^\top\boldsymbol{\beta}_j^*\right|\right\},\sum_{j=1}^K\mathbb{E}\left\{\|\mathbf{X}_t\|\cdot\max_i\|\boldsymbol{\beta}_i^*-\boldsymbol{\beta}_j^*\|\right\}\right\}\\
&\le\sqrt{\lambda_{\max}}\frac{\pi_t}{K}\cdot\min\left\{\sum_{j=1}^K\sum_{i=1}^K\|\boldsymbol{\beta}_i^*-\boldsymbol{\beta}_j^*\|,\sqrt{d}\sum_{j=1}^K\max_i\|\boldsymbol{\beta}_i^*-\boldsymbol{\beta}_j^*\|\right\}.
\end{aligned}$$

记事件 $\mathcal{E}_{t_1}:=\left\{\|\boldsymbol{\beta}_i^{(t)}-\boldsymbol{\beta}_i^*\|^2\le C^*\frac{1}{\lambda_{\min}}\frac{d+\delta_t}{t+C_bd}\sigma^2,\quad i\in[K],\ \text{对所有}\ t\in[t_1]\right\}$。定理 4 证明了

$$\mathbb{P}(\mathcal{E}_{t_1})\ge1-c'KC_b\exp(-cd)-K\exp(-cd)\sum_{s=0}^{t_1}\exp(-c\delta_s).$$

记事件 $\mathcal{E}^{(t)}:=\left\{\|\boldsymbol{\beta}_i^{(t)}-\boldsymbol{\beta}_i^*\|^2\le C^*\frac{1}{\lambda_{\min}}\frac{d+\delta_t}{t+C_bd}\sigma^2,\quad i\in[K]\right\}$。那么

$$\begin{aligned}
\mathbf{R}_2=&\underbrace{(1-\pi_t)\mathbb{E}\left\{\left(\max_{i\in[K]}\mathbf{X}_t^\top\boldsymbol{\beta}_i^*-\mathbf{X}_t^\top\boldsymbol{\beta}_{a_t}^*\right)\cdot\mathbb{I}\left\{\mathcal{E}^{(t)}\right\}\right\}}_{\mathbf{R}_{21}}\\
&+\underbrace{(1-\pi_t)\mathbb{E}\left\{\left(\max_{i\in[K]}\mathbf{X}_t^\top\boldsymbol{\beta}_i^*-\mathbf{X}_t^\top\boldsymbol{\beta}_{a_t}^*\right)\cdot\mathbb{I}\{\mathcal{E}^{(t)C}\}\right\}}_{\mathbf{R}_{22}}.
\end{aligned}$$

为简便起见，记为 $a_t^*\in\arg\max_i\mathbf{X}^\top\boldsymbol{\beta}_i^*$。注意到 $\mathbf{X}_t^\top\boldsymbol{\beta}_{a_t}^{(t)}\ge\max_i\mathbf{X}_t^\top\boldsymbol{\beta}_i^{(t)}$，于是 $\mathbf{R}_2$ 满足，

$$\begin{aligned}
\mathbf{R}_{21}&\le(1-\pi_t)\cdot\mathbb{E}\left\{\left\{\mathbf{X}_t^\top\boldsymbol{\beta}_{a_t^*}^*-\mathbf{X}_t^\top\boldsymbol{\beta}_{a_t^*}^{(t)}+\mathbf{X}_t^\top\boldsymbol{\beta}_{a_t}^{(t)}-\mathbf{X}_t^\top\boldsymbol{\beta}_{a_t}^*\right\}\cdot\mathbb{I}\{\mathcal{E}^{(t)}\}\right\}\\
&\le(1-\pi_t)\cdot\mathbb{E}\left\{\left|\mathbf{X}_t^\top\boldsymbol{\beta}_{a_t^*}^*-\mathbf{X}_t^\top\boldsymbol{\beta}_{a_t^*}^{(t)}\right|\cdot\mathbb{I}\{\mathcal{E}^{(t)}\}+\left|\mathbf{X}_t^\top\boldsymbol{\beta}_{a_t}^{(t)}-\mathbf{X}_t^\top\boldsymbol{\beta}_{a_t}^*\right|\cdot\mathbb{I}\{\mathcal{E}^{(t)}\}\right\}\\
&\le2(1-\pi_t)\cdot\min\left\{\sum_{i=1}^K\mathbb{E}\left|\mathbf{X}_t^\top\left(\boldsymbol{\beta}_i^*-\boldsymbol{\beta}_i^{(t)}\right)\right|\cdot\mathbb{I}\{\mathcal{E}^{(t)}\},\mathbb{E}\left|\|\mathbf{X}\|\cdot\max_i\|\boldsymbol{\beta}_i^{(t)}-\boldsymbol{\beta}_i^*\|\cdot\mathbb{I}\{\mathcal{E}^{(t)}\}\right|\right\}.
\end{aligned}$$

因此，考虑到事件 $\mathcal{E}^{(t)}$，我们有

$$\mathbf{R}_{21} \leq 2(1-\pi_t)\sqrt{C^*}\sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}}\sqrt{\frac{d+\delta_t}{t+C_b d}}\,\sigma\,\min\left\{K,\,\sqrt{d}\right\}.$$

同理，我们有

$$\mathbf{R}_{22} \leq (1-\pi_t)\min\left\{\sum_{j=1}^{K}\sum_{i=1}^{K}\mathbb{E}\left|\mathbf{X}_t^\top\left(\boldsymbol{\beta}_i^*-\boldsymbol{\beta}_j^*\right)\right|, \mathbb{E}\|\mathbf{X}_t\|\cdot\max_{i,j}\|\boldsymbol{\beta}_i^*-\boldsymbol{\beta}_j^*\|\right\}\cdot\mathbb{P}(\mathcal{E}^{(t)C}).$$

还需注意，$\sum_{t=0}^{t_1}\mathbb{P}(\mathcal{E}^{(t)C}) = 1-\mathbb{P}(\mathcal{E}_{t_1}) \leq c'KC_b\exp(-cd) + K\exp(-cd)\sum_{s=0}^{t_1}\exp(-c\delta_s)$。因此，我们有

$$\begin{aligned}\mathsf{Regret}(t_1) \leq \sqrt{\lambda_{\max}}\frac{\sum_{s=0}^{t_1}\pi_s}{K}\cdot\min\left\{\sum_{j=1}^{K}\sum_{i=1}^{K}\|\boldsymbol{\beta}_i^*-\boldsymbol{\beta}_j^*\|, \sqrt{d}\sum_{j=1}^{K}\max_i\|\boldsymbol{\beta}_i^*-\boldsymbol{\beta}_j^*\|\right\}\\ + C\sqrt{C^*}\sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}}\sigma\min\left\{K,\sqrt{d}\right\}\sum_{s=0}^{t_1}\sqrt{\frac{d+\delta_s}{s+C_b d}}\\ + K\sqrt{\lambda_{\max}}\exp(-cd)\min\left\{\sum_{j=1}^{K}\sum_{i=1}^{K}\mathbb{E}\|\boldsymbol{\beta}_i^*-\boldsymbol{\beta}_j^*\|, \sqrt{d}\max_{i,j}\|\boldsymbol{\beta}_i^*-\boldsymbol{\beta}_j^*\|\right\}\cdot\left(c'C_b+\sum_{s=0}^{t_1}\exp(-c\delta_s)\right)\end{aligned}$$

接下来考虑 $t > t_1$ 的遗憾。对于 $t > t_1$，$\mathbf{R}_1$ 的界与 $t \leq t_1$ 时的情况相同。对于 $\mathbf{R}_2$，的分析，我们记

$$\mathcal{E}^{(t)} := \left\{\|\boldsymbol{\beta}_i^{(t)}-\boldsymbol{\beta}_i^*\|^2 \leq C^*\frac{1}{\lambda_{\min}}\frac{d+\delta_t}{t+C_b d}\sigma^2,\ i\in\mathcal{A};\ \|\boldsymbol{\beta}_i^{(t)}-\boldsymbol{\beta}_i^*\|^2 \leq \frac{h^2}{4},\ i\in\mathcal{A}^c\right\}.$$

Then 根据引理 5，我们知道在上述事件下，$a_t \in \mathcal{A}$。因此，我们有 e

$$\begin{aligned}&\mathbb{E}\left\{\left(\max_{i\in[K]}\mathbf{X}_t^\top\boldsymbol{\beta}_i^*-\mathbf{X}_t^\top\boldsymbol{\beta}_{a_t}^*\right)\cdot\mathbb{I}\left\{\mathcal{E}^{(t)}\right\}\right\}\\ &\quad\leq\min\left\{\sum_{i\in\mathcal{A}}\mathbb{E}\left|\mathbf{X}_t^\top\left(\boldsymbol{\beta}_i^*-\boldsymbol{\beta}_i^{(t)}\right)\right|\cdot\mathbb{I}\{\mathcal{E}^{(t)}\}, \mathbb{E}\left|\|\mathbf{X}\|\cdot\max_{i\in\mathcal{A}}\|\boldsymbol{\beta}_i^{(t)}-\boldsymbol{\beta}_i^*\|\cdot\mathbb{I}\{\mathcal{E}^{(t)}\}\right|\right\}\\ &\quad\leq\sqrt{C^*}\sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}}\sqrt{\frac{d+\delta_t}{t+C_b d}}\,\sigma\,\min\left\{K,\,\sqrt{d}\right\}.\end{aligned}$$

$\mathbb{E}\left\{\left(\max^{i\in[K]}\mathbf{X}_t^\top\boldsymbol{\beta}_i^* - \mathbf{X}_t^\top\boldsymbol{\beta}_{a_t}^*\right)\cdot\mathbb{I}\left\{\mathcal{E}^{(t)C}\right\}\right\}$的分析与 $t < t_1$情形相同。因此综上，我们有

$$\begin{aligned}\mathsf{Regret}(T) \le \sqrt{\lambda_{\max}}\frac{\sum_{s=0}^T \pi_s}{K}\cdot\min\left\{\sum_{j=1}^K\sum_{i=1}^K\|\boldsymbol{\beta}_i^*-\boldsymbol{\beta}_j^*\|, \sqrt{d}\sum_{j=1}^K\max_i\|\boldsymbol{\beta}_i^*-\boldsymbol{\beta}_j^*\|\right\}\\ + C\sqrt{C^*}\sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}}\sigma\min\left\{K,\sqrt{d}\right\}\sum_{s=0}^T\sqrt{\frac{d+\delta_s}{s+C_b d}}\\ + \sqrt{\lambda_{\max}}\min\left\{\sum_{j=1}^K\sum_{i=1}^K\mathbb{E}\|\boldsymbol{\beta}_i^*-\boldsymbol{\beta}_j^*\|, \sqrt{d}\max_{i,j}\|\boldsymbol{\beta}_i^*-\boldsymbol{\beta}_j^*\|\right\}\\ \times\left(c'C_b\exp(-cd) + K\sum_{s=0}^{t_1}\exp(-c(\delta_s+d)) + |\mathcal{A}|\sum_{s=t_1+1}^T\exp(-c(\delta_s+d)) + |\mathcal{A}^c|\sum_{s=t_1+1}^T\mathbb{I}\{\pi_s\neq 0\}\right),\end{aligned}$$

至此，证明完成。 □

C.5 命题2的证明

为简洁起见，我们引入一些记号。记

$$\begin{aligned}\boldsymbol{\Sigma}_{i,t} &:= \mathbb{E}\left\{\mathbb{I}\{a_t=i\}\cdot\mathbf{X}_t\mathbf{X}_t^\top\middle|\mathcal{F}_t\right\}\\ &= \frac{\pi_t}{K}\mathbb{E}\left\{\mathbf{X}_t\mathbf{X}_t^\top\middle|\mathcal{F}_t\right\} + (1-\pi_t)\cdot\mathbb{E}\left\{\mathbf{X}_t\mathbf{X}_t^\top\cdot\mathbb{I}\left\{\mathbf{X}_t\in\mathcal{X}\left(i,\left\{\boldsymbol{\beta}_j^{(t)}\right\}_{i\in[K]}\right)\right\}\middle|\mathcal{F}_t\right\},\end{aligned}$$

and

$$\boldsymbol{\Sigma}_{i,t}^* := \frac{\pi_t}{K}\mathbb{E}\left\{\mathbf{X}_t\mathbf{X}_t^\top\middle|\mathcal{F}_t\right\} + (1-\pi_t)\cdot\mathbb{E}\left\{\mathbf{X}_t\mathbf{X}_t^\top\cdot\mathbb{I}\left\{\mathbf{X}_t\in\mathcal{U}_i^*\right\}\middle|\mathcal{F}_t\right\}.$$

矩阵 $\boldsymbol{\Sigma}_{i,\ t}^*$ 满足以下方程

$$\left(\frac{\pi_t}{K}+(1-\pi_t)\right)\cdot\lambda_{\min}\cdot\mathbf{I}_d \preceq \boldsymbol{\Sigma}_{i,t}^* \preceq \lambda_{\max}\cdot\mathbf{I}_d, \tag{23}$$

并且我们有如下均值收敛

$$\lim_{t\to+\infty}\mathbb{E}\left\|\boldsymbol{\Sigma}_{i,t}^* - \boldsymbol{\Sigma}_i(\pi^*)\right\| = 0. \tag{24}$$

此外，$\boldsymbol{\Sigma}_{i,\ t} - \boldsymbol{\Sigma}_{i,\ t}^*$ 可以写成，

$$\left\|\boldsymbol{\Sigma}_{i,t}-\boldsymbol{\Sigma}_{i,t}^*\right\| = (1-\pi_t)\left\|\mathbb{E}\left\{\mathbf{X}_t\mathbf{X}_t^\top\cdot\left(\mathbb{I}\left\{\mathbf{X}\in\mathcal{X}_t\left(i,\left\{\boldsymbol{\beta}_j^{(t)}\right\}_{i\in[K]}\right)\right\} - \mathbb{I}\left\{\mathbf{X}_t\in\mathcal{U}_i^*\right\}\right)\middle|\mathcal{F}_t\right\}\right\|$$

在假设 7 和引理 4下，$\left\|\boldsymbol{\Sigma}_{i,\ t} - \boldsymbol{\Sigma}_{i,\ t}^*\right\|$ 可以由以下内容给出上界，

$$\left\|\boldsymbol{\Sigma}_{i,t}-\boldsymbol{\Sigma}_{i,t}^*\right\| \le \begin{cases}(1-\pi_t)\cdot\kappa_0\cdot\max_{i\in\mathcal{A}}\|\boldsymbol{\beta}_i^{(t)}-\boldsymbol{\beta}_i^*\|, & \max_{i\in[K]}\|\boldsymbol{\beta}_i^{(t)}-\boldsymbol{\beta}_i^*\|\le h_0\\ (1-\pi_t)\cdot\lambda_{\max}, & \text{for all } \{\boldsymbol{\beta}_i\}_{i\in[K]}\end{cases}. \tag{25}$$

我们首先证明以下命题。

命题4。 在与命题2相同的假设和条件下，有

(1) 若 $\pi^* = 0$，则在事件 $\cup_{t=t_1}\{\max \|\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i^*\| \le \frac{h}{2}\}$ 下，对于 $i \in \mathcal{A}$，下式成立，

$$\mathbb{E}\|\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i^*\|^2 \le C\frac{d}{t + C_b d}\frac{\sigma^2}{\lambda_{\min}},$$

(2) 若 $\pi^* > 0$ 且 $C_a \ge 4K/\pi^*$，则对所有 $i \in [K]$，下式成立，

$$\mathbb{E}\|\boldsymbol{\beta}_i^{(t+1)} - \boldsymbol{\beta}_i^*\|^2 \le \left(\frac{N + 1 + C_b d}{t + 1 + C_b d}\right)^{C_a\pi^*/K}\|\boldsymbol{\beta}_i^{(N)} - \boldsymbol{\beta}_i^*\|^2 + \frac{Cd}{t + 1 + C_b d}\frac{\sigma^2}{\lambda_{\min}},$$

其中 N 是某个整数，使得对于所有 $t \ge N$， $\pi_t \ge \pi^*/2$。

证明。 我们先证明 $\pi^* = 0$ 情形，然后讨论 $\pi^* > 0$ 方案。

当 $\pi^* = 0$。对于 $i \in \mathcal{A}$，命题 3 证明了

$$\begin{aligned}\mathbb{E}\left\{\|\boldsymbol{\beta}_i^{(t+1)} - \boldsymbol{\beta}_i^*\|^2 \middle| \mathcal{F}_t\right\} \le{}& \|\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i^*\|^2 - 2\eta_t\frac{\pi_t}{K}\lambda_{\min}\|\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i^*\|^2 + 2\eta_t^2 d\lambda_{\max}^2\|\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i^*\|^2 \\ &- 2\eta_t(1 - \pi_t)\lambda_{\min}\|\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i^*\|^2 \cdot \mathbb{I}\left\{\max_j \|\boldsymbol{\beta}_j^{(t)} - \boldsymbol{\beta}_i^*\| \le \frac{h}{2}\right\} + 2\eta_t^2 d\lambda_{\max}\sigma^2.\end{aligned}$$

对上述等式两边分别取期望，得到

$$\begin{aligned}\mathbb{E}\left\{\|\boldsymbol{\beta}_i^{(t+1)} - \boldsymbol{\beta}_i^*\|^2\right\} \le{}& \left(1 - 2\eta_t\frac{\pi_t}{K}\lambda_{\min} + 2\eta_t^2 d\lambda_{\max}^2\right)\mathbb{E}\left\{\|\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i^*\|^2\right\} \\ &- 2\eta_t(1 - \pi_t)\lambda_{\min}\mathbb{E}\left\{\|\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i^*\|^2 \cdot \mathbb{I}\left\{\max_j \|\boldsymbol{\beta}_j^{(t)} - \boldsymbol{\beta}_i^*\| \le \frac{h}{2}\right\}\right\} + 2\eta_t^2 d\lambda_{\max}\sigma^2.\end{aligned}$$

对于 $t \le t_1$，根据探索率方案，我们有 $\pi_t \ge \pi_\circ$。因此，我们有

$$\begin{aligned}\mathbb{E}\left\{\|\boldsymbol{\beta}_i^{(t+1)} - \boldsymbol{\beta}_i^*\|^2\right\} \le{}& \left(1 - 2\frac{C_a}{t + C_b d}\frac{\pi}{K} + 2\frac{C_a^2 d}{(t + C_b d)^2}\frac{\lambda_{\max}^2}{\lambda_{\min}^2}\right)\mathbb{E}\left\{\|\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i^*\|^2\right\} \\ &+ 2\frac{C_a^2 d}{(t + C_b d)^2}\frac{\lambda_{\max}}{\lambda_{\min}^2}\sigma^2.\end{aligned}$$

根据 C_b 的选取，我们有

$$2\frac{C_a^2 d}{(t + C_b d)^2}\frac{\lambda_{\max}^2}{\lambda_{\min}^2} \le \frac{C_a}{t + C_b d}\frac{\pi}{K}.$$

因此，我们有

$$\begin{aligned}&\mathbb{E}\left\{\|\boldsymbol{\beta}_i^{(t+1)} - \boldsymbol{\beta}_i^*\|^2\right\} \le \left(1 - 2\frac{C_a}{t + C_b d}\frac{\pi}{K}\right)\mathbb{E}\left\{\|\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i^*\|^2\right\} + 2\frac{C_a^2 d}{(t + C_b d)^2}\frac{\lambda_{\max}}{\lambda_{\min}^2}\sigma^2 \\ &\le \prod_{l=0}^{t}\left(1 - 2\frac{C_a}{l + C_b d}\frac{\pi}{K}\right)\mathbb{E}\left\{\|\boldsymbol{\beta}_i^{(0)} - \boldsymbol{\beta}_i^*\|^2\right\} + 2\sum_{k=0}^{t}\prod_{l=k}^{t}\left(1 - 2\frac{C_a}{l + C_b d}\frac{\pi}{K}\right)\frac{C_a^2 d}{(k + C_b d)^2}\frac{\lambda_{\max}}{\lambda_{\min}^2}\sigma^2.\end{aligned}$$

根据引理16 和引理17，我们有

$$\mathbb{E}\left\{\|\boldsymbol{\beta}_i^{(t+1)}-\boldsymbol{\beta}_i^*\|^2\right\}\leq\frac{Cd}{t+1+C_bd}\frac{\sigma^2}{\lambda_{\min}}.$$

另一方面，对于 $t\geq t_1$，在满足条件 $\cup_{t=t_1}^{+\infty}\{\max_{j\in[k]}\|\boldsymbol{\beta}_j^{(t)}-\boldsymbol{\beta}_j^*\|\leq\frac{h}{2}\}$ 时，我们有

$$\mathbb{E}\left\{\|\boldsymbol{\beta}_i^{(t+1)}-\boldsymbol{\beta}_i^*\|^2\right\}\leq\left(1-2\frac{C_a}{t+C_bd}\frac{1}{K}+2\frac{C_a^2d}{(t+C_bd)^2}\frac{\lambda_{\max}^2}{\lambda_{\min}^2}\right)\mathbb{E}\left\{\|\boldsymbol{\beta}_i^{(t)}-\boldsymbol{\beta}_i^*\|^2\right\}$$
$$+2\frac{C_a^2d}{(t+C_bd)^2}\frac{\lambda_{\max}}{\lambda_{\min}^2}\sigma^2.$$

类似地，我们有 $\mathbb{E}\left\{\|\boldsymbol{\beta}_i^{(t+1)}-\boldsymbol{\beta}_i^*\|^2\right\}\leq\frac{Cd}{t+1+C_bd}\frac{\sigma^2}{\lambda_{\min}}$。

当 $\pi^*>0$ 时，对于所有 $i\in[K]$，我们有

$$\mathbb{E}\left\{\|\boldsymbol{\beta}_i^{(t+1)}-\boldsymbol{\beta}_i^*\|^2\right\}\leq\left(1-2\frac{C_a}{t+C_bd}\frac{\pi_t}{K}+2\frac{C_a^2d}{(t+C_bd)^2}\frac{\lambda_{\max}^2}{\lambda_{\min}^2}\right)\mathbb{E}\left\{\|\boldsymbol{\beta}_i^{(t)}-\boldsymbol{\beta}_i^*\|^2\right\}$$
$$+2\frac{C_a^2d}{(t+C_bd)^2}\frac{\lambda_{\max}}{\lambda_{\min}^2}\sigma^2.$$

在这种情况下，$\lim_{t\to+\infty}\pi_t=\pi^*$，并且存在 $N>0$，使得对于所有 $t\geq N$，我们有 $\pi_t>\frac{1}{2}\pi^*$。然后，通过累加上述不等式，我们有

$$\begin{aligned}\mathbb{E}\left\{\|\boldsymbol{\beta}_i^{(t+1)}-\boldsymbol{\beta}_i^*\|^2\right\}&\leq\prod_{l=N}^{t}\left(1-\frac{C_a}{l+C_bd}\frac{\pi^*}{K}\right)\mathbb{E}\left\{\|\boldsymbol{\beta}_i^{(N)}-\boldsymbol{\beta}_i^*\|^2\right\}\\&\quad+2\sum_{k=N}^{t}\prod_{l=k}^{t}\left(1-\frac{C_a}{l+C_bd}\frac{\pi^*}{K}\right)\frac{C_a^2d}{(k+C_bd)^2}\frac{\lambda_{\max}}{\lambda_{\min}^2}\sigma^2\\&\leq\left(\frac{N+1+C_bd}{t+1+C_bd}\right)^{C_a\pi^*/K}\|\boldsymbol{\beta}_i^{(N)}-\boldsymbol{\beta}_i^*\|^2+\frac{Cd}{t+1+C_bd}\frac{\sigma^2}{\lambda_{\min}},\end{aligned}$$

其中最后一行使用了引理17 和 16。

□

命题 5. 在与命题2相同的假设和条件下，我们有

(1) 若 $\pi^*=0$，则在事件 $\cup_{t=t_1}\{\text{最大 }\|\boldsymbol{\beta}_i^{(t)}-\boldsymbol{\beta}_i^*\|\leq\frac{h}{2}\}$ 下，对 $i\in\mathcal{A}$，以下结论成立，

$$\mathbb{E}\|\boldsymbol{\beta}_i^{(t)}-\boldsymbol{\beta}_i^*\|^4\leq C\left(\frac{d}{t+C_bd}\right)^2\frac{\sigma^4}{\lambda_{\min}^2},$$

(2) 若 $\pi^*>0$ 且 $C_a\geq 4K/\pi^*$，则对所有 $i\in[K]$，以下结论成立，

$$\mathbb{E}\|\boldsymbol{\beta}_i^{(t+1)}-\boldsymbol{\beta}_i^*\|^4\leq\left(\frac{N+1+C_bd}{t+1+C_bd}\right)^{2C_a\pi^*/K}\|\boldsymbol{\beta}_N-\boldsymbol{\beta}^*\|^4+C\left(\frac{d}{t+1+C_bd}\right)^2\frac{\sigma^4}{\lambda_{\min}^2},$$

其中 N 是某个整数，使得对所有 $t\geq N$, $\pi_t\geq\pi^*/2$均成立。

为了证明命题 5，我们需要根据 $\boldsymbol{\beta}_i^{(t+1)}$ 的更新展开四阶矩。命题 5 的证明与命题 4类似，因此省略。

证明 命题 2的证明。 在命题 2的条件下，命题 4 证明

$$\|\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i^*\|^2 = O_{\mathbb{P}}\left(\frac{1}{t}\right).$$

回顾一下，更新为：

$$\begin{aligned}\boldsymbol{\beta}_i^{(t+1)} - \boldsymbol{\beta}_i^* &= \boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i^* - \eta_t \cdot \mathbb{I}\{a_t = i\} \cdot (\mathbf{X}_t^\top \boldsymbol{\beta}_i^{(t)} - Y_t)\mathbf{X}_t \\ &= \left(\mathbf{I} - \eta_t \cdot \mathbb{I}\{a_t = i\} \cdot \mathbf{X}_t\mathbf{X}_t^\top\right)\left(\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i^*\right) + \eta_t \cdot \mathbb{I}\{a_t = i\} \cdot \xi_t \cdot \mathbf{X}_t,\end{aligned}$$

可将其分解为：

$$\begin{aligned}\boldsymbol{\beta}_i^{(t+1)} - \boldsymbol{\beta}_i^* = &\left(\mathbf{I} - \eta_t \cdot \mathbb{E}\left\{\mathbb{I}\{a_t = i\} \cdot \mathbf{X}_t\mathbf{X}_t^\top \middle| \mathcal{F}_t\right\}\right)\left(\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i^*\right) \\ &- \eta_t \cdot \left(\mathbb{I}\{a_t = i\} \cdot \mathbf{X}_t\mathbf{X}_t^\top - \mathbb{E}\left\{\mathbb{I}\{a_t = i\} \cdot \mathbf{X}_t\mathbf{X}_t^\top | \mathcal{F}_t\right\}\right)\left(\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i^*\right) \\ &+ \eta_t \cdot \mathbf{X}_t \cdot \mathbb{I}\{a_t = i\} \cdot \xi_t.\end{aligned}$$

使用上述定义的符号，更新可进一步写为：

$$\begin{aligned}\boldsymbol{\beta}_i^{(t+1)} - \boldsymbol{\beta}_i^* = &\left(\mathbf{I} - \eta_t \cdot \boldsymbol{\Sigma}_{i,t}^*\right)\left(\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i^*\right) + \eta_t \cdot \mathbb{I}\{a_t = i\} \cdot \xi_t \cdot \mathbf{X}_t \\ &- \eta_t\left(\boldsymbol{\Sigma}_{i,t} - \boldsymbol{\Sigma}_{i,t}^*\right)\left(\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i^*\right) - \eta_t \cdot \left(\mathbb{I}\{a_t = i\} \cdot \mathbf{X}_t\mathbf{X}_t^\top - \boldsymbol{\Sigma}_{i,t}\right)\left(\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i^*\right).\end{aligned}$$

我们从 $\boldsymbol{\beta}_i^{(0)}$ 开始累积上述更新，得到

$$\begin{aligned}\boldsymbol{\beta}_i^{(t+1)} - \boldsymbol{\beta}_i^* = &\underbrace{\prod_{s=1}^{t}\left(\mathbf{I} - \eta_s \cdot \boldsymbol{\Sigma}_{i,s}^*\right)\left(\boldsymbol{\beta}_i^{(0)} - \boldsymbol{\beta}_i^*\right)}_{\mathbf{B}_1} + \underbrace{\sum_{s=0}^{t}\prod_{l=s+1}^{t}\left(\mathbf{I} - \eta_l \cdot \boldsymbol{\Sigma}_{i,l}^*\right)\eta_s \cdot \xi_s \cdot \mathbf{X}_s \cdot \mathbb{I}\{a_s = i\}}_{\mathbf{B}_2} \\ &\underbrace{- \sum_{s=0}^{t}\prod_{l=s+1}^{t}\left(\mathbf{I} - \eta_l \cdot \boldsymbol{\Sigma}_{i,l}^*\right)\eta_s \cdot \left(\boldsymbol{\Sigma}_{i,s} - \boldsymbol{\Sigma}_{i,s}^*\right)\left(\boldsymbol{\beta}_i^{(s)} - \boldsymbol{\beta}_i^*\right)}_{\mathbf{B}_3} \\ &\underbrace{- \sum_{s=0}^{t}\prod_{l=s+1}^{t}\left(\mathbf{I} - \eta_l \cdot \boldsymbol{\Sigma}_{i,l}^*\right)\eta_s \cdot \left(\mathbb{I}\{a_s = i\} \cdot \mathbf{X}_s\mathbf{X}_s^\top - \boldsymbol{\Sigma}_{i,s}\right)\left(\boldsymbol{\beta}_i^{(s)} - \boldsymbol{\beta}_i^*\right)}_{\mathbf{B}_4}\end{aligned}$$

值得注意的是，对于 $s = 0$、1、2、.....、t，矩阵 $\boldsymbol{\Sigma}_{i,s}^*$可能没有共同的特征向量。接下来，我们分别分析 $i \in \mathcal{A}$ 对应的各项 $\mathbf{B}_1$、.....、$\mathbf{B}_4$。

$\mathbf{B}_1$ 的界。首先考虑这些矩阵的乘积,

$$0 \preceq \prod_{s=1}^{t} \left(\mathbf{I} - \eta_s \cdot \mathbf{\Sigma}_{i,s}^*\right),$$

并且在事件 $\{\mathcal{E}_t\}$ 下，我们有

$$\begin{aligned}
\left\| \prod_{s=1}^{t} \left(\mathbf{I} - \eta_s \cdot \mathbf{\Sigma}_{i,s}^*\right) \right\| &\leq \exp\left(\sum_{s=1}^{t} \log\left(\left\| \mathbf{I} - \eta_s \cdot \mathbf{\Sigma}_{i,s}^* \right\| \right) \right) \leq \exp\left(- \sum_{s=1}^{t} \eta_s \cdot \lambda_{\min}(\mathbf{\Sigma}_{i,s}^*) \right) \\
&\leq \exp\left(- \sum_{s=1}^{t_1} \frac{C_a}{s + C_b d} \cdot \frac{\pi_s}{2K} - \sum_{s=t_1+1}^{t} \frac{C_a}{s + C_b d} \frac{1}{K} \right) \\
&\leq \exp\left(- \frac{C_a}{4K} \ln\left(\frac{t_1 + C_b d}{C_b d} \right) - \frac{C_a}{K} \ln\left(\frac{t + C_b d}{t_1 + C_b d} \right) \right),
\end{aligned}$$

其中，第一行使用了方程 23 ，最后一行是该序列的积分界。上述方程给出了 $\|\mathbf{B}_1\|$ 的一个上界,

$$\|\mathbf{B}_1\| \leq \left(\frac{C_b d}{t + C_b d} \right)^{C_a / 4K} \|\boldsymbol{\beta}_i^{(0)} - \boldsymbol{\beta}_i^*\|,$$

这表明，由于 $C_a \geq 4K$, $\mathbf{B}_1 = O\left(\frac{1}{t}\right)$。

对 B_2 的分析。注意，B_2 是 $t+1$ 鞅差之和。我们将证明$\sqrt{t}\mathrm{B}_2$ 在分布上收敛到一个多元高斯分布。为此，我们使用 Cramér-Wold定理（Cramér 和 Wold，1936）和鞅中心极限定理（Brown，1971；Durrett，2019）。对于任意 d 维向量 $\mathrm{v} = (v_1, \ldots, v_d)$，我们将证明 $\sqrt{t}\mathrm{v}^\top \mathrm{B}_2$ 渐近服从高斯分布。命题5 验证了 $\mathrm{v}^\top \mathrm{B}_2$ 满足 Lindeberg条件。我们只需证明条件方差的收敛。具体而言，由于 $\xi_s \perp \mathrm{X}_s | \mathcal{F}_s$，我们有

$$\begin{aligned}
&\mathbb{E}\left\{ \left(\mathbf{v}^\top \prod_{l=s+1}^{t} \left(\mathbf{I} - \eta_l \cdot \mathbf{\Sigma}_{i,l}^*\right) \eta_s \cdot \xi_s \cdot \mathbf{X}_s \cdot \mathbb{I}\{a_s = i\} \right)^2 \middle| \mathcal{F}_s \right\} \\
&\qquad = \eta_s^2 \sigma_s^2 \mathbf{v}^\top \prod_{l=s+1}^{t} \left(\mathbf{I} - \eta_l \cdot \mathbf{\Sigma}_{i,l}^*\right) \mathbf{\Sigma}_{i,s} \prod_{l=s+1}^{t} \left(\mathbf{I} - \eta_l \cdot \mathbf{\Sigma}_{i,l}^*\right) \mathbf{v},
\end{aligned}$$

其中 $\sigma_s^2 := \mathbb{E}\{\xi_s^2 | \mathcal{F}_s\}$. 接下来，我们考察条件方差之和，并将证明其依概率收敛

$$\begin{aligned}
V_{\mathbf{B}2} := \sum_{s=0}^{t} &\mathbb{E}\left\{ \left(\mathbf{v}^\top \prod_{l=s+1}^{t} \left(\mathbf{I} - \eta_l \cdot \mathbf{\Sigma}_{i,l}^*\right) \eta_s \cdot \xi_s \cdot \mathbf{X}_s \cdot \mathbb{I}\{a_s = i\} \right)^2 \middle| \mathcal{F}_s \right\} \\
&= \sum_{s=0}^{t} \eta_s^2 \sigma_s^2 \mathbf{v}^\top \prod_{l=s+1}^{t} \left(\mathbf{I} - \eta_l \cdot \mathbf{\Sigma}_{i,l}^*\right) \mathbf{\Sigma}_{i,s} \prod_{l=s+1}^{t} \left(\mathbf{I} - \eta_l \cdot \mathbf{\Sigma}_{i,l}^*\right) \mathbf{v}.
\end{aligned}$$

我们将方差 $V_{\mathbf{B}_2}$ 分解为两项，

$$V_{\mathbf{B}2} = \sum_{s=0}^{t} \eta_s^2 \sigma_s^2 \mathbf{v}^\top \prod_{l=s+1}^{t} \left(\mathbf{I} - \eta_l \cdot \mathbf{\Sigma}_{i,l}^*\right) \mathbf{\Sigma}_{i,s}^* \prod_{l=s+1}^{t} \left(\mathbf{I} - \eta_l \cdot \mathbf{\Sigma}_{i,l}^*\right) \mathbf{v} \\ + \sum_{s=0}^{t} \eta_s^2 \sigma_s^2 \mathbf{v}^\top \prod_{l=s+1}^{t} \left(\mathbf{I} - \eta_l \cdot \mathbf{\Sigma}_{i,l}^*\right) \left(\mathbf{\Sigma}_{i,s} - \mathbf{\Sigma}_{i,s}^*\right) \prod_{l=s+1}^{t} \left(\mathbf{I} - \eta_l \cdot \mathbf{\Sigma}_{i,l}^*\right) \mathbf{v}.$$

我们首先证明上述方程中的第二项趋于 $o(1/t)$。具体而言，我们有以下上界：

$$\left| \sum_{s=0}^{t} \eta_s^2 \sigma_s^2 \mathbf{v}^\top \prod_{l=s+1}^{t} \left(\mathbf{I} - \eta_l \cdot \mathbf{\Sigma}_{i,l}^*\right) \left(\mathbf{\Sigma}_{i,s} - \mathbf{\Sigma}_{i,s}^*\right) \prod_{l=s+1}^{t} \left(\mathbf{I} - \eta_l \cdot \mathbf{\Sigma}_{i,l}^*\right) \mathbf{v} \right| \\ \leq \sigma^2 \|\mathbf{v}\|^2 \sum_{s=0}^{t} \left(\frac{C_a}{s + C_b d}\right)^2 \left(\frac{s + C_b d}{t + C_b d}\right)^{2C_a/3K} \left\|\mathbf{\Sigma}_{i,s} - \mathbf{\Sigma}_{i,s}^*\right\|. \quad (26)$$

我们首先使用命题 $\mathbb{E}\left\|\mathbf{\Sigma}_{i\text{对其进行界定，}s} - \mathbf{\Sigma}_{i,\ s}^*\right\|$ 即4，

$$\begin{aligned} &\mathbb{E}\left\|\mathbf{\Sigma}_{i,s} - \mathbf{\Sigma}_{i,s}^*\right\| \\ &\leq \mathbb{E}\left\|\mathbf{\Sigma}_{i,s} - \mathbf{\Sigma}_{i,s}^*\right\| \cdot \mathbb{I}\left\{\max_j \|\boldsymbol{\beta}_j^{(s)} - \boldsymbol{\beta}_j^*\| \leq h\right\} + \mathbb{E}\left\|\mathbf{\Sigma}_{i,s} - \mathbf{\Sigma}_{i,s}^*\right\| \cdot \mathbb{I}\left\{\max_j \|\boldsymbol{\beta}_j^{(s)} - \boldsymbol{\beta}_j^*\| \geq h\right\} \\ &\leq \kappa_0 \mathbb{E} \max_{j \in \mathcal{A}} \|\boldsymbol{\beta}_j^{(s)} - \boldsymbol{\beta}_i^*\| + \lambda_{\max} \frac{\mathbb{E} \max_{j \in [K]} \|\boldsymbol{\beta}_j^{(s)} - \boldsymbol{\beta}_j^*\|}{h} \leq \frac{C}{\sqrt{s}} \end{aligned}$$

因此，当 $t \to +\infty$时，式 26 右侧的期望为 $O\left(t^{-3/2}\right)$。由此，我们完成了对 $V_{\mathbf{B}_2}$ 的第二项为 $O\left(t^{-3/2}\right)$ 的证明。随后，我们分析 $V_{\mathbf{B}_2}$ 的第一项，并证明其在 $\lim_{t \to +\infty} \pi_t = \pi^* \in [0, 1]$ 下收敛，以及在 $\lim_{t \to +\infty} \sigma_t^2 = \sigma_*^2$的依概率收敛下收敛。

$$\left| \eta_s^2 \sigma_s^2 \mathbf{v}^\top \prod_{l=s+1}^{t} \left(\mathbf{I} - \eta_l \cdot \mathbf{\Sigma}_{i,l}^*\right) \mathbf{\Sigma}_{i,s}^* \prod_{l=s+1}^{t} \left(\mathbf{I} - \eta_l \cdot \mathbf{\Sigma}_{i,l}^*\right) \mathbf{v} \right. \\ \left. - \eta_s^2 \sigma_*^2 \mathbf{v}^\top \prod_{l=s+1}^{t} \left(\mathbf{I} - \eta_l \cdot \mathbf{\Sigma}_i(\pi^*)\right) \mathbf{\Sigma}_i(\pi^*) \prod_{l=s+1}^{t} \left(\mathbf{I} - \eta_l \cdot \mathbf{\Sigma}_i(\pi^*)\right) \mathbf{v} \right| \\ \leq \eta_s^2 \|\mathbf{v}\|^2 |\sigma_s^2 - \sigma_*^2| \lambda_{\max} \left(\frac{s + C_b d}{t + C_b d}\right)^{2C_a/3K} + 2(t-s) \eta_s^3 \|\mathbf{v}\|^2 \lambda_{\max} \left(\frac{s + C_b d}{t + C_b d}\right)^{2C_a/3K} \left\|\mathbf{\Sigma}_i(\pi^*) - \mathbf{\Sigma}_{i,s}^*\right\| \\ + \eta_s^2 \sigma^2 \|\mathbf{v}\|^2 \lambda_{\max} \left(\frac{s + C_b d}{t + C_b d}\right)^{2C_a/3K} \left\|\mathbf{\Sigma}_i(\pi^*) - \mathbf{\Sigma}_{i,s}^*\right\|.$$

对于任意 $\varepsilon > 0$，假设6 保证存在 t_ε 使得对所有 $s \geq t_\varepsilon$，

$$\mathbb{E}|\sigma_s^2 - \sigma_*^2| \leq \varepsilon, \quad \mathbb{E}\left\|\mathbf{\Sigma}_i(\pi^*) - \mathbf{\Sigma}_{i,s}^*\right\| \leq \varepsilon.$$

因此，将上述方程相加并取期望，对于所有充分大的 t，我们有

$$\mathbb{E}\left|\sum_{s=0}^{t}\eta_s^2\sigma_s^2\mathbf{v}^\top\prod_{l=s+1}^{t}\left(\mathbf{I}-\eta_l\cdot\mathbf{\Sigma}_{i,l}^*\right)\mathbf{\Sigma}_{i,s}^*\prod_{l=s+1}^{t}\left(\mathbf{I}-\eta_l\cdot\mathbf{\Sigma}_{i,l}^*\right)\mathbf{v}\right.$$
$$\left.-\sum_{s=0}^{t}\eta_s^2\sigma_*^2\mathbf{v}^\top\prod_{l=s+1}^{t}\left(\mathbf{I}-\eta_l\cdot\mathbf{\Sigma}_i(\pi^*)\right)\mathbf{\Sigma}_i(\pi^*)\prod_{l=s+1}^{t}\left(\mathbf{I}-\eta_l\cdot\mathbf{\Sigma}_i(\pi^*)\right)\mathbf{v}\right|\leq C\frac{\varepsilon}{t},$$

这意味着

$$\lim_{t\to+\infty}t\mathbb{E}\left|\sum_{s=0}^{t}\eta_s^2\sigma_s^2\mathbf{v}^\top\prod_{l=s+1}^{t}\left(\mathbf{I}-\eta_l\cdot\mathbf{\Sigma}_{i,l}^*\right)\mathbf{\Sigma}_{i,s}^*\prod_{l=s+1}^{t}\left(\mathbf{I}-\eta_l\cdot\mathbf{\Sigma}_{i,l}^*\right)\mathbf{v}\right.$$
$$\left.-\sum_{s=0}^{t}\eta_s^2\sigma_*^2\mathbf{v}^\top\prod_{l=s+1}^{t}\left(\mathbf{I}-\eta_l\cdot\mathbf{\Sigma}_i(\pi^*)\right)\mathbf{\Sigma}_i(\pi^*)\prod_{l=s+1}^{t}\left(\mathbf{I}-\eta_l\cdot\mathbf{\Sigma}_i(\pi^*)\right)\mathbf{v}\right|=0.$$

因此，我们依概率收敛，

$$\lim_{t\to+\infty}t\sum_{s=0}^{t}\eta_s^2\sigma_s^2\mathbf{v}^\top\prod_{l=s+1}^{t}\left(\mathbf{I}-\eta_l\cdot\mathbf{\Sigma}_{i,l}^*\right)\mathbf{\Sigma}_{i,s}^*\prod_{l=s+1}^{t}\left(\mathbf{I}-\eta_l\cdot\mathbf{\Sigma}_{i,l}^*\right)\mathbf{v}$$
$$=\lim_{t\to+\infty}t\sum_{s=0}^{t}\eta_s^2\sigma_*^2\mathbf{v}^\top\prod_{l=s+1}^{t}\left(\mathbf{I}-\eta_l\cdot\mathbf{\Sigma}_i(\pi^*)\right)\mathbf{\Sigma}_i(\pi^*)\prod_{l=s+1}^{t}\left(\mathbf{I}-\eta_l\cdot\mathbf{\Sigma}_i(\pi^*)\right)\mathbf{v}.\quad(27)$$

因此，为了得到 $\lim_{t\to+\infty}tV_{\mathbf{B}_2}$，我们只需研究公式 27 右侧。注意，$\mathrm{I}-\eta_l\cdot\mathbf{\Sigma}_i(\pi^*)$ 和 $\mathbf{\Sigma}_i(\pi^*)$ 具有相同的特征向量，因此可以同时对角化。回忆一下，$\mathbf{\Sigma}_i(\pi^*)$ 的奇异值分解为 $\mathbf{\Sigma}_i(\pi^*)=\mathrm{U}_i(\pi^*)\mathbf{\Lambda}_i(\pi^*)\mathrm{U}_i(\pi^*)^\top$，其中 $\mathrm{U}_i(\pi^*)$ 是一个正交矩阵，而 $\mathbf{\Lambda}_i(\pi^*)=\mathrm{diag}(\lambda_1(\pi^*)\ldots,\lambda_d(\pi^*))$ 是一个对角矩阵。于是，公式 27 右侧的矩阵部分可改写为

$$\mathbf{U}(\pi^*)^\top\sum_{s=0}^{t}\eta_s^2\prod_{l=s+1}^{t}\left(\mathbf{I}-\eta_l\cdot\mathbf{\Sigma}_i(\pi^*)\right)\mathbf{\Sigma}_i(\pi^*)\prod_{l=s+1}^{t}\left(\mathbf{I}-\eta_l\cdot\mathbf{\Sigma}_i(\pi^*)\right)\mathbf{U}(\pi^*)$$
$$=\begin{pmatrix}\sum_{s=0}^{t}\eta_s^2\prod_{l=s+1}^{t}(1-\eta_l\lambda_1(\pi^*))^2\lambda_1(\pi^*) & & \\ & \ddots & \\ & & \sum_{s=0}^{t}\eta_s^2\prod_{l=s+1}^{t}(1-\eta_l\lambda_d(\pi^*))^2\lambda_d(\pi^*)\end{pmatrix}.$$

等价地，我们只需分析对角元素的收敛。为便于表述，我们记 λ_j 表示 $\lambda_j(\pi^*)$。首先，我们对函数 $\log(\lambda_j)$ 进行泰勒展开

函数 $\log(1-x)$ 的结果，并得出

$$\begin{aligned}\sum_{s=0}^{t}\eta_s^2\prod_{l=s+1}^{t}(1-\eta_l\lambda_j)^2\lambda_j &= \sum_{s=0}^{t}\eta_s^2\exp\left(2\sum_{l=s+1}^{t}\log\left(1-\eta_l\lambda_j\right)\right)\lambda_j \\ &= \sum_{s=0}^{t}\eta_s^2\exp\left(-2\sum_{l=s+1}^{t}\eta_l\lambda_j+O\left(\sum_{l=s+1}^{t}\eta_l^2\lambda_j^2\right)\right)\lambda_j.\end{aligned}$$

接下来，我们将步长的取值代入方程。具体而言，求和项为

$$\begin{aligned}\sum_{l=s+1}^{t}\eta_l\lambda_j &= \frac{C_a\lambda_j}{\lambda_{\min}}\sum_{l=s+1}^{t}\frac{1}{l+C_bd} = \frac{C_a\lambda_j}{\lambda_{\min}}\sum_{l=s+1}^{t}\frac{1}{l+C_bd} \\ &= \frac{C_a\lambda_j}{\lambda_{\min}}\int_{s+C_bd+1}^{t+C_bd+1}\frac{1}{x}\mathrm{d}x+\frac{C_a\lambda_j}{\lambda_{\min}}O\left(\sum_{l=s+1}^{t}\left(\frac{1}{l+C_bd}\right)^2\right) \\ &= \frac{C_a\lambda_j}{\lambda_{\min}}\log\left(\frac{t+C_bd+1}{s+C_bd+1}\right)+\frac{C_a\lambda_j}{\lambda_{\min}}O\left(\frac{1}{s}\right),\end{aligned}$$

其中，第二行使用了 $f(\tau)=\int_a^\tau x^{-1}\mathrm{d}x$ 的泰勒展开。此外，同样地，平方项可以界定为：

$$O\left(\sum_{l=s+1}^{t}\eta_l^2\lambda_j^2\right)=O\left(\frac{1}{s}\right).$$

结合上述两个方程，可得

$$\begin{aligned}\sum_{s=0}^{t}\eta_s^2\prod_{l=s+1}^{t}(1-\eta_l\lambda_j)^2\lambda_j &= \sum_{s=0}^{t}\eta_s^2\exp\left(-2\frac{C_a\lambda_j}{\lambda_{\min}}\log\left(\frac{t+C_bd+1}{s+C_bd+1}\right)+O\left(\frac{1}{s}\right)\right) \\ &= \sum_{s=0}^{t}\eta_s^2\left(\frac{s+C_bd+1}{t+C_bd+1}\right)^{2C_a\lambda_j/\lambda_{\min}}\exp\left(O\left(\frac{1}{s}\right)\right) \\ &= \sum_{s=0}^{t}\eta_s^2\left(\frac{s+C_bd+1}{t+C_bd+1}\right)^{2C_a\lambda_j/\lambda_{\min}}\left(1+O\left(\frac{1}{s}\right)\right),\end{aligned}$$

其中最后一个等式是将 $\exp(x)$ 在 $x=0$ 处进行泰勒展开得到的。接下来，我们分别界定这两项。第一项可以通过对函数$\int_a^\tau x^\alpha\mathrm{d}x$进行泰勒展开来分析，

$$\begin{aligned}\sum_{s=0}^{t}\eta_s^2\left(\frac{s+C_bd+1}{t+C_bd+1}\right)^{2C_a\lambda_j/\lambda_{\min}} &= \frac{C_a^2}{\lambda_{\min}^2}\left(\frac{1}{t+C_bd+1}\right)^{2C_a\lambda_j/\lambda_{\min}}\sum_{s=0}^{t}(s+C_bd+1)^{2C_a\lambda_j/\lambda_{\min}-2} \\ &= \frac{C_a^2}{\lambda_{\min}^2}\left(\frac{1}{t+C_bd+1}\right)^{2C_a\lambda_j/\lambda_{\min}}\left(\int_{C_bd+1}^{t+C_bd+1}x^{2C_a\lambda_j/\lambda_{\min}-2}\mathrm{d}x+O\left(t^{2C_a\lambda_j/\lambda_{\min}-2}\right)\right) \\ &= \frac{C_a^2}{\lambda_{\min}^2}\frac{1}{2C_a\lambda_j/\lambda_{\min}-1}\frac{1}{t+C_bd+1}+O\left(\left(\frac{1}{t+C_bd+1}\right)^2\right).\end{aligned}$$

以类似的方式，我们可以界定 $O(1/s)$ 项，其量级为 $O(1/t^2)$。综上，我们得到方程 27 和 $tV_{\mathbf{B}_2}$,

$$\lim_{t\to+\infty} tV_{\mathbf{B}_2} = \lim_{t\to+\infty} t\sigma_*^2 \sum_{s=0}^{t} \eta_s^2 \mathbf{v}^\top \prod_{l=s+1}^{t} \left(\mathbf{I} - \eta_l \cdot \boldsymbol{\Sigma}_i(\pi^*)\right) \boldsymbol{\Sigma}_i(\pi^*) \prod_{l=s+1}^{t} \left(\mathbf{I} - \eta_l \cdot \boldsymbol{\Sigma}_i(\pi^*)\right) \mathbf{v}$$
$$= v^\top \mathbf{U} \begin{pmatrix} \frac{C_a\sigma_*^2}{\lambda_{\min}} \frac{C_a\lambda_1/\lambda_{\min}}{2C_a\lambda_1/\lambda_{\min}-1} & & \\ & \ddots & \\ & & \frac{C_a\sigma_*^2}{\lambda_{\min}} \frac{C_a\lambda_d/\lambda_{\min}}{2C_a\lambda_d/\lambda_{\min}-1} \end{pmatrix} \mathbf{U}^\top \mathbf{v}.$$

因此，根据鞅中心极限定理（Brown，1971年）和 Cramér-Wold 定理（Cramér和 Wold，1936年），我们得出结论，项 $\sqrt{t}\mathbf{B}_2$ 依分布收敛于均值为零、协方差由命题 2定义的多维高斯分布。

B的界$_3$。我们仍然先考虑矩阵乘积。对于 $t > s > t_1$，采用与第 C节类似的技巧，我们有

$$\left\| \prod_{l=s+1}^{t} \left(\mathbf{I} - \eta_l \cdot \boldsymbol{\Sigma}_{i,l}^*\right) \right\| \leq \exp\left(-C_a \ln\left(\frac{t + C_b d}{s + C_b d}\right)\right) = \left(\frac{s + C_b d}{t + C_b d}\right)^{C_a}.$$

另一方面，对于 $t > t_1 > s$ 的技巧，矩阵乘积为

$$\left\| \prod_{l=s+1}^{t} \left(\mathbf{I} - \eta_l \cdot \boldsymbol{\Sigma}_{i,l}^*\right) \right\| \leq \exp\left(-\frac{C_a}{K} \ln\left(\frac{t_1 + C_b d}{s + C_b d}\right) - C_a \ln\left(\frac{t + C_b d}{t_1 + C_b d}\right)\right)$$
$$\leq \left(\frac{s + C_b d}{t_1 + C_b d}\right)^{C_a/K} \cdot \left(\frac{t_1 + C_b d}{t + C_b d}\right)^{C_a} \cdot \mathbf{I}_d.$$

然后我们可以对 $\|\mathbf{B}_3\|$ 的范数进行界定，

$$\|\mathbf{B}_3\| \leq \sum_{s=0}^{t} \frac{(s + C_b d)^{C_a/K-1}}{(t + C_b d)^{C_a/K}} \|\boldsymbol{\Sigma}_{i,s} - \boldsymbol{\Sigma}_{i,s}^*\| \cdot \|\boldsymbol{\beta}_i^{(s)} - \boldsymbol{\beta}_i^*\|.$$

此外，方程 25 为 $\|\boldsymbol{\Sigma}_{i,\,s} - \boldsymbol{\Sigma}_{i,\,s}^*\|$提供了界，并且在命题 4的假设下，我们还可以得到

$$\mathbb{E}\left\{\|\mathbf{B}_3\|\right\} \leq \sum_{s=0}^{t_1} \frac{(s + C_b d)^{C_a/K-1}}{(t + C_b d)^{C_a/K}} \cdot \lambda_{\max} \cdot \mathbb{E}\left\{\|\boldsymbol{\beta}_i^{(t)} - \boldsymbol{\beta}_i^*\|\right\}$$
$$+ \sum_{s=t_1+1}^{t} \frac{(s + C_b d)^{C_a/K-1}}{(t + C_b d)^{C_a/K}} \cdot \kappa_0 \cdot \mathbb{E}\left\{\|\boldsymbol{\beta}_i^{(s)} - \boldsymbol{\beta}_i^*\|^2 \cdot \mathbb{I}\left\{\|\boldsymbol{\beta}_i^{(s)} - \boldsymbol{\beta}_i^*\| \leq h\right\}\right\}$$
$$+ \sum_{s=t_1+1}^{t} \frac{(s + C_b d)^{C_a/K-1}}{(t + C_b d)^{C_a/K}} \cdot \lambda_{\max} \cdot \mathbb{E}\left\{\|\boldsymbol{\beta}_i^{(s)} - \boldsymbol{\beta}_i^*\| \cdot \mathbb{I}\left\{\|\boldsymbol{\beta}_i^{(s)} - \boldsymbol{\beta}_i^*\| \geq h\right\}\right\}$$

这意味着 $\|\mathbf{B}_3\| \leq O_{\mathbb{P}}(1/t)$。

$\mathbf{B}_4$的界。值得注意的是，$\mathbf{B}_4$ 是鞅差之和。我们考虑 $\mathbb{E}\|\mathbf{B}_4\|^2$,

$$\mathbb{E}\left\{\|\mathbf{B}_4\|^2\right\} = \sum_{s=0}^{t}\prod_{l=s+1}^{t}\left\|\mathbf{I}-\eta_l\cdot\boldsymbol{\Sigma}_{i,l}^*\right\|^2\cdot\eta_s^2\cdot\mathbb{E}\left\|\left(\mathbb{I}\{a_s=i\}\cdot\mathbf{X}_s\mathbf{X}_s^\top-\boldsymbol{\Sigma}_{i,s}\right)\left(\boldsymbol{\beta}_i^{(s)}-\boldsymbol{\beta}_i^*\right)\right\|^2$$
$$\leq \sum_{s=0}^{t}\prod_{l=s+1}^{t}\left\|\mathbf{I}-\eta_l\cdot\boldsymbol{\Sigma}_{i,l}^*\right\|^2\cdot\eta_s^2\cdot\lambda_{\max}^2\mathbb{E}\left\|\boldsymbol{\beta}_i^{(s)}-\boldsymbol{\beta}_i^*\right\|^2.$$

因此，根据命题 4、引理 17 和引理 16，我们有

$$\|\mathbf{B}_4\| = O_{\mathbb{P}}\left(\frac{1}{t}\right).$$

因此，总而言之，在命题 2的条件下，$\|\mathbf{B}_i\| = O_{\mathbb{P}}(1/t)$ 对于 $i=1$，3、4 以及 $\sqrt{t}\mathbf{B}_2$依分布收敛到多元高斯分布。

□

C.6 引理6的证明

首先，我们有

$$\mathbb{P}\left(\left\|\frac{1}{t}\sum_{l=1}^{t}\left\{\mathbf{X}_l\mathbf{X}_l^\top-\mathbb{E}\left\{\mathbf{X}_l\mathbf{X}_l^\top|\mathcal{F}_l\right\}\right\}\right\|_F^2\geq\varepsilon\right)\leq\frac{1}{t^2\varepsilon}\sum_{l=1}^{t}\mathbb{E}\left\|\mathbf{X}_l\mathbf{X}_l^\top-\mathbb{E}\left\{\mathbf{X}_l\mathbf{X}_l^\top|\mathcal{F}_l\right\}\right\|_F^2,$$

这意味着 $\frac{1}{t}\sum_{l=1}^{t}\mathbf{X}_l\mathbf{X}_l^\top-\frac{1}{t}\sum_{l=1}^{t}\mathbb{E}\left\{\mathbf{X}_l\mathbf{X}_l^\top|\mathcal{F}_l\right\}$ 依概率收敛于零。另一方面，由假设 6，我们有 $\frac{1}{t}\sum_{l=1}^{t}\mathbb{E}\left\{\mathbf{X}_l\bar{\mathbf{X}}_l^\top|\mathcal{F}_l\right\}$ 依概率收敛到 $\bar{\boldsymbol{\Sigma}}^*$。另外两个论证可类似地证明。

D 技术引理

引理 9 (次指数鞅的伯恩斯坦不等式)。 设 $Z_1, Z_2, \ldots, Z_T$ 是关于 $\mathcal{F}_0, \ldots, \mathcal{F}_T$ 的实值鞅差序列。假设 $Z_t|\mathcal{F}_{t-1}$ 是次指数的，且其 *Orlicz*范数几乎处处有界于 $\|Z_t|\mathcal{F}_{t-1}\|_{\Psi_1}\leq K_t$，其中 K_t 是常数，则

$$\mathbb{P}\left(\sum_{t=1}^{T}Z_t\geq s\right)\leq\exp\left(-\min\left\{\frac{s^2}{C\sum_{t=1}^{T}K_t^2},\frac{s}{2\max_t K_t}\right\}\right).$$

证明。 根据马尔可夫不等式，有

$$\begin{aligned}\mathbb{P}\left(\sum_{t=1}^{T} Z_t \geq s\right) &\leq \exp(-\lambda s)\mathbb{E}\left\{\exp\left(\lambda \sum_{t=1}^{T} Z_t\right)\right\} \\ &= \exp(-\lambda s)\mathbb{E}\left\{\mathbb{E}\left\{\exp\left(\lambda \sum_{i=1}^{T} Z_t\right)\middle|\mathcal{F}_{T-1}\right\}\right\} \\ &= \exp(-\lambda s)\mathbb{E}\left\{\exp\left(\lambda \sum_{t=1}^{T-1} Z_t\right)\mathbb{E}\left\{\exp\left(\lambda Z_T\right)\middle|\mathcal{F}_{T-1}\right\}\right\}.\end{aligned}$$

对于 $\lambda \leq \frac{1}{\text{最大值}_t \|Z_t|\mathcal{F}_{t-1}\|_{\Psi_1}}$，条件期望满足

$$\mathbb{E}\left\{\exp\left(\lambda Z_T\right)|\mathcal{F}_{T-1}\right\} \leq \exp\left(C\lambda^2 \|Z_T|\mathcal{F}_{T-1}\|_{\Psi_1}^2\right).$$

分别对 $X_{T-1}, \ldots, X_1$ 应用该过程，然后得到

$$\mathbb{P}\left(\sum_{t=1}^{T} Z_t \geq s\right) \leq \exp\left(-\lambda s + C\lambda^2 \sum_{t=1}^{T} K_t^2\right).$$

然后代入 λ 的取值

$$\lambda = \min\left\{\frac{s}{2C\sum_{t=1}^{T} K_t^2}, \frac{c}{\max_t K_t}\right\}$$

到尾界中，得到

$$\mathbb{P}\left(\sum_{t=1}^{T} Z_t \geq s\right) \leq \exp\left(-\min\left\{\frac{s^2}{C\sum_{t=1}^{T} K_t^2}, \frac{s}{2\max_t K_t}\right\}\right).$$

□

引理10 (基于随机向量 Orlicz 范数的伯恩斯坦不等式**)**。 假设 $\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_T \in \mathbb{R}^d$ 是均值为零的向量，即， $\mathbb{E}[\mathbf{Z}_t] = 0$ 和 $\|\mathbf{Z}_t\|_{\Psi_1}$ 存在。那么，对于任意 $v > 0$，以下结论成立：

$$\mathbb{P}\left(\left\|\sum_{t=1}^{T} \mathbf{Z}_t\right\| \geq 2v\right) \leq \exp\left(Cd - \min\left\{\frac{v^2}{\sum_{t=1}^{T} \|\mathbf{Z}_t\|_{\Psi_1}^2}, \frac{v}{\max_t \|\mathbf{Z}_t\|_{\Psi_1}}\right\}\right),$$

其中 C 是某个常数。

证明。 我们注意到

$$\left\|\sum_{t=1}^{T} \mathbf{Z}_t\right\| = \sup_{\mathbf{u}\in\mathbb{S}^{d-1}} \langle \sum_{t=1}^{T} \mathbf{Z}_t, \mathbf{u}\rangle.$$

我们将使用 ε-网来界定欧几里得范数。此外，对于固定的 $\mathbf{u}$ 以及任意 $v>0$，我们有

$$\mathbb{P}\left(\left|\left\langle\sum_{t=1}^{T}\mathbf{Z}_t,\mathbf{u}\right\rangle\right|\geq v\right)\leq 2\exp\left(-\min\left\{\frac{v^2}{\sum_{t=1}^{T}\|\langle\mathbf{Z}_t,\mathbf{u}\rangle\|_{\Psi_1}^2},\frac{v}{\max_t\|\langle\mathbf{Z}_t,\mathbf{u}\rangle\|_{\Psi_1}}\right\}\right) \tag{28}$$

$$\leq 2\exp\left(-\min\left\{\frac{v^2}{\sum_{t=1}^{T}\|\mathbf{Z}_t\|_{\Psi_1}^2},\frac{v}{\max_t\|\mathbf{Z}_t\|_{\Psi_1}}\right\}\right), \tag{29}$$

其中使用了随机向量 Orlicz范数的定义。设 $\mathcal{N}$ 是 $1/2$的一个 $\mathbb{S}^{d-1}$ -网，其基数为 $|\mathcal{N}|\leq 5^d$，并对 29 在 $\mathcal{N}$上取并集，

$$\mathbb{P}\left(\max_{\mathbf{u}\in\mathcal{N}}\left|\left\langle\sum_{t=1}^{T}\mathbf{Z}_t,\mathbf{u}\right\rangle\right|\geq v\right)\leq 2|\mathcal{N}|\exp\left(-\min\left\{\frac{v^2}{\sum_{t=1}^{T}\|\mathbf{Z}_t\|_{\Psi_1}^2},\frac{v}{\max_t\|\mathbf{Z}_t\|_{\Psi_1}}\right\}\right).$$

对于任意$\mathbf{u}\in\mathbb{S}^{d-1}$，存在$\mathbf{x}\in\mathcal{N}$，使得 $\|\mathbf{u}-\mathbf{x}\|\leq 1/2$。因此，我们有

$$\begin{aligned}\left\|\sum_{t=1}^{T}\mathbf{Z}_t\right\|&=\sup_{\mathbf{u}\in\mathbb{S}^{d-1}}\left|\left\langle\sum_{t=1}^{T}\mathbf{Z}_t,\mathbf{u}\right\rangle\right|\leq\max_{\mathbf{u}\in\mathcal{N}}\left|\left\langle\sum_{t=1}^{T}\mathbf{Z}_t,\mathbf{u}\right\rangle\right|+\sup_{\mathbf{u}:\,\|\mathbf{u}\|\leq\varepsilon}\left|\left\langle\sum_{t=1}^{T}\mathbf{Z}_t,\mathbf{u}\right\rangle\right|\\&=\max_{\mathbf{u}\in\mathcal{N}}\left|\left\langle\sum_{t=1}^{T}\mathbf{Z}_t,\mathbf{u}\right\rangle\right|+\frac{1}{2}\left\|\sum_{t=1}^{T}\mathbf{Z}_t\right\|,\end{aligned}$$

这表明

$$\left\|\sum_{t=1}^{T}\mathbf{Z}_t\right\|=\sup_{\mathbf{u}\in\mathbb{S}^{d-1}}\left|\left\langle\sum_{t=1}^{T}\mathbf{Z}_t,\mathbf{u}\right\rangle\right|\leq 2\max_{\mathbf{u}\in\mathcal{N}}\left|\left\langle\sum_{t=1}^{T}\mathbf{Z}_t,\mathbf{u}\right\rangle\right|.$$

因此，我们有

$$\mathbb{P}\left(\left\|\sum_{t=1}^{T}\mathbf{Z}_t\right\|\geq 2v\right)\leq 2\times 5^d\exp\left(-\min\left\{\frac{v^2}{\sum_{t=1}^{T}\|\mathbf{Z}_t\|_{\Psi_1}^2},\frac{v}{\max_t\|\mathbf{Z}_t\|_{\Psi_1}}\right\}\right),$$

由此完成证明。 □

引理11 (向量鞅伯恩斯坦不等式)。 设$\mathbf{Z}_1,\mathbf{Z}_2,\ldots,\mathbf{Z}_T$为 d维鞅差序列，相对于 $\mathcal{F}_0,\mathcal{F}_1,\ldots,\mathcal{F}_T$，即 $\mathbb{E}[\mathbf{Z}_t|\mathcal{F}_{t-1}]=0$ 几乎处处成立，并且假设对于某个常数 K_t，$\|\mathbf{Z}_t|\mathcal{F}_{t-1}\|_{\Psi_1}\leq K_t$ 几乎处处成立。那么，对于任意 $v>0$，我们有

$$\mathbb{P}\left(\left\|\sum_{t=1}^{T}\mathbf{Z}_t\right\|\geq 2v\right)\leq\exp\left(Cd-\min\left\{\frac{v^2}{\sum_{t=1}^{T}K_t^2},\frac{v}{\max_t K_t}\right\}\right),$$

其中 C 是某个常数。

证明。 证明过程与引理 10 类似，此外还会用到引理 9 。注意，对于任意固定的 $\mathbf{u}$ 以及任意 $v>0$, 引理 9 表明

$$\mathbb{P}\left(\left|\left\langle\sum_{t=1}^{T}\mathbf{Z}_t,\mathbf{u}\right\rangle\right|\geq v\right)\leq 2\exp\left(-\min\left\{\frac{v^2}{\sum_{t=1}^{T}K_t^2},\frac{v}{\max_t K_t}\right\}\right).$$

然后，采用与引理 10类似的步骤，可得

$$\mathbb{P}\left(\left\|\sum_{t=1}^{T}\mathbf{Z}_t\right\|\geq 2v\right)\leq 2\times 5^d\exp\left(-\min\left\{\frac{v^2}{\sum_{t=1}^{T}K_t^2},\frac{v}{\max_t K_t}\right\}\right),$$

至此，证明完成。 □

引理12。 假设 $\mathbf{Z}_1,\mathbf{Z}_2,\ldots,\mathbf{Z}_T$是相对于 $\mathcal{F}_0,\mathcal{F}_1,\ldots,\mathcal{F}_T$ 的 d 维鞅差，即 $\mathbb{E}[\mathbf{Z}_t|\mathcal{F}_{t-1}]=0$ 几乎必然成立，并且假设对于某个常数 K_t，有 $\sup_{|\mathcal{S}|\leq s,\mathcal{S}\subseteq[d]}\|\mathcal{H}_{\mathcal{S}}(\mathbf{Z}_t)|\mathcal{F}_{t-1}\|_{\Psi_1}\leq K_t$ 几乎必然成立。那么对于任意 $v>0$，有

$$\mathbb{P}\left(\sup_{|\mathcal{S}|\leq s,\,\mathcal{S}\subseteq[d]}\left\|\mathcal{H}_{\mathcal{S}}\left(\sum_{t=1}^{T}\mathbf{Z}_t\right)\right\|\geq 2v\right)\leq\exp\left(Cs\log(ed/s)-\min\left\{\frac{v^2}{\sum_{t=1}^{T}K_t^2},\frac{v}{\max_t K_t}\right\}\right),$$

其中 C 是某个常数。

证明。 引理 11 表明，对于任意固定的 $\mathcal{S}$,

$$\mathbb{P}\left(\left\|\mathcal{H}_{\mathcal{S}}\left(\sum_{t=1}^{T}\mathbf{Z}_t\right)\right\|\geq 2v\right)\leq 2\times 5^s\exp\left(-\min\left\{\frac{v^2}{\sum_{t=1}^{T}K_t^2},\frac{v}{\max_t K_t}\right\}\right).$$

然后对所有 $\mathcal{S}$ 取并集，得到

$$\mathbb{P}\left(\left\|\mathcal{H}_{\mathcal{S}}\left(\sum_{t=1}^{T}\mathbf{Z}_t\right)\right\|\geq 2v\right)\leq\binom{d}{s}\times 2\times 5^s\exp\left(-\min\left\{\frac{v^2}{\sum_{t=1}^{T}K_t^2},\frac{v}{\max_t K_t}\right\}\right),$$

其中 $\binom{d}{s}\leq(ed/s)^s$ 完成了证明。 □

引理13. 假设 $g_1,\ldots,g_n$ 是零均值次高斯随机变量，且长度为 n 的向量 $\mathbf{g}=(g_1,\ldots,g_n)$ $s-\Psi_2$ 的 *Orlicz*范数至多为 σ，即 $\sup_{\mathcal{S}\subseteq[n],\,|\mathcal{S}|\leq s}\|[\mathbf{g}]_{\mathcal{S}}\|_{\Psi_2}\leq\sigma$。记 $(g_{(1)},\ldots,g_{(n)})$ 为 $(|g_1|,\ldots,|g_n|)$ 的非递增重排。那么

$$\mathbb{P}\left(\sqrt{\sum_{j=1}^{s}\left(g_{(j)}\right)^2}\geq C\sqrt{s\log(en/s)}\sigma+u\sigma\right)\leq\exp(-cu^2)$$

对于所有 $t>0$和 $s\in\{1,\ldots,p\}$。

证明。 根据该假设，对于任意固定的集合 $\mathcal{S} \subseteq [n]$ 和 $|\mathcal{S}| = s$，我们有

$$\mathbb{P}\left(\sqrt{\sum_{i\in\mathcal{S}}(g_i)^2} \geq C\sqrt{s}\sigma + u\right) \leq \exp\left(-\frac{cu^2}{\sigma^2}\right).$$

然后，我们对 $[n]$ 的所有 s 子集取均匀分布。值得注意的是

$$\#\{\mathcal{S}:\ |\mathcal{S}| = s,\ \mathcal{S} \subseteq [n]\} = \binom{n}{s} \leq \left(\frac{en}{s}\right)^s.$$

于是我们有

$$\mathbb{P}\left(\max_{|\mathcal{S}|=s,\ \mathcal{S}\subseteq[n]} \sqrt{\sum_{i\in\mathcal{S}}(g_i)^2} \geq C\sqrt{s}\sigma + u\right) \leq \left(\frac{en}{s}\right)^s \exp\left(-\frac{cu^2}{\sigma^2}\right).$$

这意味着

$$\mathbb{P}\left(\max_{|\mathcal{S}|=s,\ \mathcal{S}\subseteq[n]} \sqrt{\sum_{i\in\mathcal{S}}(g_i)^2} \geq C\sqrt{s\log(en/s)}\sigma + u\sigma\right) \leq \exp\left(-cu^2\right),$$

由此完成证明。 □

引理 14. 设 $\mathbf{\Sigma}_1$、……、 $\mathbf{\Sigma}_n$ 是对称正定矩阵，其最小特征值大于 $\lambda_{\min} > 0$，即 $\lambda_{\min}(\mathbf{\Sigma}_i) \geq \lambda_{\min}$。于是我们有 $\lambda_{\min}(\sum_{i=1}^n \mathbf{\Sigma}_i) \geq n\lambda_{\min}$。

证明。 值得注意的是， $\lambda_{\min}(\sum_{i=1}^n \mathbf{\Sigma}_i) = \inf_{\mathbf{u}\in\mathbb{S}^{d-1}} \mathbf{u}^\top(\sum_{i=1}^n \mathbf{\Sigma}_i)\mathbf{u}$。另一方面，对于任意$\mathbf{u} \in \mathbb{S}^{d-1}$，我们有 $\mathbf{u}^\top\mathbf{\Sigma}_i\mathbf{u} \geq \lambda_{\min}$。因此，我们有 $\inf_{\mathbf{u}\in\mathbb{S}^{d-1}} \mathbf{u}^\top(\sum_{i=1}^n \mathbf{\Sigma}_i)\mathbf{u} \geq \sum_{i=1}^n \inf_{\mathbf{u}\in\mathbb{S}^{d-1}} \mathbf{u}^\top\mathbf{\Sigma}_i\mathbf{u} \geq n\lambda_{\min}$，由此完成证明。 □

引理 15. 假设 $\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_T$ 是关于 $\mathcal{F}_0, \mathcal{F}_1, \ldots, \mathcal{F}_T$ 的 d 维鞅差，即 $\mathbb{E}[\mathbf{Z}_t|\mathcal{F}_{t-1}] = 0$ 几乎必然成立，并且假设对于某个常数 K_t， $\|\mathbf{Z}_t|\mathcal{F}_{t-1}\|_{\Psi_1} \leq K_t$ 几乎必然成立。那么对于任意 $v > 0$，我们有

$$\mathbb{P}\left(\left\|\sum_{t=1}^T \mathbf{Z}_t\right\|_\infty \geq 2v\right) \leq d\exp\left(-\min\left\{\frac{v^2}{C\sum_{t=1}^T K_t^2}, \frac{v}{\max_t K_t}\right\}\right),$$

其中 C 是某个常数。

证明。 对于任意 $i = 1, \ldots, d$，项 $\sum_{t=1}^T [\mathbf{Z}_t]_i$ 是鞅差之和，其中 $\|[\mathbf{Z}_t]_i\|_{\Psi_1} \leq K_t$。根据引理 9，对于任意 v，有：

$$\mathbb{P}\left(\left|\sum_{t=1}^T [\mathbf{Z}_t]_i\right| \geq 2v\right) \leq \exp\left(-\min\left\{\frac{v^2}{C\sum_{t=1}^T K_t^2}, \frac{v}{\max_t K_t}\right\}\right).$$

对所有 $i = 1,\ldots,d$ 取均匀分布，得到

$$\mathbb{P}\left(\max_{i=1,\ldots,d}\left|\sum_{t=1}^{T}[\mathbf{Z}_t]_i\right| \geq 2v\right) \leq d\exp\left(-\min\left\{\frac{v^2}{C\sum_{t=1}^{T}K_t^2}, \frac{v}{\max_t K_t}\right\}\right),$$

由此完成证明。 □

引理 16. 对于任意正值 α、$\beta > 0$ 以及任意正整数 $t \geq s$，有

$$\prod_{l=s}^{t}\left(1-\frac{\alpha}{l+\beta}\right) \leq \left(\frac{s+\beta}{t+1+\beta}\right)^{\alpha}.$$

证明。 由于 $\log(1-x) \leq x$，我们有

$$\prod_{l=s}^{t}\left(1-\frac{\alpha}{l+\beta}\right) = \exp\left(\sum_{l=s}^{t}\log\left(1-\frac{\alpha}{l+\beta}\right)\right) \leq \exp\left(-\sum_{l=s}^{t}\frac{\alpha}{l+\beta}\right).$$

此外，单调递减函数满足

$$\sum_{l=s}^{t}\frac{\alpha}{l+\beta} \geq \int_{s}^{t+1}\frac{\alpha}{x+\beta}\mathrm{d}x = \alpha\log\left(\frac{t+1+\beta}{s+\beta}\right).$$

因此，得到

$$\prod_{l=s}^{t}\left(1-\frac{\alpha}{l+\beta}\right) \leq \left(\frac{s+\beta}{t+1+\beta}\right)^{\alpha}.$$

□

引理 17. 对于任意正数 α、$\beta > 0$ 以及任意正整数 $t \geq s$，我们有

$$\sum_{l=s}^{t}(l+\beta)^{\alpha} \leq \frac{1}{\alpha+1}(t+1+\beta)^{\alpha+1}.$$