具有低阶矩信息的厚尾线性老虎机算法

摘　要

作为在线学习范式的一种，老虎机模型被广泛应用于建模序列决策问题，在广告投放，医学实验设计等领域应用广泛。特别的，老虎机模型将可选择的决策抽象为老虎机的摇臂，将决策的影响抽象为老虎机的即时收益，算法利用收益信息指导决策，旨在最大化一段时间内的累计收益。为了充分利用摇臂的决策信息，线性老虎机在一般多臂老虎机的基础上，将摇臂表示为 d 维的语意向量，即时收益的期望表示为摇臂向量的线性函数，增强了老虎机模型的表达能力。高效的线性老虎机算法需要权衡探索与利用，即兼顾对历史决策信息的利用和对未知的决策的尝试，而其中的关键在于利用历史信息对未知参数估计。

在金融数据等应用场景中，即时奖励服从厚尾分布，有更大的概率取到极端值，这也导致现有大部分基于次高斯分布奖励的老虎机算法不再适用。针对具有低阶矩信息的厚尾线性老虎机问题，本文做出了如下贡献：

本文首次提出针对具有低阶矩信息的厚尾线性老虎机算法 PHO。该算法利用伪 Huber 损失函数对线性模型未知参数进行估计，基于参数估计对应的置信椭球，采用乐观主义原则选取新的决策。本文进一步对文中提出的 PHO 算法进行理论分析，证明了算法具有

$$\tilde{O}\left(d\min\left\{\sqrt{\sum_{t=1}^{T}v_t^2}\cdot\left(\sum_{t=1}^{T}\frac{1}{v_t^2}\right)^{(2-p)/2p},T^{1/p}\right\}\right)$$

的大概率遗憾上界。相较于基于截断和中位数平均的厚尾线性老虎机算法，本文提出的 PHO 算法有更小的遗憾上界。

关键词：厚尾线性老虎机　伪 Huber 损失　弗里德曼集中不等式

Robust linear bandit algorithm under heavy tail reward and known moments

Abstract

As an online learning model, bandit models can be used to model sequential decision problems and is widely used in areas such as advertising placement and medical experiment design. Specifically, the bandit model models available decisions as the arms of a bandit, and the impact of decisions is modeled as the immediate reward of the bandit. The algorithm utilizes reward information to guide decision-making, aiming to maximize the cumulative reward over a period of time. To utilize the context information of the arms, the linear bandit model represents the context information of each arm as a d-dimensional vectors, and the expected immediate reward is modeled as a linear function of the arm vector, enhancing the expressive power of the bandit model. Efficient linear bandit algorithms need to balance exploration and exploitation, i.e., balancing the utilization of historical decision information and the attempt of unknown decisions, and the key lies in utilizing historical information to estimate unknown parameters.

In many real world applications such as finance, the immediate reward follows heavy tail distribution with higher probability taking extreme values, which leads to the failure of current bandit learning algorithms under the assumption of sub-Gaussian noise. Aiming at solving heavy-tail linear bandit problem finite moment information, with This paper made following contributions:

This paper first proposed the PHO algorithm for heavy tail linear bandit with moment information. The PHO algorithm utilize Psudo-Huber loss to estimate the unknown parameter vector in linear bandit model. With the confidence elipsoid came along with the estimator, the algorithms pick new actions using the optimism in the face of uncertainty. The paper followed by an rigorous theoretical analysis of the regret of PHO algorithm and established an high probability upper bound of:

$$\tilde{O}\left(d\min\left\{\sqrt{\sum_{t=1}^{T}v_t^2}\cdot\left(\sum_{t=1}^{T}\frac{1}{v_t^2}\right)^{(2-p)/2p},T^{1/p}\right\}\right)$$

Comparing with algorithms based on truncation and median of means, the PHO algorithm enjoys tighter upper bound of regret. .

Key words: Heavy tail linear bandit Psudo-Huber loss Freedman inequality

目录

一　绪论

(一)　研究背景及文献综述

老虎机模型

老虎机模型作为一类经典的在线机器学习问题，不同于传统的静态的监督学习任务。其可视为决策者与老虎机之间的动态博弈问题，模型数据收集与模型拟合交替进行，互相影响，而非在事先给定的数据集上拟合模型。由于其建模动态交互的特点，老虎机模型通常用以建模序列决策问题，在广告投放 [11]，医学实验 [3] 等领域中有广泛的应用 [2]。

最经典的 K 臂老虎机问题，由 Tompson[16] 首先提出。其假设老虎机有 K 个摇臂，决策者每轮 (t 轮) 选择一个摇臂，获取一定的即时收益 r_t。每个摇臂的收益为符合特定的分布的随机变量，决策者不知道各摇臂收益的分布情况，仅能通过每次摇臂获取摇臂收益的一次观测。算法的目标是最大化决策者在 T 次摇臂过程中获得的累计收益。老虎机模型的特殊之处在于摇臂决策与收益观测的双向影响关系，即摇臂决定当前轮即时收益，新的收益观测也会增加决策者对摇臂未知收益分布的认知，从而影响决策者后续的摇臂决策。因此，为了最大化累计收益，高效的老虎机算法需要在探索和利用之间权衡，既尽可能多的尝试不同的摇臂以全面地了解不同摇臂收益分布，又要充分利用已有信息，选择经验收益更大的摇臂以最大化即时收益。

尽管 K 臂老虎机模型简洁易懂，其无法利用摇臂的语意信息，仅能通过即时收益对摇臂收益分布进行推断。而利用语意信息的老虎机模型在 K 臂老虎机模型上引入语意信息，增强了老虎机模型的表达能力。例如在新闻推荐问题中 [11] 一文将网站当天每条待推荐的新闻视为多臂老虎机的一只臂，新闻的特征（类别，长度等）以及被推荐者的特征表示为一个 d 维的语意向量，用户的点击行为作为即时收益。由于用户点击行为一定程度上取决于新闻语意向量，算法可以利用语意向量以获得更高点击率。利用语意信息的老虎机模型的决策不仅取决于历史决策和收益，还取决于摇臂的语意信息，因而具有更强的灵活性。

特别的，根据决策者在每轮 t 可选行动集合 $\mathcal{D}_t$ 的不同（每轮可选行动数 $|\mathcal{D}_t|$ 是否有限，是否相同，每轮行动集合 $\mathcal{D}_t$ 是否完全相同）可分为不同的变种。例如在 [11] 一文中，作者利用线性老虎机模型对网站新闻推送进行建模，模型假设每轮可选行动数（网站待显示新闻数目）相同但每轮行动集合不相同（每轮待推送新闻内容不完全相同）。而 [17] 一文中模型假设不同轮数可选行动集合完全相同且可选行动数有限。而本文采用与 [1] 相同的假设，即假设每轮可选行动集合无限且相同。

在利用语意信息的老虎机模型中，最常见的模型为式1所示的线性老虎机，即假设即时收益是摇臂语意向量的线性函数加上一个随机误差项。其中 r_t 为 t 时刻即时收益，θ^* 为模型未知参数，ϵ_t 为模型误差项。在线性老虎机模型的基础上，[5] 一文利用广义线性模型（GLM）对即时收益进行建模。[14] 一文则使用了一般函数（GFA）并用欧式维度对函数复杂度进行刻画。其中线性老虎机为最特殊的一类模型，但从线性老虎机模型向后两者的推广并不复杂，故本文采用线性老虎机模型。

$$r_t = \langle \phi_t, \theta^* \rangle + \epsilon_t \tag{1}$$

乐观主义决策原则

为了权衡探索和利用，主流的老虎机算法都采用了基于面对不确定性的乐观主义决策原则，即假设老虎机环境和最好的估计一样好 [9]。在 K 臂老虎机问题中最经典的算法框架，置信区间上界（UCB）算法由 [8] 一文首先提出。其在每一轮选择摇臂时，不是贪婪地选取收益均值估计量最大的摇臂，而是选取置信区间上界最大的摇臂。式2给出了基本的 UCB 算法，其中 a_t 为 t 轮选择的摇臂，$\widehat{\mu^i}_{t-1}$ 为 $t-1$ 轮结束时对第 i 个摇臂收益期望的估计量，$N_t(i)$ 为前 t 轮中摇臂 i 被选择的总次数。因此2式第一部分可以视为利用，倾向于选取摇臂收益期望估计量最大的摇臂。第二部分可以

视为探索，选择历史数据少的摇臂。

$$a_t = \arg\max_{i \in K} \left(\widehat{\mu^i}_{t-1} + \mathcal{O}(\frac{1}{\sqrt{N_t(i)}}) \right) \tag{2}$$

在基于语义信息的线性老虎机模型中，UCB 算法得到了推广，具体有不同的名称，如 LinUCB,OFUL 等。基本思想是构建未知参数 θ^* 估计量 θ_{t-1} 的置信椭球 $\mathcal{C}_{t-1}$，使得 θ^* 大概率落在置信椭球内。在未知参数 θ^* 在置信椭球的条件下，选择可能给出最大奖励的摇臂，即通过如式3的双优化问题选择摇臂。这样的选择原则兼顾了探索与利用。一方面将未知参数限制在以估计量 θ_{t-1} 为中心的置信椭球内，利用对未知参数的先验知识最大化即时期望。另一方面，鼓励算法选择先验知识较少的摇臂（置信椭球轴长更长的方向的摇臂）。

$$\phi_t = \arg\max_{\phi_t \in \mathcal{D}_t, \theta \in \mathcal{C}_{t-1}} \langle \phi_t, \theta \rangle \tag{3}$$

算法遗憾界与厚尾分布

为了评估不同老虎机算法的优劣，通常采用模型最大期望累计奖励与算法期望累计奖励的差作为评判，这个差值也被称为遗憾（Regret）。形式上可以表示为式4，其中 $\phi^* = \arg\max_{\phi} \langle \phi, \theta^* \rangle$。针对不同线性老虎机模型假设（随机误差的分布不同，可选行动集合不同），一系列工作尝试优化老虎机算法并给出算法遗憾的上界。另一方面，针对不同的模型假设，另一系列工作给出了算法遗憾的下界。

$$\sum_{t=1}^{T} |\langle \phi^*, \theta^* \rangle - \langle \phi_t, \theta^* \rangle| \tag{4}$$

异方差随机误差

异方差模型假设随机误差 ϵ_t 的分布具有不同方差（或不同的次高斯常数），在 PCA，回归分析等领域都有广泛的应用，但异方差假设下的老虎机模型相关工作较少 [14]。其中 [7] 一文最先研究异方差噪声条件下的老虎机模型。其假设随机误差 ϵ_t 来自系数为 ρ_t 的次高斯分布，其中 ρ_t 被视为老虎机的输出，可以被观测到。而 [18] 一文假设噪声有全局上界 $\epsilon_t \leq R$，且每一轮可以观测到噪声的方差 v_t^2。通过加权岭回归对未知参数 θ^* 进行估计，并给出了算法的遗憾上界 $\mathcal{O}\left(R\sqrt{dT} + d\sqrt{\sum_{t=1}^{T} v_t^2}\right)$

厚尾分布

线性老虎机模型的工作大多假设随机误差服从次高斯分布或进行有界性假设。然而在金融等场景下，数据的分布为厚尾分布，并不满足以上分布假设。为解决厚尾分布下线性老虎机问题，大部分工作采用了均值中位数法（median of means）或者截断法 (truncation)。其中 [13] 首先研究厚尾分布下的线性老虎机问题，假设随机误差的 p 阶矩有限提出了利用均值中位数和截断法的两种算法，并分别达到了 $\tilde{O}(dT^{\frac{p+1}{2p}})$ 和 $\tilde{O}(\sqrt{d}T^{\frac{2p-1}{3p-2}} + dT^{\frac{p}{3p-2}})$ 的遗憾上界。之后 [15] 同样利用均值中位数法和截断法但采用了更加精细的设计，分别达到了 $dT^{\frac{1}{1+\epsilon}}$ 的遗憾上界。[17] 针对 K 臂线性老虎机做了进一步算法优化，达到了 $\sqrt{dK}T^{\frac{1}{1+\epsilon}}$ 的遗憾上界，其中 K 为摇臂数。

除此之外 [12] 一文首先研究厚尾分布下异方差的线性老虎机模型。不同于前三篇文章，该文没有采用均值中位数法或截断法，而是通过伪 Huber 对厚尾分布中的极端值进行控制。但仅研究了 $p = 2$ 即随机误差 2 阶矩有限这一特殊情况，并且获得了 $\tilde{O}(d\sqrt{T})$ 的遗憾上界。特别的，本文将采用与 [12] 一文类似的假设，但研究 $p \in [1, 2]$ 这个更一般的情况。

（二） 研究意义及创新之处

- 本文设计了首个针对具有可观测的有限 p 阶矩误差的厚尾老虎机模型的算法。

- 本文得到了 $\tilde{O}\left(d \cdot \min\left\{\sqrt{\sum_{t=1}^{T} v_t^2} \cdot \left(\sum_{t=1}^{T} \frac{1}{v_t^2}\right)^{2-p/2p}, T^{1/p}\right\}\right)$ 的遗憾上界。在误差同分布以及 $p=1$ 这两种特殊情况下与先前工作中最好的算法具有同阶的遗憾上界。
- 本文提出的算法框架充分利用误差的 p 矩信息，在 $v_t \in [\frac{1}{T}, c]$ 的范围内可以获得比误差同分布条件下的算法有更好的收敛上界。

（三） 本文结构安排

本文结构安排为: 对具有已知低阶矩信息的线性老虎机模型进行描述 → 提出利用低阶矩信息的 PHO 算法 → 对 PHO 算法遗憾上界推导 → 对遗憾上界的分析特别的，详细的理论证明过程在附录部分。

二 符号定义及模型假设

（一） 符号定义

本文用 $\langle a, b\rangle$ 表示向量 a, b 的内积，$\|\cdot\|_2$ 表示向量的 l_2 范数。给定正定矩阵 $H \in \mathbb{R}^{d\times d}$ 和向量 x，定义 $\|x\|_H = \sqrt{x^T H x}$。$Ball_d(B)$ 表示 $\mathbb{R}^d$ 空间内半径为 B 的球体。给定正整数 K，定义 $[K] = \{1, 2, \ldots, K\}$。针对适应于 $\mathcal{F}_t$ 的随机过程 X_t，采用 $\mathbb{E}_t[X_t], \mathbb{V}_t[X_t]$ 表示条件期望和条件方差 $\mathbb{E}[X_t|\mathcal{F}], \mathbb{V}_t[X_t|\mathcal{F}_{t-1}]$。

（二） 厚尾线性老虎机决策模型

本文研究问题可视为决策者（智能体）与厚尾线性老虎机（环境）间的 T 轮博弈问题，算法旨在设计摇臂策略，最大化决策者在 T 轮博弈中的期望累计收入（reward）。首先介绍厚尾线性老虎机环境:

定义 2.1 (厚尾线性老虎机). *$\{\mathcal{D}_t\}$ 表示 t 时刻的决策集即 t 时刻决策者可使用的摇臂，并对决策集做出以下有界性假设:*

$$\forall \phi \in \mathcal{D}_t \quad L_l \le \|\phi\|_2 \le L_u \tag{5}$$

决策者在 t 轮选择摇臂 $\phi_t \in \mathcal{D}_t$ 后，可以从厚尾线性老虎机观测到即时收益 r_t 和条件 p 阶矩 m_t。其中即时收益 r_t 满足如下的线性模型:

$$r_t = \langle \phi_t, \theta^* \rangle + \epsilon_t \tag{6}$$

其中 $\|\theta^*\|_2 \le B$ 为老虎机环境的未知参数，$\epsilon \in \mathbb{R}$ 是鞅差噪声，其对应的滤子 (filtration)$\{\mathcal{F}_t\} = \{r_t.m_t, \ldots, r_1, m_1\}$ 为截至 t 轮，决策者可见的全部上下文信息。ϕ_t, m_t 均为 $\mathcal{F}_{t-1}$ 可测，且给定滤子环境噪声期望为 0，p 阶矩为 m_t。即 $\mathbb{E}[\epsilon_t|\mathcal{F}_{t-1}] = 0$，$\mathbb{E}[|\epsilon_t|^p] = m_t$。

定义 2.2 (摇臂策略). *本问题考虑总轮数为 T 轮的决策问题，定义摇臂策略 $\Phi = \{\phi_1, \ldots, \phi_T\} = \{\pi_1(\mathcal{F}_0), \ldots, \pi_T(\mathcal{F}_{T-1})\}$ 为决策者在 T 轮博弈中所采取的摇臂策略。策略旨在最大化 T 轮博弈的期望累计收益 $\sum_{t=1}^{T}\langle \phi_t, \theta^*\rangle$。定义 $\Phi^* = \{\phi_1^*, \ldots, \phi_T^*\} = \arg\max_{\Phi} \sum_{t=1}^{T}\langle \phi_t, \theta^*\rangle$ 为最优摇臂策略，$R(T)$ 为策略 Φ 的遗憾 (regret):*

$$R(T) = \sum_{t=1}^{T} \left[\langle \phi_t^*, \theta^*\rangle - \langle \phi_t, \theta^*\rangle\right] \tag{7}$$

则可用 $R(T)$ 衡量摇臂策略 Φ 的优劣。

三　利用 p 阶矩信息的线性老虎机算法 (PHO)

本节将提出利用伪 Huber 损失函数 (Psudo-Huber loss) 的乐观主义 (Optimism) 算法 (PHO)，以解决具有 p 阶矩信息的线性老虎机的决策问题。

设定 $H_0 = \lambda I, \beta_0 = \sqrt{\lambda}B, c_0 = \frac{1}{6\sqrt{\ln\frac{2}{\delta}}}, \sigma_{\min} = \frac{1}{\sqrt{T}}$

for *t=1 to T* **do**

- 利用式9构建 θ^* 的 $1-\delta$ 置信区间 $\mathcal{C}_{t-1}$;
- 求解乐观主义的决策 $(\phi_t, \cdot) = \arg\max_{\phi\in\mathcal{D}_t, \theta\in\mathcal{C}_{t-1}} \langle\phi, \theta\rangle$;
- 决策者摇 ϕ_t 臂，获得观测 (r_t, v_t);
- 计算 $\sigma_t, \omega_t, \tau_t$;
- 最小化损失函数，计算 θ^* 的估计量 θ_t;
- 计算 β_t，设置 $H_t = H_{t-1} + \frac{\phi_t}{\sigma_t}\frac{\phi_t^\top}{\sigma_t}$;

end

Algorithm 1: PHO 算法流程

本文提出的 PHO 算法1可以分为两个部分：基于乐观主义原则的决策（步骤 4-5），位置参数 θ^* 的估计和置信区间的构建（步骤 6-8，3）。

每轮 (t 轮) 算法首先基于上一轮对未知参数 θ^* 的估计 θ_t,以及所有历史观测 $\{r_1, v_1, \ldots, r_{t-1}, v_{t-1}\}$ 构建未知参数 θ^* 的置信区间 $\mathcal{C}_{t-1}$。使得未知参数 θ^* 以大概率落入置信区间 $\mathcal{C}_{t-1}$ 内。进而，算法利用置信区间获得 θ^* 的乐观主义估计 $\tilde{\theta} = \arg\max_{\theta\in\mathcal{C}_{t-1}} \left(\max_{\phi_t\in\mathcal{D}_t}\langle\phi_t, \tilde{\theta}\rangle\right)$。基于乐观主义的估计 $\tilde{\theta}$，选取可以最大化预期收益的摇臂 $\phi_t = \arg\max_{\phi\in\mathcal{D}_t}\langle\phi, \tilde{\theta}\rangle$。

获取 t 轮的观测 r_t, v_t 后，通过最小化伪 Huber 损失函数获得 t 轮对 θ^* 的估计 θ_t。其中步骤 3 所构建的置信区间为：

$$\mathcal{C}_{t-1} = \left\{\theta \in Ball_d(B) : \|\theta - \theta_{t-1}\|_{H_{t-1}} \le \beta_t\right\} \tag{8}$$

$$\beta_t = 32\left[\frac{b^p\kappa\eta^{\frac{2-p}{p}}}{\tau_0^p} + \eta^{\frac{2-p}{p}}\sqrt{\kappa\tau_0^{2-p}b^p\ln\frac{1}{\delta}} + \tau_0\eta^{\frac{2-p}{p}}\ln\frac{1}{\delta}\right]$$

$$\kappa = 2d\ln\left(1 + \frac{T^2L^2}{d\lambda}\right) \quad \eta = \sqrt{1 + \frac{K^2}{L_l^2}(\lambda + \sum_{t=1}^{T}\frac{L_u^2}{v_t^2})} \tag{9}$$

可以发现，给定总轮数 T 后，β_t 为常数，而 $H_{t-1} \succeq 0$ 且随轮数 t 单调递增。所以置信区间 $\mathcal{C}$ 的半径（欧式空间）随 t 单调递减。即随着观测数的增多，对未知参数 θ^* 的估计更加准确，置信区间变小，智能体减少探索（Exploration）更多利用（Exploit）以往数据进行决策。

步骤 5 中构建 σ_t:

$$\sigma_t = \max\left\{v_t, \sigma_{\min}, \frac{\|\phi_t\|_{H_{t-1}^{-1}}}{c_0}, \frac{168\sqrt{L_uB}\|\phi_t\|_{H_{t-1}^{-1}}^{1/2}}{d^{\frac{1}{4}}}\right\} \tag{10}$$

从第一项可以看出 σ_t 是任意轮数条件方差 v_t 的上界，第二项 σ_t 是一个小的正数，确保 σ_t 非零，在后续我们将其设为 $\frac{1}{\sqrt{T}}$。第三，四项是为了控制损失函数二阶导 $\nabla^2 L_T$ 的下界，可在引理4.1证明过程中找到。如文献综述所说，本文采用如下的伪 Huber 损失函数。伪 Huber 损失函数是 Huber 损失函数的平滑逼近，一方面可以像 Huber loss 一样进行方差-偏差权衡，另一方面对于稳健系数 τ，

未知参数 θ 处处可导，拥有更好的优化性质：

$$l_\tau(x) = \tau(\sqrt{\tau^2 + x^2} - \tau) \tag{11}$$

进而步骤 7 通过最小化经验伪 Huber 损失函数获得 θ^* 的估计值：

$$\theta_t = \underset{\theta \in Ball_d(B)}{\arg\min} \quad L_t(\theta) = \frac{\lambda}{2}\|\theta\|^2 + \sum_{t=1}^{T} l_{\tau_t}\left(\frac{y_t - \langle \phi_t, \theta \rangle}{\sigma_t}\right)$$

在伪 Huber 损失函数 l_τ 中，稳健系数 τ_t 起到权衡估计量方差和偏差的作用。在独立同分布的样本的参数估计问题中，不同观测有相同权重，稳健系数 $\tau_t = \tau$ 为固定常数。而在本文的线性老虎机模型中，不同观测有相关关系。以第 t 轮为例，摇臂动作 ϕ_t 取决于前 $t-1$ 轮观测 $\{r_1, v_1, \ldots, r_{t-1}, v_{t-1}\}$，进而导致 t 轮获取的观测 r_t, v_t 与前 $t-1$ 轮观测相关 $\{r_1, v_1, \ldots, r_{t-1}, v_{t-1}\}$ 因而稳健系数 τ_t 与样本产生的轮数有关，如下所示可以分为 τ_0 和 $\left(\frac{1+w_t^2}{w_t^2}\right)^{\frac{2}{p}}$ 两部分。

$$w_t = \left\|\frac{\phi_t}{\sigma_t}\right\|_{H_{t-1}^{-1}} \quad \tau_t = \tau_0 \cdot \left(\frac{1+w_t^2}{w_t^2}\right)^{\frac{2}{p}} \quad \tau_0 = \max\{\sqrt{d}, \kappa^{\frac{1}{p}} b\} \tag{12}$$

其中 $\left(\frac{1+w_t^2}{w_t^2}\right)^{\frac{2}{p}} > 1$ 且随 t 单调递增，即新产生的样本有对于损失函数有更大的影响，因为旧样本相较于新样本与当前轮的相关性更低。

四　理论分析

(一)　算法遗憾上界

在本节中将给出 PHO 策略遗憾的大概率上界，并给出证明，首先对遗憾 $R(T)$ 分解：

$$\begin{aligned}
R(T) &= \sum_{t=1}^{T} [\langle \phi_t^*, \theta^* \rangle - \langle \phi_t, \theta^* \rangle] \\
&\leq \sum_{t=1}^{T} \left[\sup_{\tilde{\phi}_t \in \mathcal{D}_t, \theta_t \in \mathcal{C}_{t-1}} \langle \tilde{\phi_t}, \theta_t \rangle - \langle \phi_t, \theta^* \rangle \right] \\
&= \sum_{t=1}^{T} \left[\sup_{\theta_t \in \mathcal{C}_{t-1}} \langle \phi_t, \theta_t \rangle - \langle \phi_t, \theta^* \rangle \right] \\
&\leq \sum_{t=1}^{T} \|\phi_t\|_{H_{t-1}^{-1}} \cdot \sup_{\theta_t \in \mathcal{C}_{t-1}} \|\theta_t - \theta^*\|_{H_{t-1}}
\end{aligned} \tag{13}$$

最后一个不等式可以通过变形和柯西不等式获得。后面证明的关键是在每一轮构建未知参数 θ^* 的狭窄置信区间，从上述分解可看出，更紧的置信区间可以获得更紧的上界。利用向量的中值定理可以得到下式：

$$\nabla L_T(\theta_T) - \nabla L_T(\theta^*) = \int_0^1 \nabla^2 L_T((1-\alpha)\theta^* + \alpha\theta^*) d\alpha \cdot (\theta_T - \theta^*) \tag{14}$$

因此 $\|\theta_t - \theta^*\|_{H_{t-1}}$ 的上界问题可以转化为求 L_T 一阶导的上界和其二阶导的下界。

引理 4.1. 假设对于任意轮数 t，$\mathbb{E}[|z_t(\theta^*)|^p|\mathcal{F}_{t-1}] \leq b^p$ 均成立。当稳健参数 τ_0 满足以下条件时:

$$\tau_0 \geq \max\{\sqrt{d}, \kappa^{\frac{1}{p}} b\}$$

对于任意 $\|\theta\|_2 \leq B$，$\nabla^2 L_2(\theta)$ 以至少 $1-2\delta$ 概率拥有如下下界:

$$\frac{1}{4} H_T \preceq \nabla^2 L_T(\theta)$$

由于 $\|(1-\alpha)\theta^* + \alpha\theta^*\|_2 \leq B$，结合14及引理4.1可得:

$$\langle \theta_T - \theta^*, \nabla L_T(\theta_T) - \nabla L_T(\theta^*) \rangle \geq \frac{1}{4} \|\theta_T - \theta^*\|_{H_T}^2$$

由于 L_T 是凸函数，θ^T 为其鞍点，由带约束凸优化问题的一阶稳定性条件可知，对于任意 $\theta \in Ball_d(B)$:

$$\langle \nabla L_T(\theta_T), \theta_T - \theta \rangle \leq 0$$

由于 $\|\theta^*\|_2 \leq B$:

$$\begin{aligned}
\frac{1}{4} \|\theta_T - \theta^*\|_{H_T}^2 &\leq \langle \theta_T - \theta^*, -\nabla L_T(\theta^*) \rangle + \langle \theta_T - \theta^*, \nabla L_T(\theta_T) \rangle \\
&\leq \langle \theta_T - \theta^*, -\nabla L_T(\theta^*) \rangle \\
&\leq \|\theta_T - \theta^*\|_{H_T} \|\nabla L_T(\theta^*)\|_{H_T^{-1}}
\end{aligned}$$

从而可以得到:

$$\|\theta_T - \theta^*\|_{H_T} \leq 4 \|\nabla L_T(\theta^*)\|_{H_T^{-1}} \tag{15}$$

引理 4.2. 假设对于任意轮数 $T \geq 1$，有 $\mathbb{E}_t[|z_t(\theta^*)|^p] \leq b^p$，则以不小于 $1-\delta$ 的概率有下式成立:

$$\|\nabla L_T(\theta^*)\|_{H_T^{-1}} \leq \sqrt{\lambda} B + 8 \left[\frac{b^p \kappa \eta^{\frac{2-p}{p}}}{\tau_0^p} + \eta^{\frac{2-p}{p}} \sqrt{\kappa \tau_0^{2-p} b^p \ln \frac{1}{\delta}} + \tau_0 \eta^{\frac{2-p}{p}} \ln \frac{1}{\delta} \right] \tag{16}$$

因此以不小于 $1-\delta$ 的概率有:

$$\|\theta_T - \theta^*\|_{H_T} \leq \beta_T = 4\sqrt{\lambda} B + 32 \left[\frac{b^p \kappa \eta^{\frac{2-p}{p}}}{\tau_0^p} + \eta^{\frac{2-p}{p}} \sqrt{\kappa \tau_0^{2-p} b^p \ln \frac{1}{\delta}} + \tau_0 \eta^{\frac{2-p}{p}} \ln \frac{1}{\delta} \right] \tag{17}$$

进而有:

$$\sup_{\theta \in \mathcal{C}_t} \|\theta - \theta^*\|_{H_t} \leq \sup_{\theta \in \mathcal{C}_t} \|\theta - \theta_t\|_{H_t} + \|\theta_t - \theta^*\|_{H_t} \leq 2\beta_t$$

由于 β_t 随 t 单调递增:

$$Reg(T) \leq 2\beta_T \sum_{t=1}^{T} \|\phi_t\|_{H_{t-1}^{-1}}$$

定理 4.3. 设置 σ_t 满足:

$$\sigma_t = \max \left\{ v_t, \sigma_{\min}, \frac{\|\phi_t\|_{H_{t-1}^{-1}}}{c_0}, \frac{168\sqrt{L_u B} \|\phi_t\|_{H_{t-1}^{-1}}^{1/2}}{d^{\frac{1}{4}}} \right\}$$

则有：

$$\sum_{t=1}^{T}\|\phi_t\|_{H_{t-1}^{-1}} \leq \sqrt{\kappa}\cdot\sqrt{\sum_{t=1}^{T}v_t^2+T\sigma_{\min}^2}+\frac{L_u\kappa}{c_0^2\sqrt{\lambda}}+\frac{L_uB\kappa}{168^2\sqrt{d}}$$

因此：

$$Reg(T)\leq 2\beta_T\left(\sqrt{\kappa}\cdot\sqrt{\sum_{t=1}^{T}v_t^2+T\sigma_{\min}^2}+\frac{L_u\kappa}{c_0^2\sqrt{\lambda}}+\frac{L_uB\kappa}{168^2\sqrt{d}}\right) \tag{18}$$

特别的，由于 $\sigma_t \geq v_t$，引理4.1, 4.2中假设 $\mathbb{E}_t[|z_t(\theta^*)|^p] \leq b^p$ 成立且 b 可取为 1。进而，取 $\tau_0 \geq \max\{\sqrt{d}, \kappa^{\frac{1}{p}}b\}$。取 $\sigma_{min}=\frac{1}{\sqrt{T}}$ 可以得到下面不等式中第一项，将 σ_{min} 取为常数可以得到下面的不等式中第二项。

$$Reg(T)\leq \tilde{O}\left(d\cdot\min\left\{\sqrt{1+\sum_{t=1}^{T}v_t^2}\cdot\left(1+\sum_{t=1}^{T}\frac{1}{v_t^2}\right)^{\frac{2-p}{2p}},\sqrt{T+\sum_{t=1}^{T}v_t^2}\cdot(1+T)^{\frac{2-p}{2p}}\right\}\right) \tag{19}$$

（二） 理论结果的解读

式19给出了 PHO 算法的遗憾上界，其中遗憾上界随对摇臂维度 d 线性增加，与总轮数 T 的关系与每轮误差的 p 阶矩有关。

- 当不同轮之间误差的 p 阶矩为相同常数时，模型假设退化为与 [15] 一文相同。本文 PHO 算法达到了与其文中 MENU，TOFU 算法相同的遗憾上界 $\tilde{O}(dT^{\frac{1}{p}})$。
- 对于 $p=2$ 即假设模型具有误差二阶矩信息时，本文模型退化为与 [12] 一文相同的模型。本文 PHO 算法的遗憾上界与其文中的 adaOFUL 算法具有相同的阶 $\tilde{O}(d\sqrt{T})$。
- 特别的，当 v_t 具有阶 $v_t \in \left[O(\frac{1}{\sqrt{T}}), O(1)\right]$ 时。模型遗憾上界阶比 $dT^{\frac{1}{p}}$ 更低。其中 $v_t=O(\frac{1}{\sqrt{T}})$ 时 $Reg(T)\leq\tilde{O}(dT^{\frac{2-p}{p}})$

五 结论与展望

（一） 研究结论

在大数据时代，老虎机模型以及其背后的一系列在线学习算法具有直观重要的意义。尽管老虎机模型及其变种获得了研究者的关注，但现有研究大多集中于次高斯奖励的条件下。特别地，具有低阶矩信息的线性老虎机模型从未被研究过。本文基于伪 Huber 损失函数以及乐观主义决策原则，设计出了 PHO 算法，填补了该领域的空白。

针对现有工作对具有低阶矩信息的线性老虎机研究的缺失。本文提出了相应的 PHO 算法，并给出了 PHO 算法遗憾上界的理论推导。通过与现有工作的对比，证明了 PHO 算法遗憾上界在特殊情况下（p=2 或误差同分布）与现有工作具有相同阶，在特定情况下有比现有工作更低阶的的遗憾上界。此外，本文为低阶线性老虎机问题提供了在截断法和中位数平均法之外的基于伪 Huber 损失函数的高效算法。

（二） 未来展望

在本文中仅研究了线性老虎机这一简化模型。在未来研究中可以讲 PHO 算法推广至线性马尔可夫决策过程问题（Linear MDP）模型。[12][18] 都证明了这种推广可行并且相对容易。另一方面，

线性老虎机是一般函数老虎机甚至是基于深度学习模型的老虎机的特例。未来可以研究基于一般函数逼近的老虎机/马尔可夫决策过程问题。这样的推广有助于对一般的强化学习问题或现代深度强化学习问题提供理论工具。

另一方面本文提出的 PHO 算法的理论上界存在进一步优化的可能。具体而言，当 $v_t = 0$ 时，老虎机模型不再具有随机性。根据线性回归理论，在 d 次摇臂后可以获得未知参数 θ^* 的显式解。即模型遗憾上界应该可由 d 控制，而与 T 无关。[12] 一文算法的遗憾上界在 $v_t = 0$ 时即满足要求。而 PHO 算法遗憾上界在 $v_t = 0$ 时依然有 $T^{\frac{1}{p}}$ 的阶，因此未来应着力于进一步优化本文的理论证明过程。

A 附录

(一) 证明过程所需不等式及符号定义

首先，给出以下符号的定义：

$$H_t = \lambda I + \frac{\phi_t}{\sigma_t}\frac{\phi_t^\top}{\sigma_t} \quad w_t = \|\frac{\phi_t}{\sigma_t}\|_{H_{t-1}^{-1}} \quad \forall t \in [T] \tag{20}$$

$$z_t(\theta) = \frac{y_t - \langle \phi_t, \theta \rangle}{\sigma_t} \quad z_t^* = z_t(\theta^*) \tag{21}$$

为了方便阅读，将以下定义，符号与量重新介绍如下：

$$L_l \le \|\phi_t\|_2 \le L_u \quad \|\theta^*\|_2 \le B \tag{22}$$

$$\sigma_t = \max\left\{ v_t, \sigma_{\min}, \frac{\|\phi_t\|_{H_{t-1}^{-1}}}{c_0}, \frac{168\sqrt{L_u B}\|\phi_t\|_{H_{t-1}^{-1}}^{1/2}}{d^{\frac{1}{4}}} \right\} \quad \sigma_t \le K \tag{23}$$

从 σ_t 的定义可知，σ_t 显然有界，这里定义其上界为 K。下面引理A.1，A.2给出了本文证明大概率上下界的主要理论工具，弗里德曼不等式。

引理 A.1 (弗里德曼集中不等式). *[6]* 如果随机过程 $\{X_t\}_{t\in[T]}$ 适应于滤子 $\{\mathcal{F}_t\}_{t\in[T]}$，且满足 $\mathbb{E}[X_t|\mathcal{F}_{t-1}] = 0, |X_t| \le M, \sum_{t=1}^T \mathbb{E}[X_t^2|\mathcal{F}_{t-1}] \le V$，其中 M, V 为正常数，则至少有 $1-\delta$ 的概率：

$$\sum_{t=1}^T X_t \le \sqrt{2Vln\frac{1}{\delta}} + \frac{2M}{3}ln\frac{1}{\delta}$$

弗里德曼集中不等式说明鞅过程有界，且条件二阶矩有限，则鞅和以大概率聚集在鞅的均值和周围。特别的，如 X_t 相互独立，则弗里德曼不等式可变为伯恩斯坦集中不等式 *[4]*。

引理 A.2 (利用方差信息的弗里德曼集中不等式). *[10]* 假设随机过程 $\{X_t\}_{t\in[T]}$ 适应于滤子 $\mathcal{F}_t$，$\mathbb{E}[X_t|\mathcal{F}_{t-1}] = 0, |X_t| \le M, \sum_{t=1}^T \mathbb{V}_t[X_t] \le V$ 其中 M, V 为正常数，则至少有 $1-\delta$ 的概率：

$$|\sum_{t=1}^T X_t| \le 3\sqrt{\sum_{t=1}^T \mathbb{V}_t[X_t] \cdot \ln\frac{2K}{\delta}} + 5M\ln\frac{2K}{\delta}$$

其中 $K = 1 + \lceil 2\ln_2 \frac{V}{M} \rceil$

引理A.3，A.5分别给出了与 $w_t = \phi_t/\sigma_t$ 相关的两个量的大概率上界和下界。

引理 A.3 (与 w_t 有关的上界，见 [1] 引理 11). 假设随机过程 $\{x_t \in \mathbb{R}^d\}_{t\in[T]}$ 对于任意的时刻 t 有 $\|x_t\|_2 \le L$。令 $Z_t = \sum_{s=1}^t x_t x_t^\top + \lambda I$ 则：

$$\sum_{t=1}^T \min\{1, \|x_t\|_{Z_{t-1}^{-1}}^2\} \le 2d\ln\left(\frac{d\lambda + TL^2}{d\lambda}\right)$$

代入 $x_t = w_t = \frac{\phi_t}{\sigma_t}$ 有：

$$\sum_{t=1}^T \min\{1, w_t^2\} = \sum_{t=1}^T \min\{1, \left\|\frac{\phi_t}{\sigma_t}\right\|_{H_{t-1}^{-1}}\} \le 2d\ln\left(1 + \frac{TL^2}{d\lambda\sigma_{\min}^2}\right) \tag{24}$$

为简化符号，记 $\kappa = 2d\ln\left(1+\frac{TL^2}{d\lambda\sigma_{\min}^2}\right)$

引理 A.4 (矩阵特征值有关的性质)**.** 本引理不证明地给出以下矩阵特征值有关的性质。

1. 如果实矩阵 $A \in \mathbb{R}^{n\times n}$ 特征值为 $\{\lambda_1,\ldots,\lambda_n\}$，则：

$$\lambda(\alpha I + A) = \{\alpha+\lambda_1,\ldots,\alpha+\lambda_n\}$$

2. $A, B \in \mathbb{R}^{n\times n}$ 为实对称阵则：

$$\lambda_{\max}(A+B) \le \lambda_{\max}(A) + \lambda_{\max}(B)$$

3. 任意实向量 $x \in \mathbb{R}^d$ 有：

$$Rank(xx^\top) = 1 \quad \lambda(x) = \{\|x\|_2^2, 0\}$$

4. 对于实对称阵 $A \in \mathbb{R}^{n\times n}$，实向量 $x \in \mathbb{R}^{n\times n}$：

$$\lambda_{\min}(A)\|x\|_2^2 \le x^\top Ax \le \lambda_{\max}(A)\|x\|_2^2$$

5. 对于可逆矩阵 A，任一特征值 $\lambda(A)$ 有：

$$\lambda(A^{-1}) = \lambda^{-1}(A)$$

引理 A.5 (w_t 的下界)**.**

$$\begin{aligned}
w_t^2 &= \frac{1}{\sigma_t^2}\cdot\sigma_t^\top\left[\lambda I + \sum_{t=1}^{T-1}\frac{\phi_t\phi_t^\top}{\sigma_t^2}\right]^{-1}\phi_t \\
&\ge \frac{\|\phi_t\|_2^2}{\sigma_t^2}\lambda_{\min}\left(\left[\lambda I + \sum_{t=1}^{T-1}\frac{\phi_t\phi_t^\top}{\sigma_t^2}\right]^{-1}\right) \\
&\ge \frac{L_l^2}{K^2}\lambda_{\max}^{-1}\left(\lambda I + \sum_{t=1}^{T-1}\frac{\phi_t\phi_t^\top}{\sigma_t^2}\right) \\
&\ge \frac{L_l^2}{K^2}\frac{1}{\lambda + \sum_{t=1}^{T}\frac{1}{\sigma_t^2}\lambda_{\max}(\phi_t\phi_t^\top)} \\
&\ge \frac{L_l^2}{K^2}\frac{1}{\lambda + \sum_{t=1}^{T}\frac{\|\phi_t\|_2^2}{\min\{\sigma_{\min}^2, v_t^2\}}} \ge \frac{L_l^2}{K^2}\frac{1}{\lambda + \sum_{t=1}^{T}\frac{L_u^2}{\min\{\sigma_{\min}^2, v_t^2\}}}
\end{aligned}$$

其中第二个不等式利用了*A.4*中第 *5* 条性质，第三个不等式利用了 σ 的上界 ϕ_t 的下界及*A.4*中的第 *4* 条性质，第四个不等式分别利用了*A.4*中的第 *1,2* 条性质。最后一个不等式利用了*A.4*中第三条性质。

(二) 引理4.1的证明

通过计算可得：

$$\nabla^2 L_T(\theta) = \lambda I + \sum_{t=1}^{T}\left(\frac{\tau_t}{\sqrt{\tau_t^2 + z_t^2(\theta)}}\right)^3\frac{\phi_t\phi_t^\top}{\sigma_t^2}$$

进而可以将 $\nabla^2 L_T(\theta)$ 拆分为两项：

$$\nabla^2 L_T(\theta) = H_T - \sum_{t=1}^{T}\left[1-\left(\frac{\tau_t}{\sqrt{\tau_t^2+z_t^2(\theta)}}\right)^3\right]\frac{\phi_t\phi_t^\top}{\sigma_t^2} + \sum_{t=1}^{T}\left[\left(\frac{\tau_t}{\sqrt{\tau_t^2+z_t^2(\theta)}}\right)^3 - \left(\frac{\tau_t}{\sqrt{\tau_t^2+z_t^2(\theta^*)}}\right)^3\right]\frac{\phi_t\phi_t^\top}{\sigma_t^2} \tag{25}$$

记上式第二项为 $-\nabla_1^2$，第三项为 ∇_2^2，下面将分别证明 ∇_1^2 的大概率上界和 ∇_2^2 的大概率下界。

∇_1^2 的上界

对于任意的 d 维向量 $v \in \mathbb{R}^d$：

$$\begin{aligned}
v^\top \nabla_1^2 v &= \sum_{t=1}^{T}\left[1-\left(\frac{\tau_t}{\sqrt{\tau_t^2+z_t^2(\theta)}}\right)^3\right]\langle\frac{\phi_t}{\sigma_t}, v\rangle^2 \\
&\le 3\sum_{t=1}^{T}\left[1-\frac{\tau_t}{\sqrt{\tau_t^2+z_t^2(\theta)}}\right]\langle\frac{\phi_t}{\sigma_t}, v\rangle^2 \\
&\le 3\sum_{t=1}^{T}\left[1-\frac{\tau_t}{\sqrt{\tau_t^2+z_t^2(\theta)}}\right]\cdot\sup_{t\in[T]}\langle\frac{\phi_t}{\sigma_t}, v\rangle^2 \\
&\le 3\sum_{t=1}^{T}\left[1-\frac{\tau_t}{\sqrt{\tau_t^2+z_t^2(\theta)}}\right]\cdot\sup_{t\in[T]}\left\|\frac{\phi_t}{\sigma_t}\right\|_{H_t^{-1}}\cdot\|v\|_{H_t}^2 \\
&\le 3\sum_{t=1}^{T}\left[1-\frac{\tau_t}{\sqrt{\tau_t^2+z_t^2(\theta)}}\right]\cdot\sup_{t\in[T]}\left\|\frac{\phi_t}{\sigma_t}\right\|_{H_t^{-1}}\cdot\|v\|_{H_T}^2 \\
&= 3\sum_{t=1}^{T}\left[1-\frac{\tau_t}{\sqrt{\tau_t^2+z_t^2(\theta)}}\right]\cdot\sup_{t\in[T]}\frac{w_t^2}{1+w_t^2}\cdot\|v\|_{H_T}^2
\end{aligned}$$

其中第二个不等式利用了不等式：

$$\forall x\in[0,1]\quad 1-x^3\le 3(1-x)$$

由算法流程图1中，$H_t = \lambda I + \sum_{t=1}^{T}\frac{\phi_t}{\sigma_t}\frac{\phi_t^\top}{\sigma_t}$。因此：

$$\forall t\le T\quad H_T \succeq H_t$$

故第五个不等式成立。最后一个不等式利用了如下的变形：

$$\left\|\frac{\phi_t}{\sigma_t}\right\|_{H_t^{-1}} = \frac{\phi_t^\top}{\sigma_t}\left(H_{t-1}^{-1} - \frac{H_{t-1}^{-1}\frac{\phi_t}{\sigma_t}\frac{\phi_t^\top}{\sigma_t}H_{t-1}^{-1}}{1+\frac{\phi_t^\top}{\sigma_t}H_{t-1}^{-1}\frac{\phi_t}{\sigma_t}}\right)\frac{\phi_t}{\sigma_t} = w_t^2 - \frac{w_t^4}{1+w_t^2} = \frac{w_t^2}{1+w_t^2}$$

由 $v^\top\nabla_1^2 v \le 3\sum_{t=1}^{T}\left[1-\frac{\tau_t}{\sqrt{\tau_t^2+z_t^2(\theta)}}\right]\cdot\sup_{t\in[T]}\frac{w_t^2}{1+w_t^2}\cdot\|v\|_{H_T}^2$ 以及 v 的任意性可知：

$$\nabla_1^2 \preceq 3\sup_{t\in[T]}\frac{w_t^2}{1+w_t^2}\cdot H_T\sum_{t=1}^{T}\left[1-\frac{\tau_t}{\sqrt{\tau_t^2+z_t^2(\theta)}}\right] \tag{26}$$

在上式中，仅有 $=\sum_{t=1}^{T}\left[1-\frac{\tau_t}{\sqrt{\tau_t^2+z_t^2(\theta)}}\right]$ 部分具有随机性。为简化符号，记 $X_t = 1-\frac{\tau_t}{\sqrt{\tau_t^2+z_t^2(\theta)}}$。并采用利用方差信息的弗里德曼集中不等式A.2获得 $\sum_{t=1}^{T}X_t$ 的大概率上界。为利用弗里德曼不等式，

需分别给出 $|X_t|$，$\sum_{t=1}^{T}\mathbb{E}[X_t]$，以及 $\sum_{t=1}^{T}\mathbb{V}_t[X_t]$ 的上界。显然有 $|X_t|\leq 1$，其次对于任意轮数 t：

$$\begin{aligned}\mathbb{E}_t X_t &= \mathbb{E}_t\left[1-\frac{\tau_t}{\sqrt{\tau_t^2+z_t^2(\theta)}}\right]\\ &= \mathbb{E}_t\left[\frac{z_t^2(\theta^*)}{\sqrt{\tau_t^2+z_t^2(\theta^*)}(\sqrt{\tau_t^2+z_t^2(\theta^*)}+\tau_t)}\right]\\ &\leq \mathbb{E}_t\left[\frac{z_t^p(\theta^*)z_t^{2-p}(\theta^*)}{z_t^{2-p}(\theta^*)\tau_t^{p-1}\cdot 2\tau_t}\right]\\ &= \frac{1}{2\tau_t^p}\mathbb{E}_t[z_t^p]\leq\frac{b^p}{2\tau_t^p}=\frac{b^p}{2\tau_0^p}\frac{w_t^2}{1+w_t^2}\end{aligned}$$

其中 $\mathbb{E}_t[z_t^p]\leq b^p$ 由引理的假设而来。因此可以获得 $\sum_{t=1}^{T}\mathbb{E}[X_t]$ 的上界：

$$\sum_{t=1}^{T}\mathbb{E}[X_t]\leq\frac{b^p}{2\tau_0^p}\sum_{t=1}^{T}\frac{w_t^2}{1+w_t^2}\leq\frac{b^p}{2\tau_0^p}\sum_{t=}^{T}\min\{1,w_t^2\}\leq\frac{\kappa b^p}{2\tau_0^p}\tag{27}$$

最后由于：

$$\mathbb{V}_t[X_t]\leq\mathbb{E}_t[X_t^2]\leq\mathbb{E}_t[X_t]$$

因此：

$$\sum_{t=1}^{T}\mathbb{V}_t[X_t]\leq\sum_{t=1}^{T}\mathbb{E}[X_t]\leq\frac{\kappa b^p}{2\tau_0^p}$$

在这里，要求 $\tau_0^p\geq\kappa b^p$，即有 $\sum_{t=1}^{T}\mathbb{V}_t[X_t]\leq 1$ 成立。利用A.2，下式以至少 $1-\delta$ 概率成立：

$$\begin{aligned}\sum_{t=1}^{T}X_t &\leq \frac{\kappa b^p}{2\tau_0^p}+3\sqrt{\frac{\kappa b^p\ln\frac{2}{\delta}}{2\tau_0^p}}+5\ln\frac{2}{\delta}\\ &\leq \frac{1}{2}+3\sqrt{\ln\frac{2}{\delta}}+5\ln\frac{2}{\delta}\\ &\leq 9\ln\frac{2}{\delta}\end{aligned}$$

其中最后一个不等式是因为 δ 远小于 1，故可以认为 $\ln\frac{2}{\delta}\geq 1$。设 $c_0=\frac{1}{6\sqrt{\ln\frac{2}{\delta}}}$ 则有：

$$\sum_{t=1}^{T}X_t\leq\frac{1}{4c_0^2}\tag{28}$$

下面证明 $\sup_{t\in[T]}\frac{w_t^2}{1+w_t^2}$ 的上界。通过设置 $\sigma_t\geq\frac{1}{c_0^2}\cdot\|\phi_t\|_{H_{t-1}^{-1}}$，可得：

$$\sup_{t\in[T]}\frac{w_t^2}{1+w_t^2}\leq\sup_{t\in[T]}w_t^2\leq c_0^2\tag{29}$$

将式28，29代入式26可得：

$$\nabla_1^2\preceq\frac{1}{4}H_T\tag{30}$$

∇_2^2 的下界：首先给出 $\left[\left(\frac{\tau_t}{\sqrt{\tau_t^2+z_t^2(\theta)}}\right)^3-\left(\frac{\tau_t}{\sqrt{\tau_t^2+z_t^2(\theta^*)}}\right)^3\right]$ 的一个对任意轮数 t 成立的下界。而

这个下界可以通过证明 $\left|\left(\frac{\tau_t}{\sqrt{\tau_t^2+z_t^2(\theta)}}\right)^3-\left(\frac{\tau_t}{\sqrt{\tau_t^2+z_t^2(\theta^*)}}\right)^3\right|$ 的上界得到。

$$\begin{aligned}&\left|\left(\frac{\tau_t}{\sqrt{\tau_t^2+z_t^2(\theta)}}\right)^3-\left(\frac{\tau_t}{\sqrt{\tau_t^2+z_t^2(\theta^*)}}\right)^3\right|\leq 3\left|\left(\frac{\tau_t}{\sqrt{\tau_t^2+z_t^2(\theta)}}\right)-\left(\frac{\tau_t}{\sqrt{\tau_t^2+z_t^2(\theta^*)}}\right)\right|\\&\leq\frac{3\tau_t}{\sqrt{\tau_t^2+z_t^2(\theta)}\sqrt{\tau_t^2+z_t^2(\theta^*)}}\frac{|z_t^2(\theta)-z_t^2(\theta^*)|}{\sqrt{\tau_t^2+z_t^2(\theta)}+\sqrt{\tau_t^2+z_t^2(\theta^*)}}\end{aligned}\tag{31}$$

上式可由简单的代数运算得出。进一步，由 $z_t(\theta)=$ 的定义21，可知 $z_t(\theta)=z_t(\theta^*)+\langle\frac{\phi_t}{\sigma_t},\theta-\theta^*\rangle$，因此对于任意的 $c>0$：

$$\begin{aligned}z_t^2(\theta)&\leq\left(1+\frac{1}{c}\right)z_t^2(\theta^*)+(1+c)\langle\frac{\phi_t}{\sigma_t},\theta-\theta^*\rangle^2\\z_t^2(\theta^*)&\leq\left(1+\frac{1}{c}\right)z_t^2(\theta)+(1+c)\langle\frac{\phi_t}{\sigma_t},\theta-\theta^*\rangle^2\end{aligned}$$

进而通过比较 $z_t^2(\theta^*)$ 与 $z_t^2(\theta)$ 大小，有：

$$|z_t^2(\theta)-z_t^2(\theta^*)|\leq\frac{1}{c}\min\{z_t^2(\theta),z_t^2(\theta^*)\}+(1+c)\langle\frac{\phi_t}{\sigma_t},\theta-\theta^*\rangle^2\tag{32}$$

将32式代入31，可得：

$$\begin{aligned}&\left|\left(\frac{\tau_t}{\sqrt{\tau_t^2+z_t^2(\theta)}}\right)^3-\left(\frac{\tau_t}{\sqrt{\tau_t^2+z_t^2(\theta^*)}}\right)^3\right|\\&\leq\frac{3\tau_t}{\tau_t^2+\min\{z_t^2(\theta),z_t^2(\theta^*)\}}\frac{\frac{1}{c}\min\{z_t^2(\theta),z_t^2(\theta^*)\}}{2\sqrt{\tau_t^2+\min\{z_t^2(\theta),z_t^2(\theta^*)\}}}+\frac{3(1+c)}{2\tau_t^2}\langle\frac{\phi_t}{\sigma_t},\theta-\theta^*\rangle^2\\&\leq\frac{3}{2c}+\frac{3(1+c)}{2\tau_t^2}\langle\frac{\phi_t}{\sigma_t},\theta-\theta^*\rangle^2\\&\leq\frac{3}{2c}+\frac{6(1+c)}{2\tau_t^2}\frac{L^2B^2}{\sigma_t^2}\\&\leq\frac{3}{2c}+\frac{6(1+c)}{2\tau_t^2}\frac{w_t^2L^2B^2}{\sigma_t^2}\leq\frac{3}{2c}+\frac{(1+c)d}{4\times7\tau_0^2}\end{aligned}\tag{33}$$

其中第三个不等式利用了柯西不等式以及 $\|\phi_t\|\leq L_u$，$\theta,\theta^*\in Ball_d(B)$：

$$\langle\frac{\phi_t}{\sigma_t},\theta-\theta^*\rangle\leq\|\frac{\phi_t}{\sigma_t}\|_2(\|\theta\|_2+\|\theta^*\|_2)\leq\frac{2L_uB}{\sigma_t}$$

最后一个不等式则利用了 $\sigma_t\geq168\sqrt{L_uB}\|\phi_t\|_{H_{t-1}}^{1/2}/d^{1/4}$，因此：

$$\nabla_2^2\succeq\sum_{t=1}^T\left[\left(\frac{\tau_t}{\sqrt{\tau_t^2+z_t^2(\theta)}}\right)^3-\left(\frac{\tau_t}{\sqrt{\tau_t^2+z_t^2(\theta^*)}}\right)^3\right]\frac{\phi_t\phi_t^\top}{\sigma_t^2}\succeq-\left(\frac{3}{2c}+\frac{(1+c)d}{4\times7\tau_0^2}\right)\sum_{t=1}^T\frac{\phi_t\phi_t^\top}{\sigma_t^2}\tag{34}$$

进一步，令 $c=6$，$\tau_0\geq\sqrt{d}$，有：

$$\nabla_2^2\succeq-\left(\frac{1}{4}+\frac{1}{4}\right)\sum_{t=1}^T\frac{\phi_t\phi_t^\top}{\sigma_t^2}=-\frac{1}{2}\sum_{t=1}^T\frac{\phi_t\phi_t^\top}{\sigma_t^2}\tag{35}$$

将 ∇_1^2, ∇_2^2 代入 $\nabla^2 L_T(\theta)$:

$$\begin{aligned}\nabla^2 L_T(\theta) &\succeq H_T - \frac{1}{4}H_T - \frac{1}{2}\sum_{t=1}^{T}\frac{\phi_t\phi_t^\top}{\sigma_t^2} \\ &\succeq \left(1-\frac{1}{4}\right)\lambda I + (1-\frac{1}{4}-\frac{1}{2})\frac{\phi_t\phi_t^\top}{\sigma_t^2} \\ &\succeq \frac{1}{4}\lambda I + \frac{1}{4}\frac{\phi_t\phi_t^\top}{\sigma_t^2} = \frac{1}{4}H_T\end{aligned}$$

(三) 引理4.2的证明

通过计算可得:

$$\nabla L_T(\theta^*) = \lambda\theta^* - \sum_{t=1}^{T}\frac{\tau_t z_t(\theta)}{\sqrt{\tau_t^2 + z_t(\theta^*)^2}}\frac{\phi_t}{\sigma_t}$$

因此利用三角不等式有:

$$\|\nabla L_T(\theta^*)\|_{H_T^{-1}} \le \|\lambda\theta^*\|_{H_T^{-1}} + \left\|\sum_{t=1}^{T}\frac{\tau_t z_t(\theta^*)}{\sqrt{\tau_t^2 + z_t(\theta^*)^2}}\frac{\phi_t}{\sigma_t}\right\|_{H_T^{-1}} \tag{36}$$

记 $d_T = \sum_{t=1}^{T}\frac{\tau_t z_t(\theta^*)}{\sqrt{\tau_t^2+z_t(\theta^*)^2}}\frac{\phi_t}{\sigma_t}$，下面首先证明 $\|d_T\|_{H_T^{-1}}$，即36式中第二项的大概率上界。利用 Woodbury 矩阵等式有:

$$\begin{aligned}H_T^{-1} &= \left(H_{T-1} + \frac{\phi_t\phi_t^\top}{\sigma_T^2}\right)^{-1} \\ &= H_{T-1}^{-1} - H_{T-1}^{-1}\frac{\phi_t}{\sigma_t}(1 + \frac{\phi_t^\top}{\sigma_t}H_{T-1}\frac{\phi_t}{\sigma_t})^{-1}\frac{\phi_t^\top}{\sigma_t}H_{T-1}^{-1} \\ &= H_{T-1}^{-1} - \frac{H_{T-1}^{-1}\phi_T\phi_T^\top H_{T-1}^{-1}}{\sigma_T^2(1+w_T^2)}\end{aligned} \tag{37}$$

因此:

$$\|d_T\|_{H_T^{-1}}^2 = \left(d_{T-1} + \frac{\tau_T z_T^*}{\sqrt{\tau_T^2 + (x_T^*)^2}}\right)^\top H_T^{-1}\left(d_{T-1} + \frac{\tau_T z_T^*}{\sqrt{\tau_T^2 + (x_T^*)^2}}\right)$$

将37代入，并化简有:

$$\begin{aligned}\|d_T\|_{H_T^{-1}}^2 &= \|d_{T-1}\|_{H_{T-1}^{-1}}^2 - \frac{1}{1+w_t^2}\left(\frac{d_{T-1}^\top H_{T-1}^{-1}\phi_T}{\sigma_T}\right)^2 \\ &\quad + \frac{2\tau_T z_T^*}{\sqrt{\tau_T^2 + (z_T^*)^2}}\frac{d_{T-1}^\top H_T^{-1}\phi_T}{\sigma_T} + \frac{\tau_T^2(z_T^*)^2}{\tau_T^2 + (z_T^*)^2}\frac{\phi_T^\top H_T^{-1}\phi_T}{\sigma_T^2} \\ &\le \|d_{T-1}\|_{H_{T-1}^{-1}}^2 + \frac{2\tau_T z_T^*}{\sqrt{\tau_T^2 + (z_T^*)^2}}\frac{d_{T-1}^\top H_T^{-1}\phi_T}{\sigma_T} + \frac{\tau_T^2(z_T^*)^2}{\tau_T^2 + (z_T^*)^2}\frac{\phi_T^\top H_T^{-1}\phi_T}{\sigma_T^2}\end{aligned} \tag{38}$$

将式37代入式38，进一步化简有:

$$\|d_T\|_{H_T^{-1}}^2 \le \|d_{T-1}\|_{H_{T-1}^{-1}}^2 + \frac{2\tau_T z_T^*}{\sqrt{\tau_T^2 + (z_T^*)^2}}\frac{1}{1+w_T^2}\frac{d_{T-1}^\top H_{T-1}^{-1}\phi_T}{\sigma_T} + \frac{\tau_T^2(z_T^*)^2}{\tau_T^2 + (z_T^*)^2}\frac{w_t^2}{1+w_t^2}$$

通过对 t 迭代有:

$$\|d_T\|^2_{H_T^{-1}} \le \sum_{t=1}^{T} \frac{2\tau_t z_t^*}{\sqrt{\tau_t^2+(z_t^*)^2}} \frac{1}{1+w_t^2} \frac{d_{t-1}^\top H_{t-1}^{-1}\phi_t}{\sigma_t} + \sum_{t=1}^{T} \frac{\tau_t^2(z_t^*)^2}{\tau_t^2+(z_t^*)^2} \frac{w_t^2}{1+w_t^2}$$

代入式36，并利用 $\|\lambda\theta^*\|_{H_T^{-1}} \le \sqrt{\lambda}\|\theta^*\|_{\lambda^{-1}I} \le \sqrt{\lambda}B$ 有:

$$\|d_T\|^2_{H_T^{-1}} \le \sqrt{\lambda}B + \sum_{t=1}^{T} \frac{2\tau_t z_t^*}{\sqrt{\tau_t^2+(z_t^*)^2}} \frac{1}{1+w_t^2} \frac{d_{t-1}^\top H_{t-1}^{-1}\phi_t}{\sigma_t} + \sum_{t=1}^{T} \frac{\tau_t^2(z_t^*)^2}{\tau_t^2+(z_t^*)^2} \frac{w_t^2}{1+w_t^2} \tag{39}$$

下面采用数学归纳法给出 $\|d_T\|_{H_T^{-1}}$ 具有大概率上界:

$$\|d_T\|_{H_T^{-1}} \le \sqrt{\lambda}B + 8\left[\frac{b^p\kappa\eta^{\frac{2-p}{p}}}{\tau_0^p} + \eta^{\frac{2-p}{p}}\sqrt{\kappa\tau_0^{2-p}b^p\ln\frac{1}{\delta}} + \tau_0\eta^{\frac{2-p}{p}}\ln\frac{1}{\delta}\right] = \alpha_T$$

记随机事件 $A_n = \{\|d_n\|_{H_n^{-1}} \le \alpha_n\}$，其中 $\alpha_0 = 0$。则有 $\|d_T\|_{H_T^{-1}} = 0 = \alpha_0$，即 A_0 成立。

进而假设事件 $A_0 \dots A_{T-1}$ 均成立，尝试证明 A_T 成立。为简化符号，将39第二项记为 ∇_1^1，第三项纪为 ∇_2^1，下文将在 $A_0 \dots A_{T-1}$ 成立条件下分别证明 ∇_1^1 和 ∇_2^1 的大概率上界，进而证明 $\|d_T\|_{H_T^{-1}} \le \alpha_T$ 大概率成立。

事件 $A_0 \dots A_{T-1}$ 成立时 ∇_1^1 的上界:

$$\nabla_1^1 = \sum_{t=1}^{T} \frac{2\tau_t z_t^*}{\sqrt{\tau_t^2+(z_t^*)^2}} \frac{1_{A_{t-1}}}{1+w_t^2} \frac{d_{t-1}^\top H_{t-1}^{-1}\phi_t}{\sigma_t}$$

记 $X_t = \frac{2\tau_t z_t^*}{\sqrt{\tau_t^2+(z_t^*)^2}} \frac{1_{A_{t-1}}}{1+w_t^2}$，$Y_t = X_t - \mathbb{E}_t[X_t]$。尝试利用弗里德曼集中不等式A.1证明 X_t 的大概率上界，即分别证明 $|X_t|$，$\sum_{t=1}^T \mathbb{E}_t[X_t]$ 和 $\sum_{t=1}^T \mathbb{E}_t[X_t^2]$，首先:

$$\left|\frac{1_{A_{t-1}} d_{t-1}^\top H_{t-1}^{-1}\phi_t}{\sigma_t}\right| \le \|1_{A_{t-1}} d_{t-1}\|_{H_{t-1}^{-1}} \cdot \left\|\frac{\phi_t}{\sigma_t}\right\|_{H_{t-1}^{-1}} \le \alpha_{t-1} w_t$$

因此:

$$|X_t| \le \tau_t\alpha_{t-1} \cdot \frac{2w_t}{1+w_t^2} \le \tau_0 \left(\frac{\sqrt{1+w_t^2}}{w_t}\right)^{\frac{2}{p}} \cdot \frac{2w_t}{1+w_t^2} \cdot \alpha_{t-1} \le 2\tau_0 \left(\frac{\sqrt{1+w_t^2}}{w_t}\right)^{\frac{2-p}{p}} \alpha_{t-1}$$

利用A.5有:

$$\frac{\sqrt{1+w_t^2}}{w_t} = \sqrt{1+\frac{1}{w_t^2}} \le \sqrt{1+\frac{K^2}{L_l^2}(\lambda + \sum_{t=t}^{T} \frac{L_u^2}{\min\{\sigma_{\min}^2, v_t^2\}})} \tag{40}$$

为简化符号，定义 $\eta = \sqrt{1+\frac{K^2}{L_l^2}(\lambda + \sum_{t=t}^{T} \frac{L_u^2}{\min\{\sigma_{\min}^2, v_t^2\}})}$，故:

$$|X_t| \le 2\tau_0\eta^{\frac{2-p}{p}}\alpha_{t-1}$$

同样有：

$$\begin{aligned}\mathbb{E}_t[Y_t^2] &\le \mathbb{E}_t[X_t^2] \\ &= \mathbb{E}_t\left[\left(\frac{2w_t}{1+w_t^2}\right)^2 \cdot 1_{A_{t-1}} d_{t-1}\alpha_{t-1}\frac{\tau_t^2(z_t^*)^2}{\tau_t^2+(z_t^*)^2}\right] \\ &\le \left(\frac{2w_t}{1+w_t^2}\right)^2 \alpha_{t-1}^2 \mathbb{E}_t\left[\frac{|z_t^*|^p|z_t^*|s^{2-p}\tau_t^2}{\tau_t^2+(z_t^*)^2}\right] \\ &\le \left(\frac{2w_t}{1+w_t^2}\right)^2 \alpha_{t-1}^2 b^p \tau_t^{2-p} \\ &\le \min\{1,2w_t\}^2 \alpha_{t-1}^2 b^p \tau_0^{2-p}\left(\frac{\sqrt{1+w_t^2}}{w_t}\right)^{\frac{2(2-p)}{p}}\end{aligned}$$

其中第四个不等式利用了 $\mathbb{E}_t[|z_t^*|^p] \le b^p$。利用引理A.3，有：

$$\sum_{t=1}^T \mathbb{E}_t[X_t^2] \le 4\sum_{t=1}^T \min\{1,w_t^2\}\alpha_{t-1}^2\tau_0^{2-p}b^p\eta^{\frac{2(2-p)}{p}} \le 4\kappa\tau_0^{2-p}b^p\eta^{\frac{2(2-p)}{p}}\max_{t\in[T]}\alpha_t^2$$

另一方面，利用 $\mathbb{E}_t[z_t^*]=0$，有：

$$\begin{aligned}&\left|\mathbb{E}_t\left[\frac{\tau_t z_t^*}{\sqrt{\tau_t^2+(z_t^*)^2}}\right]\right| = \left|\mathbb{E}_t\left[\left(\frac{\tau_t}{\sqrt{\tau_t^2+(z_t^*)^2}}-1\right)z_t^*\right]\right| \\ &\le \mathbb{E}_t\left[\frac{(z_t^*)^3}{\sqrt{\tau_t^2+(z_t^*)^2}\left(\tau_t+\sqrt{\tau_t^2+(z_t^*)^2}\right)}\right] \le \frac{b^p}{\tau_t^{p-1}}\end{aligned}$$

因此：

$$\begin{aligned}\left|\sum_{t=1}^T \mathbb{E}_t[X_t]\right| &\le \sum_{t=1}^T \frac{b^p}{\tau_t^{p-1}}\frac{w_t}{1+w_t^2}\alpha_{t-1} \\ &\le \sup_{t\in[T]}\alpha_t\frac{b^p}{\tau_0^p}\sum_{t=1}^T\left(\frac{w_t}{\sqrt{1+w_t^2}}\right)^{\frac{2(p-1)}{p}}\frac{w_t}{1+w_t^2} \\ &\le \sup_{t\in[T]}\alpha_t\frac{b^p}{\tau_0^p}\sum_{t=1}^T\left(\frac{\sqrt{1+w_t^2}}{w_t}\right)^{\frac{2-p}{p}}\frac{w_t^2}{1+w_t^2} \\ &\le \sup_{t\in[T]}\alpha_t\frac{b^p}{\tau_0^p}\kappa\eta^{\frac{2-p}{p}}\end{aligned}$$

其中最后一个不等式因为 $\sum_{t=1}^T \frac{w_t^2}{1+w_t^2} \le \sum_{t=1}^t \min\{1,w_t^2\} \le \kappa$，$\frac{\sqrt{1+w_t^2}}{w_t} \le \eta$。利用弗里德曼集中不等式A.1，下式以至少 $1-\delta$ 概率成立：

$$\begin{aligned}\nabla_1^1 = \sum_{t=1}^T X_t &\le \left|\sum_{t=1}^T \mathbb{E}_t[X_t]\right| + 4\sup_{t\in[T]}\alpha_t\left(\sqrt{\kappa\tau_0^{2-p}b^p\eta^{\frac{2(2-p)}{p}}\ln\frac{1}{\delta}}+\frac{\tau_0\eta^{\frac{2-p}{p}}}{3}\ln\frac{1}{\delta}\right) \\ &\le 4\sup_{t\in[T]}\alpha_t\left(\frac{b^p\kappa\eta^{\frac{2-p}{p}}}{4\tau_0^p}+\eta^{\frac{2-p}{p}}\sqrt{\kappa\tau_0^{2-p}b^p\ln\frac{1}{\delta}}+\frac{\tau_0\eta^{\frac{2-p}{p}}}{3}\ln\frac{1}{\delta}\right)\end{aligned}$$

事件 $A_0 \dots A_{T-1}$ 成立时 ∇_2^1 的上界：

$$\nabla_2^1 = \sum_{t=1}^{T} \frac{\tau_t^2 (z_t^*)^2}{\tau_t^2 + (z_t^*)^2} \frac{w_t^2}{1+w_t^2}$$

记 $M_t = \frac{\tau_t^2 (z_t^*)^2}{\tau_t^2 + (z_t^*)^2} \frac{w_t^2}{1+w_t^2}, N_t = M_t - \mathbb{E}_t[M_t]$。首先证明 $|N_t|$ 的上界：

$$|M_t| \le \tau_t^2 \frac{w_t^2}{1+w_t^2} \le \tau_0^2 \quad |N_t| \le \max\{|M_t|, |\mathbb{E}_t[M_t]|\} \le \tau_0^2$$

之后证明 $\sum_{t=1}^{T} \mathbb{E}_t[N_t^2]$ 的上界：

$$\begin{aligned}
\mathbb{E}_t[N_t^2] \le \mathbb{E}_t[M_t^2] &\le \left(\frac{w_t^2}{1+w_t^2}\right)^2 \mathbb{E}_t\left[\left(\frac{\tau_t^2 (z_t^*)^2}{\tau_t^2 + (z_t^*)^2}\right)^2\right] \\
&\le \left(\frac{w_t^2}{1+w_t^2}\right)^2 \mathbb{E}_t[(z_t^*)^p (\tau_t^*)^{4-p}] \\
&\le \tau_0^{4-p} b^p \left(\frac{\sqrt{1+w_t^2}}{w_t}\right)^{\frac{2(4-p)}{p}} \left(\frac{w_t^2}{1+w_t^2}\right)^2 \\
&\le \tau_0^{4-p} b^p \left(\frac{\sqrt{1+w_t^2}}{w_t}\right)^{\frac{4(2-p)}{p}} \frac{w_t^2}{1+w_t^2} \\
&\le \tau_0^{4-p} b^p \eta^{\frac{4(2-p)}{p}} \frac{w_t^2}{1+w_t^2}
\end{aligned}$$

其中第三个不等式利用了 $\mathbb{E}_t[|z_t^*|^p] \le b^p$，因此：

$$\sum_{t=1}^{T} \mathbb{E}_t[N_t^2] \le \tau_0^{4-p} b^p \eta^{\frac{4(2-p)}{p}} \sum_{t=1}^{T} \min\{1, w_t^2\} \le \tau_0^{4-p} b^p \eta^{\frac{4(2-p)}{p}} \kappa$$

上式成立是因为，$\sum_{t=1}^{T} \frac{w_t^2}{1+w_t^2} \le \sum_{t=1}^{t} \min\{1, w_t^2\} \le \kappa$。最后利用类似技术，证明 $\sum_{t=1}^{T} \mathbb{E}_t[N_t]$ 的上界：

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{E}_t[N_t] &\le \sum_{t=1}^{T} \frac{w_t^2}{1+w_t^2} \mathbb{E}_t[\frac{\tau_t^2 (z_t^*)^2}{\tau_t^2 + (z_t^*)^2}] \\
&\le \sum_{t=1}^{T} \frac{w_t^2}{1+w_t^2} b^p \tau_t^{2-p} \\
&\le \sum_{t=1}^{T} \frac{w_t^2}{1+w_t^2} b^p \tau_0^{2-p} \left(\frac{\sqrt{1+w_t^2}}{w_t}\right)^{\frac{2(2-p)}{p}} \\
&\le \kappa b^p \tau_0^{2-p} \eta^{\frac{2(2-p)}{p}}
\end{aligned}$$

利用弗里德曼集中不等式A.1，下式以至少 $1-\delta$ 概率成立：

$$\begin{aligned}\nabla_2^1 = \sum_{t=1}^{T} N_t &\le \left|\sum_{t=1}^{T} \mathbb{E}_t[X_t]\right| + \frac{2\tau_0^2}{3}\ln\frac{1}{\delta} + \sqrt{2\kappa\tau_0^{4-p}\eta^{\frac{4(2-p)}{p}} b^p \ln\frac{1}{\delta}} \\ &\le \kappa b^p \tau_0^{2-p}\eta^{\frac{2(2-p)}{p}} + \frac{2\tau_0^2}{3}\ln\frac{1}{\delta} + \sqrt{2\kappa\tau_0^{4-p}\eta^{\frac{4(2-p)}{p}} b^p \ln\frac{1}{\delta}} \\ &\le \kappa b^p \tau_0^{2-p}\eta^{\frac{2(2-p)}{p}} + \tau_0^2 \ln\frac{1}{\delta}\cdot\eta^{\frac{2(2-p)}{p}} + 2\sqrt{\kappa\tau_0^{4-p}\eta^{\frac{4(2-p)}{p}} b^p \ln\frac{1}{\delta}} \\ &= \left(\eta^{\frac{2-p}{p}}\sqrt{\kappa b^p \tau_0^{2-p}} + \eta^{\frac{2-p}{p}}\sqrt{\tau_0^p \ln\frac{1}{\delta}}\right)^2\end{aligned}$$

证明 $\|d_T\|_{H_T^{-1}} \le \alpha_T$：

之前在事件 $A_0 \dots A_{T-1}$ 成立条件下分别证明了：

$$\nabla_1^1 \le 4\sup_{t\in[T]}\alpha_t\left(\frac{b^p\kappa\eta^{\frac{2-p}{p}}}{4\tau_0^p} + \eta^{\frac{2-p}{p}}\sqrt{\kappa\tau_0^{2-p}b^p\ln\frac{1}{\delta}} + \frac{\tau_0\eta^{\frac{2-p}{p}}}{3}\ln\frac{1}{\delta}\right) \tag{41}$$

以及：

$$\nabla_2^1 \le \left(\eta^{\frac{2-p}{p}}\sqrt{\kappa b^p\tau_0^{2-p}} + \eta^{\frac{2-p}{p}}\sqrt{\tau_0^p\ln\frac{1}{\delta}}\right)^2 \tag{42}$$

因此对于任意轮数 t，设：

$$\alpha_t = 8\left[\frac{b^p\kappa\eta^{\frac{2-p}{p}}}{\tau_0^p} + \eta^{\frac{2-p}{p}}\sqrt{\kappa\tau_0^{2-p}b^p\ln\frac{1}{\delta}} + \tau_0\eta^{\frac{2-p}{p}}\ln\frac{1}{\delta}\right] \tag{43}$$

首先，注意到 α_t 中仅有 κ，η 两项与 t 有关。一方面 κ，η 随 t 单调递增，另一方面 κ，η 两项在 α_t 中为系数幂次均为正的多项式项。因此 α_t 随 t 单调递增，故对于任意的 T 有：

$$\max_{t\in[T]}\alpha_t = \alpha_T \tag{44}$$

另一方面，由于 $\tau_0 \ge \sqrt{d} \ge 1$，$\ln\frac{1}{\delta} \ge 1$，有：

$$\nabla_1^1 \le \frac{\alpha_T^2}{2} \quad \nabla_2^1 \le \frac{\alpha_T^2}{2}$$

因此：

$$\|d_T\|_{H_T^{-1}} = \sqrt{\nabla_1^1 + \nabla_2^1} \le \alpha_T$$

因此通过上述证明，在 $A_0, \dots, A_{T-1}$ 成立的假设下证明了 A_T 成立，进而完成数学归纳法。即任意 T，以至少 $1-\delta$ 的概率有：

$$\|d_T\|_{H_T^{-1}} \le \alpha_T \tag{45}$$

进而有：

$$\nabla L_T(\theta^*) \le \|\lambda\theta^*\|_{\lambda I} + \alpha_T \le \sqrt{\lambda}B + 8\left[\frac{b^p\kappa\eta^{\frac{2-p}{p}}}{\tau_0^p} + \eta^{\frac{2-p}{p}}\sqrt{\kappa\tau_0^{2-p}b^p\ln\frac{1}{\delta}} + \tau_0\eta^{\frac{2-p}{p}}\ln\frac{1}{\delta}\right] \tag{46}$$

（四） 引理4.3的证明

因为 $w_t = \left\| \frac{\phi_t}{\sigma_t} \right\|_{H_{t-1}^{-1}}$ 且有 $\sigma_t \geq \|\phi_t\|_{H_{t-1}^{-1}}/c_0, c_0 \leq 1$:

$$\sum_{t=1}^{T} \|\phi_t\|_{H_{t-1}^{-1}} = \sum_{t=1}^{T} \sigma_t w_t = \sum_{t=1}^{T} \sigma_t \min\{1, w_t\}$$

由于有:

$$\sigma_t = \max\left\{ v_t, \sigma_{\min}, \frac{\|\phi_t\|_{H_{t-1}^{-1}}}{c_0}, \frac{168\sqrt{L_u B}\|\phi_t\|_{H_{t-1}^{-1}}^{1/2}}{d^{\frac{1}{4}}} \right\}$$

根据 σ_t 的取值，将 $[T]$ 分为三个子集，其中:

$$\mathcal{P}_1 = \{t \in [T] : \sigma_t \in \{v_t, \sigma_{\min}\}\}, \mathcal{P}_2 = \left\{ t \in [T] : \sigma_t = \frac{\|\phi_t\|_{H_{t-1}^{-1}}}{c_0} \right\},$$
$$\mathcal{P}_3 = \left\{ t \in [T] : \sigma_t = \frac{168\sqrt{L_u B}\|\phi_t\|_{H_{t-1}^{-1}}^{1/2}}{d^{\frac{1}{4}}} \right\}$$

首先对于 $t \in \mathcal{P}_1$ 有:

$$\begin{aligned}
\sum_{t\in\mathcal{P}_1} \sigma_t \min\{1, w_t\} &\leq \sum_{t=1}^{T} \max\{v_t, \sigma_{\min}\} \min\{1, w_t\} \\
&\leq \sqrt{\sum_{t=1}^{T} (v_t^2 + \sigma_{\min}^2)} \sqrt{\sum_{t=1}^{T} \min\{1, w_t^2\}} \\
&\leq \sqrt{\kappa} \cdot \sqrt{\sum_{t=1}^{T} v_t^2 + T\sigma_{\min}^2}
\end{aligned} \tag{47}$$

对于 $t \in \mathcal{P}_2$ 有 $w_t = c_0 \leq 1$，因此:

$$\begin{aligned}
\sum_{t\in\mathcal{P}_2} \sigma_t \min\{1, w_t\} &= \sum_{t\in\mathcal{P}_2} \sigma_t w_t^2 = \frac{\sup_{t\in\mathcal{P}_2} \sigma_t}{c_0} \sum_{t\in\mathcal{P}_2} w_t^2 \\
&\leq \frac{\sup_{t\in[T]} \|\phi_t\|_{H_{t-1}^{-1}}}{c_0^2} \cdot \sum_{t\in[T]} \min\{1, w_t^2\} \\
&\leq \frac{L_u \kappa}{c_0^2 \sqrt{\lambda}}
\end{aligned} \tag{48}$$

其中最后一个不等式因为 $\|\phi_t\|_{H_{t-1}^{-1}} \leq \frac{1}{\sqrt{\lambda}}\|\phi_t\|_2 \leq \frac{L_u}{\sqrt{\lambda}}$

最后对于 $t \in \mathcal{P}_3$,有 $L^2 B^2 w_t^2 168^4 = d\sigma_t^2$,进而可得 $\sigma_t = L_u B w_t / 168^2\sqrt{d} = L_u B \min\{1, w_t\}/168^2\sqrt{d}$, 进而有:

$$\sum_{t\in[T]} \sigma_t \min\{1, w_t\} = \frac{L_u B}{168^2\sqrt{d}} \sum_{t\in[T]} \min\{1, w_t^2\} \leq \frac{L_u B \kappa}{168^2\sqrt{d}} \tag{49}$$

将式47，48和49合并，有：

$$\begin{aligned}\sum_{t=1}^{T}\|\phi_t\|_{H_{t-1}^{-1}} &= \sum_{t=1}^{T}\sigma_t \min\{1, w_t\} \\ &= \sum_{t\in\mathcal{P}_1}\sigma_t \min\{1, w_t\} + \sum_{t\in\mathcal{P}_2}\sigma_t \min\{1, w_t\} + \sum_{t\in\mathcal{P}_3}\sigma_t \min\{1, w_t\} \\ &\leq \sqrt{\kappa}\cdot\sqrt{\sum_{t=1}^{T} v_t^2 + T\sigma_{\min}^2} + \frac{L_u\kappa}{c_0^2\sqrt{\lambda}} + \frac{L_u B\kappa}{168^2\sqrt{d}}\end{aligned}$$

参考文献

[1] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

[2] Djallel Bouneffouf, Irina Rish, and Charu Aggarwal. Survey on applications of multi-armed and contextual bandits. In *2020 IEEE Congress on Evolutionary Computation (CEC)*, pages 1–8. IEEE, 2020.

[3] Giuseppe Burtini, Jason Loeppky, and Ramon Lawrence. A survey of online experiment design with the stochastic multi-armed bandit. *arXiv preprint arXiv:1510.00757*, 2015.

[4] Xiequan Fan. Freedman's inequality with non-bounded martingale differences. *arXiv preprint arXiv:1404.4776*, 2014.

[5] Sarah Filippi, Olivier Cappe, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: The generalized linear case. *Advances in Neural Information Processing Systems*, 23, 2010.

[6] David A Freedman. On tail probabilities for martingales. *the Annals of Probability*, pages 100–118, 1975.

[7] Johannes Kirschner and Andreas Krause. Information directed sampling and bandits with heteroscedastic noise. In *Conference On Learning Theory*, pages 358–384. PMLR, 2018.

[8] Tze Leung Lai, Herbert Robbins, et al. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

[9] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[10] Gen Li, Changxiao Cai, Yuxin Chen, Yuantao Gu, Yuting Wei, and Yuejie Chi. Is q-learning minimax optimal? a tight sample complexity analysis. *arXiv preprint arXiv:2102.06548*, 2021.

[11] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670, 2010.

[12] Xiang Li and Qiang Sun. Variance-aware robust reinforcement learning with linear function approximation with heavy-tailed rewards. *arXiv preprint arXiv:2303.05606*, 2023.

[13] Andres Munoz Medina and Scott Yang. No-regret algorithms for heavy-tailed linear bandits. In *International Conference on Machine Learning*, pages 1642–1650. PMLR, 2016.

[14] Daniel Russo and Benjamin Van Roy. Eluder dimension and the sample complexity of optimistic exploration. *Advances in Neural Information Processing Systems*, 26, 2013.

[15] Han Shao, Xiaotian Yu, Irwin King, and Michael R Lyu. Almost optimal algorithms for linear stochastic bandits with heavy-tailed payoffs. *Advances in Neural Information Processing Systems*, 31, 2018.

[16] William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

[17] Bo Xue, Guanghui Wang, Yimu Wang, and Lijun Zhang. Nearly optimal regret for stochastic linear bandits with heavy-tailed payoffs. *arXiv preprint arXiv:2004.13465*, 2020.

[18] Dongruo Zhou, Quanquan Gu, and Csaba Szepesvari. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Conference on Learning Theory*, pages 4532–4576. PMLR, 2021.