arXiv:2603.03778v1 [cs.LG] 4 Mar 2026

无奖励的逆上下文老虎机：通过后缀模仿从非平稳学习者中学习

Yuqi Kong Xiao Zhang Weiran Shen 中国人民大学高瓴人工智能学院{kongyuqi,zhangx89,shenweiran}@ruc.edu.cn

摘要

我们研究逆上下文老虎机（ICB）问题：学习者试图优化策略，而无法获取学习者奖励、只能观察其动作的观察者则旨在恢复潜在的问题参数。在学习过程中，学习者的行为会自然地从探索转向利用，产生非平稳动作数据，给观察者带来重大挑战。为解决这一问题，我们提出一种简单有效的框架，称为两阶段后缀模仿。该框架丢弃初始预热阶段的数据，仅使用后续模仿阶段的数据进行经验风险最小化。我们推导出一个预测决策损失界，明确刻画了预热长度选择所引起的偏差-方差权衡。尽管存在严重的信息缺失，我们证明，无奖励观察者仍能达到 $\tilde{O}(1/\sqrt{N})$ 的收敛速率，与完全奖励感知学习者的渐近效率相匹配。这一结果表明，被动观察者仅凭动作就能有效地揭示最优策略，达到与学习者自身相当的表现。

—

1 引言

上下文老虎机已成为不确定环境下序贯决策的基础框架，并在推荐系统和临床试验等领域发挥了重要作用 [21, 5]。经典研究聚焦于设计能够最大化累积奖励的算法。然而，在现代应用中，仅凭性能往往不足以满足需求：智能体可能需要验证系统为何偏好某些动作而非其他动作，以及系统是否仍在探索。当奖励存在延迟、属于隐私信息或无法观测时，这一问题尤其棘手，因为此时交互日志是唯一的记录。我们将这一问题称为逆上下文老虎机：给定由某种老虎机算法生成的交互历史，观察者能否在无法获取奖励信号的情况下，恢复其背后的最优策略?

这一视角与逆向强化学习（IRL）和行为克隆（BC）相关，它们同样旨在从示范中解释行为。然而，一个根本性的错配使标准逆向强化学习方法难以奏效。传统的逆向强化学习通常假设决策由平稳最优专家生成。相比之下，老虎机日志由一个不断适应和演进的学习智能体生成，该智能体通过学习过程从一无所知逐步积累经验。此处若应用标准行为克隆算法会产生问题：它们不加区分地将这些初期探索性决策视为最优示范，进而模仿含噪且次优的数据。

“学习一个学习者”这一范式提出了两个核心技术挑战。首先，观察者以无奖励方式运行：它观测上下文和学习智能体所选的动作，

但无法观测实际奖励。其次，诱导数据集是非平稳的：随着学习者不断学习，标签质量会随时间变化，因此早期轮次的数据噪声可能远高于后期。朴素地汇总所有观测数据，可能会让早期低质量数据占据主导地位，进而降低恢复的策略的质量。

为应对这些挑战，我们提出了一个简单而有效的两阶段后缀模仿框架。其核心洞见反直觉却至关重要：更少的数据可能是更好的数据。因为忽略探索前缀可以显著提高模仿的信噪比。我们的理论分析揭示了一个出人意料且令人鼓舞的结果。我们证明，即使在严格的信息缺失下，无奖励观察者仍能达到与学习者相当的遗憾界性能（推论1）。

我们的主要贡献总结如下：

- 我们形式化了逆上下文老虎机设定，其中观察者在没有奖励反馈的情况下从一个非平稳学习者那里进行学习。我们提出了两阶段后缀模仿框架，引入一种丢弃策略，以应对探索所导致的分布偏移。

- 我们对所提出的方法进行了有限样本分析，表明观察者能够以趋于零的预测遗憾恢复最优策略。至关重要的是，我们证明了观察者的效率对于学习者的遗憾率具有稳健性。

- 我们进行了大量实验，以验证我们的框架。我们发现，尽管信息严重缺失，观察者的参数估计误差仍会收敛到与完全具备奖励感知能力的学习者相当的水平。

2 相关工作

我们的工作处于上下文老虎机、从示范中进行逆向学习以及弱监督学习的交叉领域。我们着重说明，逆向情境赌博机（ICB）的设定与标准表述有何不同：观察者只能看到情境–动作日志（不包含奖励），而示范由一个因探索而表现出非平稳行为的学习智能体生成。

上下文老虎机。上下文老虎机对带有辅助信息和赌博机反馈的重复决策进行建模。LinUCB [14], 、适用于线性模型的汤普森采样（LinTS） [3], 以及 OFUL [1] 等经典算法，在标准假设下展现出优异的实证表现，并保证无遗憾 [14, 13]。这类研究假设决策者能够观测奖励，并侧重于设计能够最大限度降低遗憾的策略。与之不同，我们研究的是逆向问题：给定此类学习者生成的情境–动作轨迹，外部观测者旨在恢复一条可部署的决策规则，使其在相同的最优性基准下表现良好——且无需任何奖励反馈。

逆强化学习与模仿学习。 ICB 与逆向强化学习（IRL）和模仿学习（IL）相关，这些方法旨在根据示范来解释或复现行为 [18, 2, 25]。一个关键区别在于信息结构和数据生成过程。许多 IRL/IL 方法假设示范由一个稳定的（近似）最优专家生成，并且通常利用奖励、环境动态或额外监督来处理次优性 [26, 19, 15]。尽管近期研究通过从带噪声或不完美的示范中学习来放宽最优性假设 [6, 10, 24], ，许多方法仍然需要奖励标注 [22] 或针对轨迹的外部偏好/排序信号 [12, 7]。我们的

设定严格无奖励：观察者只能看到交互日志。我们不依赖奖励或排序，而是利用“学习一个学习者”所特有的一种结构性特征：随着探索逐渐让位于更稳定的行为，标签质量会随时间提升，从而促成基于后缀的模仿。

逆向上下文老虎机与行为演变。 越来越多关于逆向上下文老虎机的研究旨在恢复行为如何随时间演变，并强调非平稳策略及其动态的可解释性 [11, 23]。我们采取更为务实的立场。我们并不重构学习者的完整演变过程，而是专注于从观测轨迹中恢复单一可部署策略。具体而言，我们丢弃初始预热前缀并执行后缀模仿，并为所得策略提供基于泛化的性能保证。目标并非穷尽性能极限，而是表明一种简单、可实施的流程已能在无奖励观测设定中带来强劲且可预测的收益。

3 预备知识

3.1 线性上下文老虎机

我们考虑具有未知参数 $\theta^* \in \mathbb{R}^d$的随机线性上下文老虎机问题。不失一般性，我们假设 $\|\theta^*\|_2 \le 1$。令 $\mathcal{A} = [K]$ 为有限可能臂集。在每一轮次 $t \in [N]$（其中N为时间范围），环境揭示可行臂集 $\mathcal{A}_t \subseteq \mathcal{A}$ 及其关联特征向量 $X_t = \{x_a \in \mathbb{R}^d : a \in \mathcal{A}_t\}$。我们假设对所有 t、a均有 $\|x_a\|_2 \le 1$。我们假设上下文序列 $\{(\mathcal{A}_t, X_t)\}_{t\ge1}$从上下文空间上的分布 D 中独立同分布采样。在观测到上下文后，学习者在轮次 t 选择臂 $a \in \mathcal{A}_t$，随后观测到奖励：$r_t(a) = \langle x_a, \theta^* \rangle + \xi_t$，其中 ξ_t 为噪声项。令 $\mathcal{F}_{t-1}$ 表示在观测 r_t之前的观测历史。我们假设在给定 $\mathcal{F}_{t-1}$的条件下，$\{\xi_t\}_{t\ge1}$ 为条件 R-次高斯，即 $\mathbb{E}[\xi_t|\mathcal{F}_{t-1}] = 0$ 且$\mathbb{E}[\exp(\lambda\xi_t)|\mathcal{F}_{t-1}] \le \exp(\lambda^2R^2/2)$。这是随机线性老虎机文献中的标准假设。轮次 t 的最优臂为任意：$a_t^* \in \arg\max_{a\in\mathcal{A}_t}\langle x_a, \theta^*\rangle$，并列时按固定规则打破。

定义 1. 第 t 轮的瞬时遗憾定义为最优臂与所选臂的期望奖励之差 a_t:

$$\Delta_t(a_t) := \langle x_{a_t^*}, \theta^* \rangle - \langle x_{a_t}, \theta^* \rangle \ge 0.$$

考虑有限臂集合 $\mathcal{A}$ 和固定参数 θ^*，我们引入以下一致间隔条件，以便于分析错误界并确保最优动作的唯一性。

定义 2. 我们将最小间隔 $\Delta_{\min}$ 定义为所有可能上下文下最优臂与最佳次优臂之间的最小奖励差:

$$\Delta_{\min} := \inf_{(\mathcal{A}_t, X_t)\in\mathrm{supp}(\mathcal{D})} \left[\min_{a\in\mathcal{A}_t\setminus\{a_t^*\}} \Delta_t(a) \right].$$

策略 $\pi = \{\pi_t\}_{t=1}^N$ 由决策规则组成：在每一轮次 t，给定当前可行集和特征（$\mathcal{A}_t$，X_t），该规则 π_t 输出可行臂上的一个分布，随后将动作抽取为 $a_t \sim \pi_t(\cdot|\mathcal{A}_t, X_t)$。对于将（$\mathcal{A}_t$，$X_t$）映射为动作 a_t的任意策略 π

在 $\mathcal{A}_t$ 中，将其时域为 N 的 N-轮次遗憾定义为

$$R_N(\pi) := \sum_{t=1}^{N} \left(\langle x_{a_t^*}, \theta^* \rangle - \langle x_{a_t}, \theta^* \rangle \right).$$

3.2 逆向上下文老虎机

我们已经形式化定义了完整奖励反馈下的环境和最优性概念。现在，我们考虑两个智能体与同一环境进行交互的情形。

学习者。学习者可以完整获取奖励。在每一轮 $t = 1, 2, \ldots\ldots, N$中，学习者观察 $(\mathcal{A}_t, X_t)$，选择 $\hat{a}_t \in \mathcal{A}_t$，随后观察奖励 r_t。学习者可以使用任意自适应上下文老虎机算法，且 $\hat{a}_t$ 可能取决于完整的交互历史。

观察者。相比之下，观察者以无奖励方式运行。它是监控学习者行为的外部参与方。具体而言，观察者可以获取学习者生成的上下文-动作对序列，即观察到 $(\mathcal{A}_t, X_t, \hat{a}_t)$，但无法观察奖励 r_t。与学习者不同，观察者在 N 轮交互期间是被动实体，不会直接产生遗憾。观察者试图恢复~环境参数，即估计一个向量 θ，使其与真实参数θ^*一致；这对于理解学习者为何偏好某些动作而非其他动作至关重要。我们通过预测遗憾来量化这种恢复能力：

定义 3（预测遗憾）。对于任意将 $\mathcal{A}_t, X_t$ 映射到 $\mathcal{A}_t$ 中某个臂的确定性策略 π_θ，定义其预测遗憾为：

$$\rho(\pi_\theta) := \mathbb{E}_{(\mathcal{A}_t, X_t) \sim D} \left[\langle x_{a_t^*}, \theta^* \rangle - \langle x_{\pi_\theta(\mathcal{A}_t, X_t)}, \theta^* \rangle \right].$$

观察者不与环境交互。预测遗憾是一种理论上的~构造，用作参数识别的代理指标。较小的 $\rho(\pi_\theta)$ 值意味着，θ 能够成功捕捉与分布 $\mathcal{D}$ 相关的 θ^* 决策边界。

尽管观察者的最终目标是恢复真实的环境参数 θ^*，但由于固有的可识别性问题，该问题从根本上来说是不适定的。在无法获得实际奖励的情况下，通常不可能区分那些诱导出相同臂排序的参数。事实上，整整一个参数等价类都可能同样合理地解释学习者的行为。因此，我们不再要求恢复精确的真实值，而是采用预测遗憾这一更务实的目标，以准确捕捉底层决策边界，并在功能上复现最优策略。

4 两阶段后缀模仿

观察者接收学习者交互历史的完整序列 $D_{\text{obs}}^{1:N} = \{(\mathcal{A}_t, X_t, \hat{a}_t)\}_{t=1}^{N}$。一种朴素策略是利用所有可用数据进行训练。然而，这忽略了学习者行为的非平稳性：在早期阶段，学习者主要进行探索，其动作往往是次优的；而在后期阶段，学习者则收敛至最优策略。为解决这一问题，我们提出一种简单而有效的策略，并将其命名为两阶段模仿策略：观察者有策略地丢弃历史的早期部分，并利用高质量轨迹进行学习。尽管其十分简单，我们在定理 5 中表明，该方法足以获得强有力的理论保证。

形式上，我们基于动态截断点$T(N)$将总时域 N 划分为两个阶段：

第一阶段：预热期。 对于轮次 $t = 1, 2, \ldots\ldots, T(N)$，观察者将数据 $D^{1:T(N)}{}_{\text{obs}}$视为不可靠样本，并将其排除在训练集之外。在此阶段，学习者执行标准的在线无悔算法。通过利用奖励反馈，学习者改进其策略，逐渐趋近最优臂 a_t^*。长度 $T(N)$ 是一个取决于时域 N 的参数。

第二阶段：模仿。 在此阶段，假设学习者已达到足够的准确度（满足假设 1）。观察者收集轮次$t = T(N) + 1, \ldots\ldots, N$期间的数据，观察者收集观测数据集 $D^{T(N)+1:N}{}_{\text{obs}}$，并用其构建无奖励策略 $\tilde{\pi}$。通过专注于第二阶段，我们有效地将从非平稳智能体学习这一复杂问题，转化为一个更简单的问题：从错误率有界的噪声标签中学习。

观察者在严格的信息缺失下仍能达到最优，且从未观察到任何奖励，这似乎有违直觉。然而，我们证明，在极限情形下奖励信号变得冗余，因为学习者收敛后的动作有效地编码了底层的最优决策边界（定理5）。

4.1 学习者算法

我们不将学习者限制于某一特定算法，而是考虑一类一般的自适应ˆ策略 Π。为了形式化学习者的可靠性，我们从统计学习中的Massart噪声条件 [16] 获得启发，该条件要求标签噪声与 $1/2$保持一定距离。将其适配到智能体随时间改进的非平稳设定中，我们引入动态 Massart 噪声条件：

假设 1. 我们假设学习者的表现随时间改进。具体而言，存在一个非增函数 $\eta\colon \mathbb{N} \mapsto [0, 1)$，使得对任意预热期 T，学习者在后续阶段（$t > T$）的错误概率有如下上界：

$$\sup_{t>T} \mathbb{P}(\hat{a}_t \neq a_t^* \mid \mathcal{A}_t, X_t) \leq \eta(T).$$

我们要求预热期 T 足够长，使得 $\eta(T) < 1/2$。

在关于 ξ_t的标准噪声假设下，我们证明这一假设适用于一大类标准上下文老虎机算法。考虑任意一个能够实现次线性累积遗憾的学习者，这是理性智能体的标准要求。假设学习者的策略是一种标准的无遗憾算法。由于每次错误至少会造成 $\Delta_{\min}$的遗憾，因此截至时域 N 的预期错误次数有界，从而平均错误概率也有界。这意味着标准的无遗憾算法自然满足假设1。典型示例包括 LinUCB 和 LinTS。

示例 2（LinUCB）。LinUCB [14] 会为每个臂的奖励维护一个上置信界。众所周知，它能够达到 $\tilde{O}(\sqrt{N})$ 的遗憾界。

示例 3（LinTS）。汤普森采样同样能够达到 $\tilde{O}(\sqrt{N})$的遗憾保证 [3]。因此，它也满足条件1；随着预热期T 增加，错误率会渐消失。

这一假设并非人为约束，而是学习者学习阶段的直接结果。只要学习者能够有效地探索和利用，观察者就可以依赖

算法 1 通过经验风险最小化进行后缀模仿

输入： 总时域 N，预热期 $T(N)$，第二阶段轨迹 $D^{T(N)+1:N}$观察。**输出：** 学习得到的策略 $\tilde{\pi}$。设定 $L(N) \leftarrow N - T(N)$。~ 通过最小化经验风险最小化来计算 θ：

$$\tilde{\theta} \in \arg\min_{\theta \in \mathbb{R}^d} \frac{1}{L(N)} \sum_{t=T(N)+1}^{N} \mathbf{1}[\pi_\theta(\mathcal{A}_t, X_t) \neq \hat{a}_t] \,.$$

~**返回** $\tilde{\pi}(\cdot) = \arg\max_a \langle x_a, \theta \rangle$。

基于第一阶段中日益准确的动作轨迹。

4.2 观察者算法

在第二阶段，观察者被动收集学习者的上下文-动作轨迹 $D^{T(N)+1:N}{}_{\text{obs}} = \{(\mathcal{A}_t, X_t, \hat{a}_t)\}_{t=T(N)+1}^{N}$，其中"标签"$\hat{a}_t$ 是学习者选择的臂。受预热期后的有界噪声模型 $\Pr(\hat{a}_t \neq a_t^* | \mathcal{A}_t, X_t) \leq \eta(T)$ 启发，我们将 $\hat{a}_t$ 视为最优臂的含噪代理，并通过经验风险最小化（ERM）学习策略。

具体而言，我们考虑一种由共享向量 $\theta \in \mathbb{R}^d$ 参数化的线性评分策略：$\pi_\theta(\mathcal{A}_t, X_t) = \arg\max_{a \in \mathcal{A}_t} \langle x_a, \theta \rangle$。观察者通过最小化 $D^{T(N)+1:N}{}_{\text{obs}}$ 上的经验 $0-1$ 模仿损失来拟合 θ。定义 $L(N) = N - T(N)$，损失为：

$$\tilde{\theta} \in \arg\min_{\theta} \frac{1}{L(N)} \sum_{t=T(N)+1}^{N} \mathbf{1}[\pi_\theta(\mathcal{A}_t, X_t) \neq \hat{a}_t],$$

其中，$\hat{a}_t$ 是学习者选择的臂。 $\tilde{\pi} := \pi_{\tilde{\theta}}$。最后，观察者的预测遗憾 $\rho(\tilde{\pi})$ 趋于零。

该算法表明，恢复最优策略并不需要观察结果。这将核心挑战从利用反馈进行学习转变为解读显露偏好，并证明仅凭上下文-动作轨迹就足以实现识别。

5 观察者所学策略的可证明保证

在本节中，我们对观察者通过在后缀数据集 $D^{T(N)+1:N}{}_{\text{obs}}$ 上进行经验风险最小化（ERM）所学得的策略进行有限样本分析。我们的目标是界定所学策略 $\tilde{\pi}$ 相对于观察者的有效样本量 $L(N) = N - T(N)$ 的预测遗憾（定义 3）。

从遗憾到决策错误。我们首先将遗憾与策略选择次优臂这一事件联系起来。对于任意策略 π，令 $a_t^\star \in \arg\max_{a \in \mathcal{A}_t} \langle x_a, \theta^\star \rangle$ 表示第 t 轮的一个最优臂。 π 在第 t 轮造成的瞬时遗憾为 $\Delta_t(\pi) := \langle x_{a_t^\star}, \theta^\star \rangle - \langle x_{\pi(\mathcal{A}_t, X_t)}, \theta^\star \rangle$。以下引理表明，有界特征和参数意味着奖励差距存在统一上界，进而限制每轮遗憾。令 $\mathcal{D}$ 表示在此阶段生成 $(\mathcal{A}_t, X_t)$ 的未知分布。

引理 **1.** 假设对于所有 a、t 和 $\|\theta^\star\|_2 \leq 1$，$\|x_a\|_2 \leq 1$ 均成立。那么，对于任意策略 π，预测 e

遗憾由干净风险界定：

$$\rho(\pi) \le 2 \cdot \Pr_{(\mathcal{A}_t, X_t)\sim\mathcal{D}}(\pi(\mathcal{A}_t, X_t) \neq a^*).$$

证明。由于奖励差距有界，

$$\|\langle x_{a_t^*} - x_a, \theta^*\rangle\|_2 \le \|\langle x_{a_t^*}, \theta^*\rangle\|_2 + \|\langle x_a, \theta^*\rangle\|_2 \le 2,$$

期望遗憾至多为选择次优臂概率的两倍。 □

连接干净风险与含噪风险。 在第二阶段，观察者收集带标签的观测 $D^T{}_{(N)+1:N\text{obs}} = \{(\mathcal{A}_t, X_t, \hat{a}_t)\}_{t=T(N)+1}^{N}$。我们通过两种风险来衡量观察者的表现：相对于最优臂的干净风险，以及相对于学习者动作的含噪模仿风险，

$$\mathcal{R}(\pi) := \Pr_{(\mathcal{A}_t, X_t)\sim\mathcal{D}}[\pi(\mathcal{A}_t, X_t) \neq a_t^*],$$
$$\mathcal{R}_{\text{noisy}}(\pi) := \Pr_{(\mathcal{A}_t, X_t)\sim\mathcal{D}}[\pi(\mathcal{A}_t, X_t) \neq \hat{a}_t],$$

其中 a_t^* 是最优臂，$\hat{a}_t$ 表示学习者在上下文 $(\mathcal{A}_t, X_t)$ 下的含噪选择。直观而言，$\mathcal{R}(\pi)$ 表示选择次优臂的概率，而 $\mathcal{R}_{\text{noisy}}(\pi)$ 则量化了观察者模仿学习者的效果。关键在于，在 Massart 型噪声条件下建立这两个量之间的联系。

引理 2。 令 $\pi^\star(\mathcal{A}_t, X_t) = a_t^*$。在假设 *1*（即 $\Pr[\hat{a}_t \neq a_t^* \mid \mathcal{A}_t, X_t] \le \eta(T)$ 满足 $\eta(T) < 1/2$）下，对于任意策略 π，

$$\mathcal{R}_{\text{noisy}}(\pi) - \mathcal{R}_{\text{noisy}}(\pi^\star) \ge (1 - 2\eta(T))\,\mathcal{R}(\pi).$$

证明。固定 $(\mathcal{A}_t, X_t)$。记 $\eta_t = \Pr(\hat{a}_t \neq a_t^* \mid \mathcal{A}_t, X_t) \le \eta(T)$。$\pi^*(\mathcal{A}_t, X_t) = a_t^*$。定义条件带噪损失

$$\ell_{\text{noisy}}(\pi; \mathcal{A}_t, X_t) := \Pr\big(\pi(\mathcal{A}_t, X_t) \neq \hat{a}_t \mid \mathcal{A}_t, X_t\big).$$

若 $\pi(\mathcal{A}_t, X_t) = a_t^*$，则

$$\ell_{\text{noisy}}(\pi; \mathcal{A}_t, X_t) = \ell_{\text{noisy}}(\pi^*; \mathcal{A}_t, X_t),$$

且条件超额为 0。否则，令 $b := \pi(\mathcal{A}_t, X_t) \neq a_t^*$。则

$$\begin{aligned}&\ell_{\text{noisy}}(\pi; \mathcal{A}_t, X_t) - \ell_{\text{noisy}}(\pi^*; \mathcal{A}_t, X_t)\\ =&\Pr(\hat{a}_t = a_t^* \mid \mathcal{A}_t, X_t) - \Pr(\hat{a}_t = b \mid \mathcal{A}_t, X_t).\end{aligned}$$

此外，$\Pr(\hat{a}_t = b \mid \mathcal{A}_t, X_t) \le \Pr(\hat{a}_t \neq a_t^* \mid \mathcal{A}_t, X_t) = \eta_t$，因此上述结果至少为 $(1 - \eta_t) - \eta_t = 1 - 2\eta_t$。所以，对于所有 $(\mathcal{A}_t, X_t)$，

$$\begin{aligned}&\ell_{\text{noisy}}(\pi; \mathcal{A}_t, X_t) - \ell_{\text{noisy}}(\pi^*; \mathcal{A}_t, X_t)\\ \ge&\begin{cases}0 & \text{if } \pi(\mathcal{A}_t, X_t) = a_t^*\\ 1 - 2\eta_t & \text{if } \pi(\mathcal{A}_t, X_t) \neq a_t^*\end{cases}\\ =&(1 - 2\eta_t)\mathbf{1}\big[\pi(\mathcal{A}_t, X_t) \neq a_t^*\big].\end{aligned}$$

对 $(\mathcal{A}_t, X_t, \hat{a}_t) \sim D_{\text{obs}}$ 取期望，并使用 $\eta_t \le \eta(T)$，得到

$$\mathcal{R}_{\text{noisy}}(\pi) - \mathcal{R}_{\text{noisy}}(\pi^*) \ \ge \ (1 - 2\eta(T))\,\mathcal{R}(\pi),$$

如所声称。 □

在引理 2 的条件下，对于任意策略 π，

$$\mathcal{R}(\pi) \ \le \ \frac{\mathcal{R}_{\text{noisy}}(\pi) - \mathcal{R}_{\text{noisy}}(\pi^\star)}{1 - 2\eta(T)}.$$

只需控制观察者相对于 $\pi^\star$ 的带噪模仿风险，即可得到针对最优臂的干净错误率的界。

有界噪声下的经验风险最小化泛化 回顾观察者的策略类 Π : $= \left\{\pi_\theta \ \mathcal{A}_t \ X_t \ = \ {}_{a\in\mathcal{A}_t}\langle x_a \ \theta\rangle \quad \theta \in \mathbb{R}^d\right\}$ $\tilde{\theta}$ (,) : arg max, : 。在第二阶段，观察者使用学习者所选的臂作为模仿标签，通过经验风险最小化进行拟合。将经验带噪模仿风险定义为

$$\tilde{\mathcal{R}}_{\text{noisy}}(\pi) := \frac{1}{L(N)} \sum_{t=T+1}^{N} \mathbf{1}\big[\pi(\mathcal{A}_t, X_t) \neq \hat{a}_t\big],$$

˜其中 $L(N) = N - T(N)$ 是第二阶段的时域，且 $\tilde{\pi} \in \arg\min_{\pi\in\Pi} \mathcal{R}_{\text{noisy}}(\pi)$。为了控制 $\mathcal{R}_{\text{noisy}}$ 与总体带噪风险 $\mathcal{R}_{\text{noisy}}(\pi)$ 之间的˜泛化间隙，我们使用 Natarajan 维数 [17], ，这是多类别 0-1 学习的一种标准复杂度度量。

定义 4 （Natarajan 维数）。 如果存在两种不同的动作 $a_{i,0}, a_{i,1} \in \mathcal{A}_i$ ，使得对每个 $i \in [m]$ ，对于集合 $S = \{(\mathcal{A}_1, X_1) \ldots, (\mathcal{A}_m, X_m)\}$ 中的任意 $v \in \{0,1\}^m$，都存在满足以下条件的 $\theta_v \in \mathbb{R}^d$ ，则称该集合被策略类 Π N-打散：

$$\arg\max_{a\in\mathcal{A}_i} \langle x_a, \theta_v\rangle = a_{i,v_i}, \forall i \in [m].$$

Π 的 Natarajan 维数记为 $d_{\text{Nat}}(\Pi)$，是能够被 N-打散的此类集合的最大 m 。

回顾一下，我们的观察者在此前第 Π 节4.2中定义的线性评分策略类上最小化损失。现在我们来界定该类的复杂度。

引理 3。 考虑线性策略类 $\Pi := \{\pi_\theta(\cdot) = \arg\max_{a\in\mathcal{A}} \langle x_a, \theta\rangle \mid \theta \in \mathbb{R}^d\}$。根据先前关于线性多类别预测器的 *Natarajan* 维数的研究结果， *[9]*, 具有 K 个臂时，其复杂度 Π 有如下界：

$$d_{\text{Nat}}(\Pi) = O(d \log K).$$

引理 4. 设 Π 是一个 *Natarajan* 维数为 $d_{Nat}(\Pi)$ 的假设类。对于任意 $\delta \in (0,1)$，以至少 $1 - \delta$ 的概率，在生成大小为 L 的后缀数据集 $D_{obs}^{T+1:N}$ 时，有：

$$\begin{aligned}&\sup_{\pi\in\Pi} \left|\mathcal{R}_{\text{noisy}}(\pi) - \tilde{\mathcal{R}}_{\text{noisy}}(\pi)\right| \\ \le & C \cdot \sqrt{\frac{d_{Nat}(\Pi) \log L(N) + \log(1/\delta)}{L(N)}},\end{aligned}$$

其中 C 是一个与 d、 K 和 $L(N)$ 无关的常数。

在后续分析中，我们使用常数 C 来概括集中不等式中出现的绝对常数，从而能够专注于收敛速率。尽管引理4 的完整推导遵循标准集中技术，但为节省篇幅，我们将正式证明放在附录B.1 中。具体而言，该结果结合了有限 Natarajan 维数类别的增长函数界 [17, 9]和 Azuma-Hoeffding不等式 [4] ，以处理数据生成过程的自适应特性。标准结果针对独立同分布数据给出，而对于有界损失函数，将其推广到自适应设定（其中上下文独立同分布，但标签$\hat{}$ a_t 依赖于历史）是直接的。

泛化与遗憾界。现在，我们通过假设存在 $\theta^\star$ 使得 $\pi^\star(\mathcal{A},\ X) = \pi_{\theta^\star}(\mathcal{A},\ X) = a^\star(\mathcal{A},\ X)$ 在（$\mathcal{A}$, X）上几乎必然成立，来纳入可实现性。在参数为 $\eta(T) < \frac{1}{2}$的预热后Massart噪声下，带噪模仿超额风险通过 Massart转移不等式（引理2）控制干净风险，而经验风险最小化进一步将总体保证转化为预测遗憾界。

定理 4 (经验风险最小化器的干净风险保证)。 假设可实现性成立 $\pi^\star \in \Pi$ ，且存在预热后 *Massart*噪声，其参数为 $\eta(T) < \frac{1}{2}$，其中 $T = T(N)$ 为预热长度。那么，对于任意$\delta \in (0,\ 1)$，以至少 $1-\delta$ 的概率，

$$\mathcal{R}(\tilde{\pi}) \le \frac{C}{1-2\eta(T)}\sqrt{\frac{d\log K \cdot \log L(N) + \log(1/\delta)}{L(N)}},$$

其中 $L(N) = N - T(N)$ 是第二阶段的时域， C是一个与问题参数无关的常数。

证明。 令 $L = N - T$ 为第二阶段的时域。我们应用引理 4中的一致收敛界。定义泛化误差界 ε_L 如下:

$$\varepsilon_L := C\sqrt{\frac{d_{\mathrm{Nat}}(\Pi)\log L + \log(1/\delta)}{L}}.$$

根据引理 4，事件

$$\mathcal{E} := \left\{\sup_{\pi\in\Pi}\left|\mathcal{R}_{\mathrm{noisy}}(\pi) - \tilde{\mathcal{R}}_{\mathrm{noisy}}(\pi)\right| \le \varepsilon_L\right\}$$

成立的概率至少为 $1-\delta$。

以事件 $\mathcal{E}$ 为条件。由于 $\tilde{\pi}$ 是后缀数据集上的经验风险最小化器，我们˜ ˜有 $\mathcal{R}_{\mathrm{noisy}}(\tilde{\pi}) \le \mathcal{R}_{\mathrm{noisy}}(\pi^\star)$。因此:

$$\begin{aligned}
\mathcal{R}_{\mathrm{noisy}}(\tilde{\pi}) &\le \tilde{\mathcal{R}}_{\mathrm{noisy}}(\tilde{\pi}) + \varepsilon_L && \text{(by event } \mathcal{E}\text{)}\\
&\le \tilde{\mathcal{R}}_{\mathrm{noisy}}(\pi^\star) + \varepsilon_L && \text{(by optimality of } \tilde{\pi}\text{)}\\
&\le \mathcal{R}_{\mathrm{noisy}}(\pi^\star) + 2\varepsilon_L && \text{(by event } \mathcal{E} \text{ on } \pi^\star\text{)}.
\end{aligned}$$

这意味着超额带噪风险有如下界:

$$\mathcal{R}_{\mathrm{noisy}}(\tilde{\pi}) - \mathcal{R}_{\mathrm{noisy}}(\pi^\star) \le 2\varepsilon_L.$$

根据引理2，我们有 $\Pr(\hat{a}_t \neq a_t^* \mid \mathcal{A}_t,\ X_t) \leq \eta(T) < \frac{1}{2}$。因此，

$$\mathcal{R}_{\text{noisy}}(\tilde{\pi}) - \mathcal{R}_{\text{noisy}}(\pi^*) \geq (1 - 2\eta(T))\, R(\tilde{\pi}).$$

结合这两个表达式可得，在 $\mathcal{E}$，$R(\tilde{\pi}) \leq \frac{2\varepsilon_L}{1-2\eta(T)}$ 下。~~其中 ε_L~~可通过引理 4 和引理3具体化，从而得到：

$$\begin{aligned}\mathcal{R}(\tilde{\pi}) &\leq \frac{2C}{1-2\eta(T)}\sqrt{\frac{d_{\text{Nat}}(\Pi)\log L(N) + \log(1/\delta)}{L(N)}} \\ &\leq \frac{C'}{1-2\eta(T)}\sqrt{\frac{d\log K \cdot \log L(N) + \log(1/\delta)}{L(N)}},\end{aligned}$$

其中 C' 是另一个常数。 □

定理 4揭示了这样的直觉：一旦在给定（$\mathcal{A}$，X）的条件下，学习者在预热期后的动作正确次数多于错误次数（即 $\eta(T) < 1/2$），那么在 Π 上最小化模仿误差，就能恢复一个在大多数上下文中与最优臂一致的策略。现在，我们将无噪声决策误差界转化为累积遗憾界。考虑在时域 $\tilde{\pi}$ L 内，针对独立同分布的 $D_{\text{obs}} = \{(\mathcal{A}_t,\ X_t)\}_{t=T+1}^{N}$，部署观察者的策略 $\tilde{\pi}$。</output>

定理 5（预测迁移遗憾界）。 设 $T = T(N)$ 为第一阶段的长度。在动态 *Massart* 噪声条件（假设 *1*）下，设 $\tilde{\pi}$ 为在长度为 $L(N)$的后缀数据集上通过经验风险最小化学得的策略。以至少 $1-\delta$的概率，$\tilde{\pi}$的预测迁移遗憾满足：

$$\rho(\tilde{\pi}) \leq \frac{C}{1-2\eta(T)}\sqrt{\frac{d\log K \cdot \log L(N) + \log(1/\delta)}{L(N)}},$$

其中 C 是一个问题相关常数。

证明。该结果可直接由引理1 和定理4得出。首先，引理1证明预测遗憾不超过无噪声分类风险的两倍：$\rho(\tilde{\pi}) \leq 2\mathcal{R}(\tilde{\pi})$。定理4给出了干净风险的一个高概率界 $\mathcal{R}(\tilde{\pi})$：

$$\mathcal{R}(\tilde{\pi}) \leq \frac{C}{1-2\eta(T)}\sqrt{\frac{d\log K \cdot \log L(N) + \log(1/\delta)}{L(N)}}.$$

直接代入该界即可得到所述结果。 □

定理5 中的界揭示了一种由第一阶段时段长度 T决定的基本权衡。遗憾由两个相互竞争的项组成：第一项是学习者的标签噪声 $(1-2\eta(T))^{-1}$。由于 $\eta(\cdot)$ 是非递增的（假设1），较大的 T 会降低学习者的错误率。第二项是泛化误差 $\tilde{O}(1/\sqrt{L(N)})$，它随有效样本量 $L(N) = N - T(N)$ 变化。较大的 T 会减少可用于经验风险最小化的样本数，从而增大误差。因此，选择最优的 T 需要在标签质量与训练数据数量之间取得平衡。

推论1。 假设学习者的策略满足累积遗憾界 $\mathbb{E}[R_T(\hat{\pi})] \leq C_{alg}\sqrt{T}$。进一步假设存在唯一最优臂，且在 $\Delta_{\min} > 0$ 上具有一致间隔 $\text{supp}(\mathcal{D})$（定义 *2*）。若采用保守的预热期 $T(N) = \Theta(N^\alpha)$，其中 $\alpha \in (0,1)$，

那么对于足够大的 N，条件 $\eta(T) < 1/2$ 成立，且观察者的预测遗憾按以下方式增长：

$$\rho(\tilde{\pi}) \leq \tilde{O}\left(\sqrt{\frac{d \log K}{N}}\right).$$

证明。回想一下，每次选择次优动作，都会产生至少 $\Delta_{\min}$（定义 2）的瞬时遗憾。因此，截至时间 T 的预期错误次数上界为 $\mathbb{E}[R_T]/\Delta_{\min}$。由于错误率 $\eta(t)$ 是非递增的（假设 1），时间 T 的瞬时错误率有以下平均错误率上界：

$$\eta(T) \leq \frac{1}{T}\sum_{t=1}^{T}\mathbb{P}(\hat{a}_t \neq a_t^*) \leq \frac{\mathbb{E}[R_T(\hat{\pi})]}{T\Delta_{\min}} \leq \frac{C_{\text{alg}}}{\Delta_{\min}\sqrt{T}}.$$

接下来，采用保守的预热期 $T(N) = \Theta(N^\alpha)$，有 $\lim_{N\to\infty} T(N) = \infty$。因此，$\lim_{N\to\infty} \eta(T(N)) = 0$。于是，当 N足够大时，定理 $\eta(T) < 1/2$所要求的条件得以满足，且噪声膨胀因子 $(1-2\eta(T))^{-1}$ 趋近于 1。

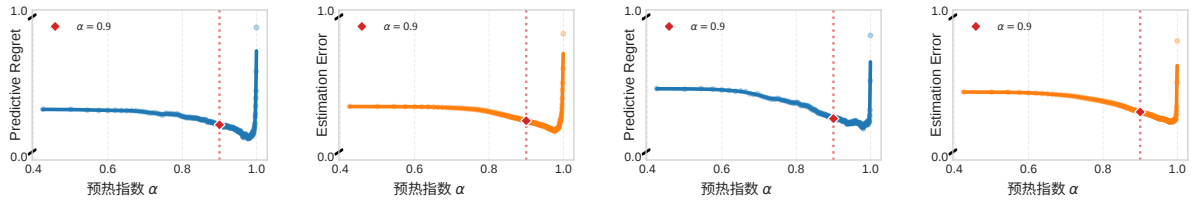
考虑有效样本量 $L(N) = N - T(N) = N - \Theta(N^\alpha)$。由于指数$\alpha < 1$，项 N 占主导，因此 $L(N) = \Theta(N)$。将这些代入定理 5 给出的界，并注意到 $\sqrt{T} = \Theta(N^{\alpha/2})$：

$$\begin{aligned}
\rho(\tilde{\pi}) &\leq \frac{C'\Delta_{\min}\Theta(N^{\alpha/2})}{\Delta_{\min}\Theta(N^{\alpha/2}) - 2C_{\text{alg}}} \cdot \\
&\quad \sqrt{\frac{d\log K \cdot \log(N - \Theta(N^\alpha)) + \log(1/\delta)}{N - \Theta(N^\alpha)}} \\
&= O(1) \cdot \tilde{O}\left(\sqrt{\frac{d\log K}{N}}\right) \\
&= \tilde{O}\left(\sqrt{\frac{d\log K}{N}}\right).
\end{aligned}$$

随着 $N \to \infty$，项 $\Theta(N^{\alpha/2})$ 占主导地位，超过常数 $2C_{\text{alg}}$，使系数趋近于 1。类似地，分母中的 N 占主导地位，超过 $\Theta(N^\alpha)$，从而恢复标准收敛速率。

□

备注 1. 直接比较观察者与学习者的表现很有启发性。标准的无遗憾学习者（例如 LinUCB 或 LinTS）通常可实现 $\tilde{O}(1/\sqrt{N})$ [8, 3] 的平均遗憾率。该推论表明，我们的观察者在预测遗憾方面也能达到 $\tilde{O}(1/\sqrt{N})$ 的渐近速率。考虑到严格的信息不对称，这一结果尤为引人注目：学习者依赖完整奖励反馈实现收敛，而观察者使用显著更少的信息，却能达到同等效率。此外，α 的选择表明，我们无需精确调整预热长度。即使选择严重偏向标签质量而非样本数量，观察者的收敛速率仍由剩余时域的线性增长所主导，从而确保最优效率。这意味着即使信息有限，观察者仍能识别出最优决策边界。



(a) LinTS-遗憾 (b) LinTS-误差 (c) LinUCB-遗憾 (d) LinUCB-误差

图1：完整时域布局下的性能比较（$d = 50$, K=200）。左侧区块展示 LinTS 的性能，右侧区块展示 LinUCB 的结果。

6 实验

我们在线性上下文多臂老虎机环境中评估所提出的后缀模仿框架。我们的首要目标是通过实证验证两阶段策略带来的权衡，并检验理论驱动的调度方案是否有效。此外，我们还旨在证明，被动观察者通过选择性事后信息，能够实现与主动学习者相近的参数识别效果。

6.1 实验设置

学习者和观察者。 学习者是一个具有完整奖励反馈的在线线性多臂老虎机智能体。我们分别报告学习者使用 LinUCB 和 LinTS 算法的结果。观察者不接收奖励，只观察 $(\mathcal{A}_t,\ X_t,\ \hat{}a_t)$。

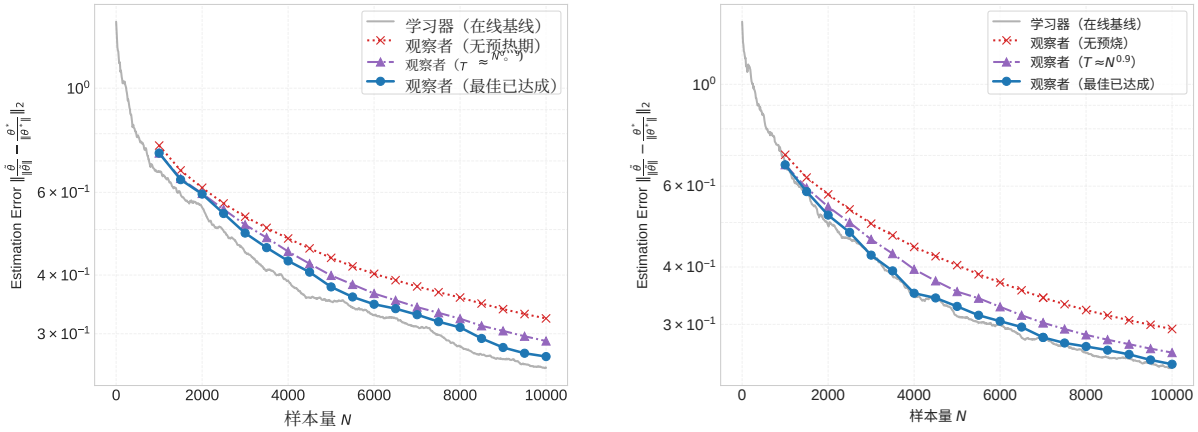
在第二阶段，它通过最小化候选集 $\pi_\theta(\mathcal{A}_t,\ X_t) = \arg\max_{a\in\mathcal{A}_t}\langle x_a,\ \theta\rangle$ 上的条件 softmax 损失，训练一个线性评分策略 $\mathcal{A}_t$，将其作为 0–1 模仿目标的凸替代损失。[1] 我们比较：（i）使用所有轮次的朴素模仿（即预热长度 $T(N) = 0$）；（ii）通过遍历 T并选择最佳值，事后确定 T的预言机预热；以及（iii）基于规则的第一阶段长度 $T(N) = N^{0.9}$，这是我们可实际实施的方案，并与预言机结果一并报告。

指标。 我们评估观察者的预测遗憾 $\rho(\tilde{\pi}) = \mathbb{E}\left[\langle x_{a_t^*}, \theta^*\rangle - \langle x_{\pi_\theta(\mathcal{A}_t,\ X_t)}\ \theta^*\rangle\right]$，该遗憾使用无噪声期望奖励 $\langle x_a,\ \theta^*\rangle$在独立同分布测试集上估计。此外，我们通过归一化参数方向误差 $\left\| \frac{\theta}{\|\tilde{\theta}\|} - \frac{\theta^*}{\|\theta^*\|} \right\|_2$评估参数恢复情况。所有结果均为多个随机种子上的平均值。

6.2 结果

预热长度的影响。 我们首先通过考察观察者在不同预热指数参数 α （$T = N^\alpha$）下的表现，验证第一阶段的必要性。如图 1 所示，预测决策损失和参数估计误差均呈现出典型的 U 形模式。由于篇幅限制，我们在正文中给出设定$d = 50$、 $K = 200$ 的结果。我们在不同维度（$d \in \{20, 50\}$）和臂集合（$K \in \{50, 100, 200\}$）上进行了广泛实验。这些设定下的结果呈现出一致趋势和稳健表现，验证了我们的理论主张。这些额外配置的详细图示见附录 D。

[1] Softma x 交叉熵是分类校准的：替代风险的超额部分可以控制 0–1 风险的超额部分；参见 Zhan g [27]. Multi Tewari 和 Bartlett [20] 讨论了类别一致性。



（a）LinTS 比较　　（b）LinUCB 比较

图 2：观察者策略与学习者基线的性能比较（$d = 50$，K=200）。图（a）展示 LinTS 的结果，图（b）展示 LinUCB 的结果。两项指标均表明，观察者（最佳已达成）的表现优于在线学习者。

在没有预热阶段时（当 $T \to 0$），观察者会遭受显著误差。这证实了我们的理论担忧：学习者的早期历史由探索性动作主导，从而引入严重的分布偏移，使观察者产生混淆。相反，当$T \to N$时，丢弃过多数据会因样本量不足而导致误差急剧上升。最优表现始终出现在中间比例处，此时观察者能够有效平衡去除有偏数据与保留足够样本以实现收敛之间的权衡。这从经验上验证了推论1，表明多种多项式预热计划都能缓解迁移遗憾。

渐近收敛与可解释性。图2 比较了不同策略的参数估计误差随时域 N 增加的变化情况。几个关键观察结果凸显了我们框架的有效性。朴素模仿的表现始终不如学习者。这表明，仅仅克隆学习者的完整轨迹，无法准确恢复真实参数，因为数据集受到了次优探索性动作的污染。值得注意的是，预言机观察者取得了与学习者自身相近的估计误差。这一反直觉的结果展示了该策略的优势：通过有选择地仅从学习者处于稳定且最优阶段的数据中学习，观察者能够过滤噪声并更准确地识别底层模型 θ^* 。

基于规则的策略最初因激进的数据剪枝而出现较高方差。然而，随着 N 增大，它迅速收敛，显著优于朴素基线，并接近预言机的性能。这证实了激进的预热计划在大样本情形下是可行且有效的。

7 结论

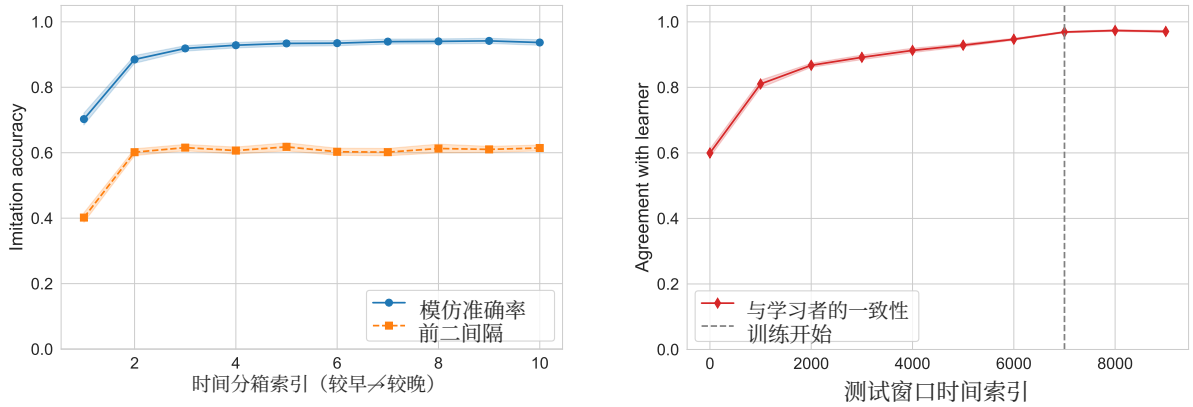
我们提出了逆上下文老虎机框架，以应对从非平稳、含噪监督中学习的挑战。我们的结果表明，信息不对称并不会妨碍最优识别。从理论上讲，我们证明了可以完全从交互轨迹中恢复底层效用函数；这一结论出乎意料地意味着，即使没有奖励反馈，也能对学习智能体进行解读。这些发现得到了详

细实验的支持。未来值得探索的方向包括将该框架扩展到非线性函数和部分监测场景。

参考文献

[1] Yasin Abbasi-Yadkori、D´avid P´al 和 Csaba Szepesv´ari。线性随机老虎机的改进算法。《神经信息处理系统进展》，24，2011年。[2] Pieter Abbeel 和 Andrew Y Ng。通过逆强化学习进行学徒式学习。载于《第二十一届国际机器学习会议论文集》，第1页，2004年。[3] Shipra Agrawal 和 Navin Goyal。面向线性收益的上下文老虎机汤普森采样。载于《国际机器学习会议》，第127–135页。机器学习研究论文集，2013年。[4] 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。[5] Yikun Ban、Yunzhe Qi 和 Jingrui He。用于个性化推荐的神经上下文老虎机。载于《2024年ACM网络会议伴随论文集》，第1246–1249页，2024年。[6] Suneel Belkhale、Yuchen Cui 和 Dorsa Sadigh。模仿学习中的数据质量。《神经信息处理系统进展》，36:80375–80395，2023年。[7] Heewoong Choi、Sangwon Jung、Hongjoon Ahn 和 Taesup Moon。离线偏好强化学习的列表式奖励估计。《机器学习研究论文集》，235:8651–8671，2024年。[8] Wei Chu、Lihong Li、Lev Reyzin 和 Robert Schapire。具有线性收益函数的上下文老虎机。载于《第十四届国际人工智能与统计会议论文集》，第208–214页。JMLR研讨会与会议论文集，2011年。[9] Amit Daniely、Sivan Sabato、Shai Ben-David 和 Shai Shalev-Shwartz。多类可学习性与经验风险最小化原则。《机器学习研究杂志》，16(1):2377–2404，2015年。[10] Huy Hoang、Tien Mai 和 Pradeep Varakantham。Sprinql：由次优示范驱动的离线模仿学习。《神经信息处理系统进展》，37:136837–136872，2024年。[11] Alihan H¨uy¨uk、Daniel Jarrett 和 Mihaela van der Schaar。逆向上下文老虎机：学习行为如何随时间演变。载于《国际机器学习会议》，第9506–9524页。机器学习研究论文集，2022年。[12] Yachen Kang、Diyuan Shi、Jinxin Liu、Li He 和 Donglin Wang。超越奖励：离线偏好引导策略优化。载于《第40届国际机器学习会议论文集》，第15753–15768页，2023年。[13] Tor Lattimore 和 Csaba Szepesvari。老虎机算法。2020年。URL https://api. semanticscholar.org/CorpusID:242172176。

[14] Lihong Li、Wei Chu、John Langford 和 Robert E Schapire。用于个性化新闻文章推荐的上下文赌博机方法。载于《第19届万维网国际会议论文集》，第661–670页，2010年。[15] Haitong Ma、Bo Dai、Zhaolin Ren、Yebin Wang 和 Na Li。通过预训练动力学表征，在任意示范上进行离线模仿学习。arXiv预印本 arXiv:2508.14383，2025年。[16] Pascal Massart 和 ´Elodie N´ed´elec。统计学习的风险界。2006年。[17] Balas K Natarajan。关于集合与函数学习。《Machine Learning》，4(1):67–97，1989年。[18] Andrew Y Ng、Stuart Russell 等。逆强化学习算法。载于Icml，第1卷，第2页，2000年。[19] Seokin Seo、Byung-Jun Lee、Jongmin Lee、HyeongJoo Hwang、Hongseok Yang 和 Kee-Eung Kim。通过稳健的平稳分布校正，缓解行为克隆中的协变量偏移。《神经信息处理系统进展》，37:109177–109201，2024年。[20] Ambuj Tewari 和 Peter L Bartlett。关于多类分类方法的一致性。《机器学习研究杂志》，8(5)，2007年。[21] Yogatheesan Varatharajah 和 Brent Berry。临床试验中用于知情决策的基于上下文赌博机的方法。《Life》，12(8):1277，2022年。[22] Tengyu Xu、Yue Wang、Shaofeng Zou 和 Yingbin Liang。具有轨迹级奖励的可证明高效离线强化学习。《IEEE信息论汇刊》，2024年。[23] Yi Xu、Weiran Shen、Jun Xu、Xiao Zhang 和 Ji-Rong Wen。IBCB：面向行为演化历史的高效逆批量上下文赌博机。《IEEE模式分析与机器智能汇刊》，(01):1–17，2026年1月。ISSN 1939-3539。doi: 10.1109/TPAMI.2026.3650796。URL https://doi.ieeecomputersociety.org/10.1109/TPAMI.2026.3650796。[24] Sheng Yue、Jiani Liu、Xingyuan Hua、Ju Ren、Sen Lin、Junshan Zhang 和 Yaoxue Zhang。如何在离线模仿学习中利用多样化示范。载于《第41届国际机器学习会议论文集》，第58037–58067页，2024年。[25] Maryam Zare、Parham M Kebria、Abbas Khosravi 和 Saeid Nahavandi。模仿学习综述：算法、最新进展与挑战。《IEEE控制论汇刊》，2024年。[26] Siliang Zeng、Chenliang Li、Alfredo Garcia 和 Mingyi Hong。当示范遇上生成式世界模型：离线逆强化学习的最大似然框架。《神经信息处理系统进展》，36:65531–65565，2023年。[27] Tong Zhang。基于凸风险最小化的分类方法的统计行为与一致性。《统计年鉴》，32(1):56–85，2004年。



(a) 可预测性　　(b) 晚期策略泛化

图3：假设1的诊断性验证：(a) 学习者的动作随时间推移变得越来越可预测；(b) 与早期测试窗口相比，经过晚期训练的观察者在晚期测试窗口中与学习者的一致性更高（基于 20 个随机种子的 95% 置信区间）。

.

假设1的验证

假设1涉及一个预热后阶段，在该阶段，学习者的行为变得越来越可靠，即有效决策噪声随预热长度增加而降低 $T(N)$。在现实世界的无奖励日志中，最优动作 a_t^* 不可观测，因此无法依据其定义验证该假设。因此，我们报告一项针对可观测推论的诊断性验证：丢弃初始前缀后，学习者的动作应能根据（$\mathcal{A}_t$， X_t）变得更加可预测，并且在相邻的晚期时间窗口之间更加稳定。

我们将轨迹划分为 J 个相等的时间分箱（前 $\rightarrow$ 后）。在每个分箱中，我们使用 80% 的样本拟合一个固定容量的逻辑回归观察器，并在其余 20% 的样本上评估模仿准确率。在 20 个独立随机种子上，模仿准确率随时间呈现显著的单调上升趋势（图3a）：时间分箱索引与平均准确率之间的斯皮尔曼相关系数为 $r = 0.9273$，且 $p = 1.12 \times 10^{-4}$。此外，将最初 30% 的分箱（早期）与最后 30% 的分箱（晚期）进行比较，晚期准确率平均提高了 0。1041，95% 置信区间为 $[0.0939, 0.1142]$。

为评估时间稳定性，我们使用轨迹最后 30% 的数据训练后期观察者策略，并在覆盖整个时域的滚动测试窗口上评估该策略与学习者的动作一致性。结果显示，后期窗口上的一致性显著高于早期窗口（图 3b）：平均差值（后期 $\rightarrow$后期 $-$后期 $\rightarrow$早期）为 0。2119，95% 置信区间为 $[0.2055, 0.2182]$。

这些诊断结果与假设 1: 的预期内容一致：经过充分的预热期后，学习者的行为变得更容易根据 $(\mathcal{A}_t,\ X_t)$ 进行预测，并且在后期各时间窗口之间更加稳定，这表明相较于早期阶段，有效行为噪声有所降低。

B 理论细节

B.1 引理证明4

证明。 将第二阶段的样本记为 $D^{T(N)+1:N}{}_{\text{obs}} = \{(\mathcal{A}_t,\ X_t,\ \hat{a}_t)\}_{t=T(N)+1}^{N}$。时域为$L(N) = N - T(N)$。对于任意 $\pi \in \Pi$，0-1 限制损失为：

$$l_t(\pi) = \mathbf{1}[\pi(\mathcal{A}_t, X_t) \neq \hat{a}_t].$$

回顾总体带噪模仿风险和经验带噪模仿风险：

$$\mathcal{R}_{\text{noisy}}(\pi) := \mathbb{E}\big[\ell_t(\pi)\big],$$

$$\tilde{\mathcal{R}}_{\text{noisy}}(\pi) := \frac{1}{L}\sum_{t=T+1}^{N} \ell_t(\pi),$$

whe其中，期望是相对于后缀阶段的数据生成过程而言的[2]。

令 $\{\mathcal{F}_t\}$ 为由截至时间 t的历史生成的自然滤过。定义鞅差序列

$$X_t(\pi) := \ell_t(\pi) - \mathbb{E}[\ell_t(\pi) \mid \mathcal{F}_{t-1}], \qquad t = T+1, \ldots, N.$$

那么，几乎必然有 $\mathbb{E}[X_t(\pi) \mid \mathcal{F}_{t-1}] = 0$ 且 $X_t(\pi) \in [-1, 1]$。因此，根据 Azuma–Hoeffding 不等式 [4]，，对任意 $\varepsilon > 0$,

$$\Pr\left(\left|\frac{1}{L}\sum_{t=T+1}^{N} X_t(\pi)\right| > \varepsilon\right) \leq 2\exp(-2L\varepsilon^2).$$

等价地,

$$\Pr\left(\left|\tilde{\mathcal{R}}_{\text{noisy}}(\pi) - \frac{1}{L}\sum_{t=T+1}^{N} \mathbb{E}[\ell_t(\pi) \mid \mathcal{F}_{t-1}]\right| > \varepsilon\right) \leq 2\exp(-2L\varepsilon^2).$$

在我们的分析所采用的有效分布视角下，时间平均条件期望与 $R_{\text{noisy}}(\pi)$ 相符，因此同一界也适用于 $\left|\tilde{\mathcal{R}}_{\text{noisy}}(\pi) - \mathcal{R}_{\text{noisy}}(\pi)\right|$。

固定实例序列 $((\mathcal{A}_{T+1},\ X_{T+1})\ldots\ldots,\ (\mathcal{A}_N,\ X_N))$。 $\{\ell_t(\pi)\}_{t=T+1}^{N}$ 的取值仅通过预测向量 $(\pi(\mathcal{A}_{T+1},\ X_t)\ldots\ldots,\ \pi(\mathcal{A}_N X_N))$ 依赖于 π。令 $\tau_\Pi(L)$ 表示 Π 在 L 个点上的多分类增长函数。对于 Natarajan 维数有限的类别，存在一个绝对常数 $c > 0$，使得

$$\tau_\Pi(L) \leq (cL)^{d_{\text{Nat}}(\Pi)}.$$

这是有限 Natarajan 维数的一个标准推论 [17, 9]。

对 Π 在后缀实例上实现的所有不同预测向量应用并集界，并利用步骤 1 中固定的π 集中不等式，可得

$$\Pr\left(\sup_{\pi\in\Pi}\left|\tilde{\mathcal{R}}_{\text{noisy}}(\pi) - \mathcal{R}_{\text{noisy}}(\pi)\right| > \varepsilon\right) \leq 2\,\tau_\Pi(L)\exp(-2L\varepsilon^2) \leq 2(cL)^{d_{\text{Nat}}(\Pi)}\exp(-2L\varepsilon^2).$$

[2]正如正文所述，上下文可能是独立同分布的，而标签 $Y_t = \hat{a}_t$ 可能依赖于历史；下面的论证使用有界损失的鞅集中界。

令右侧至多为 δ，并求解 ε，可得

$$\sup_{\pi\in\Pi}\left|\tilde{\mathcal{R}}_{\text{noisy}}(\pi)-\mathcal{R}_{\text{noisy}}(\pi)\right| \le \sqrt{\frac{d_{\text{Nat}}(\Pi)\log(cL)+\log(2/\delta)}{2L}} \;\le\; C\sqrt{\frac{d_{\text{Nat}}(\Pi)\log L+\log(1/\delta)}{L}},$$

其中 $C>0$ 吸收了绝对常数。这就证明了该引理。 □

C 实验详情

本节将全面介绍实验环境、学习者和观察者各自的超参数配置，以及用于训练观察者策略的优化过程。

C.1 环境设置

我们在合成线性上下文赌博机环境中评估我们的框架。该环境由以下规范定义：

- 上下文生成：在每个轮次 t，每个臂 $a\in[K]$的上下文向量 $x_{t,a}\in\mathbb{R}^d$ 均独立地从标准多元正态分布中抽取，即 $x_{t,a}\sim\mathcal{N}(0,I_d)$。随后，我们对每个上下文向量进行归一化，使其满足 $\|x_{t,a}\|_2\le 1$。
- 奖励生成：从单位球面 $\mathcal{S}^{d-1}$上均匀采样一个真实参数向量 $\theta^*\in\mathbb{R}^d$。臂 a在轮次 t的奖励生成方式为：$r_{t,a}=\langle x_{t,a},\theta^*\rangle+\eta_t$，其中噪声项 η_t从高斯分布 $\mathcal{N}(0,\sigma^2)$ 中采样，且满足 $\sigma=0.1$。
- 维度：除非另有说明，默认问题维度设为 $d=50$，臂的数量设为 $K=200$。交互日志的总时域为$N=10,000$。
- 软件与硬件：所有实验均使用 Python 3.9，并通过 NumPy 和 Scikit-learn 实现。

C.2 算法超参数

本节详细介绍生成数据的主动学习者和观察者模型的具体配置。

学习者配置。 我们采用两种标准的无遗憾算法作为数据生成智能体：

- LinUCB：探索参数α_{UCB} 控制置信椭球的宽度。我们设置为 $\alpha_{\text{UCB}}=0.1$，并发现这能在我们的环境中实现均衡的探索与利用权衡。置信界计算如下：$\text{UCB}_{t,a}=\langle x_{t,a},\theta_t\rangle+\alpha_{\text{UCB}}\sqrt{x_{t,a}^\top V_t^{-1}x_{t,a}}$。
- LinTS：方差参数 ν 控制后验协方差的膨胀。我们设置 $\nu=\sigma^2\cdot d$，遵循标准启发式方法以确保充分探索。在每一步，智能体采样 $\tilde{\theta}_t\sim\mathcal{N}(\hat{\theta}_t,\nu V_t^{-1})$，并选择使$\langle x_{t,a},\tilde{\theta}_t\rangle$最大化的臂。

观察 r 配置。 观察者采用两阶段后缀模仿策略：

- 预热阶段（T）：正如正文所述，我们评估三种策略 ：

  1. 朴素策略（$T = 0$）：使用完整轨迹。2. 基于规则的策略（$T \approx N^{0.9}$）：具体而言，对于 $N = 10,\ 000$，我们设置 $T = \lfloor 10000^{0.9} \rfloor = 3981$。3. 预言机（最佳）：我们遍历预热指数 $\alpha \in [0, 1]$，并报告所达到的最小误差。

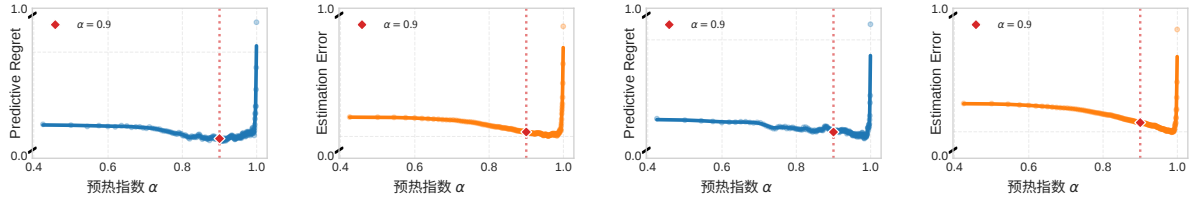
- 正则化（λ）：对于观察者的正则化最大似然估计（MLE），我们采用 L_2 正则化强度 $\lambda = 10^{-4}$。该值通过在留出的验证集 $\{10^{-3}, 10^{-4}, 10^{-5}\}$ 上进行粗略网格搜索选定。

C.3 优化细节

ˆ观察者的核心任务是通过最大化学习者在后缀数据集 $\mathcal{D}_{\text{suffix}} = \{(X_t,\ a_t)\}_{t=T+1}^{N}$ 上所选动作的似然，来估计 θ。

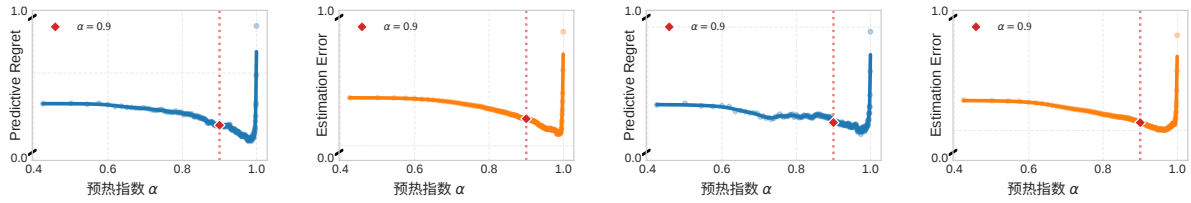
由于观察者看不到奖励，我们将该问题视为多分类任务，其中“标签”是学习者选择的臂。我们使用多项逻辑回归（Softmax 回归）模型，作为 0-1 模仿损失的凸替代损失。

D 补充实验结果



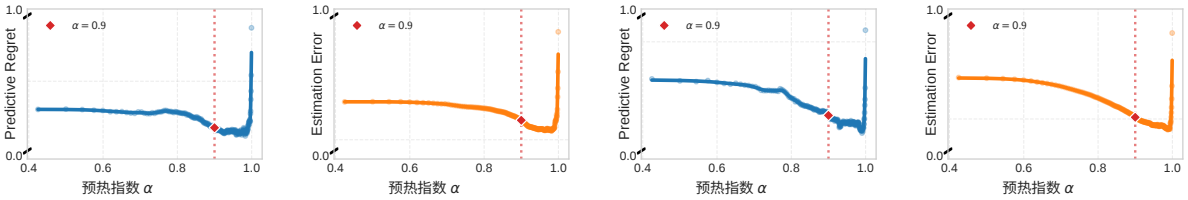
（a）LinTS-遗憾　（b）LinTS-误差　（c）LinUCB-遗憾　（d）LinUCB-误差

图4：完整时域布局下的性能比较（$d = 20,\ K = 50$）



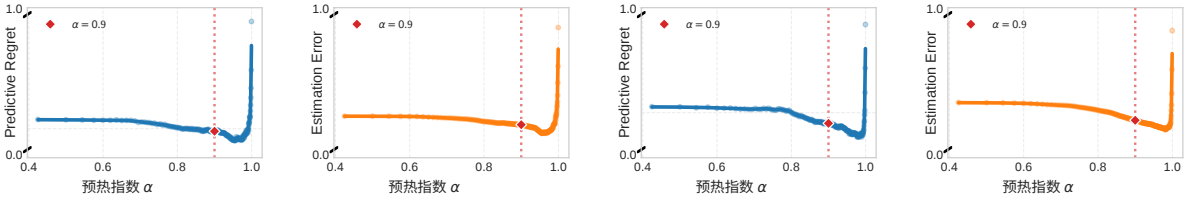
(a) LinTS-遗憾　(b) LinTS-误差　(c) LinUCB-遗憾　(d) LinUCB-误差

图5：完整时域布局下的性能比较（$d = 20,\ K = 100$）



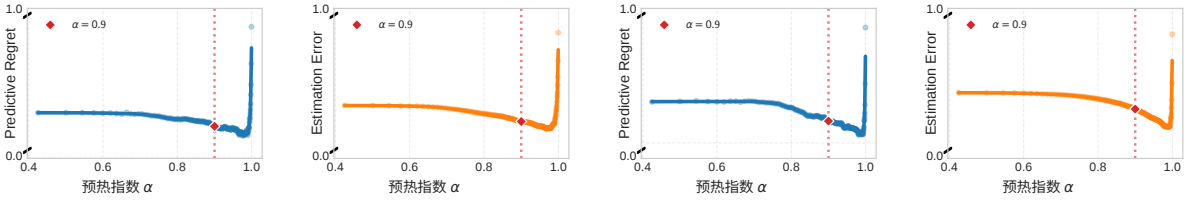
(a) LinTS-遗憾　(b) LinTS-误差　(c) LinUCB-遗憾　(d) LinUCB-误差

图 6：完整时域布局下的性能比较（$d = 20$, $K = 200$）



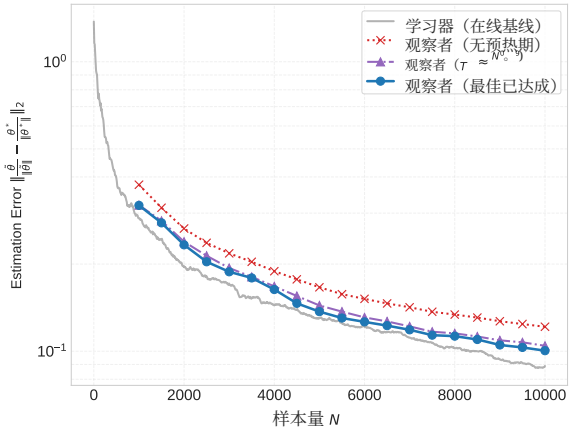
(a) LinTS-遗憾　(b) LinTS-误差　(c) LinUCB-遗憾　(d) LinUCB-误差

图 7：完整时域布局下的性能比较（$d = 50$, $K = 50$）

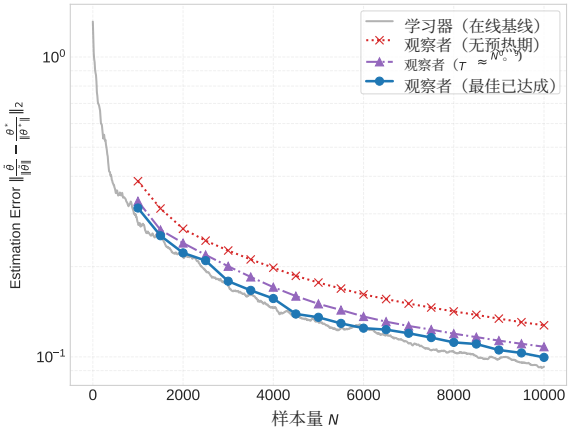


(a) LinTS-遗憾　(b) LinTS-误差　(c) LinUCB-遗憾　(d) LinUCB-误差

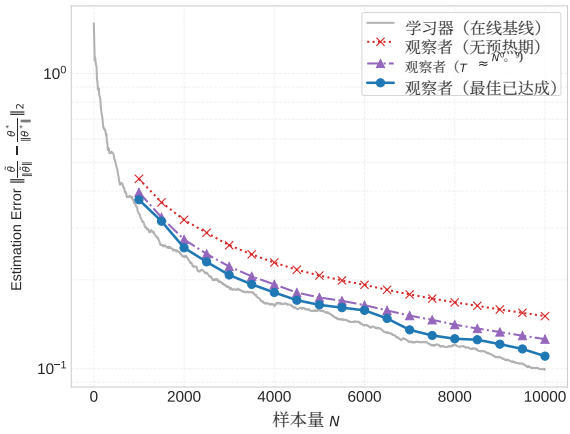
图 8：采用完整时域布局的性能比较（$d = 50$, $K = 100$）



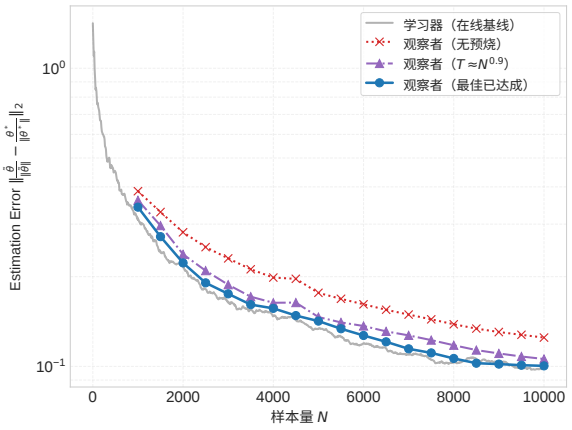
（a）LinTS（$K = 50$）



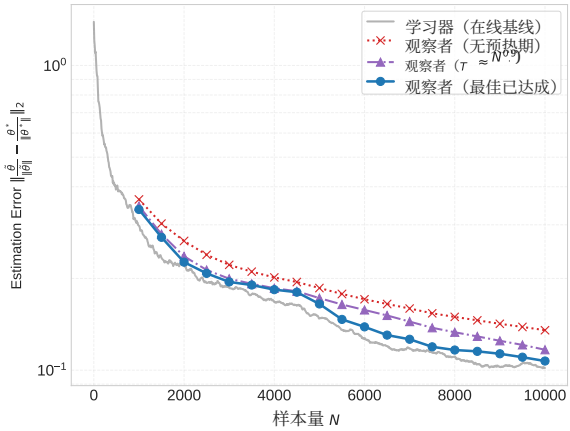
（b）LinUCB（$K = 50$）



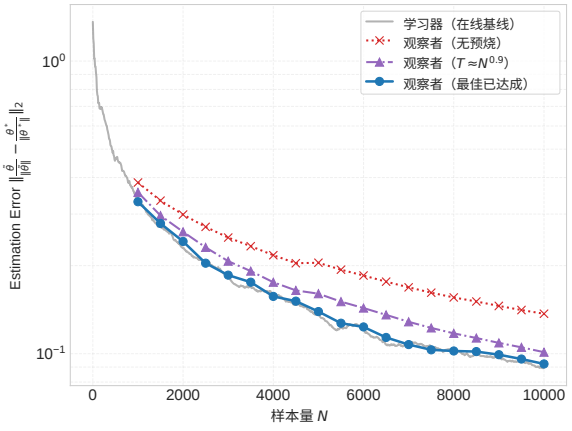
(c) LinTS ($K = 100$)



(d) LinUCB ($K = 100$)

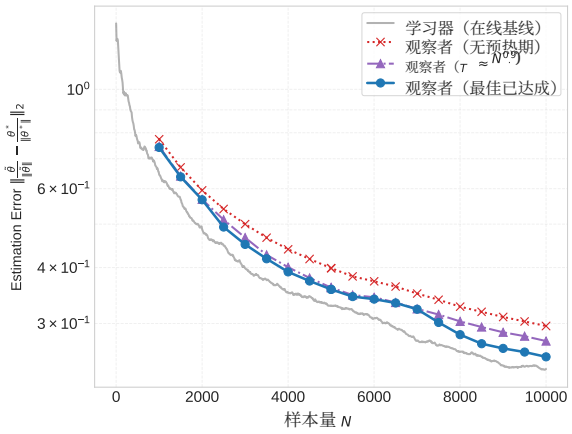


(e) LinTS ($K = 200$)

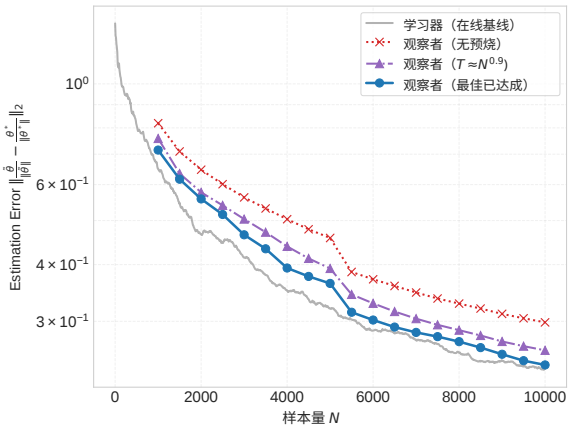


(f) LinUCB ($K = 200$)

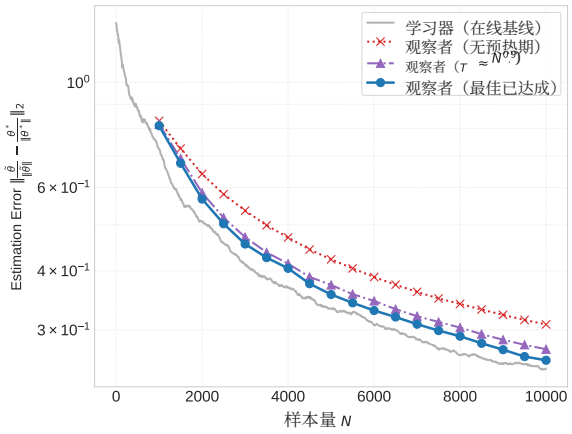
图9：不同臂数 $K \in \{50, 100, 200\}$ 下观察者策略与学习者基线的性能比较（$d = 20$）。每行对应一个固定的 K；左：LinTS，右：LinUCB。



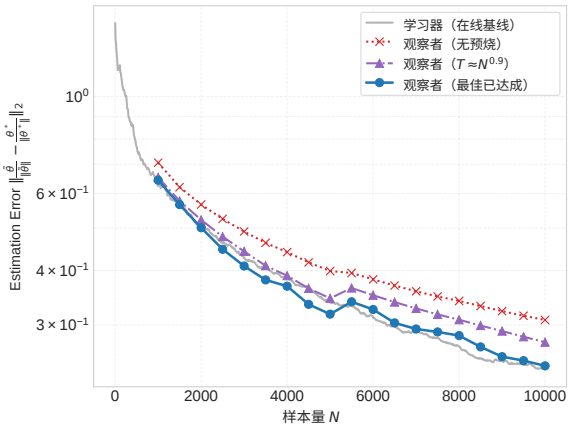
(a) LinTS（$K = 50$）



(b) LinUCB（$K = 50$）



(c) LinTS ($K = 100$)



(d) LinUCB ($K = 100$)

图10：在臂数 $K \in \{50, 100\}$ 不同时，观察者策略与学习者基线 ($d = 50$) 的性能比较。每行对应一个固定的 K；左：LinTS，右：LinUCB。