arXiv:2305.02955v1 [stat.ML] 4 May 2023

加权计数赌博机：通过重复暴露最优性克服棘手性

DhruvMalik Conor Igoe Yuanzhi Li Aarti Singh

卡内基梅隆大学机器学习系

2023年5月5日

摘要

在在线学习的推荐系统或众包应用中，人类的偏好或能力往往是算法近期动作的函数。受此启发，大量研究对这样一类设定进行了形式化：某个动作的损失取决于该动作在此前 m 个时间步中被执行的次数，其中 m 对人类记忆容量给出了上界。为了更忠实地刻画人类记忆随时间衰减的现象，我们提出了加权计数老虎机（WTB），将这一设定推广为：某个动作的损失取决于在最近 m 个时间步中该臂被选择次数的加权总和。若无进一步假设，WTB 设定难以处理。因此，我们在重复暴露最优性（REO）条件下研究这一问题。该条件受到人体生理学文献的启发，要求存在一个动作，反复执行该动作最终会比任何其他动作序列带来更小的损失。我们研究了 REO 条件下 WTB 中完全策略遗憾（CPR）的最小化问题；CPR 是最强的遗憾概念。由于 m 通常未知，我们假设只能获得其上界 M，满足 m。我们证明，对于包含 K 个动作、时域为 T 的问题，对逐次淘汰算法进行简单修改后，可实现 $\widetilde{\mathcal{O}}\left(\sqrt{KT} + (m+M)K\right)$ CPR。有趣的是，除 $(m+M)K$ 中加法因子（而非乘法因子）外，这一结果恢复了具有传统遗憾的简单随机多臂老虎机的经典保证。此外，我们还证明，在我们的设定中，任何算法都将承受 $\Omega(mK+M)$ 的加法 CPR，这表明我们的结果近乎最优。我们的算法计算高效，实验结果也证明了其实用性及其相较于自然基线方法的优越性。

1 引言

当在线学习算法部署在交互式应用中时，算法的决策会影响环境状态。反过来，这也会影响算法后续决策的质量。在推荐系统或众包等以人为中心的应用中，这一点尤为明显。例如，考虑这样一种众包场景：在每个时间步，我们都要选择一名工人执行任务，却事先不了解任何工人的能力。任务可能较为复杂，或需要进行一些微调；每名工人可能都需要经过一段校准期，反复执行任务，之后才能展现出其真正的表现。这种校准期的存在，在需要视觉运动校准的任务中已得到广泛证实 [Ada61], ，例如投掷飞镖 [WHFM20] 或投篮 [PMR+20]。因此，如果一种算法让工人轮流执行任务，却不合理地给予每名工人

校准任务的时间，就可能对其对每名工人真实能力的估计产生偏差。算法与其环境之间的这种交互，使这一场景有别于多臂老虎机等经典的非交互式框架。

为了刻画这种交互性的一个方面，在线学习领域的一项重要研究方向探讨了这样一类场景：某个动作的损失由该动作在此前 m 个时间步 [内被执行的次数来描述 HKR16, LCM17, SLC+19, SMLV20, LHK21, ABGK22, MLS22]。数量 m 通常对应于人类记忆容量或能力的上限。例如，在上述场景中， m 将表示工人对任务进行微调和校准、从而展现其真实能力所需的时间步数。

当然，这类设置只是对现实的近似。例如，心理学研究表明，人类通常对近期发生的事件记忆得更清楚 [Kla80, RVC16]。因此，如果我们在之前的 m 个时间步中执行过某个动作，那么无论该动作是在前一个时间步执行的，还是在 m个时间步之前执行的，它对当前的影响都可能有很大差异。在上述众包设置中，与工人许多时间步之前执行过任务相比，如果工人在前一个时间步执行过该任务，可能需要更短的校准期。然而，以往的形式化方法并未考虑这种差异。基于这些考量，我们做出如下贡献：

- 我们引入加权计数老虎机（WTB），通过要求动作的损失由该动作在之前 m 时间步中被执行次数的加权求和来描述，从而推广了先前的形式化定义。由于这一设定具有动态性和交互性，我们不采用传统遗憾，而是研究最强遗憾概念——完全策略遗憾（CPR）的最小化。

- 我们证明，在 WTB 中最小化 CPR 通常是难以处理的。因此，我们在额外满足重复暴露最优性（REO）的条件下研究这一问题。该条件要求存在一个动作，重复执行 m 次后，其损失会小于其他动作序列。在上述示例中，REO 意味着存在一名工人，一旦经过任务校准，其表现就会优于其他（已校准或未校准的）工人。我们通过人体生理学相关文献来阐释这一条件。

- 对于具有 K 个动作且时域为 T、满足 REO 的 WTB问题，并且在仅知道真实值 M 的上界 m 的情况下，我们证明，对经典逐次淘汰算法进行略微修改后，可达到（至多相差一个对数因子的） $\widetilde{\mathcal{O}}\left(\sqrt{KT} + (m+M)K\right)$ CPR保证。除去在$(m+M)K$中的一个加性因子外，这与随机多臂老虎机较弱的传统遗憾的下界相匹配（随机多臂老虎机是满足 REO 的 WTB 的特殊 $m = 1$ 情形）。

- 虽然人们可能希望有一种能够完全自适应于 m 且不需要此类上界 M的算法，但我们证明这是不可能的。具体而言，我们证明，任何 CPR 次线性的算法都必须依赖此类上界 M，并进一步证明，对该输入 M 的线性依赖是必要的。这意味着，在我们的设定下，任何算法所能达到的 CPR 都有一个下界 $\widetilde{\mathcal{O}}\left(\sqrt{KT} + mK + M\right)$，凸显了我们的算法接近最优。

- 通过多种数值模拟，我们证明了该（计算高效）方法的实用性，以及其相较于各种基线方法的优越性。

2 问题表述

2.1 加权计数老虎机

我们首先正式介绍加权计数老虎机：这是一种在时间范围 T 内、具有老虎机反馈的在线学习博弈，玩家可使用一个基数有限的动作集 $\mathcal{X}$ K。大量已有研究考察了这样一种情形：某个动作在任一时间步的损失，是其在此前 m 个时间步中被执行次数的函数 [HKR16、LCM17、SLC$^+$19、SMLV20、LHK21、ABGK22、MLS22]。我们将这些情形称为"计数"情形。我们的目标是推广这项研究，将其扩展到这样一种情形：某个动作的损失是其在过去 m 个时间步中被执行次数的加权计数函数。

为此，我们首先介绍一些记号。假设玩家已经进行了 t 个时间步的博弈，并且对于每个时间步 $1 \le t' \le t$，玩家执行动作 $a_{t'}$。对于一个固定动作 $x \in \mathcal{X}$，我们以逐分量的方式定义向量 $y^{t,x,m} \in \{0,1\}^m$：

$$y_i^{t,x,m} = \begin{cases} \mathbb{I}(a_{t-i+1} = x) & \text{if } t-i+1 \ge 1 \\ 0 & \text{if } t-i+1 < 1 \end{cases},$$

对于每个分量 $1 \le i \le m$。因此，向量 $y^{t,x,m}$ 存储了在之前 m 个时间步直至（并包括）当前时间步中执行动作 x 的时间步。借助这一记号，我们现在可以正式定义加权计数老虎机。

定义 1（加权计数老虎机（WTB））。如果一个在线学习博弈满足：存在一个整数 $m \ge 1$、一个向量集合 $\{w_x\}_{x\in\mathcal{X}} \subset (0,1]^m$，以及一组函数 $\{h_x\}_{x\in\mathcal{X}}$，其中每个函数都从 $\mathbb{R}$ 映射到 $[0,1]$,，使得以下条件成立，则称其为一个具有记忆容量 m 的 (m, w, h) 加权计数老虎机。对于每个 $x \in \mathcal{X}$，在时间步 t 执行动作 $a_t = x$ 所产生的期望损失为

$$h_x\left(w_x^\top y^{t,x,m}\right),$$

并且玩家会观察到一个随机观测作为反馈 $\widetilde{h}_x(w_x^\top y^{t,\ x,\ m}) \in [0,1]$，该观测独立于所有其他随机观测，并满足

$$\mathbb{E}\left[\widetilde{h}_x(w_x^\top y^{t,x,m})\right] = h_x\left(w_x^\top y^{t,x,m}\right).$$

一般来说，量 m，$\{w_x\}_{x\in\mathcal{X}}$，$\{h_x\}_{x\in\mathcal{X}}$都是未知的，玩家只能随着时间推移，通过老虎机反馈了解它们。当 $m = 1$ 时，WTB 可还原随机多臂老虎机（sMAB） [LR85, ACBF02]。然而，对于需要校准的以人为中心的应用而言，采用 $m \ge 2$ 的 WTB 往往是更好的模型。为了理解这一点，我们具体来看第 1 节中介绍的众包应用。假设要执行的任务是在飞镖靶上投掷飞镖，每位工人都是一名不同的飞镖玩家。在不了解任何玩家击中飞镖靶的真实能力的情况下，我们的目标是从 K 名玩家中找出最优秀的一位：在每个时间步选择一位玩家投掷飞镖，并观察其是否命中。乍一看，这似乎是一个经典的 sMAB 问题，其中每位玩家都有某种真实能力，而我们每次向一位玩家发出请求时，都会（随机地）观察到其真实能力。

遗憾的是，这种 sMAB 表述并未考虑飞镖选手在展现其真实表现之前所需要的校准期。视觉运动校准领域的相关文献已经证实了这一校准期的存在。例如，Wunderlich 等人

al. [WHFM20] 表明，职业飞镖选手连续投掷飞镖时，第一次投掷的准确度明显低于之后的投掷，但从第一次投掷后，表现便趋于稳定。他们将这一现象归因于热身递减 [Ada61, AW93, Ans95]„ 即特定运动技能因中断而导致表现下降，并在恢复该技能后逐渐回升。简而言之，玩家一旦“动起来”，并在第一次投掷后微调好动作参数，表现就会更好。

这一现象影响了飞镖情境下算法的设计，因为我们直到飞镖玩家第一次投掷后，才能观察到其真实表现。此外，这无法 仅靠让每位玩家各投一次以完成校准，然后假定玩家永远保持校准状态并运行标准 sMAB 算法来解决。事实上，Wunderlich 等人 [WHFM20] 证明，即使飞镖投掷中只有短暂的中断（例如玩家从靶盘上取回飞镖所需的几秒钟），也可能导致玩家“重置”，并在无意识中失去经过精细调校的动作参数。因此，sMAB 并不是这一情境的理想模型。相比之下，带有 $m \geq 2$ 的 WTB 是更贴切的模型，因为 m 描述了我们随机观察到玩家真实表现之前，玩家必须连续投掷飞镖的次数。这个动机示例（以及我们接下来将介绍的示例）中的“重置”现象意味着，若用 WTB 对该问题建模，那么 m 应当显著小于时域 T。本文将始终作此假设。

与上述计数设置 [HKR16、LCM17、SLC+19、SMLV20、LHK21、ABGK22、MLS22]，相比，WTB 对这一现象的建模更加自然；这些设置都是 WTB 的特殊情形，其中 w_x 对每个 $x \in \mathcal{X}$ 都是全 1 向量 $\vec{1}$。作为一个简化示例，假设任务是投篮球罚球，我们需要从两位玩家 x_1、 x_2 中找出表现最好的一位。考虑两种不同的玩家选择序列—— x_1、 x_2、 x_1 与 x_2、 x_1、 x_1。Phatak 等人 [PMR+20] 表明，玩家投罚球时需要至少 3 次的校准期，而且他们的投篮表现会随着每次连续罚球单调提升。这意味着，选择 x_1、 x_2、 x_1（即 x_1 投篮，然后是 x_2，再然后是 x_1）会导致 x_1 最后一投的期望表现较差，相较于选择 x_2、 x_1、 x_1 （即 x_2 投篮，然后 x_1 连投两次时的表现）。如果我们用 $m = 3$ 和 $w_x = \vec{1}$ 对此进行 WTB 建模，那么就无法区分这两种情形，因为两种情况下 x_1 在过去的 m 个时间步中都投过两次。相比之下，采用 $w \neq \vec{1}$ 的 WTB 可以为这两种情形建模出不同的损失。例如，如果我们使用模型 $w_{x_1} = [1, 1/2, 1/4]$ 和 $h_{x_1}(z) = 1 - z/3$，那么该模型表明，选择 x_2、 x_1、 x_1 会使 x_1 最后一投的期望表现优于选择 x_1、 x_2、 x_1。

更广泛地说，WTB 显著推广了先前的计数设置，使我们能够更准确地近似记忆强度随时间推移而衰减的现象；这一现象已在人类短期和长期记忆研究中得到广泛记载 [Kla80, RVC16]。这也更自然地刻画了促使人们研究计数设置的以人为中心的应用。例如，Malik 等人 [MLS22] 以推荐系统为背景阐述其研究动机，认为推荐内容会影响人类偏好，并假设数量 $m \ll T$ 限定了人类记住过去推荐内容的时长。但他们的表述并未考虑某项内容是在时长为 m 的窗口内多近被推荐的。因此，如果某项内容在过去 k 个 m 时间步内被推荐了 k 次，那么无论这 k 次推荐的先后顺序如何，他们的框架都认为由此产生的损失相同。这一限制相当明显，因为人类~当下的偏好可能只会轻微受到 $\Omega(m)$ 个时间步之前发生的推荐影响。我们的 WTB 表述更加精细，允许这 k 次推荐的不同排列顺序分别导致不同的损失。

2.2 完全策略遗憾

WTB 的一个关键性质是，某个动作造成的损失取决于算法过去采取的动作。在这类动态场景中，已有充分研究表明，常用的传统遗憾并不适合衡量算法的表现 [ADT12]。相反，通常会采用更强的策略遗憾概念 [CBDS13, ADMM18]。遵循计数设置领域的先前研究 [HKR16, LCM17, SLC+19, SMLV20, LHK21, ABGK22, MLS22], ，我们研究完全策略遗憾（CPR）的最小化；它是最强的遗憾概念。给定一个 (m, w, h)-加权计数赌博机，以及一个执行动作序列 $(a_1, a_2 \dots a_T) \in \mathcal{X}^T$ 的算法，该算法的 CPR $\mathcal{R}^{\mathrm{cp}}$ 定义为

$$\mathcal{R}^{\mathrm{cp}} = \sum_{t=1}^{T} h_{a_t}\left(w_{a_t}^{\top} y^{t,a_t,m}\right) - \min_{(x_1,x_2\dots x_T)\in\mathcal{X}^T} \sum_{t=1}^{T} h_{x_t}\left(w_{x_t}^{\top} y^{t,x_t,m}\right). \tag{1}$$

按照以往的惯例，我们将任意长度为 T 的动作序列（即 $\mathcal{X}^T$ 的一个元素）称为策略。因此，CPR 是算法所经历的累积损失，相对于 $\mathcal{X}^T$ 中最佳策略所能达到的最低损失而言。因而，最小化 CPR 等价于最小化算法的累积期望损失。值得注意的是，这一表现指标与强化学习中使用的指标相同 [JAZBJ18, WDYK20]。如果某算法的 CPR 关于 T 呈次线性，并且关于 m, K 呈多项式，则称其具有统计高效的 CPR。

先前的研究表明，对于每个 $x \in \mathcal{X}$ 都满足 $w_x = \vec{1}$ 的 WTB 情形，无需任何进一步假设，就存在一种具有统计高效 CPR 的算法 [MLS22]。遗憾的是，以下结果表明，在满足 $w_x \neq \vec{1}$ 的 WTB 中，这样的算法并不存在。

命题 1. 对于任意 $m \geq 1$，都存在一个具有 $K = 2$ 的(m, w, h)-加权计数老虎机，使得以下结论成立：任何（可能随机化的）算法，其期望 *CPR* 均满足 $\mathbb{E}\left[\mathcal{R}^{cp}\right] = \Omega\left(\min\{2^m, T\}/m\right)$。

命题 1 的证明推迟至附录C。粗略而言，证明表明，如果 $w_x \neq \vec{1}^{\sim}$，那么 h_x 可以取 $\Omega\left(2^m\right)$ 个不同的值，因此，发现最优动作序列 $^{\sim}$ 需要 $\Omega\left(2^m\right)$ 次查询。该结果表明，如果我们希望算法具有统计高效的 CPR，就必须对 WTB 设置施加结构性约束，以限制最优动作序列的集合。我们将在下文阐述并形式化这种结构。

2.3 重复暴露最优性

为说明 WTB 所建模的问题类型中所需的其他结构，我们回顾第2.1节所示的飞镖情境。值得注意的是，如果让一位玩家连续投掷飞镖，那么在第一次 投掷时，其未校准表现较差，且未必能反映其后续表现。但在第一次投掷之后的连续投掷中，Wunderlich 等人 [WHFM20] 表明，其校准后的表现会稳定下来，并优于 第一次投掷时的未校准表现。罚球 [PMR+20]也有类似的观察结果。因此，若令 $x^\star \in \mathcal{X}$ 表示校准后表现最佳的玩家，那么这意味着其校准后的表现$x^\star$ 不仅优于玩家 $x \neq x^\star$的校准后表现，也优于所有玩家的未校准表现。我们将这一见解形式化为以下条件。

定义 2（重复暴露最优性（α-REO））。一个(m, w, h)-加权计数赌博机满足参数为 α 的重复暴露最优性条件，若存在一个动作 $x^\star \in \mathcal{X}$，使得对于每个 $x \in \mathcal{X}$ 和每个 $y \in \{1\} \times \{0,1\}^{m-1}$，都有

$$h_{x^\star}\left(\|w_{x^\star}\|_1\right) \leq h_x\left(w_x^\top y\right) + \alpha.$$

因此， α-REO条件要求存在某个动作 $x^\star \in \mathcal{X}$，当连续重复执行至少 m 次时，其损失会小于其他动作序列的损失（差距不超过次优性 α）。为理解这一条件在既有研究中的意义，有两点需要说明。第一，注意到，即使我们进一步对WTB施加 α-REO条件，只要选择 $\alpha = 0$， $m = 1$，sMAB仍然是这一设定的一个特例。第二，关于计数设置的大量既有研究都聚焦于损失函数 $\{h_x\}_{x\in\mathcal{X}}$ 具有单调性的情形。例如，改进型赌博机 [HKR16]是WTB的一个特例，但需要满足诸多额外限制，其中包括（但不限于） $\{w_x\}_{x\in\mathcal{X}} = \{\vec{1}\}$ 和 $\{h_x\}_{x\in\mathcal{X}}$ 递减。我们指出，函数 $\{h_x\}_{x\in\mathcal{X}}$ 递减这一性质是0-REO条件的一个特例。

我们以心理生理学文献中已有大量实证研究的热身递减现象为依据，阐释了 REO。我们也认为，正如第 6 节所讨论的，REO 可能同样适用于推荐系统等其他交互场景。尽管如此，我们承认，我们的设定归根结底是一种数学抽象，无法完全贴合真实情况，也未能刻画以人为中心的应用所面临的诸多复杂细节。全面地形式化并研究所有这些细节超出了本文的范围，因此我们将有关未来重要研究方向的讨论留待第 6 节展开。

有了上述动机并对 REO 条件作出形式化定义后，自然会提出以下问题：

考虑任意满足 α-*REO* 的 (m, w, h)-加权计数老虎机。是否存在一种计算高效且实用的算法，能够解决这一问题并提供统计高效的 *CPR* 保证？

本文接下来的分析均致力于回答这一问题。

3 主要结果

我们给出两类结果。在第3.1节中，我们提出一种统计高效且计算高效的算法，可解决满足 REO 条件的 WTB问题。该方法只需要知道真实记忆容量 m 的一个上界 M，而其确切值往往未知。在第 3.2 节中，我们证明，不存在能够完全自适应于未知 m 的算法（即无需知道 m 的上界 $M < T$）。我们还证明，如果已知 m 的一个上界 $M < T$，那么我们的方法对 M 的依赖是最优的。

3.1 一种统计与计算高效的算法

在介绍我们的算法之前，我们先来考察一些自然的方法。由于 WTB 是强化学习（RL）问题的一个子类，我们可以尝试用强化学习算法来解决它。但即使在 $\{w_x\}_{x\in\mathcal{X}} = \{\vec{1}\}$ 的情况下，这类算法仍会遭受 $\Omega(K^m)$ CPR [ABGK22, MLS22] 的损失。我们也可以尝试将经典的 sMAB UCB算法扩展到 WTB，方法如下：将问题分成长度为 m的轮次，在每轮开始时选择使常规 UCB估计最小的动作，然后连续执行该动作 m 次，而不是只执行一次。接着我们

记录最近一次执行中观察到的损失，因为它是该动作最终损失的无偏估计，并用它来更新该动作的 UCB估计。虽然这看起来是合理的启发式方法，但每次分轮都会产生长度为 m 的开销，这可能会大幅增加遗憾。

算法 1 WTB 的逐次淘汰要求：记忆容量 m 的上界 M、时间范围 T、失败概率容忍度 $\delta \in$ （0, 1）、动作数量 K1: 定义 $S = \log_2\left(\frac{T}{4KM}+1\right)$。2: 定义 $A_1 = \mathcal{X}$。3: 定义 $n_s = KM2^s/|A_s|$ 和 $C_s = \sqrt{\frac{32}{n_s}\log\left(\frac{2KS}{\delta}\right)}$。4: 对 $s \in \{1,2\dots S\}$ 执行5: 对 $x \in A_s$ 执行6: 执行动作 x 共 $n_s \geq m$ 次，不存储任何内容。~~7: 执行动作 x 共 n_s 次，并存储 $\{h_x(\|w_x\|_1)_{s,k}\}_{k=1}^{n_s}$。8: 定义 $\widehat{\mu}_s(x) = \frac{1}{n_s}\sum_{k=1}^{n_s}\widetilde{h}_x(\|w_x\|_1)_{s,k}$。9: 结束内层循环10: 选择 $\widehat{x}_s \in \arg\min_{x\in A_s}\widehat{\mu}_s(x)$。11: 构造 $A_{s+1} = \{x \in A_s$，使得 $\widehat{\mu}_s(x) \leq \widehat{\mu}_s(\widehat{x}_s) + 2C_s\}$。12: 结束外层循环~~

—

另一种思路是将先前计数设置中的算法改造后用于我们的问题。然而，与我们的设置最为相近的先前计数设置，其 CPR 界都随 m 呈乘法增长（详情见第 5 节）。我们的关键理论贡献在于证明，得益于 REO 的额外作用，我们不仅能够解决这些计数设置中的问题，还能以仅随 m 呈加法 增长（而非乘法增长）的 CPR 保证来解决 WTB。达到这一界的算法是经过轻微修改的逐次消除（SE），详见算法 1。我们的灵感来自 Malik 等人 [MLS22], ，他们针对其计数赌博机设置对 SE 进行了改造，不过他们的修改更为复杂。相比之下，我们的修改非常简单，因为 REO 使我们只需估计每个动作的最终损失。现在我们给出主要结果，它对该算法的 CPR 进行了界定。

定理 1. 固定任意满足参数为 α 的重复暴露最优性的（m，w，h）-加权计数老虎机问题。当算法 *1* 使用输入 $M \geq m$ 和 $\delta \in$ （0，1)运行时，其完全策略遗憾以至少 $1-\delta$ 的概率满足以下上界

$$\mathcal{R}^{cp} \leq 4KM + Km\log(T) + 800\sqrt{KT\log\left(\frac{2K\log(T)}{\delta}\right)} + \alpha T. \tag{2}$$

定理的证明　　引理 1 的证明推迟至附录 A。下面我们重点介绍这一结果的几个关键方面。　.

与 sMAB 和计数设置的比较。回想一下，在经典 sMAB 中（它是通过 $m = 1$ 实现 0-REO 的 WTB 特例），任何算法都会遭受 $\widetilde{\Omega}\left(\sqrt{KT}\right)$ 传统遗憾。因此，定理 1 表明，对于更广泛的 WTB $\widetilde{\mathcal{O}}\left(\sqrt{K/T}\right)$-REO 问题类别，我们基本上可以在 CPR 上达到同样的保证，仅相差一个对数因子以及对 mK 的加性依赖。与先前针对类似计数

设置（详见第 5 节）。

效率与实用性。算法 1 计算高效且可扩展。在 T 次迭代中的总运行时间为 $\mathcal{O}(T + K\log(T))$，任一时间步所需的空间复杂度为 $\sim\mathcal{O}(K)$。这与先前针对类似计数设置的结果形成对比（详情见第 5 节）。此外，实现算法 1 无需精确了解诸如 $\{h_x\}_{x\in\mathcal{X}}$、$\{w_x\}_{x\in\mathcal{X}}$、$\alpha$ 或 m 等未知量；只需知道 m 的一个上界 M 即可。虽然算法 1 看起来需要将时间范围 T 作为输入，但我们注意到，该方法本身已经在执行倍增技巧。这意味着，对于计算机上可表示的任意 T（例如 $T \leq 2^{64} = 2^{2^6}$），并且由于 $C_s = \widetilde{\mathcal{O}}(\log\log T)$，通过简短的数值计算可以发现，将 C_s 重新定义为 $\sqrt{\frac{64}{n_s}\log(2K/\delta)}$ 并选取 $\delta < 0.009$，即可确保即使不将 T 作为输入，也能获得与公式（2）相同的 CPR 界（差异仅在于常数因子）。

各种情形下的统计最优性。在已知 m（因此 $M = m$）且 REO 在 $\alpha = 0$ 下得到满足的情形中，定理 1 的保证在一个对数因子范围内是最优的。为了解释这一点，请注意，在公式（2）的右侧，Km项由于 Malik 等人的命题 1 而无法改进。[MLS22], 而且，$\sqrt{KT}$ 项显然由于经典的 s MAB 下界而是紧的。此外，当已知 m 且 REO 在 $\alpha = \widetilde{\Theta}\left(\sqrt{mK/T}\right)$ 下得到满足时，Malik 等人的定理 2 的证明 [MLS22]表明，存在一个非平凡的 $0 < \alpha \ll 1$ 情形，在该情形下，定理 1 中对 αT 的依赖无法改进，因此定理 1 是最优的（在一个对数因子范围内）。我们注意到，目前尚不清楚公式（2）中的 αT 项对于所有 $\alpha > 0$是否都是最优的，对此进行研究是一个有趣的未来方向。对于 m 未知时的情形，我们将关于 M 最优依赖关系的研究推迟至第 3.2 节。

最佳臂识别。算法 1 也可用于识别最终损失接近 $x^\star$ 的动作。具体而言，经过 T 轮次（或 S 个周期）后，以至少 $1-\delta$ 的概率，任意动作 $x \in A_{S+1}$ 都满足 $h_x(\|w_x\|_1) \leq h_{x^\star}(\|w_{x^\star}\|_1) + 4C_s = h_{x^\star}(\|w_{x^\star}\|_1) + \widetilde{\mathcal{O}}\left(\sqrt{K/T}\right)$。

定理 1 的证明需要谨慎处理，以确保依赖关系最优，但所用技术是标准的，我们的贡献并非提出一种新颖的分析路径。相反，我们的贡献在于证明，针对规范随机多臂老虎机的经典算法可以轻松应用于解决一类范围广得多、表面上也复杂得多的问题，而这类问题在实践中有充分的动机。与我们的 WTB 可比的先前计数设置存在固有的计算和统计困难（详情参见第 5 节）。我们认为，对 REO 的形式化以及定理 1，是对一种充分有据的结构的重要识别；这种结构使得交互式以人为中心的应用中出现的问题能够得到统计和计算高效的解决方案。

3.2 对记忆容量的适应性

虽然算法 1 不需要知道真实的记忆容量 m，但确实需要知道 m 的一个上界 M。定理 1 表明，算法 1 的 CPR 随这一输入 M 线性增长；在难以对 m 给出非平凡上界的情形下，这一点并不理想。一般而言，我们希望算法能以更有利的方式（或完全不）依赖输入 M。例如，可以设计一种算法来维护真实值 m 的置信区间，并通过自适应查询来细化对 m 的估计，从而减弱或消除其

对 M 的依赖。现在我们将证明，这样的算法并不存在，即使是在 WTB 的一个特例——更简单的"计数设置"中，并且在 REO 以参数 $\alpha = 0$ 成立的情况下也是如此。

为此，我们引入一些记号。对于满足 $M \leq T$ 的任意正整数 T、M、K，令 $\mathrm{UTB}_{T,M,K}$ 表示无权计数赌博机问题的集合（即对于每个动作 x，w_x 均为全 1 向量的 WTB问题），其中每个问题的时域长度为 T、动作数量为 K、记忆容量为 $m \in \{1, 2 \dots M\}$，且满足 0-REO。对于任意可能是随机化的算法 $\mathcal{A}$ 以及任意无权计数赌博机问题 tb，令 m_{tb} 表示 tb 的记忆容量，并令 $\mathcal{R}^{\mathrm{cp}}(\mathcal{A}, \mathsf{tb})$ 表示算法 $\mathcal{A}$ 用于求解 tb 时的期望 CPR。对于满足 ϵ_1、$\epsilon_2 \in \in (0, 1)$ 且 $\epsilon_3 \in [0, \epsilon_2$ 的 $\epsilon = (\epsilon_1, \epsilon_2, \epsilon_3)$ 的一种取值，以及函数 $f\colon \mathbb{R}^2 \to \mathbb{R}$ 的一种取值，令 $\mathcal{A}_{\epsilon,f}$ 表示算法 $\mathcal{A}$ 的集合。这些算法在接收任意满足 $M \leq T$ 的正整数 T、M、K（且不接收其他信息）作为输入时，对每个问题实例 $\mathsf{tb} \in \mathrm{UTB}_{T,M,K}$，均满足

$$\mathbb{E}[\mathcal{R}^{\mathrm{cp}}(\mathcal{A}, \mathsf{tb})] \leq \min\left\{T/4, f(m_{\mathsf{tb}}, K)\left(T^{1-\epsilon_1} + T^{\epsilon_3} M^{1-\epsilon_2}\right)\right\}. \tag{3}$$

集合 $\mathcal{A}_{\epsilon,f}$ 中的算法 $\mathcal{A}$ 因此对 M 的依赖是良性的，具体来说：对于任意正整数 T, M, K，满足 $M \leq T$，以及任意问题实例 $\mathsf{tb} \in \mathrm{UTB}_{T,M,K}$，算法 $\mathcal{A}$ 事先并不知道 tb 的记忆容量 m_{tb}，只知道其上界M。尽管如此，$\mathcal{A}$在求解 tb 时的 CPR 随界 M以次线性 速度增长。不幸的是，以下结果表明这样的算法并不存在。

定理 2. 对于每个满足 ϵ_1 的 ϵ，$\epsilon_2 \in (0,\ 1)$ 且 $\epsilon_3 \in [0, \epsilon_2)$，以及每个函数 f，相应的集合 $\mathcal{A}_{\epsilon,f}$ 为空集。

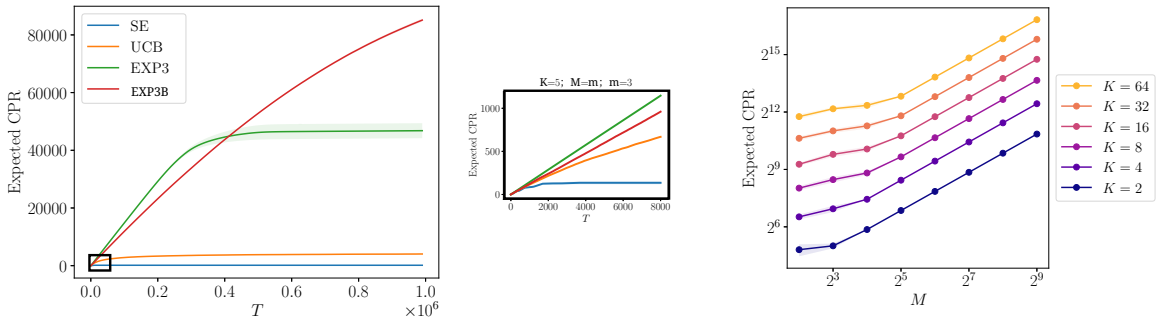
证明见附录 B。该结果揭示了"适应性代价"（例如，在不同背景下的类似结果，参见 [LC18] ），说明如果我们只有未知真实记忆容量的上界 M，那么任何算法的 CPR 都不可能同时关于 M 和 T 呈次线性增长。我们通过两个重要推论具体说明这一点。以下推论针对我们没有关于记忆容量的非平凡界的情形，或者等价地说，$M = T$。

推论 1。任取函数 $f\colon\ \mathbb{R}^2 \to \mathbb{R}$。不存在一种（可能随机化的）算法，使得对于每个 $\mathsf{tb} \in \mathrm{UTB}_{T,\ T,\ K}$，其期望 *CPR* 均有界于 $\widetilde{o}(T) f(m_{\mathsf{tb}}, K)$。

推论 1 的结果表明，不可能存在这样一种算法：对于所有满足 0-REO、时域为 T 的无加权计数赌博机实例 tb，其 CPR 关于 T 是次线性的，即使允许其对 m_{tb}, K 的依赖任意差。因此，若要获得定理 1 所提供的那类次线性 CPR 保证，就必须知道真实记忆容量的某个上界 $M < T$。下一个推论针对已知记忆容量的一个非平凡上界 $M < T$ 的情形。

推论 2。任取函数 $f\colon\ \mathbb{R}^2 \to \mathbb{R}$。不存在一种（可能随机化的）算法，能够在给定输入 M 的情况下，使得对于每个 $\mathsf{tb} \in \mathrm{UTB}_{T,\ M,\ K}$，其期望 *CPR* 均有界于 $f(m_{\mathsf{tb}}, K)(\widetilde{o}(T) + \widetilde{o}(M))$。

因此，该结果表明，对于满足 0-REO、时域为 T 的所有无加权计数赌博机实例 tb，即使允许对 K, m_{tb} 有任意差的依赖关系，我们也不能指望存在一种算法，能够对 M 和 T 都具有次线性依赖。注意，忽略对数因子，定理 1 表明，对于每个 $\mathsf{tb} \in \mathrm{UTB}_{T,M,K}$，算法 1 的 CPR 上界为 $\widetilde{\mathcal{O}}\left(\sqrt{KT} + K(M + m)\right)$。结合此前对定理 1 的讨论，推论 2 因而表明，任何算法都必须承受 $\widetilde{\Omega}\left(\sqrt{KT} + mK + M\right)$ CPR，这凸显子算法 1 对满足 0-REO 的 WTB问题近乎最优，即使在我们仅掌握未知真实记忆容量上界 M 的情况下也是如此。



(a) CPR 随时间范围 $T = 10^6$ 的变化。

(b) CPR 关于 M 的函数。

图 1：我们绘制了每种算法的期望 CPR。在两个图中，每个数据点均取自对 20 个问题实例求平均的结果，阴影区域表示均值周围的 ± 1 个标准误。(a) 中我们固定$K = 5$、 $m = 3$ 和 $M = 3$。(b) 中我们固定 $m = 4$ 和 $T = 10^6$。

4 数值结果

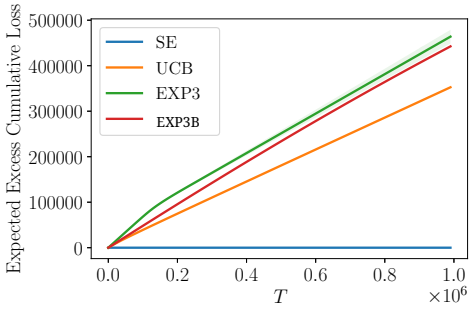
本节中，我们评估算法 1（记作 SE）在不同领域中的表现，这些领域均建模为满足 REO 的 WTB问题。在每个领域中，我们将其表现与以下三种基线方法进行比较——(A) EXP3算法 [ACBFS02], ，该算法在我们的设定中具有次线性的传统遗憾；(B) Arora 等人 [ADT12], 所述的 EXP3 批处理版本，记作 EXP3B，该算法在我们的设定中具有统计高效的CPR保证；(C) 第 3 节所述的改进 UCB算法。

4.1 无权计数赌博机上的合成损失函数

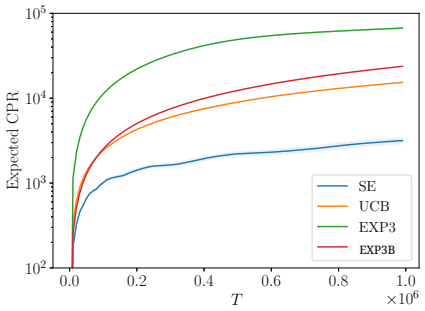
我们考虑 $\{w_x\}_{x\in\mathcal{X}} = \{\vec{1}\}$，并固定某个 $x^\star \in \mathcal{X}$。我们为每个 $x \in \mathcal{X}$ 定义 $h_x = 0.5$，但作如下修改： $h_{x^\star}(\|w_{x^\star}\|_1) = 0.35$。因此，除非我们至少执行 $x^\star$ m 次，否则损失都相同，这意味着该实例满足 0-REO。为定义反馈模型，我们要求随机变量 $\widetilde{h}_x(w_x^\top y^t, x, m)$ 服从参数为 $h_x(w_x^\top y^t, x, m)$ 的伯努利分布。当 $m = 1$ 时，这是 sMAB [Sli19], 的一个困难实例，而 UCB 显然是最优的。对于 $m > 1$，我们注意到，UCB 的变体在 $h_x \approx h_x(\|w_{x^\star}\|_1)$ 的情况下表现最佳，因为在长度为 m 的开销期间产生的损失，与反复执行某个动作所导致的最终损失几乎相同。因此，我们认为实验设计已尽可能有利于 UCB 变体。在图 1a 中，我们绘制了各方法随时间变化的期望 CPR。正如预期，SE 的表现优于所有基线方法。附录 D.1 展示了其他 m、 K、 M 取值下的类似结果，并给出了一个满足 α-REO 且 $\alpha > 0$ 的问题的结果。此外，对于相同固定的 m、 T、 K，我们还研究了 SE 的表现如何随其输入 M 变化。在图 1b 中，我们观察到，SE 的 CPR 至多是 M 的线性函数，这与定理 1 的预期一致。不过，我们也看到，在若干情况下，其增长是次线性的，因此优于定理 1 所预测的最坏情形线性增长。

4.2 加权计数老虎机上的合成损失函数

现在，我们考虑一个满足 REO 的 WTB问题，其中 $\{w_x\}_{x\in\mathcal{X}} \neq \{\vec{1}\}$。具体使用的损失函数将在附录 D.2 中讨论。由于这个问题的最优策略难以计算



（a）WTB中使用 $\{w_x\}_{x\in\mathcal{X}} \neq \{\vec{1}\}$时的额外损失。



（b）飞镖投掷锦标赛中的CPR。

图2：在（a）中，我们绘制了所描述的WTB实例中各算法相对于SE的额外期望累积损失随时间的变化，其中 $\{w_x\}_{x\in\mathcal{X}} \neq \{\vec{1}\}$ 在第 4.2 节中进行了描述，且 $K = 5$、 $m = 4$ 和$M = 4$。在（b）中，我们绘制了第 4.3 节所述模拟飞镖投掷锦标赛中各算法的期望CPR随时间的变化，并将 y轴截断至低于 10^2 的部分，以便于展示。在（a）和（b）中，数据均通过对20个问题实例取平均获得，阴影区域表示均值周围的 ± 1 标准误。

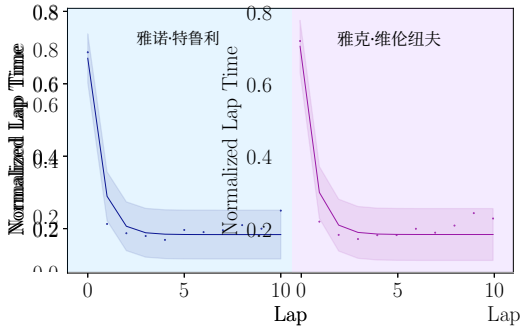
，因此 CPR 也难以计算。因此，我们没有绘制 CPR，而是绘制了各算法相对于 SE 的损失所多出的期望累积损失（因此，在任意时刻，只需对各算法的超额损失进行常数平移，即可得到 CPR）。结果如图 2a 所示，表明我们的方法优于基线方法。

4.3 模拟飞镖投掷锦标赛

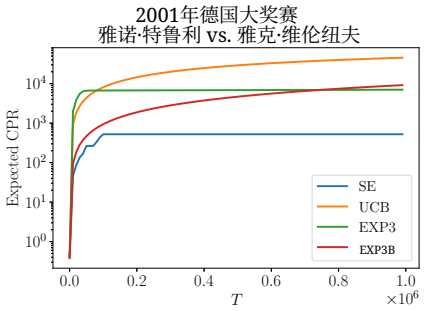
受先前研究表明运动任务中存在校准期的启发 [Ada61, PMR+20, WHFM20]，我们模拟了一场由 $K = 20$ 名玩家参与的简化飞镖投掷锦标赛。如第 2.3 节所述，Wunderlich 等人 [WHFM20] 表明，虽然玩家的第一次投掷是未校准的，不一定能反映其后续表现，但在紧接着的后续投掷中，表现会得到校准、稳定下来，并且优于第一次投掷。我们将锦标赛的每个（随机）实例建模为一个 WTB，其中包含 $m = 2$ 和任意的 w，每个玩家 $x \in \mathcal{X}$ 的期望损失函数从 $h_x(w_{x,1}) \sim \text{Unif}[0.68, 0.72]$ 以及$h_x(\|w_x\|_1) \sim \text{Unif}[0.58, 0.62]$中采样。我们根据 Wunderlich 等人 [WHFM20]，获得了这些分布的界限；他们表明，大多数玩家的平均表现集中在这些区间内。为了定义反馈模型，我们要求随机变量 $\widetilde{h}_x(w_x^\top y^{t,x,m})$服从 Bernoulii($h_x(w_x^\top y^{t,x,m})$) 分布。虽然我们的实验设计忽略了投掷飞镖时可能出现的一些现实世界中的细节（例如，未命中一次投掷可能会影响玩家下一次投掷时的信心），但我们认为，对于研究投掷飞镖达到最优表现所需的校准期而言，这是一个合理的初步模型。在图 2b 中，我们绘制了每种方法随时间变化的 CPR。正如预期，SE 的表现显著优于每种基线方法。

4.4 模拟 F1 锦标赛

在一级方程式（F1）锦标赛中，目标是从由 K 名车手组成的集合中找出速度最快的车手。锦标赛中的每位车手都必须完成若干圈。我们模拟一种改进版的 F1 锦标赛，其中 $K = 2$，并且在每个时间步选择两名车手中的一名来



(a) 概率圈速模型示意图。



(b) 时间范围 $T = 10^6$内的 CPR。

图3：在 (a) 中，我们展示了针对 2001年德国大奖赛中两位车手拟合的概率圈速模型。我们的概率模型为从 1 到 10 的每个单圈索引上的单圈时间分布设定参数。实线表示该分布在每个单圈索引上的均值。阴影区域包含该分布以均值为中心的 ± 2 个标准差。点线表示实际单圈时间。请注意，所有单圈时间都经过归一化，因此每个单圈时间都位于区间 $[0, 1]$ 内（详情见附录D.3）。在 (b) 中，我们绘制了每种算法的预期 CPR 随时间变化的函数。数据通过对 20 个问题实例取平均值获得，每个实例分别具有 $K = 2$、$m = 10$、$M = 10$ 和 $T = 10^6$，阴影区域表示均值周围的 ± 1 个标准误。

完成一圈（即在任意给定时间步，赛道上只能有一名车手）。完成一圈后，我们会观测到该车手单圈时间的一个随机实现。

值得注意的是，车手的单圈时间取决于其此前完成的圈数。为了证明这一点，我们利用1950年至 2022 [Rao22]的F1单圈时间数据，为每位F1车手拟合一个概率圈速模型（有关概率模型和数据处理的详细信息，请参见附录D.3）。在图3a中，我们展示了一对典型车手的单圈时间概率模型，并表明随着单圈索引的增加，他们的单圈时间往往会逐渐缩短。这一现象有多种合理的解释；例如，随着燃油消耗，车手所驾驶车辆的质量会逐渐减小，同时车手对赛道的熟悉程度也会不断提高。因此，我们认为，sMAB并不适合用于建模我们的F1 锦标赛。相反，将其建模为一个满足REO的WTB问题更为合适，而我们锦标赛的目标是找出校准后圈速最快的车手，而这一指标只有通过反复接触才能观测到。

我们模拟了多个改进的一级方程式锦标赛实例，每个实例包含 $K = 2$。每个实例中的两名车手经过选择，使得其校准后的表现难以区分（详见附录D.3）。此处，我们展示其中一个实例的结果。其他实例的结果见附录D.3。对于该实例，我们使用图3a所示的概率模型创建一个包含 $K = 2$ 和 $m = 10$的 WTB问题。我们记录在此前 m 个时间步中每名车手被选择的次数。选择某名车手所对应的损失由我们拟合的概率模型所参数化的分布决定。具体而言，如果我们选择车手 x，且在最近的 m 个时间步中已选择该车手 y 次，则瞬时损失将从我们拟合的概率模型针对车手 x 在单圈索引 y处所参数化的分布中采样。请注意，在此设置下，当且仅当某人多次选择较差的车手$o(T)$时，其 CPR 为 $o(T)$。在图3b中，我们绘制了该锦标赛实例中各方法随时间变化的 CPR，结果表明 SE 优于基线方法。在附录D.3所示的其他锦标赛实例中也观察到了类似结果。

5 相关工作

计数设置（ $m = T$）。大量先前研究致力于研究作为 WTB 特例的计数设置，其中 $\{w_x\}_{x\in\mathcal{X}} = \{\vec{1}\}$，并要求 $m = T$ [HKR16、LCM17、SLC+19、SMLV20、LHK21、MTPR22]。当然，为了确保问题具有可处理性，这些研究都施加了各种额外类型的假设，通常表现为对 $\{h_x\}_{x\in\mathcal{X}}$ 函数施加单调性约束。由于 $m < T$ 时， $m < T$ 会使算法设计变得复杂，因为如果某个动作未被执行，其损失就会“重置”，因此这里的结果并不适用于这种情况。由于本文主要受到 $m < T$ 的应用场景驱动，我们认为这些研究与本文并不具有直接可比性。尽管如此，我们注意到，除去 mK 中的加性项和一个对数因子后，所有这些研究中的 CPR 保证通常都比我们的定理 1 所给出的收敛速率更不理想。

具有 $m < T$的。另一类已有研究关注计数设置，这些设置是具有 $\{w_x\}_{x\in\mathcal{X}} = \{\vec{1}\}$ 的 WTB 特例；与我们一样，它们也受到此类应用的推动：$m < T$ [ABGK22、MLS22]。这些设置与我们的问题更为相似，因为它们并不要求 $m = T$。计数赌博机 [MLS22] 除了 $\{w_x\}_{x\in\mathcal{X}} = \{\vec{1}\}$ 外不作任何假设；在此情形下，任何算法都必须承受~~$\widetilde{\Omega}\left(\sqrt{mKT}\right)$~~ CPR。该研究改造了逐次淘汰算法，而我们的算法深受其启发。拥塞赌博机 [ABGK22]通过要求 $\{h_x\}_{x\in\mathcal{X}}$ 函数为递增函数，对计数赌博机作了特化；即便如此，目前已知的最佳上界仍按 $\widetilde{\mathcal{O}}\left(\sqrt{mKT}\right)$ 的规模增长。因此，这些设置中最佳的 CPR 界似乎都随 $\sqrt{m}$ 成倍增长。此外，这些设置中已知最佳算法的计算复杂度分别随 T增长， m 。相比之下，我们的 REO 条件允许设计一种计算高效的算法，其统计最优的CPR保证仅随 m 加性增长。

相关的非计数设置。 大量研究探讨了各种设置，其中每个动作的损失会以某种结构化方式随时间变化，例如遵循某个随机过程，或取决于该动作上次被执行以来经过的时间步数 [Whi81, GM11, TL12, BGZ14, BF16, KI18, BSSS19, PBG19, CCB20, CDK+20, LCCG21]。所有这些研究中损失演变的模型都不同于我们的（加权）计数设置。

策略遗憾。 许多研究针对通用的 m-记忆有界对手考察策略遗憾，其算法也适用于我们的设置 [ADT12, CBDS13, DDKP14, ADMM18, MY18]。然而，这些结果忽略了我们所考虑的特殊加权计数结构，直接应用会得到次优的 CPR 界，其表现逊于我们的定理 1。

6 讨论

在本文中，我们提出了加权计数老虎机，它对先前的计数设置进行了推广，使得每个时间步的损失成为近期被选择次数的加权求和的函数。为了确保这一具有挑战性的设置具有可处理性，我们引入了重复暴露最优性条件，并通过以人为中心的应用场景对其进行阐释：在这些场景中，要实现最佳表现，需要经过一段校准期，之后才能趋于稳定。我们证明，只需对经典的逐次淘汰算法进行简单修改，即可实现最优的完全策略遗憾保证（仅相差一个对数因子），而我们的数值结果表明，其.....

实际应用性、可扩展性以及相较于替代基线的优越性。最后，我们证明了，尽管我们的算法需要输入一个关于 m 的非平凡上界 $M < T$，但任何具有次线性 CPR 的算法都需要这样的输入，而且我们的方法对该输入 M 的依赖程度是最优的。综合而言，这意味着对于满足 0-REO 的 WTB问题，我们的算法在 CPR 方面是最优的。

我们承认，我们的工作存在一定的局限性。从理论角度来看，尽管存在一个具有非平凡 $\alpha = \widetilde{\Theta}\left(\sqrt{mK/T}\right)$ 的参数范围（如第3.1节所讨论的），在该范围内，定理 1 是最优的，且其对 α 的依赖性无法进一步改进，但尚不清楚这种依赖性对于所有 α 取值是否都是最优的。对此问题的研究是未来一个值得探索的方向。

更具体地说，我们的WTB和REO设定存在一个局限性：虽然对真实最佳表现出现之前所经历的校准期进行建模是一个合理的步骤，但它未能对许多其他细微因素进行建模，而这些因素正是最初促使我们开展这项研究的以人为中心的领域中所存在的。例如，可以设想，在强度更高的任务中，经过长时间 $m < T$反复执行某项任务后，经过校准的个体可能会开始感到疲劳。在这种情况下，对反复执行某个动作所产生的损失进行建模的最佳方式是：首先，个体经历一个校准期，其表现逐渐提升至“最佳状态”；但随后，随着疲劳不断累积，其表现最终会逐渐下降（该模型类似于 [LHK21]所考虑的 $m = T$ 设定）。我们的设定能够处理初始的校准期和表现提升阶段，但无法处理随后出现的表现下降阶段。我们认为，探索能够处理这一问题的算法是未来研究的一个重要方向。

改进模型的另一种方式，是研究更一般的情形：每个动作 x都有各自的记忆容量 m_x （而不是所有动作共用一个 m ），或者问题的记忆容量会随时间变化。我们能否设计出可以智能适应这些复杂情况的算法?

我们为 REO条件提供的具体动机示例，主要源自心理生理学文献中关于热身递减现象的研究。尽管如此，我们通常认为，REO条件也可能适用于其他交互场景，例如推荐系统。举例来说，用户可能需要多次看到某一类内容，才会确定自己对它的偏好；但如果一段时间内没有看到该内容，她就会忘记其细节，并需要经历另一个探索性校准期，重新确认自己相对于近期推荐内容对该内容的偏好。在这种情况下，目标是探索用户最终的偏好，然后反复选择用户最终最喜欢的项目。这样的场景适合应用满足 REO条件的 WTB。对推荐系统数据进行定量分析，并证明此类场景确实存在且能够用满足 REO条件的 WTB很好地建模，是一个非常有趣的未来研究方向，也可能产生广泛的实际影响。

致谢

本材料基于美国国家科学基金会研究生研究奖学金计划资助的工作，资助编号 DGE1745016。本工作还部分获得美国国家科学基金会 AIIRA 研究所资助 20216702135329 的支持。本材料中表达的任何观点、发现、结论或建议均属于作者个人，不一定反映美国国家科学基金会的观点。

A 算法 1 的分析

本节分析算法 1 的完全策略遗憾，并证明定理 1。如第 3 节所述，我们的分析总体上相当标准，不过，为了确保最终结果中的依赖关系达到最优，我们需要谨慎选择参数。因此，许多计算都与 Malik 等人 [MLS22], 的做法密切相似，但为求完整起见，我们仍给出完整论证。在正式证明定理 1 之前，我们先引入函数μ： $\mathcal{X} \to [0,1]$ 它将在证明中发挥作用。对于任意动作 $x \in \mathcal{X}$，我们定义 $\mu(x)$ 为

$$\mu(x) = h_x(\|w_x\|_1).$$

对于算法 1 外层循环中的每个轮次 $s \in \{1, 2 \dots S\}$，定义 $T_s = 2|A_s|n_s$，其中A_s， n_s 在算法 1 中定义。有了这个定义，我们现在可以正式证明定理 1。

A.1 定理 1 的证明

对于任意策略 π（这是长度为 T 的确定性动作序列），令 $\ell_t(\pi)$ 表示执行 π 时在时间步 t 遭受的期望损失。将策略 $\pi^\star$ 定义为

$$\pi^\star \in \arg\min_{\pi \in \mathcal{X}^T} \sum_{t=1}^{T} \ell_t(\pi),$$

使得 $\pi^\star$ 是最优策略（即遭受最小累积期望损失的策略）。请注意，(m、 w、 h)-加权计数赌博机的定义确保，对于任意时间步 t，都存在一个动作 $x \in \mathcal{X}$ 和 $y \in \{1\} \times \{0,1\}^{m-1}$，使得 $\ell_t(\pi^\star) = h_x(w_x^\top y)$。因此， α-REO条件确保了 that

$$\mu(x^\star) = h_{x^\star}(\|w_{x^\star}\|_1) \le h_x(w_x^\top y) + \alpha = \ell_t(\pi^\star) + \alpha.$$

特别地，这意味着

$$\sum_{t=1}^{T} \ell_t(\pi^\star) \ge T\mu(x^\star) - \alpha T. \tag{4}$$

现在令 ℓ^s 表示算法 1 在轮次 $s \in \{1, 2 \dots S\}$ 中遭受的损失。以下引理给出了算法 1 相对于 $T\mu(x^\star)$ 的累积损失上界。

引理 1. 以至少 $1 - \delta$ 的概率，算法 *1* 相对于 $T\mu(x^\star)$ 的总损失可以上界为

$$\sum_{s=1}^{S} \ell^s - T\mu(x^\star) \le 4KM + Km\log(T) + 800\sqrt{KT\log\left(\frac{2K\log(T)}{\delta}\right)}.$$

该引理 1 的证明见附录 A.2。利用引理 1 的结果，我们现在按如下方式使用它来证明定理 1。注意，根据公式 (4) 和引理 1，完全

算法 1 的策略遗憾 $\mathcal{R}^{\mathrm{cp}}$ 满足

$$\begin{aligned}\mathcal{R}^{\mathrm{cp}} &= \sum_{s=1}^{S} \ell^s - \sum_{t=1}^{T} \ell_t(\pi^\star) \\ &\leq \sum_{s=1}^{S} \ell^s - T\mu(x^\star) + \alpha T \\ &\leq 4KM + Km\log(T) + 800\sqrt{KT\log\left(\frac{2K\log(T)}{\delta}\right)} + \alpha T.\end{aligned}$$

至此，定理 1 的证明完成。 ■

A.2 引理 1 的证明

为便于证明，我们需要以下关键引理，它界定了算法 1 在每个轮次 $s \in \{1, 2 \dots S\}$ 中产生的损失。对于以下引理的陈述，请注意，完成任一轮次 $s \in \{1, 2 \dots S\}$ 总共需要 $T_s = 2|A_s|n_s$ 个时间步。

引理2。 以至少 $1-\delta$ 的概率，对于每个轮次 $s \in \{2, 3 \dots S\}$，相对于 $T_s\mu(x^\star)$ 的总损失同时满足以下界:

$$\ell^s - T_s\mu(x^\star) \leq |A_s|\left(m + 4(2n_s - m)C_{s-1}\right).$$

引理2 的证明见附录 A.3。根据引理2 的结果，我们可以保证以至少 $1-\delta$ 的概率,

$$\begin{aligned}\sum_{s=1}^{S} \ell^s - T\mu(x^\star) &= \sum_{s=1}^{S} \left(\ell^s - T_s\mu(x^\star)\right) \\ &\leq 4KM + \sum_{s=2}^{S} \left(\ell^s - T_s\mu(x^\star)\right) \\ &\leq 4KM + \sum_{s=2}^{S} |A_s|\left(m + 4(2n_s - m)C_{s-1}\right) \\ &\leq 4KM + SKm + 8\sum_{s=2}^{S} |A_s|n_sC_{s-1}.\end{aligned} \tag{5}$$

回顾算法 1 中给出的定义 $n_s = KM2^s/|A_s|$ 和 $T_s = 2|A_s|n_s$。另请注意,

$$C_{s-1} = \sqrt{\frac{32}{n_{s-1}}\log\left(\frac{2KS}{\delta}\right)} = \sqrt{\frac{32}{n_s|A_s|/(2|A_{s-1}|)}\log\left(\frac{2KS}{\delta}\right)} = \sqrt{\frac{64|A_{s-1}|}{n_s|A_s|}\log\left(\frac{2KS}{\delta}\right)}.$$

将上述关系代入公式（5）右侧的最后一项，可得

$$
\begin{aligned}
8\sum_{s=2}^{S}|A_s|n_sC_{s-1} &= 8\sum_{s=2}^{S}|A_s|n_s\sqrt{\frac{64|A_{s-1}|}{n_s|A_s|}\log\left(\frac{2KS}{\delta}\right)} \\
&= 8\sqrt{\log\left(\frac{2KS}{\delta}\right)}\sum_{s=2}^{S}|A_s|n_s\sqrt{\frac{64|A_{s-1}|}{n_s|A_s|}} \\
&= 8\sqrt{\log\left(\frac{2KS}{\delta}\right)}\sum_{s=2}^{S}|A_s|\sqrt{n_s}\sqrt{\frac{64|A_{s-1}|}{|A_s|}} \\
&= 8\sqrt{\log\left(\frac{2KS}{\delta}\right)}\sum_{s=2}^{S}|A_s|\sqrt{\frac{KM2^s}{|A_s|}}\sqrt{\frac{64|A_{s-1}|}{|A_s|}} \\
&= 8\sqrt{\log\left(\frac{2KS}{\delta}\right)}\sum_{s=2}^{S}\sqrt{KM2^s}\sqrt{64|A_{s-1}|} \\
&\le 8\sqrt{\log\left(\frac{2KS}{\delta}\right)}\sum_{s=2}^{S}K\sqrt{M2^s}\sqrt{64} \\
&\le 800\sqrt{\log\left(\frac{2KS}{\delta}\right)}K\sqrt{M}2^{S/2}.
\end{aligned}
$$

现在回顾算法 1 中 $S = \log_2\left(\frac{T}{4KM}+1\right)$ 的定义。将其代入上式，可得

$$
\begin{aligned}
8\sum_{s=2}^{S}|A_s|n_sC_{s-1} &\le 800\sqrt{\log\left(\frac{2KS}{\delta}\right)}K\sqrt{M}2^{S/2} \\
&= 800\sqrt{\log\left(\frac{2KS}{\delta}\right)}K\sqrt{M}\sqrt{\frac{T}{4KM}+1} \\
&\le 800\sqrt{\log\left(\frac{2KS}{\delta}\right)}K\sqrt{M}\sqrt{\frac{T}{KM}} \\
&= 800\sqrt{\log\left(\frac{2KS}{\delta}\right)}\sqrt{K}\sqrt{T}.
\end{aligned}
\tag{6}
$$

结合 公式（5）与公式（6），并使用上界 $S \le \log(T)$，可得结果 t. ■

A.3 引理2的证明

为便于证明，我们利用以下关键引理，它界定了每个动作 $x \in A_s$ 相对于 $\mu(x^\star)$ 的值 $\mu(x)$ 差距。

引理 3。该事件

$$\cap_{s=2}^{S}\cap_{x\in A_s}\left\{\mu(x)-\mu(x^\star)\le 4C_{s-1}\right\},$$

发生的概率至少为 $1-\delta$。

该引理 3 的证明见附录 A.4。现在让我们回到主要证明。对于任意轮次 $s > 1$ 和任意动作 $x \in A_s$，令 ℓ^s 表示算法 1 在轮次 s 中执行动作 x 共 $2n_s$ 次所经历的总损失。因此，我们有 $\ell^s = \sum_{x \in A_s} \ell^{sx}$。

注意，在单个轮次 $s > 1$ 内，对于每个 $x \in A_s$，我们执行 x $2n_s$ 次。对于执行 x 的后 $2n_s - m$ 次，动作 x 已在之前的 m 个时间步中被执行了 m 次。因此，对于执行 x 的后 $2n_s - m$ 次，执行动作 x 的期望损失为 $h_x(m) = \mu(x)$。于是，我们有

$$\ell^{sx} \leq m + (2n_s - m)\mu(x).$$

因此，对于每个 $s > 1$ 和 $x \in A_s$，我们可以使用引理 3 对其上界进行估计

$$\begin{aligned}
\ell^{sx} - 2n_s\mu(x^\star) &\leq m + (2n_s - m)\mu(x) - 2n_s\mu(x^\star) + m\mu(x^\star) \\
&\leq m + (2n_s - m)\left(\mu(x) - \mu(x^\star)\right) \\
&\leq m + 4(2n_s - m)C_{s-1}.
\end{aligned}$$

该上界对每个 $x \in A_s$ 均一致成立。回忆 $T_s = 2|A_s|n_s$，因此我们有

$$\begin{aligned}
\ell^s - T_s\mu(x^\star) &= \sum_{x \in A_s} \ell^{sx} - 2|A_s|n_s\mu(x^\star) \\
&= \sum_{x \in A_s} \left(\ell^{sx} - 2n_s\mu(x^\star)\right) \\
&\leq \sum_{x \in A_s} \left(m + 4(2n_s - m)C_{s-1}\right) \\
&= |A_s|\left(m + 4(2n_s - m)C_{s-1}\right).
\end{aligned}$$

证明完毕。 ■

A.4 引理 3 的证明

为便于证明，我们需要以下两个关键辅助结果。第一个结果界定了通过随机实现值 $\{h_x(m)_{s,k}\}$ 估计 $\mu(x)$ 时产生的 $\sim$ 误差。第二个结果表明，在运行算法 1 时，该算法基于在各轮次 $s \in \{1, 2 \ldots S\}$ 中逐次淘汰劣质动作，因此在任意轮次 s，我们都不会从可行动作集合 A_s 中淘汰 $x^\star$。

引理 4. 任取 $s \in \{1, 2 \ldots S\}$，令 B_s 表示如下事件：对于所有动作 $x \in A_s$，我们同时有

$$|\widehat{\mu}_s(x) - \mu(x)| \leq C_s.$$

那么， B_s 发生的概率至少为 $1 - \delta/S$。

引理 5. 事件 $\cap_{s=1}^S B_s$（其中事件 B_s 在引理 *4* 中定义）意味着如下事件：

$$x^\star \in \cap_{s=1}^S A_s \text{ and } \cap_{s=1}^S \left\{0 \leq \widehat{\mu}_s(x^\star) - \widehat{\mu}_s(\widehat{x}_s) \leq 2C_s\right\}.$$

引理 4 和引理 5 的证明分别见附录 A.5 和附录 A.6。

现在回到证明。根据引理 4 的结果和并集界，事件 $\cap_{s=1}^{S} B_s$发生的概率至少为 $1-\delta$。此外，引理 5 的结果表明，事件$\cap_{s=1}^{S} B_s$ 蕴含事件

$$x^\star \in \cap_{s=1}^{S} A_s. \tag{7}$$

因此，在事件 $\cap_{s=1}^{S} B_s$上，注意对于任意 $s>1$ 和任意动作 $x \in A_s$，我们都有

$$\begin{aligned}
\mu(x) - \mu(x^\star) &\overset{(i)}{\leq} \widehat{\mu}_{s-1}(x) - \mu(x^\star) + C_{s-1} \\
&\overset{(ii)}{\leq} \widehat{\mu}_{s-1}(\widehat{x}_{s-1}) - \mu(x^\star) + 3C_{s-1} \\
&\overset{(iii)}{\leq} \widehat{\mu}_{s-1}(x^\star) - \mu(x^\star) + 3C_{s-1} \\
&\overset{(iv)}{\leq} \mu(x^\star) - \mu(x^\star) + 4C_{s-1} \\
&= 4C_{s-1},
\end{aligned}$$

其中，步骤（i）由引理 4 得出，步骤（ii）由 A_s 的定义以及$x \in A_s$这一事实得出，步骤（iii）由 $\widehat{x}_{s-1}$ 的定义和式（7）得出，而步骤（iv）再次由引理 4 和式（7）得出。证明至此完成。■

A.5 引理 4 的证明

任取 $x \in \mathcal{X}$。回顾算法 1 中给出的 C_s 的定义，Hoeffding 界 [Hoe63]确保事件

$$|\widehat{\mu}_s(x) - \mu(x)| = \left| \mu(x) - \frac{1}{n_s} \sum_{k=1}^{n_s} \widetilde{h}_x(m)_{s,k} \right| \leq \sqrt{\frac{32}{n_s} \log\left(\frac{2KS}{\delta}\right)} = C_s, \tag{8}$$

发生概率至少为 $1-\delta/(KS)$。由于 $|A_s| \leq K$，因此并集界保证上述事件对所有 $x \in A_s$ 同时发生的概率至少为 $1-\delta/S$。■

A.6 引理 5 的证明

假设事件 $\cap_{s'=1}^{S} B_{s'}$ 为真。在此事件下，我们通过对 s 使用归纳法来证明该引理。首先，我们证明 $s=1$ 的基例，即 $x^\star \in A_1$ 和 $0 \leq \widehat{\mu}_1(x^\star) - \widehat{\mu}_1(\widehat{x}_1) \leq 2C_1$。然后，在归纳步骤中，我们证明，如果事件 $x^\star \in A_{s-1}$ 和 $0 \leq \widehat{\mu}_{s-1}(x^\star) - \widehat{\mu}_{s-1}(\widehat{x}_{s-1}) \leq 2C_{s-1}$ 发生，那么我们也有事件

$$x^\star \in A_s \text{ and } 0 \leq \widehat{\mu}_s(x^\star) - \widehat{\mu}_s(\widehat{x}_s) \leq 2C_s,$$

这一点也成立。

对于基例，注意到根据定义，我们可以保证 $x^\star \in A_1$。根据$\widehat{x}_1$的定义，我们知道 $0 \leq \widehat{\mu}_1(x^\star) - \widehat{\mu}_1(\widehat{x}_1)$。此外，回顾引理 4 中事件 B_1 的定义，在事件 B_1 上，我们有

$$\widehat{\mu}_1(x^\star) - \mu(x^\star) \leq C_1 \text{ and } \mu(\widehat{x}_1) - \widehat{\mu}_1(\widehat{x}_1) \leq C_1.$$

将这些等式合并，并利用 $\mu(x^\star) \le \mu(\widehat{x}_1)$ 这一事实，可得

$$\widehat{\mu}_1(x^\star) - \widehat{\mu}_1(\widehat{x}_1) \le 2C_1.$$

这验证了基例。

对于归纳步骤，假设发生了 $x^\star \in A_{s-1}$ and $0 \le \widehat{\mu}_{s-1}(x^\star) - \widehat{\mu}_{s-1}(\widehat{x}_{s-1}) \le 2C_{s-1}$。那么，根据 A_s 的定义和归纳假设，可直接推出 $x^\star \in A_s$。因此，根据 $\widehat{x}_s$ 的定义， $0 \le \widehat{\mu}_s(x^\star) - \widehat{\mu}_s(\widehat{x}_s)$成立。接着回顾引理 4 中事件 B_s 的定义，在事件 B_s 上，我们有

$$\widehat{\mu}_s(x^\star) - \mu(x^\star) \le C_s \text{ and } \mu(\widehat{x}_s) - \widehat{\mu}_s(\widehat{x}_s) \le C_s.$$

将这些方程合并起来，并利用 $\mu(x^\star) \le \mu(\widehat{x}_s)$ 确保了这一点

$$\widehat{\mu}_s(x^\star) - \widehat{\mu}_s(\widehat{x}_s) \le 2C_s.$$

这验证了 归纳步骤。正如前面所述，这足以完成证明。 ■

B 定理 2 的证明

反设该命题不成立。那么存在满足给定条件的某个 ϵ，以及某个函数 f，使得 $\mathcal{A}_{\epsilon,\ f}$ 非空。这意味着存在一个算法 $\mathcal{A}$，当输入任意正整数 T， K， M 且满足 $M \le T$ 时，该算法 $\mathcal{A}$ 满足

$$\mathbb{E}[\mathcal{R}^{\mathrm{cp}}(\mathcal{A}, \mathsf{tb})] \le \min\left\{T/4, f(m_{\mathsf{tb}}, K)\left(T^{1-\epsilon_1} + T^{\epsilon_3} M^{1-\epsilon_2}\right)\right\} \text{ for all } \mathsf{tb} \in \mathsf{UTB}_{T,M,K}. \tag{9}$$

如果 $\mathcal{A}$是随机化算法，那么这意味着存在一个具有相同性质的确定性算法。因此，不失一般性，我们可以假设 $\mathcal{A}$ 是确定性的。

固定某个整数 $K \ge 2$。选取足够大的 T， M，使以下条件同时成立

$$M < T/4 \text{ and } f(1,K)\left(T^{1-\epsilon_1} + T^{\epsilon_3} M^{1-\epsilon_2}\right) < M/2. \tag{10}$$

要看出这些条件可以同时满足，回想一下， $\epsilon_1 \in (0,\ 1)$ 且 $0 \le \epsilon_3 < \epsilon_2 < 1$。令$\gamma = \min\{\epsilon_1, \epsilon_2 - \epsilon_3\} > 0$。因此，如果我们选择 $M = T^{1-\gamma/2}$，那么由于该 M满足 $M = T^{1-\gamma/2} < T$，我们有

$$\begin{aligned} f(1,K)\left(T^{1-\epsilon_1} + T^{\epsilon_3} M^{1-\epsilon_2}\right) &< f(1,K)\left(T^{1-\epsilon_1} + T^{\epsilon_3} T^{1-\epsilon_2}\right) \\ &= f(1,K)\left(T^{1-\epsilon_1} + T^{1-(\epsilon_2-\epsilon_3)}\right) \\ &\le 2f(1,K) T^{1-\gamma}. \end{aligned}$$

因此，对于足够大的 T，对于这一 $M = T^{1-\gamma/2}$ 选择，我们有 $M < T/4$，并且还有 $f(1,K)$ $\left(T^{1-\epsilon_1} + T^{\epsilon_3} M^{1-\epsilon_2}\right) < M/2$。这表明公式（10）是可行的。

现在，我们将定义两个无权计数赌博机问题，每个问题都有 K 个动作。回顾一下，在记忆容量为 m的无权计数型老虎机问题中，在给定时间步执行某个动作所对应的损失，完全由该动作在最近 m 个时间步内被执行的次数决定。具体而言，假设在某个无权计数型老虎机问题 tb 中，我们在当前时间步执行动作 x ，且该动作在最近 m 个时间步内（包括当前时间步）总共被执行了 $1 \leq y \leq m$次。那么，存在一个函数$h_{\mathsf{tb},x}$：$\{1, 2 \dots m\} \to [0, 1]$，使得执行该动作所对应的损失表示为$h_{\mathsf{tb},x}(y)$。我们将使用这一记号来具体定义接下来要介绍的无权计数赌博机问题。

有了这一记号，我们将记忆长度为 $m_{\mathsf{tb}_A} = 1$ 的无权计数型老虎机问题 tb_A 实例化如下。对于动作 x_1，有 $h_{\mathsf{tb}_A,\ x_1} = 1/2$。对于动作 x_2，有 $h_{\mathsf{tb}_A,\ x_2} = 1$。对于其他每个动作 x，令 $h_{\mathsf{tb}_A,\ x} = 1$。我们称，当玩家执行动作 x 时，玩家几乎必然观察到 $h_{\mathsf{tb}_A,\ x}(1)$。注意，由于 $m_{\mathsf{tb}_A} = 1$，并且当我们执行任意动作时，所观察到的损失不存在随机性，因此 tb_A 确实是一个确定性的多臂老虎机问题。

现在，我们将记忆长度为 $m_{\mathsf{tb}_B} = M$ 的计数赌博机问题 tb_B 实例化如下。对于动作 x_1，我们定义 $h_{\mathsf{tb}_B,\ x_1} = 1/2$。对于动作 x_2，我们定义

$$h_{\mathsf{tb}_B, x_2}(y) = \begin{cases} 1 \text{ if } 1 \leq y < M \\ 0 \text{ if } y = M \end{cases} .$$

对于其他每个动作 x，我们定义 $h_{\mathsf{tb}_B,\ x} = 1$。再次强调，当玩家执行某个动作时，我们要求玩家对损失的观测不存在随机性。因此，tb_B 中的反馈模型是确定性的。

令问题 tb_A 和 tb_B 的时域长度为根据公式（10）选取的 T。还需注意，两个问题实例都满足参数为 $\alpha = 0$ 的 REO，因此 $\mathsf{tb}_A,\ \mathsf{tb}_B \in \mathrm{UTB}_{T,\ M,\ K}$。因此，根据我们对 $\mathcal{A}$ 确定性的假设、公式（9）以及 $m_{\mathsf{tb}_A} = 1$ 这一事实，我们有

$$\mathcal{R}^{\mathrm{cp}}(\mathcal{A}, \mathsf{tb}_A) \leq f(m_{\mathsf{tb}_A}, K) \left(T^{1-\epsilon_1} + T^{\epsilon_3} M^{1-\epsilon_2}\right) = f(1, K) \left(T^{1-\epsilon_1} + T^{\epsilon_3} M^{1-\epsilon_2}\right) < M/2 \qquad (11)$$

其中，最后一个不等式由公式（10）得出。同样，根据我们对 $\mathcal{A}$ 确定性的假设以及公式（9），我们有

$$\mathcal{R}^{\mathrm{cp}}(\mathcal{A}, \mathsf{tb}_B) \leq T/4. \qquad (12)$$

当在问题 tb_A 上运行 $\mathcal{A}$ 时，有两种情况：要么 $\mathcal{A}$ 连续 M 次选择 x_2（在其求解 tb_A、运行 T 个时间步的过程中某个时刻），要么不连续选择。

考虑第一种情况，即 $\mathcal{A}$ 在问题 tb_A 上连续 M 次选择 x_2。那么我们有$\mathcal{R}^{\mathrm{cp}}(\mathcal{A},\ \mathsf{tb}_A) \geq M/2$。这是因为 tb_A 的最优策略在每个时间步都会选择 x_1，因此 $\mathcal{A}$ 选择动作 $x \neq x_1$ 的任何时间步都会使 $\mathcal{A}$ 的 CPR 增加 $h_{\mathsf{tb}_A,\ x} - h_{\mathsf{tb}_A,\ x_1} = 1 - 1/2 = 1/2$。这与公式（11）矛盾。

考虑第二种情况：在问题 tb_A 上，$\mathcal{A}$ 连续 M 次不选择 x_2。请注意，确定性算法 $\mathcal{A}$ 可以视为一个长度为 T 的函数序列，其中第 t

个函数将过去的 $t-1$ 动作选择和损失观测映射为时间步 t所执行的动作。还要注意，在 tb_B 中选择某个动作所观测到的损失，与在 tb_A 中选择同一动作所观测到的损失不同，当且仅当该动作是 x_2，且在之前的 M 个时间步（包括当前时间步）中已执行 M 次。

因此，由于 $\mathcal{A}$ 在问题 tb_A上不会连续 M 次选择 x_2，它在应用于 tb_B时看到的损失输出序列完全相同，因此也会采取与应用于 tb_A时完全相同的动作序列。这反过来意味着它在问题 tb_B上不会连续 M 次选择 x_2。由于根据公式（10） $M < T/4$，且采取任何动作 $x \neq x_2$ 的损失始终至少为 $1/2$，因此始终采取 x_2 的策略是最优的，其累积损失为 $M-1$。所以，由于公式（10）意味着

$$T/2-(M-1) > T/2-M > T/2-T/4 = T/4,$$

我们有

$$\mathcal{R}^{\mathrm{cp}}(\mathcal{A}, \mathsf{tb}_B) \geq T/2-(M-1) > T/4.$$

这与式（12）矛盾。

无论哪种情况，我们都得出了矛盾。因此，我们已经证明，对于每个满足给定条件的 ϵ 以及每个函数 f，相应的集合 $\mathcal{A}_{\epsilon,\ f}$ 都是空集。证明至此完成。■

C 命题 1 的证明

本节将给出命题 1 的形式化证明。令 $\mathcal{X}=\{x_1, x_2\}$。令 w 逐分量定义为 $w_i = \frac{1}{2^i}$，其中每个 $1 \leq i \leq m$ 均如此。对每个 $x \in \mathcal{X}$，令 $w_x = w$。现在从 $\{0,1\}^{m-1}$ 中均匀随机抽取一个比特串 $y^\star$，其具体内容对玩家保密。定义 $h_{x_1}=1$，并定义

$$h_{x_2}(w_{x_2}^\top y^{t,x_2,m}) = h_{x_2}(w^\top y^{t,x_2,m}) = \begin{cases} 1 \text{ if } w^\top y^{t,x_2,m} \neq w^\top (1, y^\star) \\ 0 \text{ if } w^\top y^{t,x_2,m} = w^\top (1, y^\star) \end{cases}.$$

我们假设玩家所经历的损失反馈不存在随机性。这定义了一个 $(m,\ w,\ h)$-加权计数老虎机博弈。为便于后续记号表示，我们还将 $v \in \mathcal{X}^{m-1}$ 定义为

$$v_i = \begin{cases} x_2 \text{ if } y_i^\star = 1 \\ x_1 \text{ if } y_i^\star = 0 \end{cases}.$$

我们声称，按照对 w 的选择，如果 $y \neq y' \in \{1\} \times \{0,1\}^{m-1}$ 那么 $w^\top y \neq w^\top y'$。我们暂且推迟对该断言的正式证明，并利用此断言完成命题 1 的证明。关键的是，该断言意味着：当且仅当我们在时间步 t 执行动作x_2，且满足 $y^{t,\ x_2,\ m} = (1,\ y^\star)$ 时，我们才会在时间步 t 承受非单位损失。等价地，当且仅当我们在时间步 $t,\ t-1 \ldots t-m$ 的动作序列为 $(x_2,\ v)$ 时，我们才会在时间步 t 承受非单位损失。因此，循环执行 $v_{m-1},\ v_{m-2} \ldots v_1,\ x_2$ 的策略至少每隔 m 个时间步就会有一次损失为零。与此同时，识别最优策略至少与执行动作序列 $(x_2,\ v)$ 一样困难，而后者又至少与识别 $y^\star$ 一样困难。

标准的“干草堆中找针”论证 [DKWY20] 随后表明，识别出 $y^\star$ 需要$\widetilde{\Omega}(2^m)$个时间步。进而，由于循环策略 $v_{m-1},\ v_{m-2}\dots v_1,\ x_2$ 每 m 个时间步至少会有一次产生零损失，这意味着任何（可能是随机化的）$\widetilde{}$算法的预期 CPR $\mathbb{E}[\mathcal{R}^{\mathrm{cp}}]$ 下界为 $\Omega(\min\{2^m, T\}/m)$，其中期望同时取自算法的（可能的）随机化以及对 $y^\star$的采样。

现在回过头来证明我们的断言：若 $y \neq y' \in \{1\} \times \{0,1\}^{m-1}$，则 $w^\top y \neq w^\top y'$。为导出矛盾，假设 $y \neq y'$，但 $w^\top y = w^\top y'$。令 $J \subseteq \{2, 3 \dots m\}$为 $y,\ y'$不同的坐标集合，并令 $j^\star =$为 $\min J$。注意，根据假设，J非空，因此 $j^\star$定义良好。不失一般性，假设 $y_{j^\star} - y'_{j^\star} = 1$（当 $y_{j^\star} - y'_{j^\star} = -1$时的情形完全对称）。接下来注意到

$$\begin{aligned} 0 &= w^\top (y - y') \\ &= \sum_{j=1}^{m} w_j (y_j - y'_j) \\ &= \sum_{j \in J} w_j (y_j - y'_j) \\ &= w_{j^\star} + \sum_{j \neq j^\star \in J} w_j (y_j - y'_j). \end{aligned} \tag{13}$$

现在我们可以对式（13）给出下界：

$$\begin{aligned} 0 &= w_{j^\star} + \sum_{j \neq j^\star \in J} w_j (y_j - y'_j) \\ &\geq w_{j^\star} - \sum_{j \neq j^\star \in J} w_j |y_j - y'_j| \\ &= w_{j^\star} - \sum_{j \neq j^\star \in J} w_j \\ &\geq w_{j^\star} - \sum_{j=j^\star+1}^{m} w_j. \end{aligned} \tag{14}$$

现在将我们对 w 的选择代入式（14），可得

$$0 \geq w_{j^\star} - \sum_{j=j^\star+1}^{m} w_j = \frac{1}{2^{j^\star}} - \sum_{j=j^\star+1}^{m} \frac{1}{2^j} = \frac{1}{2^{j^\star}} \left(1 - \sum_{j=1}^{m-j^\star} \frac{1}{2^j}\right) > 0,$$

这显然是矛盾。因此，我们证明了如下断言：如果 $y \neq y' \in \{1\} \times \{0,1\}^{m-1}$，那么 $w^\top y \neq w^\top y'$。
■

D 扩展数值结果与细节

D.1 无权重计数赌博机

在本节中，我们首先给出无权计数型老虎机问题的更多实验结果，其中的损失函数与第4.1节所述的相同。因此，该问题

满足 0-REO。在这里，我们改变 m， K， M 的取值，并绘制各方法随时间变化的 CPR。结果见图 4。在每种情况下，我们都观察到 SE 的表现优于基线方法。这些结果还表明，算法 1 的表现对于采用保守的上界 M 来界定 m 具有稳健性。

现在，我们给出另一种无权计数型老虎机问题的结果，其中 α-REO 在 $\alpha > 0$ 下成立。对于每个动作 $x \in \mathcal{X}$，我们定义 $w_x = \vec{1}/(4m)$。我们固定一个动作 $x^\star \in \mathcal{X}$和另一个不同的动作 $x^{\star\star} \in \mathcal{X}$，然后将损失函数 $\{h_x\}_{x\in\mathcal{X}}$ 定义为

$$h_x(y^{t,x,m}) = \begin{cases} 1 - w_x^\top y^{t,x,m} - 0.15 \text{ if } x = x^\star, y^{t,x,m} = \vec{1} \\ 1 - w_x^\top y^{t,x,m} - \frac{m-1}{2m} - 0.2 \text{ if } x = x^{\star\star}, y^{t,x,m} = (1,0,0\ldots 0) \\ 1 - w_x^\top y^{t,x,m} \text{ otherwise} \end{cases} .$$

数值计算表明，该无权计数型老虎机问题实例满足α-REO，其中 $\alpha = \max\{0, -0.2 + \frac{2m-3}{4m}\}$。我们通过实验研究了算法 1 在此问题实例上的表现，并将其与基线方法进行比较，参数取值为 $m = 4$， $K = 5$， $M = 4$。由于该问题的最优策略并不明显，因此难以计算 CPR。因此，我们改为绘制各算法的期望累积损失超出 SE 损失的部分（因而，在任意时刻，只需对各算法的超额损失作常数平移，即可得到 CPR）。结果如图 5 所示，可以看出我们的方法优于其他所有方法。

D.2 加权计数赌博机

这里，我们介绍用于定义第 4.2 节中所述 WTB问题问题实例的损失函数。对于每个动作 $x \in \mathcal{X}$，我们按以下方式定义 w_x。首先，我们通过将其第 i 个坐标设为 $v_i = 1/2^i$，逐坐标定义向量 $v \in \mathbb{R}^m$。然后，对于每个 $x \in \mathcal{X}$，我们令 $w_x = v/(2\|v\|_1)$。我们固定一个动作 $x^\star \in \mathcal{X}$，随后将损失函数 $\{h_x\}_{x\in\mathcal{X}}$ 定义为

$$h_x(y^{t,x,m}) = \begin{cases} 1 - w_x^\top y^{t,x,m} - 0.15 \text{ if } x = x^\star, y^{t,x,m} = \vec{1} \\ 1 - w_x^\top y^{t,x,m} \text{ otherwise} \end{cases} .$$

因此，这个加权计数老虎机问题满足 0-REO。

D.3 模拟 F1 锦标赛

在本节中，我们详细介绍 F1单圈时间数据集、数据处理和概率模型拟合，以及用于定义有意义的 WTB问题实例的标准。本节最后给出模拟F1锦标赛的扩展结果。

D.3.1 数据集与数据处理

如第 4.4 节所述，我们使用 1950年-2022 [Rao22]的首次进站前 F1单圈时间数据，来学习适用于不同车手和比赛的概率圈速模型，并据此定义 WTB实例。该数据集包含来自 1120 场比赛和 858 名车手的比赛数据。需要注意的是，在每场比赛中，实际参赛的车手子集各不相同。对于本文而言，数据集中的每个条目都是一个形如（车手 ID，比赛 ID，单圈时间）的元组，其中车手 ID $\in \{1, \ldots, 858\}$、比赛 ID $\in \{1, \ldots, 1120\}$，而单圈时间是一个元组列表，每个元组的形式为（单圈时间，进站）。单圈时间表示官方记录的秒数时间，

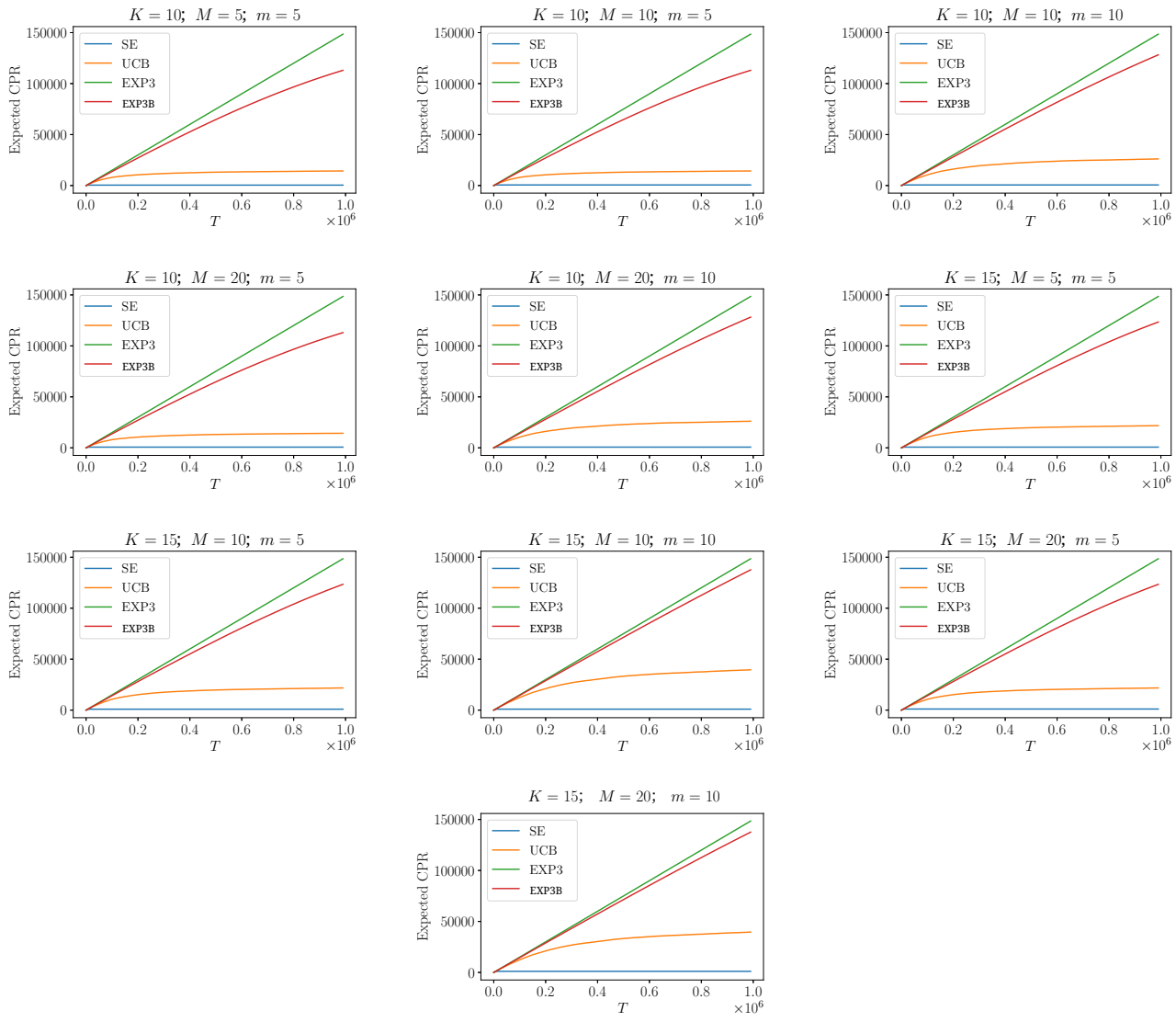


图 4：我们绘制了各算法在第4.1节所述的无权计数型老虎机问题上的期望 CPR，其中 m，K，M 的取值各不相同。所有图中的每个数据点均为对 20 个问题实例取平均所得，阴影区域表示均值上下 ± 1 个标准误的范围。

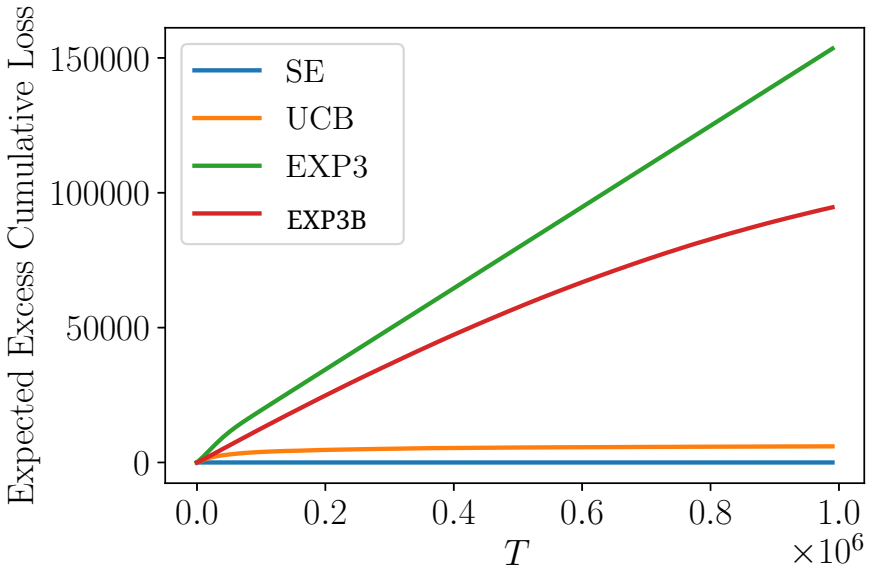


图 5：我们绘制了附录 D.1 中所述的未加权计数老虎机问题实例中，各算法相对于 SE 的期望累积损失超出量随时间的变化，参数为 $m = 4$，$K = 5$，$M = 4$。请注意，该问题实例满足 α-REO，且 $\alpha = 0.1125$。每个数据点是对 20 个问题实例取平均得到的，阴影区域表示均值周围的 ± 1 标准误。

车手在 F1 锦标赛中完成一圈所需的时间，而进站是一个二元变量，用于指示该圈是否包含一次进站。

为使不同比赛之间的单圈时间具有可比性，该数据集中的所有单圈时间均已归一化至区间 $[0,1]$。具体而言，若车手 i 在比赛 j 的第 k^{th} 圈原始单圈时间为 $\ell_{i,j,k}$，则归一化后的单圈时间为 $\tau_{i,j,k} := \dfrac{\ell_{i,j,k} - \min_{w,x} \ell_{w,j,x}}{\max_{y,z} \ell_{y,j,z} - \min_{w,x} \ell_{w,j,x}}$。

此外，对于每场比赛 j，我们会筛选单圈时间数据，仅保留数据量足以支持模型拟合的符合条件车手。在固定的比赛 j 中，如果数据集包含某位车手在首次进站前至少连续 8 个单圈时间数据点，则认为该车手符合条件。确定比赛 j 中符合条件的车手集合后，我们会缩短比赛j 中所有车手的单圈时间数据序列，使其长度与比赛 j 中最短的序列一致。例如，若比赛 j 中有 3 位符合条件的车手，其中车手 a 在进站前完成 8 圈，车手 b 完成 9 圈，车手 c 完成 10 圈，那么在学习车手 b 和 c 的单圈时间模型时，我们只使用他们最开始的 8 个单圈时间。

D.3.2 概率模型拟合

我们使用归一化后的单圈时间数据，通过最大似然法为每个符合条件的车手-比赛组合拟合单圈时间的概率模型。具体而言，我们假设，在比赛 j 中，车手 i 的k^{th} 归一化单圈时间（记为 $\tau_{i,j,k}$）服从如下分布

$$\mathcal{N}\left(\beta_{i,j} \exp(-k\alpha_{i,j}) - |\gamma_{i,j}|k,\ \sigma_{i,j}^2\right).$$

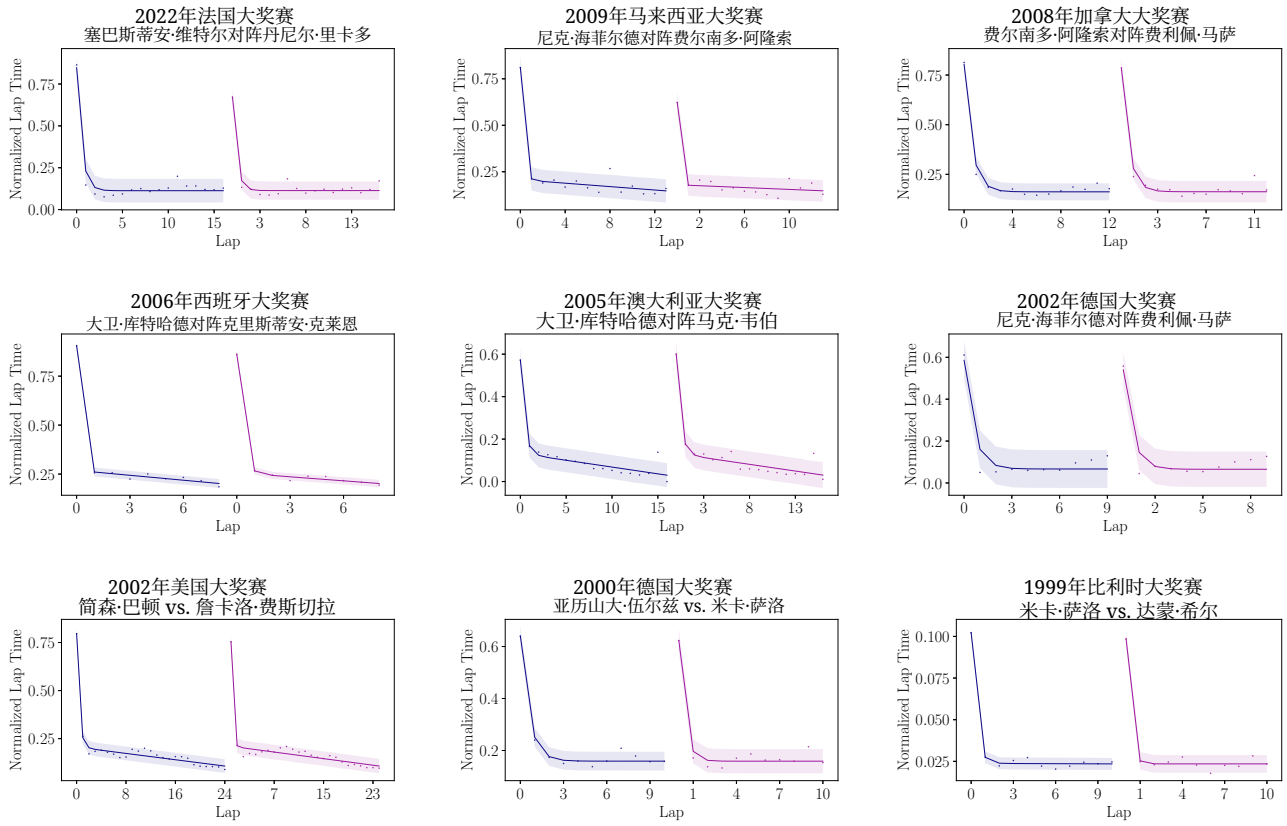


图6：在每个图中，我们选取一对车手和一场特定的比赛，并展示拟合得到的概率圈速模型。在每个图中，我们的概率模型针对每个单圈索引，对单圈时间的分布进行参数化（请注意，不同图中的单圈索引可能不同）。实线表示每个单圈索引对应的该分布均值。阴影区域以分布均值为中心，包含该分布的 ± 2 个标准差。虚线点表示实际的单圈时间（我们注意到，这些单圈时间几乎总是位于均值上下 2 个标准差的范围内）。请注意，所有单圈时间均已归一化，因此每个单圈时间都位于区间 $[0, 1]$ 内。

因此，对于每个车手-比赛 (i, j) 组合，我们的模型需要拟合 4 个参数 $(\alpha_{i,j}, \beta_{i,j}, \gamma_{i,j}, \sigma_{i,j})$。我们使用所有归一化的首次进站前数据（至少 8 个数据点），通过 scipy.optimize.curve fit 采用最大似然法拟合这些参数；该函数利用了 MINPACK [V+20] 中实现的 Levenberg-Marquardt 算法。

总体而言，我们的概率模型规定，对于每位车手 i 和每场比赛 j，单圈时间最初呈指数下降，随后呈线性下降。具体来说，随着单圈索引 k 增加，车手 i 在比赛 j 中的预期单圈时间以由 $\alpha_{i,j}$ 决定的速率下降（初始条件为 $\beta_{i,j}$），但绝不会低于 $-|\gamma_{i,j}|k$（我们的 k 值很小，因此单圈时间会趋于一个正数）。车手 i 在比赛 j 中的单圈时间方差为 $\sigma_{i,j}$（注意，该量与 k无关）。图6展示了9场比赛的归一化单圈时间数据及学习得到的概率模型（第10场比赛的概率模型见图3a）。从图中可以看出，该概率模型对（连续）单圈时间的分布进行了合理近似。

D.3.3 构造有意义的计数型赌博机实例

为了测试 SE，我们进一步筛选车手，使得车手单圈时间模型所定义的赌博机问题具有实质性的挑战性。具体而言，我们将符合条件的 TB车手对定义为参加同一场比赛的两名车手，且两者的终端平均单圈时间难以区分。特别地，记车手 i （分别为 i'）的终端预期单圈时间为 μ_i（分别为 $\mu_{i'}$）。那么，如果以下条件成立，则认为车手 i 和 i' 构成一个符合条件的 TB车手对：

- 车手 i 和 i' 在同一场比赛 j中参赛，
- $\mu_i \in [\mu_{i'} - \sigma^2_{i',j}, \mu_{i'} + \sigma^2_{i',j}]$，
- $\mu_{i'} \in [\mu_i - \sigma^2_{i,j}, \mu_i + \sigma^2_{i,j}]$。

请注意，这些条件实质上意味着，该问题至少与区分两个彼此“接近”的高斯分布的均值一样困难。有十对符合条件的TB车手对，我们据此定义了10个TB实例。具体而言，这对车手在比赛 j 中的单圈时间模型定义了一个TB实例中的损失，其中每位车手对应一个臂。在时间步 t “选择”车手 i 参加比赛 j 时，其（随机）瞬时损失取决于车手 i 在之前的 m 个时间步内于比赛 j 中完成单圈的次数，或者等价地，取决于记为 k的计数。因此，这一瞬时损失为 $\tau_{i,\ j,\ k} \sim \mathcal{N}(\beta_{i,\ j}\exp(-k\alpha_{i,\ j}) - |\gamma_{i,\ j}|k,\ \ \sigma^2_{i,\ j})$。

D.3.4 扩展结果

现在，我们展示模拟F1锦标赛全部 10 个实例的结果。每个实例的构建方式与第 4.4 节所述实例类似。我们选取了 10 组不同的车手组合及其参加的比赛。对于每组车手组合和每场比赛，我们使用F1单圈时间数据 [Rao22]，按照上文所述拟合概率圈速模型。在图6中，我们展示了所选各组车手组合的单圈时间概率模型，并表明随着单圈索引增加，他们的单圈时间往往会缩短。

我们使用图6中所示的概率模型，模拟这场锦标赛的 9 个实例，其中 $K = 2$ 和 m 等于相应图中单圈索引的数量。第 10 场比赛的结果已在第 4.4 节中给出。对于每个实例，我们都会统计在之前的 m 个时间步中，该实例中的每位车手被选中的次数。选择某位车手所产生的损失由我们拟合的概率模型所参数化的分布决定。具体而言，如果我们选择车手 x，并且在最近的 m 个时间步中已选择过他们 y 次，那么瞬时损失便从我们为单圈索引 y处的车手 x 拟合的概率模型所参数化的分布中采样。每个实例中的两位车手均经过挑选，使得他们校准后的表现难以区分，如上文所述。请注意，在此设定下，当且仅当选择较差的车手 $o(T)$多次时，才会获得 $o(T)$ CPR。在图7中，我们绘制了这场锦标赛 10 个实例中各方法的 CPR 随时间变化的曲线；每种情况下，SE 的表现都优于基线方法。

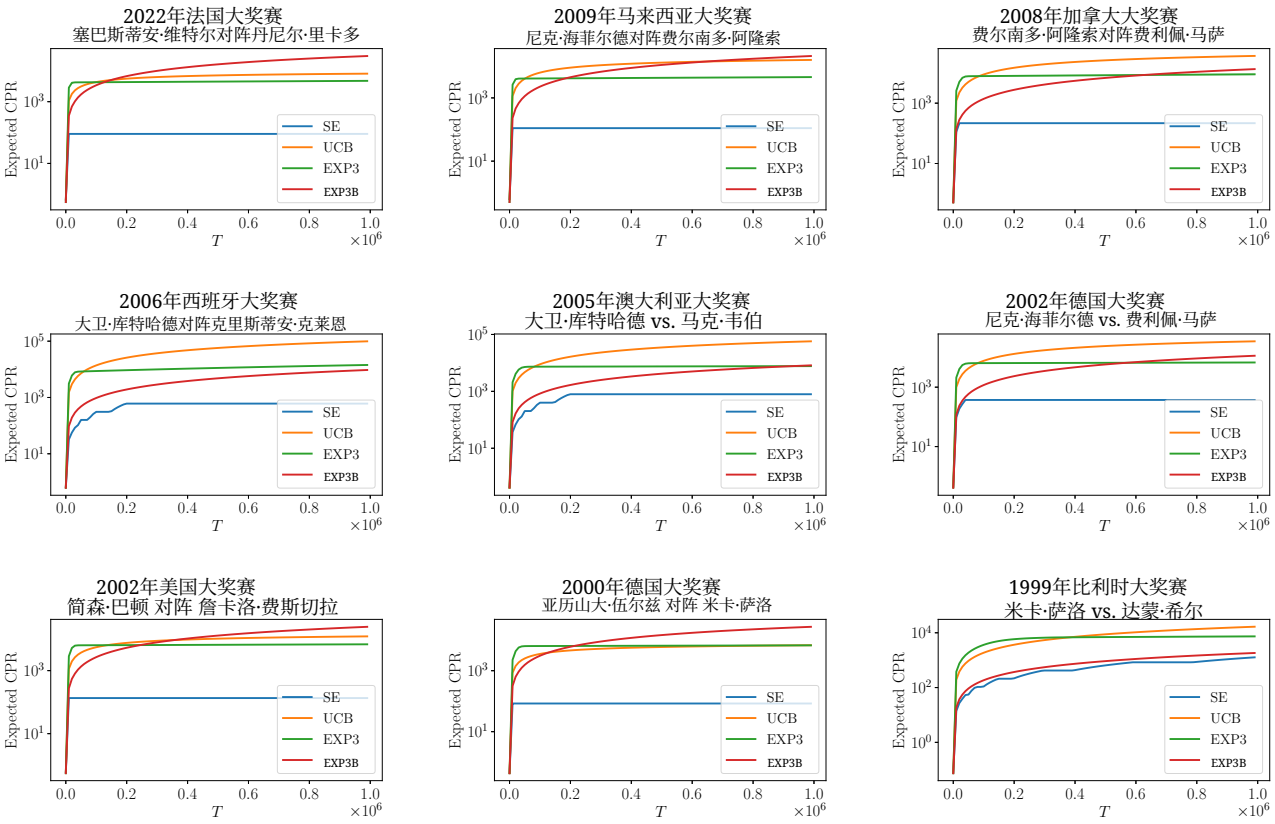


图7：对于图6中的每个图，我们模拟一个 WTB实例，并比较各种算法的表现。我们将每种算法的期望 CPR 描绘为时间的函数。数据是对 20 个问题实例取平均得到的，每个实例都有 $K = 2$ 和 $T = 10^6$。阴影区域表示均值周围的 ± 1 标准误。图中使用的值 m 等于图6中对应图的单圈索引长度。我们设定 M 等于 m。

参考文献

[ABGK22] Pranjal Awasthi、Kush Bhatia、Sreenivas Gollapudi 和 Kostas Kollias。拥塞老虎机：通过短期重置实现最优路由。载于国际机器学习会议，2022年。[ACBF02] Peter Auer、Nicol`o Cesa-Bianchi 和 Paul Fischer。多臂赌博机问题的有限时间分析。机器学习，47(2–3):235–256，2002年。[ACBFS02] Peter Auer、Nicol`o Cesa-Bianchi、Yoav Freund 和 Robert E. Schapire。非随机多臂赌博机问题。《*SIAM*计算杂志》，32(1):48–77，2002年。[Ada61] Jack A. Adams。遗忘的第二个方面：热身衰减综述。《心理学通报》，58:257–273，1961年。

[ADMM18] Raman Arora、Michael Dinitz、Teodor V. Marinov 和 Mehryar Mohri。《重复博弈中的策略遗憾》。载于《神经信息处理系统进展》，2018年。[ADT12] Raman Arora、Ofer Dekel 和 Ambuj Tewari。《针对自适应对手的在线赌博机学习：从遗憾到策略遗憾》。载于《国际机器学习会议》，2012年。[Ans95] Mark H. Anshel。《考察热身递减与插入的开放式和封闭式运动任务之间的函数关系：对练习策略的启示》。《运动科学杂志》，13：247–256，1995年。[AW93] Mark H. Anshel 和 Craig A. Wrisberg。《减少网球发球表现中的热身递减》。《运动与锻炼心理学杂志》，15(3)：290–303，1993年。[BF16] Djallel Bouneffouf 和 Raphael F´eraud。《具有已知趋势的多臂老虎机问题》。《神经计算》，205(C)：16–21，2016年。[BGZ14] Omar Besbes、Yonatan Gur 和 Assaf Zeevi。《奖励非平稳的随机多臂老虎机问题》。载于《神经信息处理系统进展》，2014年。[BSSS19] Soumya Basu、Rajat Sen、Sujay Sanghavi 和 Sanjay Shakkottai。《阻塞老虎机》。载于《神经信息处理系统进展》，2019年。[CBDS13] Nicol`o Cesa-Bianchi、Ofer Dekel 和 Ohad Shamir。《具有切换成本及其他自适应对手的在线学习》。载于《神经信息处理系统进展》，2013年。[CCB20] Leonardo Cella 和 Nicol´o Cesa-Bianchi。《收益依赖于延迟的随机老虎机》。载于《国际人工智能与统计会议》，第1168–1177页，2020年。[CDK+20] Corinna Cortes、Giulia DeSalvo、Vitaly Kuznetsov、Mehryar Mohri 和 Scott Yang。《基于差异度的非静置休息型老虎机算法》。arXiv预印本 arXiv:1710.10657，2020年。[DDKP14] Ofer Dekel、Jian Ding、Tomer Koren 和 Yuval Peres。《具有切换成本的老虎机：$T^{2/3}$ 遗憾》。载于《计算理论研讨会》，2014 年。[DKWY20] Simon S. Du、Sham M. Kakade、Ruosong Wang 和 Lin F. Yang。《良好的表征是否足以实现样本高效的强化学习？》载于《国际学习表征会议》，2020年。[GM11] Aur´elien Garivier 和 Eric Moulines。《关于切换老虎机问题的上置信界策略》。载于《国际算法学习理论会议》，2011年。[HKR16] Hoda Heidari、Michael Kearns 和 Aaron Roth。《改进型和衰退型老虎机的紧致策略遗憾界》。载于《国际人工智能联合会议》，2016年。

[Hoe63] Wassily Hoeffding。《有界随机变量之和的概率不等式》。《美国统计学会期刊》，58(301)：13–30，1963年。[JAZBJ18] Chi Jin、Zeyuan Allen-Zhu、S´ebastien Bubeck 和 Michael I Jordan。《Q学习是否已被证明高效？》载于《神经信息处理系统进展》，2018年。[KI18] R. Kleinberg 和 N. Immorlica。《充能老虎机》。载于 IEEE《计算机科学基础研讨会》，2018年。[Kla80] Roberta L. Klatzky。《人类记忆：结构与过程》。Freeman，1980年。[LC18] Andrea Locatelli 和 Alexandra Carpentier。《x臂老虎机对平滑度的自适应性》。载于《学习理论会议》，2018年。[LCCG21] Pierre Laforgue、Giulia Clerici、Nicol`o Cesa-Bianchi 和 Ran Gilad-Bachrach。《用LSD奖励打破老虎机常规：对饱和效应和季节性的最后切换依赖分析》。arXiv预印本 arXiv:2110.11819，2021年。[LCM17] Nir Levine、Koby Crammer 和 Shie Mannor。《衰退老虎机》。载于《神经信息处理系统进展》，2017年。[LHK21] David Lindner、Hoda Heidari 和 Andreas Krause。《通过策略遗憾应对机器学习决策的长期影响》。载于《国际人工智能联合会议》，2021年。[LR85] T.L Lai 和 Herbert Robbins。《渐近高效的自适应分配规则》。《应用数学进展》，6(1)：4–22，1985年。[MLS22] Dhruv Malik、Yuanzhi Li 和 Aarti Singh。《计数老虎机的完全策略遗憾界》。载于《学习理论会议》，2022年。[MTPR22] Alberto Maria Metelli、Francesco Trov`o、Matteo Pirola 和 Marcello Restelli。《随机上升老虎机》。载于《国际机器学习会议》，2022年。[MY18] Mehryar Mohri 和 Scott Yang。《与基于自动机的专家序列竞争》。载于《国际人工智能与统计会议》，2018年。[PBG19] Ciara Pike-Burke 和 Steffen Grunewalder。《恢复型老虎机》。载于《神经信息处理系统进展》，2019年。[PMR+20] Ashwin Phatak、Utkarsh Mujumdar、Robert Rein、Fabian Wunderlich、Marc Garnica 和 Daniel Memmert。《每一次投掷都有进步——对NBA精英运动员连续实时篮球罚球的校准与热身递减进行的研究》。《德国运动与体育研究期刊》，50：273–279，2020年。[Rao22] Rohan Rao。《一级方程式世界锦标赛（1950年–2022年）》。https://www.kaggle.com/datasets/rohanrao/formula-1-world-championship-1950-2020，2022年。[RVC16] Timothy J. Ricker、Evie Vergauwe 和 Nelson Cowan。《即时记忆衰减理论：从 Brown（1958）到今天（2014）》。《实验心理学季刊》，69(10)：1969–1995，2016年。

[SLC+19] Julien Seznec、Andrea Locatelli、Alexandra Carpentier、Alessandro Lazaric 和 Michal Valko。衰退老虎机并不比随机老虎机更难。载于国际人工智能与统计会议，2019 年。[Sli19] Aleksandrs Slivkins。多臂老虎机导论。机器学习基础与趋势，12(1-2):1–286，2019 年。[SMLV20] Julien Seznec、Pierre Menard、Alessandro Lazaric 和 Michal Valko。适用于静息与非静息衰退老虎机的单一算法。载于国际人工智能与统计会议，2020 年。[TL12] Cem Tekin 和 Mingyan Liu。静息与非静息老虎机的在线学习。*IEEE* 信息论汇刊，58(8):5588–5611，2012 年。[V+20] Pauli Virtanen 等。SciPy 1.0：Python 科学计算的基础算法。自然方法，17:261–272，2020 年。[WDYK20] Ruosong Wang、Simon S Du、Lin Yang 和 Sham Kakade。长时域强化学习是否比短时域强化学习更难？载于神经信息处理系统进展，2020 年。[WHFM20] Fabian Wunderlich、Herbert Heuer、Philip Furley 和 Daniel Memmert。高水平飞镖中的序列位置曲线：视觉运动校准对投掷准确性的影响。心理学研究，84:2057–2064，2020 年。[Whi81] Peter Whittle。臂获取老虎机。概率年刊，9(2):284–292，1981 年。