arXiv:2601.01069v1 [cs.LG] 3 Jan 2026

重新审视非平稳参数化老虎机与马尔可夫决策过程的加权策略

Jing Wang、Peng Zhao，*IEEE*会员，以及周志华，*IEEE*会士

摘要—**近年来，非平稳参数化老虎机受到广泛关注。应对非平稳性的原则性方法有三种，包括滑动窗口、加权和重启策略。由于许多非平稳环境呈现渐进漂移模式，加权策略常用于现实世界应用。然而，先前的理论研究表明，这种策略的分析更加复杂，且相应算法在计算效率或统计性能方面存在不足。本文重新审视非平稳参数化老虎机的加权策略。在线性老虎机（LB）中，我们发现，这一缺陷源于遗憾分析不充分，进而导致算法设计过于复杂。我们提出一种**改进的分析框架**，既简化了推导，更重要的是，得出了一种更简单的基于权重的算法；该算法与基于窗口和重启的算法同样高效，同时保持与先前研究相同的遗憾界。此外，我们的新框架还可用于改进其他参数化老虎机的遗憾界，包括广义线性老虎机（GLB）和自协调老虎机（SCB）。例如，我们设计了一种简单的加权GLB算法，其遗憾为** $\widetilde{\mathcal{O}}(k_\mu^{5/4}c_\mu^{-3/4}d^{3/4}P_T^{1/4}T^{3/4})$**，改进了先前工作的** $\widetilde{\mathcal{O}}(k_\mu^2 c_\mu^{-1}d^{9/10}P_T^{1/5}T^{4/5})$ **界，其中** k_μ **和** c_μ **刻画奖励模型的非线性，** P_T **衡量非平稳性，** d **和** T **分别表示维度和时间范围。此外，我们将该框架扩展至采用函数逼近的非平稳马尔可夫决策过程（MDP），重点研究线性混合马尔可夫决策过程和多项逻辑（MNL）混合马尔可夫决策过程。对于这两类问题，我们均提出了基于加权策略的算法，并利用分析框架建立了动态遗憾保证。**

索引词—**动态遗憾、非平稳多臂老虎机、折扣因子、在线马尔可夫决策过程、函数逼近。**

一、引言

NON-STATIONARY 参数化老虎机用于建模序贯决策问题，其中每个臂的奖励分布由未知的时变参数以结构化方式决定。近年来，由于其在许多现实世界中的非平稳在线应用（如推荐系统） [12], [13]中的重要意义，这类问题受到了广泛研究 [1]–[11] 。这项研究与强化学习（RL）的理论基础也有着密切联系，尤其是在结合函数逼近的分幕式马尔可夫决策过程（MDPs） [14]–[17]的背景下。在这些设定中，参数化老虎机经常被用于建模跨回合的奖励和转移动力学。此外，当这些底层动力学在不同回合之间表现出非平稳行为时，非平稳参数化老虎机技术自然能够扩展到捕捉奖励和转移的非平稳动力学 [18], [19]。

线性老虎机（LB）是参数化老虎机的一种基本形式：在时刻 t拉取某一臂的期望奖励，是该臂的特征向量X_t与未知参数 θ_t的内积，即 $\mathbb{E}[r_t \,|\, X_t] = X_t^\top \theta_t$。此外，广义线性老虎机（GLB）作为 LB 的推广而提出，用于建模更广泛的奖励函数（例如二元奖励），其期望奖励遵循广义线性模型，即 $\mathbb{E}[r_t \mid X_t] = \mu(X_t^\top \theta_t)$，其中 $\mu(\cdot)$是逆链接函数。此外，LB 和 GLB 在采用函数逼近的马尔可夫决策过程（MDP）中有重要应用。一个具有代表性的实例是线性混合马尔可夫决策过程，它采用 LB 对奖励函数和转移动态进行建模。在此基础上，人们提出了多项逻辑（MNL）混合马尔可夫决策过程，以解决线性函数无法对概率建模的局限。通过采用 MNL 多臂老虎机（GLB 的一种特殊情形），它能够有效地对状态转移概率建模，并确保分布有效。值得注意的是，非平稳模型允许上述模型中的参数 θ_t 随时间变化；因此，我们使用动态遗憾 [20], [21]来评估算法性能。衡量参数变化强度的典型非平稳性度量有两种：（i）在逐渐漂移的情形中，使用路径长度 $P_T = \sum_{t=2}^{T} \|\theta_{t-1} - \theta_t\|_2$衡量底层参数的累积变化量；（ii）在分段平稳情形中， Γ_T 表示在 T 轮次内参数变化的次数。

应对非平稳性有三种有理论依据的方法：滑动窗口、加权和重启策略。采用滑动窗口策略时，学习者维护一个时间窗口，其中包含最新的观测数据，以丢弃过时数据。采用加权策略时，学习者为最新数据赋予更高权重，为较早的数据赋予较低权重，从而逐步遗忘过时数据。采用重启策略时，学习者按照一定周期重启算法，以丢弃过时数据。目前，针对非平稳（广义）线性老虎机和采用线性函数逼近的分幕式马尔可夫决策过程，已知的最佳结果来自 [8]，：其提出了一种极小极大最优算法，该算法由一个非平稳性检测器和一个在近似平稳环境中表现良好的基础算法组成。每当检测器判定非平稳性超过某个限值时，算法就会

J. Wang、P. Zhao 和 Z.-H. Zhou 任职于南京大学新型软件技术国家重点实验室和人工智能学院，中国南京 210023。（电子邮箱：{wangjing,zhaop,zhouzh}@lamda.nju.edu.cn）。本文部分内容曾发表于第26届人工智能与统计国际会议论文集（AISTATS），2023年。稿件于2024年12月28日收到；修订于2025年8月6日。（通讯作者：Peng Zhao）

0000–0000 © 2025年 IEEE

表 I：在不同的非平稳赌博机和马尔可夫决策过程设置下，我们的动态遗憾界与基于权重的算法先前已知的最佳结果的比较。下文中，k_μ/c_μ 刻画 GLB/SCB 中的非线性（对于线性赌博机，该值为 1），κ 表示 MNL 混合马尔可夫决策过程中的非线性；d 为维度，H 为马尔可夫决策过程设置中一个回合的长度，路径长度 P_T 和变化次数 Γ_T 分别是漂移情形和分段平稳情形的非平稳性度量，而总路径长度 Δ 度量漂移马尔可夫决策过程情形中的非平稳性。

设置	先前工作	我们的结果
漂移线性赌博机	$\widetilde{\mathcal{O}}\left(d^{7/8}P_T^{1/4}T^{3/4}\right)$ [2]	$\widetilde{\mathcal{O}}\left(d^{3/4}P_T^{1/4}T^{3/4}\right)$ [定理 1]
漂移GLB	$\widetilde{\mathcal{O}}\left(\frac{k_\mu^2}{c_\mu}d^{9/10}P_T^{1/5}T^{4/5}\right)$ [6]	$\widetilde{\mathcal{O}}\left(\frac{k_\mu^{5/4}}{c_\mu^{3/4}}d^{3/4}P_T^{1/4}T^{3/4}\right)$ [定理 2]
漂移 SCB	$\widetilde{\mathcal{O}}\left(\frac{k_\mu^2}{c_\mu}d^{9/10}P_T^{1/5}T^{4/5}\right)$ [6]	$\widetilde{\mathcal{O}}\left(\frac{k_\mu^{5/4}}{c_\mu^{1/2}}d^{3/4}P_T^{1/4}T^{3/4}\right)$ [定理 3]
分段平稳 SCB	$\widetilde{\mathcal{O}}\left(\frac{1}{c_\mu^{1/3}}d^{2/3}\Gamma_T^{1/3}T^{2/3}\right)$ [7]	$\widetilde{\mathcal{O}}\left(d^{2/3}\Gamma_T^{1/3}T^{2/3}\right)$ [定理 4]
漂移线性混合马尔可夫决策过程	$\widetilde{\mathcal{O}}\left(Hd\Delta^{1/4}T^{3/4}\right)$ [18]	$\widetilde{\mathcal{O}}\left(Hd\Delta^{1/4}T^{3/4}\right)$ [定理 5]
漂移多项式逻辑混合马尔可夫决策过程	—	$\mathcal{O}\left(\kappa^{-1}Hd\Delta^{1/4}T^{3/4}\right)$ [定理 6]

自行重启以应对非平稳性。从这个意义上说，他们的算法可视为一种自适应重启算法。他们的算法基于 RestartUCB 算法 [3]，并结合了精心设计的非平稳性检测器和多尺度探索，能够针对线性老虎机（LB）和 GLB 实现 $\widetilde{\mathcal{O}}(\min\{\sqrt{\Gamma_T T}, P_T^{1/3}T^{2/3}\})$ 最优动态遗憾，并针对分幕式马尔可夫决策过程实现 $\widetilde{\mathcal{O}}(\min\{\sqrt{\Gamma_T T}, \Delta^{1/3}T^{2/3}\})$ 动态遗憾，其中 Δ表示总路径长度，包含奖励函数和转移函数的参数变化量之和。

在现实场景中，环境的分布变化往往呈现逐渐漂移的模式 [22]–[27],在这种情况下，软加权策略在应对非平稳性方面（从经验上看）可能比硬重启策略更具优势，这一点可从赌博机学习 [2], [3],[9], 、概念漂移分类 [28], [29], 以及自适应系统辨识 [30], [31]中观察到。因此，为非平稳参数化老虎机设计一种自适应加权算法 将极具吸引力。该算法通过赋予权重来降低过去数据的重要性，并根据环境自适应地设定权重。为此，我们考察了现有基于加权策略的非平稳参数化老虎机方法，并（令人惊讶地）发现，与重启算法等其他策略相比，当前结果存在不自然的 差距，而且从 GLB 到线性赌博机的遗憾分析也不自然地 发生了转变。

这些不自然的现象促使我们重新审视非平稳参数化老虎机 [2], [6], [7]的加权策略的算法设计与遗憾分析。事实上，关键在于对加权估计器进行估计误差分析，通常可将其分解为两部分——一部分是由参数漂移导致的偏差部分，另一部分是由随机噪声导致的方差部分。一般而言，偏差部分由非平稳策略控制，而方差部分则通过精心设计的集中不等式来处理。 [2]首次对线性老虎机（LB）的基于权重的算法进行了分析。在偏差分析中，他们引入了虚拟窗口大小，以控制偏差，从而模仿滑动窗口策略 [1] 的分析。在方差分析方面， [2] 提出了自归一化集中不等式的加权版本，该版本需要专门设计的局部范数。这增加了分析的复杂性，因为此前研究的滑动窗口 [1] 和重启策略 [3] 可以直接应用标准自归一化集中不等式 [32]。因此，他们不得不使用不同的局部范数来控制偏差部分和方差部分，导致算法设计出现意料之外的低效，也使分析变得复杂。对于线性老虎机，这使得算法 D-LinUCB [2] 需要额外维护一个协方差矩阵，作为加权版本自归一化集中不等式所用的局部范数；其效率低于基于窗口和重启的算法 [1], [3]。

由于 GLB 的结构更加丰富且复杂，这一针对加权策略的分析框架在 GLB 中引发了更为严重的问题。具体而言， [6] 研究了漂移 GLB，并设计了一种高度复杂的投影操作，沿用 [2]的思路来控制偏差和方差部分，以模拟滑动窗口分析，最终获得了 $\widetilde{\mathcal{O}}(d^{9/10}P_T^{1/5}T^{4/5})$ 的动态遗憾。不幸的是，该方法无法恢复漂移 LB（GLB 的一个特例）中基于权重的算法所享有的 $\widetilde{\mathcal{O}}(d^{7/8}P_T^{1/4}T^{3/4})$ 界 [2]。随后，[7]研究了非平稳自协调赌博机（SCB），这是 GLB 的一个具有许多吸引人结构的子类。他们只能在分段平稳设定下进行分析，却因难以使用传统分析方法界定偏差，而无法处理更具挑战性的漂移设定。此外，由于 LB 的自归一化集中不等式的加权版本 [2]无法扩展到 SCB 设定，他们进一步专门为 SCB 重新设计了一个加权版本，并以平稳 SCB 的自归一化集中不等式 [33],为基础，这给分析带来了大量额外复杂性。因此，他们在论文中提出了两个开放问题：（i）如何将基于权重的算法扩展到漂移 SCB；以及（ii）如何复现平稳

SCB [34] 中的近期进展，以改善非平稳 SCB 对 c_μ 的依赖。

我们的结果。 本文重新审视了非平稳参数化老虎机和马尔可夫决策过程中的加权策略。我们发现，早期针对加权策略的分析框架可能并不恰当，因为它依赖于仿照滑动窗口分析，以及专门设计的加权自归一化集中不等式；后者要求使用不同的 局部范数来界定偏差和方差部分，而针对每一种新场景进一步设计这种集中工具的加权版本，又带来了显著且不必要的分析复杂性。因此，尚无统一的分析框架能够直接应用于不同场景。为此，我们提出了加权策略的改进的分析框架 。在该框架中，我们为偏差部分提出了一种新的分析方法，而方差部分分析仅依赖标准自归一化集中不等式 [32] ，无需额外的加权版本，并且可以使用同一个单一的局部范数来分析偏差和方差部分。这一改进简化了加权策略的分析，也使该方法更广泛地适用于其他决策场景。此外，它还为算法设计带来了多项益处，包括提高线性赌博机的效率，以及解决 GLB 和 SCB 中遇到的投影问题。此外，我们的分析框架并不局限于老虎机场景，还可扩展到在线马尔可夫决策过程（MDP）场景。本文将结果扩展至两类基本的马尔可夫决策过程：（i）非平稳线性混合马尔可夫决策过程；（ii）非平稳多项逻辑（ MNL）混合马尔可夫决策过程。表 I 总结了我们的主要结果，并将其与加权算法目前已知的最佳结果进行了比较。具体而言，基于改进的分析框架，我们取得了以下结果：（i）对于线性赌博机，我们的方法只需维护一个协方差矩阵，而非两个，同时仍能达到与 [2] 相同的遗憾；（ii）对于 GLB，我们的方法获得了 $\widetilde{\mathcal{O}}(k_\mu^{5/4}c_\mu^{-3/4}d^{3/4}P_T^{1/4}T^{3/4})$ 遗憾界，其 d、 P_T 和 T 的阶与线性赌博机情形相匹配；（iii）对于 SCB，我们获得了 $\widetilde{\mathcal{O}}(k_\mu^{5/4}c_\mu^{-1/2}d^{3/4}P_T^{1/4}T^{3/4})$ 遗憾界；对于分段平稳 SCB，我们的方法获得了 $\widetilde{\mathcal{O}}(d^{2/3}\Gamma_T^{1/3}T^{2/3})$ 遗憾界，能够消除 c_μ^{-1} 的影响，从而解决 [7] 提出的第二个开放问题；（iv）对于线性混合马尔可夫决策过程，我们获得了 $\widetilde{\mathcal{O}}(Hd\Delta^{1/4}T^{3/4})$ 遗憾界，遗憾与重启策略达到的 [18] 相同；（v）对于 MNL 混合马尔可夫决策过程，我们首次在文献中建立了 $\widetilde{\mathcal{O}}(Hd\Delta^{1/4}T^{3/4})$ 动态遗憾界。

与我们较早的会议版本 [11],相比，本扩展版本给出了更多结果，同时采用了更简单、更清晰的分析，并改进了表述。首先，与我们的会议版本 [11]相比，本扩展版本进一步简化了分析。早期方法 [2]依赖三个关键组成部分：额外的协方差矩阵、加权自归一化集中不等式以及加权势能引理。在我们的会议版本 [11],中，我们不再需要维护额外的协方差矩阵。在本扩展版本中，我们更进一步，证明标准自归一化集中不等式足以用于分析加权策略。作为结果，加权策略分析所必需的唯一组成部分就是加权势能引理。像先前研究 [2], [11],中那样维护两个协方差矩阵并使用加权自归一化集中不等式，都是不必要的。这一简化使我们的分析和算法都更加简单且更具普适性。其次，这一简化使我们的方法更具可扩展性，也更容易推广到其他老虎机设置。早期工作 [7] 和我们的会议版本 [11] 在将该方法适配到新设置（例如 SCB）时，都需要设计新的加权自归一化集中不等式，这限制了它们的普适性。我们的改进分析消除了这一需求，使同一分析框架能够应用于不同问题，而无需针对具体问题推导加权集中不等式结果。第三，我们将结果扩展到带函数逼近的在线马尔可夫决策过程的两个基本设置：线性混合马尔可夫决策过程和多项逻辑回归（MNL）混合马尔可夫决策过程。值得注意的是，我们为 MNL 混合马尔可夫决策过程给出了首个动态遗憾保证，展现了我们改进后的加权策略分析框架的有效性和广泛适用性。

II. 相关工作

线性老虎机。 非平稳线性老虎机问题最早由 [1]研究。他们建立了 $\Omega(d^{2/3}P_T^{1/3}T^{2/3})$ 的极小极大下界，随后提出了基于滑动窗口策略的 SW-UCB 算法。然后， [2] 提出了基于加权策略的 D-LinUCB 算法， [3] 提出了基于重启策略的 RestartUCB 算法。值得注意的是，这三项工作证明了 $\widetilde{\mathcal{O}}(d^{2/3}P_T^{1/3}T^{2/3})$ 遗憾界，但正如 [35]所指出的，其遗憾分析中存在一个细微的技术漏洞。在修复该技术漏洞后，上述三种算法均可达到 $\widetilde{\mathcal{O}}(d^{7/8}P_T^{1/4}T^{3/4})$ 遗憾界 [35], [36]。然而，为了实现这一结果，这三种算法都要求在算法实现开始时将路径长度 P_T 作为输入，这并不理想。为了解决这一问题， [1] 提出了多臂老虎机之上的多臂老虎机（ BOB）策略，将其作为一种元算法来学习未知参数 P_T；该策略可以与上述算法结合，从而消除对这一先验知识的要求。随后， [8] 提出了 MASTER 算法，该算法具有理论上最优的 $\widetilde{\mathcal{O}}(\min\{d\sqrt{\Gamma_T T}, dP_T^{1/3}T^{2/3}\})$ 遗憾界，同样无需预先知道环境的非平稳程度（即 Γ_T 和 P_T），但要求满足固定臂集合假设。最近，非平稳（线性）多臂老虎机领域也取得了一些新进展 [10], [37]–[39]。此外， [40, 备注通过重启对抗性线性老虎机算法，绕过了上述技术漏洞。然而，需要注意的是，这仅适用于线性老虎机，并且要求满足固定臂集合假设以及已知 P_T。

广义线性老虎机。GLB 问题最早由 [41] 提出。他们提出了 GLM-UCB 算法，实现了 $\widetilde{\mathcal{O}}(k_\mu c_\mu^{-1}d\sqrt{T})$ 遗憾界，其中 k_μ、 c_μ 是问题相关常数， k_μ/c_μ 表示广义线性模型的非线性程度。 [6] 将平稳 GLB 推广到漂移情形，并提出了

BVD-GLM-UCB 算法，其遗憾界为 $\widetilde{\mathcal{O}}(k_\mu^2 c_\mu^{-1} d^{9/10} P_T^{1/5} T^{4/5})$。[33] 研究了 GLB 的一个特定实例，称为逻辑回归多臂老虎机（LogB）。他们首先指出，在 GLB 设定下，问题相关常数 $1/c_\mu$ 在某些情况下可能非常大，例如 LogB。随后，他们提出了 Logistic-UCB-1 算法，其遗憾界为 $\widetilde{\mathcal{O}}(c_\mu^{-1/2} d\sqrt{T})$，以及 Logistic-UCB-2 算法，其遗憾界为 $\widetilde{\mathcal{O}}(d\sqrt{T} + c_\mu^{-1})$。之后，[34] 为逻辑回归多臂老虎机建立了 $\Omega(d\sqrt{\mu'(X_*^\top \theta_*)T})$ 遗憾下界，并提出了最优算法 OFULog。[7] 将逻辑回归多臂老虎机推广到自协调赌博机，并考虑了分段平稳情形；其算法的遗憾界为 $\widetilde{\mathcal{O}}(c_\mu^{-1/3} d^{2/3} \Gamma_T^{1/3} T^{2/3})$。为处理 P_T 未知的情形，[6] 将 BVD-GLM-UCB 与 BOB 策略结合，提出了一种无参数算法，但最终结果仍非最优。与此同时，黑盒算法 [8] 能够自适应地重启平稳算法 GLM-UCB [41]，并达到 $\widetilde{\mathcal{O}}(\min\{k_\mu c_\mu^{-1}\sqrt{\Gamma_T T},\ k_\mu^{4/3} c_\mu^{-1} d P_T^{1/3} T^{2/3}\})$ 遗憾界；就 P_T 和 T 而言，该结果与非平稳线性多臂老虎机的下界相匹配，因此对于非平稳广义线性赌博机而言是最优的，因为线性多臂老虎机是广义线性赌博机的一个特例（即 $\mu(x) = x$）。近期，广义线性老虎机在效率以及关于非线性程度的遗憾最优性方面取得了显著进展。最新进展请参阅 [42]。尽管如此，这些结果关注的仍是静态遗憾设定。

函数逼近马尔可夫决策过程。近年来，采用函数逼近的强化学习受到广泛关注 [16]–[18], [43], [44], ，其中有两种基本方法：线性函数逼近和广义线性函数逼近。在线性函数逼近方面，线性混合马尔可夫决策过程由 [43], [44], 首次提出，是一种具有代表性的模型。他们提出了 UCRL-VTR 算法，其遗憾界为 $\widetilde{\mathcal{O}}(d\sqrt{H^3 T})$，其中 H 为回合时域。在此基础上，[18] 将平稳线性混合马尔可夫决策过程扩展到漂移情形，并为非平稳线性混合马尔可夫决策过程建立了 $\Omega(d^{5/6}\Delta^{1/3} H^{2/3} T^{2/3})$ 的极小极大下界，随后提出 SW-LSVI-UCB 算法，其遗憾界为 $\widetilde{\mathcal{O}}(Hd\Delta^{1/4} T^{3/4})$，其中 Δ 量化了底层参数的累积变化量。为更好地捕捉转移动态的概率特性，[16] 研究了一类广义函数逼近模型，并提出了 MNL 混合马尔可夫决策过程，该模型利用多项逻辑函数对转移进行建模。他们提出了 UCRL-MNL 算法，其遗憾界为 $\widetilde{\mathcal{O}}(\kappa^{-1} d\sqrt{H^3 T})$，其中 κ 表示 MNL 模型的非线性，H 为回合时域，K 为回合总数。随后，[17] 进一步为 MNL 混合马尔可夫决策过程取得了 $\widetilde{\mathcal{O}}(d\sqrt{H^3 T} + \kappa^{-1} d^2 H^2)$ 的遗憾界。

III. 线性多臂老虎机

本节首先介绍非平稳线性赌博机的问题设定，并描述我们的 LB-WeightUCB 算法及其理论保证。随后，我们概述引理 1 的证明，以详细说明所提出的分析框架。值得注意的是，我们的算法在不依赖专门设计的加权自归一化集中不等式的情况下，达到了已知最佳基于权重的算法 [2] 的相同遗憾界，并且效率更高。

A. 问题设定

在每一轮 t，学习者从可行集 $\mathcal{X} \subseteq \mathbb{R}^d$ 中选择一个臂 X_t，并获得奖励 r_t，使得

$$r_t = X_t^\top \theta_t + \eta_t, \tag{1}$$

其中 $\theta_t \in \mathbb{R}^d$ 是未知的时变参数，η_t 是 R-次高斯噪声。学习者的目标是最小化以下（伪）动态遗憾：

$$\text{D-REG}_T = \sum_{t=1}^{T} \max_{\mathbf{x}\in\mathcal{X}} \mathbf{x}^\top \theta_t - \sum_{t=1}^{T} X_t^\top \theta_t, \tag{2}$$

它是相对于具有未知参数完整信息的最优策略的累积遗憾。这里我们考虑漂移情形，并使用路径长度 $P_T = \sum_{t=2}^{T} \|\theta_{t-1} - \theta_t\|_2$ 作为非平稳性度量。请注意，本文聚焦于固定臂集 $\mathcal{X}$。时变臂集 $\mathcal{X}_t$ 不会给我们的基于权重的算法带来任何额外困难。唯一的区别在于，动态遗憾（2）中的最优比较基准需要更新为 $\max_{\mathbf{x}\in\mathcal{X}_t} \mathbf{x}^\top \theta_t$，而臂选择步骤（6）则在时变臂集 $\mathcal{X}_t$ 上执行，而不是在 $\mathcal{X}$ 上执行。这不会影响分析。为简单起见，本文采用固定臂集设定。

我们采用以下标准的有界性假设 [1]–[3], [32]。

假设 1。假设可行集和未知参数均有界：$\forall \mathbf{x} \in \mathcal{X}$、$\|\mathbf{x}\|_2 \le L$，并且对所有 $t \in [T]$ 均有 $\theta_t \in \Theta$成立，其中 $\Theta \triangleq \{\theta \mid \|\theta\|_2 \le S\}$。

B. 算法与遗憾保证

我们提出了 LB-WeightUCB 算法，该算法在享有更高效率的同时，能够达到与先前方法相同的遗憾保证。我们首先给出所采用的估计量，然后通过改进的分析框架推导其估计误差上界，这是算法设计和遗憾分析的关键。基于估计误差界，我们提出选择准则，并最终给出其动态遗憾的理论保证。

估计量。我们采用了一种类似于 D-LinUCB 的加权正则化最小二乘估计量 [2]，该估计量 θ_t 是以下问题的解，

$$\min_{\theta}\ \lambda \|\theta\|_2^2 + \sum_{s=1}^{t-1} w_{t-1,s} \left(X_s^\top \theta - r_s\right)^2, \tag{3}$$

其中，$\lambda > 0$ 是正则化系数，$\forall t \in [T], s \in [t-1]$，$w_{t-1,s}$ 是加权因子。为处理非平稳性，我们设定 $w_{t,s} = \gamma^{t-s}$，其中 $\gamma \in (0,1)$ 是折扣因子。这种方法对较早的数据赋予较低的权重，同时对较新的数据赋予较高的权重，从而更好地适应随时间发生的变化。显然，θ_t 存在闭式解 $\widehat{\theta}_t = V_{t-1}^{-1}(\sum_{s=1}^{t-1} w_{t-1,s} r_s X_s)$，

算法 1LB-WeightUCB

要求： 时间范围 T、折扣因子 γ、置信度 δ，正则化项 λ、参数 S、L 和 R

1: 设置 $V_0 = \lambda I_d, \theta_1 = \mathbf{0}$ 并根据（5）计算 β_0
2: **对于** $t = 1, 2, ..., T$ **do**
3: 选择 $X_t = \arg\max_{\mathbf{x}\in\mathcal{X}} \left\{ \langle \mathbf{x}, \widehat{\theta}_t \rangle + \beta_{t-1} \|\mathbf{x}\|_{V_{t-1}^{-1}} \right\}$
4: 接收奖励 r_t
5: 更新 $V_t = \gamma V_{t-1} + X_t X_t^\top + (1-\gamma)\lambda I_d$
6: 计算 $\widehat{\theta}_{t+1}$ 根据 (3) 和 β_t 根据 (5)，其中 $w_{t,s} = \gamma^{t-s}$
7: **</style结束**

其中，$V_t = \lambda I_d + \sum_{s=1}^{t} w_{t,s} X_s X_s^\top, V_0 = \lambda I_d$ 是协方差矩阵。请注意，这个闭式解还可以进一步转化为递归公式，使得$V_t = \gamma V_{t-1} + X_t X_t^\top + (1-\gamma)\lambda I_d$，其中我们设定 $w_{t,s} = \gamma^{t-s}$。这样无需存储历史数据即可在线更新，这是加权策略相较于滑动窗口策略的另一项重要计算优势。

上置信界。对于估计量 (3)，我们给出以下估计误差界。值得注意的是，这与先前结果 [2, 附录 B.3 中定理 2] 证明的第二步和第三步不同。这一差异是我们的算法效率得以提升的关键，稍后将对此进行讨论。

引理 1。对于任意 $\mathbf{x} \in \mathcal{X}$, $\gamma \in (0, 1)$ 以及 $\delta \in (0, 1)$，以至少 $1-\delta$ 的概率，以下结论对所有 $t \in [T]$ 均成立

$$\left|\mathbf{x}^\top(\widehat{\theta}_t - \theta_t)\right| \leq L^2 \sqrt{\frac{d}{\lambda}} \sum_{p=1}^{t-1} \sqrt{\sum_{s=1}^{p} w_{t-1,s} \|\theta_p - \theta_{p+1}\|_2} + \beta_{t-1} \|\mathbf{x}\|_{V_{t-1}^{-1}}, \tag{4}$$

其中 β_t 是由以下方式设定的置信区域半径

$$\beta_t = \sqrt{\lambda} S + R \sqrt{2 \log \frac{1}{\delta} + d \log \left(1 + \frac{L^2 \sum_{s=1}^{t} w_{t,s}}{\lambda d} \right)}. \tag{5}$$

引理 1 的证明见附录 A-B。根据引理 1，我们可以将臂选择准则指定为

$$X_t = \arg\max_{\mathbf{x}\in\mathcal{X}} \left\{ \langle \mathbf{x}, \widehat{\theta}_t \rangle + \beta_{t-1} \|\mathbf{x}\|_{V_{t-1}^{-1}} \right\}. \tag{6}$$

整体算法总结于算法 1 中。从算法 1 第 5 行的更新过程中可以看出，我们的算法只需维护一个单一的 协方差矩阵$V_{t-1} \in \mathbb{R}^{d\times d}$。相比之下， [2] 中提出的算法的选择准则是

$$X_t = \arg\max_{\mathbf{x}\in\mathcal{X}} \left\{ \langle \mathbf{x}, \widehat{\theta}_t \rangle + \beta_{t-1} \|\mathbf{x}\|_{V'^{-1}_{t-1} \widetilde{V}_{t-1} V'^{-1}_{t-1}} \right\},$$

其中， β_{t-1} 与我们选择准则 (6) 中的类似，
$V'_{t-1} = \lambda I_d + \sum_{s=1}^{t-1} \gamma^{t-s-1} X_s X_s^\top \in \mathbb{R}^{d\times d}$ $\widetilde{V}_{t-1} = \lambda I_d +$ 则是 $\sum_{s=1}^{t-1} \gamma^{2(t-s-1)} X_s X_s^\top \in \mathbb{R}^{d\times d}$
一个额外的协方差矩阵。因此，我们的算法比他们的算法更高效，因为它只需维护一个协方差矩阵，而非两个。这是因为，在引理 1 的分析中，我们只使用 V_{t-1}^{-1} 作为局部范数，来分析偏差和方差部分，而 [2] 的算法则需要分别使用 l_2-范数和 $V_{t-1}^{-1} \widetilde{V}_{t-1} V_{t-1}^{-1}$-范数来控制偏差部分和方差部分。在 III-C 节中，我们概述了引理 1 的分析框架，更详细的讨论见附录 A-A。此外，我们证明了我们的算法取得的遗憾与 [2] 的算法相同（在 d方面甚至略好）。

定理1. 设加权因子为 $w_{t,s} = \gamma^{t-s}$，其中 $\gamma \in (1/T, 1)$，并设 $\lambda = d$，则 ***LB-WeightUCB***（算法 *1*）的动态遗憾以至少 $1 - 1/T$ 的概率满足以下界：

$$\text{D-REG}_T \leq \widetilde{\mathcal{O}} \left(\frac{1}{(1-\gamma)^{3/2}} P_T + d(1-\gamma)^{1/2} T \right).$$

此外，将折扣因子最优地设为 $\gamma = 1 - \max\{1/T, \sqrt{P_T/(dT)}\}$ 后，***LB-WeightUCB*** 可保证

$$\text{D-REG}_T \leq \begin{cases} \widetilde{\mathcal{O}} \left(d^{3/4} P_T^{1/4} T^{3/4} \right) & \textit{when } P_T \geq d/T, \\ \widetilde{\mathcal{O}}(d\sqrt{T}) & \textit{when } P_T < d/T. \end{cases}$$

与先前研究 [1]–[3], 相比，当 $P_T \geq d/T$ 时，我们的方法将 $\widetilde{\mathcal{O}}(d^{7/8} P_T^{1/4} T^{3/4})$ 改进为 $\widetilde{\mathcal{O}}(d^{3/4} P_T^{1/4} T^{3/4})$。我们指出，这种维度依赖性的改进完全归功于对折扣因子进行了比 [2], 所采用的方法更精细的调优；后者没有将维度纳入调优。通过更精细的调优，他们的算法及其遗憾界也能得到改进。定理 1 的证明见附录 A-C。

此外，请注意，折扣因子 γ 的最优选择需要提前知道 P_T。为了在未知 P_T 的情况下获得无参数结果，我们的算法可以与 BOB 策略 [1] 相结合，并达到一个 $\widetilde{\mathcal{O}}(d^{3/4} P_T^{1/4} T^{3/4})$ 界。我们在附录 H 中给出了 LB-WeightUCB 的 BOB 版本及详细的遗憾分析。然而，这个界并非最优；受 [8], 启发，可以基于我们的结果设计一种自适应加权算法，在无需 P_T 先验知识的情况下进一步实现最优动态遗憾。这极具挑战性，因为在每一轮 $t \in [T]$，我们只能收到一个数据对 (X_t, r_t)，不足以让学习者实时更新折扣因子 γ_t。同时，MASTER 算法 [8] 可以视为自适应加权算法的一个特例，因为它只包含两种情况：将 $\gamma_t = 0$ 设为在时间 t 重启，以及将 $\gamma_t = 1$ 设为继续运行。然而，对于自适应加权算法，折扣因子 γ_t 的选择可以在 $[0, 1]$，中连续变化，这比二元决策更困难。我们将此留作未来研究的一项重要开放问题。此外，我们注意到，我们的方法可以处理时变臂集情形，而 MASTER 依赖于固定臂集合假设。在时变臂集下能否实现最优遗憾，仍不清楚。

C. 改进的分析框架

本节给出引理 1（加权线性老虎机的估计误差分析）的证明概要，同时也介绍了我们提出的分析框架。

证明概要。 根据模型假设 (1) 和估计器 (3)，估计误差可以分为两部分：

$$\widehat{\theta}_t - \theta_t = \underbrace{V_{t-1}^{-1}\left(\sum_{s=1}^{t-1} w_{t-1,s} X_s X_s^\top (\theta_s - \theta_t)\right)}_{\texttt{bias part}} + \underbrace{V_{t-1}^{-1}\left(\sum_{s=1}^{t-1} w_{t-1,s}\eta_s X_s - \lambda\theta_t\right)}_{\texttt{variance part}},$$

其中，偏差部分由参数漂移导致，而方差部分则源于随机噪声。接着，根据柯西-施瓦茨不等式，对于任意$\mathbf{x} \in \mathcal{X}$，

$$|\mathbf{x}^\top(\widehat{\theta}_t - \theta_t)| \le \|\mathbf{x}\|_{V_{t-1}^{-1}} (A_t + B_t), \tag{7}$$

其中 $A_t = \| \sum_{s=1}^{t-1} w_{t-1,\ s} X_s X_s^\top (\theta_s - \theta_t) \|_{V_{t-1}^{-1}}$ 和 $B_t = \| \sum_{s=1}^{t-1} w_{t-1,\ s}\eta_s X_s - \lambda\theta_t \|_{V_{t-1}^{-1}}$。

为（7）选择合适的局部范数，是简化并改进估计误差分析的关键。请注意，先前的分析 [2] 不得不使用不同的局部范数：在偏差部分使用 l_2-范数，在方差部分使用 $V_{t-1}'^{-1}\widetilde{V}_{t-1}V_{t-1}'^{-1}$-范数，即

$$|\mathbf{x}^\top(\widehat{\theta}_t - \theta_t)| \le \|\mathbf{x}\|_2 A_t' + \|\mathbf{x}\|_{V_{t-1}'^{-1}\widetilde{V}_{t-1}V_{t-1}'^{-1}} B_t', \tag{8}$$

其中，我们有 $A_t' = \|V_{t-1}'^{-1} \sum_{s=1}^{t-1} \gamma^{t-s-1} X_s X_s^\top (\theta_s - \theta_t)\|_2$，$B_t' = \| \sum_{s=1}^{t-1} \gamma^{t-s-1}\eta_s X_s - \lambda\theta_t \|_{\widetilde{V}_{t-1}^{-1}}$ $V_t' = \lambda I_d +$和 $\sum_{s=1}^{t-1}\gamma^{t-s} X_s X_s^\top$，$\widetilde{V}_t = \lambda I_d + \sum_{s=1}^{t}\gamma^{2(t-s)} X_s X_s^\top$。由于需要使用滑动窗口分析来分析偏差部分，他们必须使用 l_2-范数来得到 A_t' 的形式。对于方差部分，为了使用加权自归一化集中不等式，他们使用 $V_{t-1}'^{-1}\widetilde{V}_{t-1}V_{t-1}'^{-1}$-范数控制 $\mathbf{x}$项，从而可以用 $\widetilde{V}_{t-1}^{-1}$对 B_t' 项进行范数化。

作为改进，我们直接使用相同的 V_{t-1}^{-1}-范数来控制两部分，这得益于我们对偏差部分采用的新分析方法，以及对方差部分采用的改进分析方法。

偏差部分分析。偏差部分分析的关键步骤是提取底层参数的变化量，如下所示，

$$\begin{aligned} A_t &\le L\sum_{p=1}^{t-1}\sum_{s=1}^{p} w_{t-1,s} \|X_s\|_{V_{t-1}^{-1}} \|\theta_p - \theta_{p+1}\|_2 \\ &\le L\sqrt{d}\sum_{p=1}^{t-1}\sqrt{\sum_{s=1}^{p} w_{t-1,s}} \|\theta_p - \theta_{p+1}\|_2 . \end{aligned}$$

基于此，我们可以得到与路径长度 $P_T = \sum_{t=2}^{T} \|\theta_{t-1} - \theta_t\|_2$ 相关的偏差部分上界。上述论证的完整证明见引理 7。

方差部分分析。 关键在于分析以下带有加权因子的自归一化项$w_{t-1,\ s}$，

$$\begin{aligned} B_t &\le \left\|\sum_{s=1}^{t-1} w_{t-1,s}\eta_s X_s\right\|_{V_{t-1}^{-1}} + \sqrt{\lambda}S \\ &= \left\|\sum_{s=1}^{t-1} \widetilde{\eta}_s \widetilde{X}_s\right\|_{V_{t-1}^{-1}} + \sqrt{\lambda}S. \end{aligned}$$

这里请注意，$V_t = \lambda I_d + \sum_{s=1}^{t} \widetilde{X}_s \widetilde{X}_s^\top$，其中我们定义了 $\widetilde{\eta}_s \triangleq \sqrt{w_{t-1,\ s}}\eta_s$ 和 $\widetilde{X}_s \triangleq \sqrt{w_{t-1,\ s}} X_s$。值得注意的是，对于所有 $t \in [T]$ 和 $s \in [t-1]$，均有 $|w_{t-1,\ s}| \le 1$，这确保了 $\widetilde{\eta}_s$ 仍然是 R-次高斯的，具体论证可参见引理 8。因此，我们可以直接应用自归一化集中不等式（定理7）来控制方差项，而无需使用 [2] 的定理1中提出的自归一化集中不等式的加权版本。

结合对偏差部分和方差部分的分析，我们即可完成引理 1 的证明。 □

加权策略的估计误差分析首先需要将估计误差分解为偏差部分和方差部分，然后分别进行分析。 [2] 使用不同的局部范数来分解估计误差，模仿滑动窗口分析处理偏差项，并专门为方差项设计了加权自归一化集中不等式。我们改进的分析框架表明，这种复杂处理并无必要。偏差和方差可以使用相同的局部范数进行分解，同时针对加权策略专门分析偏差部分；方差项则不再需要专门设计的集中不等式或额外的局部范数。得到估计误差界后，我们接着进行遗憾分析，其中需要使用加权势能引理来界定遗憾。

加权势能引理。 术语 $\|\mathbf{x}\|_{V_{t-1}^{-1}}$ 在 (7) 中会在遗憾分析的方差部分产生一个求和项 $\sum_{t=1}^{T} \|X_t\|_{V_{t-1}^{-1}}$。由于 $V_{t-1} = \lambda I_d + \sum_{s=1}^{t-1} w_{t-1,s} X_s X_s^\top$包含加权因子，我们无法直接应用标准势能引理。因此，我们需要在加权因子为$w_{t,\ s} = \gamma^{t-s}$的遗憾分析中使用加权势能引理（见引理 9），使得

$$\sum_{t=1}^{T} \|X_t\|_{V_{t-1}^{-1}} = \widetilde{\mathcal{O}}\left(T\sqrt{d\log\frac{1}{\gamma}}\right). \tag{9}$$

接下来需要恰当地选择折扣因子 γ，以使遗憾界中的偏差项和方差项达到良好平衡。较小的 γ 对应更快的遗忘速度，这有助于减小非平稳性导致的偏差。然而，较小的 γ 也会使方差增大，具体表现为 (9) 中的关键项 $\mathcal{O}(\sqrt{\log\frac{1}{\gamma}})$。有关如何最优地选择 γ 以平衡这两部分的详情，请参阅附录 A-C。

总而言之，对于非平稳线性赌博机分析，加权策略与重启策略或滑动窗口策略一样简单，唯一的区别在于，遗憾分析需要用到加权势能引理，而无需更复杂的偏差结果。

注释 1。 我们分析框架中的关键步骤 (7) 也解决了 GLB 中的投影问题。具体而言，执行投影步骤后，偏差-方差分解只能在 V_{t-1}^{-1}-范数下进行。为适应先前的分析 (8)，[6]必须在算法中引入一种高度复杂的投影操作

而我们的框架由于对偏差部分和方差部分使用相同的 V_{t-1}^{-1}-范数，已经满足了这一条件。

IV. 广义线性老虎机

本节中，我们将加权策略应用于漂移GLB。与目前已知用于漂移GLB的最佳基于权重的算法相比， [6], 我们的算法更简单，同时具有更优的动态遗憾。此外，我们还研究了GLB中的一个重要类别，即自协调老虎机（SCB），并进一步改进了该设定下的理论保证。

A. 问题设定

GLB 假设存在一个逆链接函数 $\mu: \mathbb{R} \to \mathbb{R}$ 使得$r_t = \mu(X_t^\top \theta_t) + \eta_t$，其中 $\theta_t \in \mathbb{R}^d$是未知参数，并且可能随时间变化。与线性赌博机类似，我们将 GLB 的动态遗憾 定义如下:

$$\text{D-Reg}_T = \sum_{t=1}^{T} \left(\max_{\mathbf{x} \in \mathcal{X}} \mu(\mathbf{x}^\top \theta_t) - \mu(X_t^\top \theta_t) \right). \tag{10}$$

在 GLB 下，我们采用与线性赌博机相同的假设，包括 R-次高斯噪声、可行集和未知回归参数的有界性（假设 1）。此外，我们遵循逆链接函数 [6], [41], [45]的标准有界性假设。

假设 2. 逆链接函数 μ: $\mathbb{R} \to \mathbb{R}$ 是 k_μ-利普希茨连续的，并且连续可微，满足

$$c_\mu \triangleq \inf_{\{\theta \in \Theta, \mathbf{x} \in \mathcal{X}\}} \mu'(\theta^\top \mathbf{x}) > 0, \quad \Theta = \{\theta \mid \|\theta\|_2 \le S\}.$$

先前研究 [3], [40] 定义了一个类似的参数 $\widetilde{c}_\mu \triangleq \inf_{\{\theta \in \mathbb{R}^d, \mathbf{x} \in \mathcal{X}\}} \mu'(\theta^\top \mathbf{x}) > 0$，并得到了随 $1/\widetilde{c}_\mu$变化的遗憾上界。显然， $\widetilde{c}_\mu$ 小于我们定义的 c_μ（且可能小得多），因为c_μ 定义在 Θ 上，而 $\widetilde{c}_\mu$ 定义在 $\mathbb{R}$上。因此， $\widetilde{c}_\mu$ 不太适合出现在遗憾上界中。

B. 算法与遗憾保证

我们提出了 GLB-WeightUCB，与先前基于权重的算法 [6]相比，它是一种更简单且具有更好理论保证的算法。其关键改进得益于我们改进的分析框架，该框架与简单的投影步骤兼容。

估计量。在第 t次迭代时，我们首先采用准最大似然估计量（QMLE），不考虑向可行域的投影。具体而言，估计量 θ_t 是下列加权正则化方程的解:

$$\lambda c_\mu \theta + \sum_{s=1}^{t-1} w_{t-1,s} \left(\mu(X_s^\top \theta) - r_s \right) X_s = 0. \tag{11}$$

与估计量 (3) 类似，我们设置 $w_{t,\,s} = \gamma^{t-s}$，其中 $\gamma \in (0,\ 1)$ 是折扣因子。鉴于 θ_t 可能不属于可行集 Θ，且 c_μ 是定义在参数 $\theta \in \Theta$上的，我们需要执行以下投影步骤

$$\widetilde{\theta}_t = \arg\min_{\theta \in \Theta} \|g_t(\widehat{\theta}_t) - g_t(\theta)\|_{V_{t-1}^{-1}}, \tag{12}$$

算法 2GLB-WeightUCB

要求： 时间范围 T、折扣因子 γ、置信度 δ，
正则化项 λ、链接函数 μ、参数 S、 L和 R

设置 $V_0 = \lambda I_d, \widehat{\theta}_1 = \mathbf{0}$，计算 k_μ，c_μ 和 $\bar{\beta}_0$ 根据 (14)
对于 $t = 1, 2, ..., T$ **do**
  if $\|\widehat{\theta}_t\|_2 \le S$ **然后**
    let $\theta_t = \widehat{\theta}_t$
  else
    进行投影并得到 $\widetilde{\theta}_t$ 根据 (12)
  结束 if
  根据（15）选择 X_t
  获得奖励 r_t
  更新 $V_t = \gamma V_{t-1} + X_t X_t^\top + (1-\gamma)\lambda I_d$
  计算 $\widehat{\theta}_{t+1}$ 根据（11）计算 θ_{t+1}，并使用 $w_{t,\,s} = \gamma^{t-s}$
  计算 $\bar{\beta}_t$ 根据 (14)，$w_{t,\,s} = \gamma^{t-s}$
结束循环

where $V_t = \lambda I_d + \sum_{s=1}^{t} w_{t,s} X_s X_s^\top$ and $g_t(\theta)$ is

$$g_t(\theta) \triangleq \lambda c_\mu \theta + \sum_{s=1}^{t-1} w_{t-1,s} \mu(X_s^\top \theta) X_s. \tag{13}$$

然而，先前工作 [6]无法像平稳 GLB 或分段平稳 GLB 那样，在漂移情形下进行同样简单的投影，因为它们使用不同的局部范数分别度量偏差部分和方差部分，以进行估计误差分析。因此，它们必须设计复杂的投影，以确保偏差部分和方差部分能够由不同的局部范数度量（参见 [6, 第 4.1], 节以及我们在附录 B-A 中的重述）。

我们的改进的分析框架与这一投影操作兼容，这得益于我们的分析框架对偏差部分和方差部分采用了相同的局部范数。

上置信界。 对于采用投影 (12) 的估计量 (11)，我们构造如下估计误差界。

引理 2。 对于任意 $\mathbf{x} \in \mathcal{X}$、 $\gamma \in (0、1)$ 以及 $\delta \in (0、1)$，以至少 $1-\delta$ 的概率，对所有 $t \in [T]$，以下结论成立

$$\begin{aligned} &\left| \mu(\mathbf{x}^\top \widetilde{\theta}_t) - \mu(\mathbf{x}^\top \theta_t) \right| \\ &\le \frac{2k_\mu}{c_\mu} \left(\sum_{p=1}^{t-1} C(p) \|\theta_p - \theta_{p+1}\|_2 + \bar{\beta}_{t-1} \|\mathbf{x}\|_{V_{t-1}^{-1}} \right), \end{aligned}$$

其中 $C(p) \triangleq k_\mu L^2 \sqrt{\frac{d}{\lambda}} \sqrt{\sum_{s=1}^{p} w_{t-1,\,s}}$且 $\bar{\beta}_t$ 是由下式设定的置信区域半径

$$\sqrt{\lambda} c_\mu S + R \sqrt{2 \log \frac{1}{\delta} + d \log \left(1 + \frac{L^2 \sum_{s=1}^{t} w_{t,s}}{\lambda d} \right)}. \tag{14}$$

引理 2 的证明见附录 B-B。然后，根据引理 2，我们可以将臂选择准则表述为

$$X_t = \arg\max_{\mathbf{x} \in \mathcal{X}} \left\{ \mu(\mathbf{x}^\top \widetilde{\theta}_t) + \frac{2k_\mu}{c_\mu} \bar{\beta}_{t-1} \|\mathbf{x}\|_{V_{t-1}^{-1}} \right\}. \tag{15}$$

整体算法总结于算法 2 中。

请注意，估计方程 (11) 和置信半径 (14) 与 [6] 中算法 1 使用的相同。

但重要的是，两种方法最终得到的（投影后）估计量有显著差异。借助更简单的投影操作和我们改进的分析框架，我们可以立即为基于权重的算法获得改进的遗憾保证。

定理2。令加权因子为 $w_{t,s}=\gamma^{t-s}$，其中 $\gamma\in(1/T,1)$，并令 $\lambda=d/c_\mu^2$，则 *GLB-WeightUCB*（算法 *2*）的遗憾以至少 $1-1/T$ 的概率满足

$$\text{D-REG}_T\le\widetilde{\mathcal{O}}\left(k_\mu^2\frac{1}{(1-\gamma)^{3/2}}P_T+\frac{k_\mu}{c_\mu}d(1-\gamma)^{1/2}T\right).$$

通过最优设置 $\gamma=1-\max\{1/T,\sqrt{k_\mu c_\mu P_T/(dT)}\}$，*GLB-WeightUCB* 可实现以下动态遗憾：

$$\text{D-REG}_T\le\begin{cases}\widetilde{\mathcal{O}}\left(\frac{k_\mu^{5/4}}{c_\mu^{3/4}}d^{3/4}P_T^{1/4}T^{3/4}\right) & \textit{when } P_T\ge\frac{d}{k_\mu c_\mu T},\\ \widetilde{\mathcal{O}}\left(\frac{k_\mu}{c_\mu}d\sqrt{T}\right) & \textit{when } 0\le P_T<\frac{d}{k_\mu c_\mu T}.\end{cases}$$

与 BVD-GLM-UCB（已知最优的漂移 GLB 加权算法）相比，[6], 着眼于对 d、P_T 和 T 的依赖关系，可以看出，我们的方法将遗憾从 $\widetilde{\mathcal{O}}(d^{9/10}P_T^{1/5}T^{4/5})$ 降至 $\widetilde{\mathcal{O}}(d^{3/4}P_T^{1/4}T^{3/4})$。此外，我们的结果还改善了他们关于 c_μ 的依赖关系，改进幅度从 c_μ^{-1} 提高到 $c_\mu^{-3/4}$。

C. 自协调赌博机

本节研究自协调赌博机（SCB），这是 GLB 的一个重要子类，具有许多吸引人的结构。对于 SCB，奖励的分布属于典范指数族：$\mathbb{P}_\theta[r\,|\mathbf{x}]=\exp(r\mathbf{x}^\top\theta-b(\mathbf{x}^\top\theta)+c(r))$，其中 $b(\cdot)$ 是二次连续可微函数，而 $c(\cdot)$ 是实值函数。利用指数族的良好性质，我们有 $\mathbb{E}[r\,|\,\mathbf{x}]=b'(\mathbf{x}^\top\theta)$ 和 $\text{Var}[r\,|\,\mathbf{x}]=b''(\mathbf{x}^\top\theta)$，其中 b' 表示函数 b 的一阶导数，而 b'' 表示其二阶导数。接着，我们可以引入（逆）链接函数 $\mu(\cdot)\triangleq b'(\cdot)$，使得

$$\mathbb{E}[r_t\,|\,X_t]=\mu(X_t^\top\theta_t),\text{Var}[r_t\,|\,X_t]=\mu'(X_t^\top\theta_t).\tag{16}$$

SCB 要求链接函数满足 $|\mu''|\le\mu'$，这通常称为广义自协调性质。我们进一步引入记号 $\eta_t=r_t-\mu(X_t^\top\theta_t)$ 来表示噪声。SCB 成功地对许多重要的现实世界应用进行建模，并捕捉奖励结构。例如，选择 $\mu(x)=(1+e^{-x})^{-1}$ 可得到逻辑回归多臂老虎机（LogB），它常用于对推荐系统中的二元反馈奖励进行建模 [46]–[48]。

我们采用与线性赌博机和 GLB 相同的若干标准假设，包括可行集和未知参数有界（假设1），以及关于（逆）链接函数的非线性度量（假设 2）。此外，与 [7], 类似，我们还需要对奖励的有界性作出假设。为便于分析，令 $L=1$，这意味着 $\|\mathbf{x}\|_2\le1$ 对所有$\mathbf{x}\in\mathcal{X}$均成立。

假设 3. 每一轮获得的奖励均满足$0\le r_t\le m$，对所有 $t\in[T]$ 以及某个常数 $m>0$均成立。

算法。我们提出了 SCB-WeightUCB 算法。与 GLB 相比，我们在投影和遗憾分析中使用了一种新的局部范数，这是改进 c_μ^{-1} 阶数的关键。在迭代 t 中，我们首先采用与 GLB 相同的最大似然估计量，其定义见式 (11)。与 GLB 不同的是，这里我们使用一种新的局部范数，将其投影到可行集 Θ 上，

$$\widetilde{\theta}_t=\arg\min_{\theta\in\Theta}\left\|g_t(\widehat{\theta}_t)-g_t(\theta)\right\|_{H_t^{-1}(\theta)},\tag{17}$$

其中，$g_t(\theta)$ 与式（13）相同，而 $H_t(\theta)$ 定义为

$$H_t(\theta)\triangleq\lambda c_\mu I_d+\sum_{s=1}^{t-1}w_{t-1,s}\mu'(X_s^\top\theta)X_sX_s^\top.\tag{18}$$

值得注意的是，与 V_t 相比，$H_t(\theta)$ 取决于沿动力学的函数曲率，因此能够捕捉更多局部信息。将此投影步骤与定理 8 中重述的标准自归一化集中不等式相结合，可以在遗憾中消去一个常数 $c_\mu^{-1/2}$。对于采用投影（17）的估计量（11），我们构造如下估计误差界。

引理 3。对于任意$\mathbf{x}\in\mathcal{X}$、$\gamma\in(0、1)$以及 $\delta\in(0、1)$，以至少 $1-\delta$ 的概率，以下结论对所有 $t\in[T]$ 均成立

$$\left|\mu(\mathbf{x}^\top\widetilde{\theta}_t)-\mu(\mathbf{x}^\top\theta_t)\right|\le\frac{\sqrt{4+8S}k_\mu}{\sqrt{c_\mu}}\left(\sum_{p=1}^{t-1}\frac{C(p)}{\sqrt{c_\mu}}\|\theta_p-\theta_{p+1}\|_2+\widetilde{\beta}_{t-1}\|\mathbf{x}\|_{V_{t-1}^{-1}}\right),$$

其中 $C(p)\triangleq k_\mu L^2\sqrt{\frac{d}{\lambda}}\sqrt{\sum_{s=1}^{p}w_{t-1,s}}$，且 $\widetilde{\beta}_t$ 是由以下方法设定的置信区域半径

$$\widetilde{\beta}_t=\frac{\sqrt{\lambda c_\mu}}{2m}+\frac{2m}{\sqrt{\lambda c_\mu}}\left(\log\frac{1}{\delta}+d\log2\right)+\frac{dm}{\sqrt{\lambda c_\mu}}\log\left(1+\frac{L^2k_\mu\sum_{s=1}^{t}w_{t,s}}{\lambda c_\mu d}\right)+\sqrt{\lambda c_\mu}S.\tag{19}$$

引理 3 的证明见附录 C-A。根据引理 3，我们可以选择臂 X_t，使得

$$\arg\max_{\mathbf{x}\in\mathcal{X}}\left\{\mu(\mathbf{x}^\top\widetilde{\theta}_t)+2\sqrt{1+2S}\frac{k_\mu}{\sqrt{c_\mu}}\widetilde{\beta}_{t-1}\|\mathbf{x}\|_{V_{t-1}^{-1}}\right\}.\tag{20}$$

我们的 SCB 算法（命名为 SCB-WeightUCB）遵循与算法 2，θ_t 相同的步骤，区别在于 是通过式 (17) 计算的，β_{t-1} 是通过式 (19) 计算的，而 X_t 是通过式 (20) 计算的。此外，我们对 SCB-WeightUCB 算法有如下保证。

定理3. 对于所有 $\gamma\in(1/T,\,1)$，$\lambda=d\log(T)/c_\mu$，*SCB-WeightUCB* 的动态遗憾以至少 $1-1/T$ 的概率有如下上界：

$$\text{D-REG}_T\le\widetilde{\mathcal{O}}\left(\frac{k_\mu^2}{\sqrt{c_\mu}}\frac{1}{(1-\gamma)^{3/2}}P_T+\frac{k_\mu}{\sqrt{c_\mu}}d(1-\gamma)^{1/2}T\right).$$

通过设置 $\gamma=1-$最大值$\{1/T,\sqrt{k_\mu P_T/(dT)}\}$，我们得到

$$\text{D-REG}_T\le\begin{cases}\widetilde{\mathcal{O}}\left(\frac{k_\mu^{5/4}}{c_\mu^{1/2}}d^{3/4}P_T^{1/4}T^{3/4}\right) & \textit{when } P_T\ge\frac{d}{k_\mu T},\\ \widetilde{\mathcal{O}}\left(\frac{k_\mu}{c_\mu^{1/2}}d\sqrt{T}\right) & \textit{when } 0\le P_T<\frac{d}{k_\mu T}.\end{cases}$$

与 GLB 相比，借助自协调性质，我们将 c_μ 的阶从 c_μ^{-1} 提升至 $c_\mu^{-1/2}$。在近似平稳的环境中（P_T 足够小），我们的结果能够恢复 LogUCB1 算法 [33] 的性能。定理3的证明见附录C-B。

此外，对于分段平稳SCB，我们提出了 SCB-PW-WeightUCB算法，该算法摆脱了c_μ 的影响，因此直接改进了 [7]。

定理4。对于所有 $\gamma \in (1/2, 1)$、$D = \log(T)/\log(1/\gamma)$和 $\lambda = d\log(T)/c_\mu$，*SCB-PW-WeightUCB*的遗憾以至少 $1 - 1/T$ 的概率有如下界：

$$\text{D-REG}_T \le \widetilde{\mathcal{O}}\left(\frac{1}{1-\gamma}\Gamma_T + \frac{1}{\sqrt{1-\gamma}} + d\sqrt{(1-\gamma)T}\right).$$

通过设置 $\gamma = 1 - \max\{1/T, (\Gamma_T/(dT))^{2/3}\}$，我们得到

$$\text{D-REG}_T \le \begin{cases} \widetilde{\mathcal{O}}\left(d^{2/3}\Gamma_T^{1/3}T^{2/3}\right) & \textit{when } \Gamma_T \ge d/\sqrt{T}, \\ \widetilde{\mathcal{O}}\left(d\sqrt{T}\right) & \textit{when } 0 \le \Gamma_T < d/\sqrt{T}. \end{cases}$$

完整的算法和分析见附录D。

V. 线性混合马尔可夫决策过程

本节将加权策略应用于非平稳线性混合马尔可夫决策过程，并介绍我们的WeightUCRL算法及其理论保证。我们的算法实现了与之前基于重启的算法 [18]相同的遗憾界。

A. 问题设定

我们基于此前对分幕式非平稳马尔可夫决策过程 [18] 的定义，并给出相应的学习协议。

分幕式非平稳马尔可夫决策过程。一个分幕式马尔可夫决策过程由元组 $M =$定义，其中$\mathcal{S}$、$\mathcal{A}$、H、$\mathbb{P}$、r；其中，$\mathcal{S}$是状态空间；$\mathcal{A}$ 是动作空间；H 是每个回合的长度；$\mathbb{P} = \{\mathbb{P}_h^k\}_{h\in[H],\ k\in[K]}$，其中 $\mathbb{P}_h^k$：$\mathcal{S}\times\mathcal{A}\times\mathcal{S} \to [0,1]$ 是在第 k个回合的第 h 步的转移概率；以及$r = \{r_h^k\}_{h\in[H],\ k\in[K]}$，其中 r_h^k：$\mathcal{S}\times\mathcal{A} \to [0,1]$ 是在第 k个回合的第 h步的奖励函数。策略定义为 $\pi = \{\pi_h^k\}_{h\in[H],\ k\in[K]}$，其中每个 π_h^k：$\mathcal{S} \to \Delta(\mathcal{A})$都是一个函数，将状态 s 映射为第 k个回合第 h 阶段动作空间$\mathcal{A}$ 上的分布。

学习协议。在每个回合 k 开始时，学习者选择一个策略 $\pi^k = \{\pi_h^k\}_{h=1}^H$。在每个阶段 $[H]$ s_1^k，从初始阶段开始，学习者观察状态 s_h^k，选择一个从 $\pi_h^k(s_h^k)$ 中采样的动作 a_h^k，获得奖励 $r_h^k\ s_h^k\ a_h^k$ $h \in [H]$ $s_{h+1}^k \sim (\ ,\ )$，并转移到下一状态 $(\ ,\ )$。当到达 时，该回合结束；此时不采取任何动作，奖励等于零。对于任意策略 $\pi = \{\pi_h^k\}_{h\in[H],\ k\in[K]}$ 以及 $(s,\ a) \in \mathcal{S}\times\mathcal{A}$，我们将动作价值函数 $Q^{k,\ }{}_h{}^\pi$ 和价值函数 $V^{k,\ }{}_h{}^\pi$ 定义为

$$Q_h^{k,\pi}(s,a) = \mathbb{E}\left[\sum_{h'=h}^{H} r_{h'}^k\left(s_{h'}^k, \pi_{h'}^k\left(s_{h'}^k\right)\right) \,\middle|\, s_h^k = s, a_h^k = a\right]$$
$$V_h^{k,\pi}(s) = \mathbb{E}_{a\sim\pi_h^k(\cdot\,|\,s)}\left[Q_h^{k,\pi}(s,a)\right].$$

我们定义 $\forall V\colon\ \mathcal{S} \to \mathbb{R}$, $[\mathbb{P}_h^k V](s,\ a) = \mathbb{E}_{s'\sim\mathbb{P}_h^k(\cdot\,|\,s,\ a)} V(s')$, 策略 π对应的贝尔曼方程为

$$Q_h^{k,\pi}(s,a) = r_h^k(s,a) + \left[\mathbb{P}_h^k V_{h+1}^{k,\pi}\right](s,a)$$
$$V_h^{k,\pi}(s) = \mathbb{E}_{a\sim\pi_h^k(\cdot\,|\,s)}\left[Q_h^{k,\pi}(s,a)\right], \quad V_{H+1}^{k,\pi} = 0.$$

学习者的目标是最小化以下动态遗憾，

$$\text{D-REG}_T = \sum_{k=1}^{K} V_1^{k,\pi_*^k}\left(s_1^k\right) - \sum_{k=1}^{K} V_1^{k,\pi^k}\left(s_1^k\right), \tag{21}$$

其中，为了与多臂老虎机的记号保持一致，我们记 $T \triangleq H\cdot K$。动态遗憾衡量了学习者的策略与最优策略之间的差异，即 $\pi_*^k = \arg\max_\pi V^{k,\ }{}_1{}^\pi(s_1^k)$。

非平稳线性混合马尔可夫决策过程。如果一个马尔可夫决策过程实例$M = (\mathcal{S},\ \mathcal{A},\ H,\ \mathbb{P},\ r)$ 是线性混合马尔可夫决策过程，则存在已知的特征映射 $\phi\colon\ \mathcal{S}\times\mathcal{A}\to\mathbb{R}^d$和 $\psi\colon\ \mathcal{S}\times\mathcal{A}\times\mathcal{S}\to\mathbb{R}^d$，使得对于任意 $k\in[K],\ h\in[H]$，存在未知向量$\theta_h^k \in \mathbb{R}^d$ 和 $\mathbf{w}_h^k \in \mathbb{R}^d$，满足

$$r_h^k(s,a) = \left\langle \phi(s,a), \theta_h^k \right\rangle$$
$$\mathbb{P}_h^k(s' \mid s,a) = \left\langle \psi(s' \mid s,a), \mathbf{w}_h^k \right\rangle, \tag{22}$$

这里我们考虑漂移情形，其中使用路径长度 $P_T^\theta = \sum_{k=2}^K\sum_{h=1}^H \left\|\theta_h^{k-1} - \theta_h^k\right\|_2$来度量 θ_h^k和 $P_T^{\mathbf{w}} = \sum_{k=2}^K\sum_{h=1}^H \left\|\mathbf{w}_h^{k-1} - \mathbf{w}_h^k\right\|_2$的非平稳性，并使用 $\mathbf{w}_h^k$来度量其非平稳性，同时定义 $\Delta = P_T^\theta + P^{\mathbf{w}T}$为总路径长度。我们采用以下标准有界性假设 [18]。

假设4. 假定可行集和未知参数有界：$\forall s\in\mathcal{S}, a\in\mathcal{A}, \|\phi(s,a)\| \le L_\phi$; 对于任意有界函数 $V\colon\ \mathcal{S}\to[0,1]$ 和 $\forall s\in\mathcal{S}, a\in\mathcal{A}$, $\|\psi_V(s,a)\|_2 \le L_\psi$，其中 $\psi_V(s,a) \triangleq \sum_{s'\in\mathcal{S}}\psi(s'|s,a)V(s')$; $\theta_h^k \in \Theta$ 对所有 $k\in[K], h\in[H]$均成立，其中 $\Theta \triangleq \{\theta \mid \|\theta\|_2 \le S_\theta\}$ 且 $\mathbf{w}_h^k \in \mathcal{W}$ 对所有 $k\in[K], h\in[H]$ 均成立，其中 $\mathcal{W} \triangleq \{\mathbf{w} \mid \|\mathbf{w}\|_2 \le S_{\mathbf{w}}\}$。

B. 算法与遗憾保证

本节提出 WeightUCRL 算法。我们首先介绍所采用的奖励估计器和转移估计器，然后通过改进的分析框架推导其估计误差上界，这是算法设计和遗憾分析的关键。基于估计误差界，我们提出选择准则，并最终给出其动态遗憾的理论保证。

奖励估计器。在第 k 个回合的第 h 阶段，我们使用以下加权最小二乘估计器 $\widehat{\theta}_h^k$ 来估计未知参数 θ_h^k，该参数是以下表达式的最小化解

$$\frac{\lambda_\theta}{2}\left\|\theta\right\|_2^2+\sum_{j=1}^{k-1} w_{k-1,j}\left(r_h^j(s_h^j,a_h^j)-\phi(s_h^j,a_h^j)^\top\theta\right)^2,$$

其中，$\lambda_\theta>0$ 为正则化系数，$w_{k-1,\,j}$ 为加权因子。与估计量（3）类似，我们设定 $w_{k\,j}=\gamma^{k-j}$，$\widehat{\theta}_h^k$ 其中（）为折扣因子。显然，$\gamma\in 0,1$ 其存在闭式解：

$$\widehat{\theta}_h^k=\left(\Lambda_h^{k-1}\right)^{-1}\left(\sum_{j=1}^{k-1} w_{k-1,j} r_h^j(s_h^j,a_h^j)\phi(s_h^j,a_h^j)\right),\quad(23)$$

其中，$\Lambda_h^k=\lambda_\theta I_d+\sum_{j=1}^k w_{k,\,j}\phi(s_h^j,\ a_h^j)\phi(s_h^j,\ a_h^j)^\top$。

转移估计器。首先注意到，根据模型假设（22），我们有

$$\begin{aligned}\left[\mathbb{P}_h^k V_{h+1}^k\right](s_h^k,a_h^k)&=\sum_{s'}\left\langle\psi(s'\mid s_h^k,a_h^k),\mathbf{w}_h^k\right\rangle V_{h+1}^k(s')\\&=\left\langle\sum_{s'}\psi(s'\mid s_h^k,a_h^k)V_{h+1}^k(s'),\mathbf{w}_h^k\right\rangle\\&=\left\langle\psi_{h+1}^k\left(s_h^k,a_h^k\right),\mathbf{w}_h^k\right\rangle,\end{aligned}\quad(24)$$

为简便起见，我们记 $\psi_{h+1}^k(s,\ a)\triangleq\psi_{V_{h+1}^k}(s,\ a)$。根据假设4，我们知道 $\forall s\in\mathcal{S},\ a\in\mathcal{A},\ \left\|\psi_{h+1}^k(s,\ a)\right\|_2\le HL_\psi$。我们采用以下加权最小二乘估计器 $\widehat{\ }\mathbf{w}_h^k$ 来估计未知参数$\mathbf{w}_h^k$，它是下式的最小化解

$$\sum_{j=1}^{k-1}\alpha_{k-1,j}\left(\langle\psi_{h+1}^j(s_h^j,a_h^j),\mathbf{w}\rangle-V_{h+1}^j(s_{h+1}^j)\right)^2+\frac{\lambda_{\mathbf{w}}}{2}\left\|\mathbf{w}\right\|_2^2,$$

其中 $\lambda_{\mathbf{w}}>0$ 是正则化系数，$\alpha_{k-1,\ j}$ 是加权因子，我们设 $\alpha_{k,\ j}=\gamma^{k-j}$，其中 $\gamma\in(0,\ 1)$ 是折扣因子。对于该最小二乘估计器，$\widehat{\ }\mathbf{w}_h^k$ 有闭式解：

$$\left(\Sigma_h^{k-1}\right)^{-1}\left(\sum_{j=1}^{k-1}\alpha_{k-1,j}V_{h+1}^j(s_{h+1}^j)\psi_{h+1}^j\left(s_h^j,a_h^j\right)\right),\quad(25)$$

其中 $\Sigma_h^k=\lambda_{\mathbf{w}}I_d+\sum_{j=1}^k\alpha_{k,\ j}\psi_{h+1}^j(s_h^j,\ a_h^j)\psi_{h+1}^j(s_h^j,\ a_h^j)^\top$。

上置信界。对于估计量 (23) 和 (25)，我们分别给出估计误差界。

引理4。对于任意 $s\in\mathcal{S}$、$a\in\mathcal{A}$，以下结论对所有 $k\in[K]$、$h\in[H]$ 均成立，

$$\left|\phi(s,a)^\top\left(\widehat{\theta}_h^k-\theta_h^k\right)\right|\le\Gamma_{h,\theta}^{k-1}+\beta_\theta\left\|\phi(s,a)\right\|_{\left(\Lambda_h^{k-1}\right)^{-1}},$$

其中 $\Gamma_{h,\theta}^{k-1}\triangleq L_\phi\sqrt{\frac{d}{\lambda_\theta}}\sum_{p=1}^{k-1}\sqrt{\sum_{j=1}^p w_{k-1,\,j}}\left\|\theta_h^p-\theta_h^{p+1}\right\|_2$、$\beta_\theta\triangleq\sqrt{\lambda_\theta}S_\theta$。

引理5。对于任意 $s\in\mathcal{S}$、$a\in\mathcal{A}$ 和 $\delta\in(0,\ 1)$，以至少 $1-\delta$ 的概率，以下结论对所有 $k\in[K]$、$h\in[H]$ 均成立

$$\left|\psi_{h+1}^k(s,a)^\top\left(\widehat{\mathbf{w}}_h^k-\mathbf{w}_h^k\right)\right|$$

算法 3 WeightUCRL

Require: episode number K, time horizon H, discounted factor γ, confidence δ, regularizer $\lambda_\theta,\lambda_{\mathbf{w}}$, parameters $S_\theta,S_{\mathbf{w}},L_\phi,L_\psi$

Initialize $\left\{\pi_h^0\right\}_{h=1}^H$ as uniform distribution policies, $\left\{Q_h^0\right\}_{h=1}^H$ as zero functions.
Set $\forall h\in[H],\Lambda_h^0=\lambda_\theta I_d,\Sigma_h^0=\lambda_{\mathbf{w}}I_d$ for $k=1$
for $k=1,2,...,K$ **do**
  Receive the initial state s_1^k
  Initialize V_{H+1}^k as zero function
  for $h=H,H-1,...,1$ **do**
    $\psi_{h+1}^k(\cdot,\cdot)=\sum_{s'}\psi(s'\mid\cdot,\cdot)V_{h+1}^k(s')$
    Update $\Lambda_h^k=\gamma\Lambda_h^{k-1}+\phi(s_h^k,a_h^k)\phi(s_h^k,a_h^k)^\top+(1-\gamma)\lambda_\theta I_d$
    Update $\Sigma_h^k=\psi_{h+1}^k\left(s_h^k,a_h^k\right)\psi_{h+1}^k\left(s_h^k,a_h^k\right)^\top+\gamma\Sigma_h^{k-1}+(1-\gamma)\lambda_{\mathbf{w}}I_d$
    Compute $\widehat{\theta}_h^k$ by (23) and $\widehat{\mathbf{w}}_h^k$ by (25)
    Compute optimistic value function $Q_h^k(s,a)$ and $V_h^k(s)$ by (26)
  end for
  for $h=1,...,H$ **do**
    Choose policy as $\pi_h^k(s)=\arg\max_{a\in\mathcal{A}}Q_h^k(s,a)$
    Take action $a_h^k\sim\pi_h^k(s_h^k)$, then observe the reward $r_h^k(s_h^k,a_h^k)$ and receive the next state s_{h+1}^k
  end for
end for

$$\le\Gamma_{h,\mathbf{w}}^{k-1}+\beta_{\mathbf{w}}^{k-1}\left\|\psi_{h+1}^k(s,a)\right\|_{\left(\Sigma_h^{k-1}\right)^{-1}},$$

$\Gamma_{h,\ \mathbf{w}}^{k-1}\triangleq H^2L_\psi^2\sqrt{\frac{d}{\lambda_{\mathbf{w}}}}\sum_{p=1}^{k-1}\sqrt{\sum_{j=1}^p\alpha_{k-1,\ j}}\left\|\mathbf{w}_h^p-\mathbf{w}_h^{p+1}\right\|_2$，且 $\beta_{\mathbf{w}}^k$ 是由以下定义的置信区域的半径

$$H\sqrt{\frac{1}{2}\log\frac{1}{\delta}+\frac{d}{4}\log\left(1+\frac{H^2L_\psi^2\sum_{j=1}^k\alpha_{k,j}}{\lambda_{\mathbf{w}}d}\right)}+\sqrt{\lambda_{\mathbf{w}}}S_{\mathbf{w}}.$$

引理4和引理5的证明见附录 E-A、E-B。

臂选择。根据引理4和引理5，我们定义如下乐观价值函数 $Q_h^k(s,a)$ 和 $V_h^k(s)$，

$$\begin{aligned}Q_h^k(s,a)\triangleq\min\Bigg\{&H,\phi(s,a)^\top\widehat{\theta}_h^k+\beta_\theta\left\|\phi(s,a)\right\|_{\left(\Lambda_h^{k-1}\right)^{-1}}\\&+\psi_{h+1}^k(s,a)^\top\widehat{\mathbf{w}}_h^k+\beta_{\mathbf{w}}^{k-1}\left\|\psi_{h+1}^k(s,a)\right\|_{\left(\Sigma_h^{k-1}\right)^{-1}}\Bigg\}\\V_h^k(s)\triangleq\max_{a\in\mathcal{A}}Q_h^k(s,a)&=\mathbb{E}_{a\sim\pi_h^k(\cdot\mid s)}\left[Q_h^k(s,a)\right].\end{aligned}\quad(26)$$

在状态 s_h^k 下，我们可以将策略 $\pi_h^k(s_h^k)$ 的动作选择准则指定为 $a_h^k=\arg\max_{a\in\mathcal{A}}Q_h^k(s_h^k,a)$。整体算法总结于算法 3。我们证明了该算法满足以下遗憾保证。

定理 5。设 $T=KH$, $\delta=1/(4T)$、$\lambda_\theta=d$ 和 $\lambda_{\mathbf{w}}=H^2d$, $\forall k,j\in[K]$, $w_{k,j}=\alpha_{k,j}=\gamma^{k-j}$, $\gamma\in(1/K,1)$，则

动态遗憾 D-REG$_T$ 以至少$1-1/T$ 的概率被界定为

$$\mathcal{O}\left(Hd\left(\frac{1}{(1-\gamma)^{3/2}}\Delta+HK\sqrt{1-\gamma}\right)+H^{3/2}d\sqrt{HK}\right),$$

此外，通过将折扣因子最优地设置为 $\gamma=1-$ ——— 最大 $\left\{1/K,\sqrt{\Delta/T}\right\}$，我们有

$$\text{D-REG}_T\leq\begin{cases}\widetilde{\mathcal{O}}\left(Hd\Delta^{1/4}T^{3/4}\right) & \textit{when } \Delta\geq H/K,\\ \widetilde{\mathcal{O}}\left(dH^{3/2}\sqrt{T}\right) & \textit{when } \Delta<H/K.\end{cases}$$

定理 5 的证明见附录 E-C。与先前工作 [18], 相比，我们的结果在非平稳环境下对动态遗憾实现了相同阶的遗憾保证。此外，在近似平稳设定下，我们的结果恢复了针对平稳环境所建立的理论保证 [43], [44]。

VI. 多项逻辑混合马尔可夫决策过程

本节将探讨另一类马尔可夫决策过程，即在非平稳设定下的多项逻辑（MNL）混合马尔可夫决策过程。我们介绍 MNL-WeightUCRL 算法，该算法将加权策略应用于非平稳 MNL 混合马尔可夫决策过程，并为非平稳 MNL 混合马尔可夫决策过程提供首个理论保证。

A. 问题设定

为了解决线性函数逼近无法保证分布有效性这一局限，近期提出了一类新的模型，称为 MNL 混合马尔可夫决策过程 [16], [17]。MNL 混合马尔可夫决策过程与线性混合马尔可夫决策过程具有相同的分幕式非平稳马尔可夫决策过程结构和学习协议，其目标是最小化动态遗憾 (21)。二者的关键区别在于对转移概率的建模假设。下面，我们给出 MNL 混合马尔可夫决策过程的形式化定义。

定义 *1*（可达状态）。对于任意 $(k,h,s,a)\in[K]\times[H]\times\mathcal{S}\times\mathcal{A}$，我们将“可达状态”定义为这样一组状态：在单次转移内，从第 k 个回合的阶段 h 采取动作 a 后，可从状态 s 到达这些状态，即 $\mathcal{S}_h^k(s,a)\triangleq\{s'\in\mathcal{S}\mid\mathbb{P}_h^k(s'\mid s,a)>0\}$。此外，我们定义 $S_h^k(s,a)\triangleq\left|\mathcal{S}_h^k(s,a)\right|$，并进一步将 $U\triangleq\max_{(k,h,s,a)}S_h^k(s,a)$ 定义为可达状态的最大数量。

非平稳 MNL 混合马尔可夫决策过程。$M=(\mathcal{S},\mathcal{A},H,\mathbb{P},r)$ 是一个 MNL 混合马尔可夫决策过程，当且仅当存在已知的特征映射$\phi\colon\mathcal{S}\times\mathcal{A}\to\mathbb{R}^d$ 和 $\psi\colon\mathcal{S}\times\mathcal{A}\times\mathcal{S}\to\mathbb{R}^d$，并且对于任意$k\in[K]$，$h\in[H]$，存在未知向量 $\theta_h^k\in\mathbb{R}^d$ 和$\mathbf{w}_h^k\in\mathbb{R}^d$，使得

$$\begin{aligned}r_h^k(s,a)&=\left\langle\phi(s,a),\theta_h^k\right\rangle\\ \mathbb{P}_h^k(s'\mid s,a)&=\frac{\exp\left(\psi(s'\mid s,a)^\top\mathbf{w}_h^k\right)}{\sum_{\widetilde{s}\in\mathcal{S}_h^k(s,a)}\exp\left(\psi(\widetilde{s}\mid s,a)^\top\mathbf{w}_h^k\right)}.\end{aligned}\tag{27}$$

我们采用 MNL 混合马尔可夫决策过程 [16], [17] 的以下标准假设。

假设5。假设可行集和未知参数均有界：$\forall s\in\mathcal{S}$，$a\in\mathcal{A}$，$\|\phi(s,a)\|\leq L_\phi$;$\forall s's\in\mathcal{S}$，$a\in\mathcal{A}$，$\|\psi(s'\mid s,a)\|\leq L_\psi$；$\theta_h^k\in\Theta$对所有 $k\in[K]$，$h\in[H]$ 均成立，其中 $\Theta\triangleq\{\theta\mid\|\theta\|_2\leq S_\theta\}$且 $\mathbf{w}_h^k\in\mathcal{W}$ 对所有 $k\in[K]$，$h\in[H]$ 均成立，其中$\mathcal{W}\triangleq\{\mathbf{w}\mid\|\mathbf{w}\|_2\leq S_\mathbf{w}\}$。

假设6。存在 $0<\kappa<1$，使得对于所有 $(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]$ 和 s'，$s''\in\mathcal{S}_h^k(s,a)$，均有 $\inf_{\mathbf{w}\in\mathcal{W}}p_{s,a}^{s'}(\mathbf{w})p_{s,a}^{s''}(\mathbf{w})\geq\kappa$。

B. 算法与遗憾保证

我们提出了 MNL-WeightUCRL 算法，它为非平稳 MNL 混合马尔可夫决策过程首次提供了动态遗憾保证。我们首先介绍该方法中使用的估计量，并通过改进的分析框架推导其估计误差上界；这是算法设计和遗憾分析的关键。在此基础上，我们提出选择准则，并最终给出其动态遗憾的理论保证。

奖励估计器。由于我们使用与式 (22) 相同的线性函数来建模奖励函数，因此这里仍使用估计量 (23) 来估计未知参数 θ_h^k。

转移估计器。对于回合 k 中的轨迹 $\{(s_h^k,a_h^k)\}_{h=1}^H$，我们定义变量：$y_h^k\in\{0,1\}^{S_h^k}$，其中$y_h^k(s')=\mathbb{1}(s_{h+1}^k=s')$，对于 $s'\in\mathcal{S}_h^k\triangleq\mathcal{S}_h^k(s_h^k,a_h^k)$ 和$S_h^k\triangleq\left|\mathcal{S}_h^k\right|$。此外，我们记 $p_h^k(\psi(s'\mid s,a)^\top\mathrm{w})=\exp(\psi(s'\mid s,a)^\top\mathrm{w})/\sum_{\widetilde{s}\in\mathcal{S}_h^k(s,a)}\exp(\psi(\widetilde{s}\mid s,a)^\top\mathrm{w})$。~~我们定义 $\bar{\psi}_h^k(s')\triangleq\psi(s'\mid s_h^k,a_h^k)$~~。那么，$y_h^k$是一个来自多项分布的样本：

$$\text{multinomial}\left(1,\left[p_h^k(\bar{\psi}_h^k(s_1)^\top\mathbf{w}),\ldots,p_h^k(\bar{\psi}_h^k(s_{N_h^k})^\top\mathbf{w})\right]\right).$$

我们使用以下加权最大似然估计量（最大似然估计）$\widehat{\ }\mathbf{w}_h^k$ 来估计未知参数$\mathbf{w}_h^k$，该参数是下式的最小化解

$$\frac{\lambda_\mathbf{w}\kappa}{2}\|\mathbf{w}\|_2^2+\sum_{j=1}^{k-1}\alpha_{k-1,j}\sum_{s'\in\mathcal{S}_h^j}-y_h^j(s')\log p_h^j(\bar{\psi}_h^j(s')^\top\mathbf{w}),$$

其中 $\lambda_\mathbf{w}>0$ 是正则化系数，$\alpha_{k-1,j}$ 是加权因子。我们设定$\alpha_{k,j}=\gamma^{k-j}$，其中 $\gamma\in(0,1)$是折扣因子。具体而言，估计量 $\widehat{\ }\mathbf{w}_h^k$ 是下列方程的解：

$$\sum_{j=1}^{k-1}\alpha_{k-1,j}\sum_{s'\in\mathcal{S}_h^j}\left(p_h^j(\bar{\psi}_h^j(s')^\top\mathbf{w})-y_h^j(s')\right)\bar{\psi}_h^j(s')+\lambda_\mathbf{w}\kappa\mathbf{w}=0.\tag{28}$$

鉴于 $\widehat{\ }\mathbf{w}_h^k$ 可能不属于可行集 $\mathcal{W}$，且 κ是定义在参数 $\mathbf{w}\in\mathcal{W}$上的，我们需要执行以下投影步骤

$$\widetilde{\mathbf{w}}_h^k=\arg\min_{\mathbf{w}\in\mathcal{W}}\left\|g_h^k(\widehat{\mathbf{w}}_h^k)-g_h^k(\mathbf{w})\right\|_{(\bar{\Sigma}_h^{k-1})^{-1}},\tag{29}$$

算法 4 MNL-WeightUCRL

输入： 回合编号 K、时间范围 H、折扣的因子 γ、置信度 δ、正则化项 λ_θ，$\lambda_{\mathbf{w}}$、参数 $S_\theta, S_{\mathbf{w}}, L_\phi, L_\psi$

1: 初始化 $\{\pi_h^0\}_{h=1}^H$ as 均匀的 分布 策略，$\{\bar{Q}_h^0\}_{h=1}^H$ 作为零函数。
2: **对于** $k = 1, 2, ..., K$ **do**
3: 接收初始状态 s_1^k
4: 将 $\bar{V}_{H+1}^k$ 初始化为零函数
5: 设置 $\forall h \in [H]$, $\Lambda_h^0 = \lambda_\theta I_d$, $\bar{\Sigma}_h^0 = \lambda_{\mathbf{w}} I_d$
6: **对于** $h = H, H-1, ..., 1$ **执行**
7: 更新 $\Lambda_h^k = \gamma \Lambda_h^{k-1} + \phi(s_h^k, a_h^k)\phi(s_h^k, a_h^k)^\top + (1-\gamma)\lambda_\theta I_d$
8: 更新 $\bar{\Sigma}_h^k = \sum_{s' \in \mathcal{S}_h^j} \bar{\psi}_h^j(s'|s_h^k, a_h^k)\, \bar{\psi}_h^j(s'|s_h^k, a_h^k)^\top + \gamma \bar{\Sigma}_h^{k-1} + (1-\gamma)\lambda_{\mathbf{w}} I_d$
9: 根据 (23) 计算 $\widehat{\theta}_h^k$，并根据 (28) 计算 $\widehat{\mathbf{w}}_h^k$
10: **如果** $\|\widehat{\mathbf{w}}_h^k\|_2 \le S_{\mathbf{w}}$，**则**
11: 让 $\widetilde{\mathbf{w}}_h^k = \widehat{\mathbf{w}}_h^k$
12: **else**
13: 进行投影，并根据 (29) 得到 $\widetilde{\mathbf{w}}_h^k$
14: **结束条件判断**
15: 计算乐观价值函数 $\bar{Q}_h^k(s,a)$ 和 $\bar{V}_h^k(s)$，根据式 (30)
16: **结束循环**
17: **对于** $h = 1, ..., H$，**执行**
18: 选择策略为 $\pi_h^k(s) = \arg\max_{a \in \mathcal{A}} \bar{Q}_h^k(s,a)$
19: 执行动作 $a^k{}_h \sim \pi_h^k(s_h^k)$，然后观测奖励 $r_h^k(s_h^k, a_h^k)$ 并接收下一状态 $s^k{}_{h+1}$
20: **结束循环**
21: **结束循环**

其中 $\bar{\Sigma}_h^k = \lambda_{\mathbf{w}} I_d + \sum_{j=1}^k \alpha_{k,\,j} \sum_{s' \in \mathcal{S}_h^j} \bar{\psi}_h^j(s') \bar{\psi}_h^j(s')^\top$ 和 $g_h^k(\mathbf{w})$ 定义为

$$g_h^k(\mathbf{w}) \triangleq \lambda_{\mathbf{w}} \kappa \mathbf{w} + \sum_{j=1}^{k-1} \alpha_{k-1,j} \sum_{s' \in \mathcal{S}_h^j} p_h^j(\bar{\psi}_h^j(s')^\top \mathbf{w}) \bar{\psi}_h^j(s').$$

上置信界。 估计量 (23) 可直接应用引理4构造上置信界。对于估计量 (28)，我们给出如下估计误差界。为简洁起见，对于任意函数 V，我们定义：$\mathcal{S} \to \mathbb{R}$，$\sum_{s' \in \mathcal{S}_h^k} p_h^k \psi s' \mid s a \quad {}_h^k V s' \quad \left[\widetilde{\mathbb{P}}_h^k V\right] s a$，$= (,) \sum_{s' \in \mathcal{S}_h^k} p_h^k \psi s' \mid s a \top {}_h^k V s' \quad (,) (,) \mathbf{w} ( )$，$(,) (,) \mathbf{w} ( )$。

引理6。 对于任意 $\mathbf{x} \in \mathcal{X}$，以及 $\delta \in (0, 1)$，$\forall k, j \in [K], \alpha_{k,\,j} \le 1$，以至少 $1-\delta$ 的概率，对所有 $k \in [K], h \in [H]$，以下结论成立

$$\left| \left[\widetilde{\mathbb{P}}_h^k V\right](s,a) - \left[\mathbb{P}_h^k V\right](s,a) \right| \le \frac{H}{\kappa} \left(\Gamma_{h,\mathbf{w}}^{k-1} + \bar{\beta}_{\mathbf{w}}^{k-1} \max_{s' \in \mathcal{S}_h^k} \|\psi(s' \mid s,a)\|_{(\bar{\Sigma}_h^{k-1})^{-1}} \right),$$

$\Gamma_{h,\mathbf{w}}^{k-1} \triangleq L_\psi^2 \sqrt{\frac{d}{\lambda_{\mathbf{w}}}} \sum_{p=1}^{k-1} \sqrt{\sum_{j=1}^p \alpha_{k-1,j}} \left\| \mathbf{w}_h^p - \mathbf{w}_h^{p+1} \right\|_2$, *and* $\bar{\beta}_{\mathbf{w}}^k$ 是由以下定义的置信区域的半径

$$\sqrt{\frac{1}{2}\log\frac{1}{\delta} + \frac{d}{4}\log\left(1 + \frac{UL_\psi^2 \sum_{j=1}^k \alpha_{k,j}}{\lambda_{\mathbf{w}} d}\right)} + \sqrt{\lambda_{\mathbf{w}}}\kappa S_{\mathbf{w}}.$$

引理6的证明见附录 F-A。

动作选择。 根据引理4和引理6，我们构造乐观价值函数 $\bar{Q}_h^k(s,a)$、$\bar{V}_h^k(s)$，以及如下动作选择准则，

$$\begin{aligned}
\bar{Q}_h^k(s,a) &= \min\Bigg\{ H, \phi(s,a)^\top \widehat{\theta}_h^k + \beta_\theta \|\phi(s,a)\|_{(\Lambda_h^{k-1})^{-1}} \\
&\quad + [\widetilde{\mathbb{P}}_h^k \bar{V}_{h+1}^k](s,a) + \frac{H}{\kappa} \bar{\beta}_{\mathbf{w}}^{k-1} \max_{s' \in \mathcal{S}_h^k} \|\bar{\psi}_h^k(s')\|_{(\bar{\Sigma}_h^{k-1})^{-1}} \Bigg\} \\
\bar{V}_h^k(s) &= \max_{a \in \mathcal{A}} \bar{Q}_h^k(s,a) = \mathbb{E}_{a \sim \pi_h^k(\cdot \mid s)} \left[\bar{Q}_h^k(s,a)\right] \\
\pi_h^k(s) &= \arg\max_{a \in \mathcal{A}} \bar{Q}_h^k(s,a).
\end{aligned} \tag{30}$$

整体算法总结于算法 4。我们证明了该算法具有如下遗憾保证。

定理6。 设 $\delta = 1/(4T)$、$\lambda_\theta = d$ 和 $\lambda_{\mathbf{w}} = d$, $\forall k, j \in [K], w_{k,\,j} = \alpha_{k,j} = \gamma^{k-j}$, $\gamma \in (1/K, 1)$，动态遗憾D-REG$_T$ 以至少 $1 - 1/T$ 的概率有界，具体为

$$\mathcal{O}\left(\frac{Hd}{\kappa} \left(\frac{1}{(1-\gamma)^{3/2}} \Delta + HK\sqrt{1-\gamma} \right) + H^{3/2} d \sqrt{HK} \right).$$

此外，通过将折扣因子最优地设为 $1 - \frac{\quad}{\quad}$ $\gamma =$ 最大值 $\left\{1/K, \sqrt{\Delta/T}\right\}$，我们有

$$\text{D-REG}_T \le \begin{cases} \widetilde{\mathcal{O}}\left(\kappa^{-1} H d \Delta^{1/4} T^{3/4}\right) & \textit{when } \Delta \ge H/K, \\ \widetilde{\mathcal{O}}\left(\kappa^{-1} d H^{3/2} \sqrt{T}\right) & \textit{when } \Delta < H/K. \end{cases}$$

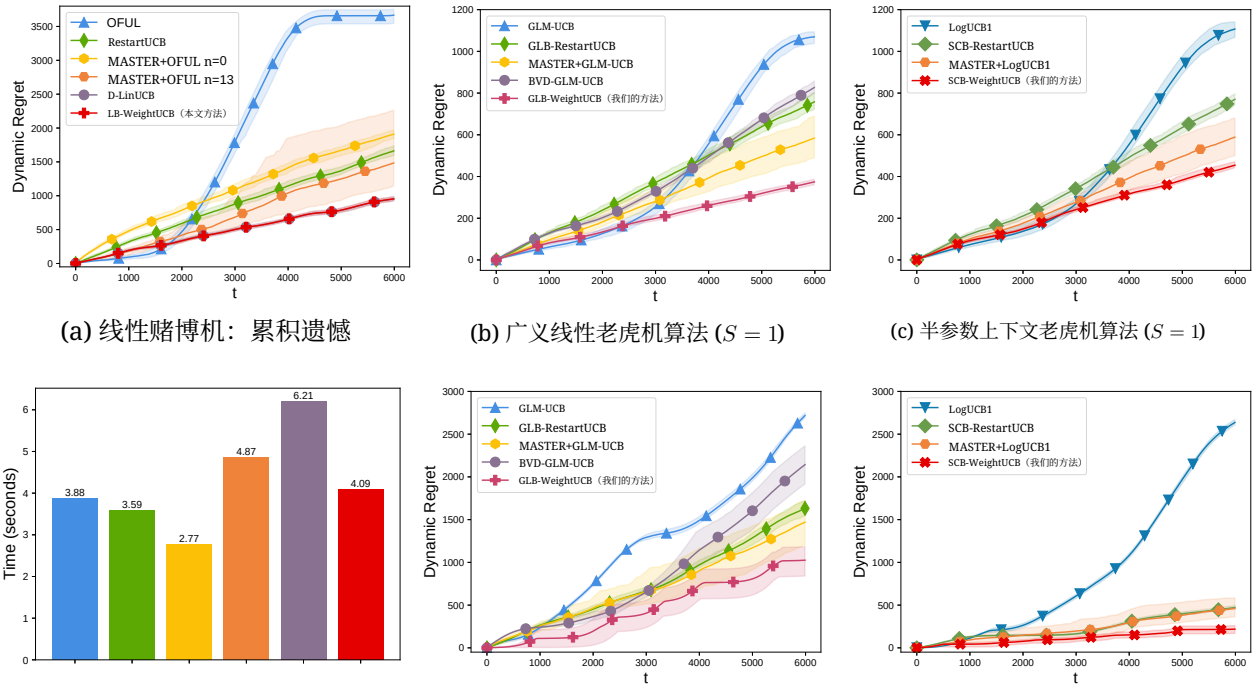
定理6的证明见附录 F-B。

VII. 实验

在本节中，我们进一步通过实证分析考察所提出算法的性能。我们分别针对漂移线性赌博机和 GLB 展示了两个合成实验。对于每个实验，我们将特征空间的维度设为 $d = 2$，轮次数设为 $T = 6000$，臂数设为 $n = 50$。每个臂的特征均从正态分布 $\mathcal{N}(0, 1)$中采样，随后重新缩放以满足 $L = 1$。我们将时变参数 θ_t 初始化为$[1, 0]$，并使其沿单位圆均匀地逆时针旋转，在 T 轮内从 0 旋转一整圈至 2π，最终返回起点 $[1, 0]$。

A. 线性赌博机

设置。 我们考虑线性模型 $r_t = X_t^\top \theta_t + \eta_t$，其中随机噪声 η_t 在每个时刻 t 独立地从正态分布 $\mathcal{N}(0, 1)$ 中抽取。我们将所提出的 LB-WeightUCB 算法与以下算法进行比较：(a) 静态算法 OFUL [32]；(b) 基于重启的算法 RestartUCB [3]；(c) 基于权重的算法 D-LinUCB [2]；以及 (d) 自适应重启算法



(a) 线性赌博机：累积遗憾 (b) 广义线性老虎机算法 ($S = 1$) (c) 半参数上下文老虎机算法 ($S = 1$)

(d) 线性赌博机：平均运行时间 (e) 广义线性老虎机算法 ($S = 5$) (f) 半参数上下文老虎机算法 ($S = 5$)

图 1：广义线性多臂老虎机实验。

MASTER+OFUL [8]。在这些算法中，RestartUCB 和 D-LinUCB 都需要关于 P_T 的先验知识，而 MASTER+OFUL 不需要。尽管 MASTER+OFUL 的适用条件较弱，我们仍将其纳入比较，因为它能以更快的收敛速率实现理论上最优的遗憾，因此是一个重要的基准。由于 P_T 可计算，我们为 LB-WeightUCB 和 D-LinUCB 设置折扣因子 $\gamma = 1 - \max\{1/T, \sqrt{P_T/(dT)}\}$，并将窗口大小 W 和重启周期H设为 $W = H = d^{1/4}\sqrt{T}/(1 + P_T)$。对于 MASTER，参数 n 表示多尺度探索参数的初始值（参见 [8]中的过程 1 的输入），而原始 MASTER 算法将其初始值设为 0（即$n = 0, 1, ...$）。然而，较小的 n 初始值会导致频繁重启，从而使性能变差。为此，我们尝试使用较大的 $n = 13$ 初始值，这在我们的实验中显著改善了性能。

结果。实验结果为 20 次独立试验的平均值。图 1a 展示了累积动态遗憾的表现，阴影区域表示实验结果的方差。图 1d 报告了每次运行的平均时间，每次运行包含 6000 轮。我们的 LB-WeightUCB 算法表现与 D-LinUCB 相当，但效率显著更高，加速比超过 1.5 倍。图 1a 还显示，经过精细调优的 nMASTER+OFUL（$n = 13$）表现优于 RestartUCB，而原版 MASTER+OFUL（$n = 0$）表现较差，这是因为其在初期重启过于频繁。然而，较大的初始值 n 会带来更高的时间开销，因为每次重启时，MASTER+OFUL 都需要执行一次过程 1，从而产生 $\mathcal{O}(n2^n)$ 的时间复杂度。更重要的是，在缓慢演变的环境中，无论是自适应重启（MASTER+OFUL）还是周期性重启（RestartUCB），表现都不如我们的加权策略。

B. 广义线性老虎机

设置。在 GLB 实验中，我们采用逻辑回归模型，即奖励满足 $r_t \sim$ 伯努利分布($\mu(X_t^\top \theta_t)$)，其中逻辑函数为 $\mu(x) = (1 + e^{-x})^{-1}$。我们分别考虑 $S = 1$ 和 $S = 5$两种情况。我们将提出的 GLB-WeightUCB 和 SCB-WeightUCB 算法与以下算法进行性能比较：(a) GLM-UCB，适用于 GLB 的静态算法 [41]；(b) LogUCB1，适用于 LogB 的静态算法 [33]；(c) BVD-GLM-UCB，适用于 GLB 的基于权重的算法 [6]；(d) GLB-RestartUCB，适用于 GLB 的重启算法 [3]；(e) SCB-RestartUCB，适用于 SCB 的重启算法 [3]；(f) MASTER+GLM-UCB，适用于 GLB 的自适应重启算法 [8]；以及 (g) MASTER+LogUCB1，适用于 LogB 的自适应重启算法 [8]。我们为 GLB-WeightUCB 设置折扣因子$\gamma = 1 - \max\{1/T, \sqrt{c_\mu P_T/(dT)}\}$，为 BVD-GLM-UCB 设置$\gamma = 1 - (P_T/(\sqrt{dT}))^{2/5}$，并为 SCB-WeightUCB 设置$\gamma = 1 - \max\{1/T, \sqrt{P_T/(dT)}\}$。我们为 GLB-RestartUCB 和 SCB-RestartUCB 设置相同的重启周期 $H = d^{1/4}\sqrt{T}/(1 + P_T)$。我们为 GLM-UCB、BVD-GLM-UCB、GLB-RestartUCB 和 MASTER+GLM-UCB 设置正则化项$\lambda = d$，为 GLB-WeightUCB 设置 $\lambda = d/c_\mu^2$

为 LogUCB1、SCB-RestartUCB、MASTER+LogUCB1 和 SCB-WeightUCB 设置 $\lambda = d\log T/c_\mu$。请注意，对于 LogB，$k_\mu = 1/4 < 1$，因此我们无需控制 k_μ 的阶数。对于两个 MASTER 算法，我们设置 $n = 13$。

结果。我们在图 1 中展示了 20 次独立试验的实验平均累积动态遗憾结果。当 S 较小时（$S = 1$，$c_\mu^{-1} \approx 5$），所有基于权重的算法都优于静态算法，且我们的 GLB-WeightUCB 和 SCB-WeightUCB 优于 BVD-GLM-UCB。当 S 较大时（$S = 5$，$c_\mu^{-1} \approx 152$），SCB-WeightUCB 显著优于 GLB-WeightUCB，这表明考虑自协调性质至关重要（请回想，LogB 是 SCB 的一个实例）。相较之下，BVD-GLM-UCB 的性能大幅下降，因为它没有考虑 c_μ^{-1}问题。与线性赌博机类似，GLB 的实验结果也表明，在缓慢演变的环境中，加权策略在经验上优于（自适应）重启策略。具体而言，我们观察到，GLB-WeightUCB 始终优于 MASTER+GLM-UCB，而 SCB-WeightUCB 始终优于 MASTER+LogUCB1。

VIII. 结论

本文重新审视了三种非平稳参数化老虎机模型（LB、GLB、SCB）和两种非平稳马尔可夫决策过程设置（线性混合马尔可夫决策过程、MNL 混合马尔可夫决策过程）中的基于权重的算法。我们发现，先前工作的不足源于对估计误差分析不够充分。因此，我们提出了一种改进的分析框架，使估计误差分析中的偏差部分和方差部分能够使用相同的局部范数。该框架使三种老虎机模型和两种强化学习模型的算法更加高效，改进了 GLB 和 SCB 设置下的遗憾界，并首次建立了 MNL 混合马尔可夫决策过程的动态遗憾界。

我们工作的意义在于，如今，对于非平稳参数化老虎机和马尔可夫决策过程，基于权重的算法在计算效率和遗憾保证两方面都已能与重启类算法相媲美。需要注意的是，目前基于窗口、基于重启和基于权重的算法只能达到 $\widetilde{\mathcal{O}}(P_T^{1/4}T^{3/4})$ 的遗憾界，无法匹敌 MASTER 算法这一自适应重启策略 [8] 所达到的最优速率 $\widetilde{\mathcal{O}}(P_T^{1/3}T^{2/3})$。借鉴这一已知的最佳结果，鉴于加权策略在渐变环境中尤为有效，而这类环境在现实世界应用中十分常见，因此亟需设计自适应基于权重的算法，在无需预先了解环境非平稳性的情况下实现最优动态遗憾界。目前的 $\Omega(P_T^{1/3}T^{2/3})$ 下界是在固定臂集合假设 [40] 下建立的。MASTER 算法 [8] 在相同假设下达到了这一速率，因此，对于固定臂集情形，$\Theta(P_T^{1/3}T^{2/3})$ 是极小极大最优速率。然而，时变臂集下的极小极大速率仍未解决。

在这项工作中，我们采用 $P_T = \sum_{t=2}^{T} \|\theta_{t-1} - \theta_t\|_2$ 作为衡量渐变环境的指标。然而，这一指标可能无法足够精确地捕捉环境中的渐变，因为它也可能涵盖其他类型的变化量，例如突变和非静止变化 [49], [50]。这或许可以解释为何基于权重的算法在理论上并未表现出显著优势，却在渐变环境的实验中比基于重启的算法表现出色。为克服这一局限，未来研究可以探索对渐变进行更精细的刻画，借鉴索博列夫类或霍尔德类 [51]背后的思想，或采用其他信息论工具 [10]。

附录 A LB-WEIGHTUCB 分析

本节将分析 LB-WeightUCB 算法。在附录 A-A 中，我们回顾 [2] 提出的 D-LinUCB 算法，并重述其估计误差分析。在附录 A-B 中，我们针对所提出的 LB-WeightUCB 算法给出自己的估计误差分析，该分析由引理 1 概述。最后，在附录 A-C 中，我们将分析动态遗憾，如定理1所述。

A. 回顾 *D-LinUCB* 算法的估计误差分析

在这一部分，我们回顾 D-LinUCB 算法此前的估计误差分析。该算法使用的估计量与我们的估计量 (3) 相同。第一步是将估计误差分解为偏差部分和方差部分，其中偏差部分表示参数漂移造成的误差，方差部分表示随机噪声造成的误差。基于奖励模型假设和估计量（与 (1) 和 (3) 相同），D-LinUCB 算法的估计误差可分解为

$$
\begin{aligned}
\widehat{\theta}_t - \theta_t &= V_{t-1}'^{-1}\left(\sum_{s=1}^{t-1}\gamma^{t-s-1} r_s X_s\right) - \theta_t \\
&= V_{t-1}'^{-1}\left(\sum_{s=1}^{t-1}\gamma^{t-s-1}\left(X_s^\top\theta_s + \eta_s\right)X_s\right) \\
&\quad - V_{t-1}'^{-1}\left(\lambda I_d + \sum_{s=1}^{t-1}\gamma^{t-s-1}X_s X_s^\top\right)\theta_t \\
&= V_{t-1}'^{-1}\left(\sum_{s=1}^{t-1}\gamma^{t-s-1}X_s X_s^\top\theta_s + \sum_{s=1}^{t-1}\gamma^{t-s-1}\eta_s X_s\right) \\
&\quad - V_{t-1}'^{-1}\left(\lambda I_d + \sum_{s=1}^{t-1}\gamma^{t-s-1}X_s X_s^\top\right)\theta_t \\
&= \underbrace{V_{t-1}'^{-1}\left(\sum_{s=1}^{t-1}\gamma^{t-s-1}X_s X_s^\top(\theta_s - \theta_t)\right)}_{\texttt{bias part}} \\
&\quad + \underbrace{V_{t-1}'^{-1}\left(\sum_{s=1}^{t-1}\gamma^{t-s-1}\eta_s X_s - \lambda\theta_t\right)}_{\texttt{variance part}},
\end{aligned} \tag{31}
$$

其中 $V'_t = \lambda I_d + \sum_{s=1}^{t-1} \gamma^{t-s} X_s X_s^\top$。随后， [2] 对偏差部分和方差部分分别使用不同的局部范数（我们稍后会解释使用不同局部范数的原因），如下所示

$$|\mathbf{x}^\top(\widehat{\theta}_t - \theta_t)| \leq \|\mathbf{x}\|_2 A'_t + \|\mathbf{x}\|_{V'^{-1}_{t-1}\widetilde{V}_{t-1}V'^{-1}_{t-1}} B'_t, \tag{32}$$

其中 $\widetilde{V}_t = \lambda I_d + \sum_{s=1}^{t} \gamma^{2(t-s)} X_s X_s^\top$和

$$A'_t = \left\| V'^{-1}_{t-1} \sum_{s=1}^{t-1} \gamma^{t-s-1} X_s X_s^\top (\theta_s - \theta_t) \right\|_2$$
$$B'_t = \left\| \sum_{s=1}^{t-1} \gamma^{t-s-1} \eta_s X_s - \lambda\theta_t \right\|_{\widetilde{V}^{-1}_{t-1}}.$$

对于偏差部分， [2] 通过引入虚拟窗口大小 D，将其按照时间线划分为两部分，

$$A'_t \leq \underbrace{\left\| \sum_{s=t-D}^{t-1} V'^{-1}_{t-1} \gamma^{t-s-1} X_s X_s^\top (\theta_s - \theta_t) \right\|_2}_{\texttt{virtual window}} + \underbrace{\left\| \sum_{s=1}^{t-D-1} V'^{-1}_{t-1} \gamma^{t-s-1} X_s X_s^\top (\theta_s - \theta_t) \right\|_2}_{\texttt{small term}},$$

第一项可以视为一个虚拟窗口，其中包含在时间 $t-D$ 之后获得的最新数据；由于它对应于窗口策略估计误差中的偏差部分，因此可以直接通过 SW-UCB 的分析 [1] 进行分析，这也是他们在偏差部分使用 l_2-范数的原因。第二项反映了时间 $t-D$之前获得的过时数据所产生的影响。由于当 $s \leq t-D-1$时， γ^{t-s-1}会非常小，因此这一小项会被第一个虚拟窗口项所支配，这意味着偏差部分实际上由虚拟窗口大小 D所控制。

对于方差部分， [2]将先前的自归一化集中不等式 [32, 定理 1] 扩展到加权~版本。该集中不等式要求使用 V_t 作为局部范数。为此， [2]将方差部分拆分为

$$\left| \mathbf{x}^\top V'^{-1}_{t-1} \left(\sum_{s=1}^{t-1} \gamma^{t-s-1} \eta_s X_s - \lambda\theta_t \right) \right| \leq \|\mathbf{x}\|_{V'^{-1}_{t-1}\widetilde{V}_{t-1}V'^{-1}_{t-1}} C'_t,$$

其中

$$\begin{aligned} C'_t &= \left\| V'^{-1}_{t-1} \left(\sum_{s=1}^{t-1} \gamma^{t-s-1} \eta_s X_s - \lambda\theta_t \right) \right\|_{V'^{-1}_{t-1}\widetilde{V}^{-1}_{t-1}V'^{-1}_{t-1}} \\ &= \left\| \sum_{s=1}^{t-1} \gamma^{t-s-1} \eta_s X_s - \lambda\theta_t \right\|_{\widetilde{V}^{-1}_{t-1}} \\ &\leq \left\| \sum_{s=1}^{t-1} \gamma^{t-s-1} \eta_s X_s \right\|_{\widetilde{V}^{-1}_{t-1}} + \sqrt{\lambda} S. \end{aligned}$$

那么，项 $\|\sum_{s=1}^{t-1} \gamma^{t-s-1} \eta_s X_s\|_{\widetilde{V}^{-1}_{t-1}}$ 可以通过加权版本的自归一化集中不等式来界定。最后，基于这一分析，D-LinUCB 需要采用以下动作选择准则，该准则仅取决于方差部分，因为偏差部分不包含 $\mathbf{x}$，

$$X_t = \arg\max_{\mathbf{x}\in\mathcal{X}} \left\{ \langle \mathbf{x}, \widehat{\theta}_t \rangle + \beta_{t-1} \|\mathbf{x}\|_{V'^{-1}_{t-1}\widetilde{V}_{t-1}V'^{-1}_{t-1}} \right\},$$

其中 β_{t-1} 是 B'_t 的上界，与式 (5) 相同。根据这一选择准则可以看出，D-LinUCB~在算法运行期间的第 t 轮需要维护两个协方差矩阵，即 V'_t和 V_t。

在下一节中，我们将给出估计误差上界的证明。我们的分析与 D-LinUCB 的分析之间的差异主要始于步骤 (32)，这是分析的关键步骤；我们的新分析框架使我们能够在偏差部分和方差部分使用相同的局部范数。

B. 引理 *1* 的证明

证明。沿用式 (31) 中相同的推导，LB-WeightUCB 算法的估计误差也可以分解为

$$\widehat{\theta}_t - \theta_t = \underbrace{V^{-1}_{t-1} \left(\sum_{s=1}^{t-1} w_{t-1,s} X_s X_s^\top (\theta_s - \theta_t) \right)}_{\texttt{bias part}} + \underbrace{V^{-1}_{t-1} \left(\sum_{s=1}^{t-1} w_{t-1,s} \eta_s X_s - \lambda\theta_t \right)}_{\texttt{variance part}}.$$

因此，根据柯西-施瓦茨不等式，对于任意 $\mathbf{x} \in \mathcal{X}$,

$$\left| \mathbf{x}^\top \left(\widehat{\theta}_t - \theta_t \right) \right| \leq \|\mathbf{x}\|_{V^{-1}_{t-1}} (A_t + B_t), \tag{33}$$

其中

$$A_t = \left\| \sum_{s=1}^{t-1} w_{t-1,s} X_s X_s^\top (\theta_s - \theta_t) \right\|_{V^{-1}_{t-1}}$$
$$B_t = \left\| \sum_{s=1}^{t-1} w_{t-1,s} \eta_s X_s - \lambda\theta_t \right\|_{V^{-1}_{t-1}}.$$

上述两项可以分别进行界定，如下列两个引理所述，

引理 7。 对于任意 $t \in [T]$, 有

$$\left\| \sum_{s=1}^{t-1} w_{t-1,s} X_s X_s^\top (\theta_s - \theta_t) \right\|_{V^{-1}_{t-1}} \leq L\sqrt{d} \sum_{p=1}^{t-1} \sqrt{\sum_{s=1}^{p} w_{t-1,s}} \|\theta_p - \theta_{p+1}\|_2.$$

引理 8。 对于任意 $\delta \in (0,1)$，以至少 $1-\delta$ 的概率，以下结论对所有 $t \in [T]$ 均成立，

$$\left\| \sum_{s=1}^{t-1} w_{t-1,s} \eta_s X_s - \lambda\theta_t \right\|_{V^{-1}_{t-1}} \leq \sqrt{\lambda} S + R\sqrt{2\log\frac{1}{\delta} + d\log\left(1 + \frac{L^2 \sum_{s=1}^{t-1} w_{t-1,s}}{\lambda d}\right)},$$

根据不等式 (33)、引理 7、引理 8、可行集的有界性假设，以及对于任意$\mathbf{x}$，$\|\mathbf{x}\|_{V_{t-1}^{-1}} \le \|\mathbf{x}\|_2/\sqrt{\lambda}$（因为 $V_{t-1} \succeq \lambda I_d$）对于任意 $\mathbf{x} \in \mathcal{X}$、$\gamma \in (0,1)$ 和 $\delta \in (0,1)$，以至少 $1-\delta$ 的概率，以下结论对所有 $t \in [T]$ 均成立，

$$
\begin{aligned}
&|\mathbf{x}^\top(\widehat{\theta}_t - \theta_t)| \\
&\le L^2\sqrt{\frac{d}{\lambda}}\sum_{p=1}^{t-1}\sqrt{\sum_{s=1}^{p} w_{t-1,s}}\,\|\theta_p-\theta_{p+1}\|_2 + \beta_{t-1}\|\mathbf{x}\|_{V_{t-1}^{-1}},
\end{aligned}
$$

其中，$\beta_t \triangleq \sqrt{\lambda}S + R\sqrt{2\log\frac{1}{\delta} + d\log\left(1+\frac{L^2\sum_{s=1}^{t} w_{t-1,s}}{d\lambda}\right)}$ 是 LB-WeightUCB 中使用的置信半径。因此，证明完毕。 □

引理 *7* 的证明。第一步是提取参数 θ_t 的变化量，如下所示，

$$
\begin{aligned}
&\left\|\sum_{s=1}^{t-1} w_{t-1,s}X_sX_s^\top(\theta_s-\theta_t)\right\|_{V_{t-1}^{-1}} \\
&= \left\|\sum_{s=1}^{t-1} w_{t-1,s}X_sX_s^\top\sum_{p=s}^{t-1}(\theta_p-\theta_{p+1})\right\|_{V_{t-1}^{-1}} \\
&= \left\|\sum_{p=1}^{t-1}\sum_{s=1}^{p} w_{t-1,s}X_sX_s^\top(\theta_p-\theta_{p+1})\right\|_{V_{t-1}^{-1}} \\
&\le \sum_{p=1}^{t-1}\left\|\sum_{s=1}^{p} w_{t-1,s}X_s\|X_s\|_2\|\theta_p-\theta_{p+1}\|_2\right\|_{V_{t-1}^{-1}} \\
&\le L\sum_{p=1}^{t-1}\sum_{s=1}^{p} w_{t-1,s}\,\|X_s\|_{V_{t-1}^{-1}}\,\|\theta_p-\theta_{p+1}\|_2,
\end{aligned}
$$

并且项 $\sum_{s=1}^{p} w_{t-1,s}\,\|X_s\|_{V_{t-1}^{-1}}$ 还可以进一步推导为关于折扣因子 γ 的表达式，如下所示，

$$
\begin{aligned}
\sum_{s=1}^{p} w_{t-1,s}\,\|X_s\|_{V_{t-1}^{-1}} &\le \sqrt{\sum_{s=1}^{p} w_{t-1,s}}\sqrt{\sum_{s=1}^{p} w_{t-1,s}\,\|X_s\|^2_{V_{t-1}^{-1}}} \\
&\le \sqrt{d}\sqrt{\sum_{s=1}^{p} w_{t-1,s}}. \qquad (34)
\end{aligned}
$$

上述倒数第二步由柯西-施瓦茨不等式得出。此外，最后一步是令$A_s = \sqrt{w_{t-1,s}}X_s$ 和 $U_{t-1} = V_{t-1}$，并应用引理25得出的。因此，证明完毕。 □

引理 *8* 的证明。

$$
\begin{aligned}
&\left\|\sum_{s=1}^{t-1} w_{t-1,s}\eta_sX_s - \lambda\theta_t\right\|_{V_{t-1}^{-1}} \\
&\le \left\|\sum_{s=1}^{t-1} w_{t-1,s}\eta_sX_s\right\|_{V_{t-1}^{-1}} + \sqrt{\lambda}S.
\end{aligned}
$$

我们定义 $\widetilde{\eta}_s \triangleq \sqrt{w_{t-1,s}}\eta_s$ 和 $\widetilde{X}_s \triangleq \sqrt{w_{t-1,s}}X_s$，并注意到 $\forall t \in [T], s \in [t-1], |w_{t-1,s}| \le 1$，那么 $\widetilde{\eta}_s$ 仍然是 R-次高斯的，因此根据定理7，我们有

$$
\begin{aligned}
&\left\|\sum_{s=1}^{t-1} w_{t-1,s}\eta_sX_s\right\|_{V_{t-1}^{-1}} \\
&\le \sqrt{2R^2\log\left(\frac{\det(V_{t-1})^{\frac{1}{2}}\det(V_0)^{-\frac{1}{2}}}{\delta}\right)}.
\end{aligned}
$$

然后，根据引理 26 和 $\det(V_0) = \lambda^d$，我们有

$$
\begin{aligned}
&\left\|\sum_{s=1}^{t-1} w_{t-1,s}\eta_sX_s\right\|_{V_{t-1}^{-1}} \\
&\le R\sqrt{2\log\frac{1}{\delta} + d\log\left(1+\frac{L^2\sum_{s=1}^{t-1} w_{t-1,s}}{d\lambda}\right)}.
\end{aligned}
$$

至此，证明完成。 □

C. 定理*1*的证明

证明。令 $X_t^* \triangleq \arg\max_{\mathbf{x}\in\mathcal{X}} \mathbf{x}^\top\theta_t$。根据引理 1 以及 X_t^*，$X_t \in \mathcal{X}$ 这一事实，以下每一项以至少 $1-\delta$ 的概率成立，

$$
\begin{aligned}
\forall t \in [T], X_t^{*\top}\theta_t &\le X_t^{*\top}\widehat{\theta}_t + \beta_{t-1}\|X_t^*\|_{V_{t-1}^{-1}} \\
&\quad + L^2\sqrt{\frac{d}{\lambda}}\sum_{p=1}^{t-1}\sqrt{\sum_{s=1}^{p} w_{t-1,s}}\,\|\theta_p-\theta_{p+1}\|_2, \\
\forall t \in [T], X_t^{\top}\theta_t &\ge X_t^{\top}\widehat{\theta}_t - \beta_{t-1}\|X_t\|_{V_{t-1}^{-1}} \\
&\quad - L^2\sqrt{\frac{d}{\lambda}}\sum_{p=1}^{t-1}\sqrt{\sum_{s=1}^{p} w_{t-1,s}}\,\|\theta_p-\theta_{p+1}\|_2.
\end{aligned}
$$

根据并集界，以下结论以至少 $1-2\delta$, $\forall t \in [T]$ 的概率成立，

$$
\begin{aligned}
&X_t^{*\top}\theta_t - X_t^\top\theta_t \\
&\le X_t^{*\top}\widehat{\theta}_t - X_t^\top\widehat{\theta}_t + 2L^2\sqrt{\frac{d}{\lambda}}\sum_{p=1}^{t-1}\sqrt{\sum_{s=1}^{p} w_{t-1,s}}\,\|\theta_p-\theta_{p+1}\|_2 \\
&\quad + \beta_{t-1}(\|X_t^*\|_{V_{t-1}^{-1}} + \|X_t\|_{V_{t-1}^{-1}}) \\
&\le 2L^2\sqrt{\frac{d}{\lambda}}\sum_{p=1}^{t-1}\sqrt{\sum_{s=1}^{p} w_{t-1,s}}\,\|\theta_p-\theta_{p+1}\|_2 + 2\beta_{t-1}\|X_t\|_{V_{t-1}^{-1}},
\end{aligned}
$$

其中，最后一步根据臂选择准则 (6) 得出，因此

$$
X_t^{*\top}\widehat{\theta}_t + \beta_{t-1}\|X_t^*\|_{V_{t-1}^{-1}} \le X_t^{\top}\widehat{\theta}_t + \beta_{t-1}\|X_t\|_{V_{t-1}^{-1}}.
$$

因此，以至少 $1-2\delta$ 的概率，以下动态遗憾界成立，并且可分为两部分，

$$\begin{aligned}\text{D-REG}_T &= \sum_{t=1}^{T}\left(X_t^{*\top}\theta_t - X_t^{\top}\theta_t\right)\\ &\le \underbrace{2L^2\sqrt{\frac{d}{\lambda}}\sum_{t=1}^{T}\sum_{p=1}^{t-1}\sqrt{\sum_{s=1}^{p} w_{t-1,s}}\left\|\theta_p-\theta_{p+1}\right\|_2}_{\texttt{bias part}}\\ &\quad+\underbrace{2\beta_T\sum_{t=1}^{T}\|X_t\|_{V_{t-1}^{-1}}}_{\texttt{variance part}},\end{aligned}$$

其中 $\beta_T = \sqrt{\lambda}S + R\sqrt{2\log\frac{1}{\delta}+d\log\left(1+\frac{L^2(1-\gamma^{2T})\lambda d}{(1-\gamma^2)}\right)}$ 是置信半径。

现在，我们分别推导偏差部分和方差部分的上界。

偏差部分。注意到 $w_{t-1,s}=\gamma^{t-s-1}$，其中 $\gamma\in(1/T,1)$。对于偏差部分，我们需要提取路径长度 P_T，并说明折扣因子 γ 对 P_T的控制。

$$\begin{aligned}&2L^2\sqrt{\frac{d}{\lambda}}\sum_{t=1}^{T}\sum_{p=1}^{t-1}\sqrt{\sum_{s=1}^{p} w_{t-1,s}}\left\|\theta_p-\theta_{p+1}\right\|_2\\ &=2L^2\sqrt{\frac{d}{\lambda}}\sum_{p=1}^{T-1}\sum_{t=p+1}^{T}\sqrt{\sum_{s=1}^{p} w_{t-1,s}}\left\|\theta_p-\theta_{p+1}\right\|_2\\ &=2L^2\sqrt{\frac{d}{\lambda}}\sum_{p=1}^{T-1}\sum_{t=p+1}^{T}\gamma^{\frac{t-1}{2}}\sqrt{\sum_{s=1}^{p}\gamma^{-s}}\left\|\theta_p-\theta_{p+1}\right\|_2\\ &=2L^2\sqrt{\frac{d}{\lambda}}\sum_{p=1}^{T-1}\frac{\gamma^{\frac{p}{2}}-\gamma^{\frac{T}{2}}}{1-\gamma^{\frac{1}{2}}}\sqrt{\frac{\gamma^{-p}-1}{1-\gamma}}\left\|\theta_p-\theta_{p+1}\right\|_2\\ &\le 2L^2\sqrt{\frac{d}{\lambda}}\sum_{p=1}^{T-1}\frac{\gamma^{\frac{p}{2}}\gamma^{-\frac{p}{2}}}{(1-\gamma^{\frac{1}{2}})\frac{1+\gamma^{\frac{1}{2}}}{2}\sqrt{1-\gamma}}\left\|\theta_p-\theta_{p+1}\right\|_2\\ &\le 4L^2\sqrt{\frac{d}{\lambda}}\frac{1}{(1-\gamma)^{3/2}}P_T.\end{aligned}$$

因此，对于偏差部分，我们有

$$\begin{aligned}&2L^2\sqrt{\frac{d}{\lambda}}\sum_{t=1}^{T}\sum_{p=1}^{t-1}\sqrt{\sum_{s=1}^{p} w_{t-1,s}}\left\|\theta_p-\theta_{p+1}\right\|_2\\ &\le 4L^2\sqrt{\frac{d}{\lambda}}\frac{1}{(1-\gamma)^{3/2}}P_T.\end{aligned}\tag{35}$$

方差部分。首先，利用柯西-施瓦茨不等式，我们有 $2\beta_T\sum_{t=1}^{T}\|X_t\|_{V_{t-1}^{-1}} \le 2\beta_T\sqrt{T\sum_{t=1}^{T}\|X_t\|^2_{V_{t-1}^{-1}}}$. 接着根据引理 9（加权势能引理），取 $w_{t,\ t}=\gamma^{t-t}=1$，$c=w_{t,\ s}/w_{t,\ s-1}=\gamma$，可得以下上界：

$$\begin{aligned}2\beta_T\sum_{t=1}^{T}\|X_t\|_{V_{t-1}^{-1}} &\le 2\beta_T\sqrt{2\max\{1,\frac{L^2}{\lambda}\}dT}\\ &\cdot\sqrt{T\log\frac{1}{\gamma}+\log\left(1+\frac{L^2\sum_{s=1}^{T}w_{T,s}}{d\lambda}\right)}.\end{aligned}\tag{36}$$

注意到，加权线性赌博机与标准线性赌博机在分析上的主要区别在于需要进行路径长度分析，以及需要使用加权势能引理。此外，我们有 $\sum_{s=1}^{t}w_{t,s}=\frac{1-\gamma^t}{1-\gamma}\le\frac{1}{1-\gamma}$。结合偏差部分和方差部分的上界，并取置信水平$\delta=1/(2T)$，根据并集界，我们以至少 $1-1/T$的概率得到以下动态遗憾界，

$$\begin{aligned}&\text{D-REG}_T \le 4L^2\sqrt{\frac{d}{\lambda}}\frac{1}{(1-\gamma)^{3/2}}P_T+2\beta_T\sqrt{2\max\left\{1,\frac{L^2}{\lambda}\right\}dT}\\ &\cdot\sqrt{T\log\frac{1}{\gamma}+\log\left(1+\frac{L^2}{\lambda d(1-\gamma)}\right)},\\ &\beta_T = \sqrt{\lambda}S+R\sqrt{2\log T+2\log 2+d\log\left(1+\frac{L^2(1-\gamma^T)}{\lambda d(1-\gamma)}\right)}.\end{aligned}$$

由于遗憾界包含一项 $T\sqrt{\log(1/\gamma)}$，我们不能令 γ 接近 0，因此设 $\gamma\ge 1/T$，并有 $\log(1/\gamma)\le C(1-\gamma)$，其中 $C=\log T/(1-1/T)$。随后，忽略时间范围 T中的对数因子，并令 $\lambda=d$，最终得到

$$\text{D-REG}_T \le \widetilde{\mathcal{O}}\left(\frac{1}{(1-\gamma)^{3/2}}P_T+d(1-\gamma)^{1/2}T\right).\tag{37}$$

当 $P_T<d/T$（对应于较小程度的非平稳性）时，我们直接令 $\gamma=1-1/T$，并得到一个 $\widetilde{\mathcal{O}}(d\sqrt{T})$ 的遗憾界。此外，对于非退化情形（$P_T\ge d/T$），我们将折扣因子最优地设为 $1-\gamma=\sqrt{P_T/(dT)}$，并得到一个 $\widetilde{\mathcal{O}}(d^{3/4}P_T^{1/4}T^{3/4})$ 的遗憾界，从而完成证明。 □

引理 9 (加权版本势引理)。假设 $V_t=\sum_{s=1}^{t}w_{t,\ s}X_sX_s^{\top}+\lambda I_d$，$V_0=\lambda I_d$，权重满足：$\forall t\in[T], s\in[t-1], w_{t,\ s}/w_{t-1,\ s}=c\le 1, \forall t,s\in[T], w_{t,\ s}\in(0,\ 1)\ w_{t,\ t}=1$ 且对所有$t\ge 1$，$\|X_t\|_2\le L$，则以下不等式成立，

$$\begin{aligned}&\sum_{t=1}^{T}\|w_{t,t}X_t\|^2_{V_{t-1}^{-1}}\\ &\le 2\max\left\{1,\frac{L^2}{\lambda}\right\}d\left(T\log\frac{1}{c}+\log\left(1+\frac{L^2\sum_{s=1}^{T}w_{T,s}}{d\lambda}\right)\right).\end{aligned}$$

引理 *9* 的证明。

$$\begin{aligned}V_t &= \sum_{s=1}^{t}w_{t,s}X_sX_s^{\top}+\lambda I_d\\ &= \sum_{s=1}^{t-1}w_{t,s}X_sX_s^{\top}+w_{t,t}X_tX_t^{\top}+\lambda I_d\end{aligned}$$

$$
\begin{aligned}
&= c\sum_{s=1}^{t-1} w_{t-1,s} X_s X_s^\top + X_t X_t^\top + \lambda I_d \qquad (w_{t,t}=1)\\
&\succeq c\left(\sum_{s=1}^{t-1} w_{t-1,s} X_s X_s^\top + X_t X_t^\top + \lambda I_d\right), \qquad (c<1)\\
&= c\left(V_{t-1} + X_t X_t^\top\right)\\
&= cV_{t-1}^{1/2}\left(I_d + V_{t-1}^{-1/2} X_t X_t^\top V_{t-1}^{-1/2}\right) V_{t-1}^{1/2},
\end{aligned}
$$

对等式两边取行列式，得到:

$$
\begin{aligned}
\det(V_t) &\geq \det(cV_{t-1}) \det\left(I_d + V_{t-1}^{-1/2} X_t X_t^\top V_{t-1}^{-1/2}\right)\\
\det(V_t) &\geq c^d \det(V_{t-1})\left(1 + \|X_t\|^2_{V_{t-1}^{-1}}\right)\\
\log\det(V_t) &\geq d\log c + \log\det(V_{t-1}) + \log\left(1 + \|X_t\|^2_{V_{t-1}^{-1}}\right)\\
d\log\frac{1}{c} + \log\frac{\det(V_t)}{det(V_{t-1})} &\geq \log\left(1 + \|X_t\|^2_{V_{t-1}^{-1}}\right)
\end{aligned}
$$

因此，从 1 求和至 T，并进行望远镜求和，可得:

$$
\begin{aligned}
& dT\log\frac{1}{c} + \log\left(\frac{\det(V_T)}{\det(V_0)}\right)\\
\geq & \sum_{t=1}^{T}\log\left(1 + \|X_t\|^2_{V_{t-1}^{-1}}\right)\\
\geq & \sum_{t=1}^{T}\log\left(1 + \frac{1}{\max\{1, L^2/\lambda\}}\|X_t\|^2_{V_{t-1}^{-1}}\right)\\
\geq & \frac{1}{2\max\{1, L^2/\lambda\}}\sum_{t=1}^{T}\|X_t\|^2_{V_{t-1}^{-1}}
\end{aligned}
$$

因此，我们有

$$
\begin{aligned}
&\sum_{t=1}^{T}\|X_t\|^2_{V_{t-1}^{-1}}\\
&\leq 2\max\left\{1, \frac{L^2}{\lambda}\right\}\cdot\left(dT\log\frac{1}{c} + \log\left(\frac{\det(V_T)}{\det(V_0)}\right)\right).
\end{aligned}
$$

最后，利用引理 26 以及 $\det(V_0) = \lambda^d$ 这一事实，即可完成证明。 □

附录 B GLB-WeightUCB 分析

本节对 GLB-WeightUCB 算法进行分析。在附录 B-A 中，我们回顾 GLB 的投影问题，并重述 [6] 中的 BVD-GLM-UCB 算法。在附录 B-B 中，我们给出 GLB-WeightUCB 算法估计误差上界的证明（即引理 2）。最后，在附录 B-C 中，我们给出定理 2 所述动态遗憾上界的证明。

A. 回顾 *BVD-GLM-UCB* 算法的投影步骤

如第 IV-B 节所述，GLB 的主要难点在于$\widehat{}$最大似然估计或准极大似然估计得到的估计量 θ_t 可能不属于可行集 Θ ，而c_μ 是定义在参数$\theta \in \Theta$上的。在平稳环境下，[41]通过引入如下投影步骤克服了这一难点:

$$
\widetilde{\theta}_t = \arg\min_{\theta\in\Theta}\|g_t(\widehat{\theta}_t) - g_t(\theta)\|_{V_{t-1}^{-1}}, \tag{38}
$$

其中，$V_t = \lambda I_d + \sum_{s=1}^{t} X_s X_s^\top$ $g_t\ \theta = \lambda c_\mu\theta +$以及 ( ) $\sum_{s=1}^{t-1}\mu\ X_s^\top\theta\ X_s$ $\gamma = 1$ ( ) 是静态版本（通过设置 得到）。根据准极大似然估计，我们知道

$$
g_t(\widehat{\theta}_t) = \lambda c_\mu\widehat{\theta}_t + \sum_{s=1}^{t-1}\mu(X_s^\top\widehat{\theta}_t)X_s = \sum_{s=1}^{t-1} r_s X_s, \tag{39}
$$

然后根据中值定理，我们知道

$$
g_t(\theta_1) - g_t(\theta_2) = G_t(\theta_1,\theta_2)(\theta_1-\theta_2), \tag{40}
$$

其中 $G_t(\theta_1,\theta_2) \triangleq \int_0^1\nabla g_t(s\theta_2+(1-s)\theta_1)\mathrm{d}s \in \mathbb{R}^{d\times d}$。注意，对于任意 $\theta\in\Theta$， g_t 的梯度满足

$$
\nabla g_t(\theta) = \lambda c_\mu I_d + \sum_{s=1}^{t-1}\mu'(X_s^\top\theta)X_s X_s^\top \succeq c_\mu V_{t-1},
$$

这显然意味着 $\forall\theta_1,\theta_2\in\Theta, G_t(\theta_1,\theta_2) \succeq c_\mu V_{t-1}$。通过这一投影步骤， [41] 可以分析如下形式的估计误差，

$$
\begin{aligned}
&|\mu(\mathbf{x}^\top\widetilde{\theta}_t) - \mu(\mathbf{x}^\top\theta_t)| \leq k_\mu|\mathbf{x}^\top(\widetilde{\theta}_t - \theta_t)|\\
&= k_\mu|\mathbf{x}^\top G_t^{-1}(\theta_t,\widetilde{\theta}_t)(g_t(\widetilde{\theta}_t) - g_t(\theta_t))|\\
&\leq k_\mu\|\mathbf{x}\|_{G_t^{-1}(\theta_t,\widetilde{\theta}_t)}\|g_t(\widetilde{\theta}_t) - g_t(\theta_t)\|_{G_t^{-1}(\theta_t,\widetilde{\theta}_t)}\\
&\leq \frac{k_\mu}{c_\mu}\|\mathbf{x}\|_{V_{t-1}^{-1}}\|g_t(\widetilde{\theta}_t) - g_t(\theta_t)\|_{V_{t-1}^{-1}}\\
&\leq \frac{2k_\mu}{c_\mu}\|\mathbf{x}\|_{V_{t-1}^{-1}}\|g_t(\widehat{\theta}_t) - g_t(\theta_t)\|_{V_{t-1}^{-1}},
\end{aligned}
$$

其中最后一步来自投影步骤。在$\widehat{}$执行投影步骤后，术语 $g_t(\theta_t) - g_t(\theta_t)$ 是未进行投影的最大似然估计的估计误差。注意，在分段平稳情形下， [7]也可以使用这一投影步骤。 [6] 认为，这两项先前研究之所以能够使用这一投影操作，主要是因为它们采用了平稳或分段平稳设定。他们提到，对于漂移情形，估计误差总是被分为偏差（跟踪误差）和方差（学习误差）部分，而这一简单的投影操作忽略了偏差部分；若要适应这两个偏差来源，就需要对其进行推广。在分析中，问题在于，投影步骤之后，估计误差$\widehat{}$项 $g_t(\theta_t) - g_t(\theta_t)$ 需要被分解为偏差部分和方差部分，并且 [6] 需要对偏差部分使用 l_2-范数，对方差部分使用 V_{t-1}^{-1} 。但整个估计误差已经由 V_{t-1}^{-1}进行范数化，这意味着他们无法对偏差部分使用先前对窗口策略的分析。

为此， [6] 针对漂移广义线性老虎机提出了 BVD-GLM-UCB 算法，如算法 5 所示，并设计了一个新的投影步骤来解决这一问题。具体而言，在每一轮 t，第一步是构造置信集 $\mathcal{E}_t^\delta(\theta)$，以表示随机噪声的影响。

$$
\mathcal{E}_t^\delta(\theta) := \left\{\theta'\in\mathbb{R}^d \,\middle|\, \|g_t(\theta') - g_t(\theta)\|_{\widetilde{V}_t^{-1}} \leq \bar{\beta}_t(\delta)\right\}. \tag{41}
$$

第二步是找到一个与可行集相交的置信集 $\mathcal{E}_t^\delta(\theta_t^p)\widehat{}$，而 θ_t^p 与 θ_t 之间的差距代表参数漂移的影响。

$$
\begin{aligned}
\theta_t^p \in & \arg\min_{\theta\in\mathbb{R}^d}\left\|g_t(\theta) - g_t(\widehat{\theta}_t)\right\|_{V_t^{-2}}\\
& \text{s.t } \Theta\cap\mathcal{E}_t^\delta(\theta) \neq \emptyset
\end{aligned} \tag{42}
$$

Algorithm 5 BVD-GLM-UCB [6]

Require: time horizon T, discounted factor γ, confidence δ, regularizer λ, inverse link function μ, parameters S, L and R

1: Set $V_0 = \lambda I_d$, $\widehat{\theta}_1 = \mathbf{0}$ and compute k_μ and c_μ
2: **for** $t = 1, 2, ..., T$ **do**
3: Solving $\theta_t^p \in \arg\min_{\theta\in\mathbb{R}^d} \left\{ \left\| g_t(\theta) - g_t(\widehat{\theta}_t) \right\|_{V_t^{-2}} \text{ s.t } \Theta \cap \mathcal{E}_t^\delta(\theta) \neq \emptyset \right\}$
4: Select $\widetilde{\theta}_t \in \Theta \cap \mathcal{E}_t^\delta(\theta_t^p)$ where $\mathcal{E}_t^\delta(\theta) := \left\{ \theta' \in \mathbb{R}^d \,\middle|\, \|g_t(\theta') - g_t(\theta)\|_{\widetilde{V}_t^{-1}} \leq \bar{\beta}_t(\delta) \right\}$
5: Get $\bar{\beta}_{t-1} = R\sqrt{2\log\frac{1}{\delta} + d\log\left(1 + \frac{L^2(1-\gamma^{2t-2})}{\lambda d(1-\gamma^2)}\right)} + \sqrt{\lambda}c_\mu S$
6: Get $X_t = \arg\max_{\mathbf{x}\in\mathcal{X}} \left\{ \mu(\mathbf{x}^\top\widetilde{\theta}_t) + \frac{2k_\mu}{c_\mu}\bar{\beta}_{t-1}\|\mathbf{x}\|_{V_{t-1}^{-1}} \right\}$
7: Receive the reward r_t
8: Update $V_t = \gamma V_{t-1} + X_t X_t^\top + (1-\gamma)\lambda I_d$, $\widetilde{V}_t = \gamma^2 V_{t-1} + X_t X_t^\top + (1-\gamma^2)\lambda I_d$
9: Compute $\widehat{\theta}_{t+1}$ by $\sum_{s=1}^t \gamma^{t-s}\left(\mu(X_s^\top\theta) - r_s\right)X_s + \lambda c_\mu\theta = 0$
10: **end for**

通过求解优化问题 (42) 得到解 θ_t^p 后，第三步是从 $\Theta \cap \mathcal{E}_t^\delta(\theta_t^p)$ 中选取 $\widetilde{\theta}_t$。基于这一步投影， [6] 可以将投影前的偏差部分和方差部分分离如下，

$$
\begin{aligned}
&|\mu(\mathbf{x}^\top\widetilde{\theta}_t) - \mu(\mathbf{x}^\top\theta_t)| \leq k_\mu|\mathbf{x}^\top(\widetilde{\theta}_t - \theta_t)| \\
&= k_\mu|\mathbf{x}^\top G_t^{-1}(\theta_t, \widetilde{\theta}_t)(g_t(\widetilde{\theta}_t) - g_t(\theta_t))| \\
&\leq k_\mu|\mathbf{x}^\top G_t^{-1}(\theta_t, \widetilde{\theta}_t)(g_t(\widetilde{\theta}_t) - g_t(\theta_t^p) + g_t(\theta_t^p) - g_t(\widehat{\theta}_t) \\
&\quad + g_t(\widehat{\theta}_t) - g_t(\bar{\theta}_t) + g_t(\bar{\theta}_t) - g_t(\theta_t))| \\
&\leq \underbrace{k_\mu|\mathbf{x}^\top G_t^{-1}(\theta_t, \widetilde{\theta}_t)(g_t(\widetilde{\theta}_t) - g_t(\theta_t^p) + g_t(\widehat{\theta}_t) - g_t(\bar{\theta}_t))|}_{\texttt{bias part}} \\
&\quad + \underbrace{k_\mu|\mathbf{x}^\top G_t^{-1}(\theta_t, \widetilde{\theta}_t)(g_t(\theta_t^p) - g_t(\widehat{\theta}_t) + g_t(\bar{\theta}_t) - g_t(\theta_t))|}_{\texttt{variance part}}.
\end{aligned}
$$

他们的偏差-方差分解促使他们在算法和分析中采用不同的局部范数来界定偏差部分和方差部分。值得注意的是，由于投影步骤的复杂性（见（41）和（42）），整体算法相当复杂，不太适合实际应用；此外，算法运行期间的每一轮都需要维护 ˜两个协方差矩阵 V_t 和 V_t （这是由于构造了置信区域（41））。在下一节中，我们将说明，通过改进的分析框架，平稳广义线性赌博机（38）中使用的简单投影足以应对漂移GLB。

B. 引理 *2* 的证明

证明。根据估计方程（11），我们知道

$$
\begin{aligned}
g_t(\widehat{\theta}_t) &= \lambda c_\mu\widehat{\theta}_t + \sum_{s=1}^{t-1} w_{t-1,s}\mu(X_s^\top\widehat{\theta}_t)X_s \\
&= \sum_{s=1}^{t-1} w_{t-1,s} r_s X_s,
\end{aligned} \tag{43}
$$

a然后根据中值定理，我们知道 此时

$$
g_t(\theta_1) - g_t(\theta_2) = G_t(\theta_1, \theta_2)(\theta_1 - \theta_2), \tag{44}
$$

其中 $G_t(\theta_1, \theta_2) \triangleq \int_0^1 \nabla g_t(s\theta_2 + (1-s)\theta_1)\mathrm{d}s \in \mathbb{R}^{d\times d}$。注意，对于任意 $\theta \in \Theta$， g_t 的梯度为

$$
\nabla g_t(\theta) = \lambda c_\mu I_d + \sum_{s=1}^{t-1} w_{t-1,s}\mu'(X_s^\top\theta)X_s X_s^\top \succeq c_\mu V_{t-1},
$$

这显然意味着 $\forall\theta_1, \theta_2 \in \Theta, G_t(\theta_1, \theta_2) \succeq c_\mu V_{t-1}$。

根据假设 2，对 g_t 应用中值定理（44）和投影（12），我们有

$$
\begin{aligned}
&|\mu(\mathbf{x}^\top\widetilde{\theta}_t) - \mu(\mathbf{x}^\top\theta_t)| \leq k_\mu|\mathbf{x}^\top(\widetilde{\theta}_t - \theta_t)| \\
&= k_\mu|\mathbf{x}^\top G_t^{-1}(\theta_t, \widetilde{\theta}_t)(g_t(\widetilde{\theta}_t) - g_t(\theta_t))| \\
&\leq k_\mu\|\mathbf{x}\|_{G_t^{-1}(\theta_t,\widetilde{\theta}_t)}\|g_t(\widetilde{\theta}_t) - g_t(\theta_t)\|_{G_t^{-1}(\theta_t,\widetilde{\theta}_t)} \\
&\leq \frac{k_\mu}{c_\mu}\|\mathbf{x}\|_{V_{t-1}^{-1}}\|g_t(\widetilde{\theta}_t) - g_t(\theta_t)\|_{V_{t-1}^{-1}} \\
&\leq \frac{2k_\mu}{c_\mu}\|\mathbf{x}\|_{V_{t-1}^{-1}}\|g_t(\widehat{\theta}_t) - g_t(\theta_t)\|_{V_{t-1}^{-1}},
\end{aligned}
$$

然后，根据模型假设、函数 g_t （13）和 $\sim g_t$（θ_t）（43），我们有，

$$
\begin{aligned}
&g_t(\theta_t) - g_t(\widehat{\theta}_t) \\
&= \lambda c_\mu\theta_t + \sum_{s=1}^{t-1} w_{t-1,s}\mu(X_s^\top\theta_t)X_s - \sum_{s=1}^{t-1} w_{t-1,s} r_s X_s \\
&= \lambda c_\mu\theta_t + \sum_{s=1}^{t-1} w_{t-1,s}\mu(X_s^\top\theta_t)X_s \\
&\quad - \sum_{s=1}^{t-1} w_{t-1,s}(\mu(X_s^\top\theta_s) + \eta_s)X_s \\
&= \underbrace{\sum_{s=1}^{t-1} w_{t-1,s}(\mu(X_s^\top\theta_t) - \mu(X_s^\top\theta_s))X_s}_{\texttt{bias part}} \\
&\quad + \underbrace{\lambda c_\mu\theta_t - \sum_{s=1}^{t-1} w_{t-1,s}\eta_s X_s}_{\texttt{variance part}}.
\end{aligned}
$$

然后，根据柯西-施瓦茨不等式，我们知道对于任意$\mathbf{x} \in \mathcal{X}$,

$$
\left|\mu(\mathbf{x}^\top\widetilde{\theta}_t) - \mu(\mathbf{x}^\top\theta_t)\right| \leq \frac{2k_\mu}{c_\mu}\|\mathbf{x}\|_{V_{t-1}^{-1}}(C_t + D_t), \tag{45}
$$

其中

$$C_t = \left\| \sum_{s=1}^{t-1} w_{t-1,s}(\mu(X_s^\top \theta_t) - \mu(X_s^\top \theta_s)) X_s \right\|_{V_{t-1}^{-1}}$$

$$D_t = \left\| \sum_{s=1}^{t-1} w_{t-1,s} \eta_s X_s - \lambda c_\mu \theta_t \right\|_{V_{t-1}^{-1}}.$$

这两项可以分别进行界定，如以下引理所述。

引理10. 对于任意 $t \in [T]$，我们有

$$\begin{aligned} &\left\| \sum_{s=1}^{t-1} w_{t-1,s}(\mu(X_s^\top \theta_t) - \mu(X_s^\top \theta_s)) X_s \right\|_{V_{t-1}^{-1}} \\ &\le L k_\mu \sqrt{d} \sum_{p=1}^{t-1} \sqrt{\sum_{s=1}^{p} w_{t-1,s}} \left\| \theta_p - \theta_{p+1} \right\|_2 . \end{aligned} \tag{46}$$

引理11. 对于任意 $\delta \in (0,1)$，以至少 $1-\delta$ 的概率，以下结论对所有 $t \in [T]$ 均成立，

$$\begin{aligned} &\left\| \sum_{s=1}^{t-1} w_{t-1,s} \eta_s X_s - \lambda c_\mu \theta_t \right\|_{V_{t-1}^{-1}} \le \sqrt{\lambda} c_\mu S \\ &+ R \sqrt{2 \log \frac{1}{\delta} + d \log \left(1 + \frac{L^2 \sum_{s=1}^{t-1} w_{t-1,s}}{\lambda d} \right)}. \end{aligned} \tag{47}$$

根据不等式 (45)、引理10、引理11 以及可行集的有界性假设，对于任意$\mathbf{x} \in \mathcal{X}$, $\gamma \in (0,1)$，$\delta \in (0,1)$，以至少 $1-\delta$ 的概率，以下结论对所有 $t \in [T]$ 均成立，

$$\begin{aligned} \left| \mu(\mathbf{x}^\top \widetilde{\theta}_t) - \mu(\mathbf{x}^\top \theta_t) \right| &\le \frac{2k_\mu}{c_\mu} \|\mathbf{x}\|_{V_{t-1}^{-1}} \\ &\cdot \left(L k_\mu \sqrt{d} \sum_{p=1}^{t-1} \sqrt{\sum_{s=1}^{p} w_{t-1,s}} \left\| \theta_p - \theta_{p+1} \right\|_2 + \bar{\beta}_{t-1} \right) \\ \le \frac{2k_\mu}{c_\mu} \Bigg(L^2 k_\mu \sqrt{\frac{d}{\lambda}} &\sum_{p=1}^{t-1} \sqrt{\sum_{s=1}^{p} w_{t-1,s}} \left\| \theta_p - \theta_{p+1} \right\|_2 \\ &+ \bar{\beta}_{t-1} \|\mathbf{x}\|_{V_{t-1}^{-1}} \Bigg), \end{aligned}$$

其中 $\bar{\beta}_t \triangleq \sqrt{\lambda} c_\mu S + R \sqrt{2 \log \frac{1}{\delta} + d \log \left(1 + \frac{L^2 \sum_{s=1}^{t} w_{t,s}}{\lambda d} \right)}$ 是 GLB-WeightUCB 中使用的置信半径。因此，证明完成。 □

引理*10*的证明。 这里我们需要提取时变参数的变化量 θ_t

$$\begin{aligned} &\left\| \sum_{s=1}^{t-1} w_{t-1,s}(\mu(X_s^\top \theta_t) - \mu(X_s^\top \theta_s)) X_s \right\|_{V_{t-1}^{-1}} \\ &\le \left\| \sum_{s=1}^{t-1} w_{t-1,s} \sum_{p=s}^{t-1} (\mu(X_s^\top \theta_p) - \mu(X_s^\top \theta_{p+1})) X_s \right\|_{V_{t-1}^{-1}} \\ &= \left\| \sum_{p=1}^{t-1} \sum_{s=1}^{p} w_{t-1,s} \alpha(X_s, \theta_p, \theta_{p+1}) (X_s^\top \theta_p - X_s^\top \theta_{p+1}) X_s \right\|_{V_{t-1}^{-1}} \\ &\le \sum_{p=1}^{t-1} \left\| \sum_{s=1}^{p} w_{t-1,s} \alpha(X_s, \theta_p, \theta_{p+1}) X_s \|X_s\|_2 \|\theta_p - \theta_{p+1}\|_2 \right\|_{V_{t-1}^{-1}} \\ &\le L \sum_{p=1}^{t-1} \sum_{s=1}^{p} w_{t-1,s} |\alpha(X_s, \theta_p, \theta_{p+1})| \, \|X_s\|_{V_{t-1}^{-1}} \|\theta_p - \theta_{p+1}\|_2 \\ &\le L k_\mu \sum_{p=1}^{t-1} \sum_{s=1}^{p} w_{t-1,s} \|X_s\|_{V_{t-1}^{-1}} \|\theta_p - \theta_{p+1}\|_2 . \end{aligned}$$

其中第四个方程归因于中值定理，其中 $\alpha(\mathbf{x}\ \theta_1, \theta_2) = \int_0^1 \mu'(v\mathbf{x}^\top \theta_2 + (1-v)x^\top \theta_1) \mathrm{d}v$: $\mu(X_s^\top \theta_p) - \mu(X_s^\top \theta_{p+1}) = \alpha(X_s, \theta_p, \theta_{p+1})(X_s^\top \theta_p - X_s^\top \theta_{p+1})$. 接下来，项 $\sum_{s=1}^{p} w_{t-1,s} \|X_s\|_{V_{t-1}^{-1}}$ 的界的推导与附录 A-B 中的不等式 (34) 相同，因此证明完成。 □

引理的证明 $\widetilde{\eta}_s \triangleq$. $\sqrt{w_{t-1,s}} \eta_s$ $\widetilde{X}_s \triangleq \sqrt{w_{t-1,s}} X_s$ $\forall t \in$ 与线性情形相同，我们定义 ,和 ., 并注意到 $[T], s \in [t-1], |w_{t-1,s}| \le 1$, 那么 $\widetilde{\eta}_s$ 仍然是 R-次高斯的，由定理7可得

$$\begin{aligned} D_t &= \left\| \sum_{s=1}^{t-1} w_{t-1,s} \eta_s X_s - \lambda c_\mu \theta_t \right\|_{V_{t-1}^{-1}} \\ &\le \left\| \sum_{s=1}^{t-1} w_{t-1,s} \eta_s X_s \right\|_{V_{t-1}^{-1}} + \|\lambda c_\mu \theta_t\|_{V_{t-1}^{-1}} \\ &\le \left\| \sum_{s=1}^{t-1} \widetilde{\eta}_s \widetilde{X}_s \right\|_{V_{t-1}^{-1}} + \sqrt{\lambda} c_\mu S \\ &\le R \sqrt{2 \log \frac{1}{\delta} + d \log \left(1 + \frac{L^2 \sum_{s=1}^{t-1} w_{t-1,s}}{d\lambda} \right)} + \sqrt{\lambda} c_\mu S. \end{aligned}$$

这里的证明与附录 A-B 中引理 8 的证明相同，唯一的区别是第二项中多了一个 c_μ 。 □

C. 定理*2*的证明

证明。 令 $X_t^* \triangleq \arg\max_{\mathbf{x} \in \mathcal{X}} \mu(\mathbf{x}^\top \theta_t)$。根据引理 2以及 $X_t^*, X_t \in \mathcal{X}$这一事实，以下各项成立的概率至少为 $1-\delta$,

$$\begin{aligned} &\forall t \in [T], \mu(X_t^{*\top} \theta_t) \le \mu(X_t^{*\top} \widetilde{\theta}_t) \\ &+ \frac{2k_\mu}{c_\mu} \Bigg(L^2 k_\mu \sqrt{\frac{d}{\lambda}} \sum_{p=1}^{t-1} \sqrt{\sum_{s=1}^{p} w_{t-1,s}} \left\| \theta_p - \theta_{p+1} \right\|_2 \\ &\qquad + \bar{\beta}_{t-1} \|X_t^*\|_{V_{t-1}^{-1}} \Bigg), \\ &\forall t \in [T], \mu(X_t^\top \theta_t) \ge \mu(X_t^\top \widetilde{\theta}_t) \\ &- \frac{2k_\mu}{c_\mu} \Bigg(L^2 k_\mu \sqrt{\frac{d}{\lambda}} \sum_{p=1}^{t-1} \sqrt{\sum_{s=1}^{p} w_{t-1,s}} \left\| \theta_p - \theta_{p+1} \right\|_2 \\ &\qquad + \bar{\beta}_{t-1} \|X_t\|_{V_{t-1}^{-1}} \Bigg). \end{aligned}$$

根据并集界，以下结论成立的概率至少为 $1-2\delta$: $\forall t \in [T]$

$$\begin{aligned}
&\mu(X_t^{*\top}\theta_t) - \mu(X_t^\top\theta_t) \le \mu(X_t^{*\top}\widetilde{\theta}_t) - \mu(X_t^\top\widetilde{\theta}_t) \\
&\quad + \frac{2k_\mu}{c_\mu}\left(\bar{\beta}_{t-1}\|X_t^*\|_{V_{t-1}^{-1}} + \bar{\beta}_{t-1}\|X_t\|_{V_{t-1}^{-1}}\right) \\
&\quad + \frac{4L^2k_\mu^2}{c_\mu}\sqrt{\frac{d}{\lambda}}\sum_{p=1}^{t-1}\sqrt{\sum_{s=1}^{p} w_{t-1,s}}\left\|\theta_p - \theta_{p+1}\right\|_2 \\
&\le \frac{4L^2k_\mu^2}{c_\mu}\sqrt{\frac{d}{\lambda}}\sum_{p=1}^{t-1}\sqrt{\sum_{s=1}^{p} w_{t-1,s}}\left\|\theta_p - \theta_{p+1}\right\|_2 \\
&\quad + \frac{4k_\mu}{c_\mu}\bar{\beta}_{t-1}\|X_t\|_{V_{t-1}^{-1}},
\end{aligned}$$

其中，最后一步来自臂选择准则（15），因此

$$\begin{aligned}
&\mu(X_t^{*\top}\widetilde{\theta}_t) + \frac{2k_\mu}{c_\mu}\bar{\beta}_{t-1}\|X_t^*\|_{V_{t-1}^{-1}} \\
&\le \mu(X_t^\top\widetilde{\theta}_t) + \frac{2k_\mu}{c_\mu}\bar{\beta}_{t-1}\|X_t\|_{V_{t-1}^{-1}}.
\end{aligned}$$

因此，以下动态遗憾界以至少 $1-2\delta$ 的概率成立，并且可以分为两部分，

$$\begin{aligned}
\text{D-REG}_T &= \sum_{t=1}^{T}\max_{\mathbf{x}\in\mathcal{X}}\mu(\mathbf{x}^\top\theta_t) - \mu(X_t^\top\theta_t) \\
&\le \underbrace{\frac{4L^2k_\mu^2}{c_\mu}\sqrt{\frac{d}{\lambda}}\sum_{t=1}^{T}\sum_{p=1}^{t-1}\sqrt{\sum_{s=1}^{p} w_{t-1,s}}\left\|\theta_p - \theta_{p+1}\right\|_2}_{\texttt{bias part}} \\
&\quad + \underbrace{\frac{4k_\mu}{c_\mu}\bar{\beta}_T\sum_{t=1}^{T}\|X_t\|_{V_{t-1}^{-1}}}_{\texttt{variance part}}.
\end{aligned}$$

其中 $\bar{\beta}_t = \sqrt{\lambda}c_\mu S + R\sqrt{2\log\frac{1}{\delta} + d\log\left(1 + \frac{L^2(1-\gamma^{2t})}{\lambda d(1-\gamma^2)}\right)}$ 是置信半径。

现在我们推导这两部分的上界。

偏差部分。 类似于不等式（35）的证明，我们有

$$\begin{aligned}
&\frac{4L^2k_\mu^2}{c_\mu}\sqrt{\frac{d}{\lambda}}\sum_{t=1}^{T}\sum_{p=1}^{t-1}\sqrt{\sum_{s=1}^{p} w_{t-1,s}}\left\|\theta_p - \theta_{p+1}\right\|_2 \\
&\le \frac{8L^2k_\mu^2}{c_\mu}\sqrt{\frac{d}{\lambda}}\frac{1}{(1-\gamma)^{3/2}}P_T.
\end{aligned}$$

方差部分。 与不等式 (36) 的证明类似，令 $C_T^{\text{GLB}} \triangleq \frac{4k_\mu}{c_\mu}\bar{\beta}_T\sqrt{2\max\{1, L^2/\lambda\}\,dT}$，我们有

$$\begin{aligned}
&\frac{4k_\mu}{c_\mu}\bar{\beta}_T\sqrt{T}\sqrt{\sum_{t=1}^{T}\|X_t\|^2_{V_{t-1}^{-1}}} \\
&\le C_T^{\text{GLB}}\sqrt{T\log\frac{1}{\gamma} + \log\left(1 + \frac{L^2}{\lambda d(1-\gamma)}\right)}.
\end{aligned}$$

算法 6 SCB-WeightUCB

需要： 时间范围 T、折扣因子 γ、置信度 δ，正则化项 λ、逆链接函数 μ、参数 S, L 和 m

设置 $V_0 = \lambda I_d, \widehat{\theta}_1 = \mathbf{0}$ 并计算 k_μ 和 c_μ
对于 $t = 1, 2, ..., T$ **do**
  if $\|\widehat{\theta}_t\|_2 \le S$ **然后**
    let $\theta_t = \widehat{\theta}_t$
  else
    执行投影并得到 $\widetilde{\theta}_t$ 根据（17）
  结束 if
  Compute $\widetilde{\beta}_{t-1}$ 根据（19）
  选择 X_t 根据（20）
  获取奖励 r_t
  更新 $V_t = \gamma V_{t-1} + X_t X_t^\top + (1-\gamma)\lambda I_d$
  计算 $\widehat{\theta}_{t+1}$ 根据 (11)
结束循环

结合偏差部分和方差部分的上界，并令 $\delta = 1/(2T^2)$，我们以至少 $1 - 1/T$ 的概率得到以下遗憾界，

$$\begin{aligned}
\text{D-REG}_T &\le \frac{8L^2k_\mu^2}{c_\mu}\sqrt{\frac{d}{\lambda}}\frac{1}{(1-\gamma)^{3/2}}P_T \\
&\quad + C_T^{\text{GLB}}\sqrt{T\log\frac{1}{\gamma} + \log\left(1 + \frac{L^2}{\lambda d(1-\gamma)}\right)}.
\end{aligned}$$

其中 $\bar{\beta}_t = R\sqrt{4\log T + 2\log 2 + d\log\left(1 + \frac{L^2}{(1-\gamma^t)\lambda d(1-\gamma)}\right)} + \sqrt{\lambda}c_\mu S$。我们设 $\gamma \ge 1/T$ 和 $\lambda = d/c_\mu^2$，并得到，

$$\text{D-REG}_T \le \widetilde{\mathcal{O}}\left(k_\mu^2\frac{1}{(1-\gamma)^{3/2}}P_T + \frac{k_\mu}{c_\mu}d(1-\gamma)^{1/2}T\right).$$

当 $P_T < d/(k_\mu c_\mu T)$ 时，我们设置 $\gamma = 1 - 1/T$，并获得一个 $\widetilde{\mathcal{O}}(k_\mu c_\mu^{-1}d\sqrt{T})$ 遗憾界。当 $P_T \ge d/(k_\mu c_\mu T)$ 时，我们将 γ 最优地设为 $1-\gamma = \sqrt{k_\mu c_\mu P_T/(dT)}$，并获得一个 $\widetilde{\mathcal{O}}(k_\mu^{5/4}c_\mu^{-3/4}d^{3/4}P_T^{1/4}T^{3/4})$ 遗憾界。注意，如果 $k_\mu < 1\sqrt{dT}$，我们只需令 $1-\gamma = \sqrt{c_\mu P_T/(\,)}$，此时遗憾界变为 $\widetilde{\mathcal{O}}(k_\mu^2 c_\mu^{-3/4}d^{3/4}P_T^{1/4}T^{3/4})$。

附录 C：SCB-WEIGHTUCB 分析

本节首先在算法 6 中介绍 SCB-WeightUCB 算法。接着，在附录 C-A 中，我们给出 SCB-WeightUCB 算法估计误差上界的证明（引理 3）。最后，在附录C-B中，我们给出动态遗憾上界的证明（定理3）。

A. 引理 3 的证明

证明。 根据估计方程 (11)，我们知道

$$\begin{aligned}
g_t(\widehat{\theta}_t) &= \lambda c_\mu\widehat{\theta}_t + \sum_{s=1}^{t-1} w_{t-1,s}\mu(X_s^\top\widehat{\theta}_t)X_s \\
&= \sum_{s=1}^{t-1} w_{t-1,s}r_s X_s,
\end{aligned} \tag{48}$$

然后根据中值定理，我们知道

$$g_t(\theta_1) - g_t(\theta_2) = G_t(\theta_1, \theta_2)(\theta_1 - \theta_2), \tag{49}$$

其中 $G_t(\theta_1, \theta_2) \triangleq \int_0^1 \nabla g_t(s\theta_2+(1-s)\theta_1)\mathrm{d}s \in \mathbb{R}^{d\times d}$。注意，对于任意 $\theta \in \Theta$, g_t 的梯度为

$$\nabla g_t(\theta) = \lambda c_\mu I_d + \sum_{s=1}^{t-1} w_{t-1,s}\mu'(X_s^\top\theta)X_sX_s^\top \succeq c_\mu V_{t-1},$$

这显然意味着 $\forall\theta_1, \theta_2 \in \Theta$, $G_t(\theta_1, \theta_2) \succeq c_\mu V_{t-1}$ 且$\forall\theta$, $H_t(\theta) \succeq c_\mu V_{t-1}$，其中 $H_t(\theta)$ 定义为

$$H_t(\theta) \triangleq \lambda c_\mu I_d + \sum_{s=1}^{t-1} w_{t-1,s}\mu'(X_s^\top\theta)X_sX_s^\top. \tag{50}$$

根据假设 2、g_t 上的中值定理 (44)、投影 (17) 和引理 28，我们有

$$\begin{aligned}
&|\mu(\mathbf{x}^\top\widetilde{\theta}_t) - \mu(\mathbf{x}^\top\theta_t)| \le k_\mu|\mathbf{x}^\top(\widetilde{\theta}_t - \theta_t)| \\
&= k_\mu|\mathbf{x}^\top G_t^{-1}(\theta_t, \widetilde{\theta}_t)(g_t(\widetilde{\theta}_t) - g_t(\theta_t))| \\
&\le k_\mu\|\mathbf{x}\|_{G_t^{-1}(\theta_t,\widetilde{\theta}_t)}\|g_t(\widetilde{\theta}_t) - g_t(\theta_t)\|_{G_t^{-1}(\theta_t,\widetilde{\theta}_t)} \\
&\le k_\mu\|\mathbf{x}\|_{G_t^{-1}(\theta_t,\widetilde{\theta}_t)}\Big(\|g_t(\widetilde{\theta}_t) - g_t(\widehat{\theta}_t)\|_{G_t^{-1}(\theta_t,\widetilde{\theta}_t)} \\
&\qquad + \|g_t(\widehat{\theta}_t) - g_t(\theta_t)\|_{G_t^{-1}(\theta_t,\widetilde{\theta}_t)}\Big) \\
&\le \sqrt{1+2S}k_\mu\|\mathbf{x}\|_{G_t^{-1}(\theta_t,\widetilde{\theta}_t)}\Big(\|g_t(\widetilde{\theta}_t) - g_t(\widehat{\theta}_t)\|_{H_t^{-1}(\widetilde{\theta}_t)} \\
&\qquad + \|g_t(\widehat{\theta}_t) - g_t(\theta_t)\|_{H_t^{-1}(\theta_t)}\Big) \\
&\le 2\sqrt{1+2S}\frac{k_\mu}{\sqrt{c_\mu}}\|\mathbf{x}\|_{V_{t-1}^{-1}}\|g_t(\widehat{\theta}_t) - g_t(\theta_t)\|_{H_t^{-1}(\theta_t)},
\end{aligned} \tag{13}$$

那么，根据模型假设 (16)、函数 $\widehat{g}_t$ 和 $g_t(\theta_t)$ (48)，我们有，

$$\begin{aligned}
&g_t(\theta_t) - g_t(\widehat{\theta}_t) \\
&= \lambda c_\mu\theta_t + \sum_{s=1}^{t-1} w_{t-1,s}\mu(X_s^\top\theta_t)X_s - \sum_{s=1}^{t-1} w_{t-1,s}r_sX_s \\
&= \lambda c_\mu\theta_t + \sum_{s=1}^{t-1} w_{t-1,s}\mu(X_s^\top\theta_t)X_s \\
&\quad - \sum_{s=1}^{t-1} w_{t-1,s}(\mu(X_s^\top\theta_s) + \eta_s)X_s \\
&= \sum_{s=1}^{t-1} w_{t-1,s}(\mu(X_s^\top\theta_t) - \mu(X_s^\top\theta_s))X_s + \lambda c_\mu\theta_t \\
&\quad - \sum_{s=1}^{t-1} w_{t-1,s}\eta_sX_s,
\end{aligned}$$

那么，根据柯西-施瓦茨不等式，我们有

$$\begin{aligned}
&\left|\mu(\mathbf{x}^\top\widetilde{\theta}_t) - \mu(\mathbf{x}^\top\theta_t)\right| \\
&\le 2\sqrt{1+2S}\frac{k_\mu}{\sqrt{c_\mu}}\|\mathbf{x}\|_{V_{t-1}^{-1}}(E_t + F_t),
\end{aligned} \tag{51}$$

其中

$$\begin{aligned}
E_t &= \left\|\sum_{s=1}^{t-1} w_{t-1,s}(\mu(X_s^\top\theta_t) - \mu(X_s^\top\theta_s))X_s\right\|_{H_t^{-1}(\theta_t)} \\
F_t &= \left\|\sum_{s=1}^{t-1} w_{t-1,s}\eta_sX_s - \lambda c_\mu\theta_t\right\|_{H_t^{-1}(\theta_t)}.
\end{aligned}$$

这两项可以分别进行界定。

引理12。对于任意 $t \in [T]$，我们有

$$\begin{aligned}
&\left\|\sum_{s=1}^{t-1} w_{t-1,s}(\mu(X_s^\top\theta_t) - \mu(X_s^\top\theta_s))X_s\right\|_{H_t^{-1}(\theta_t)} \\
&\le L\frac{k_\mu}{\sqrt{c_\mu}}\sqrt{d}\sum_{p=1}^{t-1}\sqrt{\sum_{s=1}^{p} w_{t-1,s}}\left\|\theta_p - \theta_{p+1}\right\|_2.
\end{aligned} \tag{52}$$

引理13。对于任意 $\delta \in (0,1)$，以至少 $1-\delta$ 的概率，对于所有 $t \in [T]$，我们有：

$$\begin{aligned}
&\left\|\sum_{s=1}^{t-1} w_{t-1,s}\eta_sX_s - \lambda c_\mu\theta_t\right\|_{H_t^{-1}(\theta_t)} \\
&\le \frac{\sqrt{\lambda c_\mu}}{2m} + \frac{2m}{\sqrt{\lambda c_\mu}}\log\frac{1}{\delta} + \frac{2m}{\sqrt{\lambda c_\mu}}d\log(2) + \sqrt{\lambda c_\mu}S \\
&\quad + \frac{dm}{\sqrt{\lambda c_\mu}}\log\left(1 + \frac{L^2k_\mu\sum_{s=1}^{t-1} w_{t-1,s}}{\lambda c_\mu d}\right),
\end{aligned} \tag{53}$$

根据不等式 (51)、引理10 和引理11，以及可行集的有界性假设，对于任意 $\mathbf{x} \in \mathcal{X}$, $\delta \in (0,1)$，以至少 $1-\delta$ 的概率，对于所有 $t \in [T]$，我们有：

$$\begin{aligned}
&\left|\mu(\mathbf{x}^\top\widetilde{\theta}_t) - \mu(\mathbf{x}^\top\theta_t)\right| \\
&\le 2\sqrt{1+2S}\frac{k_\mu}{\sqrt{c_\mu}}\|\mathbf{x}\|_{V_{t-1}^{-1}}\Bigg(\widetilde{\beta}_{t-1} \\
&\qquad + L\frac{k_\mu}{\sqrt{c_\mu}}\sqrt{d}\sum_{p=1}^{t-1}\sqrt{\sum_{s=1}^{p} w_{t-1,s}}\left\|\theta_p - \theta_{p+1}\right\|_2\Bigg) \\
&\le 2\sqrt{1+2S}\frac{k_\mu}{\sqrt{c_\mu}}\Bigg(\widetilde{\beta}_{t-1}\|\mathbf{x}\|_{V_{t-1}^{-1}} \\
&\qquad + L^2\frac{k_\mu}{\sqrt{\lambda c_\mu}}\sqrt{d}\sum_{p=1}^{t-1}\sqrt{\sum_{s=1}^{p} w_{t-1,s}}\left\|\theta_p - \theta_{p+1}\right\|_2\Bigg),
\end{aligned}$$

其中，$\widetilde{\beta}_t \triangleq \frac{dm}{\sqrt{\lambda c_\mu}}\left(1 + \frac{L^2k_\mu\sum_{s=1}^{t-1} w_{t-1,s}}{\lambda c_\mu d}\right) + \frac{\sqrt{\lambda c_\mu}}{2m} + \log\frac{1}{\delta} + \frac{2m}{\sqrt{\lambda c_\mu}}d\log(2) + \sqrt{\lambda c_\mu}S$

~~log log~~( ) 是 ~~SCB-Weight~~UCB 中使用的置信半径。因此，我们完成了证明。 □

引理*12*的证明。由于 $\forall\theta, H_t(\theta) \succeq c_\mu V_{t-1}$，我们有

$$\begin{aligned}
&\left\|\sum_{s=1}^{t-1} w_{t-1,s}(\mu(X_s^\top\theta_t) - \mu(X_s^\top\theta_s))X_s\right\|_{H_t^{-1}(\theta_t)} \\
&\le \frac{1}{\sqrt{c_\mu}}\left\|\sum_{s=1}^{t-1} w_{t-1,s}(\mu(X_s^\top\theta_t) - \mu(X_s^\top\theta_s))X_s\right\|_{V_{t-1}^{-1}}.
\end{aligned}$$

然后使用引理10，证明完毕。 □

引理证明 13 $\widetilde{\eta}_s \triangleq \sqrt{w_{t-1,s}}\eta_s$ $\widetilde{X}_s \triangleq \sqrt{w_{t-1,s}}X_s$。我们定义 、、 $\forall t \in [T]$ $s \in [t-1]$ $|w_{t-1,s}| \leq 1$ 并注意到、、，那么 $\widetilde{\eta}_s$ 的方差为 $\widetilde{\sigma}_s$，且其被 1 所界定，于是我们有

$$\begin{aligned} F_t &= \left\| \sum_{s=1}^{t-1} w_{t-1,s}\eta_s X_s - \lambda c_\mu \theta_t \right\|_{H_t^{-1}(\theta_t)} \\ &\leq \left\| \sum_{s=1}^{t-1} w_{t-1,s}\eta_s X_s \right\|_{H_t^{-1}(\theta_t)} + \sqrt{\lambda c_\mu} S \\ &= \left\| \sum_{s=1}^{t-1} \widetilde{\eta}_s \widetilde{X}_s \right\|_{\widetilde{H}_t^{-1}(\theta_t)} + \sqrt{\lambda c_\mu} S \end{aligned}$$

其中 $\widetilde{H}_t(\theta) \triangleq \frac{\lambda c_\mu}{m^2} I_d + \sum_{s=1}^{t-1} \frac{\mu'(X_s^\top \theta)}{m^2} \widetilde{X}_s \widetilde{X}_s^\top$。根据模型假设 (16)，我们知道 $\widetilde{\sigma}_t^2 = \frac{\mathbb{E}[\widetilde{\eta}_t^2 | \mathcal{F}_{t-1}]}{m^2} \leq \mathbb{E}[\eta_t^2 | \mathcal{F}_t] =$ 方差 $\frac{[r_t|\mathcal{F}_t]}{m^2} = \mu'(X_t\theta_t)\, m^2$，然后由自协调赌博机的自归一化集中不等式 [33, 定理 1]（在定理 8 中重述），我们可以得到第一项 $\left\| \sum_{s=1}^{t-1} \widetilde{\eta}_s \widetilde{X}_s \right\|_{\widetilde{H}_t^{-1}(\theta_t)}$ 的界如下：

$$\begin{aligned} &\left\| \sum_{s=1}^{t-1} w_{t-1,s}\widetilde{\eta}_s X_s \right\|_{H_t^{-1}(\theta_t)} \\ &\leq \frac{\sqrt{\lambda c_\mu}}{2m} + \frac{2m}{\sqrt{\lambda c_\mu}} \log\left(\frac{\det(H_t)^{1/2}}{\delta (\lambda c_\mu)^{d/2}} \right) + \frac{2m}{\sqrt{\lambda c_\mu}} d \log(2), \end{aligned}$$

于是我们有

$$\begin{aligned} \left\| \sum_{s=1}^{t-1} \gamma^{-s} \eta_s X_s \right\|_{\widetilde{H}_t^{-1}(\theta_t)} &\leq \frac{\sqrt{\lambda c_\mu}}{2m} + \frac{2m}{\sqrt{\lambda c_\mu}} \log \frac{1}{\delta} \\ + \frac{dm}{\sqrt{\lambda c_\mu}} \log\left(1 + \frac{L^2 k_\mu \sum_{s=1}^{t-1} w_{t-1,s}}{\lambda c_\mu d} \right) &+ \frac{2m}{\sqrt{\lambda c_\mu}} d \log(2). \end{aligned}$$

因此，我们得到了 F_t 项的上界。 □

B. 定理*3*的证明

证明。令 $X_t^* \triangleq \arg\max_{\mathbf{x} \in \mathcal{X}} \mu(\mathbf{x}^\top \theta_t)$。根据引理 2以及 $X_t^*, X_t \in \mathcal{X}$这一事实，以下各式均以至少 $1-\delta$ 的概率成立，

$$\begin{aligned} &\forall t \in [T], \mu(X_t^{*\top}\theta_t) \\ &\leq \mu(X_t^{*\top}\widetilde{\theta}_t) + 2\sqrt{1+2S} \frac{k_\mu}{\sqrt{c_\mu}} \left(\widetilde{\beta}_{t-1} \|X_t^*\|_{V_{t-1}^{-1}} \right. \\ &\quad \left. + L^2 \frac{k_\mu}{\sqrt{\lambda c_\mu}} \sqrt{d} \sum_{p=1}^{t-1} \sqrt{\sum_{s=1}^{p} w_{t-1,s}} \left\| \theta_p - \theta_{p+1} \right\|_2 \right), \\ &\forall t \in [T], \ \mu(X_t^\top \theta_t) \\ &\geq \mu(X_t^\top \widetilde{\theta}_t) - 2\sqrt{1+2S} \frac{k_\mu}{\sqrt{c_\mu}} \left(\widetilde{\beta}_{t-1} \|X_t\|_{V_{t-1}^{-1}} \right. \\ &\quad \left. + L^2 \frac{k_\mu}{\sqrt{\lambda c_\mu}} \sqrt{d} \sum_{p=1}^{t-1} \sqrt{\sum_{s=1}^{p} w_{t-1,s}} \left\| \theta_p - \theta_{p+1} \right\|_2 \right). \end{aligned}$$

根据并集界，以下结论以至少 $1-2\delta$ 的概率成立： $\forall t \in [T]$

$$\begin{aligned} &\mu(X_t^{*\top}\theta_t) - \mu(X_t^\top \theta_t) \leq \mu(X_t^{*\top}\widetilde{\theta}_t) - \mu(X_t^\top \widetilde{\theta}_t) \\ &+ 2\sqrt{1+2S} \left(\frac{2L^2 k_\mu^2}{c_\mu} \sqrt{\frac{d}{\lambda}} \sum_{p=1}^{t-1} \sqrt{\sum_{s=1}^{p} w_{t-1,s}} \left\| \theta_p - \theta_{p+1} \right\|_2 \right. \\ &\left. + \frac{k_\mu}{\sqrt{c_\mu}} \left(\widetilde{\beta}_{t-1} \|X_t^*\|_{V_{t-1}^{-1}} + \widetilde{\beta}_{t-1} \|X_t\|_{V_{t-1}^{-1}} \right) \right) \\ &\leq \frac{4\sqrt{1+2S} L^2 k_\mu^2}{c_\mu} \sqrt{\frac{d}{\lambda}} \sum_{p=1}^{t-1} \sqrt{\sum_{s=1}^{p} w_{t-1,s}} \left\| \theta_p - \theta_{p+1} \right\|_2 \\ &+ \frac{4\sqrt{1+2S} k_\mu}{\sqrt{c_\mu}} \widetilde{\beta}_{t-1} \|X_t\|_{V_{t-1}^{-1}}, \end{aligned}$$

其中，最后一步由臂选择准则 (20) 得出，因此

$$\begin{aligned} &\mu(X_t^{*\top}\widetilde{\theta}_t) + 2\sqrt{1+2S} \frac{k_\mu}{\sqrt{c_\mu}} \widetilde{\beta}_{t-1} \|X_t^*\|_{V_{t-1}^{-1}} \\ &\leq \mu(X_t^\top \widetilde{\theta}_t) + 2\sqrt{1+2S} \frac{k_\mu}{\sqrt{c_\mu}} \widetilde{\beta}_{t-1} \|X_t\|_{V_{t-1}^{-1}}. \end{aligned}$$

因此，以下动态遗憾界以至少 $1-2\delta$ 的概率成立，并可分为两部分，

$$\begin{aligned} \text{D-REG}_T &= \sum_{t=1}^{T} \mu(X_t^{*\top}\theta_t) - \mu(X_t^\top \theta_t) \\ &\leq \underbrace{\frac{4\sqrt{1+2S} L^2 k_\mu^2}{c_\mu} \sqrt{\frac{d}{\lambda}} \sum_{t=1}^{T} \sum_{p=1}^{t-1} \sqrt{\sum_{s=1}^{p} w_{t-1,s}} \left\| \theta_p - \theta_{p+1} \right\|_2}_{\texttt{bias part}} \\ &\quad + \underbrace{\frac{4\sqrt{1+2S} k_\mu}{\sqrt{c_\mu}} \widetilde{\beta}_T \sum_{t=1}^{T} \|X_t\|_{V_{t-1}^{-1}}}_{\texttt{variance part}}. \end{aligned}$$

where $\widetilde{\beta}_t = \frac{dm}{\sqrt{\lambda c_\mu}} \log\left(1 + \frac{L^2 k_\mu (1-\gamma^t)}{\lambda c_\mu d (1-\gamma)}\right) + \frac{2m}{\sqrt{\lambda c_\mu}} \log \frac{1}{\delta} + \frac{\sqrt{\lambda c_\mu}}{2m} + \frac{2m}{\sqrt{\lambda c_\mu}} d \log(2) + \sqrt{\lambda c_\mu} S$ is the confidence radius.

Now we derive the upper bound for these two parts.

偏差部分。类似于不等式 (35) 的证明，我们有

$$\begin{aligned} &\frac{4\sqrt{1+2S} L^2 k_\mu^2}{c_\mu} \sqrt{\frac{d}{\lambda}} \sum_{t=1}^{T} \sum_{p=1}^{t-1} \sqrt{\sum_{s=1}^{p} w_{t-1,s}} \left\| \theta_p - \theta_{p+1} \right\|_2 \\ &\leq \frac{8\sqrt{1+2S} L^2 k_\mu^2}{c_\mu} \sqrt{\frac{d}{\lambda}} \frac{1}{(1-\gamma)^{3/2}} P_T. \end{aligned}$$

方差部分。首先使用柯西-施瓦茨不等式，我们知道

$$\begin{aligned} &\frac{4\sqrt{1+2S} k_\mu}{\sqrt{c_\mu}} \widetilde{\beta}_T \sum_{t=1}^{T} \|X_t\|_{V_{t-1}^{-1}} \\ &\leq \frac{4\sqrt{1+2S} k_\mu}{\sqrt{c_\mu}} \widetilde{\beta}_T \sqrt{T} \sqrt{\sum_{t=1}^{T} \|X_t\|_{V_{t-1}^{-1}}^2}. \end{aligned}$$

对于术语 $\sqrt{\sum_{t=1}^{T}\|X_t\|^2_{V_{t-1}^{-1}}}$，我们可以使用引理 9 对其进行界定。令 $C_T^{\mathrm{SCB}} \triangleq \frac{4\sqrt{1+2S}k_\mu}{\sqrt{c_\mu}}\widetilde{\beta}_T\sqrt{2\max\{1, L^2/\lambda\}dT}$，则

$$\frac{4\sqrt{1+2S}k_\mu}{\sqrt{c_\mu}}\widetilde{\beta}_T\sqrt{T}\sqrt{\sum_{t=1}^{T}\|X_t\|^2_{V_{t-1}^{-1}}} \le C_T^{\mathrm{SCB}}\sqrt{T\log\frac{1}{\gamma}+\log\left(1+\frac{L^2}{\lambda d(1-\gamma)}\right)}.$$

结合偏差部分和方差部分的上界，并令 $\delta = 1/(2T)$，我们以至少 $1-1/T$ 的概率得到如下遗憾界，

$$\text{D-REG}_T \le \frac{8\sqrt{1+2S}L^2k_\mu^2}{c_\mu}\sqrt{\frac{d}{\lambda}}\frac{1}{(1-\gamma)^{3/2}}P_T + C_T^{\mathrm{SCB}}\sqrt{T\log\frac{1}{\gamma}+\log\left(1+\frac{L^2}{\lambda d(1-\gamma)}\right)}.$$

其中 $\widetilde{\beta}_t = \frac{dm}{\sqrt{\lambda c_\mu}}\log\left(1+\frac{L^2k_\mu}{\lambda c_\mu d(1-\gamma^t)(1-\gamma)}\right) + \frac{2m}{\sqrt{\lambda c_\mu}}\log(2T) + \frac{\sqrt{\lambda c_\mu}}{2m} + \frac{2m}{\sqrt{\lambda c_\mu}}d\log(2) + \sqrt{\lambda c_\mu}S$。由于遗憾界中存在 $T\sqrt{\log(1/\gamma)}$ 项，这意味着我们不能令 γ 接近 0，因此设 $\gamma \ge 1/T$，则有 $\log(1/\gamma) \le C(1-\gamma)$，其中 $C = \log T/(1-1/T)$。接着，忽略时间范围 T 中的对数因子，并令 $\lambda = d\log(T)/c_\mu$，最终得到：

$$\text{D-REG}_T \le \widetilde{\mathcal{O}}\left(\frac{k_\mu^2}{\sqrt{c_\mu}}\frac{1}{(1-\gamma)^{3/2}}P_T + \frac{k_\mu}{\sqrt{c_\mu}}d(1-\gamma)^{1/2}T\right).$$

当 $P_T < d/(k_\mu T)$（对应于少量非平稳性）时，我们只需设置 $\gamma = 1-1/T$，即可获得 $\widetilde{\mathcal{O}}(k_\mu c_\mu^{-1/2}d\sqrt{T})$ 遗憾界。此外，对于 $P_T \ge d/(k_\mu T)$ 这一非退化情形，我们将折扣因子设为最优值 $1-\gamma = \sqrt{k_\mu P_T/(dT)}$，并得到 $\widetilde{\mathcal{O}}(k_\mu^{5/4}c_\mu^{-1/2}d^{3/4}P_T^{1/4}T^{3/4})$ 遗憾界，从而完成证明。

□

附录D 分段平稳SCB

本节研究分段平稳环境下的SCB，我们的工作是对 [7] 的直接改进。接下来，我们将首先提出SCB-PW-WeightUCB算法，然后分析置信集。最后，我们给出动态遗憾上界的证明。

A. SCB-PW-WeightUCB算法

受 [34], 启发，我们对 [7]进行了直接改进。正如 [7], 对 $D \ge 1$所做的那样，定义 $\mathcal{T}(D) = \{1 \le t \le T,$ 使得 $\theta_s = \theta_t$ 对于 $t-D \le s \le t-1\}$成立。当 t 距离上一个最近的变化点至少 D 步时，$t \in \mathcal{T}(D)$。但不同之处在于，[7] 将 D 视为分析参数，而我们将 D 作为可调算法参数。请注意，这里的 D 不是虚拟窗口大小，而是算法对环境保持平稳时长的估计。

算法7 SCB-PW-WeightUCB

要求： 时间范围 T、折扣因子 γ、置信度 δ，正则化项 λ、逆链接函数 μ、参数 S、L 和 m、变化置信度 D

设置 $\widehat{\theta}_0 = \mathbf{0}$ 并计算 k_μ 和 c_μ

对于 $t = 1, 2, 3, ..., T$ **执行**

  计算 $(X_t, \widetilde{\theta}_t) = \arg\max_{\mathbf{x}\in\mathcal{X},\ \theta\in\mathcal{C}_t(\delta)} \mu(\mathbf{x}^\top\theta)$

  选择 X_t 并获得奖励 r_t

  计算 θ_{t+1} 根据 (11)

结束循环

估计量。在第 t次迭代时，我们采用与漂移情形中相同的最大似然估计量（11），其中$w_{t,\,s} = \gamma^{t-s}$ 。

置信集。我们进一步为实数 θ_t 构造置信集。对于 $\delta \in (0,1)$，定义如下：

$$\mathcal{C}_t(\delta) \triangleq \left\{\theta\in\Theta \;\middle|\; \|g_t(\theta) - g_t(\widehat{\theta}_t)\|_{H_t^{-1}(\theta)} \le \rho_t\right\},$$

where $\rho_t = \frac{2L^2Sk_\mu}{\sqrt{\lambda c_\mu}}\frac{\gamma^D}{1-\gamma} + \frac{Lm}{\sqrt{\lambda c_\mu}}\frac{\gamma^D}{1-\gamma} + \breve{\beta}_t$ and $\breve{\beta}_t = \frac{dm}{\sqrt{\lambda c_\mu}}\log\left(1+\frac{L^2k_\mu(1-\gamma^{2D})}{\lambda c_\mu d(1-\gamma)}\right) + \frac{\sqrt{\lambda c_\mu}}{2m} + \frac{2m}{\sqrt{\lambda c_\mu}}\log\frac{1}{\delta} + \frac{2m}{\sqrt{\lambda c_\mu}}d\log(2) + \sqrt{\lambda c_\mu}S.$

引理14。对于任意 $\delta\in(0,1)$，以至少 $1-\delta$ 的概率，有 $\forall t\in\mathcal{T}(D)$ $\theta_t \in \mathcal{C}_t(\delta)$。

$$\mathcal{C}_t(\delta) = \left\{\theta\in\Theta \mid \|g_t(\theta) - g_t(\widehat{\theta}_t)\|_{H_t^{-1}(\theta)} \le \frac{2L^2Sk_\mu}{\sqrt{\lambda c_\mu}}\frac{\gamma^D}{1-\gamma} + \frac{Lm}{\sqrt{\lambda c_\mu}}\frac{\gamma^D}{1-\gamma} + \breve{\beta}_t\right\},$$

where $\breve{\beta}_t = \frac{dm}{\sqrt{\lambda c_\mu}}\log\left(1+\frac{L^2k_\mu(1-\gamma^{2D})}{\lambda c_\mu d(1-\gamma)}\right) + \frac{\sqrt{\lambda c_\mu}}{2m} + \frac{2m}{\sqrt{\lambda c_\mu}}\log\frac{1}{\delta} + \frac{2m}{\sqrt{\lambda c_\mu}}d\log(2) + \sqrt{\lambda c_\mu}S.$

引理14的证明见附录D-B。

选择准则。前文讨论的漂移情形下的算法采用基于奖励的选择准则。但这里我们采用如下基于参数的选择准则：

$$(X_t, \widetilde{\theta}_t) = \arg\max_{\mathbf{x}\in\mathcal{X},\theta\in\mathcal{C}_t(\delta)} \mu(\mathbf{x}^\top\theta). \tag{54}$$

参数型选择准则与奖励型选择准则之间的主要区别已在 [34] 的第3.2节中讨论。整体算法总结见算法7。

B. 引理14的证明

证明。根据模型假设 (16)、函数 g_t (13) 和 $g_t(\theta_t)$(48)，我们有：

$$g_t(\theta_t) - g_t(\widehat{\theta}_t) = \lambda c_\mu\theta_t + \sum_{s=1}^{t-1}\gamma^{t-s-1}\mu(X_s^\top\theta_t)X_s - \sum_{s=1}^{t-1}\gamma^{t-s-1}r_sX_s$$

$$
\begin{aligned}
&= \lambda c_\mu \theta_t + \sum_{s=1}^{t-1} \gamma^{t-s-1} \mu(X_s^\top \theta_t) X_s \\
&\quad - \sum_{s=1}^{t-1} \gamma^{t-s-1} (\mu(X_s^\top \theta_s) + \eta_s) X_s \\
&= \sum_{s=1}^{t-1} \gamma^{t-s-1} (\mu(X_s^\top \theta_t) - \mu(X_s^\top \theta_s)) X_s + \lambda c_\mu \theta_t \\
&\quad - \sum_{s=1}^{t-1} \gamma^{t-s-1} \eta_s X_s.
\end{aligned}
$$

那么，

$$
\begin{aligned}
&\|g_t(\theta_t) - g_t(\widehat{\theta}_t)\|_{H_t^{-1}(\theta_t)} \\
&= \Bigg\| \sum_{s=1}^{t-1} \gamma^{t-s-1} (\mu(X_s^\top \theta_t) - \mu(X_s^\top \theta_s)) X_s \\
&\qquad\qquad + \lambda c_\mu \theta_t - \sum_{s=1}^{t-1} \gamma^{t-s-1} \eta_s X_s \Bigg\|_{H_t^{-1}(\theta_t)} \\
&\le \underbrace{\left\| \sum_{s=1}^{t-1} \gamma^{t-s-1} (\mu(X_s^\top \theta_t) - \mu(X_s^\top \theta_s)) X_s \right\|_{H_t^{-1}(\theta_t)}}_{\text{TERM (A)}} \\
&\quad + \underbrace{\left\| \sum_{s=1}^{t-D-1} \gamma^{t-s-1} \eta_s X_s \right\|_{H_t^{-1}(\theta_t)}}_{\text{TERM (B)}} \\
&\quad + \underbrace{\left\| \sum_{s=t-D}^{t-1} \gamma^{t-s-1} \eta_s X_s - \lambda c_\mu \theta_t \right\|_{H_t^{-1}(\theta_t)}}_{\text{TERM (C)}}.
\end{aligned}
$$

术语 (a)。 由于 $t \in \mathcal{T}(D)$，我们有

$$
\begin{aligned}
&\left\| \sum_{s=1}^{t-1} \gamma^{t-s-1} (\mu(X_s^\top \theta_t) - \mu(X_s^\top \theta_s)) X_s \right\|_{H_t^{-1}(\theta_t)} \\
&= \left\| \sum_{s=1}^{t-D-1} \gamma^{t-s-1} (\mu(X_s^\top \theta_t) - \mu(X_s^\top \theta_s)) X_s \right\|_{H_t^{-1}(\theta_t)} \\
&\le \left\| \sum_{s=1}^{t-D-1} \gamma^{t-s-1} k_\mu X_s^\top (\theta_t - \theta_s) X_s \right\|_{H_t^{-1}(\theta_t)} \\
&\le \sum_{s=1}^{t-D-1} \gamma^{t-s-1} k_\mu \|X_s\|_2 \|(\theta_t - \theta_s)\|_2 \|X_s\|_{H_t^{-1}(\theta_t)} \\
&\le \frac{2L^2 S k_\mu}{\sqrt{\lambda c_\mu}} \frac{\gamma^D}{1-\gamma}.
\end{aligned}
$$

术语 (b)。

$$
\begin{aligned}
\left\| \sum_{s=1}^{t-D-1} \gamma^{t-s-1} \eta_s X_s \right\|_{H_t^{-1}(\theta_t)} &\le \sum_{s=1}^{t-D-1} \gamma^{t-s-1} m \|X_s\|_{H_t^{-1}(\theta_t)} \\
&\le \frac{Lm}{\sqrt{\lambda c_\mu}} \sum_{s=1}^{t-D-1} \gamma^{t-s-1} \\
&\le \frac{Lm}{\sqrt{\lambda c_\mu}} \frac{\gamma^D}{1-\gamma}.
\end{aligned}
$$

术语 (c)。我们定义 $\widetilde{\eta}_s \triangleq \underbrace{\sqrt{\gamma^{t-s-1}}\eta_s, \widetilde{X}_s}_{|m|} \triangleq \sqrt{\gamma^{t-s-1}} X_s$，并注意到 $\forall t \in [T], s \in [t-1], |\gamma^{t-s-1}| \le 1$，于是 $\widetilde{\eta}_s$ $\frac{1}{\lambda c_\mu / m^2 I_d + \sum_{s=1}^{t-1} \mu' \frac{X_s^\top \theta}{m^2} \widetilde{X}_s \widetilde{X}_s^\top}$ $\widetilde{\sigma}_s$ $\widetilde{H}_t\ \theta$ $\triangleq$ 有方差为 的 $\widetilde{H}_{t-D:t}\ \theta$ 上界。令 $(\,)\,()$ 且 $(\,)$ $= \lambda c_\mu I_d + \sum_{s=t-D}^{t-1} \mu' \frac{X_s^\top \theta}{m^2} \widetilde{X}_s \widetilde{X}_s^\top$ $()$，

$$
\begin{aligned}
&\left\| \sum_{s=t-D}^{t-1} \gamma^{t-s-1} \eta_s X_s - \lambda c_\mu \theta_t \right\|_{H_t^{-1}(\theta_t)} \\
&\le \left\| \sum_{s=t-D}^{t-1} \gamma^{t-s-1} \eta_s X_s \right\|_{H_t^{-1}(\theta_t)} + \sqrt{\lambda c_\mu} S \\
&\le \left\| \sum_{s=t-D}^{t-1} \widetilde{\eta}_s \widetilde{X}_s \right\|_{\widetilde{H}_{t-D:t}^{-1}(\theta_t)} + \sqrt{\lambda c_\mu} S,
\end{aligned}
$$

其中 $\widetilde{H}_t(\theta) \succeq \widetilde{H}_{t-D:t}(\theta)$。接下来，我们需要使用定理 8 中重述的自归一化界 [33, Theorem 1]，来界定项 $\|\sum_{s=t-D}^{t-1} \widetilde{\eta}_s X_s\|_{\widetilde{H}_{t-D:t}^{-1}(\theta_t)}$，与引理13的证明类似，我们有

$$
\begin{aligned}
&\left\| \sum_{s=t-D}^{t-1} \gamma^{-s} \eta_s X_s \right\|_{\widetilde{H}_{t-D:t}^{-1}(\theta_t)} \\
&\le \frac{\sqrt{\lambda c_\mu}}{2m} + \frac{2m}{\sqrt{\lambda c_\mu}} \log \left(\frac{\det\left(\widetilde{H}_{t-D:t}\right)^{1/2}}{\delta (\lambda c_\mu)^{d/2}} \right) + \frac{2m}{\sqrt{\lambda c_\mu}} d \log(2) \\
&\le \frac{\sqrt{\lambda c_\mu}}{2m} + \frac{2m}{\sqrt{\lambda c_\mu}} \log \frac{1}{\delta} + \frac{dm}{\sqrt{\lambda c_\mu}} \log \left(1 + \frac{L^2 k_\mu (1-\gamma^{2D})}{\lambda c_\mu d (1-\gamma)} \right) \\
&\quad + \frac{2m}{\sqrt{\lambda c_\mu}} d \log(2).
\end{aligned}
$$

Let $\breve{\beta}_t \triangleq \frac{dm}{\sqrt{\lambda c_\mu}} \log \left(1 + \frac{L^2 k_\mu (1-\gamma^{2D})}{\lambda c_\mu d (1-\gamma)} \right) + \frac{2m}{\sqrt{\lambda c_\mu}} \log \frac{1}{\delta} + \frac{\sqrt{\lambda c_\mu}}{2m} + \frac{2m}{\sqrt{\lambda c_\mu}} d \log(2) + \sqrt{\lambda c_\mu} S$, finally we have,

$$
\begin{aligned}
&\|g_t(\theta_t) - g_t(\widehat{\theta}_t)\|_{H_t^{-1}(\theta_t)} \\
&\le \frac{2L^2 S k_\mu}{\sqrt{\lambda c_\mu}} \frac{\gamma^D}{1-\gamma} + \frac{Lm}{\sqrt{\lambda c_\mu}} \frac{\gamma^D}{1-\gamma} + \breve{\beta}_t,
\end{aligned}
$$

由此完成证明。 □

C. 定理4的证明

证明。令 $R_t = \mu(X_t^{*\top}\theta_t) - \mu(X_t^\top\theta_t)$

$$\begin{aligned}\text{D-REG}_T = \sum_{t=1}^{T} R_t &= \sum_{t\notin\mathcal{T}(D)} R_t + \sum_{t\in\mathcal{T}(D)} R_t \\ &= \Gamma_T D + \sum_{t\in\mathcal{T}(D)} R_t.\end{aligned}$$

对于 $t \in \mathcal{T}(D)$，根据选择准则 (54),

$$\begin{aligned}&R_t = \mu(X_t^{*\top}\theta_t) - \mu(X_t^\top\theta_t)\\ &\le \mu(X_t^\top\widetilde{\theta}_t) - \mu(X_t^\top\widehat{\theta}_t) + \mu(X_t^\top\widehat{\theta}_t) - \mu(X_t^\top\theta_t)\\ &\le \alpha(X_t,\widetilde{\theta}_t,\widehat{\theta}_t)\left|X_t^\top\left(\widetilde{\theta}_t-\widehat{\theta}_t\right)\right| + \alpha(X_t,\theta_t,\widehat{\theta}_t)\left|X_t^\top\left(\theta_t-\widehat{\theta}_t\right)\right|\\ &\le \sqrt{1+2S}\\ &\quad\cdot\left(\alpha(X_t,\widetilde{\theta}_t,\widehat{\theta}_t)\|X_t\|_{G_t^{-1}(\widetilde{\theta}_t,\widehat{\theta}_t)}\left\|g_t(\widetilde{\theta}_t)-g_t(\widehat{\theta}_t)\right\|_{H_t^{-1}(\widetilde{\theta}_t)}\right.\\ &\quad\left.+\,\alpha(X_t,\theta_t,\widehat{\theta}_t)\|X_t\|_{G_t^{-1}(\theta_t,\widehat{\theta}_t)}\left\|g_t(\theta_t)-g_t(\widehat{\theta}_t)\right\|_{H_t^{-1}(\theta_t)}\right).\end{aligned}$$

其中 $\alpha(\mathbf{x}\ \theta_1,\theta_2) = \int_0^1 \mu'(v\mathbf{x}^\top\theta_2 + (1-v)x^\top\theta_1)\mathrm{d}v$，最后一个倒数第二个不等式由中值定理$\mu(\mathbf{x}^\top\theta_1) - \mu(\mathbf{x}^\top\theta_2) = \alpha(\mathbf{x}\ \theta_1,\theta_2)(\mathbf{x}^\top\theta_1 - \mathbf{x}^\top\theta_2)$ 得出。由于 ~~that~~ $\theta_t \in \mathcal{C}_t(\delta)$，且以至少 $1-\delta$ 的概率，$\forall t \in[T], \theta_t \in \mathcal{C}_t(\delta)$，并由并集界可知，以下动态遗憾界以至少 $1-\delta$ 的概率成立，

$$\begin{aligned}\sum_{t\in\mathcal{T}(D)} R_t &\le \sum_{t\in\mathcal{T}(D)} \sqrt{1+2S}\big(\alpha(X_t,\widetilde{\theta}_t,\widehat{\theta}_t)\|X_t\|_{G_t^{-1}(\widetilde{\theta}_t,\widehat{\theta}_t)}\rho_t\\ &\quad+\alpha(X_t,\theta_t,\widehat{\theta}_t)\|X_t\|_{G_t^{-1}(\theta_t,\widehat{\theta}_t)}\rho_t\big)\\ &\le \sqrt{1+2S}\rho_T\left(\sum_{t\in\mathcal{T}(D)}\alpha(X_t,\widetilde{\theta}_t,\widehat{\theta}_t)\|X_t\|_{G_t^{-1}(\widetilde{\theta}_t,\widehat{\theta}_t)}\right.\\ &\quad\left.+\sum_{t\in\mathcal{T}(D)}\alpha(X_t,\theta_t,\widehat{\theta}_t)\|X_t\|_{G_t^{-1}(\theta_t,\widehat{\theta}_t)}\right).\end{aligned}$$

现在我们尝试推导项 ~ $\sum_{t\in\mathcal{T}(D)}\alpha(X_t,\theta_t,\theta_t)\|X_t\|_{G_t^{-1}(\widetilde{\theta}_t,\widehat{\theta}_t)}$ 的上界。

根据 g_t (13) 的定义，我们有

$$\begin{aligned}&g_t(\theta_1) - g_t(\theta_2)\\ &= \lambda c_\mu(\theta_1-\theta_2) + \sum_{s=1}^{t-1}\gamma^{t-s-1}(\mu(X_s^\top\theta_1) - \mu(X_s^\top\theta_2))X_s\\ &= \lambda c_\mu(\theta_1-\theta_2) + \sum_{s=1}^{t-1}\gamma^{t-s-1}\alpha(X_s,\theta_1,\theta_2)X_s^\top X_s(\theta_1-\theta_2)\\ &= \left(\lambda c_\mu + \sum_{s=1}^{t-1}\gamma^{t-s-1}\alpha(X_s,\theta_1,\theta_2)X_s^\top X_s\right)(\theta_1-\theta_2).\end{aligned}$$

然后根据 G_t (49) 的定义，我们知道 $G_t(\theta_1,\theta_2) = \lambda c_\mu + \sum_{s=1}^{t-1}\gamma^{t-s-1}\alpha\ X_s\ \theta_1\ \theta_2\ X_s^\top X_s.$ =(,),

这意味着 ~~$G_t(\widetilde{\theta}_t,\widehat{\theta}_t) = \lambda c_\mu I_d + \sum_{s=1}^{t-1}\gamma^{t-s-1}\alpha(X_s,\widetilde{\theta}_t,\widehat{\theta}_t)X_sX_s^\top$~~，如果令 $\widetilde{X}_s = \sqrt{\alpha(X_s,\widetilde{\theta}_t,\widehat{\theta}_t)}X_s$，那么

$$\begin{aligned}&\sum_{t\in\mathcal{T}(D)}\alpha(X_t,\widetilde{\theta}_t,\widehat{\theta}_t)\|X_t\|_{G_t^{-1}(\widetilde{\theta}_t,\widehat{\theta}_t)}\\ &\le \sqrt{\sum_{t=1}^{T}\alpha(X_t,\widetilde{\theta}_t,\widehat{\theta}_t)}\sqrt{\sum_{t=1}^{T}\alpha(X_t,\widetilde{\theta}_t,\widehat{\theta}_t)\|X_t\|^2_{G_t^{-1}(\widetilde{\theta}_t,\widehat{\theta}_t)}}\\ &\le \sqrt{k_\mu T}\sqrt{\sum_{t=1}^{T}\left\|\widetilde{X}_t\right\|^2_{G_t^{-1}(\widetilde{\theta}_t,\widehat{\theta}_t)}}.\end{aligned}$$

然后对于项 $\sqrt{\sum_{t=1}^{T}\|\widetilde{X}_t\|^2_{G_t^{-1}(\widetilde{\theta}_t,\widehat{\theta}_t)}}$，我们可以直接使用引理 9 对其进行界定，令 $C_T^{\text{PWSCB}} \triangleq \sqrt{2k_\mu}\max\{1, L^2k_\mu/(\lambda c_\mu)\}dT$，我们有

$$\begin{aligned}&\sqrt{k_\mu T}\sqrt{\sum_{t=1}^{T}\left\|\widetilde{X}_t\right\|^2_{G_t^{-1}(\widetilde{\theta}_t,\widehat{\theta}_t)}}\\ &\le C_T^{\text{PWSCB}}\sqrt{T\log\frac{1}{\gamma} + \log\left(1+\frac{L^2k_\mu}{\lambda c_\mu d(1-\gamma)}\right)}.\end{aligned}$$

^用相同的方法可以给项 $\sum_{t\in\mathcal{T}(D)}\alpha(X_t,\theta_t,\theta_t)\|X_t\|_{G_t^{-1}(\theta_t,\widehat{\theta}_t)}$ 建立上界，得到:

$$\begin{aligned}&\sum_{t\in\mathcal{T}(D)}\alpha(X_t,\theta_t,\widehat{\theta}_t)\|X_t\|_{G_t^{-1}(\theta_t,\widehat{\theta}_t)}\\ &\le C_T^{\text{PWSCB}}\sqrt{T\log\frac{1}{\gamma} + \log\left(1+\frac{L^2k_\mu}{\lambda c_\mu d(1-\gamma)}\right)}.\end{aligned}$$

结合这两个上界并令 $\delta = 1/T$，我们以至少 $1-1/T$ 的概率得到如下遗憾界:

$$\begin{aligned}&\text{D-REG}_T \le \Gamma_T D\\ &+ 2\sqrt{1+2S}\rho_T C_T^{\text{PWSCB}}\sqrt{T\log\frac{1}{\gamma} + \log\left(1+\frac{L^2k_\mu}{\lambda c_\mu d(1-\gamma)}\right)},\end{aligned}$$

where $\rho_t = \frac{2L^2Sk_\mu}{\sqrt{\lambda c_\mu}}\frac{\gamma^D}{1-\gamma} + \frac{Lm}{\sqrt{\lambda c_\mu}}\frac{\gamma^D}{1-\gamma} + \breve{\beta}_t$ and $\breve{\beta}_t = \frac{dm}{\sqrt{\lambda c_\mu}}\log\left(1+\frac{L^2k_\mu(1-\gamma^{2D})}{\lambda c_\mu d(1-\gamma)}\right) + \frac{\sqrt{\lambda c_\mu}}{2m} + \frac{2m}{\sqrt{\lambda c_\mu}}\log(T) + \frac{2m}{\sqrt{\lambda c_\mu}}d\log(2) + \sqrt{\lambda c_\mu}S$. Since there is a $T\sqrt{\log(1/\gamma)}$ term in the regret bound, which means that we cannot let γ close to 0, so we set $\gamma \ge 1/2$, then we have $\log(1/\gamma) \le 2\log(2)(1-\gamma)$. Then, we set $D = \log(T)/\log(1/\gamma)$, noticing that $0 < 1/\gamma - 1 < 1$ and using $\log(1+x) \ge x/2$ for $0 < x < 1$, we have

$$\log\frac{1}{\gamma} = \log(1+1/\gamma-1) \ge \frac{1-\gamma}{2\gamma}.$$

因此，我们有 $D \le {}^{2\gamma}\!/_{\log(T)1-\gamma}$。然后，忽略时间范围 T 中的对数因子，并令 $\lambda = d\log(T)/c_\mu$，最终得到:

$$\begin{aligned}&\text{D-REG}_T\\ &\le \widetilde{\mathcal{O}}\left(\frac{1}{1-\gamma}\Gamma_T + \left(\frac{1}{\sqrt{d}}\frac{1}{1-\gamma}\frac{1}{T} + \sqrt{d}\right)\sqrt{d(1-\gamma)}T\right)\\ &\le \widetilde{\mathcal{O}}\left(\frac{1}{1-\gamma}\Gamma_T + \frac{1}{\sqrt{1-\gamma}} + d\sqrt{(1-\gamma)}T\right).\end{aligned}$$

当 $\Gamma_T < d/\sqrt{T}$ ¯(这对应于少量的非平稳性）时，我们只需设置_$\gamma = 1-1/T$，即可获得 $\widetilde{\mathcal{O}}(d\sqrt{T})$ 遗憾界。此外，对于 $\Gamma_T > d/\sqrt{T}$ 的非退化情形，我们将折扣

因子最优地设为 $1-\gamma = (\Gamma_T/(dT))^{2/3}$，从而达到 $\widetilde{\mathcal{O}}(d^{2/3}\Gamma_T^{1/3}T^{2/3})$ 遗憾，由此完成证明。 □

APPENDIX E
ANALYSIS OF WEIGHTUCRL

A. 引理4的证明

证明。固定 $h \in [H]$，根据奖励模型假设（22）和估计量（23），奖励估计的估计误差可分解为

$$\begin{aligned}
&\widehat{\theta}_h^k - \theta_h^k \\
&= \left(\Lambda_h^{k-1}\right)^{-1}\left(\sum_{j=1}^{k-1} w_{k-1,j} r_h^j(s_h^j,a_h^j)\phi(s_h^j,a_h^j)\right) - \theta_h^k \\
&= \left(\Lambda_h^{k-1}\right)^{-1}\left(\sum_{j=1}^{k-1} w_{k-1,j}\phi(s_h^j,a_h^j)^\top\theta_h^j\phi(s_h^j,a_h^j)\right) - \theta_h^k \\
&= \left(\Lambda_h^{k-1}\right)^{-1}\left(\sum_{j=1}^{k-1} w_{k-1,j}\phi(s_h^j,a_h^j)\phi(s_h^j,a_h^j)^\top\theta_h^j\right) \\
&\quad - \left(\Lambda_h^{k-1}\right)^{-1}\left(\lambda_\theta I_d + \sum_{j=1}^{k-1} w_{k-1,j}\phi(s_h^j,a_h^j)\phi(s_h^j,a_h^j)^\top\right)\theta_h^k \\
&= \underbrace{\left(\Lambda_h^{k-1}\right)^{-1}\left(\sum_{j=1}^{k-1} w_{k-1,j}\phi(s_h^j,a_h^j)\phi(s_h^j,a_h^j)^\top\left(\theta_h^j-\theta_h^k\right)\right)}_{\texttt{bias part}} \\
&\quad - \underbrace{\left(\Lambda_h^{k-1}\right)^{-1}\lambda_\theta\theta_h^k}_{\texttt{variance part}}.
\end{aligned}$$

然后，根据柯西-施瓦茨不等式，我们知道，对于任意 $s\in\mathcal{S}$, $a\in\mathcal{A}$,

$$\left|\phi(s,a)^\top\left(\widehat{\theta}_h^k-\theta_h^k\right)\right| = \|\phi(s,a)\|_{\left(\Lambda_h^{k-1}\right)^{-1}}\left(A_h^k+B_h^k\right), \tag{55}$$

其中

$$A_h^k = \left\|\sum_{j=1}^{k-1} w_{k-1,j}\phi(s_h^j,a_h^j)\phi(s_h^j,a_h^j)^\top\left(\theta_h^j-\theta_h^k\right)\right\|_{\left(\Lambda_h^{k-1}\right)^{-1}},$$

$$B_h^k = \left\|\lambda_\theta\theta_h^k\right\|_{\left(\Lambda_h^{k-1}\right)^{-1}}.$$

上述两项可以分别进行界定，

术语 A_h^k。 第一步是提取k的变化量 parameter θ_h as follows,

$$\begin{aligned}
&\left\|\sum_{j=1}^{k-1} w_{k-1,j}\phi(s_h^j,a_h^j)\phi(s_h^j,a_h^j)^\top\left(\theta_h^j-\theta_h^k\right)\right\|_{\left(\Lambda_h^{k-1}\right)^{-1}} \\
&= \left\|\sum_{j=1}^{k-1} w_{k-1,j}\phi(s_h^j,a_h^j)\phi(s_h^j,a_h^j)^\top\sum_{p=j}^{k-1}\left(\theta_h^p-\theta_h^{p+1}\right)\right\|_{\left(\Lambda_h^{k-1}\right)^{-1}} \\
&= \left\|\sum_{p=1}^{k-1}\sum_{j=1}^{p} w_{k-1,j}\phi(s_h^j,a_h^j)\phi(s_h^j,a_h^j)^\top\left(\theta_h^p-\theta_h^{p+1}\right)\right\|_{\left(\Lambda_h^{k-1}\right)^{-1}} \\
&\le \sum_{p=1}^{k-1}\left\|\sum_{j=1}^{p} w_{k-1,j}\phi(s_h^j,a_h^j)\left\|\phi(s_h^j,a_h^j)\right\|_2\left\|\theta_h^p-\theta_h^{p+1}\right\|_2\right\|_{\left(\Lambda_h^{k-1}\right)^{-1}} \\
&\le L_\phi\sum_{p=1}^{k-1}\sum_{j=1}^{p} w_{k-1,j}\left\|\phi(s_h^j,a_h^j)\right\|_{\left(\Lambda_h^{k-1}\right)^{-1}}\left\|\theta_h^p-\theta_h^{p+1}\right\|_2,
\end{aligned}$$

以及术语 $\sum_{j=1}^{p} w_{k-1,\ j}\left\|\phi(s_h^j,\ a_h^j)\right\|_{(\Lambda_h^{k-1})^{-1}}$ 可以进一步推导出关于权重 $w_{k-1,\ j}$ 的表达式，如下:

$$\begin{aligned}
&\sum_{j=1}^{p} w_{k-1,j}\left\|\phi(s_h^j,a_h^j)\right\|_{\left(\Lambda_h^{k-1}\right)^{-1}} \\
&\le \sqrt{\sum_{j=1}^{p} w_{k-1,j}}\sqrt{\sum_{j=1}^{p} w_{k-1,j}\left\|\phi(s_h^j,a_h^j)\right\|^2_{\left(\Lambda_h^{k-1}\right)^{-1}}} \\
&\le \sqrt{d}\sqrt{\sum_{j=1}^{p} w_{k-1,j}}.
\end{aligned}$$

上面的第一步由柯西–施瓦茨不等式得出。此外，最后一步由引理25得出，令$X_j = \sqrt{w_{k-1,j}}\phi(s_h^j,a_h^j)$ 和 $U_{k-1} = \Lambda_h^{k-1}$，这意味着对于术语 1，我们有

$$\begin{aligned}
&\left\|\sum_{j=1}^{k-1} w_{k-1,j}\phi(s_h^j,a_h^j)\phi(s_h^j,a_h^j)^\top\left(\theta_h^j-\theta_h^k\right)\right\|_{\left(\Lambda_h^{k-1}\right)^{-1}} \\
&\le L_\phi\sqrt{d}\sum_{p=1}^{k-1}\sqrt{\sum_{j=1}^{p} w_{k-1,j}}\left\|\theta_h^p-\theta_h^{p+1}\right\|_2,
\end{aligned} \tag{56}$$

术语 B_h^k。

$$\left\|\lambda_\theta\theta_h^k\right\|_{\left(\Lambda_h^{k-1}\right)^{-1}} \le \frac{\lambda_\theta}{\sqrt{\lambda_{\min}(\Lambda_h^{k-1})}}\left\|\theta_h^k\right\|_2 \le \sqrt{\lambda_\theta}S_\theta. \tag{57}$$

将方程 (56) 和方程 (57) 代入方程 (55)，得到

$$\begin{aligned}
&\left\|\widehat{\theta}_h^k-\theta_h^k\right\|_{\Lambda_h^{k-1}} \\
&\le L_\phi\sqrt{d}\sum_{p=1}^{k-1}\sqrt{\sum_{j=1}^{p} w_{k-1,j}}\left\|\theta_h^p-\theta_h^{p+1}\right\|_2 + \sqrt{\lambda_\theta}S_\theta,
\end{aligned}$$

进一步地，我们有

$$\begin{aligned}
&\left|\phi(s,a)^\top\left(\widehat{\theta}_h^k-\theta_h^k\right)\right| \\
&\le \|\phi(s,\ a)\|_{\left(\Lambda_h^{k-1}\right)^{-1}} \\
&\quad\cdot\left(L_\phi\sqrt{d}\sum_{p=1}^{k-1}\sqrt{\sum_{j=1}^{p} w_{k-1,j}}\left\|\theta_h^p-\theta_h^{p+1}\right\|_2 + \sqrt{\lambda_\theta}S_\theta\right) \\
&\le L_\phi^2\sqrt{\frac{d}{\lambda_\theta}}\sum_{p=1}^{k-1}\sqrt{\sum_{j=1}^{p} w_{k-1,j}}\left\|\theta_h^p-\theta_h^{p+1}\right\|_2 \\
&\qquad + \beta_\theta\|\phi(s,a)\|_{\left(\Lambda_h^{k-1}\right)^{-1}},
\end{aligned}$$

where $\|\phi(s,a)\|_{\left(\Lambda_h^{k-1}\right)^{-1}} \le \|\phi(s,a)\|_2/\sqrt{\lambda_\theta}$ and $\beta_\theta \triangleq \sqrt{\lambda_\theta}S_\theta$, which completes the proof. □

B. 引理*5*的证明

证明。 固定 $h \in [H]$，根据模型假设 (22) 和估计量 (25)，我们有

$$\begin{aligned}\widehat{\mathbf{w}}_h^k &= \left(\Sigma_h^{k-1}\right)^{-1}\left(\sum_{j=1}^{k-1}\alpha_{k-1,j}V_{h+1}^j(s_{h+1}^j)\psi_{h+1}^j\left(s_h^j,a_h^j\right)\right)\\ &= \left(\Sigma_h^{k-1}\right)^{-1}\left(\sum_{j=1}^{k-1}\alpha_{k-1,j}\left(\psi_{h+1}^j\left(s_h^j,a_h^j\right)^\top\mathbf{w}_h^j+\eta_{h+1}^j\right)\right.\\ &\qquad\left.\cdot\,\psi_{h+1}^j\left(s_h^j,a_h^j\right)\right),\end{aligned} \tag{58}$$

其中，我们将噪声定义为 $\eta_{h+1}^j \triangleq V_{h+1}^j\left(s_{h+1}^j\right) - \left[\mathbb{P}_h^j V_{h+1}^j\right]\left(s_h^j,a_h^j\right)$，并且有

$$\begin{aligned}\mathbf{w}_h^k &= \left(\Sigma_h^{k-1}\right)^{-1}\left(\lambda_{\mathbf{w}}I_d\right.\\ &\left.+\sum_{j=1}^{k-1}\alpha_{k-1,j}\psi_{h+1}^j\left(s_h^j,a_h^j\right)\psi_{h+1}^j\left(s_h^j,a_h^j\right)^\top\right)\mathbf{w}_h^k,\end{aligned} \tag{59}$$

结合方程 (58) 和方程 (59)，可得估计误差

$$\begin{aligned}&\widehat{\mathbf{w}}_h^k-\mathbf{w}_h^k\\ &= \underbrace{\left(\Sigma_h^{k-1}\right)^{-1}\left(\sum_{j=1}^{k-1}\alpha_{k-1,j}\psi_{h+1}^j\left(s_h^j,a_h^j\right)\psi_{h+1}^j\left(s_h^j,a_h^j\right)^\top\right.}_{\texttt{bias part}}\\ &\qquad\underbrace{\left.\cdot\left(\mathbf{w}_h^j-\mathbf{w}_h^k\right)\right)}_{\texttt{bias part}}\\ &\quad+\underbrace{\left(\Sigma_h^{k-1}\right)^{-1}\left(\sum_{j=1}^{k-1}\alpha_{k-1,j}\eta_{h+1}^j\psi_{h+1}^j\left(s_h^j,a_h^j\right)-\lambda_{\mathbf{w}}\mathbf{w}_h^k\right)}_{\texttt{variance part}}.\end{aligned}$$

然后，根据柯西-施瓦茨不等式，对于任意 $s \in \mathcal{S}$，$a \in \mathcal{A}$，我们有

$$\begin{aligned}&\left|\psi_{h+1}^k(s,a)^\top\left(\widehat{\mathbf{w}}_h^k-\mathbf{w}_h^k\right)\right|\\ &\le \left\|\psi_{h+1}^k(s,a)\right\|_{\left(\Sigma_h^{k-1}\right)^{-1}}\left(C_h^k+D_h^k\right),\end{aligned} \tag{60}$$

其中

$$C_h^k = \left\|\sum_{j=1}^{k-1}\alpha_{k-1,j}\cdot\psi_{h+1}^j\left(s_h^j,a_h^j\right)\psi_{h+1}^j\left(s_h^j,a_h^j\right)^\top\left(\mathbf{w}_h^j-\mathbf{w}_h^k\right)\right\|_{\left(\Sigma_h^{k-1}\right)^{-1}}$$

$$D_h^k = \left\|\sum_{j=1}^{k-1}\alpha_{k-1,j}\eta_{h+1}^j\psi_{h+1}^j\left(s_h^j,a_h^j\right)-\lambda_{\mathbf{w}}\mathbf{w}_h^k\right\|_{\left(\Sigma_h^{k-1}\right)^{-1}}.$$

The above two terms can be bounded separately, as summarized in the following two lemmas,

引理15。 对于任意 $k \in [K]$，我们有

$$C_h^k \le HL_\psi\sqrt{d}\sum_{p=1}^{k-1}\sqrt{\sum_{j=1}^{p}\alpha_{k-1,j}}\left\|\mathbf{w}_h^p-\mathbf{w}_h^{p+1}\right\|_2.$$

引理16。 如果 $\forall k,j \in [K], \alpha_{k-1,\,j} \le 1$，对于任意 $\delta \in (0,1)$，以至少 $1-\delta$ 的概率，以下结论对所有$k \in [K]$均成立，

$$D_h^k \le H\sqrt{\frac{1}{2}\log\frac{1}{\delta}+\frac{d}{4}\log\left(1+\frac{H^2L_\psi^2\sum_{j=1}^{k-1}\alpha_{k-1,j}}{\lambda_{\mathbf{w}}d}\right)}+\sqrt{\lambda_{\mathbf{w}}}S_{\mathbf{w}}.$$

根据不等式 (60)、引理15、引理16 和 $\left\|\psi_{h+1}^k(s,a)\right\|_{\left(\Sigma_h^{k-1}\right)^{-1}} \le \left\|\psi_{h+1}^k(s,a)\right\|_2/\sqrt{\lambda_{\mathbf{w}}} \le HL_\psi/\sqrt{\lambda_{\mathbf{w}}}$，以至少 $1-\delta$ 的概率，对所有 $k \in [K]$，以下结论成立，

$$\begin{aligned}&\left|\psi_{h+1}^k(s,a)^\top\left(\widehat{\mathbf{w}}_h^k-\mathbf{w}_h^k\right)\right|\\ &\le \Gamma_{h,\mathbf{w}}^{k-1}+\beta_{\mathbf{w}}^{k-1}\left\|\psi_{h+1}^k(s,a)\right\|_{\left(\Sigma_h^{k-1}\right)^{-1}},\end{aligned}$$

其中

$$\Gamma_{h,\mathbf{w}}^{k-1} \triangleq H^2L_\psi^2\sqrt{\frac{d}{\lambda_{\mathbf{w}}}}\sum_{p=1}^{k-1}\sqrt{\sum_{j=1}^{p}\alpha_{k-1,j}}\left\|\mathbf{w}_h^p-\mathbf{w}_h^{p+1}\right\|_2$$

$$\beta_{\mathbf{w}}^{k-1} \triangleq H\sqrt{\frac{1}{2}\log\frac{1}{\delta}+\frac{d}{4}\log\left(1+\frac{H^2L_\psi^2\sum_{j=1}^{k-1}\alpha_{k-1,j}}{\lambda_{\mathbf{w}}d}\right)}+\sqrt{\lambda_{\mathbf{w}}}S_{\mathbf{w}},$$

至此，证明完成。 □

引理*15*的证明。 第一步是提取参数$\mathbf{w}_h^k$ 的变化量，如下所示，

$$\begin{aligned}&\left\|\sum_{j=1}^{k-1}\alpha_{k-1,j}\psi_{h+1}^j\left(s_h^j,a_h^j\right)\cdot\psi_{h+1}^j\left(s_h^j,a_h^j\right)^\top\left(\mathbf{w}_h^j-\mathbf{w}_h^k\right)\right\|_{\left(\Sigma_h^{k-1}\right)^{-1}}\\ &= \left\|\sum_{j=1}^{k-1}\alpha_{k-1,j}\psi_{h+1}^j\left(s_h^j,a_h^j\right)\cdot\psi_{h+1}^j\left(s_h^j,a_h^j\right)^\top\sum_{p=j}^{k-1}\left(\mathbf{w}_h^p-\mathbf{w}_h^{p+1}\right)\right\|_{\left(\Sigma_h^{k-1}\right)^{-1}}\\ &= \left\|\sum_{p=1}^{k-1}\sum_{j=1}^{p}\alpha_{k-1,j}\psi_{h+1}^j\left(s_h^j,a_h^j\right)\cdot\psi_{h+1}^j\left(s_h^j,a_h^j\right)^\top\left(\mathbf{w}_h^p-\mathbf{w}_h^{p+1}\right)\right\|_{\left(\Sigma_h^{k-1}\right)^{-1}}\\ &\le \sum_{p=1}^{k-1}\left\|\sum_{j=1}^{p}\alpha_{k-1,j}\psi_{h+1}^j\left(s_h^j,a_h^j\right)\cdot\psi_{h+1}^j\left(s_h^j,a_h^j\right)^\top\left(\mathbf{w}_h^p-\mathbf{w}_h^{p+1}\right)\right\|_{\left(\Sigma_h^{k-1}\right)^{-1}}\end{aligned}$$

$$\leq HL_{\psi}\sum_{p=1}^{k-1}\sum_{j=1}^{p}\alpha_{k-1,\ j}\underbrace{\left\|\psi_{h+1}^{j}\left(s_h^j,a_h^j\right)\right\|}_{\|(\Sigma_h^k)^{-1}\|^{-1}}\cdot\left\|\mathbf{w}_h^p-\mathbf{w}_h^{p+1}\right\|_2,$$

以及项 $\sum_{j=1}^{p}\alpha_{k-1,\ j}\left\|\psi_{h+1}^{j}\left(s_h^j,a_h^j\right)\right\|_{(\Sigma_h^{k-1})^{-1}}$，可以进一步推导出关于权重 $\alpha_{k-1,\ j}$ 的表达式，如下所示，

$$\begin{aligned}&\sum_{j=1}^{p}\alpha_{k-1,j}\left\|\psi_{h+1}^{j}\left(s_h^j,a_h^j\right)\right\|_{\left(\Sigma_h^{k-1}\right)^{-1}}\\&\leq\sqrt{\sum_{j=1}^{p}\alpha_{k-1,j}}\sqrt{\sum_{j=1}^{p}\alpha_{k-1,j}\left\|\psi_{h+1}^{j}\left(s_h^j,a_h^j\right)\right\|_{\left(\Sigma_h^{k-1}\right)^{-1}}^2}\\&\leq\sqrt{d}\sqrt{\sum_{j=1}^{p}\alpha_{k-1,j}}.\end{aligned}$$

上文中，倒数第二步依据柯西-施瓦茨不等式成立。此外，令 $X_j=\sqrt{\alpha_{k-1,\ j}}\psi_{h+1}^{j}\left(s_h^j,a_h^j\right)$ 和 $U_{k-1}=\Sigma_h^{k-1}$，最后一步由引理25得出。因此，我们完成了证明。 □

引理*16*的证明。

$$\begin{aligned}&\left\|\sum_{j=1}^{k-1}\alpha_{k-1,j}\eta_{h+1}^{j}\psi_{h+1}^{j}\left(s_h^j,a_h^j\right)-\lambda_{\mathbf{w}}\mathbf{w}_h^k\right\|_{\left(\Sigma_h^{k-1}\right)^{-1}}\\&\leq\left\|\sum_{j=1}^{k-1}\alpha_{k-1,j}\eta_{h+1}^{j}\psi_{h+1}^{j}\left(s_h^j,a_h^j\right)\right\|_{\left(\Sigma_h^{k-1}\right)^{-1}}+\left\|\lambda_{\mathbf{w}}\mathbf{w}_h^k\right\|_{\left(\Sigma_h^{k-1}\right)^{-1}}\\&\leq\left\|\sum_{j=1}^{k-1}\alpha_{k-1,j}\eta_{h+1}^{j}\psi_{h+1}^{j}\left(s_h^j,a_h^j\right)\right\|_{\left(\Sigma_h^{k-1}\right)^{-1}}+\sqrt{\lambda_{\mathbf{w}}}S_{\mathbf{w}}.\end{aligned}$$

令 $\widetilde{\eta}_{h+1}^{j}\triangleq\sqrt{\alpha_{k-1,\ j}}\eta_{h+1}^{j}$ 和 $X_j\triangleq\sqrt{\alpha_{k-1,\ j}}\psi_{h+1}^{j}\left(s_h^j,a_h^j\right)$，则我们注意到，由于奖励 $r\in[0,1]$，且 $\alpha_{k-1,\ j}\leq 1$，噪声 $\widetilde{\eta}_{h+1}^{j}$ 有界于：

$$\widetilde{\eta}_{h+1}^{j}=\sqrt{\alpha_{k-1,j}}\left(V_{h+1}^{j}\left(s_{h+1}^{j}\right)-\left[\mathbb{P}_hV_{h+1}^{j}\right]\left(s_h^j,a_h^j\right)\right)\leq H,$$

根据引理23，我们发现噪声 $\widetilde{\eta}_{h+1}^{j}$ 是 $\frac{H}{2}$ -次高斯的。然后，根据定理7，以至少 $1-\delta$ 的概率，以下结论对所有 $k\in[K]$ 均成立。

$$\begin{aligned}&\left\|\sum_{j=1}^{k-1}\widetilde{\eta}_{h+1}^{j}X_j\right\|_{\left(\Sigma_h^{k-1}\right)^{-1}}\\&\leq\sqrt{\frac{H^2}{2}\log\left(\frac{\det(\Sigma_h^{k-1})^{\frac{1}{2}}\det(\Sigma_{0,h})^{-\frac{1}{2}}}{\delta}\right)}\end{aligned}$$

其中

$$\begin{aligned}\det(\Sigma_h^{k-1})&\leq\left(\frac{\operatorname{trace}(\Sigma_h^{k-1})}{d}\right)^d\\&=\left(\frac{d\lambda_{\mathbf{w}}+\sum_{j=1}^{k-1}\left\|\alpha_{k-1,j}\psi_{h+1}^{j}\left(s_h^j,a_h^j\right)\right\|_2^2}{d}\right)^d\\&=\left(\frac{d\lambda_{\mathbf{w}}+H^2L_{\psi}^2\sum_{j=1}^{k-1}\alpha_{k-1,j}}{d}\right)^d\\\det(\Sigma_{0,h})&\leq{\lambda_{\mathbf{w}}}^d,\end{aligned}$$

因此我们有

$$\begin{aligned}&\left\|\sum_{j=1}^{k-1}\alpha_{k-1,j}\eta_{h+1}^{j}\psi_{h+1}^{j}\left(s_h^j,a_h^j\right)\right\|_{\left(\Sigma_h^{k-1}\right)^{-1}}\\&\leq H\sqrt{\frac{1}{2}\log\frac{1}{\delta}+\frac{d}{4}\log\left(1+\frac{H^2L_{\psi}^2\sum_{j=1}^{k-1}\alpha_{k-1,j}}{\lambda_{\mathbf{w}}d}\right)}.\end{aligned}$$

由此完成证明。 □

C. 定理 *5* 的证明

证明。 为证明该定理，我们首先引入以下引理

引理17。 我们将模型预测误差定义为

$$E_h^k(s,a)=r_h^k(s,a)+\mathbb{P}_h^kV_{h+1}^k(s,a)-Q_h^k(s,a),\qquad(61)$$

那么，以至少 $1-2\delta$ 的概率，以下结论对所有 $k\in[K]$、$h\in[H]$ 和 $\forall s\in\mathcal{S}$、$a\in\mathcal{A}$ 均成立：

$$\begin{aligned}&-2\beta_{\theta}\left\|\phi(s,a)\right\|_{\left(\Lambda_h^{k-1}\right)^{-1}}-2\beta_{\mathbf{w}}^{k-1}\left\|\psi_{h+1}^{k}(s,a)\right\|_{\left(\Sigma_h^{k-1}\right)^{-1}}\\&-\Gamma_{h,\theta}^{k-1}-\Gamma_{h,\mathbf{w}}^{k-1}\leq E_h^k(s,a)\leq\Gamma_{h,\theta}^{k-1}+\Gamma_{h,\mathbf{w}}^{k-1}.\end{aligned}$$

我们还可以通过以下引理将动态遗憾与模型预测误差联系起来。

引理18。 对于策略 $\{\pi_h^k\}_{h\in[H],\ k\in[K]}$，其中 $a_h^k=\arg\max_{a\in\mathcal{A}}Q_h^k(s_h^k,a)$，以及最优策略和 $\delta\in(0,1)$，以下分解以至少 $1-2\delta$ 的概率成立，

$$\begin{aligned}\text{D-REG}_T\leq&\sum_{k=1}^{K}\sum_{h=1}^{H}\left(\mathbb{E}_{\pi_{*,h}^k}\left[E_h^k(s_h^k,a_h^k)\right]-E_h^k(s_h^k,a_h^k)\right)\\&+4H\sqrt{2T\log(1/\delta)}.\end{aligned}$$

基于引理18，并注意到 $\forall s \in \mathcal{S}$、$a \in \mathcal{A}$,$\left|E_h^k(s, a)\right| \leq 2H$，我们有

$$
\begin{aligned}
&\mathbb{E}_{\pi_{*,h}^k}\left[E_h^k(s_h^k, a_h^k)\right] - E_h^k(s_h^k, a_h^k) \\
\leq\ & \min\left\{4H, 2\Gamma_{h,\theta}^{k-1} + 2\Gamma_{h,\mathbf{w}}^{k-1} + 2\beta_\theta \left\|\phi(s_h^k, a_h^k)\right\|_{\left(\Lambda_h^{k-1}\right)^{-1}}\right. \\
&\left. + 2\beta_{\mathbf{w}}^{k-1} \left\|\psi_{h+1}^k\left(s_h^k, a_h^k\right)\right\|_{\left(\Sigma_h^{k-1}\right)^{-1}}\right\} \\
\leq\ & 2\Gamma_{h,\theta}^{k-1} + 2\Gamma_{h,\mathbf{w}}^{k-1} + \min\left\{4H, 2\beta_\theta \left\|\phi(s_h^k, a_h^k)\right\|_{\left(\Lambda_h^{k-1}\right)^{-1}}\right\} \\
&+ \min\left\{4H, 2\beta_{\mathbf{w}}^{k-1} \left\|\psi_{h+1}^k\left(s_h^k, a_h^k\right)\right\|_{\left(\Sigma_h^{k-1}\right)^{-1}}\right\} \\
\leq\ & 2\Gamma_{h,\theta}^{k-1} + 2\Gamma_{h,\mathbf{w}}^{k-1} + 4H\beta_\theta \min\left\{1, \left\|\phi(s_h^k, a_h^k)\right\|_{\left(\Lambda_h^{k-1}\right)^{-1}}\right\} \\
&+ 4\beta_{\mathbf{w}}^{k-1} \min\left\{1, \left\|\psi_{h+1}^k\left(s_h^k, a_h^k\right)\right\|_{\left(\Sigma_h^{k-1}\right)^{-1}}\right\},
\end{aligned}
$$

最后一个不等式由 $\beta_\theta > 1$，$\beta_{\mathbf{w}}^{k-1} \geq H$ 得出。因此，以至少 $1-4\delta$ 的概率，

$$
\begin{aligned}
\text{D-REG}_T \leq\ & \underbrace{4H\beta_\theta \sum_{k=1}^{K}\sum_{h=1}^{H} \min\left\{1, \left\|\phi(s_h^k, a_h^k)\right\|_{\left(\Lambda_h^{k-1}\right)^{-1}}\right\}}_{\texttt{variance part1}} \\
&+ \underbrace{4\sum_{k=1}^{K}\sum_{h=1}^{H} \beta_{\mathbf{w}}^{k-1} \min\left\{1, \left\|\psi_{h+1}^k\left(s_h^k, a_h^k\right)\right\|_{\left(\Sigma_h^{k-1}\right)^{-1}}\right\}}_{\texttt{variance part2}} \\
&+ \underbrace{2\sum_{k=1}^{K}\sum_{h=1}^{H} \Gamma_{h,\theta}^{k-1} + 2\sum_{k=1}^{K}\sum_{h=1}^{H} \Gamma_{h,\mathbf{w}}^{k-1}}_{\texttt{bias part}} + 4H\sqrt{2T\log(1/\delta)}.
\end{aligned}
$$

偏差。现在我们设 $w_{k,\,j} = \gamma^{k-j}$，$\gamma \in (0,\ 1)$，

$$
\begin{aligned}
&2\sum_{k=1}^{K}\sum_{h=1}^{H} \Gamma_{h,\theta}^{k-1} \\
=\ & 2L_\phi^2\sqrt{\frac{d}{\lambda_\theta}} \sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{p=1}^{k-1} \sqrt{\sum_{j=1}^{p} w_{k-1,j}} \left\|\theta_h^p - \theta_h^{p+1}\right\|_2 \\
=\ & 2L_\phi^2\sqrt{\frac{d}{\lambda_\theta}} \sum_{p=1}^{K-1}\sum_{h=1}^{H}\sum_{k=p+1}^{K} \sqrt{\sum_{j=1}^{p} w_{k-1,j}} \left\|\theta_h^p - \theta_h^{p+1}\right\|_2 \\
=\ & 2L_\phi^2\sqrt{\frac{d}{\lambda_\theta}} \sum_{p=1}^{K-1}\sum_{h=1}^{H}\sum_{k=p+1}^{K} \gamma^{\frac{k-1}{2}} \sqrt{\sum_{j=1}^{p} \gamma^{-j}} \left\|\theta_h^p - \theta_h^{p+1}\right\|_2 \\
=\ & 2L_\phi^2\sqrt{\frac{d}{\lambda_\theta}} \sum_{p=1}^{K-1}\sum_{h=1}^{H} \frac{\gamma^{\frac{p}{2}} - \gamma^{\frac{K}{2}}}{1-\gamma^{\frac{1}{2}}} \sqrt{\frac{\gamma^{-p}-1}{1-\gamma}} \left\|\theta_h^p - \theta_h^{p+1}\right\|_2 \\
\leq\ & 4L_\phi^2\sqrt{\frac{d}{\lambda_\theta}} \frac{1}{(1-\gamma)^{3/2}} \sum_{p=1}^{K-1}\sum_{h=1}^{H} \left\|\theta_h^p - \theta_h^{p+1}\right\|_2, \qquad (62)
\end{aligned}
$$

然后我们设定 $\alpha_{k,j} = \gamma^{k-j}, \gamma \in (0,1)$，并且有

$$
\begin{aligned}
&2\sum_{k=1}^{K}\sum_{h=1}^{H} \Gamma_{h,\mathbf{w}}^{k-1} \\
=\ & 2H^2L_\psi^2\sqrt{\frac{d}{\lambda_{\mathbf{w}}}} \sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{p=1}^{k-1} \sqrt{\sum_{j=1}^{p} \alpha_{k-1,j}} \left\|\mathbf{w}_h^p - \mathbf{w}_h^{p+1}\right\|_2 \\
=\ & 2H^2L_\psi^2\sqrt{\frac{d}{\lambda_{\mathbf{w}}}} \sum_{p=1}^{K-1}\sum_{h=1}^{H}\sum_{k=p+1}^{K} \sqrt{\sum_{j=1}^{p} \alpha_{k-1,j}} \left\|\mathbf{w}_h^p - \mathbf{w}_h^{p+1}\right\|_2 \\
=\ & 2H^2L_\psi^2\sqrt{\frac{d}{\lambda_{\mathbf{w}}}} \sum_{p=1}^{K-1}\sum_{h=1}^{H}\sum_{k=p+1}^{K} \gamma^{\frac{k-1}{2}} \sqrt{\sum_{j=1}^{p} \gamma^{-j}} \left\|\mathbf{w}_h^p - \mathbf{w}_h^{p+1}\right\|_2 \\
=\ & 2H^2L_\psi^2\sqrt{\frac{d}{\lambda_{\mathbf{w}}}} \sum_{p=1}^{K-1}\sum_{h=1}^{H} \frac{\gamma^{\frac{p}{2}} - \gamma^{\frac{K}{2}}}{1-\gamma^{\frac{1}{2}}} \sqrt{\frac{\gamma^{-p}-1}{1-\gamma}} \left\|\mathbf{w}_h^p - \mathbf{w}_h^{p+1}\right\|_2 \\
\leq\ & 4H^2L_\psi^2\sqrt{\frac{d}{\lambda_{\mathbf{w}}}} \frac{1}{(1-\gamma)^{3/2}} \sum_{p=1}^{K-1}\sum_{h=1}^{H} \left\|\mathbf{w}_h^p - \mathbf{w}_h^{p+1}\right\|_2, \qquad (63)
\end{aligned}
$$

方差。对于方差部分 1，我们有

$$
\begin{aligned}
&4H\beta_\theta \sum_{k=1}^{K}\sum_{h=1}^{H} \min\left\{1, \left\|\phi(s_h^k, a_h^k)\right\|_{\left(\Lambda_h^{k-1}\right)^{-1}}\right\} \\
\leq\ & 4H^2\beta_\theta\sqrt{K}\sqrt{\sum_{k=1}^{K} \min\left\{1, \left\|\phi(s_h^k, a_h^k)\right\|_{\left(\Lambda_h^{k-1}\right)^{-1}}^2\right\}}.
\end{aligned}
$$

根据引理 9（势引理），并令 $X_k = \phi(s_h^k,\ a_h^k)$、$U_k = \Lambda_h^{k-1}$，我们知道 $\forall h \in [H]$，因此有

$$
\begin{aligned}
&\sum_{k=1}^{K} \min\left\{1, \left\|\phi(s_h^k, a_h^k)\right\|_{\left(\Lambda_h^{k-1}\right)^{-1}}^2\right\} \\
\leq\ & 2d\left(K\log\frac{1}{\gamma} + \log\left(1 + \frac{L_\phi^2}{\lambda_\theta d(1-\gamma)}\right)\right),
\end{aligned}
$$

因此有

$$
\begin{aligned}
&4H\beta_\theta \sum_{k=1}^{K}\sum_{h=1}^{H} \min\left\{1, \left\|\phi(s_h^k, a_h^k)\right\|_{\left(\Lambda_h^{k-1}\right)^{-1}}\right\} \\
\leq\ & 4H^2\beta_\theta\sqrt{K}\sqrt{2d\left(K\log\frac{1}{\gamma} + \log\left(1 + \frac{L_\phi^2}{\lambda_\theta d(1-\gamma)}\right)\right)}.
\end{aligned}
$$

对于方差部分 2，我们有

$$
\begin{aligned}
&4\sum_{k=1}^{K}\sum_{h=1}^{H} \beta_{\mathbf{w}}^{k-1} \min\left\{1, \left\|\psi_{h+1}^k\left(s_h^k, a_h^k\right)\right\|_{\left(\Sigma_h^{k-1}\right)^{-1}}\right\} \\
\leq\ & 4\beta_{\mathbf{w}}^K \sum_{k=1}^{K}\sum_{h=1}^{H} \min\left\{1, \left\|\psi_{h+1}^k\left(s_h^k, a_h^k\right)\right\|_{\left(\Sigma_h^{k-1}\right)^{-1}}\right\} \qquad (64) \\
\leq\ & 4H\beta_{\mathbf{w}}^K\sqrt{K}\sqrt{\sum_{k=1}^{K} \min\left\{1, \left\|\psi_{h+1}^k\left(s_h^k, a_h^k\right)\right\|_{\left(\Sigma_h^{k-1}\right)^{-1}}^2\right\}}
\end{aligned}
$$

根据势引理，我们知道 $\forall h \in [H]$，因此有

$$
\begin{aligned}
&\sum_{k=1}^{K} \min\left\{1, \left\|\psi_{h+1}^k\left(s_h^k, a_h^k\right)\right\|_{\left(\Sigma_h^{k-1}\right)^{-1}}^2\right\} \\
\leq\ & 2d\left(K\log\frac{1}{\gamma} + \log\left(1 + \frac{H^2L_\psi^2}{\lambda_{\mathbf{w}} d(1-\gamma)}\right)\right),
\end{aligned}
$$

因此有

$$4\sum_{k=1}^{K}\sum_{h=1}^{H}\beta_{\mathbf{w}}^{k-1}\min\left\{1,\left\|\psi_{h+1}^{k}\left(s_h^k,a_h^k\right)\right\|_{\left(\Sigma_h^{k-1}\right)^{-1}}\right\}$$
$$\leq 4H\beta_{\mathbf{w}}^{K}\sqrt{K}\sqrt{2d\left(K\log\frac{1}{\gamma}+\log\left(1+\frac{H^2L_\psi^2}{\lambda_{\mathbf{w}}d(1-\gamma)}\right)\right)}.$$

由于遗憾界中存在一项 $HK\sqrt{\log(1/\gamma)}$，因此不能令 γ 接近 0，所以我们设 $\gamma\geq 1/K$，并有 $\log(1/\gamma)\leq C(1-\gamma)$，其中 $C=\log K/(1-1/K)$。我们设定$\lambda_\theta=d$，且 $\lambda_{\mathbf{w}}=H^2d$。结合偏差部分和方差部分的上界，并以置信水平 $\delta=1/(4T)$，根据并集界可得以下动态遗憾界，其成立概率至少为 $1-1/T$，

$$\begin{aligned}&\text{D-REG}_T\\&\leq\mathcal{O}\Big(\frac{1}{(1-\gamma)^{3/2}}P_T^{\theta}+H\frac{1}{(1-\gamma)^{3/2}}P_T^{\mathbf{w}}+HdHK\sqrt{1-\gamma}\\&\quad+H^{3/2}d\sqrt{HK}\Big)\\&\leq\mathcal{O}\left(Hd\left(\frac{1}{(1-\gamma)^{3/2}}\Delta+HK\sqrt{1-\gamma}\right)+H^{3/2}d\sqrt{HK}\right)\\&\overset{\gamma=}{}\end{aligned}$$

此外，将折扣因子最优地设为$1-\max\left\{1/K,\ \sqrt{\Delta/T}\right\}$，可得

$$\text{D-REG}_T\leq\begin{cases}\widetilde{\mathcal{O}}\left(Hd\Delta^{1/4}T^{3/4}\right) & \text{when } \Delta\geq H/K,\\ \widetilde{\mathcal{O}}\left(dH^{3/2}\sqrt{T}\right) & \text{when } \Delta<H/K.\end{cases}\tag{65}$$

□

引理*17*的证明。 我们首先考虑 E_h^k的上界。根据 Q_h^k (26)的定义、模型假设 (22) 和方程 (24)，可得 $\forall a\in\mathcal{A},\ s\in\mathcal{S}$，

$$\begin{aligned}&r_h^k(s,a)+\left[\mathbb{P}_h^kV_{h+1}^k\right](s,a)-Q_h^k(s,a)\\&=r_h^k(s,a)+\left[\mathbb{P}_h^kV_{h+1}^k\right](s,a)-\phi(s,a)^\top\widehat{\theta}_h^k-\psi_{h+1}^k(s,a)^\top\widehat{\mathbf{w}}_h^k\\&\quad-\beta_\theta\left\|\phi(s,a)\right\|_{\left(\Lambda_h^{k-1}\right)^{-1}}-\beta_{\mathbf{w}}^{k-1}\left\|\psi_{h+1}^k(s,a)\right\|_{\left(\Sigma_h^{k-1}\right)^{-1}}\\&=\phi(s,a)^\top\left(\theta_h^k-\widehat{\theta}_h^k\right)+\psi_{h+1}^k(s,a)^\top\left(\mathbf{w}_h^k-\widehat{\mathbf{w}}_h^k\right)\\&\quad-\beta_\theta\left\|\phi(s,a)\right\|_{\left(\Lambda_h^{k-1}\right)^{-1}}-\beta_{\mathbf{w}}^{k-1}\left\|\psi_{h+1}^k(s,a)\right\|_{\left(\Sigma_h^{k-1}\right)^{-1}}\\&\leq\Gamma_{h,\theta}^{k-1}+\Gamma_{h,\mathbf{w}}^{k-1},\end{aligned}$$

其中，最后一个不等式来自引理4和引理5。同样地，我们可以得到 E_h^k、 $\forall a\in\mathcal{A},\ s\in\mathcal{S}$的下界，

$$\begin{aligned}&Q_h^k(s,a)-r_h^k(s,a)-\left[\mathbb{P}_h^kV_{h+1}^k\right](s,a)\\&=\phi(s,a)^\top\widehat{\theta}_h^k+\beta_\theta\left\|\phi(s,a)\right\|_{\left(\Lambda_h^{k-1}\right)^{-1}}+\psi_{h+1}^k(s,a)^\top\widehat{\mathbf{w}}_h^k\\&\quad+\beta_{\mathbf{w}}^{k-1}\left\|\psi_{h+1}^k(s,a)\right\|_{\left(\Sigma_h^{k-1}\right)^{-1}}-r_h^k(s,a)-\left[\mathbb{P}_h^kV_{h+1}^k\right](s,a)\\&=\phi(s,a)^\top\left(\widehat{\theta}_h^k-\theta_h^k\right)+\psi_{h+1}^k(s,a)^\top\left(\widehat{\mathbf{w}}_h^k-\mathbf{w}_h^k\right)\\&\quad+\beta_\theta\left\|\phi(s,a)\right\|_{\left(\Lambda_h^{k-1}\right)^{-1}}+\beta_{\mathbf{w}}^{k-1}\left\|\psi_{h+1}^k(s,a)\right\|_{\left(\Sigma_h^{k-1}\right)^{-1}}\\&\leq\Gamma_{h,\theta}^{k-1}+\Gamma_{h,\mathbf{w}}^{k-1}+2\beta_\theta\left\|\phi(s,a)\right\|_{\left(\Lambda_h^{k-1}\right)^{-1}}\\&\quad+2\beta_{\mathbf{w}}^{k-1}\left\|\psi_{h+1}^k(s,a)\right\|_{\left(\Sigma_h^{k-1}\right)^{-1}},\end{aligned}$$

至此，证明完成。 □

引理*18*的证明。我们首先分解单步动态遗憾，

$$\begin{aligned}&V_1^{k,\pi_*^k}\left(s_1^k\right)-V_1^{k,\pi^k}\left(s_1^k\right)\\&=\underbrace{V_1^{k,\pi_*^k}\left(s_1^k\right)-V_1^k\left(s_1^k\right)}_{\text{TERM (1)}}+\underbrace{V_1^k\left(s_1^k\right)-V_1^{k,\pi^k}\left(s_1^k\right)}_{\text{TERM (2)}}.\end{aligned}$$

术语 (1)。首先有 $\forall s\in\mathcal{S}$，

$$\begin{aligned}&V_h^{k,\pi_*^k}(s)-V_h^k(s)\\&=\mathbb{E}_{a\sim\pi_{*,h}^k(s)}\left[Q_h^{k,\pi_*^k}(s,a)\right]-\mathbb{E}_{a\sim\pi_h^k(s)}\left[Q_h^k(s,a)\right]\\&=\mathbb{E}_{a\sim\pi_{*,h}^k(s)}\left[Q_h^{k,\pi_*^k}(s,a)\right]-\mathbb{E}_{a\sim\pi_{*,h}^k(s)}\left[Q_h^k(s,a)\right]\\&\quad+\mathbb{E}_{a\sim\pi_{*,h}^k(s)}\left[Q_h^k(s,a)\right]-\mathbb{E}_{a\sim\pi_h^k(s)}\left[Q_h^k(s,a)\right]\\&\leq\mathbb{E}_{a\sim\pi_{*,h}^k(s)}\left[Q_h^{k,\pi_*^k}(s,a)-Q_h^k(s,a)\right],\end{aligned}$$

其中，最后一个不等式来自$\pi_h^k(s)=\arg\max_{a\in\mathcal{A}}Q_h^k(s,\ a)$。于是我们有

$$\begin{aligned}&Q_h^{k,\pi_*^k}(s,a)-Q_h^k(s,a)\\&=r_h^k(s,a)+\left[\mathbb{P}_h^kV_{h+1}^{k,\pi_*^k}\right](s,a)-r_h^k(s,a)-\left[\mathbb{P}_h^kV_{h+1}^k\right](s,a)\\&\quad+r_h^k(s,a)+\left[\mathbb{P}_h^kV_{h+1}^k\right](s,a)-Q_h^k(s,a)\\&=\left[\mathbb{P}_h^k\left(V_{h+1}^{k,\pi_*^k}-V_{h+1}^k\right)\right](s,a)+E_h^k(s,a),\end{aligned}$$

其中，最后一个等式来自模型预测误差的定义。为简化记号，我们定义算子 $\mathbb{J}_h^kf(s)=\left\langle f(x,\ \cdot)\pi_{*,\ h}^k(\cdot\mid s)\right\rangle$，于是我们有

$$\begin{aligned}&V_h^{k,\pi_*^k}(s)-V_h^k(s)\\&\leq\mathbb{J}_h^k\mathbb{P}_h^k\left(V_{h+1}^{k,\pi_*^k}-V_{h+1}^k\right)(s,a)+\mathbb{J}_h^kE_h^k(s,a),\end{aligned}$$

递归展开上述不等式可得

$$\begin{aligned}&V_1^{k,\pi_*^k}\left(s_1^k\right)-V_1^k\left(s_1^k\right)\\&\leq\left(\prod_{h=1}^{H}\mathbb{J}_h^k\mathbb{P}_h^k\right)\left(V_{H+1}^{k,\pi_*^k}(s_{H+1}^k)-V_{H+1}^k(s_{H+1})\right)\\&\quad+\sum_{h=1}^{H}\left(\prod_{j=1}^{H}\mathbb{J}_j^k\mathbb{P}_j^k\right)\mathbb{J}_h^kE_h^k(s_h^k,a_h^k)\\&\leq\sum_{h=1}^{H}\mathbb{E}_{\pi_{*,h}^k}\left[E_h^k(s_h^k,a_h^k)\right],\end{aligned}$$

其中，最后一个不等式是由 $\forall\pi,V^{k,\pi}_{,H+1}($ $\cdot\ =0\quad\cdot\ =0$ $),V_{H+1}^k()$ 得出的。于是我们有

$$\begin{aligned}&\sum_{k=1}^{K}V_1^{k,\pi_*^k}\left(s_h^k\right)-V_1^k\left(s_h^k\right)\\&\leq\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{\pi_{*,h}^k}\left[E_h^k(s_h^k,a_h^k)\right].\end{aligned}\tag{66}$$

术语 (2). 根据方程 (61)，我们有 $\forall s \in \mathcal{S}, a \in \mathcal{A}$,

$$\begin{aligned}E_h^k(s,a) &= r_h^k(s,a) + \left[\mathbb{P}_h^k V_{h+1}^k\right](s,a) - Q_h^k(s,a)\\ &= r_h^k(s,a) + \left[\mathbb{P}_h^k V_{h+1}^k\right](s,a) - Q_h^{k,\pi^k}(s,a)\\ &\quad + Q_h^{k,\pi^k}(s,a) - Q_h^k(s,a)\\ &= \left(\left[\mathbb{P}_h^k V_{h+1}^k\right](s,a) - \left[\mathbb{P}_h^k V_{h+1}^{k,\pi^k}\right](s,a)\right)\\ &\quad + Q_h^{k,\pi^k}(s,a) - Q_h^k(s,a).\end{aligned}$$

应用该等式，我们进一步得到

$$\begin{aligned}&V_h^k\left(s_h^k\right) - V_h^{k,\pi^k}\left(s_h^k\right)\\ &= \mathbb{E}_{a\sim\pi_h^k(s_h^k)}\left[Q_h^k(s_h^k,a)\right] - \mathbb{E}_{a\sim\pi_h^k(s_h^k)}\left[Q_h^{k,\pi^k}(s_h^k,a)\right]\\ &\quad + E_h^k - E_h^k\\ &= \mathbb{E}_{a\sim\pi_h^k(s_h^k)}\left[Q_h^k(s_h^k,a)\right] - \mathbb{E}_{a\sim\pi_h^k(s_h^k)}\left[Q_h^{k,\pi^k}(s_h^k,a)\right]\\ &\quad + \left(\left[\mathbb{P}_h^k V_{h+1}^k\right]\left(s_h^k,a_h^k\right) - \left[\mathbb{P}_h^k V_{h+1}^{k,\pi^k}\right]\left(s_h^k,a_h^k\right)\right)\\ &\quad + Q_h^{k,\pi_h^k}\left(s_h^k,a_h^k\right) - Q_h^k\left(s_h^k,a_h^k\right) - E_h^k\left(s_h^k,a_h^k\right).\end{aligned}$$

我们定义

$$\begin{aligned}\mathcal{M}_{h,V}^k &= \left(\left[\mathbb{P}_h^k V_{h+1}^k\right]\left(s_h^k,a_h^k\right) - \left[\mathbb{P}_h^k V_{h+1}^{k,\pi_h^k}\right]\left(s_h^k,a_h^k\right)\right)\\ &\quad - \left(V_{h+1}^k\left(s_{h+1}^k\right) - V_{h+1}^{k,\pi^k}\left(s_{h+1}^k\right)\right)\\ \mathcal{M}_{h,Q}^k &= \mathbb{E}_{a\sim\pi_h^k(s)}\left[Q_h^k(s_h^k,a) - Q_h^{k,\pi_h^k}(s_h^k,a)\right]\\ &\quad - \left(Q_h^k\left(s_h^k,a_h^k\right) - Q_h^{k,\pi_h^k}\left(s_h^k,a_h^k\right)\right),\end{aligned}$$

于是我们有

$$\begin{aligned}&V_h^k\left(s_h^k\right) - V_h^{\pi^k}\left(s_h^k\right) - \left(V_{h+1}^k\left(s_{h+1}^k\right) - V_{h+1}^{\pi^k}\left(s_{h+1}^k\right)\right)\\ &= \mathcal{M}_{h,V}^k + \mathcal{M}_{h,Q}^k - E_h^k\left(s_h^k,a_h^k\right).\end{aligned}$$

对 $k \in [K]$ 和 $h \in [H]$求和，由于 $V_{H+1}^k = 0$, $V^{k}{}_{,H+1}^{\pi} = 0$，我们有

$$\begin{aligned}&\sum_{k=1}^K V_1^k\left(s_1^k\right) - V_1^{\pi^k}\left(s_1^k\right)\\ &\le \sum_{k=1}^K\sum_{h=1}^H\left(-E_h^k\left(s_h^k,a_h^k\right)\right) + \sum_{k=1}^K\sum_{h=1}^H \mathcal{M}_{h,V}^k + \sum_{k=1}^K\sum_{h=1}^H \mathcal{M}_{h,Q}^k.\end{aligned}$$

由于 $\mathcal{M}_{h,V}^k$ 和 $\mathcal{M}_{h,Q}^k$ 是有界于 $2H$的鞅差分，因此根据引理 24，以至少 $1-\delta$的概率，以下结论均成立:

$$\begin{aligned}\sum_{k=1}^K\sum_{h=1}^H \mathcal{M}_{h,V}^k &\le 2H\sqrt{2T\log(1/\delta)},\\ \sum_{k=1}^K\sum_{h=1}^H \mathcal{M}_{h,Q}^k &\le 2H\sqrt{2T\log(1/\delta)},\end{aligned}$$

其中 $T = KH$。那么，以至少 $1-2\delta$ 的概率，以下结论成立:

$$\begin{aligned}&\sum_{k=1}^K\left(V_1^k\left(s_1^k\right) - V_1^{\pi^k}\left(s_1^k\right)\right)\\ &\le \sum_{k=1}^K\sum_{h=1}^H\left(-E_h^k\left(s_h^k,a_h^k\right)\right) + 4H\sqrt{2T\log(1/\delta)}.\end{aligned} \tag{67}$$

结合方程 (66) 和方程 (67)，可得

$$\begin{aligned}&V_1^{k,\pi_*^k}\left(s_1^k\right) - V_1^{k,\pi^k}\left(s_1^k\right)\\ &= \sum_{k=1}^K\sum_{h=1}^H\left(\mathbb{E}_{\pi_{*,h}^k}\left[E_h^k(s_h^k,a_h^k)\right] - E_h^k\left(s_h^k,a_h^k\right)\right)\\ &\quad + 4H\sqrt{2T\log(1/\delta)}.\end{aligned}$$

□

附录F MNL-WEIGHTUCRL 分析

A. 引理6的证明

证明。 根据 [17],的引理 2，我们知道

$$\begin{aligned}&\left|\left[\widetilde{\mathbb{P}}_h^k V\right](s,a) - \left[\mathbb{P}_h^k V\right](s,a)\right|\\ &= \left|\sum_{s'\in\mathcal{S}_h^k} p_h^k(\psi(s'\mid s,a)^\top \widetilde{\mathbf{w}}_h^k)V(s')\right.\\ &\quad \left. - \sum_{s'\in\mathcal{S}_h^k} p_h^k(\psi(s'\mid s,a)^\top \mathbf{w}_h^k)V(s')\right|\\ &\le H\max_{s'\in\mathcal{S}_h^k}\left|\psi(s'\mid s,a)^\top\left(\widetilde{\mathbf{w}}_h^k - \mathbf{w}_h^k\right)\right|.\end{aligned}$$

我们首先构造以下引理。

引理19。 对于任意$\mathbf{x} \in \mathcal{X}$，以及 $\delta \in (0,1)$， $\forall k$、$j \in [K]$、$\alpha_{k,j} \le 1$，以至少 $1-\delta$ 的概率，以下结论对所有 $k \in [K]$、$h \in [H]$ 均成立

$$\begin{aligned}&\left|\psi(s'\mid s,a)^\top\left(\widetilde{\mathbf{w}}_h^k - \mathbf{w}_h^k\right)\right|\\ &\le \frac{1}{\kappa}\left(\Gamma_{\mathbf{w}}^{k-1} + \bar{\beta}_{\mathbf{w}}^{k-1}\left\|\psi(s'\mid s,a)\right\|_{\left(\bar{\Sigma}_h^{k-1}\right)^{-1}}\right).\end{aligned}$$

$\Gamma_{\mathbf{w}}^{k-1} \triangleq L_\psi^2\sqrt{\frac{d}{\lambda_{\mathbf{w}}}}\sum_{p=1}^{k-1}\sqrt{\sum_{j=1}^p \alpha_{k-1,j}}\left\|\mathbf{w}_h^p - \mathbf{w}_h^{p+1}\right\|_2$，且 $\bar{\beta}_{\mathbf{w}}^k$ 是由以下方式设定的置信区域的半径

$$\begin{aligned}\bar{\beta}_{\mathbf{w}}^k \triangleq &\sqrt{\frac{1}{2}\log\frac{1}{\delta} + \frac{d}{4}\log\left(1 + \frac{UL_\psi^2\sum_{j=1}^k \alpha_{k,j}}{\lambda_{\mathbf{w}} d}\right)}\\ &+ \sqrt{\lambda_{\mathbf{w}}}\kappa S_{\mathbf{w}}.\end{aligned}$$

于是我们有

$$\begin{aligned}&\left|\left[\widetilde{\mathbb{P}}_h^k V\right](s,a) - \left[\mathbb{P}_h^k V\right](s,a)\right|\\ &\le \frac{H}{\kappa}\left(\Gamma_{\mathbf{w}}^{k-1} + \bar{\beta}_{\mathbf{w}}^{k-1}\max_{s'\in\mathcal{S}_h^k}\left\|\psi(s'\mid s,a)\right\|_{\left(\bar{\Sigma}_h^{k-1}\right)^{-1}}\right),\end{aligned}$$

其中 $\Gamma_{\mathbf{w}}^{k-1} \triangleq L_\psi^2\sqrt{\frac{d}{\lambda_{\mathbf{w}}}}\sum_{p=1}^{k-1}\sqrt{\sum_{j=1}^p \alpha_{k-1,j}}\left\|\mathbf{w}_h^p - \mathbf{w}_h^{p+1}\right\|_2$，$\bar{\beta}_{\mathbf{w}}^k$ 是所设置信任区域的半径

$$\begin{aligned}\bar{\beta}_{\mathbf{w}}^k \triangleq &\sqrt{\frac{1}{2}\log\frac{1}{\delta} + \frac{d}{4}\log\left(1 + \frac{UL_\psi^2\sum_{j=1}^k \alpha_{k,j}}{\lambda_{\mathbf{w}} d}\right)}\\ &+ \sqrt{\lambda_{\mathbf{w}}}\kappa S_{\mathbf{w}}.\end{aligned}$$

□

引理*19*的证明。固定 $h \in [H]$，根据估计量 (28)，我们有

$$\begin{aligned} g_h^k(\widehat{\mathbf{w}}_h^k) &= \lambda_{\mathbf{w}}\kappa\widehat{\mathbf{w}}_h^k + \sum_{j=1}^{k-1}\alpha_{k-1,j}\sum_{s'\in\mathcal{S}_h^j} p_h^j(\bar{\psi}_h^j(s')^\top\widehat{\mathbf{w}}_h^k)\bar{\psi}_h^j(s') \\ &= \sum_{j=1}^{k-1}\alpha_{k-1,j}\sum_{s'\in\mathcal{S}_h^j} y_h^j(s')\bar{\psi}_h^j(s') \\ &= \sum_{j=1}^{k-1}\alpha_{k-1,j}\sum_{s'\in\mathcal{S}_h^j}\left(p_h^j(\bar{\psi}_h^j(s')^\top\mathbf{w}_h^j) + \eta_h^j(s')\right)\bar{\psi}_h^j(s'), \end{aligned}$$

其中我们定义 $\eta_h^j(s') \triangleq y_h^j(s') - p_h^j(\bar{\psi}_h^j(s')^\top\mathbf{w}_h^j)$。根据中值定理，我们知道

$$g_h^k(\mathbf{w}_1) - g_h^k(\mathbf{w}_2) = G_h^k(\mathbf{w}_1 - \mathbf{w}_2), \tag{68}$$

其中 $G_h^k \quad _1 - \quad _2 \triangleq \int_0^1 \nabla g_h^k \quad s \quad _2 + \ 1 - s \quad _1 \quad s \ \in \mathcal{W}$ $\mathbb{R}^{d\times d}$ (ww) (w( )w) d 。注意，对于任意 $\mathbf{w}$， g_h^k 的梯度为

$$\begin{aligned} &\nabla g_h^k(\mathbf{w}) = \lambda_{\mathbf{w}}\kappa I_d \\ &+ \sum_{j=1}^{k-1}\alpha_{k-1,j}\Bigg(\sum_{s'\in\mathcal{S}_h^j} p_h^j(\bar{\psi}_h^j(s')^\top\mathbf{w})\bar{\psi}_h^j(s')\bar{\psi}_h^j(s')^\top \\ &- \sum_{s'\in\mathcal{S}_h^j}\sum_{s''\in\mathcal{S}_h^j} p_h^j(\bar{\psi}_h^j(s')^\top\mathbf{w})p_h^j(\bar{\psi}_h^j(s'')^\top\mathbf{w})\bar{\psi}_h^j(s')\bar{\psi}_h^j(s'')^\top\Bigg) \end{aligned}$$

根据 [17], 的引理5，我们知道

$$\begin{aligned} &\nabla g_h^k(\mathbf{w}) \\ &\succeq \lambda_{\mathbf{w}}\kappa I_d + \kappa\sum_{j=1}^{k-1}\alpha_{k-1,j}\sum_{s'\in\mathcal{S}_h^j}\bar{\psi}_h^j(s')\bar{\psi}_h^j(s')^\top = \kappa\bar{\Sigma}_h^{k-1}, \end{aligned}$$

这显然意味着 $\forall \mathbf{w}_1, \mathbf{w}_2 \in \mathcal{W}, G_h^k(\mathbf{w}_1, \mathbf{w}_2) \succeq \kappa\bar{\Sigma}_h^{k-1}$。根据假设6，对 g_h^k 应用中值定理（68）和投影（29），我们有

$$\begin{aligned} &\left|\psi(s' \mid s,a)^\top\left(\widetilde{\mathbf{w}}_h^k - \mathbf{w}_h^k\right)\right| \\ &= \left|\psi(s' \mid s,a)^\top\left(G_h^k(\widetilde{\mathbf{w}}_h^k, \mathbf{w}_h^k)\right)^{-1}\left(g_h^k(\widetilde{\mathbf{w}}_h^k) - g_h^k(\mathbf{w}_h^k)\right)\right| \\ &\leq \|\psi(s' \mid s,a)\|_{\left(G_h^k(\widetilde{\mathbf{w}}_h^k, \mathbf{w}_h^k)\right)^{-1}} \\ &\quad \cdot \left\|g_h^k(\widetilde{\mathbf{w}}_h^k) - g_h^k(\mathbf{w}_h^k)\right\|_{\left(G_h^k(\widetilde{\mathbf{w}}_h^k, \mathbf{w}_h^k)\right)^{-1}} \\ &\leq \frac{1}{\kappa}\|\psi(s' \mid s,a)\|_{\left(\bar{\Sigma}_h^{k-1}\right)^{-1}}\left\|g_h^k(\widetilde{\mathbf{w}}_h^k) - g_h^k(\mathbf{w}_h^k)\right\|_{\left(\bar{\Sigma}_h^{k-1}\right)^{-1}} \\ &\leq \frac{1}{\kappa}\|\psi(s' \mid s,a)\|_{\left(\bar{\Sigma}_h^{k-1}\right)^{-1}}\left\|g_h^k(\widehat{\mathbf{w}}_h^k) - g_h^k(\mathbf{w}_h^k)\right\|_{\left(\bar{\Sigma}_h^{k-1}\right)^{-1}}, \end{aligned}$$

我们进一步有

$$\begin{aligned} &g_h^k(\widehat{\mathbf{w}}_h^k) - g_h^k(\mathbf{w}_h^k) \\ &= \sum_{j=1}^{k-1}\alpha_{k-1,j}\sum_{s'\in\mathcal{S}_h^j}\left(p_h^j(\bar{\psi}_h^j(s')^\top\mathbf{w}_h^j) + \eta_h^j(s')\right)\bar{\psi}_h^j(s') \\ &\quad - \lambda_{\mathbf{w}}\kappa\mathbf{w}_h^k - \sum_{j=1}^{k-1}\alpha_{k-1,j}\sum_{s'\in\mathcal{S}_h^j} p_h^j(\bar{\psi}_h^j(s')^\top\mathbf{w}_h^k)\bar{\psi}_h^j(s') \\ &= \underbrace{\sum_{j=1}^{k-1}\alpha_{k-1,j}\sum_{s'\in\mathcal{S}_h^j}\left(p_h^j(\bar{\psi}_h^j(s')^\top\mathbf{w}_h^j) - p_h^j(\bar{\psi}_h^j(s')^\top\mathbf{w}_h^k)\right)\bar{\psi}_h^j(s')}_{\texttt{bias part}} \\ &\quad \underbrace{- \lambda_{\mathbf{w}}\kappa\mathbf{w}_h^k + \sum_{j=1}^{k-1}\alpha_{k-1,j}\sum_{s'\in\mathcal{S}_h^j}\eta_h^j(s')\bar{\psi}_h^j(s')}_{\texttt{variance part}}. \end{aligned}$$

然后，根据柯西-施瓦茨不等式，我们知道对于任意 $s \in \mathcal{S}$, $a \in \mathcal{A}$,

$$\begin{aligned} &\left|\psi(s' \mid s,a)^\top\left(\widehat{\mathbf{w}}_h^k - \mathbf{w}_h^k\right)\right| \\ &\leq \frac{1}{\kappa}\|\psi(s' \mid s,a)\|_{\left(\bar{\Sigma}_h^{k-1}\right)^{-1}}\left(E_h^k + F_h^k\right), \end{aligned} \tag{69}$$

其中

$$\begin{aligned} E_h^k &= \Bigg\|\sum_{j=1}^{k-1}\alpha_{k-1,j}\sum_{s'\in\mathcal{S}_h^j}\Big(p_h^j(\bar{\psi}_h^j(s')^\top\mathbf{w}_h^j) \\ &\qquad - p_h^j(\bar{\psi}_h^j(s')^\top\mathbf{w}_h^k)\Big)\bar{\psi}_h^j(s')\Bigg\|_{\left(\bar{\Sigma}_h^{k-1}\right)^{-1}} \\ F_h^k &= \left\|-\lambda_{\mathbf{w}}\kappa\mathbf{w}_h^k + \sum_{j=1}^{k-1}\alpha_{k-1,j}\sum_{s'\in\mathcal{S}_h^j}\eta_h^j(s')\bar{\psi}_h^j(s')\right\|_{\left(\bar{\Sigma}_h^{k-1}\right)^{-1}}. \end{aligned}$$

上述两项可以分别进行界定，如以下两个引理所述，

引理20。 对于任意 $k \in [K]$，有

$$E_h^k \leq L_\psi\sqrt{d}\sum_{p=1}^{k-1}\sqrt{\sum_{j=1}^{p}\alpha_{k-1,j}}\left\|\mathbf{w}_h^p - \mathbf{w}_h^{p+1}\right\|_2.$$

引理21。 若 $\forall k \in [K], \forall j \in [k-1], \alpha_{k-1,j} \leq 1$，对于任意 $\delta \in (0,1)$，以至少 $1-\delta$ 的概率，以下结论对所有 $k \in [K]$ 均成立，

$$\begin{aligned} F_h^k \leq \sqrt{\frac{1}{2}\log\frac{1}{\delta} + \frac{d}{4}\log\left(1 + \frac{UL_\psi^2\sum_{j=1}^{k-1}\alpha_{k-1,j}}{\lambda_{\mathbf{w}}d}\right)} \\ + \sqrt{\lambda_{\mathbf{w}}}\kappa S_{\mathbf{w}}. \end{aligned}$$

基于不等式（69）、引理20、引理21，以及 $\|\psi(s' \mid s,a)\|_{\left(\bar{\Sigma}_h^{k-1}\right)^{-1}} \leq \|\psi(s' \mid s,a)\|_2/\sqrt{\lambda_{\mathbf{w}}} \leq L_\psi/\sqrt{\lambda_{\mathbf{w}}}$，以至少 $1-\delta$ 的概率，以下结论对所有 $k \in [K]$ 均成立，

$$\begin{aligned} &\left|\psi(s' \mid s,a)^\top\left(\widetilde{\mathbf{w}}_h^k - \mathbf{w}_h^k\right)\right| \\ &\leq \frac{1}{\kappa}\left(\Gamma_{\mathbf{w}}^{k-1} + \bar{\beta}_{\mathbf{w}}^{k-1}\|\psi(s' \mid s,a)\|_{\left(\bar{\Sigma}_h^{k-1}\right)^{-1}}\right), \end{aligned}$$

其中

$$\begin{aligned} \Gamma_{\mathbf{w}}^{k-1} &\triangleq L_\psi^2\sqrt{\frac{d}{\lambda_{\mathbf{w}}}}\sum_{p=1}^{k-1}\sqrt{\sum_{j=1}^{p}\alpha_{k-1,j}}\left\|\mathbf{w}_h^p - \mathbf{w}_h^{p+1}\right\|_2 \\ \bar{\beta}_{\mathbf{w}}^{k-1} &\triangleq \sqrt{\frac{1}{2}\log\frac{1}{\delta} + \frac{d}{4}\log\left(1 + \frac{UL_\psi^2\sum_{j=1}^{k-1}\alpha_{k-1,j}}{\lambda_{\mathbf{w}}d}\right)} \\ &\quad + \sqrt{\lambda_{\mathbf{w}}}\kappa S_{\mathbf{w}}, \end{aligned}$$

由此完成证明。 □

引理*20*的证明。第一步是提取参数$\mathbf{w}_h^k$ 的变化量，如下所示，

$$\begin{aligned}
&\left\|\sum_{j=1}^{k-1}\alpha_{k-1,j}\sum_{s'\in\mathcal{S}_h^j}\Big(p_h^j(\bar{\psi}_h^j(s')^\top\mathbf{w}_h^j)\right.\\
&\qquad\left.-p_h^j(\bar{\psi}_h^j(s')^\top\mathbf{w}_h^k)\Big)\bar{\psi}_h^j(s')\right\|_{(\bar{\Sigma}_h^{k-1})^{-1}}\\
=&\left\|\sum_{j=1}^{k-1}\alpha_{k-1,j}\sum_{s'\in\mathcal{S}_h^j}\sum_{p=j}^{k-1}\Big(p_h^j(\bar{\psi}_h^j(s')^\top\mathbf{w}_h^p)\right.\\
&\qquad\left.-p_h^j(\bar{\psi}_h^j(s')^\top\mathbf{w}_h^{p+1})\Big)\bar{\psi}_h^j(s')\right\|_{(\bar{\Sigma}_h^{k-1})^{-1}}\\
=&\left\|\sum_{p=1}^{k-1}\sum_{j=1}^{p}\alpha_{k-1,j}\sum_{s'\in\mathcal{S}_h^j}\Big(p_h^j(\bar{\psi}_h^j(s')^\top\mathbf{w}_h^p)\right.\\
&\qquad\left.-p_h^j(\bar{\psi}_h^j(s')^\top\mathbf{w}_h^{p+1})\Big)\bar{\psi}_h^j(s')\right\|_{(\bar{\Sigma}_h^{k-1})^{-1}}.
\end{aligned}$$

$p_h^j(\bar{\psi}_h^j(s')^\top\mathbf{w})$的梯度为

$$\begin{aligned}
\nabla p_h^j(\bar{\psi}_h^j(s')^\top\mathbf{w}) &= p_h^j(\bar{\psi}_h^j(s')^\top\mathbf{w})\bar{\psi}_h^j(s')^\top\\
&\quad-p_h^j(\bar{\psi}_h^j(s')^\top\mathbf{w})\sum_{s''\in\mathcal{S}_h^j}p_h^j(\bar{\psi}_h^j(s'')^\top\mathbf{w})\bar{\psi}_h^j(s'')^\top,
\end{aligned}$$

根据中值定理，存在 $\nu\in[0,1]$ 满足

$$\begin{aligned}
&p_h^j(\bar{\psi}_h^j(s')^\top\mathbf{w}_h^p)-p_h^j(\bar{\psi}_h^j(s')^\top\mathbf{w}_h^{p+1})\\
&=p_h^j(\bar{\psi}_h^j(s')^\top\bar{\mathbf{w}}_h^p)\left(\bar{\psi}_h^j(s')^\top-\sum_{s''\in\mathcal{S}_h^j}p_h^j(\bar{\psi}_h^j(s'')^\top\bar{\mathbf{w}}_h^p)\bar{\psi}_h^j(s'')^\top\right)\\
&\quad\cdot\left(\mathbf{w}_h^p-\mathbf{w}_h^{p+1}\right)\\
&\le L_\psi p_h^j(\bar{\psi}_h^j(s')^\top\bar{\mathbf{w}}_h^p)\left(1-\sum_{s''\in\mathcal{S}_h^j}p_h^j(\bar{\psi}_h^j(s'')^\top\bar{\mathbf{w}}_h^p)\right)\\
&\quad\cdot\left\|\mathbf{w}_h^p-\mathbf{w}_h^{p+1}\right\|_2\\
&\le L_\psi\nabla p_h^j(\bar{\psi}_h^j(s')^\top\bar{\mathbf{w}}_h^p)\left\|\mathbf{w}_h^p-\mathbf{w}_h^{p+1}\right\|_2,
\end{aligned}$$

其中我们定义 $\nabla p_h^j(\bar{\psi}_h^j(s')^\top\bar{\mathbf{w}}_h^p)=p_h^j(\bar{\psi}_h^j(s')^\top\bar{\mathbf{w}}_h^p)\cdot\left(1-\sum_{s''\in\mathcal{S}_h^j}p_h^j(\bar{\psi}_h^j(s'')^\top\bar{\mathbf{w}}_h^p)\right)\bar{\psi}_h^j(s')^\top$。于是我们有

$$\begin{aligned}
&\left\|\sum_{p=1}^{k-1}\sum_{j=1}^{p}\alpha_{k-1,j}\sum_{s'\in\mathcal{S}_h^j}\Big(p_h^j(\bar{\psi}_h^j(s')^\top\mathbf{w}_h^p)\right.\\
&\qquad\left.-p_h^j(\bar{\psi}_h^j(s')^\top\mathbf{w}_h^{p+1})\Big)\bar{\psi}_h^j(s')\right\|_{(\bar{\Sigma}_h^{k-1})^{-1}}\\
&\quad\cdot\left\|\mathbf{w}_h^p-\mathbf{w}_h^{p+1}\right\|_2\\
&\le L_\psi\sum_{p=1}^{k-1}\sum_{j=1}^{p}\alpha_{k-1,j}\left|\sum_{s'\in\mathcal{S}_h^j}\nabla p_h^j(\bar{\psi}_h^j(s')^\top\bar{\mathbf{w}}_h^p)\right|\\
&\quad\cdot\left\|\max_{s'\in\mathcal{S}_h^j}\bar{\psi}_h^j(s')\right\|_{(\bar{\Sigma}_h^{k-1})^{-1}}\left\|\mathbf{w}_h^p-\mathbf{w}_h^{p+1}\right\|_2\\
&\le L_\psi\sum_{p=1}^{k-1}\sum_{j=1}^{p}\alpha_{k-1,j}\left\|\max_{s'\in\mathcal{S}_h^j}\bar{\psi}_h^j(s')\right\|_{(\bar{\Sigma}_h^{k-1})^{-1}}\left\|\mathbf{w}_h^p-\mathbf{w}_h^{p+1}\right\|_2,
\end{aligned}$$

其中最后一个不等式源于以下事实：$\left|\sum_{s'\in\mathcal{S}_h^j}\nabla p_h^j(\bar{\psi}_h^j(s')^\top\mathbf{w}_h^p)\right|\le 1$。此外，对于项$\sum_{j=1}^p\alpha_{k-1,j}\left\|\max_{s'\in\mathcal{S}_h^j}\bar{\psi}_h^j(s')\right\|_{(\bar{\Sigma}_h^{k-1})^{-1}}$，还可以进一步推导出关于权重 $\alpha_{k-1,j}$ 的表达式，如下所示，

$$\begin{aligned}
&\sum_{j=1}^{p}\alpha_{k-1,j}\left\|\max_{s'\in\mathcal{S}_h^j}\bar{\psi}_h^j(s')\right\|_{(\bar{\Sigma}_h^{k-1})^{-1}}\\
&\le\sqrt{\sum_{j=1}^{p}\alpha_{k-1,j}}\sqrt{\sum_{j=1}^{p}\alpha_{k-1,j}\left\|\max_{s'\in\mathcal{S}_h^j}\bar{\psi}_h^j(s')\right\|^2_{(\bar{\Sigma}_h^{k-1})^{-1}}}\\
&\le\sqrt{\sum_{j=1}^{p}\alpha_{k-1,j}}\sqrt{\sum_{j=1}^{p}\sum_{s'\in\mathcal{S}_h^j}\alpha_{k-1,j}\left\|\bar{\psi}_h^j(s')\right\|^2_{(\bar{\Sigma}_h^{k-1})^{-1}}}\\
&\le\sqrt{d}\sqrt{\sum_{j=1}^{p}\alpha_{k-1,j}}.
\end{aligned}$$

上文倒数第二步由柯西-施瓦茨不等式得出。此外，最后一步是因为

$$\begin{aligned}
&\sum_{j=1}^{p}\sum_{s'\in\mathcal{S}_h^j}\alpha_{k-1,j}\left\|\bar{\psi}_h^j(s')\right\|^2_{(\bar{\Sigma}_h^{k-1})^{-1}}\\
&\sum_{j=1}^{p}\sum_{s'\in\mathcal{S}_h^j}\mathrm{Tr}(\alpha_{k-1,j}\bar{\psi}_h^j(s')^\top\left(\bar{\Sigma}_h^{k-1}\right)^{-1}\bar{\psi}_h^j(s'))\\
&=\mathrm{Tr}\left(\left(\bar{\Sigma}_h^{k-1}\right)^{-1}\sum_{j=1}^{p}\sum_{s'\in\mathcal{S}_h^j}\alpha_{k-1,j}\bar{\psi}_h^j(s')\bar{\psi}_h^j(s')^\top\right)\\
&\le\mathrm{Tr}\left(\left(\bar{\Sigma}_h^{k-1}\right)^{-1}\sum_{j=1}^{k-1}\sum_{s'\in\mathcal{S}_h^j}\alpha_{k-1,j}\bar{\psi}_h^j(s')\bar{\psi}_h^j(s')^\top\right)\\
&\quad+\mathrm{Tr}\left(U_{t-1}^{-1}\lambda\sum_{i=1}^{d}\mathbf{e}_i\mathbf{e}_i^\top\right)\\
&=\mathrm{Tr}(I_d)=d.
\end{aligned}$$

至此，证明完成。 □

引理*21*的证明。

$$\left\|-\lambda_{\mathbf{w}}\kappa\mathbf{w}_h^k+\sum_{j=1}^{k-1}\alpha_{k-1,j}\sum_{s'\in\mathcal{S}_h^j}\eta_h^j(s')\bar{\psi}_h^j(s')\right\|_{(\bar{\Sigma}_h^{k-1})^{-1}}$$

$$\leq \left\|\sum_{j=1}^{k-1}\alpha_{k-1,j}\sum_{s'\in\mathcal{S}_h^j}\eta_h^j(s')\bar{\psi}_h^j(s')\right\|_{\left(\bar{\Sigma}_h^{k-1}\right)^{-1}} + \left\|\lambda_{\mathbf{w}}\kappa\mathbf{w}_h^k\right\|_{\left(\bar{\Sigma}_h^{k-1}\right)^{-1}}$$

$$\leq \left\|\sum_{j=1}^{k-1}\alpha_{k-1,j}\sum_{s'\in\mathcal{S}_h^j}\eta_h^j(s')\bar{\psi}_h^j(s')\right\|_{\left(\bar{\Sigma}_h^{k-1}\right)^{-1}} + \sqrt{\lambda_{\mathbf{w}}}\kappa S_{\mathbf{w}}.$$

我们定义 $\widetilde{\eta}_j = \sqrt{\alpha_{k-1,j}}\eta_j$ 和 $X_j = \sqrt{\alpha_{k-1,j}}\bar{\psi}_h^j(s')$，然后注意到，由于奖励 $r\in[0,1]$，且 $\alpha_{k-1,j}\leq 1$，噪声 $\widetilde{\eta}_j$ 有界：

$$\widetilde{\eta}_j = \sqrt{\alpha_{k-1,j}}\left(y_h^j(s') - p_h^j(\bar{\psi}_h^j(s')^\top\mathbf{w}_h^j)\right)\leq 1,$$

根据引理23，我们发现噪声 $\widetilde{\eta}_j$ 是1 2-次亚高斯的。然后，根据定理7，以至少 $1-\delta$ 的概率，以下结论对所有 $k\in[K]$ 均成立。

$$\left\|\sum_{j=1}^{k-1}\widetilde{\eta}_j X_j\right\|_{\left(\bar{\Sigma}_h^{k-1}\right)^{-1}} \leq \sqrt{\frac{1}{2}\log\left(\frac{\det(\bar{\Sigma}_h^{k-1})^{\frac{1}{2}}\det(\bar{\Sigma}_h^0)^{-\frac{1}{2}}}{\delta}\right)}$$

其中

$$\begin{aligned}\det(\bar{\Sigma}_h^{k-1}) &\leq \left(\frac{\mathrm{trace}(\bar{\Sigma}_h^{k-1})}{d}\right)^d \\ &= \left(\frac{d\lambda_{\mathbf{w}} + \sum_{j=1}^{k-1}\sum_{s'\in\mathcal{S}_h^j}\alpha_{k-1,j}\left\|\bar{\psi}_h^j(s')\right\|_2^2}{d}\right)^d \\ &= \left(\frac{d\lambda_{\mathbf{w}} + UL_\psi^2\sum_{j=1}^{k-1}\alpha_{k-1,j}}{d}\right)^d\end{aligned}$$

$$\det(\bar{\Sigma}_h^0)\leq(\lambda_{\mathbf{w}})^d,$$

因此我们有

$$\left\|\sum_{j=1}^{k-1}\alpha_{k-1,j}\eta_j\bar{\psi}_{h+1}^j\left(s_h^j,a_h^j\right)\right\|_{\left(\bar{\Sigma}_h^{k-1}\right)^{-1}} \leq \sqrt{\frac{1}{2}\log\frac{1}{\delta}+\frac{d}{4}\log\left(1+\frac{UL_\psi^2\sum_{j=1}^{k-1}\alpha_{k-1,j}}{\lambda_{\mathbf{w}}d}\right)}.$$

由此完成证明。 □

B. 定理*6*的证明

证明。 为证明该定理，我们首先引入以下引理

引理 22。 我们将模型预测误差定义为

$$E_h^k(s,a) = r_h^k(s,a) + \mathbb{P}_h^k\bar{V}_{h+1}^k(s,a) - \bar{Q}_h^k(s,a), \qquad (70)$$

那么，以至少 $1-2\delta$ 的概率，对于所有 $k\in[K]$、$h\in[H]$ 以及 $\forall s\in\mathcal{S}$、$a\in\mathcal{A}$，以下结论成立：

$$\begin{aligned}&-\Gamma_{h,\theta}^{k-1} - \frac{H}{\kappa}\Gamma_{h,\mathbf{w}}^{k-1} - 2\beta_\theta\left\|\phi(s,a)\right\|_{\left(\Lambda_h^{k-1}\right)^{-1}} \\ &-2\frac{H}{\kappa}\bar{\beta}_{\mathbf{w}}^{k-1}\max_{s'\in\mathcal{S}_h^k}\left\|\psi(s'\mid s,a)\right\|_{\left(\bar{\Sigma}_h^{k-1}\right)^{-1}} \leq E_h^k(s,a) \leq \Gamma_{h,\theta}^{k-1} + \frac{H}{\kappa}\Gamma_{h,\mathbf{w}}^{k-1}.\end{aligned}$$

And notice that $\forall s\in\mathcal{S},a\in\mathcal{A}$, $\left|E_h^k(s,a)\right|\leq 2H$, we have

$$\begin{aligned}&\mathbb{E}_{\pi_{*,h}^k}\left[E_h^k(s_h^k,a_h^k)\right] - E_h^k(s_h^k,a_h^k) \\ &\leq \min\left\{4H, 2\Gamma_{h,\theta}^{k-1} + 2\frac{H}{\kappa}\Gamma_{h,\mathbf{w}}^{k-1} + 2\beta_\theta\left\|\phi(s_h^k,a_h^k)\right\|_{\left(\Lambda_h^{k-1}\right)^{-1}}\right. \\ &\qquad\left. + 2\frac{H}{\kappa}\bar{\beta}_{\mathbf{w}}^{k-1}\max_{s'\in\mathcal{S}_h^k}\left\|\psi(s'\mid s_h^k,a_h^k)\right\|_{\left(\bar{\Sigma}_h^{k-1}\right)^{-1}}\right\} \\ &\leq 2\Gamma_{h,\theta}^{k-1} + 2\frac{H}{\kappa}\Gamma_{h,\mathbf{w}}^{k-1} + \min\left\{4H, 2\beta_\theta\left\|\phi(s_h^k,a_h^k)\right\|_{\left(\Lambda_h^{k-1}\right)^{-1}}\right\} \\ &\qquad + \min\left\{4H, 2\frac{H}{\kappa}\bar{\beta}_{\mathbf{w}}^{k-1}\max_{s'\in\mathcal{S}_h^k}\left\|\psi(s'\mid s_h^k,a_h^k)\right\|_{\left(\bar{\Sigma}_h^{k-1}\right)^{-1}}\right\} \\ &\leq 2\Gamma_{h,\theta}^{k-1} + 2\frac{H}{\kappa}\Gamma_{h,\mathbf{w}}^{k-1} + 4H\beta_\theta\min\left\{1, \left\|\phi(s_h^k,a_h^k)\right\|_{\left(\Lambda_h^{k-1}\right)^{-1}}\right\} \\ &\qquad + 4\frac{H}{\kappa}\bar{\beta}_{\mathbf{w}}^{k-1}\min\left\{1, \max_{s'\in\mathcal{S}_h^k}\left\|\psi(s'\mid s_h^k,a_h^k)\right\|_{\left(\bar{\Sigma}_h^{k-1}\right)^{-1}}\right\},\end{aligned}$$

根据引理18，我们可以进一步将动态遗憾与模型预测误差联系起来，以至少$1-4\delta$的概率，有：

$$\begin{aligned}\text{D-REG}_T \leq\ &\underbrace{4H\beta_\theta\sum_{k=1}^K\sum_{h=1}^H\min\left\{1,\left\|\phi(s_h^k,a_h^k)\right\|_{\left(\Lambda_h^{k-1}\right)^{-1}}\right\}}_{\texttt{variance part1}} \\ &+\underbrace{4\frac{H}{\kappa}\sum_{k=1}^K\sum_{h=1}^H\bar{\beta}_{\mathbf{w}}^{k-1}\min\left\{1,\max_{s'\in\mathcal{S}_h^k}\left\|\psi(s'\mid s_h^k,a_h^k)\right\|_{\left(\bar{\Sigma}_h^{k-1}\right)^{-1}}\right\}}_{\texttt{variance part2}} \\ &+\underbrace{2\sum_{k=1}^K\sum_{h=1}^H\Gamma_{h,\theta}^{k-1} + 2\frac{H}{\kappa}\sum_{k=1}^K\sum_{h=1}^H\Gamma_{h,\mathbf{w}}^{k-1}}_{\texttt{bias part}} + 4H\sqrt{2T\log(1/\delta)}.\end{aligned}$$

偏差。 现在我们设定 $w_{k,j}=\gamma^{k-j},\gamma\in(0,1)$，与方程 (62) 相同，我们有

$$2\sum_{k=1}^K\sum_{h=1}^H\Gamma_{h,\theta}^{k-1} \leq 4L_\phi^2\sqrt{\frac{d}{\lambda_\theta}}\frac{1}{(1-\gamma)^{3/2}}\sum_{p=1}^{K-1}\sum_{h=1}^H\left\|\theta_h^p-\theta_h^{p+1}\right\|_2,$$

然后我们设定 $\alpha_{k,j}=\gamma^{k-j},\gamma\in(0,1)$，类似于方程 (63)，我们有

$$\begin{aligned}&2\frac{H}{\kappa}\sum_{k=1}^K\sum_{h=1}^H\Gamma_{h,\mathbf{w}}^{k-1} \\ &\leq 4\frac{H}{\kappa}L_\psi^2\sqrt{\frac{d}{\lambda_{\mathbf{w}}}}\frac{1}{(1-\gamma)^{3/2}}\sum_{p=1}^{K-1}\sum_{h=1}^H\left\|\mathbf{w}_h^p-\mathbf{w}_h^{p+1}\right\|_2,\end{aligned}$$

方差。与方程（64）相同，我们有

$$\begin{aligned}&4H\beta_\theta\sum_{k=1}^K\sum_{h=1}^H\min\left\{1,\left\|\phi(s_h^k,a_h^k)\right\|_{\left(\Lambda_h^{k-1}\right)^{-1}}\right\} \\ &\leq 4H\beta_\theta\sqrt{KH}\sqrt{H2d\left(K\log\frac{1}{\gamma}+\log\left(1+\frac{L_\phi^2}{\lambda_\theta d(1-\gamma)}\right)\right)}.\end{aligned}$$

对于第二项，

$$
\begin{aligned}
&4\frac{H}{\kappa}\sum_{k=1}^{K}\sum_{h=1}^{H}\bar{\beta}_{\mathbf{w}}^{k-1}\min\left\{1,\max_{s'\in\mathcal{S}_h^k}\left\|\psi(s'\mid s_h^k,a_h^k)\right\|_{\left(\bar{\Sigma}_h^{k-1}\right)^{-1}}\right\}\\
&\le 4\frac{H}{\kappa}\bar{\beta}_{\mathbf{w}}^{K}\sum_{k=1}^{K}\sum_{h=1}^{H}\min\left\{1,\max_{s'\in\mathcal{S}_h^k}\left\|\psi(s'\mid s_h^k,a_h^k)\right\|_{\left(\bar{\Sigma}_h^{k-1}\right)^{-1}}\right\}\\
&\le 4\frac{H^2}{\kappa}\bar{\beta}_{\mathbf{w}}^{K}\sqrt{K}\sqrt{\sum_{k=1}^{K}\min\left\{1,\max_{s'\in\mathcal{S}_h^k}\left\|\psi(s'\mid s_h^k,a_h^k)\right\|^2_{\left(\bar{\Sigma}_h^{k-1}\right)^{-1}}\right\}}\\
&\le 4\frac{H^2}{\kappa}\bar{\beta}_{\mathbf{w}}^{K}\sqrt{K}\sqrt{\sum_{k=1}^{K}\min\left\{1,\sum_{s'\in\mathcal{S}_h^k}\left\|\psi(s'\mid s_h^k,a_h^k)\right\|^2_{\left(\bar{\Sigma}_h^{k-1}\right)^{-1}}\right\}}.
\end{aligned}
$$

根据引理 9（势引理），我们知道 $\forall h\in[H]$，于是有

$$
\begin{aligned}
&\sum_{k=1}^{K}\min\left\{1,\max_{s'\in\mathcal{S}_h^k}\left\|\psi(s'\mid s_h^k,a_h^k)\right\|^2_{\left(\bar{\Sigma}_h^{k-1}\right)^{-1}}\right\}\\
&\le 2d\left(K\log\frac{1}{\gamma}+\log\left(1+\frac{UL_\psi^2}{\lambda_{\mathbf{w}}d(1-\gamma)}\right)\right),
\end{aligned}
$$

因此有

$$
\begin{aligned}
&4\frac{H}{\kappa}\sum_{k=1}^{K}\sum_{h=1}^{H}\bar{\beta}_{\mathbf{w}}^{k-1}\min\left\{1,\max_{s'\in\mathcal{S}_h^k}\left\|\psi(s'\mid s_h^k,a_h^k)\right\|_{\left(\bar{\Sigma}_h^{k-1}\right)^{-1}}\right\}\\
&\le 4\frac{H^2}{\kappa}\bar{\beta}_{\mathbf{w}}^{K}\sqrt{K}\sqrt{2d\left(K\log\frac{1}{\gamma}+\log\left(1+\frac{UL_\psi^2}{\lambda_{\mathbf{w}}d(1-\gamma)}\right)\right)}.
\end{aligned}
$$

由于遗憾界中有一项 $HK\sqrt{\log(1/\gamma)}$，我们不能令 γ 接近 0，因此设 $\gamma\ge 1/K$，并有 $\log(1/\gamma)\le C(1-\gamma)$，其中 $C=\log K/(1-1/K)$。我们设$\lambda_\theta=d$，且 $\lambda_{\mathbf{w}}=d$。结合偏差部分和方差部分的上界，并以置信水平 $\delta=1/(4T)$，根据并集界，以至少 $1-1/T$ 的概率得到以下动态遗憾界，

$$
\begin{aligned}
\text{D-REG}_T &\le \mathcal{O}\Bigg(\frac{1}{(1-\gamma)^{3/2}}P_T^{\theta}+\frac{H}{\kappa}\frac{1}{(1-\gamma)^{3/2}}P_T^{\mathbf{w}}\\
&\quad+dH^2K\sqrt{1-\gamma}+\frac{dH^2K}{\kappa}\sqrt{1-\gamma}+H^{3/2}d\sqrt{HK}\Bigg)\\
&\le \mathcal{O}\left(\frac{Hd}{\kappa}\left(\frac{1}{(1-\gamma)^{3/2}}\Delta+HK\sqrt{1-\gamma}\right)+H^{3/2}d\sqrt{HK}\right).
\end{aligned}
$$

此外，通过将折扣因子最优地设置为 $1-\gamma=\max\left\{1/K,\sqrt{\Delta/T}\right\}$，我们有

$$
\text{D-REG}_T\le\begin{cases}\widetilde{\mathcal{O}}\left(\kappa^{-1}Hd\Delta^{1/4}T^{3/4}\right) & \text{when } \Delta\ge H/K,\\ \widetilde{\mathcal{O}}\left(\kappa^{-1}dH^{3/2}\sqrt{T}\right) & \text{when } \Delta<H/K.\end{cases}
$$

□

引理*22*的证明。我们首先考虑 E_h^k 的上界，根据 $\bar{Q}_h^k$ (26) 的定义、模型假设 (22) 和方程 (24)，我们有 $\forall a\in\mathcal{A},\ s\in\mathcal{S}$，

$$
\begin{aligned}
&r_h^k(s,a)+\left[\mathbb{P}_h^k\bar{V}_{h+1}^k\right](s,a)-\bar{Q}_h^k(s,a)\\
&=r_h^k(s,a)+\left[\mathbb{P}_h^k\bar{V}_{h+1}^k\right](s,a)-\phi(s,a)^\top\widehat{\theta}_h^k\\
&\quad-\beta_\theta\left\|\phi(s,a)\right\|_{\left(\Lambda_h^{k-1}\right)^{-1}}-[\widetilde{\mathbb{P}}_h^k\bar{V}_{h+1}^k](s,a)\\
&\quad-\frac{H}{\kappa}\bar{\beta}_{\mathbf{w}}^{k-1}\max_{s'\in\mathcal{S}_h^k}\left\|\psi(s'\mid s,a)\right\|_{\left(\bar{\Sigma}_h^{k-1}\right)^{-1}}\\
&=\phi(s,a)^\top\left(\theta_h^k-\widehat{\theta}_h^k\right)-\beta_\theta\left\|\phi(s,a)\right\|_{\left(\Lambda_h^{k-1}\right)^{-1}}\\
&\quad+\left(\left[\mathbb{P}_h^k\bar{V}_{h+1}^k\right](s,a)-\left[\widetilde{\mathbb{P}}_h^k\bar{V}_{h+1}^k\right](s,a)\right)\\
&\quad-\frac{H}{\kappa}\bar{\beta}_{\mathbf{w}}^{k-1}\max_{s'\in\mathcal{S}_h^k}\left\|\psi(s'\mid s,a)\right\|_{\left(\bar{\Sigma}_h^{k-1}\right)^{-1}}\\
&\le\Gamma_{h,\theta}^{k-1}+\frac{H}{\kappa}\Gamma_{h,\mathbf{w}}^{k-1},
\end{aligned}
$$

其中，最后一个不等式由引理4和引理6得出。类似地，我们可以得到 $E_h^k,\forall a\in\mathcal{A},\ s\in\mathcal{S}$ 的下界，

$$
\begin{aligned}
&\bar{Q}_h^k(s,a)-r_h^k(s,a)-\left[\mathbb{P}_h^k\bar{V}_{h+1}^k\right](s,a)\\
&=\phi(s,a)^\top\widehat{\theta}_h^k+\beta_\theta\left\|\phi(s,a)\right\|_{\left(\Lambda_h^{k-1}\right)^{-1}}+[\widetilde{\mathbb{P}}_h^k\bar{V}_{h+1}^k](s,a)\\
&\quad+\frac{H}{\kappa}\bar{\beta}_{\mathbf{w}}^{k-1}\max_{s'\in\mathcal{S}_h^k}\left\|\psi(s'\mid s,a)\right\|_{\left(\bar{\Sigma}_h^{k-1}\right)^{-1}}\\
&\quad-r_h^k(s,a)-\left[\mathbb{P}_h^k\bar{V}_{h+1}^k\right](s,a)\\
&=\phi(s,a)^\top\left(\widehat{\theta}_h^k-\theta_h^k\right)+\left(\left[\widetilde{\mathbb{P}}_h^k\bar{V}_{h+1}^k\right](s,a)-\left[\mathbb{P}_h^k\bar{V}_{h+1}^k\right](s,a)\right)\\
&\quad+\beta_\theta\left\|\phi(s,a)\right\|_{\left(\Lambda_h^{k-1}\right)^{-1}}\\
&\quad+\frac{H}{\kappa}\bar{\beta}_{\mathbf{w}}^{k-1}\max_{s'\in\mathcal{S}_h^k}\left\|\psi(s'\mid s,a)\right\|_{\left(\bar{\Sigma}_h^{k-1}\right)^{-1}}\\
&\le\Gamma_{h,\theta}^{k-1}+\frac{H}{\kappa}\Gamma_{h,\mathbf{w}}^{k-1}+2\beta_\theta\left\|\phi(s,a)\right\|_{\left(\Lambda_h^{k-1}\right)^{-1}}\\
&\quad+2\frac{H}{\kappa}\bar{\beta}_{\mathbf{w}}^{k-1}\max_{s'\in\mathcal{S}_h^k}\left\|\psi(s'\mid s,a)\right\|_{\left(\bar{\Sigma}_h^{k-1}\right)^{-1}},
\end{aligned}
$$

由此完成证明。 □

附录G 技术引理

本节给出若干有用的引理，主要涉及集中性以及自协调性质的一些导数。

A. 集中不等式

引理23 (Hoeffding 引理)。设 Z 是一个实随机变量，满足 $Z\in[a,b]$ 几乎必然成立。则

$$
\forall\lambda\in\mathbb{R}\quad\mathbb{E}\left[e^{\lambda(Z-\mathbb{E}[Z])}\right]\le\exp\left(\frac{\lambda^2(b-a)^2}{8}\right),
$$

随机变量 $(Z-\mathbb{E}[Z])$ 是$(b-a)/2$-次高斯。

引理 24 (Azuma-Hoeffding 不等式)。令 $M>0$为一个常数。令 $\{x_i\}_{i=1}^n$ 为关于滤过 $\{\mathcal{G}_i\}_i$（$\mathbb{E}[x_i\mid\mathcal{G}_i]=0$ 几乎必然成立，且 x_i 是 $\mathcal{G}_{i+1}$-可测的）的鞅差序列，使得对所有 $i\in[n]$，都有 $|x_i|\le M$ 成立

几乎必然成立。则，对于任意 $0 < \delta < 1$，以至少 $1-\delta$的概率，有

$$\sum_{i=1}^{n} x_i \le M\sqrt{2n\log(1/\delta)}.$$

定理7 (线性多臂老虎机的自归一化集中不等式 [32,定理 1])。设 $\{F_t\}_{t=0}^{\infty}$ 为一个滤过。设 $\{\eta_t\}_{t=0}^{\infty}$ 为一个实值随机过程，使得 η_t是 F_t-可测的，且对于某个 $R \ge 0$， η_t 是条件 R次高斯的，即 $\forall\lambda \in \mathbb{R}$, $\mathbb{E}\left[e^{\lambda\eta_t} \mid F_{t-1}\right] \le \exp(\frac{\lambda^2R^2}{2})$。设$\{X_t\}_{t=1}^{\infty}$为一个取值于 $\mathbb{R}^d$的随机过程，使得 X_t 是F_{t-1}-可测的。假设 V 是一个 $d\times d$ 正定矩阵。对于任意 $t \ge 0$，定义

$$V_t = V_0 + \sum_{s=1}^{t} X_s X_s^\top, \qquad S_t = \sum_{s=1}^{t} \eta_s X_s.$$

那么，对于任意 $\delta > 0$，以至少 $1-\delta$的概率，对所有$t \ge 0$,

$$\|S_t\|_{V_t^{-1}} \le \sqrt{2R^2\log\left(\frac{\det(V_t)^{\frac{1}{2}}\det(V_0)^{-\frac{1}{2}}}{\delta}\right)}.$$

定理 8（针对自一致多臂老虎机的自归一化集中不等式 [33, 定理 1]）。令 $\{\mathcal{F}_t\}_{t=0}^{\infty}$ 为一个滤过。令 $\{\eta_t\}_{t=0}^{\infty}$ 为一个鞅差序列，使得 η_t 是关于 $\mathcal{F}_t$ 可测的。令 $\{X_t\}_{t=0}^{\infty}$ 为定义在 $\mathbb{R}^d$ 上的随机过程，使得 X_t 是关于 $\mathcal{F}_{t-1}$ 可测的，且$\|X_t\|_2 \le 1$。此外，假设在给定 $\mathcal{F}_t$的条件下，我们有 $|\eta_t| \le 1$ 几乎处处成立，并记 $\sigma_t^2 = \mathbb{E}\left[\eta_t^2 \mid \mathcal{F}_{t-1}\right]$。对于任意 $t \ge 0$，定义

$$H_t = \sum_{s=1}^{t} \sigma_s^2 X_s X_s^\top + \lambda I_d, \qquad S_t = \sum_{s=1}^{t} \eta_s X_s,$$

满足 $\lambda > 0$。那么，对于任意 $\delta > 0$，以至少$1-\delta$的概率，对所有 $t \ge 0$,

$$\|S_t\|_{H_t^{-1}} \le \frac{\sqrt{\lambda}}{2} + \frac{2}{\sqrt{\lambda}}\log\left(\frac{\det(H_t)^{1/2}}{\delta\lambda^{d/2}}\right) + \frac{2}{\sqrt{\lambda}}d\log(2).$$

引理25。 假设 $U_0 = \lambda I_d$、 $U_t = U_{t-1} + A_t A_t^\top$以及$A_t \in \mathbb{R}^d$, 则

$$\forall p \in [t-1], \quad \sum_{s=1}^{p} \|A_s\|^2_{U_{t-1}^{-1}} \le d. \tag{71}$$

引理*25*的证明。

$$\begin{aligned}
&\sum_{s=1}^{p} \|A_s\|^2_{U_{t-1}^{-1}} = \sum_{s=1}^{p} \mathrm{Tr}(A_s^\top U_{t-1}^{-1} A_s) = \mathrm{Tr}\big(U_{t-1}^{-1}\sum_{s=1}^{p} A_s A_s^\top\big)\\
&\le \mathrm{Tr}\big(U_{t-1}^{-1}\sum_{s=1}^{p} A_s A_s^\top\big) + \mathrm{Tr}\big(U_{t-1}^{-1}\sum_{s=p+1}^{t-1} A_s A_s^\top\big)\\
&\quad + \mathrm{Tr}\big(U_{t-1}^{-1}\lambda\sum_{i=1}^{d}\mathbf{e}_i\mathbf{e}_i^\top\big) = \mathrm{Tr}(I_d) = d.
\end{aligned}$$

□

引理 26 $V_t = \sum_{s=1}^{t} w_{t\ s} X_s X_s^\top + \lambda I_d$ $V_0 = \lambda I_d$ $\| \ \|_2 \le L$ (行列式不等式)。 令， ， 。假设$\mathbf{x}$，我们有，

$$\det(V_t) \le \left(\lambda + \frac{L^2\sum_{s=1}^{t} w_{t,s}}{d}\right)^d.$$

证明。 现在我们有 $V_t = \sum_{s=1}^{t} w_{t,\ s} X_s X_s^\top + \lambda I_d$，对两边取迹，得到 $\mathrm{Tr}(V_t)$ 的上界

$$\begin{aligned}
\mathrm{Tr}(V_t) &= \mathrm{Tr}(\lambda I_d) + \sum_{s=1}^{t} w_{t,s}\mathrm{Tr}\left(X_s X_s^\top\right)\\
&= \lambda d + \sum_{s=1}^{t} w_{t,s}\|X_s\|_2^2 \le \lambda d + L^2\sum_{s=1}^{t} w_{t,s}.
\end{aligned} \tag{72}$$

根据行列式的定义以及 $\mathrm{Tr}(V_t)$ 的上界（72），我们可以得到 $\det(V_t)$ 的上界，

$$\begin{aligned}
\det(V_t) = \prod_{i=1}^{d}\lambda_i &\le \left(\frac{\sum_{i=1}^{d}\lambda_i}{d}\right)^d = \left(\frac{\mathrm{Tr}(V_t)}{d}\right)^d\\
&\le \left(\lambda + \frac{L^2\sum_{s=1}^{t} w_{t,s}}{d}\right)^d.
\end{aligned}$$

□

B. 自协调性质

基于（逆）链接函数 $\mu(\cdot)$的广义自协调性质，我们有以下引理，后续将用于推导引理28。

引理 27 (引理 9，出自 [33])。 对于任意 $z_1, z_2 \in \mathbb{R}$，有以下不等式：

$$\begin{aligned}
\mu'(z_1)\frac{1-\exp(-|z_1-z_2|)}{|z_1-z_2|} &\le \int_0^1 \mu'(z_1+v(z_2-z_1))\mathrm{d}v\\
&\le \mu'(z_1)\frac{\exp(|z_1-z_2|)-1}{|z_1-z_2|}.
\end{aligned}$$

此外， $\int_0^1 \mu'\ z_1+v\ z_2-z_1 \quad v \ge \mu'\ z_1\ \ 1+|z_1-((\))\mathrm{d}()()$。
$z_2|^{\ -1}$

下面的引理给出了 [33]中引理10的加权版本，可以很容易地证明。

引理 28。 其中 G_t 定义见（49）， H_t 定义见（50），以下不等式成立

$$\begin{aligned}
\forall\theta_1,\theta_2 \in \Theta, \quad & G_t(\theta_1,\theta_2) \ge (1+2S)^{-1}H_t(\theta_1),\\
& G_t(\theta_1,\theta_2) \ge (1+2S)^{-1}H_t(\theta_2).
\end{aligned}$$

引理 29 (引理 7，出自 [36])。 将 L_i 记为回合 i $L_i \triangleq$ 中累积奖励的绝对值，即 $\left|\sum_{t=\ i-1\ \Delta+1}^{i\Delta} r_t\ X_t\right|$ $_{()}()$，则

$$\Pr\left[\forall i \in [\lceil T/\Delta\rceil], L_i \le LS\Delta + 2R\sqrt{\Delta\ln\frac{T}{\sqrt{\Delta}}}\right] \ge 1-\frac{2}{T}.$$

附录 H 老虎机中的老虎机

A. BOB算法

我们将 T 轮次划分为长度相等、每回合包含 Δ 轮的回合，使得 $\Delta = \lceil d\sqrt{T} \rceil$。在每个回合中，我们使用不同的折扣因子 γ 运行 LB-WeightUCB。具体来说，γ 来自候选集 $\mathcal{W}$，

$$\mathcal{W} = \left\{ \gamma_i = 1 - d^{-\frac{1}{2}} 2^{1-i} \mid i \in [N] \right\}, \tag{73}$$

其中，$N = \lceil \log_2(T/\sqrt{d}) \rceil + 1$ 是候选值的数量；回顾一下，S 是底层回归参数范数的上界。

为了自适应地选择最优的 γ，我们将选择过程建模为一个对抗性老虎机问题。在这种表述中，每一步对应一个长度为 Δ的回合，因此总共有 $\lceil T/\Delta \rceil$ 轮次。在每一步中，老虎机算法从候选集 $\mathcal{W}$中选择一个 γ，观察 LB-WeightUCB 在相应折扣因子下获得的累积奖励，并将其用作反馈以更新其选择策略。

令 $\gamma_{\min}$($\gamma_{\max}$)为候选集 $\mathcal{W}$中的最小（最大）折扣因子，则显然可以验证

$$\gamma_{\min} = 1 - \frac{1}{\sqrt{d}}, \quad 1 > \gamma_{\max} \geq 1 - \frac{1}{T}.$$

B. 遗憾分析

定理9. *在不知道路径长度 P_T 的情况下，LB-WeightUCB 与老虎机中的老虎机机制结合，以至少 $1 - 3/T$ 的概率满足，*

$$\text{D-REG}_T = \sum_{t=1}^{T} \max_{\mathbf{x} \in \mathcal{X}} \mathbf{x}^\top \theta_t - \sum_{t=1}^{T} X_t^\top \theta_t = \widetilde{\mathcal{O}}\left(d^{3/4} P_T^{1/4} T^{3/4} \right).$$

定理 9 的证明. 我们首先对动态遗憾进行分解。令 $X_t^* \triangleq \operatorname{argmax}_{\mathbf{x} \in \mathcal{X}} \mathbf{x}^\top \theta_t$，则有

$$\begin{aligned}
\text{D-REG}_T &= \sum_{t=1}^{T} \left(X_t^{*\top} \theta_t - X_t^\top \theta_t \right) \\
&= \underbrace{\sum_{t=1}^{T} X_t^{*\top} \theta_t - \sum_{i=1}^{\lceil T/\Delta \rceil} \sum_{t=(i-1)\Delta+1}^{i\Delta} X_t^{\widetilde{\gamma}^{*}\top} \theta_t}_{\texttt{base-regret}} \\
&\quad + \underbrace{\sum_{i=1}^{\lceil T/\Delta \rceil} \sum_{t=(i-1)\Delta+1}^{i\Delta} \left(X_t^{\widetilde{\gamma}^{*}\top} \theta_t - X_t^{\gamma_i \top} \theta_t \right)}_{\texttt{meta-regret}},
\end{aligned}$$

其中，$\widetilde{\gamma}^*$ 是候选集中用于逼近最优折扣因子 $\gamma^* = 1 - \max\{1/T, \sqrt{P_T/(dT)}\}$ 的最佳折扣因子。

基础遗憾。根据（37）和并集界，我们以至少 $1 - 2N\delta$ 的概率得到，

$$\begin{aligned}
\texttt{base-regret} &= \sum_{t=1}^{T} X_t^{*\top} \theta_t - \sum_{i=1}^{\lceil T/\Delta \rceil} \sum_{t=(i-1)\Delta+1}^{i\Delta} X_t^{\widetilde{\gamma}^{*}\top} \theta_t \\
&\leq \sum_{i=1}^{\lceil T/\Delta \rceil} \widetilde{\mathcal{O}}\left(\frac{1}{(1-\widetilde{\gamma}^*)^{3/2}} P_i + d(1-\widetilde{\gamma}^*)^{1/2} \Delta \right) \\
&\leq \widetilde{\mathcal{O}}\left(\frac{1}{(1-\widetilde{\gamma}^*)^{3/2}} P_T + d(1-\widetilde{\gamma}^*)^{1/2} T \right) \\
&\leq \widetilde{\mathcal{O}}\left(\frac{1}{2^{3/2}(1-\gamma^*)^{3/2}} P_T + d(1-\gamma^*)^{1/2} T \right) \\
&\leq \widetilde{\mathcal{O}}(d^{3/4} P_T^{1/4} T^{3/4}).
\end{aligned} \tag{74}$$

我们知道，$\gamma^* = 1 - \max\{1/T, \sqrt{P_T/(dT)}\}$是最优折扣因子。由于 $\gamma^* \in [\gamma_{\min}, \gamma_{\max}]$，候选集$\mathcal{W}$ 覆盖了 γ^*。此外，由于$\mathcal{W}$采用几何间隔，因此存在某个 $\widetilde{\gamma}^* \in \mathcal{W}$ 使得

$$1 - \widetilde{\gamma}^* \leq 1 - \gamma^* \leq 2(1 - \widetilde{\gamma}^*).$$

元遗憾.元遗憾的分析沿用基于窗口的算法 [52, Proposition 1]的证明。

$$\begin{aligned}
&\texttt{meta-regret} \\
&= \sum_{i=1}^{\lceil T/\Delta \rceil} \sum_{t=(i-1)\Delta+1}^{i\Delta} \left(X_t^{\widetilde{\gamma}^{*}\top} \theta_t - X_t^{\gamma_i \top} \theta_t \right) \\
&= \sum_{i=1}^{\lceil T/\Delta \rceil} \sum_{t=(i-1)\Delta+1}^{i\Delta} \left(X_t^{\widetilde{\gamma}^{*}\top} \theta_t + \eta_t - X_t^{\gamma_i \top} \theta_t - \eta_t \right) \\
&= \sum_{i=1}^{\lceil T/\Delta \rceil} \sum_{t=(i-1)\Delta+1}^{i\Delta} \left(r_t^{\widetilde{\gamma}^{*}} - r_t^{\gamma_i} \right) \\
&= \sum_{i=1}^{\lceil T/\Delta \rceil} \left(L_i^{\widetilde{\gamma}^{*}} - L_i^{\gamma_i} \right),
\end{aligned}$$

根据引理 29，我们知道，以至少 $1 - \frac{2}{T}$ 的概率，有 $\forall i \in [\lceil T/\Delta \rceil], L_i \leq LS\Delta + 2R\sqrt{\Delta \ln \frac{T}{\sqrt{\Delta}}}$，我们定义 $L_{\max} \triangleq LS\Delta + 2R\sqrt{\Delta \ln \frac{T}{\sqrt{\Delta}}}$。我们选择 Exp3.IX [53] 作为元算法。于是，以至少 $1 - \delta$ 的概率，有

$$\begin{aligned}
\sum_{i=1}^{\lceil T/\Delta \rceil} \left(L_i^{\widetilde{\gamma}^{*}} - L_i^{\gamma_i} \right) &= \widetilde{\mathcal{O}}\left(L_{\max} \sqrt{\frac{T}{\Delta} N} \right) \\
&= \widetilde{\mathcal{O}}\left(\sqrt{\Delta T N} \right) = d^{1/2} T^{3/4}.
\end{aligned} \tag{75}$$

结合基础遗憾 (74) 和元遗憾 (75) 的上界，并利用并集界，令 $\delta = {}^{1}/_{(2N+1)T}$，我们以至少 $1 - 3/T$ 的概率得到，

$$\text{D-REG}_T = \widetilde{\mathcal{O}}\left(d^{3/4} P_T^{1/4} T^{3/4} + d^{1/2} T^{3/4} \right).$$

至此，我们完成了证明。 □

致谢

本研究得到国家自然科学基金委员会（62361146852、62206125）的资助。Peng Zhao获得了小米基金会的部分资助。

参考文献

[1] W. C. Cheung、D. Simchi-Levi和R. Zhu，“学习在非平稳性下进行优化”，载于第22届国际人工智能与统计会议（AISTATS）论文集，2019年，第1079–1087页。[2] Y. Russac、C. Vernade和O. Capp´e，“面向非平稳环境的加权线性老虎机”，载于神经信息处理系统进展第32卷（NeurIPS），2019年，第12 040–12 049页。[3] P. Zhao、L. Zhang、Y. Jiang和Z.-H. Zhou，“一种解决非平稳线性多臂老虎机问题的简单方法”，载于第23届国际人工智能与统计会议（AISTATS）论文集，2020年，第746–755页。[4] Y. Russac、O. Capp´e和A. Garivier，“非平稳广义线性老虎机算法”，arXiv预印本，第arXiv:2003.10113卷，2020年。[5] B. Kim和A. Tewari，“非平稳随机线性老虎机的随机探索”，载于第36届人工智能不确定性会议（UAI）论文集，2020年，第71–80页。[6] L. Faury、Y. Russac、M. Abeille和C. Calauz`enes，“参数漂移下广义线性老虎机的遗憾界”，arXiv预印本，第arXiv:2103.05750卷，2021年。[7] Y. Russac、L. Faury、O. Capp´e和A. Garivier，“带遗忘机制的广义线性老虎机的自协调分析”，载于第24届国际人工智能与统计会议（AISTATS）论文集，2021年，第658–666页。[8] C.-Y. Wei和H. Luo，“无需先验知识的非平稳强化学习：一种最优黑箱方法”，载于第34届学习理论会议（COLT）论文集，2021年，第4300–4354页。[9] Y. Deng、X. Zhou、B. Kim、A. Tewari、A. Gupta和N. Shroff，“面向非平稳环境的加权高斯过程老虎机”，载于第25届国际人工智能与统计会议（AISTATS）论文集，2022年，第6909–6932页。[10] Y. Liu、B. Van Roy和K. Xu，“通过预测采样进行非平稳老虎机学习”，载于第26届国际人工智能与统计会议论文集（AISTATS），2023年，第6215–6244页。[11] J. Wang、P. Zhao和Z.-H. Zhou，“重新审视非平稳参数化老虎机的加权策略”，载于第26届国际人工智能与统计会议论文集（AISTATS），第206卷，2023年，第7913–7942页。[12] S. Tomkins、P. Liao、P. V. Klasnja和S. A. Murphy，“Intelligentpooling：面向移动健康的实用汤普森采样”，Machine Learning，第110卷，第9期，第2685–2727页，2021年。[13] W. Huleihel、S. Pal和O. Shayevitz，“在非平稳环境中学习用户偏好”，载于第24届国际人工智能与统计会议（AISTATS）论文集，2021年，第1432–1440页。[14] C. Jin、Z. Yang、Z. Wang和M. I. Jordan，“具有线性函数逼近的可证明高效强化学习”，载于第33届学习理论会议（COLT）论文集，2020年，第2137–2143页。[15] A. Touati和P. Vincent，“在线性马尔可夫决策过程中的高效非平稳学习”，arXiv预印本，第arXiv:2010 12870卷，2020年。[16] T. Hwang和M. Oh，“具有多项Logit函数近似的基于模型的强化学习”，载于第37届AAAI人工智能会议（AAAI）论文集，2023年，第7971–7979页。[17] L.-F. Li、Y.-J. Zhang、P. Zhao和Z.-H. Zhou，“具有多项Logit函数近似的可证明高效强化学习”，载于神经信息处理系统进展第37卷（NeurIPS），2024年，第58 539–58 573页。[18] H. Zhong、Z. Yang、Z. Wang和C. Szepesv´ari，“乐观策略优化在非平稳马尔可夫决策过程中具有可证明的高效率”，arXiv预印本，第arXiv: 2110.08984卷，2022年。[19] L.-F. Li、P. Zhao和Z.-H. Zhou，“对抗性线性混合马尔可夫决策过程的近最优动态遗憾”，载于神经信息处理系统进展第37卷（NeurIPS），2024年，第55 858–55 883页。[20] O. Besbes、Y. Gur和A. Zeevi，“具有非平稳奖励的随机多臂老虎机问题”，神经信息处理系统进展第27卷（NeurIPS），2014年。[21] P. Zhao、Y. Zhang、L. Zhang和Z.-H. Zhou，“自适应性与非平稳性：在线凸优化中依赖于问题的动态遗憾”，Journal of Machine Learning Research，第25卷，第98:1–98:52页，2024年。[22] Z.-H. Zhou，“开放环境机器学习”，National Science Review，第9卷，第8期，2022年。[23] K. Crammer、Y. Mansour、E. Even-Dar和J. W. Vaughan，“具有概念漂移的遗憾最小化”，载于第23届学习理论会议（COLT）论文集，2010年，第168–180页。

[24] C. Chiang, C. Lee，和 C. Lu，“Beating 中的多臂老虎机 g逐渐 y 演变 g 世界”，载于《第26届学习理论年度会议论文集（COLT）》，2013年，第210–227页。[25] J. Gama、I. Zliobaite、A. Bifet、M. Pechenizkiy 和 A. Bouchachia，“概念漂移适应综述”，《ACM计算调查》，第46卷，第4期，第44:1–44:37页，2014年。[26] A. Ghosh、A. Sankararaman、K. Ramchandran、T. Javidi 和 A. Mazumdar，“非平稳匹配市场中的竞争多臂老虎机”，IEEE《信息论汇刊》，第70卷，2850，2024年4，2831，页码–。[27] S. Li、S. Zhao、Z. Cao、S. Huang 和 S. Chen，“含噪且偏移标签分布下的鲁棒领域适应”，《计算机科学前沿》，第19卷，第3期，第193310页，2025年。[28] C. Anagnostopoulos、D. K. Tasoulis、N. M. Adams、N. G. Pavlidis 和 D. J. Hand，“采用自适应遗忘的流式分类在线线性与二次判别分析”，《统计分析与数据挖掘》，第5卷，第2期，第139–166页，2012年。[29] P. Zhao、X. Wang、S. Xie、L. Guo 和 Z.-H. Zhou，“无分布假设的单遍学习”，IEEE知识与数据工程汇刊，第33卷，第951–963页，2021年。[30] L. Guo、L. Ljung 和 P. Priouret，“遗忘因子RLS算法的性能分析”，《自适应控制与信号处理国际期刊》，第7卷，第6期，第525–537页，1993年。[31] Y. Chu 和 C. M. Mak，“用于分布式估计的变遗忘因子扩散递归最小二乘算法”，《信号处理》，第140卷，第219–225页，2017年。[32] Y. Abbasi-Yadkori、D. P´al 和 C. Szepesv´ari，“线性随机多臂老虎机的改进算法”，载于《第24届神经信息处理系统进展（NIPS）》，2011年，第2312–2320页。[33] L. Faury、M. Abeille、C. Calauz`enes 和 O. Fercoq，“逻辑回归多臂老虎机的改进乐观算法”，载于《第37届国际机器学习会议（ICML）论文集》，2020年，第3052–3060页。[34] M. Abeille、L. Faury 和 C. Calauz`enes，“逻辑回归多臂老虎机的实例级极小极大最优算法”，载于《第24届人工智能与统计国际会议（AISTATS）论文集》，2021年，第3691–3699页。[35] P. Zhao 和 L. Zhang，“重新审视非平稳线性多臂老虎机”，arXiv预印本，卷号arXiv:2103.05324，2021年。[36] P. Zhao、L. Zhang、Y. Jiang 和 Z.-H. Zhou，“ 非平稳线性多臂老虎机的一种简单方法 ”，arXiv 预印本，卷号arXiv:2103 05324，2021年。[37] J. Suk 和 S. Kpotufe，“追踪多臂老虎机中最显著的臂切换”，载于《第35届学习理论会议（COLT）论文集》，2022年，第2160–2182页。[38] Y. Abbasi-Yadkori、A. Gy¨orgy 和 N. Lazi´c，“重新审视非平稳随机多臂老虎机的动态遗憾”，《Machine Learning Research期刊》，第24卷，第288期，第1–37页，2023年。[39] G. Clerici、P. Laforgue 和 N. Cesa-Bianchi，“带记忆的线性多臂老虎机：从收益递减到收益递增”，arXiv预印本，卷号arXiv:2302 08345，2023年。[40] W. C. Cheung、D. Simchi-Levi 和 R. Zhu，“对冲漂移：非平稳条件下的优化学习”，《管理科学》，第68卷，第3期，第1696–1713页，2022年。[41] S. Filippi、O. Capp´e、A. Garivier 和 C. Szepesv´ari，“参数化老虎机：广义线性情形”，载于《第23届神经信息处理系统进展（NIPS）》，2010年，第586–594页。[42] Y.-J. Zhang、S.-A. Xu、P. Zhao 和 M. Sugiyama，“广义线性老虎机：通过单遍更新实现近乎最优遗憾”，载于《神经信息处理系统进展第38卷（NeurIPS）》，2025年，即将发表。[43] Z. Jia、L. Yang、C. Szepesv´ari 和 M. Wang，“采用价值目标回归的基于模型的强化学习”，载于《第2届动力学与控制学习年度会议（L4DC）论文集》，2020年，第666–686页。[44] A. Ayoub、Z. Jia、C. Szepesv´ari、M. Wang 和 L. Yang，“采用价值目标回归的基于模型的强化学习”，载于《第37届国际机器学习会议（ICML）论文集》，2020年，第463–474页。[45] L. Li、Y. Lu 和 D. Zhou，“广义线性上下文多臂老虎机的可证明最优算法”，载于《第34届国际机器学习会议（ICML）论文集》，2017年，第2071–2080页。[46] L. Zhang、T. Yang、R. Jin、Y. Xiao 和 Z.-H. Zhou，“单比特反馈下的在线随机线性优化”，载于《第33届国际机器学习会议（ICML）论文集》，2016年，第392–401页。

[47] K.-S. Jun、A. Bhargava、R. D. Nowak 和 R. Willett，“可扩展广义线性老虎机：在线计算与哈希”，载于神经信息处理系统进展第*30*卷（*NIPS*），2017年，第99–109页。[48] S. Dong、T. Ma 和 B. V. Roy，“汤普森采样在逻辑回归多臂老虎机上的表现”，载于第*32*届学习理论会议（*COLT*）论文集，2019年，第1158–1160页。[49] C. Tekin 和 M. Liu，“休眠与非休眠老虎机的在线学习”，*IEEE*《信息论汇刊》，第58卷，第8期，第5588–5611页，2012年。[50] P. N. Karthik、V. Y. F. Tan、A. Mukherjee 和 A. Tajer，“非休眠老虎机中具有固定置信度的最优最佳臂识别”，*IEEE*《信息论汇刊》，第70卷，7384，2024年10，7349，页码–。[51] D. Baby 和 Y. Wang，“总变差有界序列的在线预测”，载于神经信息处理系统进展第*32*卷（*NeurIPS*），2019年，第11 069–11 079页。[52] W. C. Cheung、D. Simchi-Levi 和 R. Zhu，“对冲漂移：非平稳条件下的优化学习”，*arXiv*预印本，卷号arXiv:1903.01461，2019年。[53] G. Neu，“不再探索：非随机老虎机高概率遗憾界的改进”，载于第*28*届神经信息处理系统进展（*NIPS*），2015年，第3168–3176页。

Jing Wang 于2021年获得南京航空航天大学理学学士学位。目前，他正在南京大学人工智能学院攻读博士学位。他的研究兴趣主要集中在机器学习和数据挖掘领域。

Peng Zhao （IEEE会员）于2016年获得同济大学理学学士学位，并于2021年获得南京大学博士学位。他目前是南京大学人工智能学院的助理教授。他的研究兴趣在于机器学习基础，重点关注在线学习、多臂老虎机和优化。他在顶级期刊（如*Journal of MachineLearning Research* 和IEEE/ACM汇刊）以及ICML、NeurIPS和COLT等顶级会议上共发表了60多篇论文。他经常担任ICML和NeurIPS的领域主席，并担任*Machine Learning* （施普林格）的执行编辑。

周志华 （IEEE会士）于1996年、1998年和2000年分别在中国南京大学获得计算机科学理学学士、理学硕士和哲学博士学位（荣誉学位）。2001年，他加入南京大学，任助理教授，目前为教授兼副校长。他还是LAMDA团队的创始主任。他著有《集成方法：基础与算法》《进化学习：理论与算法进展》和《Machine Learning》等书籍，并在国际顶级期刊或会议论文集中发表了200多篇论文。他拥有30多项专利。他的研究兴趣主要集中于人工智能、机器学习和数据挖掘。他是美国计算机协会、AAAI、美国科学促进会等组织的会士。他曾获得多项奖励和荣誉，包括国家自然科学奖、IEEE计算机学会爱德华·J·麦克拉斯基技术成就奖以及CCF-ACM人工智能奖。他创办了亚洲机器学习会议（ACML）。他担任IJCAI理事会主席、人工智能讲义（施普林格）系列丛书编辑、《人工智能》杂志咨询委员会委员、《style id='3'>计算机科学前沿》主编，以及《style id='5'>中国科学：信息科学》副主编。