用于非平稳马尔可夫决策过程的重启式贝叶斯在线变点检测

Reda Alami[*]T
echnology 创新研究院 阿布扎比，阿拉伯
联合酋长国 reda.alami@tii.ae

Mohammed Mahfoud[*]慕尼黑
工业大学 慕尼黑，德国
mo.mahfoud@tum.de

Eric Moulines
Ecole Polytech
nique 巴黎，法国
eric.moulines@polytechnique.edu

arXiv:2304.00232v1 [cs.LG] 1 Apr 2023

摘要

我们研究非平稳强化学习（RL）环境中的学习问题，其中该环境可由分段平稳离散时间马尔可夫决策过程（MDP）完整描述。我们提出了重启贝叶斯在线变点检测算法（R-BOCPD）的一种变体，该变体适用于源自更一般的多项分布的输入流，并在误报率和检测延迟方面提供近乎最优的理论保证。基于此，我们提出了适用于状态转移核从多项分布中采样的马尔可夫决策过程的 UCRL2 算法改进版本，称为 R-BOCPD-UCRL2。我们进行了有限时间性能分析，并证明 R-BOCPD-UCRL2 具有良好的遗憾界 $\mathcal{O}\left(DO\sqrt{ATK_T\log\binom{T}{\delta}+K_T\log\frac{K_T}{\delta}\min_\ell \mathbf{KL}(\boldsymbol{\theta}^{(\ell+1)}\,\|\,\boldsymbol{\theta}^{(\ell)})}\right)$，其中 D 是定义分段平稳马尔可夫决策过程设定的马尔可夫决策过程集合中的最大 MDP 直径，O 是有限的状态数（在所有变点之间保持不变），A 是有限的动作数（在所有变点之间保持不变），K_T 是截至时域 T 的变点数，$\boldsymbol{\theta}^{(\ell)}$ 是区间 $[c_\ell, c_{\ell+1}$ 内的转移核），我们假设该转移核在状态集合 O 上服从多项分布。有趣的是，性能界不会直接随 MDP 状态转移分布和奖励的变化幅度缩放，因此也可用于模拟突变。在实践中，R-BOCPD-UCRL2 在多种合成环境场景下优于当前最先进方法。我们还提供详细的实验设置以及代码仓库（将在发表后提供），以便轻松复现实验。

1 引言

在典型的序贯在线决策场景中，决策者（我们称之为 智能体）首先观测其当前状态，并与环境进行交互。随后，它从当前状态下的可能动作集合中选择一个动作，转移到由生成其状态转移分布的随机过程所决定的另一个状态，并获得一个随机奖励，用于衡量当前决策/动作相对于其在先前状态下本可以采取的最优动作集合的质量。通过这种持续交互，智能体试图学习最优的决策方案或策略，以最大化其累积奖励。这准确描述了强化学习（RL）问题，而事实证明，强化学习对于在许多领域中建模各种重要问题非常有用。

在经典强化学习中，通常假设状态转移分布和奖励由一个在整个学习过程中保持平稳的随机过程生成。然而，在真实的在线学习环境中，这一假设相当严格。因此，有必要定义一种新的强化学习变体，通常称为非平稳强化学习。对后者的研究受到多种应用的推动，包括消费者决策建模（Xu &Yun (2020 年)), 服务提供商对客户的适应以及定价（Taylor (2018 年); Kanoria & Qian (2019 年); Bimpikiset al. (2019 年); Gurvich et al. (2019 年)), 无线通信网络（Zhou &Bambos (2015);Zhou etal. (2016年)), 流行病网络与控制（Nowzari et al. (2016年); Kiss et al. (2017年)), 库存管理（Agrawal& Jia

[*]等额贡献。

(2019 年); Huh & Rusmevichientong (2009 年)), 仅举几例。在上述应用领域中，非平稳性通常源于突发变化，而这类变化在高度敏感的环境中可能造成严重影响。

尽管在非平稳强化学习环境中准确建模潜在变化通常非常困难，但我们尤其关注探索一种变体：其中非平稳性可以完全由分段平稳的离散时间马尔可夫决策过程（MDP）建模。更具体地说，我们假设这样一种情况：相关 MDP 的状态转移分布和奖励可以在未知的预定义时间点任意变化，我们将这些时间点称为变点。通过这种方式，我们可以准确且动态地处理上述突变环境。该问题的示意图见图1。

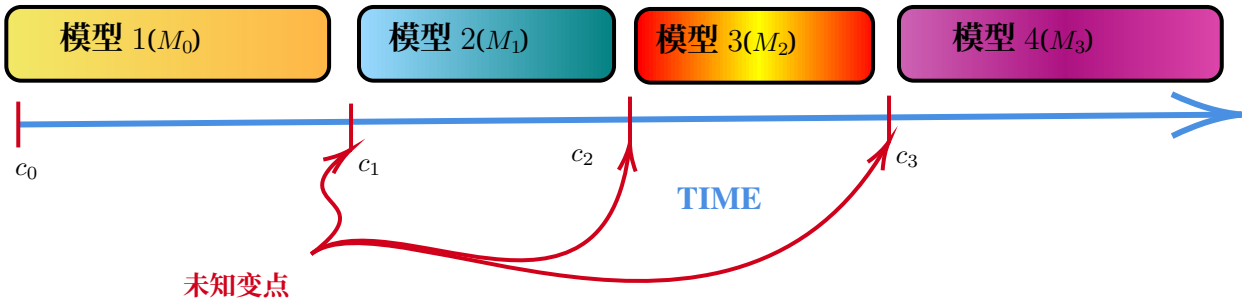

图1：分段平稳离散时间马尔可夫决策过程。从变点 c_ℓ 开始，环境由马尔可夫决策过程 M_ℓ 建模。

主要贡献： 我们的主要贡献总结如下

1. 我们将重启式贝叶斯在线变点检测算法 R-BOCPD 扩展到更一般的设置，其中在线观测流根据多项分布生成，并针对虚警率和检测延迟控制提供（近似）最优的理论保证。2. 我们提出了 UCRL2 的改进版本，将 R-BOCPD 作为检测学习环境中变化的手段，并提供近似最优的遗憾界；这些遗憾界仅依赖少量历史观测，并允许状态转移分布和奖励发生任意幅度的变化。3. 我们通过实验展示了相关结果，并将该算法的性能与当前最先进方法进行了比较。论文发表后，我们将提供一个 GitHub 仓库，以确保结果可复现。

下面，我们列出文献中与本研究问题相关的一些关键结果。

2 相关工作

为概述以往关于切换强化学习（MDP强化学习）的研究，我们首先介绍平稳的基础MDP领域取得的一些主要结果，然后阐述非平稳强化学习领域近期的一些关键贡献与模型。最后，我们还介绍一些主要来自时间序列领域的结果（大多属于算法层面）。

2.1 平稳马尔可夫决策过程中的强化学习

我们仅讨论与本文问题直接相关的贡献。具体而言，我们区分折扣奖励（Sidford 等人（2018a;b）；Wang（2020））和非折扣奖励（Auer 等人（2008b）；Azar 等人（2017）；Dann 等人（2017）；Jin 等人（2018）；Zanette 和 Brunskill（2019））两种情形。对于前一种情形，Sidford 等人（2018a;b）；Wang（2020）提出了样本复杂度近乎最优的算法。对于后一种情形，算法 $\mathcal{A}$ 的遗憾定义为其获得的奖励与最优算法获得的奖励之差，已有大量遗憾界被提出。Auer 等人（2008b）首先给出了极小极大下界 $\Omega(\sqrt{DOAT})$，其中 D 是第 3 节所定义的 MDP直径，并假设在所考虑的时间范围 T 内，马尔可夫决策过程的状态转移分布和奖励均为时不变（O 表示状态数量，A 表示动作数量）。基于上置信界，Auer 等人（ 2008b ）还提出了 UCRL2，这是一种可实现遗憾界 $\tilde{\mathrm{O}}(DO\sqrt{AT})$ 的算法。后来也有人提出了具有更优遗憾界的 UCRL2 变体，但由于它们在实践中效率较低（例如 Zhang & Ji（2019）中的计算效率问题），尽管具有极小极大最优性，本文仍不予讨论。

2.2 马尔可夫决策过程建模的非平稳强化学习

以一种相当朴素的方式，Auer 等人（2008a）已经考虑了环境中预先设定且智能体已知的变化次数 ℓ。基于此，Restarted-UCRL2 会定期重启 UCRL2。近期，若干论文提出了适用于表格型设定下非平稳强化学习的算法：Gajane 等人（2018）；Ortner 等人（2020）；Cheung 等人（2020）。这类算法假设，直到当前回合（记为 k）为止，马尔可夫决策过程在所有回合中保持不变，并根据截至 $k-1$ 的数据估计状态转移分布和奖励。如果回合 $k-1$ 和 k 之间发生变化（这通常是可能的），估计量就会产生偏差；Ortner 等人（2020）证明，这些算法会遭受与偏差范数成比例的线性遗憾。作为补救方法，Gajane 等人（2018）；Cheung 等人（2020）提出了滑动窗口方法，在该方法中，估计量更重视近期观测到的转移，并降低较早观测数据的权重。正如 Cheung 等人（2020）所示，Gajane 等人（2018）的方法会导致次优遗憾界。Cheung 等人（2020）针对 Gajane 等人（2018）提出了置信区间扩宽变体，该变体对平滑变化的马尔可夫决策过程具有更优的遗憾界，因此仅适用于这样一种情形：环境的状态转移分布被设定为至多发生初始化时假定的变化预算所允许的变化。Ortner 等人（2020）会定期重启算法，并在每次重启时丢弃过去的数据；但要获得其动态遗憾*界，它需要比 Cheung 等人（2020）掌握更多关于重启间状态转移分布和奖励变化的信息。虽然与本文考虑的设定没有直接关联，我们也希望强调 Yu 和 Mannor（2009）；Neu 等人（2010）；Arora 等人（2012）；Dick 等人（2014）；Jin 等人（2020）；Rivera Cardoso 等人（2019）的贡献：这些工作假设状态转移分布在整个学习过程中保持不变，同时允许奖励发生变化。受篇幅限制，我们决定省略其他与本文设定关联不那么直接的贡献。

2.3 在线变点检测背景

在线变点检测领域设计了变点检测算法，用于检测随机过程的分布从一种概率分布变为另一种。研究人员在多种问题设定和不同模型假设下，考察了这些算法在虚警率和检测延迟方面的最优性。Shiryaev 的工作奠定了主要理论基础（Shiryaev，1963 年）。现有的在线变点检测方法总体上分为两类：贝叶斯方法和非贝叶斯算法。前者提供检测结果的不确定性度量，而后者主要关注衡量变点前后数据统计特征的差异。若干用于在线变点检测的贝叶斯方法（Alami 等人（2020 年）；Agudelo-Espa˜na 等人（2020 年）；Knoblauch 和 Damoulas（2018 年）；Saatc¸i 等人（2010 年））基于标准的贝叶斯在线变点检测方法（Fearnhead 和 Liu（2007 年）），该方法递归地对距上次变点以来经过时间的后验概率进行建模。另一方面，非贝叶斯方法主要依赖似然比检验（Severo 和 Gama（2006 年）；Maillard（2019b）；Page（1954 年））；当出现离群值时，若最新观测的概率降低，也会导致假阳性。

3 问题表述

一个马尔可夫决策过程实例可简洁地表示为元组 $M(\mathsf{O}, \mathsf{A}\ P, R, T)$，其中 $\mathsf{O} = \{1, ..., O\}$ 表示有限状态集（$O = |\mathsf{O}|$），$\mathsf{A} = \{1, ..., A\}$ 表示有限动作集（$A = |\mathsf{A}|$），T为有限时间范围，$\{R_t\}_{t=1 o\in\mathsf{O}, a\in\mathsf{A}}^{T}$ 为分布式奖励序列。对于给定的 t 和状态-动作对（o, a），$r_t \sim R_t(o, a)$按照某个定义在 $[0, 1]$上的未知分布独立同分布地抽取。此外，我们定义状态转移分布序列 $P = \{P_t\}_{t=1}^{T}$，其中 $P_t = \{P_t(.|o, a)\}_{o\in\mathsf{O}, a\in\mathsf{A}}$；对于给定时间点 t的每个状态-动作对（o, a），$P_t(.|o, a)$是 O 上的多项式概率分布。乍看之下，将状态转移分布限定为由多项分布生成似乎有所限制，但这一设定受到多个研究领域的广泛关注。相关应用包括：对认知无线电中的碰撞进行建模、监控统计模型的性能、监控用于网络监控的探测器中的事件、多臂老虎机问题、临床试验中的实验以及推荐系统等。

为清晰地定义我们所关注的问题，下面概述需要考虑的主要假设。

非平稳性。与文献中提出的许多方法不同，我们一般性地假设，状态转移概率分布和奖励可以在未知的时间步任意变化（称为变点

*不同于我们和Gajane 等人（2018 年的做法，后者将时间 t 的遗憾定义为当前马尔可夫决策过程在时间 t 的最优策略所获得的奖励与学习者策略所获得的奖励之差，Cheung 等人（2020 年）；Ortner 等人（2020 年）考虑了另一种遗憾概念，他们同样将其称为动态遗憾。后者定义为学习者策略与可实现的最佳稳态策略之间的差异。尽管Yu 和 Mannor（2009 年）；Even-Dar 等人（2009 年）；Neu 等人（2010 年）；Dick 等人（2014 年）广泛使用这一术语，但它通常并不十分有用，因为最佳稳态策略仍可能导致不理想的奖励，尤其是在发生非光滑变化的环境中

），也就是说，这些变化不受累积到某个预先设定的变化预算 等条件的限制。更根本地说，我们仅作出一个最小假设：在整个学习过程中，有限状态集和动作集合保持不变。此外，由于它属于指数族概率分布，因此具有良好的集中性质，可平滑地融入依赖上置信界的乐观探索算法。

外生性。 我们假设变点以及奖励分布和状态转移分布的变化，均不依赖于算法先前的行为或 MDP参数历史的过滤（o_1， a_1， r_1.....， o_t， a_t， r_t）。可以假设，变点集合 $\{c_\ell\}_{\ell=1}^{K_T}$ 是通过外生过程预先生成的，其中 K_T 表示截至时间范围 T内发生的变点数量。

收敛性与界。 对于收敛性，我们自然假设奖励有界，即 $|R_t(o, a)| \leqslant$最大值$_{t,\ o,\ a} R_t(o, a) < \infty$， $\forall o \in \mathsf{O}$ $\forall a \in \mathsf{A}$ $\forall t \in \{1, ..., T\}$。

内生性。 智能体从某个任意状态 o_{init}开始。在时间 t，它观测到状态 $o_t \in \mathsf{O}$，并根据某个策略 $\pi \in \Pi$采取动作$a_t \in \mathsf{A}$。因此，它根据状态转移分布 $P(.|o_t, a_t)$ 转移到下一状态 $o_{t+1} \in \mathsf{O}$，并接收到按照 $[0, 1]$上的某个未知分布抽取的随机奖励 $r_t \sim R_t(o_t, a_t)$。这里的内生性假设将可行策略集合 Π 限制为非预见性，即策略选择仅取决于当前状态和先前观测的集合（o_1， a_1， r_1.....， o_t， a_t， r_t）。
现在我们已经介绍了所考虑的一组假设，可以开始给出本文中将用到的定义。

切换MDP问题。继 Garivier & Moulines (2011) 首次将这一名称用于多臂老虎机，随后 Gajane 等人（2018 年）将其用于马尔可夫决策过程之后，我们采用“切换MDP”这一名称来描述我们的设定。将变点时刻集合定义为 $\{c_\ell\}_{\ell=0}^{K_T}$，我们认为分段平稳MDP M在 $t \in [c_\ell, c_{\ell+1})$ 内处于配置 $M_\ell(\mathsf{O}, \mathsf{A}\ P_\ell, R_\ell, T$。因此，切换马尔可夫决策过程问题 M可完全表示为元组 $\mathbf{M} = \{\mathbb{S} = \{M_0, .., M_{K_T-1}\}, \mathcal{C} = \{c_0, .., c_{K_T}\}\}$，其中 c_0和 c_{K_T}分别是学习开始（$t = 1$）和结束（$t = T$）时刻。

马尔可夫决策过程的直径。 马尔可夫决策过程 M_ℓ 的直径定义如下：

$$D(M_\ell) = \max_{o_1, o_2 \in \mathsf{O}, o_1 \neq o_2} \min_{\pi \in \Pi} \mathbb{E}[\tau(o_1, o_2, M_\ell, \pi)]$$

其中，随机变量 $\tau(o_1, o_2, M_\ell, \pi)$ 表示从可行平稳策略集合 Π 中选取的策略 π，平均需要多少步骤才能从 o_1 转移到 o_2。特别地，我们将直径有限的马尔可夫决策过程称为连通马尔可夫决策过程。

由于本文与许多其他学习设置一样，使用遗憾 作为性能度量，因此我们给出平均奖励 的定义。对于马尔可夫决策过程 M_ℓ 的一个组成部分，在执行算法$\mathcal{A}$ 并遵循平稳策略时，其平均奖励可写为： $\rho(M_\ell, \mathcal{A}, o_{\text{init}}) := \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} r_t\right]$，其中奖励序列根据策略从$o_{\text{init}}$开始所生成的状态和动作确定。 –

假设构成$\mathbf{M}$ 的每个马尔可夫决策过程都是可沟通的，根据Puterman （2014 年），最优平均奖励不依赖于该马尔可夫决策过程的初始状态。这一结果是遗憾定义的基础，因为它使得$\mathbf{M}$ 的最优平均奖励可以分解为其各组成马尔可夫决策过程的最优平均奖励之和$\{M_\ell\}_{\ell=0}^{K_T-1}$。现在，将马尔可夫决策过程 M_ℓ的最优奖励定义如下： $\rho^\star_{M_\ell} := \max_{\pi, o \in \mathsf{O}} \rho(M_\ell, \mathcal{A}, o)$
其中，同样地， π是
由 $\mathcal{A}$生成的非预见策略。现在，我们可以定义切换马尔可夫决策过程问题的遗憾如下。

切换马尔可夫决策过程的遗憾。 算法 $\mathcal{A}$ 在切换马尔可夫决策过程问题 $\mathbf{M} = \left(\{M_\ell\}_{\ell=0}^{K_T}, \{c_\ell\}_{\ell=1}^{K_T}\right)$ 中，从某个初始状态 s 开始，在时间范围 T内的遗憾记为：

$$\mathfrak{R}(\mathbf{M}, \mathcal{A}, o, T) = \sum_{t=1}^{T} (\rho^\star_{\mathbf{M}}(t) - \mathbb{E}[r_t]) \quad \text{where} \quad \rho^\star_{\mathbf{M}}(t) := \rho^\star_{M_\ell} \text{ if } t \in [c_\ell, c_{\ell+1}).$$

4 将变点检测作为应对非平稳性的手段

本节首先设计Alami 等人（2020 年）提出的重启式贝叶斯在线变点检测器的多项式版本。随后，我们给出该算法在虚警率和检测延迟方面的数学保证。最后，我们设计R-BOCPD-UCRL2 策略，这是一个配备了 UCRL2 实例，并结合R-BOCPD 来处理分段平稳的马尔可夫决策过程。

4.1 重启 TED 多项分布的贝叶斯在线变点检测

本节研究在线变点检测问题：收集一组具有共同波动上界的独立多元随机变量序列，其均值可能在一个或多个时间点发生变化。具体而言，我们考虑一个智能体，其目标是检测在线数据流生成过程中的变化。在每个时间步 t，智能体观察数据 $x_t \sim$ 多项分布（$\mu_{1,\,t}$，……，$\mu_{O,\,t}$）：一个服从参数为（$\mu_{1,\,t}$，……，$\mu_{O,\,t}$）的多项分布的随机变量 $\in [0,1]^O (x_t \in \{1,...,O\})$，并需要判断数据流的生成过程中是否发生了变化。或者，智能体可以在每个时间步 t计算最近变点的估计值 $\widehat{c}_t$。

记号。 下文中，我们用 $n_{s:t} := t - s + 1$ 表示从时间 s 到时间t的观测数量。此外，在序列 $\mathbf{x}_{s:t} = (x_s, \ldots, x_t)$ 中观测到 o的经验频率记为 $\widehat{\mu}_{o,\,s:t} := \frac{1}{n_{s:t}} \sum_{s'=s}^{t} \mathbb{I}\{x_{s'} = o\}$。

定义 4.1（多项分布的 Kullback-Leibler 散度）。设 $\boldsymbol{\theta}^{(1)} = \left(\theta_1^{(1)}, ..., \theta_O^{(1)}\right)$ 和$\boldsymbol{\theta}^{(2)} = \left(\theta_1^{(2)}, ..., \theta_O^{(2)}\right)$ 是两个多项分布的参数，则 Multi $\left(\theta_1^{(2)}, ..., \theta_O^{(2)}\right)$ 和 Multi $\left(\theta_1^{(1)}, ..., \theta_O^{(1)}\right)$ 之间的相对熵定义如下：$\mathbf{KL}\left(\boldsymbol{\theta}^{(2)} \,\|\, \boldsymbol{\theta}^{(1)}\right) = \sum_{o=1}^{O} \theta_o^{(2)} \log \frac{\theta_o^{(2)}}{\theta_o^{(1)}}$。

定义 4.2（分段平稳多项过程）。令 T 表示数据流长度，K_T 表示截至时间 T 观测到的变点总数。假设观测 $x_t \sim$ 多项分布 $(\mu_{1,\,t}, \ldots\ldots, \mu_{O,\,t})$ 由一个分段多项过程生成，且存在一个非递减的变点序列 $(c_\ell)_{\ell \in [1, K_T]} \in \mathbb{N}^{K_T}$，满足：

$$\forall \ell \in \{0, ..., K_T\}, \forall t \in [c_\ell, c_{\ell+1}), \forall o \in \mathsf{O} \quad \mu_{o,t} = \theta_{o,\ell}, \quad \text{where: } c_0 = 1 < c_2 < ... < c_{K_T} = T. \tag{1}$$

定理 4.3（检测延迟下界）。令：$(x_r, ..., x_{c_\ell - 1}) \sim$ 多项分布 $\left(\boldsymbol{\theta}^{(l-1)} = \left(\theta_1^{(\ell-1)}, ..., \theta_O^{(\ell-1)}\right)\right)$，且 $(x_{c_\ell}, ..., x_t) \sim$ 多项分布 $\left(\boldsymbol{\theta}^{(\ell)} = \left(\theta_1^{(\ell)}, ..., \theta_O^{(\ell)}\right)\right)$，$\mathcal{A}$ 为一种在线变点检测策略，c_ℓ 为待检测的变点，r 为重启时间。假设虚警率受到如下控制：$\mathbb{P}_{\boldsymbol{\theta}^{(\ell-1)}}\left\{\exists s \in [r, \tau_c) : \mathcal{A}(x_{r:s}) = 1\right\} \leqslant \delta$，则当量 $\frac{n_{r:c_\ell}}{\log \frac{1}{\delta}} \underset{\delta \to 0}{\to} \infty$ 时，期望检测延迟 $\mathbb{E}_{\boldsymbol{\theta}^{(\ell-1)}, \boldsymbol{\theta}^{(\ell)}}[\widehat{c}_{\mathcal{A}}(x_{r:t}) - c_\ell]$ 的下界如下：$\mathbb{E}_{\boldsymbol{\theta}^{(\ell-1)}, \boldsymbol{\theta}^{(\ell)}}[\widehat{c}_{\mathcal{A}}(x_{r:t}) - c_\ell] \geqslant \left(\frac{1}{\mathbb{P}_{\boldsymbol{\theta}^{(\ell-1)}}\left\{\widehat{c}_{\mathcal{A}}(x_{r:t}) > c_\ell\right\}} KL(\boldsymbol{\theta}^{(\ell)} \,\|\, \boldsymbol{\theta}^{(\ell-1)})\right) \log \frac{1}{\delta}$。

将重启式贝叶斯在线变点检测器（R-BOCPD）扩展至多项分布。Alami 等人 (2020 年) 提出了重启式贝叶斯在线变点检测（R-BOCPD），这是贝叶斯在线变点检测器的一个剪枝版本，适用于分布均值发生变化的单变量伯努利分布样本。

本节中，我们提出将R-BOCPD 算法Alami 等人 (2020 年) 中介绍的算法扩展至多项分布。首先，我们从扩展适用于多项分布的拉普拉斯预测器入手。

定义 4.4（扩展拉普拉斯预测器）。预测器 PRED $(x_{t+1}|\mathbf{x}_{s:t})$ 接收一个序列$\mathbf{x}_{s:t} \in \{1, ..., O\}^{n_{s:t}}$ 作为输入，并按如下方式预测下一个观测值 $x_{t+1} \in \{1, ..., O\}$

$$\text{PRED}(x_{t+1}|\mathbf{x}_{s:t}) := \frac{\sum_{i=s}^{t} \mathbb{I}\{x_i = x_{t+1}\} + 1}{n_{s:t} + O} \tag{2}$$

其中 $\forall x \in \{1, ..., O\}$ PRED $(x|\emptyset) = \frac{1}{O}$对应于赋予生成 θ_c 的过程的均匀先验。

回顾 Alami 等人（2020 年）的研究，他们没有处理运行长度，而是采用预测器这一概念，即连续拉普拉斯预测器的乘积。因此，与 Alami 等人 （2020 年）一样，我们引入预测器的损失。

定义 4.5 (预测器 器损失)。 使用该预测器，预测器 s 在时间 t 的瞬时损失为 :

$$l_{s:t} := -\log \textsc{Pred}\left(x_t|\mathbf{x}_{s:t-1}\right) = -\sum_{o=1}^{O} \mathbb{I}\{x_t = o\} \log \textsc{Pred}\left(o|\mathbf{x}_{s:t-1}\right).$$

然后，令 $\widehat{L}_{s:t} := \sum_{s'=s}^{t} l_{s':t}$ 表示预测器 s 从时间 s 到时间 t 所产生的累积损失，其粗略表达式如下：

$$\widehat{L}_{s:t} := \sum_{s'=s}^{t} -\log \textsc{Pred}\left(x_t|\mathbf{x}_{s':t-1}\right) \tag{3}$$

因此，预测器权重的更新将保持不变（对于某个时间函数 $\eta_{r,\ s,\ t} \in (0,\ 1)$）。

$$\omega_{r,s:t} = \begin{cases} \frac{\eta_{r,s,t}}{\eta_{r,s,t-1}} \exp\left(-l_{s,t}\right) \omega_{r,s:t-1} & \forall s < t, \\ \eta_{r,t,t} \times \mathcal{W}_{r:t-1} & s = t\,. \end{cases} \text{with} \quad \mathcal{W}_{r:s-1} := \exp\left(-\widehat{L}_{r:s-1}\right) \text{ for some starting time } r. \tag{4}$$

最后，我们采用与Alami 等人(2020 年)相同的重启过程：

$$\textbf{\textit{Restart}}\left(x_r, ..., x_t\right) = \mathbb{I}\{\exists s \in (r, t] : \omega_{r,s:t} > \omega_{r,r:t}\} \tag{5}$$

我们在算法 R-BOCPD中描述了用于多项分布的1。

算法 1 用于多项分布的 R-BOCPD

输入： η_r、$_s$、$_t \in$ **(0, 1) 1：** $r \leftarrow 1$、ω_r、$_{1:1} \leftarrow 1$、η_r、**1**、$_1 \leftarrow 1$。**2：对于** $t = 1$，**.....，执行 3：观测** $x_t \sim$ **多项分布**（μ_1、$_t$、**.....**、μ_O、$_t$）**4：为每个预测器定义从时间** r **到时间** t **的** s**：** ω_r、$_{s:t} \leftarrow \begin{cases} \eta_r \\ \eta_{r\ s} \end{cases}$、$_{s\ t}$、、、$_{t-1}$ **exp**（$-l_{s:t}$）ω_r、$_{s:t-1}$ $\forall s < t$、$\eta_{r\ t}$、$_t \times \mathcal{W}_{r:t-1}$ $s = t$. **5：如果重启**（x_r、**.....**、x_t）$= 1$ **则** $r \leftarrow t + 1$、ω_r、$_{r:r} \leftarrow 1$、η_r、$_r$、$_r \leftarrow 1$。**6：估计最近变点：** $\widehat{c}_t \leftarrow r$。**7：结束循环**

4.2 多项分布情形下R-BOCPD 的性能保证

本节给出关于参数η_r、$_s$、$_t$ 的充分条件，以保证多项分布情形下R-BOCPD 算法的误报率控制（定理 4.6）和有限检测延迟（定理 4.8）。

4.2.1 受控误报率

定理 4.6 (虚警率)。 令： $\boldsymbol{\theta} = (\theta_1, ..., \theta_O)$ 表示多项分布 *Multi* $(\theta_1, ..., \theta_O)$的参数向量。对于 $r < t$，假设$(x_r, ..., x_t) \sim$服从 *Multi* $(\theta_1, ..., \theta_O)^{\otimes n_{r:t}}$。令 $\alpha > 1$，若 $\eta_{r,\ s,\ t}$ 足够小，使得：

$$\eta_{r,s,t} < \left(\prod_{i=1}^{O-1} \frac{\left(n_{r:s-1} + i\right)\left(n_{s:t} + i\right)}{n_{r:t} + i}\right) \times \frac{\exp\left(2b_1\right)}{\left(n_{r:s-1} n_{s:t}\right)^{\frac{O-1}{2}} \times (O-1)!} \times \left(\frac{\log(4\alpha + 2)\delta^2}{4n_{r:t} \log((\alpha + 3)\, n_{r:t})}\right)^{\alpha} \tag{6}$$

$$\textit{where: } b_1 = -\frac{O}{12} - \frac{O-1}{2} \log\left(2\pi\right) + \frac{O}{2} \log O.$$

那么，以高于 $1-\delta$ 的概率，在区间 $[r, c_\ell)$ 上不会发生虚警：

$$\mathbb{P}_{\boldsymbol{\theta}}\Big\{\exists\, t \in [r, c_\ell) : \textbf{\textit{Restart}}\,(x_r, ..., x_t) = 1\Big\} \leqslant \delta.$$

定义 4.7（相对间隔 $\Delta_{o,\ r,\ s,\ t}$）。令 $\Delta_o \in [0,1]$。预测器 s 在时间 t 的相对间隔 $\Delta_{o,\ r,\ s,\ t}$ 取以下形式（取决于 s的位置）：

$$\Delta_{o,r,s,t} = \left(\frac{n_{r:c_\ell-1}}{n_{r:s-1}}\mathbb{I}\{c_\ell \leqslant s \leqslant t\} + \frac{n_{c_\ell:t}}{n_{s:t}}\mathbb{I}\{s < c_\ell\}\right)\Delta_o.$$

4.2.2 最优检测延迟

定理 *4.8*（有限检测延迟）。令 $(x_r, ..., x_{c_\ell-1}) \sim$ 多项分布 $\left(\theta_1^{(\ell-1)}, ..., \theta_O^{(\ell-1)}\right)^{\otimes n_{r:c_\ell-1}}$，$(x_{c_\ell}, ..., x_t) \sim$ 多项分布 $\left(\theta_1^{(\ell)}, ..., \theta_O^{(\ell)}\right)^{\otimes n_{c_\ell:t}}$，$\Delta_o = \left|\theta_o^{(\ell-1)} - \theta_o^{(\ell)}\right|$：与观测 $o \in \{1, ..., O\}$ 和 $\boldsymbol{\Delta} =$ 相关的变点间隔；$(\Delta_1, ..., \Delta_O)$ 为变点间隔向量。接着，令：$f_{r,s,t} = \sum_{i=1}^{O-1} log\,(n_{r:s-1}+i) + \sum_{i=1}^{O-1} log\,\binom{n_{s:t}+i}{n_{r:t}+i} - \binom{O-1}{2} log\,\binom{n_{s:t}}{n_{r:t}} - log(O-1)!$。若 $\eta_{r,s,t}$ 足够大，使得：

$$\eta_{r,s,t} > \exp\Big(-2n_{r,s-1}\sum_{o=1}^{O}\left(\Delta_{o,r,s,t} - \mathcal{C}_{r,s,t,\delta}\right)^2 + f_{r,s,t}\Big), \tag{7}$$

那么，以不超过 $\mathfrak{D}_{\boldsymbol{\Delta},\ r,\ c_\ell}$ 的延迟检测到变点 c_ℓ（概率至少为 $1-\delta$）：

$$\mathfrak{D}_{\boldsymbol{\Delta},r,c_\ell} = \min\left\{d \in \mathbb{N}^\star : d > \frac{2}{\sum_{o=1}^{O}\left(\Delta_o - \mathcal{C}_{r,c_\ell,d+c_\ell-1,\delta}\right)^2} \times \frac{-\log\eta_{r,c_\ell,d+c_\ell-1} + f_{r,c_\ell,d+c_\ell-1}}{1 + \frac{2\left(\log\eta_{r,c_\ell,d+c_\ell-1} - f_{r,c_\ell,d+c_\ell-1}\right)}{n_{r,c_\ell-1}\times\sum_{o=1}^{O}\left(\Delta_o - \mathcal{C}_{r,c_\ell,d+c_\ell-1,\delta}\right)^2}}\right\} \tag{8}$$

$$\textit{where: } \mathcal{C}_{r,s,t,\delta} = \frac{\sqrt{2}}{2}\left(\sqrt{\frac{1+\frac{1}{n_{r:s-1}}}{n_{r:s-1}}\log\left(\frac{2\sqrt{n_{r:s}}}{\delta}\right)} + \sqrt{\frac{1+\frac{1}{n_{s:t}}}{n_{s:t}}\log\left(\frac{2n_{r:t}\sqrt{n_{s:t}+1}\log^2(n_{r:t})}{\log(2)\delta}\right)}\right). \tag{9}$$

讨论 4.9.（检测延迟的渐近行为）渐近状态对应于最后一次重启 r 与新变点 c_ℓ 之间的经过时间趋于无穷，同时虚警概率 δ 趋于零的情形。因此，我们得到：

$$\mathfrak{D}_{\boldsymbol{\Delta},r,c_\ell} \underset{n_{r,c_\ell-1}\to\infty}{\to} \frac{2\left(-\log\eta_{r,\tau_c,d+\tau_c-1} + o\big(\log\frac{1}{\delta}\big)\right)}{\sum_{o=1}^{O}\Delta_o^2} \overset{(a)}{=} \mathcal{O}\left(\frac{\log\frac{1}{\delta}}{\boldsymbol{KL}\big(\boldsymbol{\theta}^{(\ell)} \,\big\|\, \boldsymbol{\theta}^{(\ell-1)}\big)}\right) \tag{10}$$

其中，(a) 源自 Pinsker不等式，该不等式将 Kullback-Leibler散度与全变差联系起来。因此，根据定理 4.3 的陈述，R-BOCPD 的检测延迟 $\mathfrak{D}_{\boldsymbol{\Delta},\ r,\ c_\ell}$ 渐近阶最优。

讨论 4.10.（超参数 $\eta_{r,\ s,\ t}$ 的选择）为了保证虚警率和检测延迟，我们需要选择合适形式的参数 $\eta_{r,\ s,\ t}$，使其同时满足方程 *(7)* 和方程 *(6)* 中的两个条件。选择 $\eta_{r,\ s,\ t} \approx \frac{1}{n_{r:t}}$ 即可满足这两个条件。

4.3 R-BOCPD-UCRL2 策略

现在，我们提出在切换MDP环境中为 UCRL2 配备 M 和 R-BOCPD。这一方法源于能够将 M $= \{\mathbb{S} = \{M_0, .., M_{K_T-1}\}, \mathcal{C} = \{c_0, .., c_{K_T}\}\}$ 分解为独立生成的平稳时段 $(c_\ell,\ \ c_{\ell+1})$；在每个时段中，R-BOCPD 首先检测从马尔可夫决策过程 M_ℓ 切换到 $M_{\ell+1}$ 的变更时刻 c_ℓ，然后相应地重启 UCRL2。根据所给出的上置信界，以高概率保证延迟最小。我们还以类似方式确保在平稳时段内以高概率不会发生其他重启。具体方法详见算法 2。

定义 4.11. (UCRL2 框架）。我们采用与原始 UCRL2Auer 等人 (2008b) 相同的记号和学习框架。受篇幅限制，此处不再赘述，详见附录 A。

定理 4.12 (有限时间最优遗憾上界)。 以至少 $1-\delta$ 的概率，对于一个切换型马尔可夫决策过程 $\mathbf{M} = \{\mathbb{S} = \{M_0, .., M_{K_T-1}\}, \mathcal{C} = \{c_0, .., c_{K_T}\}\}$ (从某个初始状态 o_{c_0} 开始)，若其（分段）平稳时段长度至少为 1，则按第 *R-BOCPD-UCRL2* 节中定义的遗憾 *3* 满足以下上界：

$$\mathfrak{R}\left(\mathbf{M}, \textit{R-BOCPD-UCRL2}, o_{c_0}, T\right) \leqslant 34DO\sqrt{ATK_T \log\left(\frac{T}{\delta}\right)} + \sum_{\ell=0}^{K_T-1} \mathfrak{D}_{\mathbf{\Delta}_{\ell+1}, c_\ell + d_\ell, c_{\ell+1}}$$

其中，$\mathfrak{D}_{\mathbf{\Delta}_{\ell+1},\ c_\ell+d_\ell,\ c_{\ell+1}}$ 是 *R-BOCPD* 在输入流 $(o_{c_\ell+d_\ell}, ..., o_{c_{\ell+1}})$ 上针对间隔 $\mathbf{\Delta}_{\ell+1} = (\Delta_{1,\ \ell+1}, ..., \Delta_{O,\ \ell+1})$ 的检测延迟，其中 $\Delta_{o,\ \ell+1} = \left|\theta_o^{(\ell)} - \theta_o^{(\ell+1)}\right|$，$\boldsymbol{\theta}^{(\ell)} = \left(\theta_1^{(\ell)}, ..., \theta_O^{(\ell)}\right)$ 和 $\boldsymbol{\theta}^{(\ell+1)} = \left(\theta_1^{(\ell+1)}, ..., \theta_O^{(\ell+1)}\right)$ 是状态集合 O 上变点 $c_{\ell+1}$前后的状态转移核。

现在，我们给出一个推论，以刻画 R-BOCPD-UCRL2 遗憾的渐近行为。

推论 4.13 (渐近遗憾上界)。 以至少 $1-\delta$ 的概率，假设 $c_{\ell+1} - c_\ell - d_\ell \gg 1$ 的虚警概率为 $\delta_{\textit{False-Alarm}} \to 0$ （如方程 （*10*）所示），我们可以将 *R-BOCPD-UCRL2* 在 $\mathbf{M}$ 上从某个状态 o_{c_0} 开始的遗憾渐近上界写为：

$$\mathfrak{R}\left(\mathbf{M}, \textit{R-BOCPD-UCRL2}, o_{c_0}, T\right) = \mathcal{O}\left(DO\sqrt{ATK_T \log\left(\frac{T}{\delta}\right)} + \frac{K_T \log\frac{K_T}{\delta}}{\min_\ell \boldsymbol{KL}\left(\boldsymbol{\theta}^{(\ell+1)} \,\middle\|\, \boldsymbol{\theta}^{(\ell)}\right)}\right)$$

讨论 *4.14*。（上界的最优性）我们推导出了 *R-BOCPD-UCRL2* 遗憾上界的有限时间变体和渐近变体，两者的表现均优于当前最先进水平。鉴于设计 *R-BOCPD-UCRL2* 的目的在于使其能够适应快速且突然变化的非平稳强化学习环境，其遗憾界与变点前后分布之间的距离呈最优相关性。我们还强调，据我们所知，我们提出的方法是首个能够达到 $\tilde{\mathcal{O}}(T^{\frac{1}{2}})$ 遗憾的方法。此前提出的滑动窗口方法 *Gajane* 等人（*2018*）；*Cheung* 等人（*2020*）分别得到 $\tilde{\mathcal{O}}(T^{\frac{2}{3}})$ 和 $\tilde{\mathcal{O}}(T^{\frac{3}{4}})$ 的遗憾界。带重启的 *UCRL2 Auer* 等人（*2008a*）即使重启次数比 *R-BOCPD-UCRL2* 多 $T^{\frac{1}{3}} K_T^{-1}$ 次，所达到的遗憾仍仅为 $\tilde{\mathcal{O}}(T^{\frac{2}{3}})$。

5 实验

据我们所知，我们将 R-BOCPD-UCRL2 与在我们设定下表现最佳的 4 种算法进行基准比较。具体如下：

- **滑动窗口 UCRL2**(SWUCRL2，Gajane 等人 (2018 年))：采用滑动窗口方法，仅维护过滤历史中最近的 W 个时间步，其中 W 被称为窗口大小。
- **带置信度扩展的滑动窗口 UCRL2** (SWUCRL2-CW，Cheung 等人 (2020 年)) 比 *SWUCRL2*更加乐观；除了大小为 W的窗口外，还定义了一个置信度扩展参数 η，用于量化在通过上置信界实现的传统乐观探索基础上所需进行的额外乐观探索量。
- **重启 UCRL2** (Restarted-UCRL2，Auer 等人(2008a)) 定义了原始 UCRL2 的一种变体，其中后者会在步骤 $\tau_i = \lceil i^3 / K_T^2 \rceil$ 处重启，并且假设变化次数 K_T在初始化时已知。
- **原始 UCRL2**(UCRL2，Auer 等人 (2008a))。

此外，我们还考虑 UCRL2 的一个变体，按照第 3 节中的遗憾定义，该变体是遗憾最优的（因为 UCRL2在每个平稳时段都近乎最优 $[c_\ell, c_{\ell+1})$ 。其定义如下：

- **配备预言机的 UCRL2** （UCRL2 预言机）能够获知所有变化 $\{c_\ell\}_{\ell=1}^{K_T-1}$，并在初始化时就已知晓，因此会在每个 c_ℓ 为 $\ell \in \{0, ..., K_T - 1\}$时恰好重启 UCRL2。

我们在附录F中更详细地列出了具体的实验设置，以及各算法所选用的超参数。

5.1 实验结果

我们在多种合成生成的马尔可夫决策过程上评估上述算法的性能，这些过程具有不同基数的状态-动作集合。变点在时间范围 T 内（随机）生成，从而能够考察改变变点之间间隔时长对学习的影响。我们绘制各方法的累积奖励随时间变化的曲线如下

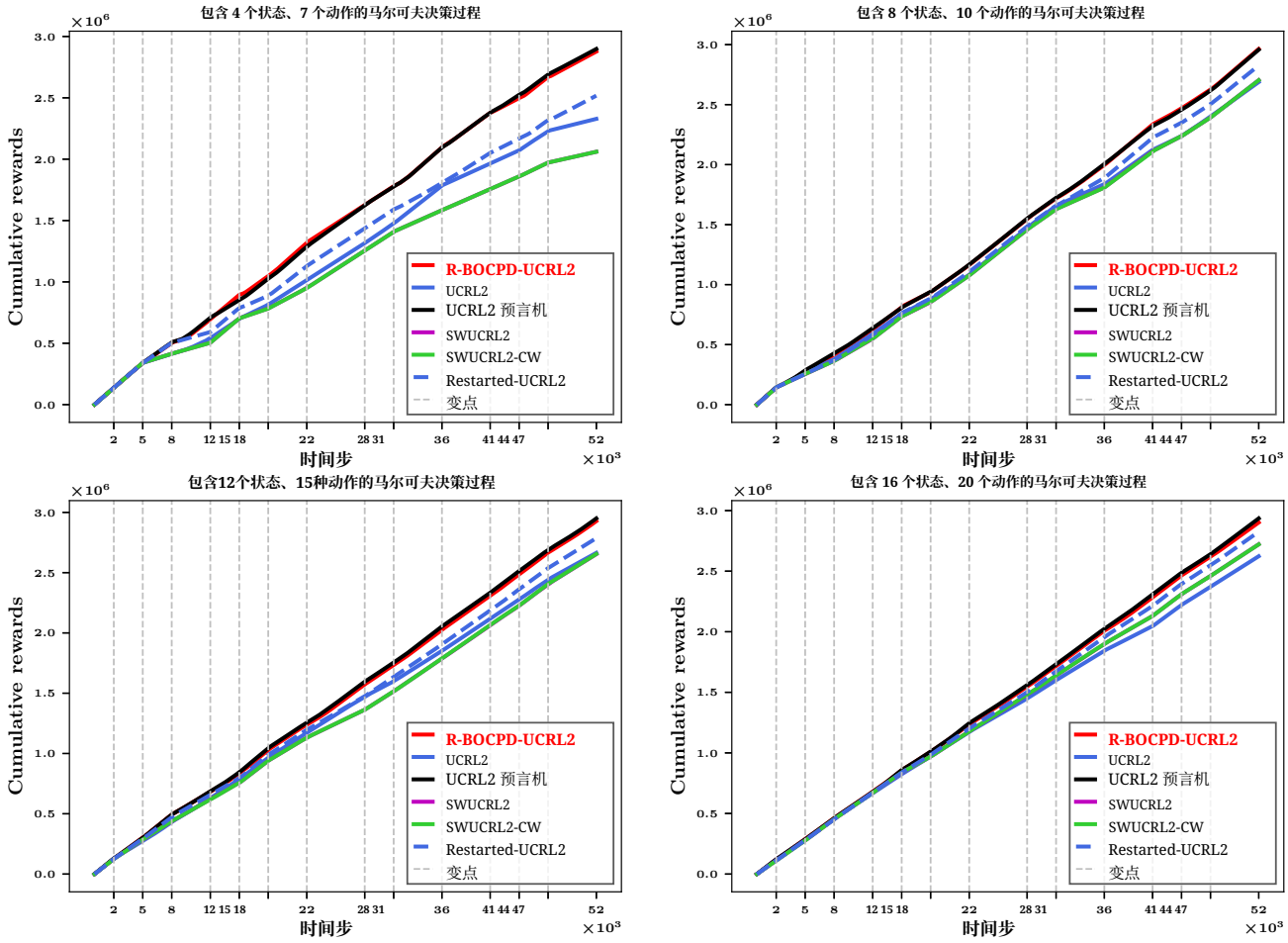

图 2：针对一系列随机变点，在不同状态-动作对上将 R-BOCPD-UCRL2 与当前最先进水平进行基准比较。变点之间的突变程度（即状态转移分布和奖励的变化幅度）也各不相同，从而能够模拟对马尔可夫决策过程产生全局和局部影响的变化。

5.2 性能评估

图 2 表明 R-BOCPD-UCRL2 在实践中（几乎）达到了遗憾最优，因为其在所定义的一般设置中的表现与能够感知变点的（最优）UCRL2 预言机的表现非常接近。我们还强调，观察到的性能具有良好的泛化性，适用于具有不同状态空间和动作空间规模、且针对状态转移分布和奖励具有不同总变化预算的各种切换马尔可夫决策过程问题。受篇幅限制，各算法性能及其关键假设的深入讨论见附录F.2。

6 讨论与总体说明

在本研究中，我们提出了重启贝叶斯在线变点检测算法（R-BOCPD-UCRL2）。这是一种配备变点检测器的基于模型的强化学习算法，可在非平稳环境中运行，而这类环境可通过离散时间分段平稳 MDP 完整刻画。我们将贝叶斯在线变点检测算法（BOCPD）的理论保

证扩展到了更一般的多项分布，并证明 R-BOCPD-UCRL2 是遗憾最优的，其渐近遗憾界为$\mathcal{O}\left(DO\sqrt{ATK_T \log\left(\frac{T}{\delta}\right) + \frac{K_T}{\log\frac{K_T}{\delta}} \min_\ell \mathbf{KL}(\boldsymbol{\theta}^{(\ell+1)} \| \boldsymbol{\theta}^{(\ell)})}\right)$。据我们所知，该算法首次在时间范围 T 内达到 $\tilde{O}(T^{\frac{1}{2}})$界。我们进一步证明了 R-BOCPD-UCRL2 在实践中的优势：与当前最先进水平相比，它所需的重启次数少得多，且无需隐式定义的输入参数（如 MDP 直径、变化预算等）。

局限性与未来工作：这里，我们重点介绍一些未纳入本文研究范围、但在作者看来仍极具潜力的研究方向。首先，我们讨论状态转移分布源自多项分布这一假设。需要指出的是，贝叶斯在线变点检测器并不一定要求输入流源自多项分布，也可以扩展到概率分布指数族中的任意分布。对于后者，关于最小检测延迟和误报率的理论保证同样适用。其次，我们指出，我们的算法倾向于检测分布发生剧烈变化的变点，例如按全变差范数衡量时变化足够大的情形。一个值得探索的方向是定义一个阈值，使变点检测器能够根据 MDP 参数变化的全局或局部性质，决定何时重启平稳强化学习算法（此处为 UCRL2）。最后，一个自然的扩展方向是提出配备变点检测器的无模型非平稳强化学习算法。

算法2 R-BOCPD-UCRL2 输入：置信参数 $\delta \in (0,1)$，$\eta_{r,s,t} \in (0,1)$ **1：设置** $\forall (o,a)$ $N_0(o,a) \leftarrow 0, V_0(o,a) \leftarrow 0, t \leftarrow 1, k \leftarrow 1$ **并观测初始状态** s_1。 **2：初始化重启时间** $r \leftarrow 1$ **3：对于每一对** (o,a)，**执行 4：初始化一个 R-BOCPD**$_{o,a}$ **过程 5：结束循环 6：对于每个回合** $k \geqslant 1$，**执行 7：初始化回合** k：**8：设置回合** k, $t_k = t$ **的开始时间 9：对于所有** $(o,a) \in \mathbf{O} \times \mathbf{A}$，**初始化回合** $k, V_k(o,a) := 0$ **的状态-动作计数。此外，对于所有状态** $o \in \mathcal{O}$ **和动作** $a \in \mathcal{A}$，**设置在回合** k **之前动作** a **在状态** o **中执行的次数；10：对于所有** $o, o' \in \mathbf{O}$ **和** $a \in \mathbf{A}$，**设置在回合** k **之前，动作** a **在状态** o **中执行时观测到的累积奖励，以及导致下一状态为** o' **的次数；11：计算估计值** $\widehat{R}_k(o,a) := {}^{R_k}{}_{(o,a)}$ **max**$_{\{1, N_k(o,a)\}}$, $\widehat{P}_k(o' \mid o,a) := {}^{P_k}{}_{(o,a,o')}$ **max**$_{\{1, N_k(o,a)\}}$ **12：计算策略** $\tilde{\pi}_k$：

$$N_k(o,a) := \#\left\{r \leqslant t < t_k : o_t = o, a_t = a\right\}.$$

$$R_k(o,a) := \sum_{t=r}^{t_k - 1} r_t \mathbb{I}\{o_t = o, a_t = a\}, \text{and } P_k(o,a,o') := \#\left\{r \leqslant t < t_k : o_t = o, a_t = a, o_{t+1} = o'\right\}.$$

13：令 $\mathcal{M}_k$ 为所有马尔可夫决策过程的集合，其状态空间为 O、动作空间为 A，且转移概率 $\tilde{P}(\cdot \mid o, a)$ 接近于 $\widehat{P}_k(\cdot \mid o, a)$，奖励 $\tilde{R}(o, a) \in [0,1]$ 接近于 $\widehat{R}_k(o, a)$，满足：

$$\left|\tilde{R}(o,a) - \widehat{R}_k(o,a)\right| \leqslant \sqrt{\frac{7\log(2OAt_k/\delta)}{2\max\{1, N_k(o,a)\}}} \text{ and } \left\|\tilde{P}(\cdot \mid o,a) - \widehat{P}_k(\cdot \mid o,a)\right\|_1 \leqslant \sqrt{\frac{14O\log(2At_k/\delta)}{\max\{1, N_k(o,a)\}}}.$$

14: 使用扩展价值迭代，求得策略 $\tilde{\pi}_k$ 和一个乐观马尔可夫决策过程 $M_k \in \mathcal{M}_k$，使得

$$\tilde{\rho}_k := \min_{o \in \mathcal{O}} \rho\left(\bar{M}_k, \tilde{\pi}_k, o\right) \geqslant \max_{M' \in \mathcal{M}_k, \pi, o'} \rho\left(M', \pi, o'\right) - \frac{1}{\sqrt{t_k}}$$

15: 执行策略 $\tilde{\pi}_k$**: 16: 当** $V_k(o_t, \tilde{\pi}_k(o_t)) <$ **max**$\{1, N_k(o_t, \tilde{\pi}_k(o_t))\}$ **时 17: 选择动作** $a_t = \tilde{\pi}_k(o_t)$，**获得奖励** r_t **并观察下一状态** o_{t+1}。 **18: 更新** $V_k(o_t, a_t) := V_k(o_t, a_t) + 1$ **19: 设置** $t := t+1$ **20: 对序列** $(o_r, ..., o_t)$ **执行变点检测检验。 21: 如果 R-BOCPD**$_{o_t,a_t}$ **重启** $(o_r, ..., o_t) = 1$ **则 22:** $\forall (o,a)$ $N_k(o,a) \leftarrow 0, V_k(o,a) \leftarrow 0, r \leftarrow t+1$。 **23: 结束如果 24: 结束当 25: 结束循环**

参考文献

Shipra Agrawal 和 Randy Jia。《具有凸成本函数的结构化马尔可夫决策过程中的学习：改进库存管理的遗憾界》。载于*2019*年*ACM*经济学与计算会议论文集，第 743–744 页，2019 年。

DiegoAgudelo-Espa˜na、Sebastian Gomez-Gonzalez、Stefan Bauer、Bernhard Sch¨olkopf 和 Jan Peters。贝叶斯在线预测变点。载于 人工智能不确定性会议，第 320–329 页。PMLR，2020 年。Reda Alami、Odalric Maillard 和 Raphael Feraud。重启式贝叶斯在线变点检测器实现了最优检测延迟。载于 Hal Daum´e III 和 Aarti Singh （编），第 *37* 届国际机器学习会议论文集，第 119 卷，收录于机器学习研究论文集，第 211–221 页。PMLR，2020 年 7 月 13–18 日。网址：https://proceedings.mlr.press/v119/alami20a.html。Raman Arora、Ofer Dekel 和 Ambuj Tewari。具有对抗性奖励和赌博机反馈的确定性马尔可夫决策过程。*arXiv* 预印本 *arXiv:1210.4843*，2012 年。Peter Auer、Thomas Jaksch 和 Ronald Ortner。强化学习的近最优遗憾界。神经信息处理系统进展，21，2008a。

Peter Auer、Thomas Jaksch 和 Ronald Ortner。强化学习的近最优遗憾界。神经信息处理系统进展，21，2008b。

Mohammad Gheshlaghi Azar、Ian Osband 和 R´emi Munos。强化学习的极小极大遗憾界。载于国际机器学习会议，第 263–272 页。PMLR，2017年。

Kostas Bimpikis、Ozan Candogan 和 Daniela Saban。网约车网络中的空间定价。运筹学，67(3)：744–769，2019 年。

Wang Chi Cheung、David Simchi-Levi 和 Ruihao Zhu。非平稳马尔可夫决策过程的强化学习：更多乐观主义的益处。载于国际机器学习会议，第 1843–1854 页。PMLR，2020 年。

Christoph Dann、Tor Lattimore 和 Emma Brunskill。统一 PAC 与遗憾：分幕式强化学习的统一 PAC 界。神经信息处理系统进展，30，2017年。

Travis Dick、Andras Gyorgy 和 Csaba Szepesvari。具有变化的成本序列的马尔可夫决策过程在线学习。载于国际机器学习会议，第 512–520 页。PMLR，2014 年。

Eyal Even-Dar、Sham MKakade 和 Yishay Mansour。在线马尔可夫决策过程。运筹学数学，34(3)：726–736，2009 年。

Paul Fearnhead 和 Zhen Liu。《多变点问题的在线推断》。英国皇家统计学会会刊：*B*辑（统计方法学），69(4):589–605，2007 年。Pratik Gajane、Ronald Ortner 和 Peter Auer。《奖励和转移任意变化的马尔可夫决策过程的滑动窗口算法》。*arXiv* 预印本 *arXiv:1805.10066*，2018 年。Aur´elien Garivier 和 Eric Moulines。《切换多臂老虎机问题的上置信界策略》。载于算法学习理论：第*22*届国际会议，*ALT 2011*，埃斯波，芬兰，*2011*年*10*月*5*日至*7*日。论文集 *22*，第 174–188 页。施普林格，2011 年。Itai Gurvich、Martin Lariviere 和 Antonio Moreno。《按需经济中的运营：具备自主排班能力的人员配置服务》。共享经济：让供给满足需求，第 249–278 页，2019 年。Woonghee Tim Huh 和 Paat Rusmevichientong。《删失需求下库存规划的非参数渐近分析》。运筹学数学，34(1):103–123，2009 年。Chi Jin、Zeyuan Allen-Zhu、Sebastien Bubeck 和 Michael I Jordan。《Q学习是否已被证明是高效的？》神经信息处理系统进展，31，2018 年。Chi Jin、Tiancheng Jin、Haipeng Luo、Suvrit Sra 和 Tiancheng Yu。《利用赌博机反馈和未知转移学习对抗性马尔可夫决策过程》。载于国际机器学习会议，第 4860–4869 页。PMLR，2020 年。

Yash Kanoria 和Pengyu Qian。《通过镜像反压进行闭合网络中的盲态动态资源分配》。*arXiv* 预印本 *arXiv:1903.02764*，2019 年。

Istv´an Z Kiss、Joel C Miller、P´eter L Simon 等。《网络流行病学数学》。沙姆：施普林格，598:31，2017年。

Jeremias Knoblauch 和 Theodoros Damoulas。《结合模型选择的时空贝叶斯在线变点检测》。载于国际机器学习会议，第2718–2727页。PMLR，2018 年。

Odalric-Ambrym Maillard。《统计序贯决策数学》。2019a。

Odalric-Ambrym Maillard。《序贯变点检测：扫描统计量的拉普拉斯集中与非渐近延迟界》。载于 Aur´elien Garivier 和 Satyen Kale（编），第*30*届国际算法学习理论会议论文集，第98卷《机器学习研究论文集》，第610–632页。PMLR，2019年3月22日至24日。网址：https://proceedings.mlr.press/v98/maillard19a.html。

Gergely Neu、Andras Antos、Andr´as Gy¨orgy 和 Csaba Szepesv´ari。《赌博机反馈下的在线马尔可夫决策过程》。神经信息处理系统进展，23，2010 年。

Cameron Nowzari、Victor MPreciado 和 GeorgeJ Pappas。《流行病的分析与控制：复杂网络中传播过程综述》。*IEEE* 控制系统杂志，36(1):26–46，2016年。

Ronald Ortner、Pratik Gajane 和 Peter Auer。《强化学习的变分遗憾界》。载于人工智能中的不确定性，第 81–90 页。PMLR，2020 年。Ewan S Page。《连续检验方案》。*Biometrika*，41(1/2):100–115，1954 年。Martin L Puterman。《马尔可夫决策过程：离散随机动态规划》。John Wiley & Sons，2014 年。

Adrian Rivera Cardoso、He Wang 和 Huan Xu。带有变化奖励的大规模马尔可夫决策过程。神经信息处理系统进展，32，2019 年。

Yunus Saatc.i、Ryan D Turner 和 Carl E Rasmussen。高斯过程变点模型。载于 第*27*届国际机器学习会议论文集（*ICML-10*），第 927–934 页，2010 年。

Milton Severo 和 Joao Gama。使用卡尔曼滤波器和累积和进行变化检测。载于 发现科学：第*9*届国际会议，*DS 2006*，巴塞罗那，西班牙，*2006*年*10*月*7*日至*10*日。论文集 *9*，第 243–254 页。施普林格，2006 年。

Albert N Shiryaev。关于快速检测问题中的最优方法。概率论及其应用，8（1）：22–46，1963 年。

Aaron Sidford、Mengdi Wang、Xian Wu、Lin F Yang 和 Yinyu Ye。使用生成模型求解折扣马尔可夫决策过程的近最优时间复杂度和样本复杂度。*arXiv* 预印本 *arXiv:1806.01492*，2018 年 a。

Aaron Sidford、Mengdi Wang、Xian Wu 和 Yinyu Ye。方差缩减值迭代及求解马尔可夫决策过程的更快算法。载于第二十九届年度 *ACM-SIAM*离散算法研讨会论文集，第 770–787 页。工业与应用数学学会，2018 年 b。

Alexander L Strehl 和 Michael L Littman。马尔可夫决策过程中基于模型的区间估计分析。计算机与系统科学杂志，74(8):1309–1331，2008 年。

Terry A Taylor。按需服务平台。制造与服务运营管理，20(4):704–720，2018 年。Mengdi Wang。随机线性规划以近线性（有时是次线性）时间求解马尔可夫决策问题。运筹学数学，45(2):517–546，2020 年。Kuang Xu 和 Se-Young Yun。具有衰减记忆的强化学习。运筹学数学，45(4): 1258–1288，2020 年。

Jia Yuan Yu 和 Shie Mannor。任意调制马尔可夫决策过程。载于第*48*届 *IEEE* 决策与控制会议（*CDC*）论文集（与 *2009* 年第*28*届中国控制会议联合举办），第 2946–2953 页。IEEE，2009 年。

Andrea Zanette 和 Emma Brunskill。利用价值函数界，在不依赖领域知识的强化学习中获得更紧的依赖问题的遗憾界。载于国际机器学习会议，第 7304–7312 页。PMLR，2019 年。

Zihan Zhang 和 Xiangyang Ji。通过评估最优偏差函数实现强化学习中的遗憾最小化。神经信息处理系统进展，32，2019 年。

Zhengyuan Zhou 和 Nicholas Bambos。固定环境和随机环境中的无线通信博弈。载于*2015*年第*54*届*IEEE*决策与控制会议（*CDC*），第1637–1642页。IEEE，2015年。

Zhengyuan Zhou、Peter Glynn 和 Nicholas Bambos。《无线通信中的功率控制重复博弈：均衡与遗憾》。载于*2016*年第*55*届*IEEE*决策与控制会议 *(CDC)*，第 3603–3610 页。IEEE，2016年。

一个 UCRL2框架

我们介绍如下学习框架和相关记号：

- $N_k(o,\ a)$ 表示在回合 k 之前，所有状态$o \in \mathsf{O}$ 和动作 $a \in \mathsf{A}$ 中，任意动作 action a 在状态 o 下执行的次数。
- $V_k(o,\ a)$ 表示截至回合 k，状态-动作对 $(o,\ a)$ 被访问的次数。
- $P_k(o,\ a,\ o')$ 是状态转移核：在状态 $o \in \mathsf{O}$ 下采取动作 $a \in \mathsf{A}$ 会使智能体转移到状态$o' \in \mathsf{O}$，其中 P_k 定义为从上一次 UCLR2 重启时间 r开始截至时间 t_k。
- $P_k(o' \mid o,\ a)$ 是根据最近的 $t_k - r + 1$ 次观测对三元组 $(o,\ a,\ o')$ 估计得到的状态转移核。
- $R_k(o,\ a)$ 是从上一次 UCLR2 重启时间r开始到时间 t_k 为止，状态-动作对 $(o,\ a)$ 的平均奖励。
- $R_k(o,\ a)$ 是根据最近 $t_k - r + 1$ 次观测估算的状态-动作对 $(o,\ a)$ 的平均奖励。
- $\mathcal{M}_k$ 是给定 $P_k(o' \mid o,\ a)$ 和 $R_k(o,\ a)$ 时在统计上合理的马尔可夫决策过程集合，其状态空间为 O，动作空间为 A。
- M_k 是从 $\mathcal{M}_k$中选出的一个乐观马尔可夫决策过程。
- $P(\cdot \mid o,\ a)$ 是 M_k 的转移核，与 $P_k(\cdot \mid o,\ a)$ 接近。
- $R(o,\ a)$ 是 M_k 的平均奖励，与 $R_k(o,\ a)$ 接近。
- $\tilde{\pi}_k$ 是一个近似最优策略，适用于 M_k 通过扩展价值迭代选取，如 Auer 等人 (2008b中所定义。

B 多项分布情形下的累积损失控制

B.1 记号与有用定义

下文中，我们用 $\Sigma_{o,s:t}$ 表示实现值 $o \in \{1, \ldots O\}$ 在序列$\mathbf{x}_{s:t}$ 中被观测到的次数，使得：

$$\Sigma_{o,s:t} = \sum_{s'=s}^{t} \mathbb{I}\{x_{s'} = o\}$$

B.2 累积损失的闭式表达

注意：

$$\forall \mathbf{x}_{s:t} \in \{1, \ldots, O\}^{n_{s:t}} \quad \widehat{L}_{s:t} := \sum_{s'=s}^{t} -\log \textsc{Pred}\left(x_t | \mathbf{x}_{s':t-1}\right) = -\log \prod_{s'=s}^{t} \textsc{Pred}\left(x_t | \mathbf{x}_{s':t-1}\right)$$

我们通过对 $n_{s:t} \in \mathbb{N}^\star$ 进行归纳法来证明：

$$\forall \mathbf{x}_{s:t} \in \{1, \ldots, O\}^{n_{s:t}} \quad \prod_{s'=s}^{t} \textsc{Pred}\left(x_t | \mathbf{x}_{s':t-1}\right) = \frac{(O-1)!}{\left(\prod_{i=1}^{O-1} (n_{s:t} + i)\right)} \times \frac{\prod_{o=1}^{O} \Sigma_{o,s:t}!}{n_{s:t}!} \tag{11}$$

方程（11）证明：

步骤 1： 对于 $n_{s:t}=1$。这意味着 $t=s$、 $x_t \in \{1,...,O\}$ 和 $\mathbf{x}_{s':\,t-1}=\emptyset$。根据预测器 PRED 的定义（|），我们得到

$$\textsc{Pred}\left(x_t|\emptyset\right) := \frac{1}{O} = \frac{(O-1)!}{\left(\prod_{i=1}^{O-1}(1+i)\right)} \times \frac{\prod_{o=1}^{O}1}{1!}$$

步骤 2： 假设对于某个 $n_{s:t} \in \mathbb{N}^{\star}$，它对应于序列 $\mathbf{x}_{s:t} \in \{1,...,O\}$，我们有

$$\prod_{s'=s}^{t}\textsc{Pred}\left(x_t|\mathbf{x}_{s':t-1}\right) = \frac{(O-1)!\prod_{o=1}^{O}\Sigma_{o,s:t}!}{(n_{s:t}+O-1)!} = \frac{(O-1)!\prod_{o=1}^{O}\Sigma_{o,s:t}!}{\left(\prod_{i=1}^{O-1}(n_{s:t}+i)\right)n_{s:t}!} \tag{12}$$

然后，注意到：

$$\prod_{s'=s}^{t+1}\textsc{Pred}\left(x_{t+1}|\mathbf{x}_{s':t}\right) = \prod_{s'=s+1}^{t+1}\textsc{Pred}\left(x_{t+1}|\mathbf{x}_{s':t}\right) \times \textsc{Pred}\left(x_{t+1}|\mathbf{x}_{s:t}\right)$$

然后，根据方程（2）中预测器的定义以及方程（12）的表述，我们得到（对于 $x_{t+1}=a \in \{1,...,O\}$）

$$
\begin{aligned}
\prod_{s'=s}^{t+1} \textsc{Pred}\left(x_{t+1}|\mathbf{x}_{s':t}\right) &= \frac{(O-1)!\prod_{o=1}^{O}\Sigma_{o,s+1:t+1}!}{\left(\prod_{i=1}^{O-1}\left(n_{s+1:t+1}+i\right)\right)n_{s+1:t+1}!} \times \frac{\sum_{i=s}^{t}\mathbb{I}\{x_i=x_{t+1}\}+1}{n_{s:t}+O} \\
&= \frac{(O-1)!\prod_{o=1}^{O}\Sigma_{o,s+1:t+1}!}{\left(\prod_{i=1}^{O-1}\left(n_{s+1:t+1}+i\right)\right)n_{s+1:t+1}!} \times \frac{\Sigma_{a,s:t}+1}{n_{s:t}+O} \\
&= \frac{(O-1)!\prod_{o=1}^{O}\Sigma_{o,s+1:t+1}!}{\left(\prod_{i=1}^{O-1}\left(n_{s:t}+i\right)\right)n_{s:t}!} \times \frac{\Sigma_{a,s:t+1}}{n_{s:t}+O} \\
&= \frac{(O-1)!\prod_{o=1}^{O}\Sigma_{o,s:t+1}!}{\left(\prod_{i=1}^{O-1}\left(n_{s:t}+i\right)\right)n_{s:t}!} \times \frac{1}{n_{s:t}+O} \\
&= \frac{(O-1)!\prod_{o=1}^{O}\Sigma_{o,s:t+1}!}{\left(\prod_{i=1}^{O-1}\left(n_{s:t+1}+i-1\right)\right)(n_{s:t+1}-1)!} \times \frac{1}{n_{s:t+1}+O-1} \\
&= \frac{(O-1)!\prod_{o=1}^{O}\Sigma_{o,s:t+1}!}{(n_{s:t+1}+O-1)\left(\prod_{i=1}^{O-1}\left(n_{s:t+1}+i-1\right)\right)(n_{s:t+1}-1)!} \\
&= \frac{(O-1)!\prod_{o=1}^{O}\Sigma_{o,s:t+1}!}{\left(\prod_{i=1}^{O-1}\left(n_{s:t+1}+i\right)\right)(n_{s:t+1})!}
\end{aligned}
$$

□

^注意，累积损失 $L_{s:t}$ 可以写成如下形式：

$$
\begin{aligned}
\widehat{L}_{s:t} &= \log\left((n_{s:t}+O-1)!\right) - \sum_{o=1}^{O}\log\left(\Sigma_{o,s:t}!\right) - \log(O-1)! \\
&= \sum_{i=1}^{O-1}\log\left(n_{s:t}+i\right) + \log\left(n_{s:t}!\right) - \sum_{o=1}^{O}\log\left(\Sigma_{o,s:t}!\right) - \log(O-1)!
\end{aligned}
$$

其中， $n!$ 表示 n 的阶乘，使得：

$$
n! = n \times (n-1) \times (n-2) \times ... \times 1
$$

然后，使用以下斯特林公式：

$$
\forall\, n \geqslant 1 \quad \sqrt{2\pi n}\left(\frac{n}{e}\right)^n \leqslant n! \leqslant \sqrt{2\pi n}\left(\frac{n}{e}\right)^n \exp\left(\frac{1}{12}\right),
$$

我们得到量 $^{n!}{}_{n_1!\ n_2!\ \ldots n_O!}$ 的上界和下界：

$$\frac{n^n}{n_1^{n_1} n_2^{n_2} \ldots n_O^{n_O}} \times \frac{\exp(b_1)}{n^{\frac{o-1}{2}}} \leqslant \frac{n!}{n_1! n_2! \ldots n_O!} \leqslant \frac{n^n}{n_1^{n_1} n_2^{n_2} \ldots n_O^{n_O}} \tag{13}$$

以及 $\sum_{i=1}^{O} n_i = n, \quad n_i \geqslant 0 \; \forall i \in \{1, \ldots, O\}$ $b_1 = -\frac{O}{12} - \frac{O-1}{2} \log (2\pi) + \frac{O}{2} \log O$

B.3 平稳观测下累积损失的上界

在推导累积损失的上界之前，应注意：

$$\Sigma_{s:t} \log \Sigma_{s:t} + \bar{\Sigma}_{s:t} \log \bar{\Sigma}_{s:t} = \Sigma_{s:t} \log \theta + \bar{\Sigma}_{s:t} \log \bar{\theta} + n_{s:t} \log n_{s:t} + n_{s:t} \boldsymbol{KL} \left(\frac{\Sigma_{s:t}}{n_{s:t}} \,||\, \theta \right). \tag{14}$$

$$\sum_{o=1}^{O} \Phi\left(\Sigma_{o,s:t}\right) = \sum_{o=1}^{O} \Sigma_{o,s:t} \log \Sigma_{o,s:t} = \sum_{o=1}^{O} \Sigma_{o,s:t} \log \theta_o + n_{s:t} \log n_{s:t} + n_{s:t} \mathbf{KL}\left(\frac{\Sigma_{1,s:t}}{n_{s:t}}, \ldots, \frac{\Sigma_{O,s:t}}{n_{s:t}} \,\middle\|\, \theta_1, \ldots, \theta_O \right) \tag{15}$$

$$\sum_{o=1}^{O} \Phi\left(\Sigma_{o,s:t}\right) = \sum_{o=1}^{O} \Sigma_{o,s:t} \log \Sigma_{o,s:t} = \sum_{o=1}^{O} \Sigma_{o,s:t} \log \theta_o + n_{s:t} \log n_{s:t} + n_{s:t} \mathbf{KL}(\widehat{\mu}_{1,s:t}, \ldots, \widehat{\mu}_{O,s:t} \parallel \theta_1, \ldots, \theta_O) \tag{16}$$

$$\sum_{o=1}^{O} \Phi\left(\Sigma_{o,s:t}\right) = \sum_{o=1}^{O} \Sigma_{o,s:t} \log \Sigma_{o,s:t} = \sum_{o=1}^{O} \Sigma_{o,s:t} \log \theta_o + \Phi\left(n_{s:t}\right) + n_{s:t} \mathbf{KL}(\widehat{\mu}_{1,s:t}, \ldots, \widehat{\mu}_{O,s:t} \parallel \theta_1, \ldots, \theta_O) \tag{17}$$

于是，累积损失的上界形式如下：

$$\widehat{L}_{s:t} \overset{(a)}{\leqslant} \Phi\left(n_{s:t}\right) - \sum_{o=1}^{O} \Phi\left(\Sigma_{o,s:t}\right) + \sum_{i=1}^{O-1} \log\left(n_{s:t} + i\right) - \log(O-1)! \tag{18}$$

$$\overset{(b)}{\leqslant} \sum_{i=1}^{O-1} \log\left(n_{s:t} + i\right) - \sum_{o=1}^{O} \Sigma_{o,s:t} \log \theta_o - n_{s:t} \mathbf{KL}(\widehat{\mu}_{1,s:t}, \ldots, \widehat{\mu}_{O,s:t} \parallel \theta_1, \ldots, \theta_O) - \log(O-1)!$$

$$\overset{(c)}{\leqslant} \sum_{i=1}^{O-1} \log\left(n_{s:t} + i\right) - \sum_{o=1}^{O} \Sigma_{o,s:t} \log \theta_o - \log(O-1)! \tag{19}$$

其中：

- (a) 利用方程 (13) 的左侧可得 $n = n_{s:\ t}$ 和 $a_o = \Sigma_{o,\ s:\ t} \ \forall o \in \{1, \ldots, O\}$。
- (b) 根据方程 (17) 的陈述可得。
- (c) 根据 Kullback-Leibler 散度始终为正这一事实可得（即 $\mathbf{KL}(\bullet \parallel \bullet) \geqslant 0$）。

B.4 平稳观测的累积损失下界

累积损失的下界具有以下形式：

$$\widehat{L}_{s:t} \overset{(a)}{\geqslant} \Phi\left(n_{s:t}\right) - \sum_{o=1}^{O} \Phi\left(\Sigma_{o,s:t}\right) + \sum_{i=1}^{O-1} \log\left(n_{s:t} + i\right) - \frac{O-1}{2} \log n_{s:t} + b_1 - \log(O-1)! \tag{20}$$

$$\overset{(b)}{\geqslant} \sum_{i=1}^{O-1} \log\left(n_{s:t} + i\right) - \sum_{o=1}^{O} \Sigma_{o,s:t} \log \theta_o - n_{s:t} \mathbf{KL}(\widehat{\mu}_{1,s:t}, \ldots, \widehat{\mu}_{O,s:t} \parallel \theta_1, \ldots, \theta_O) - \frac{O-1}{2} \log n_{s:t} + b_1 - \log(O-1)! \tag{21}$$

其中：

- （a）利用方程（13）的左侧可得，对于 $n = n_{s:t}$ 以及 $n_o = \Sigma_{o,\ s:t}\ \forall o \in \{1, ..., O\}$
- （b）根据方程（17）的表述可得。

用于推导虚警率和检测延迟的有用引理

引理 *B.1*（时间一致的 $KL(\bullet \parallel \bullet)$ 集中性）。设：$\boldsymbol{\theta} = (\theta_1, ..., \theta_O)$ 表示多项分布 $Multi\ (\theta_1, ..., \theta_O)$ 的生成参数向量。$\forall o \in \{1, ..., O\}$，令 $\widehat{\mu}_{o,t}$ 表示在序列 $(x_1, ..., x_t) \sim Multi\ (\theta_1, ..., \theta_O)^{\otimes t}$ 中观察到实现值 $o \in \{1, ..., O\}$ 的经验频率，则对于所有 $(\delta, \alpha) \in (0,1) \times (1, \infty)$，我们有：

$$\mathbb{P}_{\boldsymbol{\theta}}\underbrace{\left\{ \forall t \in \mathbb{N}^\star : \boldsymbol{KL}(\widehat{\mu}_{1,t}, ..., \widehat{\mu}_{O,t} \parallel \theta_1, ..., \theta_O) < \frac{\alpha}{t} \log \frac{\log(\alpha t) \log(t)}{\log^2(\alpha)\delta} \right\}}_{E^{(1)}_{\boldsymbol{\theta},\delta,\alpha}} \geqslant 1 - \delta$$

引理 B.2（双时间一致 **KL**（$\bullet \parallel \bullet$）浓缩）。设：$\boldsymbol{\theta} =$（$\theta_1$、...、$\theta_O$）表示多项分布 *Multi*（$\theta_1$、...、$\theta_O$）的生成参数向量。

$\forall o \in \{1, ..., O\}$，设 $\widehat{\mu}_{o,\ s:t}$ 表示在序列（x_s、...、x_t）$\sim Multi$（θ_1、...、θ_O）$^{\otimes n_{s:t}}$中观测到 o 的经验频率，则对于所有（δ、α）$\in$（0、1）$\times$（1、∞），我们有：

$$\mathbb{P}_{\boldsymbol{\theta}}\underbrace{\left\{ \forall t \in \mathbb{N}^\star, \forall s \in (r, t] : \boldsymbol{KL}(\widehat{\mu}_{1,s:t}, ..., \widehat{\mu}_{O,s:t} \parallel \theta_1, ..., \theta_O) < \frac{\alpha}{n_{s:t}} \times \log \frac{n_{r:t} \log^2(n_{r:t}) \log((\alpha+1)\, n_{s:t})}{\log(2) \log^2(\alpha)\delta} \right\}}_{E^{(2)}_{\boldsymbol{\theta},\delta,\alpha}} \geqslant 1 - \delta$$

引理 B.3（双时间一致浓缩）。设：（x_r、...x_t）$\in \{1, ..., O\}^{n_{r:t}}$ 为一组从多项分布中抽取的独立随机变量，其生成参数可以任意选择，并且 $\widehat{\mu}_{o,\ i:j}$ 表示在序列（x_i、...、x_j）中观测到 o 的经验频率。则对于所有（r、δ）$\in \mathbb{N}^\star \times$（0、1），我们得到以下控制：

$$\mathbb{P}\left\{ \exists\, t > r, s \in [r, t) : |\widehat{\mu}_{o,r:s-1} - \widehat{\mu}_{o,s:t} - \mathbb{E}\left[\widehat{\mu}_{o,r:s-1} - \widehat{\mu}_{o,s:t}\right]| \geqslant \mathcal{C}'_{r,s,t,\delta} \right\} \leqslant \delta,$$

$$\mathcal{C}'_{r,s,t,\delta} = \frac{\sqrt{2}}{2}\left(\sqrt{\frac{1 + \frac{1}{n_{r:s-1}}}{n_{r:s-1}} \log\left(\frac{2\sqrt{n_{r:s}}}{\delta}\right)} + \sqrt{\frac{1 + \frac{1}{n_{s:t}}}{n_{s:t}} \log\left(\frac{2n_{r:t}\sqrt{n_{s:t}+1}\log^2(n_{r:t})}{\log(2)\delta}\right)} \right).$$

引理 B.1、B.2和B.3 的证明超出了本文的范围，有兴趣的读者可参阅**第 3.4 节** 中Maillard（2019a）。

C 虚警率的推导

定理4.6的证明：

令 $\boldsymbol{\theta} = (\theta_1, ..., \theta_O)$ 表示多项分布 Multi $(\theta_1, ..., \theta_O)^{\otimes n_{r:t}}$ 的生成参数向量。

假设 设 $\forall t \in [r, c_\ell)$ $(x_r, ..., x_t) \sim$ Multi $(\theta_1, ..., \theta_O)^{\otimes n_{r:t}}$。证明分为三个主要步骤 s:
让我们构造一个合适的 $\eta_{r,\ s,\ t}$ 值，以确保在周期 $[r, c_\ell)$ 上控制虚警。为此，我们来控制事件：$\{\exists t > r$, 重启 $(x_r, ..., x_t) = 1\}$，这等价于事件$\{\exists t > r,\ s \in (r,\ t] : \omega_{r,\ s,\ t} \geqslant \omega_{r,\ r,\ t}\}$。

步骤 1：等价事件。首先，注意到：

$$\begin{aligned}\left\{\exists t>r,\ s\in(r,t]:\omega_{r,s,t}\geqslant\omega_{r,r,t}\right\}&\Leftrightarrow\left\{\exists t>r,\ s\in(r,t]:\quad\log\omega_{r,s,t}\geqslant\log\omega_{r,r,t}\right\}.\\&\overset{(a)}{\Leftrightarrow}\left\{\exists t>r,\ s\in(r,t]:-\log\eta_{r,s,t}\leqslant\widehat{L}_{r:t}-\widehat{L}_{s:t}-\widehat{L}_{r:s-1}\right\}\end{aligned}\tag{22}$$

其中，(a) 直接来自预测器权重的定义$\omega_{r,\ s,\ t}$ 该定义见方程（4）。

Step 2: Using the cumulative loss controls. Then, note that $\forall\delta\in(0,1)\,,\forall\alpha>1$ we have:

$$\begin{aligned}&\mathbb{P}_{\boldsymbol{\theta}}\Big\{\exists t>r,\ s\in(r,t]:\omega_{r,s,t}\geqslant\omega_{r,r,t}\Big\}\overset{(a)}{=}\mathbb{P}_{\boldsymbol{\theta}}\Big\{\exists t>r,\ s\in(r,t]:\log\omega_{r,s,t}\geqslant\log\omega_{r,r,t}\Big\}\\&\overset{(b)}{=}\mathbb{P}_{\boldsymbol{\theta}}\Big\{\exists t>r,\ s\in(r,t]:\quad-\log\eta_{r,s,t}\leqslant\widehat{L}_{r:t}-\widehat{L}_{r:s-1}-\widehat{L}_{s:t}\Big\}\\&\overset{(c)}{\leqslant}\mathbb{P}_{\boldsymbol{\theta}}\Big\{\exists t>r,\ s\in(r,t]:-\log\eta_{r,s,t}\leqslant\sum_{i=1}^{O-1}\log\left(n_{r:t}+i\right)-\sum_{i=1}^{O-1}\log\left(n_{r:s-1}+i\right)-\sum_{i=1}^{O-1}\log\left(n_{s:t}+i\right)\\&\qquad+\frac{O-1}{2}\log n_{r:s-1}+\frac{O-1}{2}\log n_{s:t}-2b_1+n_{s:t}\mathbf{KL}(\widehat{\mu}_{1,s:t},...,\widehat{\mu}_{O,s:t}\parallel\theta_1,...,\theta_O)\\&\qquad+n_{r:s-1}\mathbf{KL}(\widehat{\mu}_{1,r:s-1},...,\widehat{\mu}_{O,r:s-1}\parallel\theta_1,...,\theta_O)+\log(O-1)!\Big\}\\&\overset{(d)}{\leqslant}\mathbb{P}_{\boldsymbol{\theta}}\Big\{\exists t>r,\ s\in(r,t]:-\log\eta_{r,s,t}\leqslant\sum_{i=1}^{O-1}\log\frac{n_{r:t}+i}{\left(n_{r:s-1}+i\right)\left(n_{s:t}+i\right)}+\frac{O-1}{2}\log\left(n_{r:s-1}n_{s:t}\right)-2b_1\\&\qquad+n_{s:t}\mathbf{KL}(\widehat{\mu}_{1,s:t},...,\widehat{\mu}_{O,s:t}\parallel\theta_1,...,\theta_O)+n_{r:s-1}\mathbf{KL}(\widehat{\mu}_{1,r:s-1},...,\widehat{\mu}_{O,r:s-1}\parallel\theta_1,...,\theta_O)+\log(O-1)!\Big\}\\&\overset{(e)}{\leqslant}\frac{\delta}{2}+\mathbb{P}_{\boldsymbol{\theta}}\Big\{\exists t>r,\ s\in(r,t]:-\log\eta_{r,s,t}\leqslant\sum_{i=1}^{O-1}\log\frac{n_{r:t}+i}{\left(n_{r:s-1}+i\right)\left(n_{s:t}+i\right)}+\frac{O-1}{2}\log\left(n_{r:s-1}n_{s:t}\right)-2b_1\\&\quad+n_{s:t}\mathbf{KL}(\widehat{\mu}_{1,s:t},...,\widehat{\mu}_{O,s:t}\parallel\theta_1,...,\theta_O)+n_{r:s-1}\mathbf{KL}(\widehat{\mu}_{1,r:s-1},...,\widehat{\mu}_{O,r:s-1}\parallel\theta_1,...,\theta_O)+\log(O-1)!\bigcap E^{(1)}_{\boldsymbol{\theta},\delta/2,\alpha}\Big\}\\&\overset{(f)}{\leqslant}\frac{\delta}{2}+\mathbb{P}_{\boldsymbol{\theta}}\Big\{\exists t>r,\ s\in(r,t]:-\log\eta_{r,s,t}\leqslant\sum_{i=1}^{O-1}\log\frac{n_{r:t}+i}{\left(n_{r:s-1}+i\right)\left(n_{s:t}+i\right)}+\frac{O-1}{2}\log\left(n_{r:s-1}n_{s:t}\right)-2b_1\\&\qquad+n_{s:t}\mathbf{KL}(\widehat{\mu}_{1,s:t},...,\widehat{\mu}_{O,s:t}\parallel\theta_1,...,\theta_O)+\alpha\log\frac{2\log(\alpha n_{r:s-1})\log(n_{r:s-1})}{\log^2(\alpha)\delta}+\log(O-1)!\Big\}\\&\overset{(g)}{\leqslant}\delta+\mathbb{P}_{\boldsymbol{\theta}}\Big\{\exists t>r,\ s\in(r,t]:-\log\eta_{r,s,t}\leqslant\sum_{i=1}^{O-1}\log\frac{n_{r:t}+i}{\left(n_{r:s-1}+i\right)\left(n_{s:t}+i\right)}+\frac{O-1}{2}\log\left(n_{r:s-1}n_{s:t}\right)-2b_1\\&\qquad+n_{s:t}\mathbf{KL}(\widehat{\mu}_{1,s:t},...,\widehat{\mu}_{O,s:t}\parallel\theta_1,...,\theta_O)+\alpha\log\frac{2\log(\alpha n_{r:s-1})\log(n_{r:s-1})}{\log^2(\alpha)\delta}+\log(O-1)!\bigcap E^{(2)}_{\boldsymbol{\theta},\delta/2,\alpha}\Big\}\\&\overset{(h)}{\leqslant}\delta+\mathbb{P}_{\boldsymbol{\theta}}\Big\{\exists t>r,\ s\in(r,t]:-\log\eta_{r,s,t}\leqslant\sum_{i=1}^{O-1}\log\frac{n_{r:t}+i}{\left(n_{r:s-1}+i\right)\left(n_{s:t}+i\right)}+\frac{O-1}{2}\log\left(n_{r:s-1}n_{s:t}\right)-2b_1\\&\qquad+\alpha\log\frac{2n_{r:t}\log^2(n_{r:t})\log(\alpha n_{s:t})\log(n_{s:t})}{\log(2)\log^2(\alpha)\delta}+\alpha\log\frac{2\log(\alpha n_{r:s-1})\log(n_{r:s-1})}{\log^2(\alpha)\delta}+\log(O-1)!\Big\}\\&\overset{(i)}{\leqslant}\delta+\mathbb{P}_{\boldsymbol{\theta}}\Big\{\exists t>r,\ s\in(r,t]:-\log\eta_{r,s,t}\leqslant\log\left(\prod_{i=1}^{O-1}\frac{n_{r:t}+i}{\left(n_{r:s-1}+i\right)\left(n_{s:t}+i\right)}\times\left(n_{r:s-1}n_{s:t}\right)^{\frac{O-1}{2}}\right)-2b_1\\&\qquad+\alpha\log\frac{2n_{r:t}\log^2(n_{r:t})\log(\alpha n_{s:t})\log(n_{s:t})}{\log(2)\log^2(\alpha)\delta}+\alpha\log\frac{2\log(\alpha n_{r:s-1})\log(n_{r:s-1})}{\log^2(\alpha)\delta}+\log(O-1)!\Big\}\end{aligned}\tag{23}$$

其中：

- (a) 利用了对数函数的单调性，因此成立。
- （b）得益于方程（4）。
- （c）得益于使用方程（21）中累积损失的下界，以及方程（19）中累积损失的上界。
- （d）利用对数函数的基本性质即可得到。
- （e）利用以下性质即可得到：$\mathbb{P}\Big\{A\Big\} \leqslant \mathbb{P}\Big\{\neg B\Big\} + \mathbb{P}\Big\{A \cap B\Big\}$ 其中 $B = E_{(1)}^{\boldsymbol{\theta},\ \delta/2,\ \alpha}$。
- (f) 根据引理B.1的陈述成立。
- (g) 利用以下性质可得：$\mathbb{P}\Big\{A\Big\} \leqslant \mathbb{P}\Big\{\neg B\Big\} + \mathbb{P}\Big\{A \cap B\Big\}$ 其中 $B = E_{(2)}^{\boldsymbol{\theta},\ \delta/2,\ \alpha}$。
- (h) 根据引理B.2的陈述成立。
- (i) 利用对数函数的单调性可得。

步骤 3： 关于 $\eta_{r,\ s,\ t}$ 根据方程（23），我们推导出保证虚警控制的充分条件：$\eta_{r,\ s,\ t}$

$$
\begin{aligned}
\eta_{r,s,t} &< \left(\prod_{i=1}^{O-1} \frac{(n_{r:s-1}+i)\,(n_{s:t}+i)}{n_{r:t}+i}\right) \times \frac{\exp(2b_1)}{(n_{r:s-1}n_{s:t})^{\frac{O-1}{2}} \times (O-1)!} \\
&\times \left(\frac{\log^2(\alpha)\delta}{2\log(\alpha n_{r:s-1})\log(n_{r:s-1})} \times \frac{\log(2)\log^2(\alpha)\delta}{2n_{r:t}\log^2(n_{r:t})\log(\alpha n_{s:t})\log(n_{s:t})}\right)^{\alpha} \\
&= \left(\prod_{i=1}^{O-1} \frac{(n_{r:s-1}+i)\,(n_{s:t}+i)}{n_{r:t}+i}\right) \times \frac{\exp(2b_1)}{(n_{r:s-1}n_{s:t})^{\frac{O-1}{2}} \times (O-1)!} \times \left(\frac{\log(4\alpha)\log(2)\delta^2}{4n_{r:t}\log(\alpha n_{r:t})\log^2(n_{r:t})\log(n_{r:t})}\right)^{\alpha} \\
&= \left(\prod_{i=1}^{O-1} \frac{(n_{r:s-1}+i)\,(n_{s:t}+i)}{n_{r:t}+i}\right) \times \frac{\exp(2b_1)}{(n_{r:s-1}n_{s:t})^{\frac{O-1}{2}} \times (O-1)!} \times \left(\frac{\log(4\alpha+2)\delta^2}{4n_{r:t}\log((\alpha+3)\,n_{r:t})}\right)^{\alpha}
\end{aligned}
$$

然后，在平稳时段内以高概率不会发生虚警 $[r, c_\ell)$：

$$
\mathbb{P}_{\boldsymbol{\theta}}\Big\{\exists\, t \in [r, c_\ell) : \textbf{\textit{Restart}}\,(x_r, ..., x_t) = 1\Big\} \leqslant \delta.
$$

□

D 检测延迟的推导

定理4.8：

证明分为三个主要步骤：

步骤 1：一些预备知识 在构建检测延迟之前，我们需要介绍三个中间结果。第一个结果将数量 $\Phi(\Sigma_{o,s:t})$ 与 $\Phi(\widehat{\mu}_{o,s:t})$ 联系起来，如下所示：

$$
\forall\,(s,t): \quad \sum_{o=1}^{O} \Phi\left(\Sigma_{o,s:t}\right) - \Phi\left(n_{s:t}\right) = n_{s:t} \sum_{o=1}^{O} \Phi\left(\widehat{\mu}_{o,s:t}\right). \tag{24}
$$

使用记号： $\widehat{\boldsymbol{\mu}}_{a:b} = (\widehat{\mu}_{1,a:b},...,\widehat{\mu}_{O,a:b})\ \forall\ a < b$然后，注意到：

$$n_{r:s-1}\sum_{o=1}^{O}\Phi\left(\widehat{\mu}_{o,r:s-1}\right)+n_{s:t}\sum_{o=1}^{O}\Phi\left(\widehat{\mu}_{o,s:t}\right)-n_{r:t}\sum_{o=1}^{O}\Phi\left(\widehat{\mu}_{o,r:t}\right)=n_{r:s-1}\mathbf{KL}(\widehat{\boldsymbol{\mu}}_{r:s-1}\parallel\widehat{\boldsymbol{\mu}}_{r:t})+n_{s:t}\mathbf{KL}(\widehat{\boldsymbol{\mu}}_{s:t}\parallel\widehat{\boldsymbol{\mu}}_{r:t}).$$
$$=n_{r:s-1}\mathbf{KL}(\widehat{\mu}_{1,r:s-1},...,\widehat{\mu}_{O,r:s-1}\parallel\widehat{\mu}_{1,r:t},...,\widehat{\mu}_{O,r:t})+n_{s:t}\mathbf{KL}(\widehat{\mu}_{1,s:t},...,\widehat{\mu}_{O,s:t}\parallel\widehat{\mu}_{1,r:t},...,\widehat{\mu}_{O,r:t}) \tag{25}$$

那么，观察到：

$$\forall o\in\{1,...,O\}\quad n_{r:s-1}\left(\widehat{\mu}_{o,r:s-1}-\widehat{\mu}_{o,r:t}\right)^2+n_{s:t}\left(\widehat{\mu}_{o,s:t}-\widehat{\mu}_{o,r:t}\right)^2=\frac{n_{r:s-1}n_{s:t}}{n_{r:t}}\left(\widehat{\mu}_{o,r:s-1}-\widehat{\mu}_{o,s:t}\right)^2. \tag{26}$$

那么，我们还需要一个有用的记号$f_{r,\ s,\ t}$：

$$f_{r,s,t}=\sum_{i=1}^{O-1}\log\left(n_{r:s-1}+i\right)+\sum_{i=1}^{O-1}\log\left(\frac{n_{s:t}+i}{n_{r:t}+i}\right)-\frac{O-1}{2}\log\left(\frac{n_{s:t}}{n_{r:t}}\right)-\log(O-1)!$$

最后，根据引理B.3，对量$|\widehat{\mu}_{o,\ r:s-1}-\widehat{\mu}_{o,\ s:t}|$ 的控制形式如下：

$$\mathbb{P}_{\boldsymbol{\theta}^{(1)},\boldsymbol{\theta}^{(2)}}\Big\{\underbrace{\forall\ s\in[r:t)\quad|\widehat{\mu}_{o,r:s-1}-\widehat{\mu}_{o,s:t}|\geqslant\Delta_{o,r,s,t}-\mathcal{C}'_{r,s,t,\delta}}_{E^{(3)}_{o,r,t,\delta}}\Big\}\geqslant 1-\delta, \tag{27}$$

我们将 $\boldsymbol{\theta}^{(1)}$ 和 $\boldsymbol{\theta}^{(2)}$ 定义为针对变点 $c_{\ell+1}$的动作集 O 上的前状态转移核和后状态转移核，并将 Δ_o 定义为逐状态变化量 $\Delta_o=\left|\theta_o^{(1)}-\theta_o^{(2)}\right|$。然后，相对间隔$\Delta_{r,\ s,\ t}$ 可表示如下

$$\forall o\in\{1,..,O\}\ \Delta_{o,r,s,t}=\left|\mathbb{E}_{\boldsymbol{\theta}^{(1)},\boldsymbol{\theta}^{(2)}}\left[\widehat{\mu}_{o,r:s-1}-\widehat{\mu}_{o,s:t}\right]\right|=\begin{cases}\frac{n_{c_\ell:t}}{n_{s:t}}\left|\theta_o^{(1)}-\theta_o^{(2)}\right|=\frac{n_{c_\ell:t}}{n_{s:t}}\Delta_o & \text{if } s<c_\ell\leqslant t,\\ \frac{n_{r:c_\ell-1}}{n_{r:s-1}}\left|\theta_o^{(1)}-\theta_o^{(2)}\right|=\frac{n_{r:c_\ell-1}}{n_{r:s-1}}\Delta_o & \text{if } c_\ell\leqslant s\leqslant t.\end{cases} \tag{28}$$

多项分布的Pinsker不等式

$$\mathbf{KL}(\widehat{\mu}_{1,s:t},...,\widehat{\mu}_{O,s:t}\parallel\widehat{\mu}_{1,r:t},...,\widehat{\mu}_{O,r:t})\geqslant\frac{1}{2}\left(\sum_{o=1}^{O}|\widehat{\mu}_{o,s:t}-\widehat{\mu}_{o,r:t}|\right)^2 \tag{29}$$

步骤 2：构建检测变点的充分条件 c_ℓ设： $\boldsymbol{\theta}^{(1)}=\left(\theta_1^{(1)},...,\theta_O^{(1)}\right)\in[0,1]^O$且 $\boldsymbol{\theta}^{(2)}=\left(\theta_1^{(2)},...,\theta_O^{(2)}\right)$。首先，假设： $x_r,\ .....\ x_{c_\ell-1}\sim$多项分布 $\left(\theta_1^{(1)},...,\theta_O^{(1)}\right)$和 $x_{c_\ell},\ .....\ x_t\sim$多项分布 $\left(\theta_1^{(2)},...,\theta_O^{(2)}\right)$。然后，为了构建检测延迟，我们需要证明，在 c_ℓ之后的某个时刻，重启准则 重启 $(x_r,\ .....\ x_t)$ 被激活。换言之，我们需要构建以下保证：

$$\mathbb{P}_{\boldsymbol{\theta}^{(1)},\boldsymbol{\theta}^{(2)}}\Big\{\exists t>c_\ell:\textbf{\textit{Restart}}\left(x_r,...,x_t\right)=1\Big\}>1-\delta.$$

请注意：

$\{\forall\, t > c_\ell$: 重启 $(x_r, ..., x_t) = 0\}$ (a) $\Leftrightarrow \{\forall\, t > c_\ell, \forall s \in (r,t]$: 对数 $\omega_{r,s,t} \leqslant$ 对数 $\omega_{r,r,t}\}$. (b)
$\Leftrightarrow \{\forall\, t > c_\ell, \forall s \in (r,t]$: 对数 $\eta_{r,s,t} \leqslant \widehat{L}_{r:s-1} + \widehat{L}_{s:t} - \widehat{L}_{r:t}\}$, O (c) $\Rightarrow \{\forall\, t > c_\ell$
$, \forall s \in (r,t]$: 对数 $\eta_{r,s,t} \leqslant f_{r,s,t} - \sum_{o=1}^{O} \Phi(\Sigma_{o,r:s-1}) + \Phi(n_{r:s-1}) - \sum_{o=1}^{O} \Phi(\Sigma_{s:t}) + \Phi(n_{s:t})$
$+ \sum_{o=1}^{O} \Phi(\Sigma_{r:t}) - \Phi(n_{r:t})\}$, O (d) $\Rightarrow \{\forall\, t > c_\ell, \forall s \in (r,t]$: 对数 $\eta_{r,s,t} \leqslant f_{r,s,t} - n_{r:s-1} \sum_{o=1}^{O} \Phi$
$(\widehat{\mu}_{o,r:s-1}) - n_{s:t} \sum_{o=1}^{O} \Phi(\widehat{\mu}_{o,s:t}) + n_{r:t} \sum_{o=1}^{O} \Phi(\widehat{\mu}_{o,r:t})\}$, O (e) $\Rightarrow \{\forall\, t > c_\ell, \forall s \in (r,t]$: 对数 $\eta_{r,s,t} \leqslant f_{r,s,t} - n_{r:s-1}$ KL $(\widehat{\mu}_{1,r:s-1}, ..., \widehat{\mu}_{O,r:s-1} \,\|\, \widehat{\mu}_{1,r:t}, ..., \widehat{\mu}_{O,r:t}) - n_{s:t}$ KL $(\mu_{1,s:t}, ..., \mu_{O,s:t} \,\|\, \mu_{1,r:t}, ..., \mu_{O,r:t})\}$

$$\overset{(f)}{\Rightarrow} \left\{\forall\, t > c_\ell, \forall s \in (r,t] : \log \eta_{r,s,t} \leqslant f_{r,s,t} - \frac{n_{r:s-1}}{2} \left(\sum_{o=1}^{O} |\widehat{\mu}_{o,r:s-1} - \widehat{\mu}_{o,r:t}|\right)^2 - \frac{n_{s:t}}{2} \left(\sum_{o=1}^{O} |\widehat{\mu}_{o,s:t} - \widehat{\mu}_{o,r:t}|\right)^2 \right\}$$

$$\overset{(g)}{\Rightarrow} \left\{\forall\, t > c_\ell, \forall s \in (r,t] : \log \eta_{r,s,t} \leqslant f_{r,s,t} - \frac{n_{r:s-1}}{2} \sum_{o=1}^{O} (\widehat{\mu}_{o,r:s-1} - \widehat{\mu}_{o,r:t})^2 - \frac{n_{s:t}}{2} \sum_{o=1}^{O} (\widehat{\mu}_{o,s:t} - \widehat{\mu}_{o,r:t})^2 \right\}$$

$$\Leftrightarrow \left\{\forall\, t > c_\ell, \forall s \in (r,t] : \log \eta_{r,s,t} \leqslant f_{r,s,t} - \frac{1}{2} \sum_{o=1}^{O} \left(n_{r:s-1} (\widehat{\mu}_{o,r:s-1} - \widehat{\mu}_{o,r:t})^2 + n_{s:t} (\widehat{\mu}_{o,s:t} - \widehat{\mu}_{o,r:t})^2\right) \right\}$$

$$\overset{(h)}{\Rightarrow} \left\{\forall\, t > c_\ell, \forall s \in (r,t] : \log \eta_{r,s,t} \leqslant f_{r,s,t} - \frac{1}{2} \sum_{o=1}^{O} \frac{n_{r:s-1} n_{s:t}}{n_{r:t}} (\widehat{\mu}_{o,r:s-1} - \widehat{\mu}_{o,s:t})^2 \right\}$$

$$\Leftrightarrow \left\{\forall\, t > c_\ell, \forall s \in (r,t] : \frac{1}{2} \sum_{o=1}^{O} \frac{n_{r:s-1} n_{s:t}}{n_{r:t}} (\widehat{\mu}_{o,r:s-1} - \widehat{\mu}_{o,s:t})^2 \leqslant f_{r,s,t} - \log \eta_{r,s,t} \right\}$$

其中:

- (a) 得益于方程 (5) 中对重启过程的定义成立。
- (b) 得益于方程 (4) 中的陈述成立。
- (c) 得益于方程 (18) 中的上界以及
- (d) 得益于方程 (24) 中的陈述成立。
- (e) 根据方程 (25) 中的陈述成立。
- (f) 根据方程 (29) 成立。
- (g) 根据以下方程成立: $\left(\sum_{o=1}^{O} |\widehat{\mu}_{o,s:t} - \widehat{\mu}_{o,r:t}|\right)^2 \geqslant \sum_{o=1}^{O} (\widehat{\mu}_{o,s:t} - \widehat{\mu}_{o,r:t})^2$
- (h) 根据方程 (26) 成立。

因此，我们有:

$$\{\forall\, t > c_\ell : \textbf{\textit{Restart}}(x_r, ..., x_t) = 0\} \Rightarrow \left\{\forall\, t > c_\ell, \forall s \in (r,t] : \frac{1}{2} \sum_{o=1}^{O} \frac{n_{r:s-1} n_{s:t}}{n_{r:t}} (\widehat{\mu}_{o,r:s-1} - \widehat{\mu}_{o,s:t})^2 \leqslant f_{r,s,t} - \log \eta_{r,s,t} \right\} \tag{30}$$

然后，使用概率算子，我们得到:

$$
\begin{aligned}
&\mathbb{P}_{\boldsymbol{\theta}^{(1)},\boldsymbol{\theta}^{(2)}}\Big\{\forall\, t > c_\ell : \textbf{\textit{Restart}}\,(x_r,...,x_t) = 0\Big\} \\
&\overset{(k)}{\leqslant} \mathbb{P}_{\boldsymbol{\theta}^{(1)},\boldsymbol{\theta}^{(2)}}\Big\{\forall\, t > c_\ell, \forall s \in (r,t] : \frac{1}{2}\sum_{o=1}^{O} \frac{n_{r:s-1}n_{s:t}}{n_{r:t}}\left(\widehat{\mu}_{o,r:s-1} - \widehat{\mu}_{o,s:t}\right)^2 \leqslant f_{r,s,t} - \log\eta_{r,s,t}\Big\} \\
&\overset{(l)}{\leqslant} \mathbb{P}_{\boldsymbol{\theta}^{(1)},\boldsymbol{\theta}^{(2)}}\Big\{\neg\Big\{\bigcap_{o\in\mathcal{O}} E^{(3)}_{o,r,t,\delta'}\Big\}\Big\} \\
&+ \mathbb{P}_{\boldsymbol{\theta}^{(1)},\boldsymbol{\theta}^{(2)}}\Big\{\forall\, t > c_\ell, \forall s \in (r,t] : \frac{1}{2}\sum_{o=1}^{O} \frac{n_{r:s-1}n_{s:t}}{n_{r:t}}\left(\widehat{\mu}_{o,r:s-1} - \widehat{\mu}_{o,s:t}\right)^2 \leqslant f_{r,s,t} - \log\eta_{r,s,t} \bigcap \Big\{\bigcap_{o\in\mathcal{O}} E^{(3)}_{o,r,t,\delta'}\Big\}\Big\} \\
&\overset{(m)}{\leqslant} \mathbb{P}_{\boldsymbol{\theta}^{(1)},\boldsymbol{\theta}^{(2)}}\Big\{\bigcup_{o\in\mathcal{O}} \neg E^{(3)}_{o,r,t,\delta'}\Big\} \\
&+ \mathbb{P}_{\boldsymbol{\theta}^{(1)},\boldsymbol{\theta}^{(2)}}\Big\{\forall\, t > c_\ell, \forall s \in (r,t] : \frac{1}{2}\sum_{o=1}^{O} \frac{n_{r:s-1}n_{s:t}}{n_{r:t}}\left(\widehat{\mu}_{o,r:s-1} - \widehat{\mu}_{o,s:t}\right)^2 \leqslant f_{r,s,t} - \log\eta_{r,s,t} \bigcap \Big\{\bigcap_{o\in\mathcal{O}} E^{(3)}_{o,r,t,\delta'}\Big\}\Big\} \\
&\overset{(n)}{\leqslant} \sum_{o\in\mathcal{O}} \mathbb{P}_{\boldsymbol{\theta}^{(1)},\boldsymbol{\theta}^{(2)}}\Big\{\neg E^{(3)}_{o,r,t,\delta'}\Big\} \\
&+ \mathbb{P}_{\boldsymbol{\theta}^{(1)},\boldsymbol{\theta}^{(2)}}\Big\{\forall\, t > c_\ell, \forall s \in (r,t] : \frac{1}{2}\sum_{o=1}^{O} \frac{n_{r:s-1}n_{s:t}}{n_{r:t}}\left(\widehat{\mu}_{o,r:s-1} - \widehat{\mu}_{o,s:t}\right)^2 \leqslant f_{r,s,t} - \log\eta_{r,s,t} \bigcap \Big\{\bigcap_{o\in\mathcal{O}} E^{(3)}_{o,r,t,\delta'}\Big\}\Big\} \\
&\overset{(o)}{\leqslant} O\delta' + \mathbb{P}_{\boldsymbol{\theta}^{(1)},\boldsymbol{\theta}^{(2)}}\Big\{\forall\, t > c_\ell, \forall s \in (r,t] : \frac{1}{2}\sum_{o=1}^{O} \frac{n_{r:s-1}n_{s:t}}{n_{r:t}}\left(\Delta_{o,r,s,t} - \mathcal{C}'_{r,s,t,\delta'}\right)^2 \leqslant f_{r,s,t} - \log\eta_{r,s,t}\Big\} \\
&\Leftrightarrow O\delta' + \mathbb{P}_{\boldsymbol{\theta}^{(1)},\boldsymbol{\theta}^{(2)}}\Big\{\forall\, t > c_\ell, \forall s \in (r,t] : \underbrace{1 - \frac{f_{r,s,t} - \log\eta_{r,s,t}}{\frac{n_{r,s-1}}{2} \times \sum_{o=1}^{O}\left(\Delta_{o,r,s,t} - \mathcal{C}'_{r,s,t,\delta'}\right)^2}}_{A_{r,s,t,\delta'}} \leqslant \frac{n_{r:s-1}}{n_{r:t}}\Big\}
\end{aligned}
$$

其中：

- (k) 成立，这得益于方程（30）中的蕴含关系。
- (l) 利用以下性质成立：$\mathbb{P}\Big\{A\Big\} \leqslant \mathbb{P}\Big\{\neg B\Big\} + \mathbb{P}\Big\{A \cap B\Big\}$ 其中 $B = \bigcap_{o\in\mathcal{O}} E_{(3)\,o,\ r,\ t,\ \delta'}$。
- (m) 得益于以下事实成立：$\neg\Big\{\bigcap_{o\in\mathcal{O}} E_{(3)\,o,\ r,\ t,\ \delta'}\Big\} = \bigcup_{o\in\mathcal{O}} \neg E_{(3)\,o,\ r,\ t,\ \delta'}$。
- (n) 利用事件 $\bigcup_{o\in\mathcal{O}} \neg E_{(3)\,o,\ r,\ t,\ \delta'}$ 上的并集界成立。
- 利用方程（27）可得 (o) 成立。

T因此，为了推导检测延迟，$A_{r,\ s,\ t,\ \delta'}$ 量应满足 t.

推导检测延迟所需的$A_{r,\ s,\ t,\ \delta'}$条件：

$$
\begin{cases}
A_{r,s,t,\delta'} > 0 & \Leftrightarrow \eta_{r,s,t} > \exp\left(-\frac{n_{r,s-1}}{2} \times \sum_{o=1}^{O}\left(\Delta_{o,r,s,t} - \mathcal{C}'_{r,s,t,\delta'}\right)^2\right)\exp\left(f_{r,s,t}\right), \\
A_{r,s,t,\delta'} < 1 & \Leftrightarrow \eta_{r,s,t} < \exp\left(f_{r,s,t}\right)
\end{cases}
\tag{31}
$$

检测变点的主要含义。请注意：

$$\begin{aligned}
&\left\{\exists t > c_\ell, s \in (r,t] : 1 + \frac{\log \eta_{r,s,t} - f_{r,s,t}}{\frac{n_{r,s-1}}{2} \times \sum_{o=1}^{O} \left(\Delta_{o,r,s,t} - \mathcal{C}'_{r,s,t,\delta}\right)^2} > \frac{n_{r:s-1}}{n_{r:t}}\right\} \\
&\Leftrightarrow \mathbb{P}_{\boldsymbol{\theta}^{(1)},\boldsymbol{\theta}^{(2)}}\left\{\forall\, t > c_\ell, \forall s \in (r,t] :\ 1 - \frac{f_{r,s,t} - \log \eta_{r,s,t}}{\frac{n_{r,s-1}}{2} \times \sum_{o=1}^{O} \left(\Delta_{o,r,s,t} - \mathcal{C}'_{r,s,t,\delta'}\right)^2} \leqslant \frac{n_{r:s-1}}{n_{r:t}}\right\} = 0 \\
&\Rightarrow \mathbb{P}_{\boldsymbol{\theta}^{(1)},\boldsymbol{\theta}^{(2)}}\left\{\forall\, t > c_\ell : \textbf{\textit{Restart}}\left(x_r, ..., x_t\right) = 0\right\} \leqslant O\delta' \Leftrightarrow \mathbb{P}_{\boldsymbol{\theta}^{(1)},\boldsymbol{\theta}^{(2)}}\left\{\exists t > c_\ell : \textbf{\textit{Restart}}\left(x_r, ..., x_t\right) = 1\right\} > 1 - O\delta'.
\end{aligned} \tag{32}$$

令 $\delta = O\delta'$ 且 $\mathcal{C}_{r,s,t,\delta} = \mathcal{C}'_{r,s,t,\frac{\delta}{O}}$。

然后，利用方程 (32) 的结果，变点 c_ℓ 在时间 t (以至少 $1-\delta$ 的概率)被检测到，若存在某个 $s \in (r,t]$，使得：

$$1 + \frac{\log \eta_{r,s,t} - f_{r,s,t}}{\frac{n_{r,s-1}}{2} \times \sum_{o=1}^{O} \left(\Delta_{o,r,s,t} - \mathcal{C}_{r,s,t,\delta}\right)^2} > \frac{n_{r:s-1}}{n_{r:t}}. \tag{33}$$

令 $\boldsymbol{\Delta} = (\Delta_1, ..., \Delta_O)$ 表示变点间隔的向量。

步骤 3：检测延迟的非渐近表达式 $\mathfrak{D}_{\boldsymbol{\Delta},\, r,\, c_\ell}$ 为了构造检测延迟，我们需要确保存在 $s \in (r,\ t]$ 使得方程 (33) 得以满足。特别地，对于 $s = c_\ell$，方程 (33) 得以满足。由此，检测变点的条件 c_ℓ 可写为如下形式

$$1 + \frac{\log \eta_{r,c_\ell,t} - f_{r,c_\ell,t}}{\frac{n_{r,s-1}}{2} \times \sum_{o=1}^{O} \left(\Delta_o - \mathcal{C}_{r,s,t,\delta}\right)^2} > \frac{n_{r:c_\ell-1}}{n_{r:t}}. \tag{34}$$

为了构造延迟，我们应引入以下变量：$d = t - c_\ell + 1 = n_{c_\ell:\, t} \in \mathbb{N}^\star$。
因此，由方程 (34)，我们得到：

$$\begin{aligned}
&\left\{1 + \frac{\log \eta_{r,c_\ell,d+c_\ell-1} - f_{r,c_\ell,d+c_\ell-1}}{\frac{n_{r,s-1}}{2} \times \sum_{o=1}^{O} \left(\Delta_o - \mathcal{C}_{r,s,t,\delta}\right)^2} > \frac{n_{r:c_\ell-1}}{n_{r:c_\ell-1} + d}\right\}. \\
&\Leftrightarrow \left\{d > \frac{2}{\sum_{o=1}^{O} \left(\Delta_o - \mathcal{C}_{r,c_\ell,d+c_\ell-1,\delta}\right)^2} \times \frac{-\log \eta_{r,c_\ell,d+c_\ell-1} + f_{r,c_\ell,d+c_\ell-1}}{1 + \frac{2\left(\log \eta_{r,c_\ell,d+c_\ell-1} - f_{r,c_\ell,d+c_\ell-1}\right)}{n_{r,c_\ell-1} \times \sum_{o=1}^{O} \left(\Delta_o - \mathcal{C}_{r,c_\ell,d+c_\ell-1,\delta}\right)^2}}\right\}.
\end{aligned}$$

最后，变点 c_ℓ 被检测到（概率至少为 $1-\delta$），且延迟不超过 $\mathfrak{D}_{\boldsymbol{\Delta},\, r,\, c_\ell}$，满足：

$$\mathfrak{D}_{\boldsymbol{\Delta},r,c_\ell} = \min\left\{d \in \mathbb{N}^\star : d > \frac{2}{\sum_{o=1}^{O} \left(\Delta_o - \mathcal{C}_{r,c_\ell,d+c_\ell-1,\delta}\right)^2} \times \frac{-\log \eta_{r,c_\ell,d+c_\ell-1} + f_{r,c_\ell,d+c_\ell-1}}{1 + \frac{2\left(\log \eta_{r,c_\ell,d+c_\ell-1} - f_{r,c_\ell,d+c_\ell-1}\right)}{n_{r,c_\ell-1} \times \sum_{o=1}^{O} \left(\Delta_o - \mathcal{C}_{r,c_\ell,d+c_\ell-1,\delta}\right)^2}}\right\}.$$

□

E R-BOCPD 配备 UCRL2 的分析

我们考虑第 3 节中定义的遗憾表述。鉴于 R-BOCPD 算法所提供理论保证的性质，我们采用按变点进行分解的方式，以便分析切换*MDP* 问题 $\mathbf{M} = \{\mathbb{S} = \{M_0, .., M_{K_T-1}\}, \mathcal{C} = \{c_0, .., c_{K_T}\}\}$ 作为一系列平稳马尔可夫决策过程 M_ℓ在时间实例 $t \in [c_\ell, c_{\ell+1})$上的演变。可将其表述如下：

$$\begin{aligned}\mathfrak{R}\left(\mathbf{M}, \texttt{R-BOCPD-UCRL2}, o, T\right) &= \sum_{t=1}^{T}\left(\rho^{\star}_{\mathbf{M}_\ell}(t) - \mathbb{E}\left[r_t\right]\right) \\ &= \sum_{\ell=0}^{K_T-1}\sum_{t=c_\ell}^{c_{\ell+1}-1}\left(\rho^{\star}_{\mathbf{M}_\ell}(t) - \mathbb{E}\left[r_t\right]\right)\end{aligned}$$

记 c_ℓ 为变化发生的时刻 ℓ ，记 $t = c_\ell$ 为从 c_ℓ 开始、但不包括 $c_{\ell+1}$的时间区间，即 $t \in [c_\ell, c_\ell + 1)$，并定义 r_t 为 UCRL2 在时刻UCRL2 获得的随机奖励 t，其初始状态为 o。同样，只有利用初始状态对应的平稳时段内奖励总和/遗憾的独立性，才能进行这种分解，正如Puterman (2014 年)。

现在，记 d_ℓ 为 R-BOCPD 在给定区间 $[c_\ell, c_{\ell+1})$ 内实现的检测延迟。因此，自然可将其分解为平稳时段 $[c_\ell + d_\ell, c_{\ell+1})$ 和检测阶段 $[c_\ell, c_\ell + d_\ell)$。

E.1 检测阶段 $[c_\ell, c_\ell + d_\ell)$

在检测阶段，我们假设算法取最坏情况下的遗憾值 1，因为 r_t 是根据 $[0, 1]$中的某个未知分布采样得到的。为了最小化总遗憾，我们依赖 R-BOPCD 的最小检测延迟，将其记为

$$\mathfrak{D}_{\boldsymbol{\Delta}_\ell, r, c_\ell} = \min\left\{ d_\ell \in \mathbb{N}^\star : d_\ell > \frac{2}{\sum_{o=1}^{O}\left(\Delta_o - \mathcal{C}_{r, c_\ell, d_\ell + c_\ell - 1, \delta}\right)^2} \times \frac{-\log \eta_{r, c_\ell, d_\ell + c_\ell - 1} + f_{r, c_\ell, d_\ell + c_\ell - 1}}{1 + \frac{2\left(\log \eta_{r, c_\ell, d_\ell + c_\ell - 1} - f_{r, c_\ell, d_\ell + c_\ell - 1}\right)}{n_{r, c_\ell - 1} \times \sum_{o=1}^{O}\left(\Delta_o - \mathcal{C}_{r, c_\ell, d_\ell + c_\ell - 1, \delta}\right)^2}} \right\} \tag{35}$$

with

$$\boldsymbol{\Delta}_\ell = (\Delta_{1,\ell}, ..., \Delta_{O,\ell}) \quad \text{where} \quad \Delta_{o,\ell} = \left|\theta_o^{(\ell-1)} - \theta_o^{(\ell)}\right|$$

对于从某个时间点 $r < c_\ell$开始的输入流，当 $\delta_{1_\ell} \in$取值于 $(0, 1)$ 时，该结论以至少 $1 - \delta_{1_\ell}$的概率成立。这使我们可以如下写出从某个状态 o_{c_ℓ} 开始的 $t \in [c_\ell, c_\ell + d_\ell)$ 的遗憾

$$\begin{aligned}\mathfrak{R}\left(M_\ell, \texttt{R-BOCPD-UCRL2}, o_{c_\ell}, [c_\ell, c_\ell + d_\ell)\right) &= \sum_{t=c_\ell}^{c_\ell + d_\ell - 1}\left(\rho^{\star}_{\mathbf{M}_\ell}(t) - \mathbb{E}\left[r_t\right]\right) \\ &= (c_\ell + d_\ell - 1 - c_\ell + 1) \cdot 1 \\ &= d_\ell \\ &= \mathfrak{D}_{\boldsymbol{\Delta}_\ell, c_{\ell-1} + d_{\ell-1}, c_\ell}\end{aligned}$$

其中， $c_{\ell-1} + d_{\ell-1}$ 对应于变点 $c_{\ell-1}$ 之后延迟最长的重启时间，且其发生概率至少为 $1 - \delta_{1_{\ell-1}} - \delta_{2_{\ell-1}} + \delta_{1_{\ell-1}}\delta_{2_{\ell-1}}$； $\delta_{1_{\ell-1}}$ 对应于 R-BOCPD 的延迟超过$\mathfrak{D}_{\boldsymbol{\Delta}_{\ell-1},\ c_{\ell-2} + d_{\ell-2},\ c_{\ell-1}}$ 的概率（针对变点 $c_{\ell-1}$ ），而 $\delta_{2_{\ell-1}}$ 对应于平稳时段 $[c_{\ell-1} + d_{\ell-1}, c_\ell$上的最坏情况虚警概率）。

E.2 检测后阶段 $[c_\ell + d_\ell, c_{\ell+1})$ 与回合遗憾

基于第 3 节中关于变点生成过程的假设，我们现在分析阶段 $[c_\ell + d_\ell, c_{\ell+1})$ 中的遗憾，假设 R-BOCPD-UCRL2 以至少 $t = c_\ell + d_\ell$ 的概率恰好在时间 $1 - \delta_{1_\ell}$ 重启 $\delta_{1_\ell} \in (0,1)$。现在，为了进行类似于 Auer 等人 (2008b) 的分析，我们需要确保在 $[c_\ell + d_\ell, c_{\ell+1})$期间没有发生重启/虚警，也就是说，该时段是平稳的。鉴于 R-BOCPD 保证虚警概率有界，我们对平稳时段内重启 UCRL2 这一事件进行分解。具体而言，我们采用 UCRL2 的平稳奖励总和集中性质，如 Auer 等人 (2008b) 中所述，并结合 δ-界保证，即 R-BOCPD 的虚警概率。

为在应用 UCRL2 时对奖励总和使用集中性论证，我们注意到，在时间 t，奖励 r_{t+1} 仅通过外生过程 $\mathcal{E}$依赖于奖励 r_t（以及滤过历史（o_1、a_1、r_1, ……、o_t、a_t、r_t，下文沿用该记号））。因此，对于所有 t，给定 $\mathcal{E}$ 时，r_{t+1} 和 r_t 相互独立，即 $r_{t+1} \perp\!\!\!\perp r_t | \mathcal{E}$。根据 Hoeffding 不等式，这使我们能够写出，对于 $t \in [c_\ell + d_\ell, c_{\ell+1})$ 和 $\delta_\ell \in$（0、1)

$$
\begin{aligned}
&\mathbb{P}\Bigg[\underbrace{\sum_{t=c_\ell+d_\ell}^{c_{\ell+1}-1} r_t \leqslant \sum_{o,a} N_\ell(o,a)\bar{r}_\ell(s,a) - \sqrt{\frac{5}{8}\left(c_{\ell+1}-(c_\ell+d_\ell)\right)\log\left(\frac{8\left(c_{\ell+1}-(c_\ell+d_\ell)\right)}{\delta_\ell}\right)}\ \Bigg|\ (N_\ell(o,a))_{o,a}, \mathcal{E}}_{E_\ell^{(4)}}\Bigg] \\
&\overset{(a)}{\leqslant} \mathbb{P}_{\boldsymbol{\theta}}\left[E_\ell^{(4)} \,\middle|\, \forall t \in [c_\ell + d_\ell + 1, c_{\ell+1}) : \textbf{\textit{Restart}}\left(o_{c_\ell+d_\ell}, ..., o_t\right) = 0\right] \\
&+ \mathbb{P}_{\boldsymbol{\theta}}\left[\exists\, t \in [c_\ell + d_\ell + 1, c_{\ell+1}) : \textbf{\textit{Restart}}\left(o_{c_\ell+d_\ell}, ..., o_t\right) = 1\right] \\
&\overset{(b)}{\leqslant} \left(\frac{\delta_\ell}{8\left(c_{\ell+1}-(c_\ell+d_\ell)\right)}\right)^{5/4} + \delta_{2_\ell} \\
&< \frac{\delta_\ell}{12\left(c_{\ell+1}-(c_\ell+d_\ell)\right)^{\frac{5}{4}}} + \delta_{2_\ell}
\end{aligned}
$$

其中，(a) 源自不等式 $\mathbb{P}(A) \leqslant \mathbb{P}(A|B) + \mathbb{P}(\neg B)$，其中 $A = E_\ell^{(4)}$ 和 $B = \Big\{\forall t \in (c_\ell + d_\ell + 1、 c_{\ell+1})$: 重启 $(o_{c_\ell+d_\ell}$, ……、 $o_t) = 0\Big\}$；(b) 源自原始 UCRL2 在 r_ℓ（o、a）$= \frac{1}{N_\ell}\sum_{t=c_\ell+d_\ell}^{c_{\ell+1}-1} r_t \cdot \mathbb{I}\{o_t = o, a_t = a\}$下的 Hoeffding 不等式，同时还结合了 R-BOCPD 关于误报率的保证。

因此，我们可以将从某个状态开始的 ℓ^{th} 变点 o_{c_ℓ} 的检测后遗憾表示如下

$$
\begin{aligned}
&\mathfrak{R}\left(M_\ell, \text{R-BOCPD-UCRL2}, o_{c_\ell}, [c_\ell + d_\ell, c_{\ell+1})\right) = (c_{\ell+1} - c_\ell - d_\ell)\rho^\star_{\mathbf{M}_\ell} - \sum_{t=c_\ell+d_\ell}^{c_{\ell+1}-1} r_t \\
&< (c_{\ell+1} - c_\ell - d_\ell)\rho^\star_{\mathbf{M}_\ell} - \sum_{o,a} N_\ell(o,a)\bar{r}_\ell(o,a) + \sqrt{\frac{5}{8}(c_{\ell+1} - c_\ell - d_\ell)\log\left(\frac{8\left(c_{\ell+1}-c_\ell-d_\ell\right)}{\delta_\ell}\right)}
\end{aligned}
$$

其发生的概率至少为 $1 - \frac{\delta_\ell}{12\left(c_{\ell+1}-c_\ell-d_\ell\right)^{\frac{5}{4}}} - \delta_{2_\ell}$。与 Auer 等人（2008b）一样，我们按照回合数进行分解，并将变化区间 $[c_\ell + d_\ell, c_{\ell+1}$ 中的回合数记为 m_ℓ。因此，我们可以写成$\sum_{k=1}^{m_\ell} \nu_k = N_\ell(o,a)$ 和 $\sum_{o,a} N_\ell(o, a) = c_{\ell+1} - c_\ell - d_\ell$，其中 $\nu_k(o,a)$ 表示状态-动作对 (o,a) 在回合 k 中的最终计数。
因此，定义 $\mathfrak{R}_k(M_\ell$, R-BOCPD-UCRL2, $o_{c_\ell}, [c_\ell + d_\ell, c_{\ell+1})) := \sum_{o,a} \nu_k(o,a)\left(\rho^\star_{M_\ell} - r_\ell(o,a)\right)$，我们

可以写成

$$\mathfrak{R}\left(M_\ell, \texttt{R-BOCPD-UCRL2}, o_{c_\ell}, [c_\ell + d_\ell, c_{\ell+1})\right)$$
$$\leqslant \sum_{k=1}^{m_\ell} \mathfrak{R}_k\left(M_\ell, \texttt{R-BOCPD-UCRL2}, o_{c_\ell}, [c_\ell + d_\ell, c_{\ell+1})\right) + \sqrt{\frac{5}{8}\left(c_{\ell+1} - c_\ell - d_\ell\right)\log\left(\frac{8\left(c_{\ell+1} - c_\ell + d_\ell\right)}{\delta_\ell}\right)}$$

其概率至少为 $1 - \frac{\delta_\ell}{12\left(c_{\ell+1} - c_\ell - d_\ell\right)^{\frac{5}{4}}} - \delta_{2_\ell}$。

现在，遵循Auer 等人 (2008b) 对标准 UCRL2的分析，我们可以推导出R-BOCPD-UCRL2 在检测后时期 $[c_\ell + d_\ell, c_{\ell+1})$中，针对 $c_{\ell+1} - c_\ell - d_\ell > 1$的最终遗憾界，如下所示

$$\mathfrak{R}\left(M_\ell, \texttt{R-BOCPD-UCRL2}, o_{c_\ell}, [c_\ell + d_\ell, c_{\ell+1})\right) \leqslant 34 D_\ell O\sqrt{A\left(c_{\ell+1} - c_\ell - d_\ell\right)\log\left(\frac{c_{\ell+1} - c_\ell - d_\ell}{\delta_\ell}\right)}$$

其成立的概率至少为 $1 - \frac{\delta_\ell}{4\left(c_{\ell+1} - c_\ell - d_\ell\right)^{\frac{5}{4}}} - \delta_{2_\ell}$，其中 D_ℓ 是马尔可夫决策过程 M_ℓ 的直径，定义见 3。另请注意，$d_0 = \mathfrak{D}_{\boldsymbol{\Delta}_{,.,c_0}} = 0$作为时间实例 c_0，定义了学习的开始时间。

E.3 总遗憾界

总结前两个步骤并对各变化时期求和，可得

$$\sum_{\ell=0}^{K_T-1} \mathfrak{R}\left(M_\ell, \texttt{R-BOCPD-UCRL2}, o_{c_\ell}, [c_\ell, c_{\ell+1})\right)$$
$$\leqslant 34 O\sqrt{A}\sum_{\ell=0}^{K_T-1} D_\ell\sqrt{\left(c_{\ell+1} - c_\ell - d_\ell\right)\log\left(\frac{c_{\ell+1} - c_\ell - d_\ell}{\delta_\ell}\right)} + \sum_{\ell=0}^{K_T-1} \mathfrak{D}_{\boldsymbol{\Delta}_{\ell+1}, c_\ell + d_\ell, c_{\ell+1}} \tag{36}$$

其概率至少为 $1 - \sum_{\ell=0}^{K_T-1}\left(\frac{\delta_\ell}{4\left(c_{\ell+1} - c_\ell - d_\ell\right)^{\frac{5}{4}}} + \delta_{2_\ell}\right) - \sum_{\ell=1}^{K_T}\left(\delta_{1_{\ell-1}} + \delta_{2_{\ell-1}} - \delta_{1_{\ell-1}}\delta_{2_{\ell-1}}\right)$。现在，我们通过给出后者概率的一个界来完成证明。不失一般性，我们将置信参数固定为 $\delta_\ell = \delta_{1_\ell} = \delta_{2_\ell} := \frac{\delta}{8K_T}, \forall \ell$。因此，我们可以写出：

$$\begin{aligned}
1 - \sum_{\ell=0}^{K_T-1}\left(\frac{\delta_\ell}{4\left(c_{\ell+1} - c_\ell - d_\ell\right)^{\frac{5}{4}}} + \delta_{2_\ell}\right) - \sum_{\ell=1}^{K_T}\left(\delta_{1_{\ell-1}} + \delta_{2_{\ell-1}} - \delta_{1_{\ell-1}}\delta_{2_{\ell-1}}\right) & \\
&> 1 - \sum_{\ell=0}^{K_T-1}\left(\frac{\delta_\ell}{4\left(c_{\ell+1} - c_\ell - d_\ell\right)^{\frac{5}{4}}} + \delta_{2_\ell}\right) - \sum_{\ell=1}^{K_T}\left(\delta_{1_{\ell-1}} + \delta_{2_{\ell-1}}\right) \\
&> 1 - \frac{3\delta}{8} - \frac{\delta}{4K_T}\sum_{\ell=0}^{K_T-1}\frac{1}{\left(c_{\ell+1} - c_\ell - d_\ell\right)^{\frac{5}{4}}} \\
&> 1 - \frac{3\delta}{8} - \frac{\delta}{4K_T}\sum_{\ell=0}^{K_T-1} 1 \qquad \text{as} \quad c_{\ell+1} - c_\ell - d_\ell > 1 \\
&> 1 - \frac{3\delta}{8} - \frac{\delta}{4} = 1 - \frac{5\delta}{8} \\
&> 1 - \delta
\end{aligned}$$

现在，定义 $D := \max_{\ell} \quad D_\ell$，推导相应的遗憾归结为

$$\begin{aligned}
\mathfrak{R}\left(\mathbf{M}, \texttt{R-BOCPD-UCRL2}, o_{c_0}, T\right) &= \sum_{\ell=0}^{K_T-1} \mathfrak{R}\left(M_\ell, \texttt{R-BOCPD-UCRL2}, o_{c_\ell}, [c_\ell, c_{\ell+1})\right) \\
&\leqslant 34O\sqrt{A} \sum_{\ell=0}^{K_T-1} D_\ell \sqrt{(c_{\ell+1} - c_\ell - d_\ell) \log\left(\frac{c_{\ell+1} - c_\ell - d_\ell}{\delta_\ell}\right)} + \sum_{\ell=0}^{K_T-1} \mathfrak{D}_{\boldsymbol{\Delta}_{\ell+1}, c_\ell + d_\ell, c_{\ell+1}} \\
&= 34DO\sqrt{A} \sum_{\ell=0}^{K_T-1} \sqrt{\frac{1}{K_T} K_T \left(c_{\ell+1} - c_\ell - d_\ell\right) \log\left(\frac{K_T(c_{\ell+1} - c_\ell - d_\ell)}{\delta}\right)} + \sum_{\ell=0}^{K_T-1} \mathfrak{D}_{\boldsymbol{\Delta}_{\ell+1}, c_\ell + d_\ell, c_{\ell+1}} \\
&\leqslant 34DO\sqrt{A} \sum_{\ell=0}^{K_T-1} \sqrt{\frac{T}{K_T} \log\left(\frac{T}{\delta}\right)} + \sum_{\ell=0}^{K_T-1} \mathfrak{D}_{\boldsymbol{\Delta}_{\ell+1}, c_\ell + d_\ell, c_{\ell+1}} \\
&= 34DO\sqrt{AT K_T \log\left(\frac{T}{\delta}\right)} + \sum_{\ell=0}^{K_T-1} \mathfrak{D}_{\boldsymbol{\Delta}_{\ell+1}, c_\ell + d_\ell, c_{\ell+1}}
\end{aligned} \tag{37}$$

其成立的概率至少为 $1-\delta$。定理4.12的证明至此完成。

E.4 渐近遗憾界

基于上一节的内容以及方程（10）中的渐近检测延迟，我们将渐近遗憾界写作如下

$$\begin{aligned}
\mathfrak{R}\left(\mathbf{M}, \texttt{R-BOCPD-UCRL2}, o_{c_0}, T\right) &\leqslant 34DO\sqrt{AT K_T \log\left(\frac{T}{\delta}\right)} + \sum_{\ell=0}^{K_T-1} \mathfrak{D}_{\boldsymbol{\Delta}_{\ell+1}, c_\ell + d_\ell, c_{\ell+1}} \\
&\leqslant 34DO\sqrt{AT K_T \log\left(\frac{T}{\delta}\right)} + \sum_{\ell=0}^{K_T-1} \lim_{c_{\ell+1} - c_\ell - d_\ell \to \infty} \mathfrak{D}_{\boldsymbol{\Delta}_{\ell+1}, c_\ell + d_\ell, c_{\ell+1}} \\
&= 34DO\sqrt{AT K_T \log\left(\frac{T}{\delta}\right)} + \sum_{\ell=0}^{K_T-1} \mathcal{O}\left(\frac{\log \frac{K_T}{\delta}}{\mathbf{KL}\left(\boldsymbol{\theta}^{(\ell+1)} \,\middle\|\, \boldsymbol{\theta}^{(\ell)}\right)}\right) \\
&\leqslant 34DO\sqrt{AT K_T \log\left(\frac{T}{\delta}\right)} + \mathcal{O}\left(\frac{K_T \log \frac{K_T}{\delta}}{\min_{\ell} \mathbf{KL}\left(\boldsymbol{\theta}^{(\ell+1)} \,\middle\|\, \boldsymbol{\theta}^{(\ell)}\right)}\right)
\end{aligned}$$

同样地，假设误报率为 $1-\delta$，则其成立的概率至少为0。推论4.13的推导至此完成。

F 实验设置与讨论

鉴于 R-BOCPD-UCRL2 对状态转移分布和奖励变化的普适性，适合用于基准测试、并与当前最先进水平进行比较的环境，是一个在未知时间点上状态转移分布和奖励会发生突变的合成环境。首先，随机选取状态空间和动作空间的大小。随后，从状态集合 O 上的多项分布中抽取状态转移概率，并从 $[0,1]$，中随机抽取奖励；在此过程中，我们可以控制生成过程中的变化幅度，从而模拟状态转移分布和奖励的大幅变化及相对小幅的变化。现在，固定一组变点，并确保相邻变点之间有足够的时间间隔，在每个变点之后，将该过程重复至时间范围 $T = 50000$ 结束。我们对每个状态-动作对考虑 100 次实现值，并关注其平均累积奖励。

F.1 超参数的选择

现在，我们列出第 5 节中基于滑动窗口的算法所采用的超参数。

- **滑动窗口 UCRL2** (SWUCRL2，Gajane 等人 (2018 年)): 窗口大小按照Ga-₂jane 等人 (2018 年) $W^\star = \left(\frac{16.53}{K_T} TDO\sqrt{A \log \frac{T}{\delta}}\right)^3$。直径 D通过在针对每种状态空间和动作空间大小组合的一大范围合适取值上进行超参数搜索来估计。选择使累积奖励最大的直径。

- **带置信度扩展的滑动窗口 UCRL2** (SWUCRL2-CW，Cheung 等人(2020 年)): 同样，窗口大小和扩展因子根据Cheung 等人 (2020 年) $W^\star := 3O^{\frac{2}{3}} A^{\frac{1}{2}} T^{\frac{1}{2}} / (B_r + B_p + 1)^{\frac{1}{2}}$和 $\eta^\star := \sqrt{(B_p+1)W^\star/T}$分别最优选择。这里， B_p和 B_r均预先按照全变差意义计算。在现实的强化学习设置中， T也无法预先得知，但我们仍选择 T来初始化SWUCRL2。

F.2 讨论

现在，除了第 5 节中提供的较为一般的性能评估外，我们还将更详细地说明每种算法的运作方式，如下所示

- 滑动窗口 UCRL2（SWUCRL2，Gajane 等人（2018））：滑动窗口方法的一个关键参数是构成切换马尔可夫决策过程问题 M 的每个马尔可夫决策过程 D_ℓ 的直径 M_ℓ，它用于量化在 M_ℓ 所指定的设置下学习的难度。直径 D_ℓ（定义见第 3 节）是一个无法直接从马尔可夫决策过程中获取、但对 SWUCRL2 按照所述方式运行至关重要的参数。即使考虑 $D = \max_\ell D_\ell$，SWUCRL2 仍依赖超参数搜索来估计整体最优直径 D，这在实践中相当受限。我们还指出，当 D 选择不理想时，SWUCRL2 的表现较差。现在，考虑到最优窗口大小选择 $W^\star$ is of$\tilde{O}(A^{\frac{1}{3}} O^{\frac{2}{3}} D^{\frac{2}{3}} T^{\frac{2}{3}} K_T{}^{-\frac{1}{2}})$，即使对于相当小的马尔可夫决策过程，SWUCRL2 也需要跟踪大量观测。

- **带置信度扩展的滑动窗口 UCRL2** (SWUCRL2-CW，Cheung 等人 (2020 年))：与 SWUCRL2 不同，它不依赖于一个以不可知方式选取的参数，SWUCRL2-CW 仍然依赖于预先设定的状态转移分布和奖励变化预算，而在现实环境中，这些预算是未知的。其变体强化学习之上的多臂老虎机（BORL）无需假设已知 B_r 和 B_p，实际表现优于 SWUCRL2-CW 。
- **重启 UCRL2** (Restarted-UCRL2，Auer 等人(2008a))：尽管其表现优于滑动窗口方法，但它需要大量重启 $T^{\frac{1}{3}} K_T{}^{\frac{2}{3}}$，这对于大规模问题而言代价过高。此外，由于重启频率会随时间呈多项式下降（变化次数固定时），当变化在学习后期频繁发生且时间范围相当长时，其性能不可避免地会下降。

- **原始 UCRL2**（UCRL2，Auer 等人 （2008a）：这里将其视为基线。

F.2.1 扩展到真实环境

我们还特别关注真实的马尔可夫决策过程设置，例如 *RiverSwim* （Strehl 和 Littman （2008）），*MachineRe-placement* 和网格世界 等。在这里，模拟环境的真实变化之所以困难，是因为我们无法访问底层的状态转移分布和奖励。虽然可以改变一些通常随机选取的环境参数，但在撰写本文时，我们尚不清楚如何控制这些变量的变化过程，以及如何将其与底层状态转移分布和奖励中相当隐性的变化联系起来。