RESTARTED BAYESIAN ONLINE CHANGE-POINT DETECTION FOR NON-STATIONARY MARKOV DECISION PROCESSES

Reda Alami[*]
Technology Innovation Institute
Abu Dhabi, United Arab Emirates
reda.alami@tii.ae

Mohammed Mahfoud[*]
Technical University of Munich
Munich, Germany
mo.mahfoud@tum.de

Eric Moulines
Ecole Polytechnique
Paris, France
eric.moulines@polytechnique.edu

ABSTRACT

We consider the problem of learning in a non-stationary reinforcement learning (RL) environment, where the setting can be fully described by a piecewise stationary discrete-time Markov decision process (MDP). We introduce a variant of the Restarted Bayesian Online Change-Point Detection algorithm (R-BOCPD) that operates on input streams originating from the more general multinomial distribution and provides near-optimal theoretical guarantees in terms of false-alarm rate and detection delay. Based on this, we propose an improved version of the UCRL2 algorithm for MDPs with state transition kernel sampled from a multinomial distribution, which we call R-BOCPD-UCRL2. We perform a finite-time performance analysis and show that R-BOCPD-UCRL2 enjoys a favorable regret bound of $\mathcal{O}\left(DO\sqrt{ATK_T \log\left(\frac{T}{\delta}\right)} + \frac{K_T \log \frac{K_T}{\delta}}{\min\limits_{\ell} \mathbf{KL}\left(\boldsymbol{\theta}^{(\ell+1)} \,\|\, \boldsymbol{\theta}^{(\ell)}\right)}\right)$, where D is the largest MDP diameter from the set of MDPs defining the piecewise stationary MDP setting, O is the finite number of states (constant over all changes), A is the finite number of actions (constant over all changes), K_T is the number of change points up to horizon T, and $\boldsymbol{\theta}^{(\ell)}$ is the transition kernel during the interval $[c_\ell, c_{\ell+1})$, which we assume to be multinomially distributed over the set of states O. Interestingly, the performance bound does not directly scale with the variation in MDP state transition distributions and rewards, ie. can also model abrupt changes. In practice, R-BOCPD-UCRL2 outperforms the state-of-the-art in a variety of scenarios in synthetic environments.

1 INTRODUCTION

In a typical sequential online decision making setting, a decision maker, which we refer to as *agent*, interacts with its environment by first observing its current state. It then select an action from the set of possible actions in its state, moves to another state determined by the stochastic process that generates its state transition distribution, and receives a random reward that quantifies the quality of its current decision/action relative to the set of optimal actions it could have taken in its previous state. Through this continuous interaction, the agent attempts to learn an optimal decision-making scheme or *policy* to maximize its cumulative rewards. This accurately describes the reinforcement learning (RL) problem, which has proven useful in modeling a variety of important problems in many domains.

In classical RL, it is often assumed that state transition distributions and rewards are generated by a stochastic process that is stationary throughout the learning process. However, this assumption is quite restrictive in the context of real online learning environments. Therefore, it is necessary to define a new variant of RL, commonly referred to as non-stationary RL. The study of the latter is supported by a variety of applications in consumer decision modeling (Xu & Yun (2020)), service provider adaptation to customers, and pricing (Taylor (2018); Kanoria & Qian (2019); Bimpikis et al. (2019); Gurvich et al. (2019)), wireless communication networks (Zhou & Bambos (2015); Zhou et al. (2016)), epidemic networks and control (Nowzari et al. (2016); Kiss et al. (2017)), inventory management (Agrawal & Jia (2019); Huh & Rusmevichientong (2009)), non-stationary multi armed bandits (Alami et al. (2017); Alami & Azizi

[*] Equal Contribution.

(2020); Alami (2023; 2018); Garivier & Moulines (2011)) to name a few. In the aforementioned application areas, non-stationarity is often due to abrupt changes that can have drastic effects in highly sensitive environments.

While accurately modeling latent change in a non-stationary RL environment is generally very difficult, we are particularly interested in exploring a variant where non-stationarity can be fully modeled by a piecewise stationary discrete-time Markov decision process (MDP). More specifically, we assume a situation in which the associated MDP state transition distributions and rewards can change arbitrarily at unknown predefined time points, which we refer to as *change-points*. In this way, we can accurately and dynamically handle the abruptly changing environment mentioned above. An illustration of the problem can be found in Figure 1.

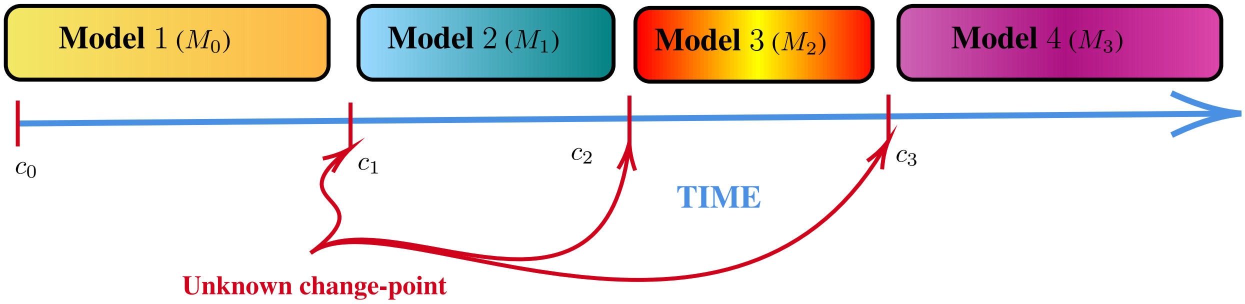

Figure 1: Piecewise Stationary Discrete-time MDP. Starting from change-point c_ℓ, the environment is modeled by MDP M_ℓ.

Key Contributions. We list a summary of our main contributions as follows

1. We extend the Restarted Bayesian Online Change-Point Detection algorithm R−BOCPD to the more general setting where the online observation stream is generated according to a multinomial distribution, and provide (near) optimal theoretical guarantees in terms of false alarm rate and detection delay control.
2. We propose an improved version of UCRL2 that incorporates R−BOCPD as a means of detecting changes in the learning environment and provides near-optimal regret bounds that rely only on a small set of past observations and allow for changes of arbitrary magnitude in both state transition distributions and rewards.
3. We demonstrate the results experimentally and compare the performance of the algorithm' with that of the state of the art.

2 Related work

To give a general overview of past works dealing with a switching RL (MDP RL), we first present some of the main results obtained in the field of stationary vanilla MDP, and then describe some recent key contributions and models in the field of non-stationary RL. Finally, we also present some results (mostly algorithmic in nature) mainly from the time series area.

2.1 RL in Stationary MDPs

We restrict ourselves to contributions that are directly relevant to our problem. In particular, we distinguish the discounted (Sidford et al. (2018a;b); Wang (2020)) and non-discounted reward cases (Auer et al. (2008b); Azar et al. (2017); Dann et al. (2017); Jin et al. (2018); Zanette & Brunskill (2019)). In the former case, Sidford et al. (2018a;b); Wang (2020) proposed near-optimal algorithms with respect to sample complexity. For the latter case, where the regret of an algorithm $\mathcal{A}$ is defined as the difference between the reward obtained by $\mathcal{A}$ and that of an optimal algorithm, numerous regret bounds have been proposed. Auer et al. (2008b) first established a minimax lower bound $\Omega(\sqrt{DOAT})$, where D is the MDP diameter as defined in Section 3, and the 'state transition distributions and rewards of the MDP are assumed to be time-invariant over the time horizon T under consideration (O denotes the number of states and A the number of actions). Based on upper confidence bounds, Auer et al. (2008b) has also proposed UCRL2, an algorithm that achieves a regret bound of $\tilde{\mathcal{O}}(DO\sqrt{AT})$. Variants of UCRL2 with improved regret bounds were also proposed later, but are omitted in this manuscript due to their lack of efficiency in practice, e.g., computational inefficiency as in Zhang & Ji (2019), despite their minimax optimality.

2.2 Non-stationary RL, as modeled by MDPs

In a rather naive way, Auer et al. (2008a) already considers the case where a predefined *known-to-agent* number of changes ℓ should occur in the environment. Based on this, `Restarted-UCRL2` periodically restarts `UCRL2`. More recently, a number of papers have developed Gajane et al. (2018); Ortner et al. (2020); Cheung et al. (2020) Algorithms for non-stationary RL in the tabular setting. Such algorithms assume that the MDP is constant over all episodes up to the current one, say k, and estimate the state transition distributions and rewards based on the data up to $k-1$. If a change occurs between episodes $k-1$ and k, which is generally possible, the estimator *biased* and Ortner et al. (2020) show that the algorithms suffer a linear regret that scales with the norm of the bias. As a remedy, Gajane et al. (2018); Cheung et al. (2020) proposes a sliding-window approach in which estimators favor recently observed transitions and penalize older ones. As shown in Cheung et al. (2020), the Gajane et al. (2018) approach leads to sub-optimal *regret* bounds.Cheung et al. (2020) proposes a *confidence widening* variant to Gajane et al. (2018) in which the regret bounds for *smoothly-changing* MDPs are more favorable and thus only handle the case where the state transition distributions of the environment are set to change up to a *variation budget* assumed at initialization. Ortner et al. (2020) periodically restarts the algorithm and discards past data at each restart, but requires much more information about the variations in state transition distributions and rewards between restarts than Cheung et al. (2020) to achieve its *dynamic regret* [*] bound. Although not directly related to the setting we consider, we would also like to highlight the contributions of Yu & Mannor (2009); Neu et al. (2010); Arora et al. (2012); Dick et al. (2014); Jin et al. (2020); Rivera Cardoso et al. (2019), where state transition distributions are assumed to be fixed throughout the learning process, while rewards are allowed to change. Finally, we highlight the contributions that tackle non-stationary RL via change detection (Banerjee et al. (2017); Alegre et al. (2021); Da Silva et al. (2006)).

2.3 Background on online Change-point detection

In the online change-point detection literature, change-point detection algorithms are designed that allows the detection of a change in the distribution of a stochastic process from one probability distribution to another. The optimality properties in term of false alarm rate and detection delay of the algorithms are studied under several problem formulations and different model assumptions. The main theoretical foundations were set up by the work of Shiryaev Shiryaev (1963). Existing online change-point detection are globally categorized into two types: Bayesian approaches and non-Bayesian algorithms. The former provide uncertainty quantities of the detection while the latter mainly focuses on measuring the discrepancy of the data statistics before and after the change-point. Several Bayesian methods for online change-point detection (Alami et al. (2020); Agudelo-España et al. (2020); Knoblauch & Damoulas (2018); Saatçi et al. (2010)) rely on the standard Bayesian Online Change-Point Detection (Fearnhead & Liu (2007)) that recursively models the posterior probability of the elapsed time since the last change. On the other hand, non-Bayesian methods mainly rely on the likelihood ratio test Severo & Gama (2006); Maillard (2019b); Page (1954) that also leads to false positive when the probability of the latest observations decrease given an outlier.

3 Problem Formulation

An instance of an MDP can be concisely specified by the tuple $M(\mathsf{O}, \mathsf{A}, P, R, T)$, where $\mathsf{O} = \{1, ..., O\}$ represents the finite set of states ($O = |\mathsf{O}|$), $\mathsf{A} = \{1, ..., A\}$ denotes the finite set of actions ($A = |\mathsf{A}|$), T is the finite time horizon and $\{R_t\}_{t=1 o\in\mathsf{O}, a\in\mathsf{A}}^{T}$ is the sequence of distribution rewards. For a given t and state-action pair (o, a), $r_t \sim R_t(o, a)$ is drawn i.i.d according to some unknown distribution on $[0, 1]$. Moreover, we define the sequence of state transition distributions $P = \{P_t\}_{t=1}^{T}$, with $P_t = \{P_t(.|o, a)\}_{o\in\mathsf{O}, a\in\mathsf{A}}$, where $P_t(.|o, a)$ is a **multinomial** probability distribution over O for each state-action pair (o, a) at a given time instance t. While constraining the state transition distributions to be generated from a multinomial distribution may seem restrictive at first glance, it enjoys a widespread interest from various research communities, from which we list a few key applications: modelling the collisions in cognitive radio, monitoring the performances of statistical models, monitoring events in probes for network supervision, the multi armed bandit problem, experiments in clinical trials and recommender systems to name a few.

To clearly define our problem of interest, we outline the main assumptions to be considered.

Non-Stationarity. In the contrary to many approaches proposed by literature, we assume in all generality that the state transition distributions and rewards can change arbitrarily at unknown time steps (referred to as *change-points*

[*] In contrast to ours and Gajane et al. (2018)'s, which defines regret at time t as the difference between the reward achieved by the active MDPs optimal policy at time t and that of the learner's policy, Cheung et al. (2020); Ortner et al. (2020) consider a different notion of regret, which they also call *dynamic regret*. The latter is defined as the difference between the learner's policy and the best achievable steady-state policy. Although the term Yu & Mannor (2009); Even-Dar et al. (2009); Neu et al. (2010); Dick et al. (2014) is widely used, it is generally not very useful because the best steady-state policy can still lead to undesirable rewards, especially in environments with nonsmooth change

thereof), i.e, the changes are not constrained to accumulate to a certain predefined *variation budget* for example. We also, more fundamentally, make the minimal assumption that the finite set of states and that of actions are constant throughout the learning process. Moreover, being part of the exponential family of probability distributions, it exhibits a favorable concentration behavior, allowing the smooth integration to optimistic exploration based algorithms relying on upper-confidence bounds.

Exogeneity. We assume that the dependence of the reward distributions on the previous states, actions and rewards is controlled by an exogenous process that generates the rewards independently given the filtration of the MDP history $(o_1, a_1, r_1, ..., o_t, a_t, r_t)$.

Convergence & Bounds. For convergence, we naturally assume bounded rewards, i. e, $|R_t(o,a)| \leqslant \max\limits_{t,o,a} R_t(o,a) < \infty, \forall o \in \mathsf{O},\ \forall a \in \mathsf{A}, \forall t \in \{1, ..., T\}$.

Endogeneity. The agent starts at some arbitrary state o_{init}. At time t, they observe state $o_t \in \mathsf{O}$ and take action $a_t \in \mathsf{A}$ according to some policy $\pi \in \Pi$. As a result, they transition to the next state $o_{t+1} \in \mathsf{O}$ according to state transition distribution $P(.|o_t, a_t)$, receiving stochastic reward $r_t \sim R_t(o_t, a_t)$ drawn according to some unknown distribution on $[0, 1]$. The endogeneity assumption here restricts the set of feasible policies Π to be *non-anticipatory*, i.e, the policy choice only depends on the current state and the set of previous observations $(o_1, a_1, r_1, ..., o_t, a_t, r_t)$.

Now that we have introduced the set of assumptions we consider, we are now ready to formulate the definitions and that will be used throughout the paper.

Switching-MDP Problem. Following a first instance of the name for Multi-Armed Bandits Garivier & Moulines (2011), and then in Gajane et al. (2018) for MDPs, we adopt the name *switching-MDP* to characterize our setting. Defining the set of change-point times as $\{c_\ell\}_{\ell=0}^{K_T}$, we consider the piecewise stationary MDP M to be in configuration $M_\ell(\mathsf{O}, \mathsf{A}, P_\ell, R_\ell, T)$ for $t \in [c_\ell, c_{\ell+1})$. Hence the switching-MDP problem M can be fully expressed by the tuple $\mathbf{M} = \{\mathbb{S} = \{M_0, .., M_{K_T-1}\}, \mathbb{C} = \{c_0, .., c_{K_T}\}\}$, where c_0 and c_{K_T} are the respective learning start ($t = 1$) and end ($t = T$) times.

Diameter of a MDP. The diameter of an MDP M_ℓ is defined as follows:

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

where the random variable $\tau\left(o_1, o_2, M_\ell, \pi\right)$ denotes the number of steps a policy π from the set of feasible stationary policies Π takes to move from o_1 to o_2 on average. In particular, we refer to an MDP with a finite diameter as a *communicating MDP*.

As we use *regret* as a performance measure throughout this article, as it is done in numerous other learning settings, we introduce the definition of the *average reward* for a constituent MDP M_ℓ of $\mathbf{M}$, given the execution of an algorithm $\mathcal{A}$ following a stationary policy, which can be written as follows: $\rho(M_\ell, \mathcal{A}, o_{\text{init}}) := \lim\limits_{T \to \infty} \frac{1}{T} \mathbb{E}\left[\sum\limits_{t=1}^{T} r_t\right]$, where the sequence of rewards is assigned following the states and actions chosen by the policy generated by $\mathcal{A}$ starting from o_{init}.

Assuming each of the MDPs that constitute $\mathbf{M}$ to be communicating, we get by virtue of Puterman (2014) that the optimal average reward does not depend on the initial state of the MDP. This result is fundamental to the definition of the regret, as it allows to decompose the optimal average reward for $\mathbf{M}$ into the sum of that for its constituent MDPs $\{M_\ell\}_{\ell=0}^{K_T-1}$. Now, defining the optimal reward for MDP M_ℓ as follows: $\rho^\star_{M_\ell} := \max\limits_{\pi, o \in \mathsf{O}} \rho(M_\ell, \mathcal{A}, o)$ where, again, π is the non-anticipatory policy generated by $\mathcal{A}$. Now we are ready to define the regret for a switching-MDP problem as follows.

Regret for a switching-MDP. The regret of an algorithm $\mathcal{A}$ for a switching-MDP problem $\mathbf{M} = \left(\{M_\ell\}_{\ell=0}^{K_T}, \{c_\ell\}_{\ell=1}^{K_T}\right)$ up to time horizon T starting from some initial state s is written as:

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

4 Change-point Detection as remedy to Non-stationarity

In this section, we start by designing the multinomial version of the Restarted Bayesian Online Change-point detector introduced in Alami et al. (2020). Then, we provide the mathematical guarantees of this algorithm in term of false alarm rate and detection delay. Finally, we design the `R-BOCPD-UCRL2` strategy, an `UCRL2` instance equipped with the `R-BOCPD` in order to handle piecewise stationaty MDPs.

4.1 Restarted Bayesian Online Change-point detection for multinomial distributions

In this section, we study the online change point detection problem, where a sequence of independent multivariate random variables with common fluctuation upper bound are collected, and the mean may change at one or multiple time points. Indeed, we consider an agent aiming at detecting changes in the generation of an online stream. At each time step t, the agent observes the datum $x_t \sim \text{Multi}\left(\mu_{1,t}, ..., \mu_{O,t}\right)$: a random variable following the multinomial distribution of parameters $\left(\mu_{1,t}, ..., \mu_{O,t}\right) \in [0,1]^O$ ($x_t \in \{1, ..., O\}$) and need to decide whether or not there is a change in the generation of the stream. Alternatively, the agent may compute at each time step t, an estimation $\widehat{c}_t$ of the last change-point.

Notations In the following, we denote $n_{s:t} := t - s + 1$, the number of observations from time s until time t. Moreover, the empirical frequency of observing o in the sequence $\mathbf{x}_{s:t} = (x_s, ..., x_t)$ is denoted as $\widehat{\mu}_{o,s:t} := \frac{1}{n_{s:t}} \sum_{s'=s}^{t} \mathbb{I}\{x_{s'} = o\}$.

Definition 4.1 (Kullback Leibler divergence for multinomial distributions)**.** Let's $\boldsymbol{\theta}^{(1)} = \left(\theta_1^{(1)}, ..., \theta_O^{(1)}\right)$ and $\boldsymbol{\theta}^{(2)} = \left(\theta_1^{(2)}, ..., \theta_O^{(2)}\right)$ be the paramereters of two multinomial distributions, then the relative entropy from Multi $\left(\theta_1^{(2)}, ..., \theta_O^{(2)}\right)$ and Multi $\left(\theta_1^{(1)}, ..., \theta_O^{(1)}\right)$ is defined as follows: $\mathbf{KL}\left(\boldsymbol{\theta}^{(2)} \parallel \boldsymbol{\theta}^{(1)}\right) = \sum_{o=1}^{O} \theta_o^{(2)} \log \frac{\theta_o^{(2)}}{\theta_o^{(1)}}$.

Definition 4.2 (Piecewise stationary multinomial process)**.** Let T denote the stream length and K_T the overall number of change-points observed until time T. We assume that the observations $x_t \sim \text{Multi}\left(\mu_{1,t}, ..., \mu_{O,t}\right)$ are generated by a piecewise multinomial process such that there exists a non-decreasing change-points sequence $\left(c_\ell\right)_{\ell \in [1, K_T]} \in \mathbb{N}^{K_T}$ verifying:

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

Remark. We make the rather classical assumption that the change-points arrive according to a Poisson process with parameter $\rho \in (0,1)$, which we refer to as the *switching-rate* and assume to be close to 0. This allows the interval in between two successive change-points to be large enough for successful finite-time detection.

Theorem 4.3 (Lower Bound for the detection delay)**.** *Let:* $(x_r, ..., x_{c_\ell - 1}) \sim$ *Multi* $\left(\boldsymbol{\theta}^{(l-1)} = \left(\theta_1^{(\ell-1)}, ..., \theta_O^{(\ell-1)}\right)\right)$ *and* $(x_{c_\ell}, ..., x_t) \sim$ *Multi* $\left(\boldsymbol{\theta}^{(\ell)} = \left(\theta_1^{(\ell)}, ..., \theta_O^{(\ell)}\right)\right)$, *$\mathcal{A}$ an online change-point detection strategy, c_ℓ the change-point to detect and r the restarting time. Assuming that the false alarm rate is controlled such that:* $\mathbb{P}_{\boldsymbol{\theta}^{(\ell-1)}}\left\{\exists s \in [r, \tau_c) : \mathcal{A}\left(\mathbf{x}_{r:s}\right) = 1\right\} \leqslant \delta$, *then as the quantity* $\frac{n_{r:c_\ell}}{|\log \delta|} \underset{\delta \to 0}{\to} \infty$, *the expected detection delay* $\mathbb{E}_{\boldsymbol{\theta}^{(\ell-1)}, \boldsymbol{\theta}^{(\ell)}}\left[\widehat{c}_{\mathcal{A}}\left(\mathbf{x}_{r:t}\right) - c_\ell\right]$ *is lower bounded as follows:* $\mathbb{E}_{\boldsymbol{\theta}^{(\ell-1)}, \boldsymbol{\theta}^{(\ell)}}\left[\widehat{c}_{\mathcal{A}}\left(\mathbf{x}_{r:t}\right) - c_\ell\right] \geqslant \left(\frac{\mathbb{P}_{\boldsymbol{\theta}^{(\ell-1)}}\left\{\widehat{c}_{\mathcal{A}}(\mathbf{x}_{r:t}) > c_\ell\right\}}{\boldsymbol{KL}\left(\boldsymbol{\theta}^{(\ell)} \parallel \boldsymbol{\theta}^{(\ell-1)}\right)}\right) \log \frac{1}{\delta}$.

Extension of the Restarted Bayesian Online Change-point detector (R-BOCPD) for multinomial distributions. Alami et al. (2020) introduced Restarted Bayesian Online Change Point Detection (`R-BOCPD`), which is a pruned version of Bayesian Online Change-point Detector applicable for univariate Bernoulli-distributed samples with changes in the mean of the distribution.

In this section, we propose to extend the `R-BOCPD` algorithm introduced in Alami et al. (2020) in order to deal with multinomial distributions. First, we start by extending the Laplace predictor for multinomial distributions.

Definition 4.4 (Extending the Laplace predictor)**.** The predictor $\textsc{Pred}\left(x_{t+1} | \mathbf{x}_{s:t}\right)$ takes as input a sequence $\mathbf{x}_{s:t} \in \left\{1, ..., O\right\}^{n_{s:t}}$ and predicts the value of the next observation $x_{t+1} \in \left\{1, ..., O\right\}$ as follows

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

where $\forall x \in \{1, ..., O\}$ $\textsc{Pred}\,(x|\emptyset) = \frac{1}{O}$ corresponds to the uniform prior given to the process generating θ_c.

Recall that in the work of Alami et al. (2020), instead of dealing with a run-length, they deal with the notion of forecaster that is a product of successive Laplace predictors. Thus, as in Alami et al. (2020), we introduce the loss of a forecaster.

Definition 4.5 (Forecaster loss)**.** Using the predictor, the instantaneous loss of the forecaster s at time t is given by:

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

Then, let $\widehat{L}_{s:t} := \sum_{s'=s}^{t} l_{s':t}$ denotes the cumulative loss incurred by the forecaster s from time s until time t which takes the following crude expression:

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

Thus, the forecaster weights update will remain the same (for some temporal function $\eta_{r,s,t} \in (0, 1)$).

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

Finally, we keep the same restart procedure as in Alami et al. (2020), namely

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

We describe the R-BOCPD algorithm for multinomial distributions in Algorithm 1.

Algorithm 1 R-BOCPD for multinomial distributions

```
Input: $\eta_{r,s,t} \in (0, 1)$
$r \leftarrow 1$, $\omega_{r,1:1} \leftarrow 1$, $\eta_{r,1,1} \leftarrow 1$.
for $t = 1, \dots$ do
    Observe $x_t \sim$ Multi $(\mu_{1,t}, ..., \mu_{O,t})$
    Define for each forecaster $s$ from time $r$ to time $t$: $\omega_{r,s:t} \leftarrow \begin{cases} \frac{\eta_{r,s,t}}{\eta_{r,s,t-1}} \exp\left(-l_{s:t}\right) \omega_{r,s:t-1} & \forall s < t, \\ \eta_{r,t,t} \times \mathcal{W}_{r:t-1} & s = t\,. \end{cases}$
    if Restart $(x_r, ..., x_t) = 1$ then $r \leftarrow t + 1$, $\omega_{r,r:r} \leftarrow 1$, $\eta_{r,r,r} \leftarrow 1$.
    Estimate the last change-point: $\widehat{c}_t \leftarrow r$.
end for
```

4.2 Performance Guarantees for the R-BOCPD in the Multinomial Case

Throughout this section, we provide sufficient conditions on the parameter $\eta_{r,s,t}$ that guarantee the false-alarm rate control (Theorem 4.6) and the finite detection delay (Theorem 4.8) for the R-BOCPD algorithm in the multinomial case.

4.2.1 Controlled False-alarm Rate

Theorem 4.6 (False-alarm rate)**.** *Let: $\boldsymbol{\theta} = (\theta_1, ..., \theta_O)$ denotes the vector of the parameters for a Multinomial distribution Multi $(\theta_1, ..., \theta_O)$. For $r < t$, assume that $(x_r, ..., x_t) \sim$ Multi $(\theta_1, ..., \theta_O)^{\otimes n_{r:t}}$. Let $\alpha > 1$, if $\eta_{r,s,t}$ is small enough such that:*

$$\eta_{r,s,t} < \left(\prod_{i=1}^{O-1} \frac{(n_{r:s-1} + i)\,(n_{s:t} + i)}{n_{r:t} + i}\right) \times \frac{\exp\left(2b_1\right)}{\left(n_{r:s-1} n_{s:t}\right)^{\frac{O-1}{2}} \times (O-1)!} \times \left(\frac{\log(4\alpha + 2)\delta^2}{4 n_{r:t} \log((\alpha + 3)\, n_{r:t})}\right)^{\alpha} \tag{6}$$
$$\textit{where: } b_1 = -\frac{O}{12} - \frac{O-1}{2} \log\left(2\pi\right) + \frac{O}{2} \log O.$$

then, with a probability higher than $1-\delta$, no false alarm occurs on the interval $[r, c_\ell)$:

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

Definition 4.7 (Relative gap $\Delta_{o,r,s,t}$)**.** Let $\Delta_o \in [0,1]$. The relative gap $\Delta_{o,r,s,t}$ for the forecaster s at time t takes the following form (depending on the position of s):

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

4.2.2 Optimal Detection Delay

Theorem 4.8 (Finite detection delay)**.** *Let $(x_r, ..., x_{c_\ell-1}) \sim$ Multi $\left(\theta_1^{(\ell-1)}, ..., \theta_O^{(\ell-1)}\right)^{\otimes n_{r:c_\ell-1}}$, $(x_{c_\ell}, ..., x_t) \sim$ Multi $\left(\theta_1^{(\ell)}, ..., \theta_O^{(\ell)}\right)^{\otimes n_{c_\ell:t}}$, $\Delta_o = \left|\theta_o^{(\ell-1)} - \theta_o^{(\ell)}\right|$: the change-point gap related to observation $o \in \{1, ..., O\}$ and $\boldsymbol{\Delta} = (\Delta_1, ..., \Delta_O)$ the vector of change-point gap. Then, let: $f_{r,s,t} = \sum_{i=1}^{O-1}\log\left(n_{r:s-1}+i\right) + \sum_{i=1}^{O-1}\log\left(\frac{n_{s:t}+i}{n_{r:t}+i}\right) - \frac{O-1}{2}\log\left(\frac{n_{s:t}}{n_{r:t}}\right) - \log(O-1)!$. If $\eta_{r,s,t}$ is large enough such that:*

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

then, the change-point c_ℓ is detected (with probability at least $1-\delta$) with delay not exceeding $\mathfrak{D}_{\boldsymbol{\Delta},r,c_\ell}$, such that:

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

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

Discussion 4.9. *(Asymptotic behavior of the detection delay) The asymptotic regime corresponds to the case where the elapsed time between the last restart r and the new change point c_ℓ tends to infinity, while the probability of false alarm δ tends to zero. Thus, we get:*

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

where (a) originates from the Pinsker inequality that relates the Kullback-Leibler divergence to the total variation. Thus, following the statement of Theorem 4.3, the detection delay $\mathfrak{D}_{\boldsymbol{\Delta},r,c_\ell}$ of the `R-BOCPD` is asymptotically order optimal.

Discussion 4.10. *(Choice of the hyperparameter $\eta_{r,s,t}$) In order to guarantee the false alarm rate and detection delay, we need to choose an appropriate form of the parameter $\eta_{r,s,t}$ that satisfies both conditions in Equation (7) and Equation (6). Choosing $\eta_{r,s,t} \approx \frac{1}{n_{r:t}}$ is satisfying both conditions.*

4.3 The `R-BOCPD-UCRL2` Strategy

Now, we propose to equip `UCRL2` in the switching-MDP setting $\mathbf{M}$ with `R-BOCPD`. This originates from the ability to decompose $\mathbf{M} = \{\mathbb{S} = \{M_0, .., M_{K_T-1}\}, \mathcal{C} = \{c_0, .., c_{K_T}\}\}$ according to independently generated stationary periods $(c_\ell, c_{\ell+1})$, along which `R-BOCPD` first detects the switch from MDP M_ℓ to $M_{\ell+1}$ at change instance c_ℓ and restarts `UCRL2` accordingly with minimal delay with high probability as quantified by the provided upper confidence bound. We also ensure that no other restarts occur during a stationary period with high probability in a similar way. The exact approach is explained in more detail in Algorithm 2.

Definition 4.11. (`UCRL2` Framework). We adopt the same notation and learning framework as in the original `UCRL2` Auer et al. (2008b), which we omit here due to space constraints. We list that in more depth in Appendix B.

Theorem 4.12 (Finite-time Optimal Regret Upper Bound)**.** *With probability at least $1-\delta$, it holds for a switching-MDP problem $\mathbf{M} = \{\mathbb{S} = \{M_0, .., M_{K_T-1}\}, \mathcal{C} = \{c_0, .., c_{K_T}\}\}$ (starting at some initial state o_{c_0}) with (piecewise) stationary periods of length at least 1 that the* `R-BOCPD-UCRL2` *regret as defined in Section 3 is bounded as follows:*

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

where $\mathfrak{D}_{\boldsymbol{\Delta}_{\ell+1}, c_\ell+d_\ell, c_{\ell+1}}$ is `R-BOCPD`*'s detection delay on input stream $(o_{c_\ell+d_\ell}, ..., o_{c_{\ell+1}})$ for the gap $\boldsymbol{\Delta}_{\ell+1} = (\Delta_{1,\ell+1}, ..., \Delta_{O,\ell+1})$ where $\Delta_{o,\ell+1} = \left|\theta_o^{(\ell)} - \theta_o^{(\ell+1)}\right|$, with $\boldsymbol{\theta}^{(\ell)} = \left(\theta_1^{(\ell)}, ..., \theta_O^{(\ell)}\right)$ and $\boldsymbol{\theta}^{(\ell+1)} = \left(\theta_1^{(\ell+1)}, ..., \theta_O^{(\ell+1)}\right)$ being the pre and post state-transition kernels over the set of states O for change-point $c_{\ell+1}$.*

Now, we introduce a corollary to characterize the asymptotic behavior of `R-BOCPD-UCRL2`'s regret.

Corollary 4.13 (Asymptotic Regret Upper Bound)**.** *With probability at least $1-\delta$, assuming $c_{\ell+1} - c_\ell - d_\ell \gg 1$ with false-alarm probability $\delta_{\textit{False-Alarm}} \to 0$ (as in Equation (10)), we can write the asymptotic upper bound for the regret of* `R-BOCPD-UCRL2` *on $\mathbf{M}$ starting at some state o_{c_0} as follows:*

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

Discussion 4.14. *(Optimality of the Upper Bound) We derived both a finite-time and an asymptotic variant of* `R-BOCPD-UCRL2`*'s regret upper bound, both comparing favorably to state-of-the-art. Given the purpose of design of* `R-BOCPD-UCRL2`*, which is to allow it to adapt to rapidly and abruptly changing non-stationary RL environments, the regret bound correlates optimally with the distance between the distributions before and after the change-point. We also highlight that our proposed approach is the first one to obtain a regret of $\tilde{\mathcal{O}}(T^{\frac{1}{2}})$ up to our knowledge. Previously proposed sliding-window approaches Gajane et al. (2018); Cheung et al. (2020) obtain a regret bound of $\tilde{\mathcal{O}}(T^{\frac{2}{3}})$ and $\tilde{\mathcal{O}}(T^{\frac{3}{4}})$ respectively.* `UCRL2` *with restarts Auer et al. (2008a), even while restarting $T^{\frac{1}{3}} K_T^{\frac{-1}{3}}$ more times than* `R-BOCPD-UCRL2`*, still only obtains a regret of $\tilde{\mathcal{O}}(T^{\frac{2}{3}})$.*

5 Experiments

We benchmark `R-BOCPD-UCRL2` against 4 algorithms that perform the best within our setting, to the best of our knowledge. We list them as follows:

- **Sliding-Window UCRL2** (`SWUCRL2`, Gajane et al. (2018)) uses a sliding-window approach to only maintain the last W time steps of the filtration history, where W is referred to as the *window-size*.
- **Sliding-Window UCRL2 with Confidence Widening** (`SWUCRL2-CW`, Cheung et al. (2020)) rely on even more *optimism* than `SWUCRL2`, where in addition to a window of size W, defines a *confidence widening* parameter η that quantifies the amount of additional optimistic exploration to be done on top of the conventional optimistic exploration realized via upper confidence bounds.
- **Restarted UCRL2** (`Restarted-UCRL2`, Auer et al. (2008a)) define a variant of vanilla `UCRL2` where the latter is restarted at steps $\tau_i = \frac{i^3}{{K_T}^2}$ and where the number of changes K_T is assumed to be known at initialization.
- **Vanilla UCRL2** (`UCRL2`, Auer et al. (2008a)).

Moreover, we consider a variant of `UCRL2`, which is regret-optimal as per our regret definition in Section 3 (as `UCRL2` is near-optimal at each stationary period $[c_\ell, c_{\ell+1})$). It is defined as follows

- **Oracle-Equipped UCRL2** (`UCRL2 Oracle`) is *aware* of all the changes $\{c_\ell\}_{\ell=1}^{K_T-1}$ already at initialization and hence restarts `UCRL2` exactly at each c_ℓ for $\ell \in \{0, ..., K_T - 1\}$.

We list the exact experimental setup along with the hyperparameters of choice of each algorithm in more detail in Appendix G.

5.1 Experimental Results

We evaluate the performance of the aforementioned algorithms on a variety of synthetically generated MDPs, with state-action sets of different cardinalities. The change-points are (randomly) generated up to time horizon T, allowing to examine the effect of changing the duration in-between change-points on learning. We plot the cumulative rewards of each approach as a function of time as follows

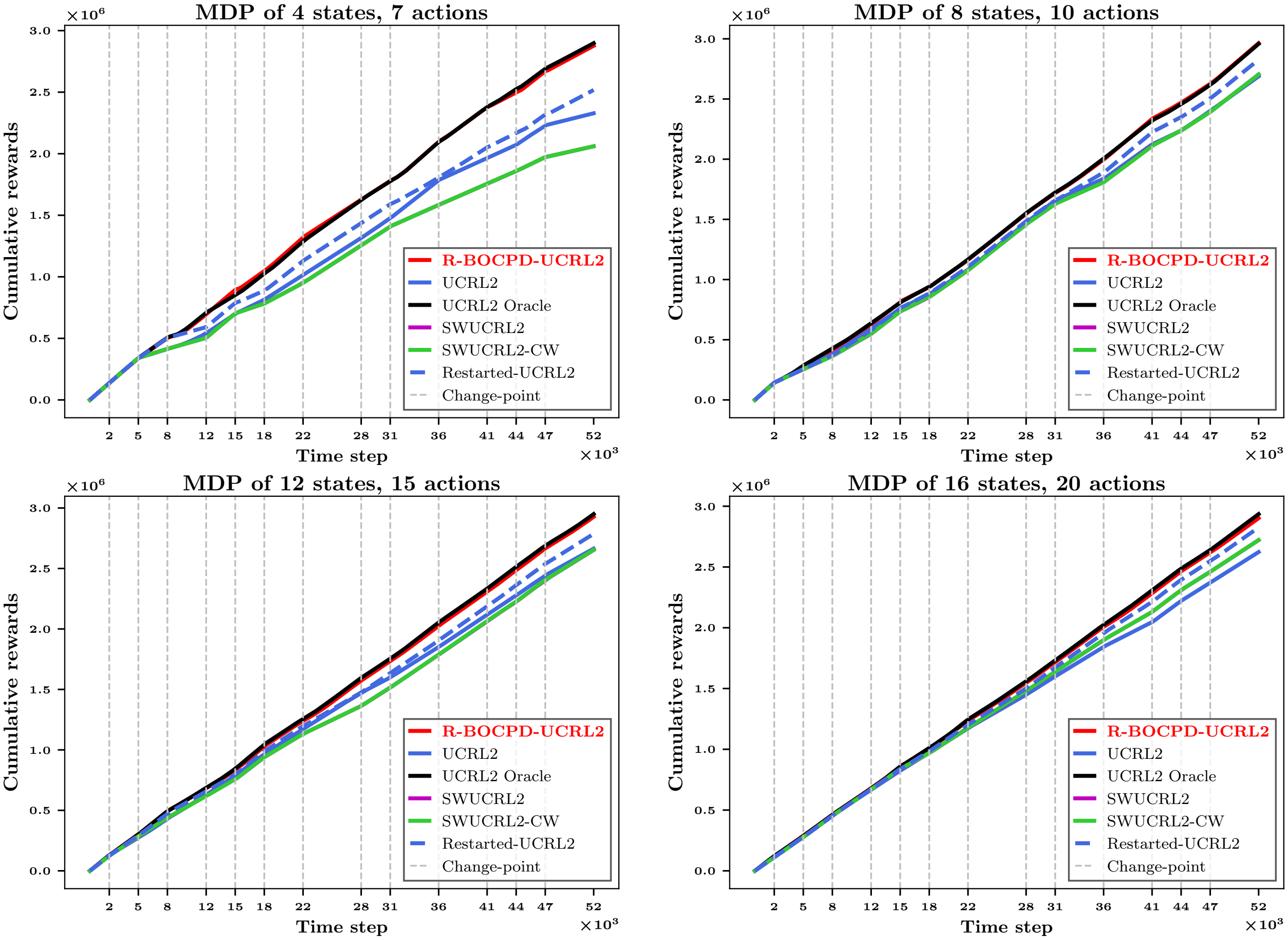

Figure 2: Benchmark of R-BOCPD-UCRL2 against state-of-the-art for various state-action pairs for a sequence of random change-points. The level of abruptness, i.e the variation of the state-transition distributions and rewards also varies among change-points, allowing to model both globally and locally induced changes to the MDPs.

5.2 Performance Evaluation

Figure 2 shows that R-BOCPD-UCRL2 is (nearly) regret-optimal in practice, as its performance in the defined general setting is very close to that of the change-point aware (optimal) UCRL2 Oracle. We also highlight that the observed performance generalizes well beyond switching-MDP problems of various state and action space sizes and different total *variation budgets* for both the state-transition distributions and rewards. Given the space constraints, a more in-depth discussion of the performance of each algorithm along with its key assumptions is provided in Appendix G.2.

6 Discussion & Overall Remarks

In this work, we proposed the Restarted Bayesian Online Change-Point Detection algorithm (R-BOCPD-UCRL2), which is a change-point detector equipped model-based RL algorithm operating on non-stationary environments that can be fully characterized via a discrete-time piecewise-stationary MDP. We extended the theoretical guarantees of the Bayesian Online Change-Point Detection algorithm (BOCPD) to the more general multinomial distribution and proved that R-BOCPD-UCRL2 is regret-optimal with an asymptotic regret bound of $\mathcal{O}\left(DO\sqrt{ATK_T\log\left(\frac{T}{\delta}\right)}+\frac{K_T\log\frac{K_T}{\delta}}{\min\limits_{\ell}\mathbf{KL}\left(\boldsymbol{\theta}^{(\ell+1)}\,\|\,\boldsymbol{\theta}^{(\ell)}\right)}\right)$, which is the first to achieve a bound of $\tilde{O}(T^{\frac{1}{2}})$ in the time horizon T up to our knowledge. We further proved the optimality of R-BOCPD-UCRL2 in practice, as it compares to state-of-the-art, with much fewer restarts and no implicitly defined input parameters (MDP diameter, variation budget etc).

Limitations and Future Work. Here, we highlight a few ideas that were not considered within the scope of this manuscript, but still would be very promising directions in the authors' opinion. First, we address the assumption that the state-transition distributions originate from a multinomial distribution. We note that the Bayesian Online Change-point Detector does not necessarily require an input stream stemming from a multinomial distributions, but can be extended to arbitrary distributions from the exponential family of probability distributions. Given the latter, the theoretical guarantees in terms of minimal detection delay and false-alarm rate do extend as well. Next, we highlight that our algorithm is biased towards detecting change-points around which distributions change in a rather radical way, i.e with large enough variation in the sense of a total variation norm for instance. Here, a good direction would be to define a threshold value to allow the change-point detector to decide when to restart the stationary RL algorithm (here UCRL2) given the global/local nature of the changes to the MDP parameters. Finally, a natural extension would also be to propose new change-point detector equipped model-free non-stationary RL algorithms.

REFERENCES

Shipra Agrawal and Randy Jia. Learning in structured mdps with convex cost functions: Improved regret bounds for inventory management. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pp. 743–744, 2019.

Diego Agudelo-España, Sebastian Gomez-Gonzalez, Stefan Bauer, Bernhard Schölkopf, and Jan Peters. Bayesian online prediction of change points. In *Conference on Uncertainty in Artificial Intelligence*, pp. 320–329. PMLR, 2020.

Réda Alami. Thompson Sampling for the non-Stationary Corrupt Multi-Armed Bandit. In *The 14th European Workshop on Reinforcement Learning*, volume 14, Lille, France, October 2018. URL https://hal.science/hal-01963539.

Reda Alami. Bayesian change-point detection for bandit feedback in non-stationary environments. In Emtiyaz Khan and Mehmet Gonen (eds.), *Proceedings of The 14th Asian Conference on Machine Learning*, volume 189 of *Proceedings of Machine Learning Research*, pp. 17–31. PMLR, 12–14 Dec 2023. URL https://proceedings.mlr.press/v189/alami23a.html.

Réda Alami and Oussama Azizi. TS-GLR: an Adaptive Thompson Sampling for the Switching Multi-Armed Bandit Problem. In *NeurIPS 2020 challenges of real world reinforcement learning workshop*, Virtual, Canada, December 2020. URL https://hal.science/hal-03628791.

Réda Alami, Odalric Maillard, and Raphael Féraud. Memory Bandits: a Bayesian approach for the Switching Bandit Problem. In *NIPS 2017 - 31st Conference on Neural Information Processing Systems*, Long Beach, United States, December 2017. URL https://hal.science/hal-01811697.

Reda Alami, Odalric Maillard, and Raphael Feraud. Restarted Bayesian online change-point detector achieves optimal detection delay. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 211–221. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/alami20a.html.

Lucas N Alegre, Ana LC Bazzan, and Bruno C da Silva. Minimum-delay adaptation in non-stationary reinforcement learning via online high-confidence change-point detection. *arXiv preprint arXiv:2105.09452*, 2021.

Raman Arora, Ofer Dekel, and Ambuj Tewari. Deterministic mdps with adversarial rewards and bandit feedback. *arXiv preprint arXiv:1210.4843*, 2012.

Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. *Advances in neural information processing systems*, 21, 2008a.

Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. *Advances in neural information processing systems*, 21, 2008b.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pp. 263–272. PMLR, 2017.

Taposh Banerjee, Miao Liu, and Jonathan P How. Quickest change detection approach to optimal control in markov decision processes with model changes. In *2017 American control conference (ACC)*, pp. 399–405. IEEE, 2017.

Kostas Bimpikis, Ozan Candogan, and Daniela Saban. Spatial pricing in ride-sharing networks. *Operations Research*, 67(3):744–769, 2019.

Wang Chi Cheung, David Simchi-Levi, and Ruihao Zhu. Reinforcement learning for non-stationary markov decision processes: The blessing of (more) optimism. In *International Conference on Machine Learning*, pp. 1843–1854. PMLR, 2020.

Bruno C Da Silva, Eduardo W Basso, Ana LC Bazzan, and Paulo M Engel. Dealing with non-stationary environments using context detection. In *Proceedings of the 23rd international conference on Machine learning*, pp. 217–224, 2006.

Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 30, 2017.

Travis Dick, Andras Gyorgy, and Csaba Szepesvari. Online learning in markov decision processes with changing cost sequences. In *International Conference on Machine Learning*, pp. 512–520. PMLR, 2014.

Eyal Even-Dar, Sham M Kakade, and Yishay Mansour. Online markov decision processes. *Mathematics of Operations Research*, 34(3):726–736, 2009.

Paul Fearnhead and Zhen Liu. On-line inference for multiple changepoint problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):589–605, 2007.

Pratik Gajane, Ronald Ortner, and Peter Auer. A sliding-window algorithm for markov decision processes with arbitrarily changing rewards and transitions. *arXiv preprint arXiv:1805.10066*, 2018.

Aurélien Garivier and Eric Moulines. On upper-confidence bound policies for switching bandit problems. In *Algorithmic Learning Theory: 22nd International Conference, ALT 2011, Espoo, Finland, October 5-7, 2011. Proceedings 22*, pp. 174–188. Springer, 2011.

Itai Gurvich, Martin Lariviere, and Antonio Moreno. Operations in the on-demand economy: Staffing services with self-scheduling capacity. *Sharing Economy: Making Supply Meet Demand*, pp. 249–278, 2019.

Woonghee Tim Huh and Paat Rusmevichientong. A nonparametric asymptotic analysis of inventory planning with censored demand. *Mathematics of Operations Research*, 34(1):103–123, 2009.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? *Advances in neural information processing systems*, 31, 2018.

Chi Jin, Tiancheng Jin, Haipeng Luo, Suvrit Sra, and Tiancheng Yu. Learning adversarial markov decision processes with bandit feedback and unknown transition. In *International Conference on Machine Learning*, pp. 4860–4869. PMLR, 2020.

Yash Kanoria and Pengyu Qian. Blind dynamic resource allocation in closed networks via mirror backpressure. *arXiv preprint arXiv:1903.02764*, 2019.

István Z Kiss, Joel C Miller, Péter L Simon, et al. Mathematics of epidemics on networks. *Cham: Springer*, 598:31, 2017.

Jeremias Knoblauch and Theodoros Damoulas. Spatio-temporal bayesian on-line changepoint detection with model selection. In *International Conference on Machine Learning*, pp. 2718–2727. PMLR, 2018.

Odalric-Ambrym Maillard. Mathematics of statistical sequential decision making. 2019a.

Odalric-Ambrym Maillard. Sequential change-point detection: Laplace concentration of scan statistics and non-asymptotic delay bounds. In Aurélien Garivier and Satyen Kale (eds.), *Proceedings of the 30th International Conference on Algorithmic Learning Theory*, volume 98 of *Proceedings of Machine Learning Research*, pp. 610–632. PMLR, 22–24 Mar 2019b. URL https://proceedings.mlr.press/v98/maillard19a.html.

Gergely Neu, Andras Antos, András György, and Csaba Szepesvári. Online markov decision processes under bandit feedback. *Advances in Neural Information Processing Systems*, 23, 2010.

Cameron Nowzari, Victor M Preciado, and George J Pappas. Analysis and control of epidemics: A survey of spreading processes on complex networks. *IEEE Control Systems Magazine*, 36(1):26–46, 2016.

Ronald Ortner, Pratik Gajane, and Peter Auer. Variational regret bounds for reinforcement learning. In *Uncertainty in Artificial Intelligence*, pp. 81–90. PMLR, 2020.

Ewan S Page. Continuous inspection schemes. *Biometrika*, 41(1/2):100–115, 1954.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Adrian Rivera Cardoso, He Wang, and Huan Xu. Large scale markov decision processes with changing rewards. *Advances in Neural Information Processing Systems*, 32, 2019.

Yunus Saatçi, Ryan D Turner, and Carl E Rasmussen. Gaussian process change point models. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pp. 927–934, 2010.

Milton Severo and Joao Gama. Change detection with kalman filter and cusum. In *Discovery Science: 9th International Conference, DS 2006, Barcelona, Spain, October 7-10, 2006. Proceedings 9*, pp. 243–254. Springer, 2006.

Albert N Shiryaev. On optimum methods in quickest detection problems. *Theory of Probability & Its Applications*, 8 (1):22–46, 1963.

Aaron Sidford, Mengdi Wang, Xian Wu, Lin F Yang, and Yinyu Ye. Near-optimal time and sample complexities for solving discounted markov decision process with a generative model. *arXiv preprint arXiv:1806.01492*, 2018a.

Aaron Sidford, Mengdi Wang, Xian Wu, and Yinyu Ye. Variance reduced value iteration and faster algorithms for solving markov decision processes. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 770–787. SIAM, 2018b.

Alexander L Strehl and Michael L Littman. An analysis of model-based interval estimation for markov decision processes. *Journal of Computer and System Sciences*, 74(8):1309–1331, 2008.

Terry A Taylor. On-demand service platforms. *Manufacturing & Service Operations Management*, 20(4):704–720, 2018.

Mengdi Wang. Randomized linear programming solves the markov decision problem in nearly linear (sometimes sublinear) time. *Mathematics of Operations Research*, 45(2):517–546, 2020.

Kuang Xu and Se-Young Yun. Reinforcement with fading memories. *Mathematics of Operations Research*, 45(4): 1258–1288, 2020.

Jia Yuan Yu and Shie Mannor. Arbitrarily modulated markov decision processes. In *Proceedings of the 48h IEEE Conference on Decision and Control (CDC) held jointly with 2009 28th Chinese Control Conference*, pp. 2946–2953. IEEE, 2009.

Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning*, pp. 7304–7312. PMLR, 2019.

Zihan Zhang and Xiangyang Ji. Regret minimization for reinforcement learning by evaluating the optimal bias function. *Advances in Neural Information Processing Systems*, 32, 2019.

Zhengyuan Zhou and Nicholas Bambos. Wireless communications games in fixed and random environments. In *2015 54th IEEE Conference on Decision and Control (CDC)*, pp. 1637–1642. IEEE, 2015.

Zhengyuan Zhou, Peter Glynn, and Nicholas Bambos. Repeated games for power control in wireless communications: Equilibrium and regret. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, pp. 3603–3610. IEEE, 2016.

A Main Algorithm

Algorithm 2 R-BOCPD-UCRL2

```
Input: A confidence parameter $\delta \in (0,1)$, $\eta_{r,s,t} \in (0,1)$
Set $\forall(o,a)\ N_0(o,a) \leftarrow 0, V_0(o,a) \leftarrow 0, t \leftarrow 1, k \leftarrow 1$ and observe initial state $s_1$.
Initialize restart time $r \leftarrow 1$
for each $(o,a)$ pair do
    Initialize a R-BOCPD$_{o,a}$ procedure
end for
for episodes $k \geqslant 1$ do
    Initialize episode $k$:
        Set the start time of episode $k$, $t_k = t$
        For all $(o,a) \in \mathsf{O} \times \mathsf{A}$ initialize the state-action counts for episode $k, V_k(o,a) := 0$ . Further, set the number
    of times any action action $a$ was executed in state $o$ prior to episode $k$ for all the states $o \in \mathcal{O}$ and actions $a \in \mathcal{A}$,
            $N_k(o,a) := \#\{r \leqslant t < t_k : o_t = o, a_t = a\}$ .
        For all $o, o' \in \mathsf{O}$ and $a \in \mathsf{A}$, set the observed cumulative rewards when action $a$ was executed in state $o$ and
    the number of times that resulted into the next state being $o'$ prior to episode $k$,
            $R_k(o,a) := \sum_{t=r}^{t_k-1} r_t \mathbb{I}\{o_t = o, a_t = a\}$, and $P_k(o,a,o') := \#\{r \leqslant t < t_k : o_t = o, a_t = a, o_{t+1} = o'\}$ .
        Compute estimates $\widehat{R}_k(o,a) := \frac{R_k(o,a)}{\max\{1,N_k(o,a)\}}$, $\widehat{P}_k(o' \mid o,a) := \frac{P_k(o,a,o')}{\max\{1,N_k(o,a)\}}$
    Compute policy $\tilde{\pi}_k$:
        Let $\mathcal{M}_k$ be the set of all MDPs with state space $\mathsf{O}$ and action space $\mathsf{A}$, and with transition probabilities
    $\tilde{P}(\cdot \mid o,a)$ close to $\widehat{P}_k(\cdot \mid o,a)$, and rewards $\tilde{R}(o,a) \in [0,1]$ close to $\widehat{R}_k(o,a)$, such that:
            $\left|\tilde{R}(o,a) - \widehat{R}_k(o,a)\right| \leqslant \sqrt{\frac{7\log(2OAt_k/\delta)}{2\max\{1,N_k(o,a)\}}}$ and $\left\|\tilde{P}(\cdot \mid o,a) - \widehat{P}_k(\cdot \mid o,a)\right\|_1 \leqslant \sqrt{\frac{14O\log(2At_k/\delta)}{\max\{1,N_k(o,a)\}}}$.
        Use extended value iteration to find a policy $\tilde{\pi}_k$ and an optimistic MDP $\bar{M}_k \in \mathcal{M}_k$ such that
            $\tilde{\rho}_k := \min_{o \in \mathcal{O}} \rho\left(\bar{M}_k, \tilde{\pi}_k, o\right) \geqslant \max_{M' \in \mathcal{M}_k, \pi, o'} \rho(M', \pi, o') - \frac{1}{\sqrt{t_k}}$
    Execute policy $\tilde{\pi}_k$:
    while $V_k(o_t, \tilde{\pi}_k(o_t)) < \max\{1, N_k(o_t, \tilde{\pi}_k(o_t))\}$ do
        Choose action $a_t = \tilde{\pi}_k(o_t)$, obtain reward $r_t$ and observe next state $o_{t+1}$.
        Update $V_k(o_t, a_t) := V_k(o_t, a_t) + 1$
        Set $t := t + 1$
        Perform a change-point detection test over the sequence $(o_r, ..., o_t)$.
        if R-BOCPD$_{o_t,a_t}$.Restart$(o_r, ..., o_t) = 1$ then
            $\forall(o,a)\ N_k(o,a) \leftarrow 0, V_k(o,a) \leftarrow 0, r \leftarrow t + 1$.
        end if
    end while
end for
```

B UCRL2 Framework

We introduce our learning framework and relevant notation as follows:

- $N_k(o,a)$ is the number of times any action action a was executed in state o up to episode k for all the states $o \in \mathsf{O}$ and actions $a \in \mathsf{A}$.
- $V_k(o,a)$ is the number of visits to state-action pair (o,a) up to episode k.

- $P_k\left(o, a, o'\right)$ is the state-transition kernel where action $a \in \mathsf{A}$ taken at state $o \in \mathsf{O}$ takes the agent to state $o' \in \mathsf{O}$, where P_k is defined up to time t_k starting from last `UCLR2` restart time r.
- $\widehat{P}_k\left(o' \mid o, a\right)$ is the estimated state-transition kernel for triple $\left(o, a, o'\right)$ from the last $t_k - r + 1$ observations.
- $R_k(o, a)$ are the mean rewards for state-action pair (o, a) up to time t_k starting from last `UCLR2` restart time r.
- $\widehat{R}_k(o, a)$ are the estimated mean rewards for state-action pair (o, a) given the last $t_k - r + 1$ observations.
- $\mathcal{M}_k$ is defined as the set of statistically plausible MDPs given $\widehat{P}_k\left(o' \mid o, a\right)$ and $\widehat{R}_k(o, a)$, with state space O and action space A.
- $\bar{M}_k$ is an optimistic MDP chosen from $\mathcal{M}_k$.
- $\tilde{P}(\cdot \mid o, a)$ is the transition kernel of $\bar{M}_k$ that is close to $\widehat{P}_k(\cdot \mid o, a)$.
- $\tilde{R}(o, a)$ are the mean rewards of $\bar{M}_k$ that are close to $\widehat{R}_k(o, a)$.
- $\tilde{\pi}_k$ is a near optimal policy for $\bar{M}_k$ chosen via extended value iteration, as defined in Auer et al. (2008b).

C Control of the cumulative loss in the multinomial case

C.1 Notation and useful definitions

In the following, we denote by $\Sigma_{o,s:t}$ the number of times the realization $o \in \{1, ...O\}$ has been observed in the sequence $\mathbf{x}_{s:t}$ such that:

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

C.2 Cumulative loss close form

Notice that:

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

Let's show by induction on $n_{s:t} \in \mathbb{N}^\star$ that:

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

Proof of Equation (11):

Step 1: For $n_{s:t} = 1$. It means that $t = s$, $x_t \in \{1, ..., O\}$ and $\mathbf{x}_{s':t-1} = \emptyset$. Using the definition of the predictor $\textsc{Pred}\left(|\right)$, we obtain

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

Step 2: Assume that for some $n_{s:t} \in \mathbb{N}^{\star}$ that corresponds to the sequence $\mathbf{x}_{s:t} \in \{1, ..., O\}$, we have

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

Then, observe that:

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

Then, using the definition of the forecaster in Equation (2) and the statement of Equation (12), we obtain (for $x_{t+1} = a \in \{1, ..., O\}$)

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

□

Notice that the cumulative loss $\widehat{L}_{s:t}$ can be written as follows:

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

where $n!$ denotes the factorial of n such that:

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

Then, using the following Stirling formula:

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

we get the upper bound and the lower bound of the quantity $\frac{n!}{n_1!n_2!....n_O!}$:

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

with $\sum_{i=1}^{O} n_i = n,\ n_i \geqslant 0\ \forall i \in \{1, ..., O\}$ and $b_1 = -\frac{O}{12} - \frac{O-1}{2}\log(2\pi) + \frac{O}{2}\log O$

C.3 Upper bound of the cumulative loss for stationary observations

Before deriving the upper bound on the cumulative loss, one should notice that:

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

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

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

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

Then, the upper bound of the cumulative loss takes the following form:

$$\begin{aligned}
\widehat{L}_{s:t} &\overset{(a)}{\leqslant} \Phi\left(n_{s:t}\right) - \sum_{o=1}^{O}\Phi\left(\Sigma_{o,s:t}\right) + \sum_{i=1}^{O-1}\log\left(n_{s:t}+i\right) - \log(O-1)! && (18)\\
&\overset{(b)}{\leqslant} \sum_{i=1}^{O-1}\log\left(n_{s:t}+i\right) - \sum_{o=1}^{O}\Sigma_{o,s:t}\log\theta_o - n_{s:t}\mathbf{KL}(\widehat{\mu}_{1,s:t}, ..., \widehat{\mu}_{O,s:t} \parallel \theta_1, ..., \theta_O) - \log(O-1)! \\
&\overset{(c)}{\leqslant} \sum_{i=1}^{O-1}\log\left(n_{s:t}+i\right) - \sum_{o=1}^{O}\Sigma_{o,s:t}\log\theta_o - \log(O-1)! && (19)
\end{aligned}$$

where:

- (a) holds using the left side of Equation (13) for $n = n_{s:t}$ and $a_o = \Sigma_{o,s:t}\ \ \forall o \in \{1, ..., O\}$.
- (b) holds thanks to the statement of Equation (17).
- (c) holds thanks to the fact that the Kullback Leibler divergence is always positive (i.e. $\mathbf{KL}(\bullet \parallel \bullet) \geqslant 0$).

C.4 Lower bound of the cumulative loss for stationary observations

The lower bound of the cumulative loss is taking the following form:

$$\begin{aligned}
\widehat{L}_{s:t} &\overset{(a)}{\geqslant} \Phi\left(n_{s:t}\right) - \sum_{o=1}^{O}\Phi\left(\Sigma_{o,s:t}\right) + \sum_{i=1}^{O-1}\log\left(n_{s:t}+i\right) - \frac{O-1}{2}\log n_{s:t} + b_1 - \log(O-1)! && (20)\\
&\overset{(b)}{\geqslant} \sum_{i=1}^{O-1}\log\left(n_{s:t}+i\right) - \sum_{o=1}^{O}\Sigma_{o,s:t}\log\theta_o - n_{s:t}\mathbf{KL}(\widehat{\mu}_{1,s:t}, ..., \widehat{\mu}_{O,s:t} \parallel \theta_1, ..., \theta_O) - \frac{O-1}{2}\log n_{s:t} + b_1 - \log(O-1)! && (21)
\end{aligned}$$

where:

- (a) holds using the left side of Equation (13) for $n = n_{s:t}$ and $n_o = \Sigma_{o,s:t} \;\; \forall o \in \{1, ..., O\}$
- (b) holds thanks to the statement of Equation (17).

Useful lemmas to derive the false alarm rate and detection delay

Lemma C.1 (Time uniform **KL**($\bullet \parallel \bullet$) concentration)**.** *Let: $\boldsymbol{\theta} = (\theta_1, ..., \theta_O)$ denotes the vector of the generative parameters for the Multinomial distribution Multi $(\theta_1, ..., \theta_O)$. $\forall o \in \{1, ..., O\}$, let $\widehat{\mu}_{o,t}$ denotes the empirical frequency of observing the realization $o \in \{1, ..., O\}$ in the sequence $(x_1, ..., x_t) \sim$ Multi $(\theta_1, ..., \theta_O)^{\otimes t}$, then for all $(\delta, \alpha) \in (0, 1) \times (1, \infty)$ we have:*

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

Lemma C.2 (Doubly-time uniform **KL**($\bullet \parallel \bullet$) concentration)**.** *Let: $\boldsymbol{\theta} = (\theta_1, ..., \theta_O)$ denotes the vector of the generative parameters for the Multinomial distribution Multi $(\theta_1, ..., \theta_O)$.*

$\forall o \in \{1, ..., O\}$, let $\widehat{\mu}_{o,s:t}$ denotes the empirical frequency of observing o in the sequence $(x_s, ..., x_t) \sim$ Multi $(\theta_1, ..., \theta_O)^{\otimes n_{s:t}}$, then for all $(\delta, \alpha) \in (0, 1) \times (1, \infty)$ we have:

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

Lemma C.3 (Doubly-time uniform concentration)**.** *Let: $(x_r, ...x_t) \in \{1, ..., O\}^{n_{r:t}}$ be a sequence of independent random variables sampled from a Multinomial distribution whose generative parameter can be chosen arbitrarily and $\widehat{\mu}_{o,i:j}$ the empirical frequency of observing o in the sequence $(x_i, ..., x_j)$. Then, for all $(r, \delta) \in \mathbb{N}^\star \times (0, 1)$, we get the following control:*

$$\mathbb{P}\Big\{\exists\, t > r, s \in [r, t) : |\widehat{\mu}_{o,r:s-1} - \widehat{\mu}_{o,s:t} - \mathbb{E}\left[\widehat{\mu}_{o,r:s-1} - \widehat{\mu}_{o,s:t}\right]| \geqslant \mathcal{C}'_{r,s,t,\delta}\Big\} \leqslant \delta,$$
$$\mathcal{C}'_{r,s,t,\delta} = \frac{\sqrt{2}}{2}\left(\sqrt{\frac{1+\frac{1}{n_{r:s-1}}}{n_{r:s-1}} \log\left(\frac{2\sqrt{n_{r:s}}}{\delta}\right)} + \sqrt{\frac{1+\frac{1}{n_{s:t}}}{n_{s:t}} \log\left(\frac{2n_{r:t}\sqrt{n_{s:t}+1}\log^2(n_{r:t})}{\log(2)\delta}\right)}\right).$$

The proof of lemmas C.1, C.2, and C.3 is beyond the scope of this manuscript, we refer the interested reader to section **3.4** of Maillard (2019a).

D Derivation of the false alarm rate

Proof of Theorem 4.6:

Let $\boldsymbol{\theta} = (\theta_1, ..., \theta_O)$ denotes the vector of the generative parameters for the Multinomial distribution Multi $(\theta_1, ..., \theta_O)^{\otimes n_{r:t}}$.

Assume that: $\forall t \in [r, c_\ell) \;\; (x_r, ..., x_t) \sim$ Multi $(\theta_1, ..., \theta_O)^{\otimes n_{r:t}}$. The proof follows three main steps:

Let us build a suitable value of $\eta_{r,s,t}$ in order to ensure the control of the false alarm on the period $[r, c_\ell)$. To this end, let us control the event: $\{\exists t > r, \texttt{Restart}\,(x_r, ..., x_t) = 1\}$ which is equivalent to the event $\{\exists t > r, \;\; s \in (r, t] : \omega_{r,s,t} \geqslant \omega_{r,r,t}\}$.

Step 1: Equivalent events. First, notice that:

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

where (a) comes directly from the definition of the forecaster weights $\omega_{r,s,t}$ stated in Equation (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\\
&\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)! \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}$$

where:

- (a) holds by using the monotonic behavior of the logarithm function.
- (b) holds thanks to Equation (4).
- (c) holds thanks to the use of the lower bound of the cumulative loss in Equation (21) and the upper bound of the cumulative loss in Equation (19).
- (d) holds by using basic properties of the logarithm function.
- (e) holds by using the property that: $\mathbb{P}\Big\{A\Big\} \leqslant \mathbb{P}\Big\{\neg B\Big\} + \mathbb{P}\Big\{A \cap B\Big\}$ where $B = E^{(1)}_{\boldsymbol{\theta},\delta/2,\alpha}$.
- (f) holds thanks to the statement of Lemma C.1.
- (g) holds by using the property that: $\mathbb{P}\Big\{A\Big\} \leqslant \mathbb{P}\Big\{\neg B\Big\} + \mathbb{P}\Big\{A \cap B\Big\}$ where $B = E^{(2)}_{\boldsymbol{\theta},\delta/2,\alpha}$.
- (h) holds thanks to the statement of Lemma C.2.
- (i) holds by using the monotonic behavior of the logarithm function.

Step 3: Sufficient condition on $\eta_{r,s,t}$ Based on Equation (23), we derive a sufficient condition on $\eta_{r,s,t}$ to guarantee the false alarm control:

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

Then, no false alarm occurs with high probability during a stationary period $[r, c_\ell)$:

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

□

E Derivation of the detection delay

Proof of Theorem 4.8:

The proof follows three main steps:

Step 1: Some preliminaries Before building the detection delay, we need to introduce three intermediate results.

The first result is to link the quantity $\Phi\left(\Sigma_{o,s:t}\right)$ to $\Phi\left(\widehat{\mu}_{o,s:t}\right)$ such that:

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

Using the notation: $\widehat{\boldsymbol{\mu}}_{a:b} = (\widehat{\mu}_{1,a:b}, ..., \widehat{\mu}_{O,a:b}) \quad \forall\, a < b$

Then, observe that :

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

Then, observe that:

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

Then, we will also need a useful notation as $f_{r,s,t}$:

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

Finally, following Lemma C.3, the control of the quantity $|\widehat{\mu}_{o,r:s-1} - \widehat{\mu}_{o,s:t}|$ takes the following form:

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

We define $\boldsymbol{\theta}^{(1)}$ and $\boldsymbol{\theta}^{(2)}$ to be the pre and post state-transition kernels over the set of actions O for change-point $c_{\ell+1}$ and Δ_o to be the per state variation $\Delta_o = \left|\theta_o^{(1)} - \theta_o^{(2)}\right|$. Then we write the relative gap $\Delta_{r,s,t}$ as follows

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

Pinsker inequality for multinomial distributions

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

Step 2: Building the sufficient conditions for detecting the change-point c_ℓ Let: $\boldsymbol{\theta}^{(1)} = \left(\theta_1^{(1)}, ..., \theta_O^{(1)}\right) \in [0,1]^O$ and $\boldsymbol{\theta}^{(2)} = \left(\theta_1^{(2)}, ..., \theta_O^{(2)}\right)$. First, assume that: $x_r, ..., x_{c_\ell - 1} \sim \text{Multi}\left(\theta_1^{(1)}, ..., \theta_O^{(1)}\right)$ and $x_{c_\ell}, ..., x_t \sim \text{Multi}\left(\theta_1^{(2)}, ..., \theta_O^{(2)}\right)$. Then, to build the detection delay, we need to prove that at some instant after c_ℓ the restart criterion $\boldsymbol{\mathit{\mathtt{Restart}}}\left(x_r, ..., x_t\right)$ is activated. In other words, we need to build the following guarantee:

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

Notice that:

$$\begin{aligned}
&\left\{\forall\, t > c_\ell : \boldsymbol{\texttt{Restart}}\left(x_r, ..., x_t\right) = 0\right\} \overset{(a)}{\Leftrightarrow} \left\{\forall\, t > c_\ell, \forall s \in (r,t] : \log \omega_{r,s,t} \leqslant \log \omega_{r,r,t}\right\}. \\
&\overset{(b)}{\Leftrightarrow} \left\{\forall\, t > c_\ell, \forall s \in (r,t] : \log \eta_{r,s,t} \leqslant \widehat{L}_{r:s-1} + \widehat{L}_{s:t} - \widehat{L}_{r:t}\right\}. \\
&\overset{(c)}{\Rightarrow} \Bigg\{\forall\, t > c_\ell, \forall s \in (r,t] : \log \eta_{r,s,t} \leqslant f_{r,s,t} - \sum_{o=1}^{O} \Phi\left(\Sigma_{o,r:s-1}\right) + \Phi\left(n_{r:s-1}\right) - \sum_{o=1}^{O} \Phi\left(\Sigma_{s:t}\right) + \Phi\left(n_{s:t}\right) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad + \sum_{o=1}^{O} \Phi\left(\Sigma_{r:t}\right) - \Phi\left(n_{r:t}\right) \Bigg\}. \\
&\overset{(d)}{\Rightarrow} \left\{\forall\, t > c_\ell, \forall s \in (r,t] : \log \eta_{r,s,t} \leqslant f_{r,s,t} - 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) \right\} \\
&\overset{(e)}{\Rightarrow} \Big\{\forall\, t > c_\ell, \forall s \in (r,t] : \log \eta_{r,s,t} \leqslant f_{r,s,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}) \\
&\qquad\qquad\qquad\qquad\qquad\qquad - 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})\Big\} \\
&\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} \left|\widehat{\mu}_{o,r:s-1} - \widehat{\mu}_{o,r:t}\right|\right)^2 - \frac{n_{s:t}}{2} \left(\sum_{o=1}^{O} \left|\widehat{\mu}_{o,s:t} - \widehat{\mu}_{o,r:t}\right|\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} \left(\widehat{\mu}_{o,r:s-1} - \widehat{\mu}_{o,r:t}\right)^2 - \frac{n_{s:t}}{2} \sum_{o=1}^{O} \left(\widehat{\mu}_{o,s:t} - \widehat{\mu}_{o,r:t}\right)^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}\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\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}} \left(\widehat{\mu}_{o,r:s-1} - \widehat{\mu}_{o,s:t}\right)^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}} \left(\widehat{\mu}_{o,r:s-1} - \widehat{\mu}_{o,s:t}\right)^2 \leqslant f_{r,s,t} - \log \eta_{r,s,t} \right\}
\end{aligned}$$

where:

- (a) holds thanks to the definition of the restart procedure in Equation (5).
- (b) holds thanks to the statement of Equation (4).
- (c) holds thanks to the upper bound in Equation (18) and
- (d) holds thanks to the statement of Equation (24).
- (e) holds thanks to the statement of Equation (25).
- (f) holds thanks to Equation (29).
- (g) holds thanks to the following equation: $\left(\sum_{o=1}^{O} \left|\widehat{\mu}_{o,s:t} - \widehat{\mu}_{o,r:t}\right|\right)^2 \geqslant \sum_{o=1}^{O} \left(\widehat{\mu}_{o,s:t} - \widehat{\mu}_{o,r:t}\right)^2$
- (h) holds thanks to Equation (26).

Thus we have:

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

Then, by using the probability operator, we obtain:

$$
\begin{aligned}
&\mathbb{P}_{\boldsymbol{\theta}^{(1)},\boldsymbol{\theta}^{(2)}}\Big\{\forall\, t > c_\ell : \textbf{\texttt{Restart}}\left(x_r, ..., x_t\right) = 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}\Big(\Delta_{o,r,s,t} - \mathcal{C}'_{r,s,t,\delta'}\Big)^2}}_{A_{r,s,t,\delta'}} \leqslant \frac{n_{r:s-1}}{n_{r:t}}\Big\}
\end{aligned}
$$

where:

- (k) holds thanks to the implication in Equation (30).
- (l) holds by using the property that: $\mathbb{P}\Big\{A\Big\} \leqslant \mathbb{P}\Big\{\neg B\Big\} + \mathbb{P}\Big\{A \cap B\Big\}$ where $B = \bigcap_{o\in\mathcal{O}} E^{(3)}_{o,r,t,\delta'}$.
- (m) holds thanks to the fact that: $\neg\Big\{\bigcap_{o\in\mathcal{O}} E^{(3)}_{o,r,t,\delta'}\Big\} = \bigcup_{o\in\mathcal{O}} \neg E^{(3)}_{o,r,t,\delta'}$.
- (n) holds thanks to the use of a union bound on the event $\bigcup_{o\in\mathcal{O}} \neg E^{(3)}_{o,r,t,\delta'}$.
- (o) holds using Equation (27).

Then, in order to derive the detection delay, some conditions on the $A_{r,s,t,\delta'}$ quantity should meet.

Conditions on $A_{r,s,t,\delta'}$ to derive the detection delay:

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

Main implication for detecting the change-point. Notice that:

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

Let $\delta = O\delta'$ and $\mathcal{C}_{r,s,t,\delta} = \mathcal{C}'_{r,s,t,\frac{\delta}{O}}$

Then, using the result of Equation (32), the change-point c_ℓ is detected at time t (with probability at least $1-\delta$) if for some $s \in (r,t]$, we have:

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

Let $\boldsymbol{\Delta} = (\Delta_1, ..., \Delta_O)$ denotes the vector of the change-point gap.

Step 3: Non-asymptotic expression of the detection delay $\mathfrak{D}_{\boldsymbol{\Delta},r,c_\ell}$ To build the detection delay, we need to ensure the existence of $s \in (r,t]$ such that Equation (33) is satisfied. In particular, Equation (33) can be satisfied for $s = c_\ell$. By this way, a condition to detect the change-point c_ℓ is written as follows

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

To build the delay, we should introduce the following variable: $d = t - c_\ell + 1 = n_{c_\ell:t} \in \mathbb{N}^\star$.

Thus from Equation (34), we obtain:

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

Finally, the change-point c_ℓ is detected (with a probability at least $1-\delta$) with a delay not exceeding $\mathfrak{D}_{\boldsymbol{\Delta},r,c_\ell}$, such that:

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

□

F R-BOCPD EQUIPPED UCRL2 ANALYSIS

We consider a formulation of the regret as defined in Section 3. Given the nature of the theoretical guarantees provided by the `R-BOCPD` algorithm, we adopt a decomposition with respect to the change points, allowing to analyze the *switching-MDP* problem $\mathbf{M} = \{\mathbb{S} = \{M_0, .., M_{K_T - 1}\}, \mathcal{C} = \{c_0, .., c_{K_T}\}\}$ as a sequence of stationary MDPs M_ℓ over time instances $t \in [c_\ell, c_{\ell+1})$. This can be formulated as follows:

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

where we denote by c_ℓ the time instance change ℓ happens, denote by $t = c_\ell$ the time instance starting at c_ℓ up to but not including $c_{\ell+1}$, i.e $t \in [c_\ell, c_\ell + 1)$, and define r_t to be the random reward `UCRL2` receives at time instant t, when starting at some initial state o. Again, this decomposition is only possible by using the independence of the sum of rewards/regrets in the stationary periods of the initial state as in Puterman (2014).

Now, denote by d_ℓ the detection delay achieved by `R-BOCPD` in a given interval $[c_\ell, c_{\ell+1})$. Hence, the natural decomposition into a stationary period $[c_\ell + d_\ell, c_{\ell+1})$ and a detection phase $[c_\ell, c_\ell + d_\ell)$.

F.1 DETECTION PHASE $[c_\ell, c_\ell + d_\ell)$

During the detection phase, we suppose the algorithm assumes the worst possible regret of 1, as r_t is sampled according to some unknown distribution in $[0, 1]$. To minimize the total regret, we rely on R-BOPCD's minimal detection delay, which we write as

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

with

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

which holds with probability at least $1 - \delta_{1_\ell}$ for $\delta_{1_\ell} \in (0, 1)$, for an input stream starting at some time instance $r < c_\ell$. This allows us to write the regret for $t \in [c_\ell, c_\ell + d_\ell)$ starting at some state o_{c_ℓ} as follows

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

where $c_{\ell-1} + d_{\ell-1}$ corresponds to the maximally delayed restart time after change-point $c_{\ell-1}$ with probability at least $1 - \delta_{1_{\ell-1}} - \delta_{2_{\ell-1}} + \delta_{1_{\ell-1}}\delta_{2_{\ell-1}}$, where $\delta_{1_{\ell-1}}$ corresponds to the probability of the `R-BOCPD` delay exceeding $\mathfrak{D}_{\mathbf{\Delta}_{\ell-1}, c_{\ell-2} + d_{\ell-2}, c_{\ell-1}}$ for change-point $c_{\ell-1}$ and $\delta_{2_{\ell-1}}$ corresponds to the worst-case false-alarm probability on the stationary period $[c_{\ell-1} + d_{\ell-1}, c_\ell)$.

F.2 Post-Detection Phase $[c_\ell + d_\ell, c_{\ell+1})$ and Episodic Regret

Relying on the assumptions about the change-point generating process in Section 3, we now analyze the regret in the phase $[c_\ell + d_\ell, c_{\ell+1})$ assuming `R-BOCPD-UCRL2` restarts `UCRL2` exactly at time $t = c_\ell + d_\ell$ with probability at least $1 - \delta_{1_\ell}$ for $\delta_{1_\ell} \in (0, 1)$. Now, to perform a similar analysis to that of Auer et al. (2008b), we need to ensure that no restarts/false-alarms happen during $[c_\ell + d_\ell, c_{\ell+1})$, i.e it is stationary. Given that `R-BOCPD` guarantees a bounded probability of false-alarm, we adopt a decomposition with regards to the event of restarting `UCRL2` during a stationary period. More precisely, we make use of the concentration characteristic of the stationary sum of rewards for `UCRL2` as in Auer et al. (2008b) along with the δ-bound guarantee in the `R-BOCPD` false-alarm probability.

To use the concentration argument for the sum of rewards when applying `UCRL2`, we note that at time instance t, reward r_{t+1} is only dependent on reward r_t (and filtration history $(o_1, a_1, r_1, ..., o_t, a_t, r_t)$ henceforth) through an exogenous process $\mathcal{E}$. Hence r_{t+1} and r_t are independent given $\mathcal{E}$ for all t, or $r_{t+1} \perp\!\!\!\perp r_t | \mathcal{E}$. This allows us to write, by virtue of Hoeffding's inequality, for $t \in [c_\ell + d_\ell, c_{\ell+1})$ and $\delta_\ell \in (0, 1)$

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

where (a) originates from the inequality $\mathbb{P}(A) \leqslant \mathbb{P}(A|B) + \mathbb{P}(\neg B)$, where $A = E_\ell^{(4)}$ and $B = \Big\{\forall t \in (c_\ell + d_\ell + 1, c_{\ell+1}) : \texttt{\textbf{\textit{Restart}}}\left(o_{c_\ell+d_\ell}, ..., o_t\right) = 0\Big\}$ and (b) originates from Hoeffding's inequality for vanilla `UCRL2` with $\bar{r}_\ell(o,a) = \frac{1}{N_\ell(o,a)} \sum_{t=c_\ell+d_\ell}^{c_{\ell+1}-1} r_t \cdot \mathbb{I}\{o_t = o, a_t = a\}$, in addition to `R-BOCPD`'s guarantee on false-alarm rate.

Thus, we can express the post-detection regret for the ℓ^{th} change-point starting at some state o_{c_ℓ} as follows

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

with probability at least $1 - \frac{\delta_\ell}{12(c_{\ell+1} - c_\ell - d_\ell)^{\frac{5}{4}}} - \delta_{2_\ell}$. As in Auer et al. (2008b), we adopt a decomposition over the number of episodes, which we denote by m_ℓ for change interval $[c_\ell + d_\ell, c_{\ell+1})$. Consequently, we can write $\sum_{k=1}^{m_\ell} \nu_k = N_\ell(o,a)$ and $\sum_{o,a} N_\ell(o,a) = c_{\ell+1} - c_\ell - d_\ell$, where $\nu_k(o,a)$ denotes the final counts of state-action pair (o,a) in episode k. Hence, defining $\mathfrak{R}_k\left(M_\ell, \texttt{R-BOCPD-UCRL2}, o_{c_\ell}, [c_\ell + d_\ell, c_{\ell+1})\right) := \sum_{o,a} \nu_k(o,a)\left(\rho^\star_{M_\ell} - \overline{r_\ell}(o,a)\right)$, we can write

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

with probability at least $1 - \frac{\delta_\ell}{12(c_{\ell+1}-c_\ell-d_\ell)^{\frac{5}{4}}} - \delta_{2_\ell}$.

Now, following the analysis of Auer et al. (2008b) for vanilla `UCRL2`, we can derive the final regret bound for `R-BOCPD-UCRL2` in post-detection period $[c_\ell + d_\ell, c_{\ell+1})$, for $c_{\ell+1} - c_\ell - d_\ell > 1$, as follows

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

which holds with probability at least $1 - \frac{\delta_\ell}{4(c_{\ell+1}-c_\ell-d_\ell)^{\frac{5}{4}}} - \delta_{2_\ell}$, where D_ℓ is the diameter of MDP M_ℓ as defined in 3. Also note that $d_0 = \mathfrak{D}_{\boldsymbol{\Delta}_{\cdot,\cdot,c_0}} = 0$ as time instance c_0 defines the start of learning.

F.3 Total Regret Bound

Wrapping up the last two steps and summing over the change periods, we can write

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

with probability at least $1 - \sum_{\ell=0}^{K_T-1}\left(\frac{\delta_\ell}{4(c_{\ell+1}-c_\ell-d_\ell)^{\frac{5}{4}}} + \delta_{2_\ell}\right) - \sum_{\ell=1}^{K_T}\left(\delta_{1_{\ell-1}} + \delta_{2_{\ell-1}} - \delta_{1_{\ell-1}}\delta_{2_{\ell-1}}\right)$. Now we conclude the proof by providing a bound for the latter probability. Without loss of generality, we fix our confidence parameters as $\delta_\ell = \delta_{1_\ell} = \delta_{2_\ell} := \frac{\delta}{8K_T}, \forall \ell$. Hence, we can write:

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

Now, defining $D := \max_{\ell} D_\ell$, deriving the corresponding regret boils down to

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

which holds with probability at least $1-\delta$. This completes the proof for Theorem 4.12.

F.4 Asymptotic Regret Bound

Now building up on the previous section and the asymptotic detection delay in Equation (10), we write the asymptotic regret bounds as follows

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

which again holds with probability at least $1-\delta$, assuming a false-alarm rate of 0. This completes the derivation of Corollary 4.13.

G Experimental Setup & Discussion

Given the generality of R-BOCPD-UCRL2 with respect to variations in the state transition distributions and rewards, a suitable environment to benchmark its performance vis-à-vis state-of-the-art is a synthetic environment where abrupt changes occur to the state transition distributions and rewards at unknown time instances. First, the sizes of the state and action spaces is chosen randomly. Then, the state transition probabilities are sampled from a multinomial distribution over the set of states O and the rewards are sampled randomly from $[0, 1]$, where we can control the variation in the generation process to be able to simulate both large changes to state-transition distributions and rewards and relatively small ones. Now, fixing a set of change-points, chosen with sufficient time difference in-between successive ones, the process is repeated after each change-point over a time horizon $T = 50000$. We consider 100 realizations of each state-action pair and are interested in the average cumulative rewards of that.

G.1 Choice of Hyperparameters

Now, we list our hyperparameter choice for the sliding-window based algorithms in Section 5.

- **Sliding-Window UCRL2** (`SWUCRL2`, Gajane et al. (2018)): The window size is chosen optimally as in Gajane et al. (2018), $W^\star = \left(\frac{16.53}{K_T} TDO\sqrt{A\log\frac{T}{\delta}}\right)^{\frac{2}{3}}$. The diameter D is estimated based on a hyperparameter search over a large range of suitable values for each combination of state space and action space sizes. The diameter maximizing the cumulative rewards was chosen.
- **Sliding-Window UCRL2 with Confidence Widening** (`SWUCRL2-CW`, Cheung et al. (2020)): Again, the window size and widening factor are chosen optimally according to Cheung et al. (2020), $W^\star := 3O^{\frac{2}{3}}A^{\frac{1}{2}}T^{\frac{1}{2}}/(B_r + B_p + 1)^{\frac{1}{2}}$ and $\eta^\star := \sqrt{(B_p + 1)W^\star/T}$ respectively. Here, B_p and B_r are computed beforehand in a total-variation sense. While, in a realistic RL setting, T is also unknown beforehand, we still choose our choice of T to initialize `SWUCRL2`.

G.2 Discussion

Now, in addition to the rather general performance evaluation provided in Section 5, we highlight in more detail how each algorithm operates as follows

- **Sliding-Window UCRL2** (`SWUCRL2`, Gajane et al. (2018)): An essential parameter of the sliding-window approach is the diameter D_ℓ of each MDP M_ℓ defining the switching-MDP problem $\mathbf{M}$, which is used to quantify the difficulty of learning in the setting specified by M_ℓ. Diameter D_ℓ, as defined in Section 3, is a parameter that cannot be accessed directly from the MDP and yet is key for `SWUCRL2` to perform as claimed. Even while considering $D = \max_\ell D_\ell$, `SWUCRL2` still relies on a hyperparameter search to estimate the overall optimal diameter D, which is quite restrictive in practice. We also highlight that `SWUCRL2` performs poorly for a suboptimal choice of D. Now, considering the optimal window-size choice $W^\star$ is of $\tilde{\mathcal{O}}(A^{\frac{1}{3}}O^{\frac{2}{3}}D^{\frac{2}{3}}T^{\frac{2}{3}}{K_T}^{-\frac{1}{2}})$, `SWUCRL2` requires to keep track of a significant number of observations even for rather small MDPs.
- **Sliding-Window UCRL2 with Confidence Widening** (`SWUCRL2-CW`, Cheung et al. (2020)): While not relying on an agnostically chosen parameter as for `SWUCRL2`, (`SWUCRL2-CW` still relies on the knowledge of predefined variation budgets for the state-transition distributions and rewards, which are unknown in a realistic setting. Its variant Bandit-over-Reinforcement Learning (`BORL`), which operates without assuming the knowledge of B_r and B_p, performs than `SWUCRL2-CW` in practice.
- **Restarted UCRL2** (`Restarted-UCRL2`, Auer et al. (2008a)): While comparing favorably to sliding-window approaches, it requires a very large number of restarts $T^{\frac{1}{3}}{K_T}^{\frac{2}{3}}$, which is quite prohibitive for large-sizes problems. In addition, since restarting frequency decreases polynomially with time (for a fixed number of changes), performance will inevitably degrade for considerably long time horizons with changes occuring frequently at the later stages of learning.
- **Vanilla UCRL2** (`UCRL2`, Auer et al. (2008a)): Here rather considered as a baseline.

G.2.1 Extension to realistic environments

We also highlight our interest in realistic MDP settings such as *RiverSwim* (Strehl & Littman (2008)), *MachineReplacement* and *GridWorld* among others. Here, the difficulty of simulating realistic changes to the environment is that we don't have access to the underlying state-transition distributions and rewards. While it is possible to alter some of the environment parameters that are typically chosen at random, it is unclear to us at the moment of writing this manuscript how to control the process of changing these variables and relate it to the rather latent changes in the underlying state-transition distributions and rewards.